

AfricaNLP 2026

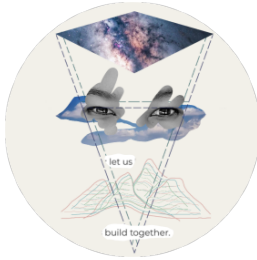
**7th Workshop on African Natural Language Processing
(AfricaNLP 2026)**

Proceedings of the Workshop

March 28, 2026

The AfricaNLP organizers gratefully acknowledge the support from the following sponsors.

Niger-Congo Tier



Nilo-Saharan Tier



©2026 Association for Computational Linguistics

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-364-7

Introduction

We are pleased to present the proceedings of the 7th Workshop on African Natural Language Processing (AfricaNLP 2026), held in Rabat, Morocco on March 28, 2026, as part of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026).

This year’s theme, “Multilingual Multimodal LLMs,” reflects the growing importance of developing language models that can process and understand African languages across multiple modalities. The workshop continues our mission to advance natural language processing research for African languages, bringing together researchers, practitioners, and stakeholders from across the continent and beyond.

We received 56 submissions spanning topics such as machine translation, speech recognition, language modeling, multimodal AI, and culturally grounded NLP for African languages. Of these, 30 papers were accepted as archival contributions and appear in these proceedings, while 14 papers were presented as non-archival contributions at the workshop. Together, the accepted papers reflect research across diverse African languages and highlight the continued growth and vibrancy of the African NLP research community.

We thank all authors for their contributions, our program committee for their thorough reviews, and our sponsors, the Masakhane Research Foundation and Microsoft, for their generous support. We also thank the EACL 2026 organizers for hosting our workshop.

With gratitude,
The AfricaNLP 2026 Organizing Committee

Organizing Committee

General Chair

Shamsuddeen Hassan Muhammad, Imperial College London

Program Chairs

Constantine Lignos, Brandeis University
Shamsuddeen Hassan Muhammad, Imperial College London
David Ifeoluwa Adelani, McGill University and Mila

Publication Chairs

Everlyn Asiko Chimoto, University of Cape Town and Lelapa AI
Idris Abdulmumin, University of Pretoria
Clemencia Siro, Centrum Wiskunde & Informatica

Sponsorship Chair

Clemencia Siro, Centrum Wiskunde & Informatica

Mentoring Chair

Sang Yun Kwon, The University of British Columbia

Communications Chairs

Millicent Ochieng, Microsoft Research Lab - Africa
David Ifeoluwa Adelani, McGill University and Mila

Publicity Chair

Jessica Ojo, Mila & McGill University

Program Committee

Reviewers

Idris Abdulmumin, Henok Biadgign Ademtew, Ibrahim Said Ahmad, Felermينو D. M. A. Ali, Lukman Jibril Aliyu, Anietie Andy, Berk Atul, Abinew Ali Ayele

Tadesse Destaw Belay, Meriem Beloucif, Emmanuel Bolarinwa, Jan Buys, Happy Buzaaba

Rendi Chevi, Everlyn Asiko Chimoto, Chiamaka Ijeoma Chukwuneke

Daryna Dementieva, Emmanuel Dorley, Bonaventure F. P. Dossou

Chris Chinenye Emezue

David Guzmán, Tajuddeen Gwadabe

Khaldi Hadjer

Oana Ignat, Amina Abubakar Imam, Sukairaj Hafiz Imam, Ahmad Ibrahim Ismail, Sheriff Issaka

Adejumobi Monjolaoluwa Joshua

Salomon Kabongo Kabenamualu, Sulaiman Kagumire, Andrew Kiprop Kipkebut, Alfred Malengo Kondoro, Sujay S Kumar, Sang Yun Kwon

Falalu Ibrahim Lawan, Eric Le Ferrand, En-Shiun Annie Lee, Senyu Li, Weiran Lin

Rahmad Mahendra, Marek Masiak, Dunstan Matekenya, Francois Meyer, Kausar Yetunde Mo-shood, Anjishnu Mukherjee

Quang Phuoc Nguyen, Gebregziabihier Nigusie

Jacki O'Neill, Millicent Ochieng, Perez Ogayo, Odunayo Ogundepo, Jessica Ojo, Ugochi Okafor, Chibuzor Okocha, Akintunde Oladipo, Flora Oladipupo, Abigail Oppong, Salomey Osei, Abraham Toluwase Owodunni

Chester Palen-Michel, Ted Pedersen, Van-Thuy Phi

Samuel Rutunda

Elizabeth Salesky, Avinash Kumar Sharma, Kathleen Siminyu, Clemencia Siro, Rui Sousa-Silva, Nirmal Surange, Jonne Sälevä

Allahsera Auguste Tapo, Atnafu Lambebo Tonja

Kosei Uemura

Eric Peter Wairagala

Debela Desalegn Yadeta, Kweku Andoh Yamoah, Seid Muhie Yimam

Miaoran Zhang

Tolúlopé Ògúnrèní

Invited Talk
**Data-Efficient Language Modelling for Low-Resource
Languages**

Francois Meyer
University of Cape Town

Abstract: Progress in language modelling has been driven by scaling data and model size, but this approach is infeasible for most African languages. In this talk, I will present our work on developing data-efficient language models – architectures and training algorithms that improve performance on limited training data. I will present examples of how linguistically informed modelling, which targets and leverages the linguistic properties of specific languages, can improve sample efficiency. Finally, I will discuss the emerging intersection between low-resource NLP and developmentally inspired NLP, exploring how insights from human language learning can help us build more efficient models.

Bio: Francois Meyer is a Lecturer in the Computer Science Department at the University of Cape Town and co-investigator in the UCT NLP research group. His research is on data-efficient language modelling and linguistically informed subword tokenisation. He completed his PhD at the University of Cape Town and previously obtained a masters in AI at the University of Amsterdam.

Invited Talk

The Emergence of Multilingual Representations: Tracing Linguistic Capabilities During Language Model Pretraining

Barbara Plank

LMU Munich

Abstract: Multilingual large language models exhibit remarkable zero-shot and cross-lingual transfer capabilities. However, most analyses focus on fully trained models, leaving limited understanding of how and when different types of linguistic information emerge, interact, and align within multilingual representation spaces during training.

In this talk, I present a series of studies investigating the training dynamics of linguistic knowledge in language models, tracing how linguistic structure and cross-lingual alignment develop over time. Studying these dynamics requires access to intermediate checkpoints, which are only available to a limited extent. Nevertheless, analyzing emerging representations opens up new avenues for diagnosing and improving multilingual LLMs. Understanding how alignment forms during pretraining is particularly important for models intended to support underrepresented and low-resource languages, where effective transfer and shared representations are crucial for performance.

Bio: Barbara Plank is Full Professor and Chair for AI and Computational Linguistics at LMU Munich, Co-director of the Center for Information and Language Processing and Head of the MaiNLP (Munich AI and NLP) lab at LMU. Barbara Plank is an ELLIS Fellow (European Laboratory for Learning and Intelligent Systems) and regularly serves in international organizations and on scientific advisory committees.

Invited Talk
**Towards Multimodal AI for African Languages and Cultures:
Lessons from Afri-MCQA**

Atnafu Lambebo Tonja
MBZUAI

Abstract: What will it take to develop multimodal AI that truly comprehends African languages and cultures? In this talk, I explore this question through lessons from Afri-MCQA, a benchmark covering 15 African languages across 12 countries. Our evaluation highlighted that current models face major challenges, such as 1) they are unable to process speech in African languages, 2) they lack cultural context, and 3) they struggle to generate culturally relevant responses, rather than merely recognizing them. I will share these insights and outline a pathway forward, emphasizing the importance of speech-first development, culturally grounded training, and cross-lingual knowledge transfer as critical steps in creating effective multimodal AI for Africa.

Bio: Atnafu Lambebo Tonja is a Postdoctoral Researcher at the Mohamed bin Zayed University of Artificial Intelligence (MBZUAI) in the UAE, where he leads projects on culturally-diverse multilingual visual question answering and multimodal machine translation. He earned his PhD in Computer Science from Instituto Politécnico Nacional in Mexico City, focusing on neural machine translation for low-resource languages. His research focuses on advancing NLP for underrepresented languages, particularly African and Ethiopian languages, through the development of multilingual language models and sustainable data curation frameworks. His work on low-resource languages, especially for African and Ethiopian languages, has been published in top NLP venues including ACL, EMNLP, and NAACL.

Invited Talk
**Beyond Parallel Data: Harnessing External Knowledge for
Low-Resource MT**

Felermino Ali
Porto University

Abstract: Translating from high-resource languages into Mozambican languages remains a pressing challenge in African NLP. The scarcity of parallel corpora, orthographic variation across dialects, and the frequent presence of loanwords and code-switching complicate the task of building robust translation systems. In this talk, I will share how we address these barriers through lexicon-guided neural machine translation. By integrating bilingual dictionaries and systematic loanword mappings directly into the training process, we move beyond data scarcity toward structured lexical enrichment. Our approach leverages dictionary entries and loanword mappings to construct sentence-specific glossaries, dynamically incorporated via input augmentation. On FLORES benchmarks, this method demonstrates clear gains: stronger lexical coverage, fewer inconsistencies, and translations that better capture contextual nuance. Beyond the technical improvements, this work points to a broader vision: advancing low-resource machine translation not only by scaling data but by intelligently bridging vocabulary gaps with structured linguistic knowledge. For Mozambican languages, this means opening pathways to more inclusive digital communication, empowering communities, and ensuring that the linguistic richness of African languages is represented in the global NLP landscape.

Bio: Felermino Ali is a researcher at MSR Africa and a PhD Candidate at the University of Porto in Portugal, focused on natural language processing (NLP) with a specialization in low-resource African languages. His work centers on building neural machine translation systems for low-resource languages and advancing methods to more effectively evaluate MT performance in low-resource settings.

Invited Talk

The knowns and unknowns of multilingual data augmentation

Julia Kreutzer

Cohere Labs

Abstract: In this talk I will present recipes for multilingual fine-tuning data augmentation that have been developed to overcome data scarcity in languages beyond English. We will then discuss what the limitations of these approaches are, and what directions are relevant for future research.

Bio: Julia Kreutzer is a Senior Research Scientist at Cohere Labs, where she focuses on research around multilingual large language models. She has a background in machine translation, with a PhD from Heidelberg University and prior work experience at Google Translate. She's passionate about advancing NLP technologies for underrepresented languages and has been part of multiple open science initiatives to work towards this goal collaboratively.

Table of Contents

<i>Dealing with the Hard Facts of Low-Resource African NLP</i> Michael Leventhal, Yacouba Diarra, Nouhoum Coulibaly, Panga Azazia Kamaté, Aymane Dem- bélé, Madani Amadou Tall and Emmanuel Elise Kone	1
<i>M-MiniGPT4: Multilingual VLLM Alignment via Translated Data</i> Seung Hun Eddie Han, Youssef Mohamed and Mohamed Elhoseiny	11
<i>InstructLR: A Scalable Approach to Create Instruction Dataset for Under-Resourced Languages</i> Mamadou K. Keita, Sebastien Diarra, Christopher M Homan and Seydou Diallo	17
<i>Leveraging CoHere Multilingual Embeddings and Inverted Softmax Retrieval for Automatic Parallel Sentence Alignment in Low-Resource Languages</i> Abubakar Auwal Khalid, Salisu Musa Borodo and Amina Abubakar Imam	37
<i>AfriCaption: Establishing a New Paradigm for Image Captioning in African Languages</i> Mardiyyah Oduwole, Prince Mireku, Fatimo Adebajo, Oluwatosin Olajide, Mahi Aminu Aliyu and Jekaterina Novikova	44
<i>Developing an English–Efik Corpus and Machine Translation System for Digitization Inclusion</i> Offiong Bassey Edet, Mbuotidem Awak, Emmanuel Ubene Oyo-Ita, Benjamin Okon Nyong and Ita Etim Bassey	56
<i>Reasoning Beyond Labels: Measuring LLM Sentiment in Low-Resource, Culturally Nuanced Contexts</i> Millicent Ochieng, Anja Thieme, Ignatius Ezeani, Risa Ueno, Samuel Chege Maina, Keshet Ron- nen, Javier Gonzalez and Jacki O’Neill	64
<i>ÒWE-Voice: An Evaluation of Monolingual and Multilingual ASR Model Using Yoruba Proverb Speech Dataset</i> Daud Abolade	82
<i>Language Choice in Nigerian Social Media Hate Speech</i> Nneoma C Udeze and Rob Voigt	88
<i>The Token Tax: Systematic Bias in Multilingual Tokenization</i> Jessica M. Lundin, Ada Zhang, Nihal Karim, Hamza Louzan, Guohao Wei, David Ifeoluwa Ade- lani and Cody Carroll	103
<i>EduNaija AI Tutor: A Multi-Agent Retrieval-Augmented Generation System for Nigerian Curriculum Education</i> Israel Olanrewaju Odeajo and Edifon Emmanuel Jimmy	113
<i>Synthetic Data Generation Pipeline for Low-Resource Swahili Sentiment Analysis: Multi-LLM Judging with Human Validation</i> Samuel Gyamfi, Alfred Malengo Kondoro, Yankı Öztürk, Richard Hans Schreiber and Vadim Borisov	116
<i>Advancing African NLP: UDMorph and flexiPipe</i> Maarten Janssen	142
<i>Linguistically Informed Evaluation of Multilingual ASR for African Languages</i> Fei-Yueh Chen, Lateef Adeleke and C. M. Downey	149

<i>Evaluating Native-Speaker Preferences on Machine Translation and Post-Edits for Five African Languages</i>	
Hiba El Oirghi, Tajuddeen Gwadabe and Marine Carpuat	163
<i>Building a Conversational AI Assistant for African Travel Services with LLMs and RAG</i>	
Grace Kevine Ngoufo, Shamsuddeen Hassan Muhammad and Kevin Jeff Fogang Fokoa	171
<i>Morphologically-informed Somali Lemmatization Corpus built with a Web-based Crowdsourcing Platform</i>	
Abdifatah Ahmed Gedi, Shafie Abdi Mohamed, Yusuf A. Yusuf, Muhidin A. Mohamed, Fuad Mire Hassan and Houssein A Assowe	179
<i>Kunnafonidilaw ka Cadeau: an ASR dataset of present-day Bambara</i>	
Michael Leventhal, Yacouba Diarra, Nouhoum Coulibaly and Panga Azazia Kamaté	190
<i>Full Fine-Tuning vs. Parameter-Efficient Adaptation for Low-Resource African ASR: A Controlled Study with Whisper-Small</i>	
Sukairaj Hafiz Imam, Muhammad Yahuza Bello, Hadiza Ali Umar, Tadesse Destaw Belay, Idris Abdulmumin, Seid Muhie Yimam and Shamsuddeen Hassan Muhammad	197
<i>Real-Time Spoken Instruction Following and Translation in Ugandan Languages</i>	
Benjamin Akera, Tim Wenjie Hu, Patrick Walukagga, Evelyn Nafula Ouma, Yiga Gilbert, Ernest Tonny Mwebaze and John Quinn	204
<i>SALT-31: A Machine Translation Benchmark Dataset for 31 Ugandan Languages</i>	
Solomon Nsumba, Benjamin Akera, Evelyn Nafula Ouma, Medadi E. Ssentanda, Deo Kawalya, Engineer Bainomugisha, Ernest Tonny Mwebaze and John Quinn	211
<i>Sample-Size Scaling of the African Languages NLI Evaluation</i>	
Anuj Tiwari, Oluwapelumi Ogunremu, Terry Oko-odion, Jesujuwon Egbewale and Hannah Sopuruchi Nwokocha	217
<i>Evaluating Yoruba Text-to-Speech Systems for Accessible Computer-Based Testing in Visually Impaired Learners</i>	
Kausar Yetunde Moshood, Victor Tolulope Olufemi, Oreoluwa Boluwatife Babatunde, Emmanuel Bolarinwa and Williams Oluwademilade	228
<i>Power Asymmetries, Bias, and AI, a Reflection of Society on Low-Resourced Languages - African Languages as Case Study</i>	
Simbiat Ajao	235
<i>Sudanese-Flores: Extending FLORES+ to Sudanese Arabic Dialect</i>	
Hadia Mohmmedosman Ahmed Samil and David Ifeoluwa Adelani	243
<i>Where Are We at with Automatic Speech Recognition for the Bambara Language?</i>	
Seydou Diallo, Yacouba Diarra, Panga Azazia Kamaté, Aboubacar Ouattara, Mamadou K. Keita and Adam Bouno Kampo	248
<i>Enhancing Automatic Speech Recognition Models for Maternal and Reproductive Health: Fine-Tuning and Real-World Evaluation in Wolof</i>	
Ertony Basilwango, Yann Le Beux, Oche David Ankeli and Pierre Herve Berdys	256
<i>Eyaa-Tom 26, Yodi-Mantissa and Lom Bench: A Community Benchmark for TTS in Local Languages</i>	
Bakoubolo Essowe Justin, Catherine Nana Nyaah Essuman, Messan Agbobli, Ahoefa Kansiwier, Eli Jean Doumeyan, Julie Pato, Notou Your Timibe, Emile KOGBEDJI Agossou and Guedela Bakouya	264

<i>Using Subword-Embeddings for Bilingual Lexicon Induction in Bantu Languages</i>	
Adrian Breiding and Alan Akbik	271
<i>AfriNLLB: Efficient Translation Models for African Languages</i>	
Yasmin Moslem, Aman Kassahun Wassie and Amanuel Gizachew Abebe.....	281

Dealing with the Hard Facts of Low-Resource African NLP

Yacouba Diarra, Nouhoum Souleymane Coulibaly, Panga Azazia Kamaté,
Madani Amadou Tall, Emmanuel Élisé Koné, Aymane Dembélé, Michael Leventhal
RobotsMali AI4D Lab, Bamako, Mali; Correspondence: research@robotsmali.org

Abstract

Creating speech datasets, models, and evaluation frameworks for low-resource languages remains challenging given the lack of a broad base of pertinent experience to draw from. This paper reports on the field collection of 612 hours of spontaneous speech in Bambara, a low-resource West African language; the semi-automated annotation of that dataset with transcriptions; the creation of several monolingual ultra-compact and small models using the dataset; and the automatic and human evaluation of their output. We offer practical suggestions for data collection protocols, annotation, and model design, as well as evidence for the importance of performing human evaluation. In addition to the main dataset, multiple evaluation datasets, models, and code are made publicly available.

1 Introduction

End-to-end ASR (E2E-ASR) systems for languages with large amounts of text data, especially English, have achieved human-level performance on several benchmarks (Xiong et al., 2016). In contrast, training E2E-ASR for low-resource languages remains challenging due to the considerable amounts of labeled data and computational resources required by modern deep learning architectures (Kaplan et al., 2020).

Until recently, African language aligned data for ASR existed only as a minute portion of large multilingual datasets, often primarily for benchmarking purposes (Ardila et al., 2020; Goyal et al., 2022), rather than for training models intended to be deployed in systems that recognize real-world speech. No open ASR model exists for the vast majority of the 2000+ languages on the continent.

In pursuit of what they call *omnilingualism*, Meta released the Massively Multilingual Speech (Pratap et al., 2023) and Omnilingual ASR (Omnilingual ASR Team et al., 2025) model suites

in 2023 and 2025. Their approach used massive self-supervised learning (Baeovski et al., 2020) and finetuning on small labeled datasets, consisting, principally, of publicly available readings of religious texts in the 2023 release, and data obtained from community-centered crowdsourced data in the 2025 release.¹ This project provided some level of ASR capability for many African and non-African languages for the first time. While a positive development, there is less than 50 hours of data for many of those languages, some with less than 10 hours, a very small fraction of the 120,710 hours on which the supervised finetuning (SFT) models were trained. The underlying self-supervised encoder (a 7B-parameter wav2vec 2.0 model) was trained on approximately 4.3M hours of unlabeled audio (Omnilingual ASR Team et al., 2025).

For Bambara, a Manding language spoken in several West African countries (primarily in Mali), with more than 15 million L1 and L2 speakers and mutual intelligibility with Malinke, Dioula, and Mandinka, which are spoken by an additional 25 million people, the development of speech recognition technology could affect a population of roughly 40 million (Eberhard et al., 2023). However, as a *low-literacy, predominantly oral language*, Bambara transcription is a hard problem: few speakers can write it, and even those who can lack the facility to do so quickly and easily (Diarra et al., 2025b).

The CMU Wilderness Multilingual Speech Dataset, a dataset of aligned sentences and audio for some 700 languages based on readings of the New Testament, is, to the best of our knowledge, the first mention of Bambara in speech corpora

¹Meta released part of their labeled corpus openly, offering spontaneous speech recordings and their transcriptions for 348 under-served languages, along with training script configuration and docs: <https://github.com/facebookresearch/omnilingual-asr/>

prepared specifically to train speech synthesis models (Black, 2019). The dataset was never released as an open resource. Jeli-ASR, a corpus of 30 hours of griot narrations with their transcriptions and French translations, has so far been the only open ASR dataset for Bambara (Diarra et al., 2022). Since its release in 2022, Jeli-ASR has given rise to derivative datasets and has supported the development of the first openly released ASR models for Bambara on Hugging Face².

The African Next Voices project (ANV), undertaken by a network of African universities and organizations, recently released what is thought to be the largest dataset of African languages for AI so far (Marivate et al., 2025), with more still to be published. The project aims to record and transcribe over 9,000 hours of speech in 18 languages across South Africa, Kenya, Nigeria, and Mali (after Bambara completed the list as a later addition).

In this paper, we present the Bambara portion of this initiative for which we have collected and annotated 612 hours of spontaneous Bambara speech collected across the southern part of Mali. We share statistics and metadata about the dataset, the collection process and the results of our ASR experiments with models finetuned and tested on a subset of 101 hours.

2 Data Collection and Annotation

In the audio recording phase, we followed an approach similar to Emezue et al., using *facilitators*: individuals with knowledge of the language—L1 or L2 speakers—whom we trained in data collection guidelines and in the use of our mobile data collection app.³ The guidelines covered requirements for the recording environment, quality checks for background noise, and management of participants’ contributions with respect to voice quality, staying on topic, and minimizing code-switching to French. Pronunciation, often an issue when recordings are based on read speech, was rarely a concern here, as we recorded spontaneous speech on familiar topics from L1/L2 contributors. In total, we collected 626.32 hours of audio and processed 612 hours to create the dataset.

²oza75 released a finetuned Whisper model in early 2024 (later taken down), followed by several releases by RobotsMali in early 2025

³We also open-source this app, a minimalist Flutter-based tool designed with a simple user interface to minimize user training time: <https://github.com/RobotsMali-AI/Africa-Voice-App>

The raw recordings were then segmented using Silero VAD’s open voice activity detection model (Silero Team, 2024), retaining on average 70% of the original duration and yielding 423 hours of speech chunks ranging from 240 milliseconds to 30 seconds. This step also removed long silences and inaudible speech from the recordings, increasing the amount of usable speech for the transcription pipeline and eliminating manual segmentation (Li et al., 2019). Almost all segments are mono-speaker, although a small number of recordings include brief facilitator speech; overlaps are rare. The segments were first pre-transcribed with RobotsMali/soloni-114m-tdt-ctc-v0, and human transcribers were tasked with reviewing and correcting these model-generated transcriptions rather than transcribing from scratch. We then finetuned RobotsMali/soloni-114m-tdt-ctc-v2 on 98 hours of human-corrected transcripts and re-transcribed the segments with this model. WER and CER metrics were used to compare the two sets of transcriptions, and human review and correction continued on the new model outputs for a period of time to assess their impact on the annotation process (Section 5.2). Further details on transcription guidelines and the labeling interface are provided in Appendices A and B.

3 The ANV Bambara Dataset

The African Next Voices Dataset is the largest open Bambara ASR corpus collected to date. Comprised of natural, spontaneous speech, often from people with great knowledge in domains such as health, agriculture, the food industry, art, and more, we have captured a profound and authentic snapshot of the Malian society and culture in relatively pure Bambara, with the habitual code-switching to French minimized. The dataset collection was also designed to limit the variance of accents and speaking styles, focusing on the southern regions of Mali, relatively close to the capital. While a truly representative dataset would include code-switching and regional accent variation, the objective fixed for ANV was to gather homogeneous data that would simplify training and provide a baseline for a group of low-resource languages with limited NLP support.

3.1 Profile of the Dataset

The dataset features 512 unique speakers from Bamako, the capital, and four localities between

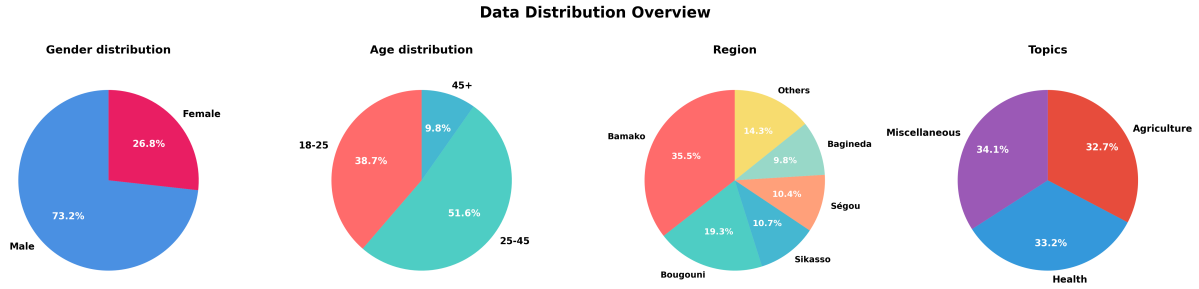


Figure 1: Statistics overview charts of the African Next Voices Bambara dataset: Age, Gender, Region and topics distribution. The first three charts are calculated with respect to the number of speakers while the topics distributions are expressed in durations. The locations represented as 'others' refer to rural areas/villages around the 5 other main regions

25 and 300 km from Bamako: Ségou, Sikasso, Bagineda, and Bougouni. Figure 1 presents an overview of the gender, age, region, and subject-matter distributions in the dataset. The “Miscellaneous” category encompasses topics ranging from education to social norms and beliefs, history, industry, art, and fashion. The speaker distribution is not gender-balanced, reflecting cultural and security constraints at the time of collection.

The average segment duration is ≈ 2 seconds. Chunks shorter than one second were not uploaded in the annotation pipeline as they often consist of formulaic expressions, discourse markers and set responses of 1 to 3 words such as *nse*, *nba* (female or male response to a salutation), *ayiwa* (a term used to express agreement or closure) or *nka* (but). These short segments are transcribed accurately by the models at much lower error rates than longer, novel utterances (Tall, 2025). They are a significant percentage of the total dataset such that including them would skew WER/CER measurements while not contributing to ASR performance. The Hugging Face dataset is divided into 3 subsets totaling 874,762 utterances totaling approximately 423 hours. Each subset contains the audio segments and two sets of accompanying transcriptions labeled either **v1**, created by soloni-v0, trained mainly on Jeli-asr, and **v2** created by soloni-v2, finetuned from soloni-v0 using 98 hours of segments and human-corrected transcriptions collected over the course of the project.

- **The 'human-corrected' subset:** A 159 hour subset (260,008 utterances) with human reviewed, corrected and validated transcriptions. This subset is the only one with a 'text' attribute containing a transcription that has gone through human review and correction.

- **The 'model-annotated' subset:** A 212 hour subset (355,571 utterances) that has model-generated transcriptions that have not been reviewed by humans. This subset has only **v1** and **v2** labels corresponding to the model used to generate transcriptions.
- **The 'short' subset:** A 52 hour subset (259,183 utterances) of duration less than one second that we have filtered out from the pool of segments to be annotated. Those short utterances are model-annotated and have **v1** and **v2** labels.

We have also released the original 612 hours dataset comprised of 1777 raw recordings ranging from 8 seconds to 1.48 hours, with all the associated segment timestamps, the anonymized metadata, the SNR quality check results and all the preprocessing code⁴.

3.2 Signal-to-Noise Ratio as a proxy to Audio Quality

We use a Voice Activity Detection (VAD) based method to estimate SNR. VAD output is used to separate the signal into two distinct regions: speech activity ($\text{vad}[n] = 1$) and voice-inactive ($\text{vad}[n] = 0$). We use VAD to estimate the speech and noise power instead of a histogram-based approach such as the standard NIST SNR method. In our setup, SNR is defined, using VAD, as the ratio between the estimated speech power (from speech activity regions) and the estimated noise power (typically the average power in silence regions) from the same recording (Vondrasek and Pollák, 2005;

⁴We have made the recordings and metadata available through Google Cloud Storage. The link will be found in the GitHub repository holding the code: <https://github.com/RobotsMali-AI/afvoices>

Silero Team, 2024.) Figure 2 shows the distribution of SNR values from the unsegmented recordings in the dataset.

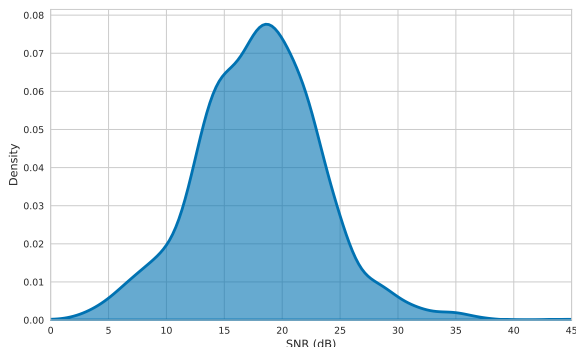


Figure 2: Density Distribution of Signal-to-Noise Ratio values in the African Next Voices Bambara Dataset. Note that the SNR values are not bounded.

The classical SNR definition yields a different value distribution than VAD-based estimates, so we follow Xu et al., who report VAD-based SNRs for several noise types and levels at 5, 10, and 20 dB and treat 30 dB as clean speech; Table 1 groups our recordings into five bands using thresholds derived from their datapoints. We prefer a VAD-based SNR here because ASR performance depends on the quality of speech frames rather than the entire waveform, including silence. Estimating signal power over VAD-identified speech regions and noise power over non-speech regions yields an SNR measure that better reflects recognition difficulty in our small corpus and allows us to retain and prioritize high-quality speech segments from otherwise noisy files.

SNR Category	Threshold (dB)	Recordings
Very Low SNR	< 0	1
Low SNR	[0, 5)	15
Medium SNR	[5, 15)	486
High SNR	[15, 25)	1135
Very High SNR	≥ 25	140
Total Audios		1777

Table 1: Distribution of Audio Recordings by Signal-to-Noise Ratio (SNR) Category.

71.75% of the recordings fall into the 'High SNR' to 'Very High SNR' categories; this indicates that the dataset consists of relatively clean audio recordings.

4 ASR Experiments

We performed experiments with a subset of our human-corrected transcribed segments to explore the potential of the ANV Bambara dataset for monolingual ASR modeling. We finetuned the models from our earlier experiments with Jeli-asr —themselves finetuned from different models of NVIDIA’s Parakeet family of English-trained ASR models and on QuartzNet— and evaluate all the models on both the test set of our experiment (Afvoices Test) and a smaller, more heterogeneous benchmark (Nyana Eval) that we also introduce in this paper. We also report on human evaluation by native speakers, comparing results obtained from the latest finetuned models and their predecessors on the Nyana Eval benchmark (section 5.2).

Experimental setup: For our experiments, we finetuned open-source models based on NVIDIA’s Parakeet family and QuartzNet.

We finetuned a 114M- and two 600M-parameter Parakeet models⁵, but as we did not perform human evaluation on the larger model, we only report on the 114M model, soloni-114m-tdt-ctc-v0, in this study. soloni-114m-tdt-ctc-v0 uses a Fast-Conformer encoder (Rekesh et al., 2023) and a hybrid decoding setup with two independent but jointly trained decoders: a Token-and-Duration Transducer (TDT) decoder —an extension of the RNN-Transducer that predicts both a token and its duration (Xu et al., 2023)—and a convolutional decoder trained with a Connectionist Temporal Classification (CTC) loss (Graves et al., 2006). This dual-decoder design makes the model particularly interesting for analyzing how the two decoding approaches behave under the same training conditions. soloni-114m-tdt-ctc-v0 also provides insight into how the architecture will perform when scaling up to the larger models in the family.

stt-bm-quartznet15x5-v0 is a finetune of NVIDIA’s 18M parameter, ultra-compact QuartzNet model, an end-to-end convolutional architecture with 1D time-channel separable convolutions (Kriman et al., 2019). This model addresses a particularly critical use case in Mali where a large portion of the population does not have access to internet connectivity. We had al-

⁵We trained the Parakeet 600M model with an auto regressive decoder and 600M model with a convolutional decoder. While these models are not further discussed in this paper, all our models with their associated metrics can be found in RobotsMali’s Hugging Face repository.

ready deployed this model in a Bambara-language ASR-based reading tutor app⁶ that runs locally in low-end smartphones.

QuartzNet is a character-based decoding model. The vocabulary for the dataset consists of 38 unique characters, including the 30 letters of the Bambara alphabet, 5 accented French vowels, whitespace, hyphen (used in some compound words) and apostrophe (largely used in writing contractions). For soloni we train a SentencePiece tokenizer with a vocab size of 512. We had 4 NVIDIA A100 GPUs with a combined 320GB of VRAM for the experiment.

Training Data: We finetuned the two models on 98 hours of voice data with human-corrected transcripts, consisting of 167,816 utterances, and tested on 3 hours (5175 samples). We implemented and applied most of the normalization steps described by Zupon et al. before training, but we did not remove any of the acoustic event tags (presented in Appendix A) as we wanted to model those events as well.

Training configurations: We first trained soloni for 110k steps on 2 GPUs, with 32 batch size, using the AdamW optimizer and Noam scheduler with learning rate scaling factor of 0.003 and a 10% warmup ratio (Vaswani et al., 2017). Then we trained the resulting model for 100k more steps on all 4 GPUs, this time with an LR scaling factor of 1.5 and a 2% warmup ratio, all with no gradient accumulation and bf16 precision.

We trained QuartzNet for 65k steps on 4 GPUs, with 64 batch size, using the Novograd optimizer (Ginsburg et al., 2020) and a Cosine LR scheduler with a 1×10^3 and 1×10^6 upper and lower bounds and 6,000 warmup steps.

5 Evaluation of the Models

We evaluated soloni-114m-tdt-ctc-v0 both with the CTC and with TDT decoders, and the QuartzNet model on the Afvoices test set and a smaller benchmark, *ɲɛna* (transliterated to Nyana for English keyboards and nyana-eval for identification on Hugging Face)⁷, that we compiled through the stratified sampling of 15 audio files from each of:

⁶Our reading tutor app using the QuartzNet model, An Be Kalan, is available for [iOS](#) and [Android](#)

⁷*ɲɛna* means “opinion”; this emphasizes the human-evaluation focus of the dataset. We also release it on Hugging Face: <https://huggingface.co/datasets/RobotsMali/nyana-eval>

- the test set of Kunkado (Diarra et al., 2025a);
- a generally unused subset of *Jeli-asr* (street interviews) that we cleaned beforehand (Diarra et al., 2022);
- crowd-sourced recordings of readings of excerpts of the books from the GAIFE project (Tapo et al., 2025).

5.1 WER evaluation

Table 2 presents the Word Error Rates and Character Error Rates of the two models before and after our finetuning experiments. We show a significant improvement for all models across all metrics and benchmarks, up to **37% WER improvement** for soloni-114m-tdt, the best performing model overall. The relatively smaller WER improvement on the more challenging Nyana Eval benchmark, containing much noisier—sometimes multi-speaker—audio from street interviews and radio recordings than the Afvoices training and test data, highlights the potential limitations of the model in real-world deployment scenarios.

We note that the CTC branch of soloni lost its edge over TDT branch on the Afvoices test set when we increased the amount of training data from the 30 hours of *Jeli-asr* to 98 hours in this experiment, confirming that, for many sequence modeling tasks, autoregressive architectures outperform non-autoregressive ones when we scale training data (Graves, 2012; Li et al., 2020).

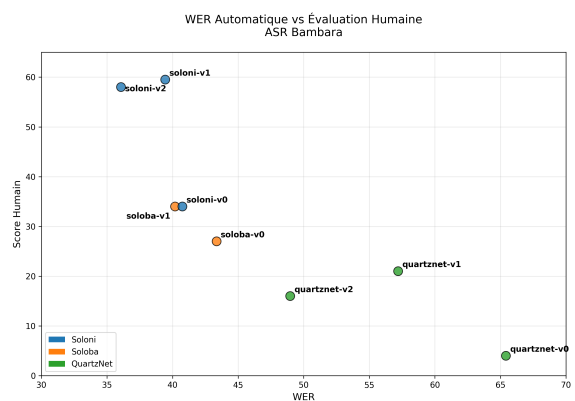


Figure 3: WER vs human evaluation. Figure from (Tall, 2025)

5.2 Human Evaluation

We performed a detailed human analysis of the outputs of several RobotsMali ASR models, including the models from this experiment, using

Model	WER (%) ↓		CER (%) ↓	
	Afvoices Test	Nyana Eval	Afvoices Test	Nyana Eval
soloni-114m (CTC)				
Unfinetuned (v0)	43.12	40.75	23.48	24.7
Finetuned (v2)	29.05	36.07	13.41	20.04
soloni-114m (TDT)				
Unfinetuned (v0)	45.52	47.1	26.68	31.27
Finetuned (v2)	28.58	38.13	12.94	22.3
Quartznet (CTC)				
Unfinetuned (v0)	73.66	65.42	37.85	30.66
Finetuned (v2)	42.57	48.97	18.70	24.22

Table 2: ASR experiment metrics: We apply the same normalization steps to our test sets and this time we remove the acoustic event tags from both the reference and the prediction before calculating the WER and CER. The values in bold highlight the best performances per metric

Nyana Eval as test data (Tall, 2025). Across models, we observed systematic difficulties with disfluencies, proper names, code-switching, and overlapping speech. The highest-rated model in this evaluation is soloni-v1, a finetune of soloni-v0 trained on RobotsMali/kunkado (Diarra et al., 2025a), a dataset composed of everyday speech. soloni-v2, finetuned on the Afvoices dataset with minimized noise, code-switching, and voice overlap, was judged slightly less robust by human evaluators on the more natural recordings in Nyana Eval, despite achieving a better WER (36.07% vs. 39.44%). Figure 3 plots the WER on Nyana Eval against the corresponding human evaluation scores.

We also report speed gains after replacing the v0 transcriptions with v2 in the human review-and-correction pipeline. The transcription team completed 45 hours of audio in approximately 800 hours of work, corresponding to a 17× real-time factor (17 hours of annotation per hour of speech). Using soloni-v2 instead of soloni-v0 yielded a 112% improvement in the rate at which human-corrected transcriptions could be produced, compared to an earlier study in which a 36× ratio was observed (Diarra et al., 2025b).

6 Conclusion

We released 612 hours of spontaneous Bambara speech and a 423-hour segmented corpus, together with metadata, VAD-based SNR estimates, transcription guidelines, a minimalist mobile recording app, multiple evaluation sets, and monolingual ASR models (an ultra-compact QuartzNet variant

and a 114M-parameter Parakeet-based model). For a low-literacy, predominantly oral language, this substantially increases the pool of publicly available resources.

Using soloni-114m-tdt-ctc-v0 in a human-model loop enabled semi-automated annotation, reducing the time required for corrected transcriptions and improving WER/CER on both in-domain and more challenging benchmarks. Human evaluation on Nyana Eval revealed systematic gaps between automatic string-based metrics and native-speaker judgments.

Taken together, these elements define a practical workflow for dealing with the hard facts of low-resource African NLP: targeted field collection with trained facilitators, noise-aware preprocessing, semi-automated annotation, and evaluation protocols that combine automatic and human measures while accounting for deployment on modest hardware. Future work will extend this approach to other Manding languages and to datasets that more fully reflect real-world speech.

Limitations

The African Next Voices (ANV) Bambara dataset was meticulously designed to provide a high-quality, clean foundation for ASR research. While successful in establishing the largest open corpus for the language, this intentional simplification inherently introduces limitations when considering the deployment of derived models in authentic, unconstrained Malian contexts. The core limitation lies in the necessary trade-off between controlled

data collection and the complex nature of the real-world environment. We can enumerate the following key limitations:

- **Acoustic Purity vs Real-World challenges:** The emphasis on clean recordings resulted in a dataset where 71.75% of the recordings were classified with 'High SNR' or 'Very High SNR'. This limits the model's exposure to challenging acoustic conditions—such as urban street noise and background voices that characterize typical Malian environments. Consequently, models trained exclusively on this corpus may exhibit a noticeable drop in performance when deployed in the real world.
- **Suppression of Code-Switching and Multilingualism:** The transcription protocol's mandate to replace code-switched terms and foreign words with the generic [cs] tag, or to force a Bambara-phonology transcription, simplifies the target vocabulary for ASR training. However, this approach sacrifices linguistic realism. The resulting models are fundamentally unprepared to transcribe the common, fluid shifts between Bambara and French, which are integral to spontaneous speech.

Ultimately, the ANV Bambara dataset represents a simplified version of the language's acoustic and linguistic reality. While this simplification provides a more stable foundation for core ASR research, it comes at the cost of real-world robustness. Practitioners seeking to deploy these models in authentic Malian contexts characterized by inherent noise, fluent code-switching, and diverse accents must anticipate the need for a targeted domain adaptation.

References

- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. **Common Voice: A massively-multilingual speech corpus**. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 4211–4215, Marseille, France. European Language Resources Association (ELRA).
- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. **wav2vec 2.0: A framework for self-supervised learning of speech representations**. *Preprint*, arXiv:2006.11477.
- Alan W Black. 2019. **Cmu wilderness multilingual speech dataset**. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975.
- Sebastien Diarra, Michael Leventhal, and Allahsera Auguste Tapo. 2022. **Robotsmali griots speech dataset, and asr**. <https://github.com/robotsmali-ai/jeli-asr/>.
- Yacouba Diarra, Nouhoum Coulibaly, Panga Azazia Kamaté, and Michael Leventhal. 2025a. **kunnafonidilaw ka cadeau: an ASR dataset to power the development of models that understand present-day bambara**. *Hugging Face Datasets*. Arxiv coming soon.
- Yacouba Diarra, Nouhoum Souleymane Coulibaly, and Michael Leventhal. 2025b. **Cost analysis of human-corrected transcription for predominately oral languages**. *Preprint*, arXiv:2510.12781.
- David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, 26th edition. SIL International, Dallas, TX.
- Chris Emezue, NaijaVoices Community, Busayo Awobade, Abraham Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, and Chris Pal. 2025. **The naijavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages**. *Preprint*, arXiv:2505.20564.
- Boris Ginsburg, Patrice Castonguay, Oleksii Hrinchuk, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, Huyen Nguyen, Yang Zhang, and Jonathan M. Cohen. 2020. **Stochastic gradient methods with layer-wise adaptive moments for training of deep networks**. *Preprint*, arXiv:1905.11286.
- 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.
- Alex Graves. 2012. **Sequence transduction with recurrent neural networks**. *Preprint*, arXiv:1211.3711.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. **Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks**. In *ICML 2006*, volume 2006, pages 369–376.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. **Scaling laws for neural language models**. *Preprint*, arXiv:2001.08361.
- Mamadou Konta and Valentin Vydrin. 2014. **Propositions pour l'orthographe du bamanankan**. *Man-denkan*, (52):3–38.

- Samuel Kriman, Stanislav Beliaev, Boris Ginsburg, Jocelyn Huang, Oleksii Kuchaiev, Vitaly Lavrukhin, Ryan Leary, Jason Li, and Yang Zhang. 2019. [Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions](#). *Preprint*, arXiv:1910.10261.
- Jinyu Li, Yu Wu, Yashesh Gaur, Chengyi Wang, Rui Zhao, and Shujie Liu. 2020. [On the comparison of popular end-to-end models for large scale speech recognition](#). *Preprint*, arXiv:2005.14327.
- Xinjian Li, Zhong Zhou, Siddharth Dalmia, Alan W. Black, and Florian Metze. 2019. [Santlr: Speech annotation toolkit for low resource languages](#). *Preprint*, arXiv:1908.01067.
- Vukosi Marivate, Ife Adebara, and Lilian Wanzare. 2025. [African languages for ai: the project that’s gathering a huge new dataset](#). The Conversation. Accessed: 2025-11-13.
- Omnilingual ASR Team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenhaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, and 13 others. 2025. [Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages](#).
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. [Scaling speech technology to 1,000+ languages](#). *Preprint*, arXiv:2305.13516.
- Dima Rekeshe, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. [Fast conformer with linearly scalable attention for efficient speech recognition](#). *Preprint*, arXiv:2305.05084.
- Silero Team. 2024. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. <https://github.com/snakers4/silero-vad>.
- Madani Amadou Tall. 2025. [Analyse comparative humaine des modèles asr bambara de robots mali](#).
- Allahsera Auguste Tapo, Nouhoum Coulibaly, Seydou Diallo, Sebastien Diarra, Christopher M Homan, Mamadou K. Keita, and Michael Leventhal. 2025. [GAIfe: Using GenAI to improve literacy in low-resourced settings](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7914–7929, Albuquerque, New Mexico. Association for Computational Linguistics.
- Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020–2022. [Label Studio: Data labeling software](#). Open source software available from <https://github.com/heartexlabs/label-studio>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *Preprint*, arXiv:1706.03762.
- Martin Vondrasek and Petr Pollák. 2005. Methods for speech snr estimation: Evaluation tool and analysis of vad dependency. *Radioengineering*, 14.
- Valentin Feodosievich Vydrin. 2022. [Vers un dictionnaire orthographique bambara](#). *Mandenkan : Bulletin Semestriel d’Études Linguistiques Mandé*, (68):59–82.
- Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Michael Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2016. [Achieving human parity in conversational speech recognition](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, PP.
- Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. 2023. [Efficient sequence transduction by jointly predicting tokens and durations](#). *Preprint*, arXiv:2304.06795.
- Haitian Xu, Zheng-Hua Tan, Paul Dalsgaard, and Børge Lindberg. 2005. [Robust speech recognition based on noise and snr classification - a multiple-model framework](#). In *Interspeech 2005*, pages 977–980.
- Andrew Zupon, Evan Crew, and Sandy Ritchie. 2021. [Text normalization for low-resource languages of africa](#). *Preprint*, arXiv:2103.15845.

A Transcription Guidelines

The transcription process for this Bambara dataset was implemented as a **First Review** task, where annotators corrected and validated pre-transcribed audio segments. Annotators were instructed to respect the following rules to simplify and standardize transcription and ensure high-quality data for ASR training and evaluation.

Language and Orthography

- Use the *standardized orthography of Bambara* (Konta and Vydrin, 2014). Correct any orthographic errors or non-standard characters (e.g., accented characters are prohibited, conforming to the ordinance of AMALAN⁸).

⁸We have been working closely with the Académie Malienne des Langues (AMALAN) and the Direction Nationale de l’Éducation Non Formelle et des Langues Nationales (DNENF-LN) prior to this project and we were able to leverage their expertise for all matters pertaining to national language standards for Bambara.

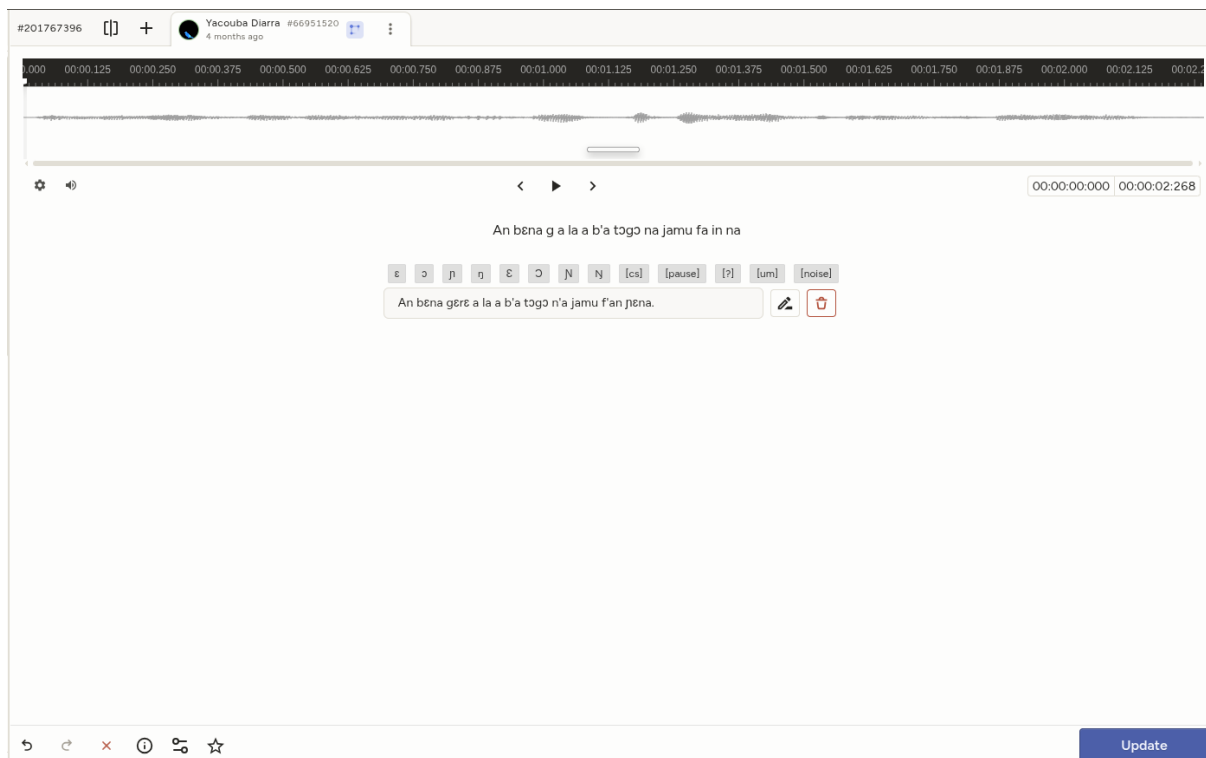


Figure 4: The Labeling Interface for the African Next Voices Bambara Transcription Project. The interface shows the original audio waveform, the automatically generated pre-transcription, and the field for human correction/validation.

- Do not correct the speaker’s grammatical errors or pronunciation mistakes.
- Elision and Mispronunciation: Transcribe exactly what is pronounced. For example, transcribe "Ne b’a fe" instead of "Ne be a fɛ" if "b’a" was pronounced.
- Proper Nouns: Must be capitalized.
- Repetitions, Disfluencies, and Stuttering: Write out repeated words/phonemes without using ellipses (e.g., "n’i ko ko ko ne kalabanci e dun?").
- Consult the Bamadaba dictionary for language references (Vydrin, 2022).
- Abbreviations and Acronyms: Transcribe them as pronounced and in *uppercase* (e.g., FIFA, BIM, ORTM). Do not use periods in acronyms.
- Spelled Words: Use hyphens to separate letters (e.g., K-E-L-E-N).
- **Code-Switching/Foreign Words:**
 - If the word exists in Bamadaba and was pronounced with the local Bambara phonology, transcribe it as written in the dictionary (e.g., Parce que → *paseke*, passerelle → *pasereli*). This rule exists because Bambara, like many African languages, has borrowed, transformed and standardized many words from the former colonial language.
 - If the word was pronounced with its original foreign pronunciation and/or is not recognized in Bamadaba, replace it with the tag [cs] (e.g., ne ka véhicule → ne ka [cs]).

Numbers, Abbreviations, and Foreign Words

- Numbers: Must be written in full letters exactly as pronounced (e.g., 35 → *bi saba ni duuru / bi saba ni Loru*).
- The Ordinal forms (e.g., *folo*, *filanan*, *sabanan*) must also be written in full letters.
- If a number is pronounced in a foreign language (e.g., French), use the code-switching tag [cs].

Acoustic Event Tags (for ASR Modeling)

Acoustic events and background sounds were retained and modeled in the final transcriptions using the following specific tags:

- Inaudible/Incomplete Speech or Overlap: Use the tag [?] for incomprehensible, non-audible speech, or speech overlaps.
- Vocalized/Disfluencies (Mouth Sounds): Use the tag [um] for sounds like *εε*, *hum*, *onh*, *ah*, *unhun*, etc.
- Long Silences: Use the tag [pause] for silences longer than 5 seconds (or longer than 3 seconds at the beginning or end of a segment). This tag was rarely used after VAD segmentation.
- Background Noise: Use the tag [noise] for all occurrences of strong background noise, including applause, coughing, laughter, phone rings, children, etc.

Punctuation

- Standard punctuation (commas, periods, question marks, etc.) should be used.

B Labeling Interface

The data annotation for the ANV Bambara project was performed using a tailored platform built on top of Label Studio (Tkachenko et al., 2020-2022). This interface facilitated the task by presenting pre-transcribed audio segments for human correction and validation. The audio files and their pre-labels were sequentially loaded from Google Cloud Storage into the labeling interface.

As shown in Figure 4, the simple interface provided the following key elements:

- Audio Segment: The interface displays the audio waveform and playback controls, allowing the annotator to listen to the segment.
- Pre-Transcription: The initial transcription was automatically generated by our ASR models.
- Correction Field: The annotator validates and corrects the automatic transcription in a designated field.
- Acoustic Event Tags: A row of buttons provides quick access to the acoustic event tags and the few Bambara characters that are not typically found on a standard keyboard.

This process of pre-transcription followed by human correction minimized human labor for segmentation while optimizing the usable speech data for the transcription pipeline.

M-MiniGPT4: Multilingual VLLM Alignment via Translated Data

Seung Hun Han ^{*†*}
MBZUAI

Youssef Mohamed ^{*}
KAUST

Mohamed Elhoseiny
KAUST

{firstname.lastname}@kaust.edu.sa

^{*} Equal Contribution [†] Work done while being an intern at KAUST

Abstract

This paper presents a Multilingual Vision Large Language Model, named M-MiniGPT4. Our model exhibits strong vision-language understanding (VLU) capabilities across 11 languages. We utilize a mixture of native multilingual and translated data to push the multilingual VLU performance of the MiniGPT4 architecture. In addition, we propose a multilingual alignment training stage that uses parallel text corpora to further enhance the multilingual capabilities of our model. M-MiniGPT4 achieves 36% accuracy on the multilingual MMMU benchmark, outperforming state-of-the-art models in the same weight class, including foundation models released after the majority of this work was completed. We open-source our models, code, and translated datasets to facilitate future research in low-resource and multilingual settings.

1 Introduction

With the rise of powerful general-purpose Large Language Models (LLMs), multimodal extensions have begun to garner significant attention. In particular, Vision Large Language Models (VLLMs) combine the reasoning capabilities of LLMs with visual perception using a vision encoder. Notable early examples of open-source VLLMs include (Zhu et al., 2023a), (Liu et al., 2023), and (Dai et al., 2023a).

However, as early LLM development focused primarily on the English language, the derived VLLMs also tended to possess limited multilingual capabilities. Consequently, the benefiting audience has been restricted to English speakers, leaving approximately 75% of the global population¹ unable to benefit from advancements in these powerful open-source VLLMs.

Recently, open-source LLMs trained on multilingual data, such as Llama 3.1 (Dubey et al., 2024), Qwen (Team, 2024), and Command R (AI, 2024), have emerged. Similarly, Qwen-VL (Bai et al., 2023), based on Qwen 2.5 (Yang et al., 2024), has demonstrated improved multilingual capabilities. However, many of these models were not developed with multilinguality as their core objective, leading to limitations in language coverage. In this work, we explore the development and benchmarking of massively multilingual VLLMs using both synthetic and human-translated data. Furthermore, we demonstrate that the plug-and-play framework of MiniGPT4 is well-suited for multilingual learning, allowing it to scale with independent advancements in text-only LLMs and vision encoders.

VLLMs are typically trained in three stages. The first stage is designed to align the vision and language modalities and involves large-scale paired image-language data. Models following stage 1 tend to show weak reasoning performance; thus, a second stage involving high-quality instruction data is required to produce performant models. While data for both stages are readily available for English, high-quality multilingual multimodal data is scarce, and parallel multilingual multimodal datasets are virtually nonexistent. To mitigate this issue, we utilize state-of-the-art translation models to translate popular vision-language datasets. We show that translated data improves the model’s multilingual performance without any notable degradation in English performance.

However, translated data does not account for the cultural and linguistic nuances that manifest only in natively collected datasets. As a result, relying solely on translation can result in sub-optimal multilingual VLLMs. To address this, we leverage parallel text corpora used for training machine translation models, as well as multilingual non-parallel text-only data, to improve the multilingual alignment of our models.

^{*}Corresponding author: eddiehunhan@gmail.com

¹cochrane.org/news/cochrane-evidence-different-languages

To summarize, our contributions are:

- We translate multiple vision-language datasets to create new multilingual resources.
- We demonstrate the use of parallel text corpora to improve the multilingual performance of VLLMs.
- We train a state-of-the-art (SOTA) multilingual VLLM based on the MiniGPT4 architecture.
- We translate the MMMU benchmark to assess the multilingual reasoning performance of VLLMs.
- We open-source all translated datasets and models to support the community.

2 Related Work

Large Language Models and Multilinguality. LLMs have emerged as a transformative force in artificial intelligence, with success attributed to advances in GPU capabilities and large-scale training data. The field witnessed a paradigm shift with GPT-3 (Brown et al., 2020), demonstrating remarkable zero-shot capabilities. This sparked the development of numerous models, including open-source alternatives such as Bloom and OPT (Scao et al., 2022; Zhang et al., 2022), and proprietary models like Chinchilla (Hoffmann et al., 2022), PaLM (Chowdhery et al., 2022), and Megatron-Turing NLG (Smith et al., 2022).

While LLaMA (Touvron et al., 2023) introduced an approach with fewer parameters but more extensive training data, and the field continues to evolve with Llama 2/3, GPT-4, and Mistral (Jiang et al., 2023), these models primarily focus on English. To address this, Multilingual Large Language Models (MLLMs) have emerged, excelling in cross-lingual transfer tasks. XLM-R (Conneau et al., 2019) pioneered cross-lingual capabilities, followed by models like Bloom and mT5 (Xue et al., 2020) that intentionally incorporate substantial non-English data. Their instruction-tuned variants, Bloomz and mT0 (Muennighoff et al., 2022), have further advanced multilingual capabilities. Our research leverages these robust cross-lingual transfer capabilities to extend Vision and Language models into multilingual applications.

VLLMs: Recent advances in vision-language integration have focused on adapting LLMs to process visual information. Early approaches like VisualGPT (Chen et al., 2022) and Flamingo (Alayrac

et al., 2022) combined pre-trained LLMs with visual features. BLIP-2 (Li et al., 2023) introduced the Q-former to bridge visual and language representations. Building upon this, MiniGPT-4 (Zhu et al., 2023b) enhanced performance by incorporating the Vicuna model. LLaVA (Liu et al., 2023) aligned a frozen image encoder with LLaMA through instruction tuning, while InstructBLIP (Dai et al., 2023b) leveraged 26 diverse datasets. While effective for English, cross-lingual capabilities in these models remain largely unexplored.

Recently, models such as PALO (Rasheed et al., 2025) have tackled the multilingual aspect of VLLMs, supporting visual reasoning for 10 languages via translated instruction datasets. Although PALO showed promising visual understanding, it performed poorly on visual reasoning benchmarks. We observed that PALO models excelled at lengthy descriptions but failed at direct question answering. In this paper, we show that this issue stems from the limited size of the PALO dataset. Accordingly, we provide a more diverse translated dataset, resulting in significantly better performance in both understanding and reasoning tasks.

3 Datasets

VLLMs require high-quality multimodal data. We used No Language Left Behind (NLLB 1.3B) (Costa-jussà et al., 2022) to translate popular vision-language datasets into 10 languages: Chinese, Hindi, Spanish, French, Arabic, Bengali, Russian, Urdu, Japanese, and Korean. Additionally, we translated the MMMU visual reasoning benchmark (Yue et al., 2024) using the same model. We utilize both V&L datasets and text-only datasets.

3.1 V&L Datasets

- **Conceptual Captions** (Sharma et al., 2018), **SBU** (Ordonez et al., 2011), and **LAION** (Schuhmann et al., 2021) are weakly-labelled datasets consisting of English image-caption pairs. They are used for Stage 1 pretraining to align vision and text modalities. Collectively, they consist of roughly 5 million instances.
- **LLaVA-Instruct** consists of 6.8 million English image-text pairs used to train the LLaVA 1.5 model (Liu et al., 2023). The text consists of conversations and responses generated by GPT-4. We translated this dataset using

NLLB to all 10 target languages. We refer to the translated version as **LAVAM**.

- **PALO** consists of 2 million image-text pairs used to train the PALO model (Rasheed et al., 2025). The dataset providers translated the LLaVA-665K dataset into 9 different languages.
- **Cambrian Image (CI)** is a high-quality multimodal dataset released for the Cambrian-1 family of models (Tong et al., 2024), consisting of approximately 58 million English image-text pairs. We translated this dataset to all 10 target languages and refer to the translated version as **CIM**.
- **WIT** is a natively multilingual dataset derived from Wikipedia, consisting of 130K image-article pairs (Srinivasan et al., 2021). We filtered over 95% of the original datapoints due to quality issues (corrupted text, stub articles). We used the NLLB-CLIP model (Visheratin, 2023) to measure caption-image cosine similarity ($S_{I,C}$) and selected only pairs where $S_{I,C} \geq 0.0$.

3.2 Text Corpora

- **Cambrian Text (CT)** is a high-quality text dataset released for the Cambrian-1 models (Tong et al., 2024), consisting of 22 million datapoints. We translated this to the target languages, naming the version **CT M**.
- **Flores** is a natively multilingual dataset derived from Wikipedia, consisting of 110K translation pairs between 11 languages (Team, 2022).
- **XStoryCloze** is a human-translated paragraph completion dataset derived from StoryCloze (Lin et al., 2021), consisting of 20K datapoints across 10 languages.

We combine Flores and XStoryCloze into **MText** in our experiments.

3.3 Multilingual MMMU Benchmark

We utilized NLLB to translate the MMMU benchmark, which is designed to evaluate the reasoning capabilities of VLLMs. We validated the quality of our translation via evaluation before and after back-translation. Specifically, we evaluated our model on the official English MMMU, translated

Model	E-MMMU	BT-MMMU
Pretrained	34.61	34.45
Finetuned	34.14	33.02

Table 1: **Performance validation via back-translation.** We report the average accuracy after back-translation from all target languages. **E-MMMU**: English MMMU; **BT-MMMU**: Back-Translated MMMU.

MMMU to the target languages, and then back to English. Finally, we evaluated our model on the back-translated version. If the translation caused significant information loss, performance should drop; however, as shown in Table 1, performance remains consistent.

4 Experiments

4.1 Model Setup

We base our model on the MiniGPT4 (Zhu et al., 2023b) architecture. To support multilinguality, we replace the Vicuna LLM with Llama 3 (Dubey et al., 2024), which demonstrates superior performance on multilingual tasks. Our training pipeline consists of three stages, each designed to enhance specific aspects of model performance.

Stage 1 aims to align the visual and language modalities. We use large-scale image-caption datasets (Conceptual Captions, SBU, and LAION). Our experiments indicate that incorporating additional datasets at this stage does not yield performance improvements; thus, this stage remains consistent with the original MiniGPT4 implementation.

Stage 2 enhances multilingual understanding by training with multilingual multimodal data. We leverage our translated datasets (ccSBU, LAION, LAVAM, PALO). We further experiment with the Cambrian Image (CI) dataset and its translated version (CI M).

Stage 3 focuses on boosting multilingual capabilities. We conduct ablation studies using CI and Cambrian Text (CT) datasets in both original and translated versions (CI M, CT M), as well as the parallel corpora used for translation (MText).

4.2 Results

Table 3 compares our model with state-of-the-art vision-language models on the MMMU and MMMU Multi benchmarks. On MMMU Multi, our model substantially outperforms other fine-tuning approaches built on the same base model, improv-

Stage 2	Stage 3	MMMU	MMMU Multi
(ccSBU, LAION, LAVAM, PALO)	-	31.02	29.83
+ CI	-	34.14	31.21
+ CI M	-	36.69	33.57
+ CI	CI M + CT M + MText	<u>37.07</u>	32.90
+ CI M	CI M + CT M + MText	37.27	<u>33.45</u>
+ CI	CI + CT	35.19	32.93
+ CI	CI + CT + MText	35.65	32.60

Table 2: Ablation Studies on Training Data Combinations. **CI**: Cambrian Image; **CT**: Cambrian Text; **M**: Translated/Multilingual version.

Model	MMMU	MMMU Multi
PALO	28.36	13.12
Qwen-VL 2.5	52.89	25.46
Our Model	37.27	33.45

Table 3: Comparison to SOTA Vision-Language Models.

ing from 13.12% (PALO) to 33.45%, despite both methods using the Llama-3 backbone. Our approach also exceeds the performance of the latest open-source foundational model, Qwen-VL 2.5, which achieves 25.46% on this benchmark, highlighting the effectiveness of the proposed method for multi-modal, multi-step reasoning.

On the standard MMMU benchmark, Qwen-VL 2.5 attains higher accuracy (52.89%) than our model. We attribute this gap primarily to differences in instruction tuning scale and data diversity, as Qwen-VL 2.5 benefits from more extensive instruction tuning than was applied in our setting.

Table 2 presents our ablation studies. Several key observations emerge:

- Adding the Cambrian Image dataset (CI) in Stage 2 improves performance on both benchmarks.
- Using the translated version (CI M) in Stage 2 yields further improvements (36.69% on MMMU and 33.57% on MMMU Multi).
- The optimal configuration combines CI M in Stage 2 with CI M + CT M + MText in Stage 3.
- Including multilingual text data (MText) in Stage 3 generally improves performance when combined with translated datasets.

These results demonstrate the effectiveness of our three-stage training approach and the impor-

tance of incorporating multilingual multimodal data to enhance cross-lingual vision-language understanding.

5 Conclusion

In this paper, we presented M-MiniGPT4, a multilingual vision-language model that demonstrates strong performance across 11 languages. Our approach leverages a three-stage training process that effectively combines native multilingual data with translated datasets to optimize cross-lingual vision-language understanding. We demonstrated that using translated vision-language data significantly improves multilingual performance and that incorporating parallel text corpora further enhances the model’s capabilities.

Our experiments show that M-MiniGPT4 achieves state-of-the-art multilingual performance on the MMMU Multi benchmark (33.45%), substantially outperforming existing models like Qwen-VL 2.5 and PALO in multilingual visual reasoning tasks. By open-sourcing our models and translated datasets, we facilitate further research in multilingual multimodal AI, making these technologies more accessible to non-English speakers worldwide.

6 Limitations

Despite promising results, M-MiniGPT4 faces several limitations:

- **Translation Nuance:** Reliance on machine translation may not fully capture cultural nuances and linguistic subtleties present in natively collected multilingual data.
- **Language Coverage:** While our model was finetuned on 11 languages, this covers only a fraction of the world’s languages.

- **Resource Disparity:** Translation quality varies, with high-resource languages (e.g., Spanish, French) benefiting from better translations compared to lower-resource languages (e.g., Bengali, Urdu).
- **Evaluation:** Our metrics may not comprehensively assess all aspects of cross-cultural understanding in visual reasoning.
- **Inherited Bias:** Reliance on pretrained LLMs inherits the biases and limitations inherent in the base models.

Future work should focus on expanding language coverage, incorporating more natively collected multilingual data, and developing nuanced evaluation frameworks for cross-cultural understanding.

References

- Cohere AI. 2024. [Cohere command r models](#). Accessed: 2024-12-24.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. [Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond](#). *Preprint*, arXiv:2308.12966.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). *Preprint*, arXiv:2005.14165.
- Jun Chen, Han Guo, Kai Yi, Boyang Li, and Mohamed Elhoseiny. 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18030–18040.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.
- 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*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023a. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#). *Preprint*, arXiv:2305.06500.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023b. [Instructblip: Towards general-purpose vision-language models with instruction tuning](#). *arXiv preprint arXiv:2305.06500*.
- 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*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. [Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models](#). *arXiv preprint arXiv:2301.12597*.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, Veselin Stoyanov, and Xian Li. 2021. [Few-shot learning with multilingual language models](#). *CoRR*, abs/2112.10668.

- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.
- 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*.
- Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. *Im2text: Describing images using 1 million captioned photographs*. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.
- Hanoona Rasheed, Muhammad Maaz, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M. Anwer, Tim Baldwin, Michael Felsberg, and Fahad S. Khan. 2025. Palo: A large multilingual multimodal language model. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV 2025)*.
- Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.
- Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernamed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.
- Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. 2022. Using deep-speed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.
- Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. *Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning*. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21*, page 2443–2449. ACM.
- NLLB Team. 2022. No language left behind: Scaling human-centered machine translation.
- Qwen Team. 2024. *Qwen2.5 technical report*. *Preprint*, arXiv:2412.15115.
- Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. *Cambrian-1: A fully open, vision-centric exploration of multimodal llms*. *Preprint*, arXiv:2406.16860.
- 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*.
- Alexander Visheratin. 2023. *Nllb-clip – train performant multilingual image retrieval model on a budget*. *Preprint*, arXiv:2309.01859.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer. *arXiv preprint arXiv:2010.11934*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023a. *Minigt-4: Enhancing vision-language understanding with advanced large language models*. *Preprint*, arXiv:2304.10592.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023b. *Minigt-4: Enhancing vision-language understanding with advanced large language models*. *arXiv preprint arXiv:2304.10592*.

InstructLR: A Scalable Approach to Create Instruction Dataset for Under-Resourced Languages

Mamadou K. KEITA¹, Sebastien Diarra², Christopher Homan¹, Seydou Diallo³

¹Rochester Institute of Technology, ²RobotsMali ³MALIBA-AI

Abstract

Effective text generation and chat interfaces for low-resource languages (LRLs) remain a challenge for state-of-the-art large language models (LLMs) to support. This is mainly due to the difficulty of curating high-quality instruction datasets for LRLs, a limitation prevalent in the languages spoken across the African continent and other regions. Current approaches, such as automated translation and synthetic data generation, frequently yield outputs that lack fluency or even orthographic consistency. In this paper, we introduce InstructLR, a novel framework designed to generate high-quality instruction datasets for LRLs. Our approach integrates LLM-driven text generation with a dual-layer quality filtering mechanism: an automated filtering layer based on retrieval-augmented-generation (RAG)-based n-shot prompting, and a human-in-the-loop validation layer. Drawing inspiration from benchmarks such as MMLU in task definition, InstructLR has facilitated the creation of three multi-domain instruction benchmarks: **ZarmaInstruct-50k**, **BambaraInstruct-50k**, and **FulfuldeInstruct-50k**.

1 Introduction

Large language models (LLMs) are proficient in many tasks, with recent models sometimes outperforming humans, *depending on the language*. They tend to perform *substantially worse* on low-resource languages (LRLs), such as those spoken across Africa and other regions, than on higher-resource languages. This performance gap is evidently due to the limited representation of these languages in pre-training and fine-tuning datasets. Although LLMs such as GPT-4 (OpenAI et al., 2024) and Gemini (Team et al., 2024) have made progress in multilingual capabilities, many LRLs remain poorly, if at all, supported.

Existing approaches to address this gap also face major limitations. Machine translation (MT)

of fine-tuning datasets from higher-resourced languages into LRLs often produces unnatural text that fails to capture language-specific nuances (Zhu et al., 2024). Synthetic data generation frequently results in hallucinated content and a lack of cultural awareness (Guo and Chen, 2024). The relatively high cost of creating human-annotated instruction data for LRLs worsens the situation.

We introduce **InstructLR**, a novel framework designed to produce high-quality instruction tuning datasets for LRLs through a combined approach that balances automation with human-in-the-loop validation. Unlike direct translation approaches that often produce unnatural outputs, InstructLR uses translation at the instruction response generation stage, where instructions—initially in a high-resource language (e.g., French)—are translated to the target LRL along with the other output components. **This allows the model to generate *contextually appropriate* responses directly in the target language (since the high resource and low resource instructions will be both embedded during the responses generation)—rather than translating complete instruction-response pairs.**

Our contributions are as follows:

- We propose **InstructLR**, a scalable pipeline that integrates LLM generation, RAG-based correction, and human-in-the-loop validation to produce high-quality instruction data for LRLs.
- We use this framework to create three 50k-scale, multi-domain instruction benchmarks: **ZarmaInstruct-50k**, **BambaraInstruct-50k**, and **FulfuldeInstruct-50k**—all under a CC-BY-SA 4.0 license—with links available at: https://huggingface.co/datasets/27Group/InstructLR_Generate_Datasets.
- We conduct experiments comparing three

training approaches: zero-shot baseline (no fine-tuning), MT-Seed baseline (fine-tuning on machine-translated instructions), and InstructLR (fine-tuning on our framework’s output). This comparison aims to isolate the effectiveness of our framework versus direct translation methods.

Our evaluation addresses three research questions: (RQ1) How do open-source LLMs perform on instruction-following tasks for these LRLs without fine-tuning? (RQ2) How much does fine-tuning on InstructLR datasets improve performance compared to MT baselines? (RQ3) How well do InstructLR-trained models generalize to downstream tasks?

Our study demonstrates that InstructLR enables effective instruction-following in previously unsupported languages, by achieving BLEU scores of 22.8 (Zarma), 30.1 (Bambara), and 28.9 (Fulfulde) compared to near-zero baseline performance. Furthermore, the framework reduces dataset creation costs by 88% through automated quality filtering while maintaining good linguistic quality, as validated by native speakers who preferred InstructLR outputs over machine-translation baselines in 78-84% of comparisons.

2 InstructLR

We designed **InstructLR** (Figure 1) to assist in creating domain-specific instruction datasets for LRLs.

InstructLR consists of multiple stages—including: seed instruction, instruction-response-pair creation, automated quality checking, human validation, and the final dataset—organized as a pipeline. In this section, we describe each stage and show how they work together to produce clean instruction data.

2.1 Seed Instruction

Topic Selection To ensure the final dataset is comprehensive and useful for training models, InstructLR starts by curating a diverse set of topics. We draw inspiration from established multi-task benchmarks like MMLU (Hendrycks et al., 2021) because they provide a structured framework of knowledge domains and reasoning skills. Our selection process targets a balanced distribution across a wide range of areas. These include STEM fields (e.g., Physics, Mathematics, Computer Science), humanities (e.g., History, Law, Philosophy),

and social sciences. The goal is to create a dataset that supports not only knowledge recall but also the development of complex reasoning abilities.

Seed Instruction Generation After gathering the topic list, seed instructions are generated in a high-resource language. This approach is a necessary adaptation of the self-instruct method (Wang et al., 2023) for the LRL context. The standard self-instruct loop is technically infeasible here, as it requires a teacher model with strong generative capabilities *in the target language* to create novel instructions—a prerequisite that current models do not meet for languages like Zarma. Our method circumvents this by using the LLM for the task it can perform well (ideation in French). The choice of the high-resource language depends on its presence in the region where the target LRL is used—e.g., French-speaking countries will use French.

The seed generation process uses a modified self-instruct method, where we design an instruction generation prompt template (see Section I.1) to produce diverse, domain-appropriate instructions. We incorporate two quality control mechanisms within the prompt: (1) We add instruction diversity by using different directive verbs—e.g., explain, describe, analyze—to prevent repetitive instructions. (2) The prompt includes guidelines to avoid output that contains hallucinations, sensitive content, or falls outside the target domain.

The output is structured in a JSONL format, where each instruction is based on one topic.

2.2 Instruction-Response Pairs

Once the curated set of seed instructions is prepared, the next step is generating instruction-response pairs in the target LRL. This is done using an LLM with some baseline capability—ability to generate mediocre, yet acceptable outputs—to generate content in the target LRL¹.

The LLM is prompted using a structured prompt template—(see Section I.2)—with specific guidelines to handle edge cases often encountered during translation between the higher-resource language and the target LRL, and other specifications such as the response length. **The seed instructions enable the model to translate the instructions to the LRL and generate responses directly in the LRL, informed by both the high-resource and LRL**

¹This phase only works if the chosen LLM has indeed a baseline ability to generate in the target LRL. Otherwise, the produced content would be hallucinated outputs.

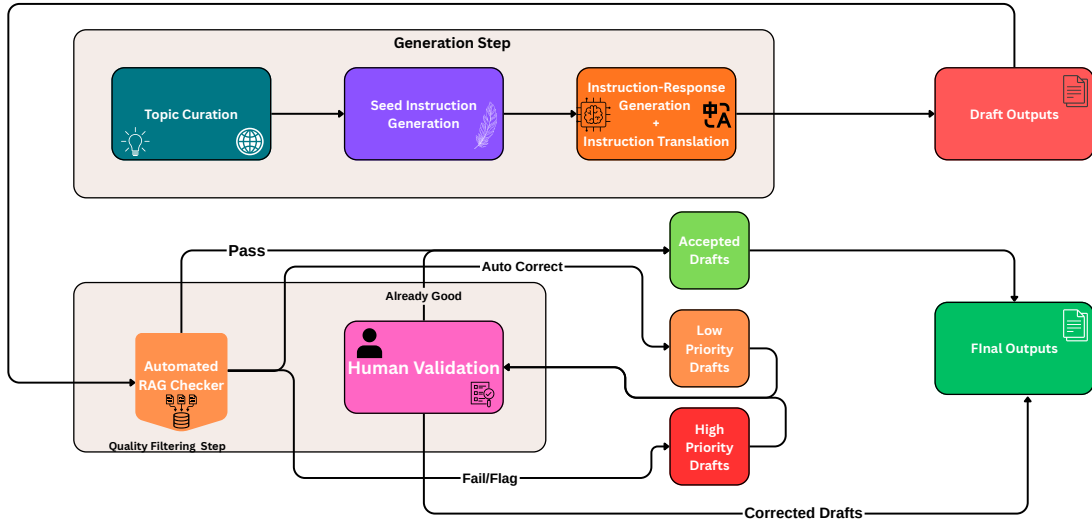


Figure 1: The InstructLR pipeline for creating high-quality instruction-tuning datasets for LRLs. The pipeline starts by the topic curation and finishes by final output.

instructions—unlike MT approaches that translate pre-existing aligned segments. The template includes explicit constraints addressing: (1) **Word adaptation:** rules for handling technical terms, proper nouns, and domain-specific vocabulary that might not have direct equivalents in the target LRL. (2) **Prioritize understandability:** guidelines to prioritize understandability and fidelity over word-for-word translation. (3) **Language specific constraints:** language specific guidelines that cannot be generalized.

For reasoning tasks, the prompt additionally requests a chain-of-thought (CoT) component in the target LRL and ensures that the generated responses include explicit reasoning steps in the LRL.

This stage outputs **drafts** structured by key metadata fields, as shown in Table 11. Each draft includes the original instruction in the high-resource language, the translated instruction in the target LRL, the generated response in the target LRL, and, for reasoning tasks, the CoT explanation—in case of reasoning tasks—in the target LRL.

2.3 Dual-Layer Quality Filtering

Raw drafts produced by an LLM often contain domain inconsistencies, fluency issues, and factual errors—particularly for LRLs with limited coverage in pretraining data. To deliver a dataset with a minimized error rate while keeping human effort affordable, we implemented a dual-layer quality pipeline that combines automated and human-driven quality assessment.

Layer 1: Automated Quality Check An automated Retrieval-Augmented Generation (RAG) checker processes the drafts using a knowledge base of clean sentences, grammar rules, and glossaries of the LRL. To ground the automated quality assessment, the RAG checker retrieves relevant information to guide the LLM’s correction suggestions, and ensures that every correction adheres to linguistic rules of the LRL. With an elaborated n -shot prompting, it suggests corrections or flags drafts for human review. When the RAG successfully corrects a draft, it is marked as “**low priority**” for human review. If the RAG flags a draft as problematic but can not propose a correction, it is marked as “**top priority**” for human review. Drafts with no detected issues are accepted as is.

The RAG component is convenient when the LLM used for checking has moderate proficiency in the LRL. For LRLs with “**no**” LLM support, alternative strategies for the automated layer would be needed; and for LRLs where LLMs are already highly proficient, simpler prompting might suffice for the automated check.

Layer 2: Human Validation A team of native speakers checks drafts flagged or corrected by the RAG system. The human validation protocol varies depending on the language. However, the main objective is to assess the grammar, orthography, and fluency. All corrected and validated drafts are then formatted as JSONL.

InstructLR is designed to be language-agnostic, requiring only minimal adaptation to target a new

LRL. The framework’s modularity allows components to be improved or replaced depending on the context.

3 Dataset Creation and Analysis

To demonstrate the effectiveness of InstructLR for generating instruction datasets, we report on our use of it to create a dataset in Zarma, a West African language spoken by over six million people (Keita et al., 2024).

3.1 Seed Instruction Creation

For this stage, we selected 20 topics—listed with descriptions in Table 10—and proceeded with instruction generation. Since Zarma coexists with French in everyday usage (Keita et al., 2024), we chose French as the primary language for generating seed instructions, and a suitable model for French: the **Mistral 7b** model (Jiang et al., 2023). We then generated French instructions per topic and equally split across the topics ($\approx 5\%$ per topic).

3.2 Draft Generation

Once we had the curated set of French seed instructions and their associated topics, we moved on to generating the first drafts of instruction-response pairs in Zarma². To achieve this, we tested several models—Gemini 2.5 Pro, GPT 4.o, and Llama 3.3 (Grattafiori et al., 2024)—to determine which one demonstrated a relatively acceptable understanding of Zarma.

We selected Gemini 2.5 Pro due to its basic understanding of Zarma. While not perfect, it outperformed other models in generating coherent Zarma texts with fewer hallucinations.

We adjusted the prompt template (see Section I.2) for Gemini and included the following specific guidelines to handle edge cases that may happen during translation between French and Zarma. These included:

Handling of nouns and loanwords: We instructed the model not to change proper nouns. For example, names of people, cities like Niamey, or countries like Niger should remain as they are, rendered in the target language’s phonetic script. Similarly, for common French loanwords already understood in Zarma, the model was prompted to keep the existing commonly used form.

Scientific or technical terms: If the input text contained scientific or technical terms that

²All 50,000 instructions were processed, and a snapshot of the outputs is shown in Table 11.

do not have a direct, commonly known equivalent in Zarma—e.g., a term like “photosynthesis” or “algorithm”—the instruction was to keep the original term unchanged. The same rules apply to things like book titles, etc. The goal was to avoid the model inventing new words that would not be understood.

Managing unknown French words: For French words in the input that the model needed to use in the output but might not have a standard equivalent or common borrowing in the target language, we allowed a process of phonetic adaptation. This means the model could “**Frenchize**” the word—writing it out in the target language’s phonetic script based on its French pronunciation. A good example of this might be the French word “**politique**,” which could be written as “**politik**” in Zarma or Bambara, if that matches how such words are typically borrowed and written phonetically. This was preferred over omitting the concept or making a potentially incorrect direct translation.

3.3 Quality Assessment

Knowledge base construction: Our RAG checker used a knowledge base of 3,000 clean sentences from the Feriji dataset (Keita et al., 2024), 20 Zarma grammar rules each followed by examples, and bilingual glossaries, all encoded with a FAISS dense index (Douze et al., 2025). This knowledge base enabled the system to contextualize and evaluate drafts with high precision.

Base model: We relied on the Gemini 2.0 flash model for our RAG. Similarly to the reason of selecting Gemini 2.5 Pro for drafts generation, the choice of the model is guided by the fact that the model already has a basics understanding of the language.

The full detail of our RAG checker is explained in Section C.

After processing the 50,000-draft dataset, **4,563** drafts were flagged as top priority—a ratio of 9.126% of the dataset—while **2,535** were successfully corrected by the RAG, considered low priority (5.07%). The remaining **42,902** drafts were accepted without correction.

3.3.1 Human Evaluation

Annotator pool: We recruited five volunteers—all native Zarma speakers with prior experience reading and writing in the language. Before starting work, annotators underwent a short training session covering: the annotation task itself, how to use the

tools, and what types of corrections are acceptable. Additionally, we assessed the inter-annotator agreement using **Krippendorff’s Alpha**, and obtained a score of **0.793** on 351 samples from the annotated sets. The results of the evaluation are presented in Table 1.

Evaluation outcomes: As shown in Table 1, among the 4,563 top-priority flagged samples, the primary issues detected were fluency problems (56.40%), followed by suffix misuse errors (24.14%) and tense consistency errors (19.46%). In the 2,535 low-priority samples, **1978** (78.028%) were already correct despite being flagged by the automated system, with the remaining **557** (21.97%) requiring only minor typographic adjustments that did not affect comprehensibility.

3.4 ZarmaInstruct-50k Dataset

Following the InstructLR pipeline, we created ZarmaInstruct-50k, the first multi-domain instruction benchmark in the Zarma language. The dataset is composed of 50,000 instruction-response pairs covering 20 different topics (as shown in Table 10). Table 1 presents statistics of ZarmaInstruct-50k.

3.5 Generalization to Bambara and Fulfulde

To validate the language-agnostic nature and scalability of our framework, we applied the full **InstructLR** pipeline to two additional West African LRLs: Bambara and Fulfulde. We maintained the core methodology used for Zarma, generating **50,000** instruction-response pairs for each language using the same seed topics and French as the high-resource language. The objective was to confirm that the framework could be effectively redeployed with minimal adaptation.

The process yielded two new large-scale benchmarks: **BambaraInstruct-50k** and **FulfuldeInstruct-50k**. Initial raw drafts generated by Gemini 2.5 Pro showed error patterns comparable to those observed in Zarma—including minor fluency issues and occasional word-level hallucinations, which highlights the need for the dual-layer quality filtering mechanism to address these errors.

More details about the generation process, raw output quality assessment, and full dataset statistics for both Bambara and Fulfulde are provided in Section D.

4 Experiments

We evaluate InstructLR through systematic experiments that assess both output quality and downstream task performance.

Experiment Setups We evaluate six open-source models across different parameter scales: Gemma-3-270M, Gemma-3-1B, Gemma-3-4B (Team et al., 2025), Llama-3.1-8B (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3, and Phi-4 (Abdin et al., 2024). For each language, we split our 50k datasets into 49,000 training pairs and 1,000 held-out test pairs for evaluation.

For the baselines, We compare against two baselines; **Zero-Shot Baseline:** Each base model evaluated on test sets without fine-tuning. **MT-Seed Baseline:** To isolate the effect of our generation pipeline, we create a controlled comparison using direct MT of our French seed instructions. We fine-tune Llama-3.1-8B (our best model across all the languages experimented before the MT one) on datasets created by translating the same 50,000 French seed instructions using MADLAD-400 (Kudugunta et al., 2023)—because MADLAD is the only known model (until this date) that supports all the three languages of this experiment. This approach avoids confusion caused by culture-specific instructions in existing datasets such as Alpaca (Taori et al., 2023).

We use unsloth (Daniel Han and team, 2023) with QLoRA (Dettmers et al., 2023) for efficient fine-tuning. Training parameters include: learning rate $2e-5$, 3 epochs, with CoT responses included as supervised targets. We ensure no overlap between training and test sets.

Automatic Evaluation Table 2 presents results on held-out test sets using BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) metrics. Zero-shot performance demonstrates limitations of current LLMs for these languages, with scores near zero across all models—which confirms that Zarma, Bambara, and Fulfulde are minimally or not covered by the models used for the trainings.

Fine-tuning on InstructLR datasets produces important improvements. The best-performing model (Llama-3.1-8B with InstructLR) achieves 22.8 BLEU on Zarma, 30.1 on Bambara, and 28.9 on Fulfulde. **These results demonstrate that our framework enables effective instruction-following capabilities in previously unsupported**

Table 1: ZarmaInstruct-50k Dataset Characteristics and Quality Assessment. *Percentage of top priority drafts (4,563). †Percentage of low priority drafts (2,535).

(a) Dataset Characteristics			(b) Quality Assessment Results		
Metric	Count	%	Metric	Count	%
<i>Instruction Distribution</i>			<i>Automated Filtering</i>		
Instructions with 1–10 tokens	1,379	2.76	Total drafts processed	50,000	100.00
Instructions with 11–20 tokens	27,655	55.31	Accepted without correction	42,902	85.80
Instructions with >20 tokens	20,966	41.93	Low priority (corrected by RAG)	2,535	5.07
			Top priority (needs human review)	4,563	9.13
<i>Response Distribution</i>			<i>Human Validation - Top Priority</i>		
Responses with <50 tokens	29,833	59.67	Major fluency errors	2,574	56.41*
Responses with 50–100 tokens	20,167	40.33	Suffix misuse errors	1,101	24.13*
Instructions with CoT reasoning	12,500	25.00	Tense consistency errors	888	19.46*
<i>Instruction Types</i>			<i>Human Validation - Low Priority</i>		
Open-ended questions	41,957	83.91	Already correct	1,978	78.03†
Definition requests	121	0.24	Minor typographic adjustments	557	21.97†
Explanation tasks	5,781	11.56			
List generation tasks	2,141	4.28			

languages.

The MT-Seed baseline underperforms InstructLR across all languages. On Zarma, InstructLR outperforms MT-Seed by 9.3 BLEU points (22.8 vs 13.5).

Human Evaluation We conduct comprehensive human evaluation with native speakers using our best-performing model (Llama-3.1-8B with InstructLR) across three evaluation protocols.

– **Pairwise Preference Evaluation** Two native speakers per language independently compared system outputs on 500 randomly selected prompts from our test sets. Evaluators chose between system outputs or marked ties when outputs were equivalent in quality.

Table 3 shows strong preference for InstructLR across all languages. Against zero-shot baselines, InstructLR wins in 89.2% of Zarma comparisons, 94.0% of Bambara comparisons, and 91.8% of Fulfulde comparisons. The high tie rates with zero-shot baselines (4-6%) reflect cases where both systems produced minimal or no valid output. When compared to MT-Seed baselines, InstructLR maintains advantages with win rates of 78.4% (Zarma), 83.6% (Bambara), and 80.8% (Fulfulde). The lower margins against MT-Seed reflect that both systems produce fluent output, but InstructLR demonstrates higher linguistic quality and appropriateness.

– **Quality Evaluation** Native speakers rated 500 responses per protocol on three quality aspects using 5-point scales: fluency, correctness, and relevance.

Table 4 demonstrates quality advantages for InstructLR across all aspects and languages. Zero-shot baselines score poorly (1.1-1.6 range) due

to their inability to generate coherent responses in these languages. MT-Seed baselines achieve moderate scores (2.1-3.3 range) but fall short of InstructLR’s performance. InstructLR achieves strong scores across languages, with Bambara and Fulfulde showing particularly high ratings (4.0-4.2 range). The slightly lower Zarma scores (2.9-3.7 range) reflect the more complex grammatical structure and our evaluation criteria during the human validation process.

4.1 Downstream Task Evaluation

To assess practical utility beyond instruction-following, we evaluate models on Named Entity Recognition (NER). We created 1,000-statement NER datasets per language with annotations for person, location, and organization entities. Models were prompted to extract entities using zero-shot prompting without task-specific fine-tuning. We evaluate using exact match accuracy and macro-averaged F1 scores.

Table 4 shows that InstructLR-trained models demonstrate strong generalization to downstream tasks. InstructLR achieves exact match scores of 41.2% (Zarma), 54.4% (Bambara), and 50.6% (Fulfulde), outperforming both zero-shot baselines (9-13% range) and MT-Seed baselines (27-37% range).

The improvements over MT-Seed baselines (13-17 percentage point gains) confirm that our quality filtering approach produces more reliable training data that enables better task generalization.

5 Discussion

Our experimental results demonstrate that InstructLR successfully creates useful instruction datasets for under-resourced languages. The experiments confirm that models fine-tuned on our

Lang.	Model	Protocol	BLEU \uparrow	R-L \uparrow	MTR \uparrow
Zarma	G-270M	Zero-Shot	0.1 \pm 0.1	1.2 \pm 0.5	0.5 \pm 0.3
	G-270M	InstructLR	12.5 \pm 1.8	18.3 \pm 2.1	15.1 \pm 1.9
	G-1B	Zero-Shot	0.2 \pm 0.1	1.4 \pm 0.6	0.6 \pm 0.3
	G-1B	InstructLR	15.8 \pm 2.0	22.1 \pm 2.5	18.4 \pm 2.2
	G-4B	Zero-Shot	0.3 \pm 0.2	1.7 \pm 0.7	0.7 \pm 0.4
	G-4B	InstructLR	18.2 \pm 2.2	25.6 \pm 2.8	21.3 \pm 2.5
	L-8B	Zero-Shot	0.3 \pm 0.2	1.8 \pm 0.8	0.8 \pm 0.4
	L-8B	MT-Seed	13.5 \pm 1.9	20.1 \pm 2.4	16.5 \pm 2.0
	L-8B	InstructLR	22.8\pm2.5	30.4\pm3.1	26.1\pm2.8
	Mistral-7B	Zero-Shot	0.2 \pm 0.1	1.5 \pm 0.6	0.6 \pm 0.3
	Mistral-7B	InstructLR	20.1 \pm 2.3	28.5 \pm 3.0	23.9 \pm 2.6
	Phi-4	Zero-Shot	0.3 \pm 0.2	1.6 \pm 0.7	0.7 \pm 0.4
	Phi-4	InstructLR	21.8 \pm 2.4	29.7 \pm 3.0	25.1 \pm 2.7
	Bambara	G-270M	Zero-Shot	0.2 \pm 0.1	1.1 \pm 0.5
G-270M		InstructLR	11.8 \pm 1.7	17.9 \pm 2.0	14.6 \pm 1.8
G-1B		Zero-Shot	0.3 \pm 0.2	1.6 \pm 0.7	0.7 \pm 0.4
G-1B		InstructLR	18.1 \pm 2.1	24.7 \pm 2.6	21.2 \pm 2.3
G-4B		Zero-Shot	0.4 \pm 0.3	1.9 \pm 0.8	0.8 \pm 0.4
G-4B		InstructLR	23.2 \pm 2.5	31.4 \pm 3.2	27.8 \pm 2.9
L-8B		Zero-Shot	0.4 \pm 0.3	2.1 \pm 0.9	0.9 \pm 0.5
L-8B		MT-Seed	21.3 \pm 2.4	29.8 \pm 3.0	25.7 \pm 2.7
L-8B		InstructLR	30.1\pm2.9	39.8\pm3.8	34.5\pm3.4
Mistral-7B		Zero-Shot	0.3 \pm 0.2	1.7 \pm 0.7	0.7 \pm 0.4
Mistral-7B		InstructLR	25.8 \pm 2.7	34.1 \pm 3.4	30.2 \pm 3.1
Phi-4		Zero-Shot	0.4 \pm 0.3	1.8 \pm 0.8	0.8 \pm 0.4
Phi-4		InstructLR	27.3 \pm 2.8	36.5 \pm 3.6	32.1 \pm 3.2
Fulfulde		G-270M	Zero-Shot	0.1 \pm 0.1	1.0 \pm 0.4
	G-270M	InstructLR	10.9 \pm 1.6	16.8 \pm 1.9	13.7 \pm 1.7
	G-1B	Zero-Shot	0.2 \pm 0.1	1.3 \pm 0.6	0.5 \pm 0.3
	G-1B	InstructLR	16.7 \pm 2.0	23.1 \pm 2.5	19.8 \pm 2.2
	G-4B	Zero-Shot	0.3 \pm 0.2	1.6 \pm 0.7	0.7 \pm 0.4
	G-4B	InstructLR	21.8 \pm 2.4	29.3 \pm 3.0	25.9 \pm 2.7
	L-8B	Zero-Shot	0.2 \pm 0.2	1.5 \pm 0.7	0.6 \pm 0.4
	L-8B	MT-Seed	19.7 \pm 2.3	28.1 \pm 2.9	24.2 \pm 2.6
	L-8B	InstructLR	28.9\pm2.8	38.2\pm3.7	33.1\pm3.3
	Mistral-7B	Zero-Shot	0.2 \pm 0.1	1.4 \pm 0.6	0.6 \pm 0.3
	Mistral-7B	InstructLR	24.3 \pm 2.6	32.7 \pm 3.3	28.9 \pm 3.0
	Phi-4	Zero-Shot	0.3 \pm 0.2	1.6 \pm 0.7	0.7 \pm 0.4
	Phi-4	InstructLR	26.1 \pm 2.7	35.0 \pm 3.5	30.8 \pm 3.1

Table 2: Results of the metric-based experiments. G = Gemma-3, L = Llama-3.1-8B, $R-L$ = ROUGE-L, MTR = METEOR.

data achieve important improvements over both zero-shot and MT baselines. Furthermore, the performance gains across three different languages—Zarma, Bambara, and Fulfulde—prove the framework’s language-agnostic design.

An important component behind the framework’s effectiveness is its dual-layer quality filtering mechanism. The automated RAG-based layer processes the majority of the data (85.8%) without human input, which directly enables the 88% cost reduction compared to full human annotation (see Section F). This balance makes large-scale dataset creation economically feasible. The quality of the resulting data is confirmed by the high performance on automatic metrics—where fine-tuning yields BLEU scores as high as 22.8 (Zarma), 30.1 (Bambara), and 28.9 (Fulfulde) from near-zero baselines.

Human evaluation further emphasizes these find-

Lang	Comparison	InstructLR	Baseline	Ties
Zarma	Zero-Shot	89.2% [86.1, 91.7]	4.4% [2.9, 6.6]	6.4% [4.5, 9.0]
	MT-Seed	78.4% [74.6, 81.8]	12.2% [9.6, 15.4]	9.4% [7.1, 12.4]
Bambara	Zero-Shot	94.0% [91.6, 95.8]	2.4% [1.4, 4.1]	3.6% [2.3, 5.6]
	MT-Seed	83.6% [80.1, 86.6]	8.0% [5.9, 10.7]	8.4% [6.3, 11.1]
Fulfulde	Zero-Shot	91.8% [89.0, 93.9]	2.8% [1.6, 4.7]	5.4% [3.6, 7.9]
	MT-Seed	80.8% [77.0, 84.1]	11.0% [8.6, 14.1]	8.2% [6.1, 11.0]

Table 3: Results of the human preferences experiment. Human evaluation and MT-Seed were carried out with our best-performing model (Llama-3.1-8B with InstructLR).

ings. Native speakers showed a strong preference for InstructLR outputs over baselines in 78-94% of comparison. Also, the model trained on ZarmaInstruct achieves a 41.2% exact match score on a zero-shot NER task, a considerable improvement over the baselines. These findings suggests the datasets from InstructLR can serve as foundational resources for real-world applications.

In sum, these findings position InstructLR as an efficient and economically friendly framework in creating multi-domain instructions dataset for LRLs, and thus opening more research opportunities for these languages.

6 Conclusion & Future Work

This paper introduces InstructLR, a framework for generating high-quality instruction datasets for low-resource languages. Our work addresses the critical data gap that limits LLM performance in these languages. Using this pipeline, we created three 50k-scale benchmarks: ZarmaInstruct-50k, BambaraInstruct-50k, and FulfuldeInstruct-50k. The framework’s dual-layer quality filter, which combines RAG-based checking with human validation, effectively corrects errors while managing costs. Our experiments demonstrate that fine-tuning on these datasets enables open-source models to follow instructions in the target languages, showing significant improvements over both zero-shot and machine-translation baselines.

Future work will focus on several key areas. We aim to reduce the framework’s dependency on commercial LLMs to increase its accessibility. Also, we plan to extend InstructLR to 12 new languages, including those with different high-resource contact languages and those with no existing LLM coverage. Finally, we will work to develop more sophisticated automated quality assessment techniques. These enhancements will target complex grammatical rules and aim to improve the detection

Table 4: Human quality ratings and downstream NER. The NER experiment was conducted with our best model from the automatic evaluation: (**Llama-3.1-8B with InstructLR**) (see Table 2)

(a) Human quality ratings.				(b) NER (exact match & macro-F1).				
Lang	Model	Fluency \uparrow	Correctness \uparrow	Relevance \uparrow	Lang	Model	Exact Match \uparrow	Macro-F1 \uparrow
Zarma	Zero-shot	1.2 [1.1, 1.3]	1.1 [1.0, 1.2]	1.3 [1.2, 1.4]	Zarma	Zero-shot	9.8% [7.2, 12.7]	21.4 [18.1, 24.7]
	MT-Seed	2.3 [2.2, 2.5]	2.1 [2.0, 2.3]	2.6 [2.5, 2.7]		MT-Seed	27.6% [23.6, 31.8]	49.3 [45.2, 53.2]
	InstructLR	3.3 [3.2, 3.4]	2.9 [2.8, 3.1]	3.7 [3.6, 3.8]		InstructLR	41.2% [36.8, 45.7]	63.8 [60.1, 67.2]
Bambara	Zero-shot	1.4 [1.3, 1.5]	1.2 [1.1, 1.3]	1.3 [1.2, 1.4]	Bambara	Zero-shot	13.0% [10.1, 16.4]	27.2 [23.9, 30.6]
	MT-Seed	3.0 [2.9, 3.2]	2.7 [2.6, 2.9]	3.3 [3.2, 3.4]		MT-Seed	36.8% [32.5, 41.3]	57.9 [54.2, 61.5]
	InstructLR	4.2 [4.0, 4.5]	4.0 [3.9, 4.1]	4.2 [4.1, 4.3]		InstructLR	54.4% [50.0, 58.7]	71.6 [68.4, 74.7]
Fulfulde	Zero-shot	1.3 [1.2, 1.4]	1.1 [1.0, 1.2]	1.2 [1.1, 1.3]	Fulfulde	Zero-shot	12.2% [9.4, 15.6]	25.9 [22.6, 29.3]
	MT-Seed	2.8 [2.7, 3.0]	2.5 [2.4, 2.7]	3.1 [3.0, 3.2]		MT-Seed	33.0% [29.0, 37.3]	55.2 [51.3, 58.9]
	InstructLR	4.1 [4.0, 4.2]	3.8 [3.7, 4.0]	4.0 [3.9, 4.1]		InstructLR	50.6% [46.2, 55.0]	69.1 [65.8, 72.2]

of factual or cultural inconsistencies.

References

- Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. [Phi-4 technical report](#). *Preprint*, arXiv:2412.08905.
- David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D’souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, and 42 others. 2021. [Masakhaner: Named entity recognition for african languages](#). *Preprint*, arXiv:2103.11811.
- Tuka Alhanai, Adam Kasumovic, Mohammad Ghassemi, Aven Zitzelberger, Jessica Lundin, and Guillaume Chabot-Couture. 2024. [Bridging the gap: Enhancing llm performance for low-resource african languages with new benchmarks, fine-tuning, and cultural adjustments](#). *Preprint*, arXiv:2412.12417.
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024. [Monolingual or multilingual instruction tuning: Which makes a better alpaca](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1347–1356, St. Julian’s, Malta. Association for Computational Linguistics.
- Michael Han Daniel Han and Unsloth team. 2023. [Unsloth](#).
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. [Deepseek-v3 technical report](#). *Preprint*, arXiv:2412.19437.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [Qlora: Efficient finetuning of quantized llms](#). *Preprint*, arXiv:2305.14314.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. [The faiss library](#). *Preprint*, arXiv:2401.08281.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Xu Guo and Yiqiang Chen. 2024. [Generative ai for synthetic data generation: Methods, challenges and the future](#). *Preprint*, arXiv:2403.04190.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). *Preprint*, arXiv:2009.03300.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Mamadou Keita, Elysabhete Ibrahim, Habibatu Alfari, and Christopher Homan. 2024. [Feriji: A French-Zarma parallel corpus, glossary & translator](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–9, Bangkok, Thailand. Association for Computational Linguistics.

- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. **Madlad-400: A multilingual and document-level large audited dataset**. *Preprint*, arXiv:2309.04662.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Youmi Ma, Sakae Mizuki, Kazuki Fujii, Taishi Nakamura, Masanari Ohi, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Koki Maeda, Kakeru Hattori, Takumi Okamoto, Shigeki Ishida, Rio Yokota, Hiroya Takamura, and Naoaki Okazaki. 2025. **Building instruction-tuning datasets from human-written instructions with open-weight large language models**. *Preprint*, arXiv:2503.23714.
- Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. **Crosslingual generalization through multitask finetuning**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. **Gpt-4 technical report**. *Preprint*, arXiv:2303.08774.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Victor Sanh, Albert Webson, Colin Raffel, Stephen H. Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, and 22 others. 2022. **Multitask prompted training enables zero-shot task generalization**. *Preprint*, arXiv:2110.08207.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. **Stanford alpaca: An instruction-following llama model**. https://github.com/tatsu-lab/stanford_alpaca.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2024. **Gemini: A family of highly capable multimodal models**. *Preprint*, arXiv:2312.11805.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. **Gemma 3 technical report**. *Preprint*, arXiv:2503.19786.
- Kosei Uemura, Mahe Chen, Alex Pejovic, Chika Maduabuchi, Yifei Sun, and En-Shiun Annie Lee. 2024. **AfriInstruct: Instruction tuning of African languages for diverse tasks**. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13571–13585, Miami, Florida, USA. Association for Computational Linguistics.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. **Superglue: A stickier benchmark for general-purpose language understanding systems**. *Preprint*, arXiv:1905.00537.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. **Glue: A multi-task benchmark and analysis platform for natural language understanding**. *Preprint*, arXiv:1804.07461.
- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khoshabi, and Hannaneh Hajishirzi. 2023. **Self-instruct: Aligning language models with self-generated instructions**. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhui Chen. 2024. **Mmlu-pro: A more robust and challenging multi-task language understanding benchmark**. *Preprint*, arXiv:2406.01574.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. **Finetuned language models are zero-shot learners**. *Preprint*, arXiv:2109.01652.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. **Multilingual machine translation with**

large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

A Limitations

While InstructLR provides a robust framework for generating instruction datasets for LRLs, we acknowledge several limitations that impact its current effectiveness and scalability.

First, our framework currently relies on commercial LLMs for the initial draft generation, as these are often the only models with even a basic capability in many LRLs. This dependency introduces a cost factor that may be a challenge for researchers. Additionally, the InstructLR pipeline requires that the target LRL is at least minimally covered by an existing LLM. For languages with no current LLM support, the framework is inapplicable without significant adaptations.

Another limitation concerns the demonstrated scope of our framework. While we successfully applied it to three distinct West African languages, all three share French as a high-resource contact language. Consequently, further work is needed to validate its effectiveness for languages with different features or writing systems.

The scope of our quality assessment also presents a limitation. The automated quality assessment and human validation layers focus primarily on grammatical correctness and fluency, not on factual accuracy. Errors in the source LLM’s knowledge could therefore propagate into the final datasets. Furthermore, the reliance on French seed instructions, even on general topics inspired by MMLU, could introduce a cultural bias toward Western or francophone perspectives. Finally, our human validation relies on small annotator pools, which may not capture the full dialectal variation within the language communities.

B Related Work

Instruction tuning for Low-Resource Languages

Instruction tuning aligns LLMs with user needs by fine-tuning on task instruction data (Ma et al., 2025). Benchmarks—like FLAN, T0, etc—provide instruction datasets for LLMs to be trained on (Wei et al., 2022; Sanh et al., 2022; Wang et al., 2024; Hendrycks et al., 2021; Wang et al., 2020, 2019). However, these advances are centered on higher-resource languages—leaving LRLs with marginal

coverage. This is particularly true for many African languages, due to the lack of task-specific data and the affordability of creating data. Recent work addresses this gap through multilingual instruction tuning. Muennighoff et al. (2023) showed that fine-tuning a multilingual model on English tasks can enable zero-shot instruction-following in other languages present only in pre-training data. Moreover, adding a small portion of multilingual data during fine-tuning yields further improvements on the target-language tasks (Muennighoff et al., 2023). Nevertheless, “severely” LRLs—particularly African languages—still lag behind, as the current benchmarks cover only relatively better-represented languages—such as Hausa or Swahili.

Several works provide instruction data specifically for African languages. For instance, Masakhane has produced datasets for tasks such as machine translation (MT) or named entity recognition (e.g., MasakhaNER supports 10 African languages (Adelani et al., 2021)). AfriInstruct integrates translation data (FLORES, MAFAND-MT for 16 languages), topic classification and summarization data (XL-Sum, etc), sentiment corpora (AfriSenti and NollySenti), and Masakhane benchmarks (NER, POS tagging) into a unified training set (Uemura et al., 2024). Yet, these are limited to a few African languages—not even half of the total languages present in the region. Our work addresses the need for scale-appropriate tools for building instruction datasets for LRL.

Synthetic Instructions Due to the lack of human-written instruction data in most LRLs, a popular alternative is synthetic instruction generation. The self-instruct framework proposed by Wang et al. (2023) demonstrated that one can create an instruction dataset by prompting a language model with a handful of seed tasks to produce new instruction–response pairs. Following this, researchers have explored extending self-instruction to other languages. For example, Chen et al. (2024) translates the Alpaca English instructions into eight languages to compare multilingual vs. monolingual instruction tuning, and finds that even machine-translated instructions can provide cross-lingual benefits.

Also, it is important to mention the recent trend of using LLMs as annotators to reduce the cost of creating LRL data. For instance, Alhanai et al. (2024) leverage GPT-4o to automate parts of their

quality assessment process by having the model score generated text on metrics such as fluency and factual consistency.

However, purely synthetic data approaches are not fully reliable in terms of quality. Model-generated instructions may contain errors, non-fluent phrasing, or cultural inappropriateness in the target LRL. Recent work highlights the need for careful control of LLM-synthesized data using strategies like rewriting the generated instructions or having multiple LLMs chat with each other to stimulate feedback dialog (Ma et al., 2025). Despite these solutions, this limitation still remains, and proves the need of human-in-the-loop approaches within these processes.

InstructLR leverages these previous approaches and combines their strengths into a unified framework for generating quality synthetic instruction data for LRLs with minimal human intervention. While self-instruction and translation approaches offer scalability, they often lack quality for LRLs. InstructLR addresses this limitation by integrating a robust LRL-aware dual-layer quality filtering process that includes RAG-based checks and human-in-the-loop validation to ensure higher fidelity and fluency.

C RAG-Based Checker Details

In this section, we provide an overview of the Retrieval-Augmented Generation (RAG) checker developed for quality assessment of Zarma text. Our system combines dense retrieval with language-model analysis to detect and correct grammatical errors and to improve textual fluency.

C.1 System Architecture

The RAG checker integrates two primary components: a retrieval module and a generation/assessment module. The retrieval module uses a knowledge base comprising 3,000 clean Zarma sentences from the Feriji dataset (Keita et al., 2024), 20 Zarma grammar rules with examples, and bilingual glossaries. These resources were encoded with a FAISS dense index (Douze et al., 2025) for efficient semantic retrieval.

For the generation component, we used the Gemini 2.0 Flash model, selected for its understanding of Zarma linguistic structures. This model processes retrieved contextual information alongside input text to perform grammar checking and correction.

Metric	Value
GLEU Score	0.8978
M ² Score	0.3400
False-Positive Rate	0.0
Fluency Assessment Score	4.3/5

Table 5: Performance metrics of the RAG-based checker on 300 Zarma test sentences

The system operates through the following workflow:

1. Input text is analyzed to identify potential error patterns.
2. Relevant grammar rules, example sentences, and vocabulary entries are retrieved from the knowledge base.
3. Retrieved context is incorporated into a prompt that guides the LLM to analyze and, if necessary, correct the text.
4. The system produces a structured assessment, including error identification and correction suggestions.

Our prompt design was important to ensure reliable performance. The prompt included instructions for recognizing proper nouns, maintaining linguistic coherence, and providing explicit reasoning for any corrections.

C.2 Evaluation Protocols

To evaluate the RAG checker, we designed a controlled test set of 300 Zarma sentences. The test set comprised 200 sentences with injected grammatical errors, created by prompting the DeepSeek v3 (DeepSeek-AI et al., 2025) LLM to break specific Zarma grammar rules, and 100 unaltered sentences that served as a gold standard for measuring false-positive rates. Each sentence was processed through the RAG analyzer, and the system’s assessments and corrections were compared with the gold references.

C.3 Evaluation Results

Table 5 presents the quantitative results of the controlled test. The average GLEU score (0.8978) reflects close n -gram alignment with the gold corrections. The M² accuracy of 0.3400 indicates that at least one suggestion matched the gold correction exactly for 34 % of the error sentences. No false

positives were recorded across the 100 correct sentences. In addition, 2 native Zarma speakers rated the outputs’ fluency at 4.3/5.

C.4 Prompt Configuration

The checker uses the following core prompt:

```

RAG ANALYZER PROMPT (EVALUATION CONFIGURATION)

You are a Zarma language expert. Analyze this potentially corrupted
Zarma sentence: "{sentence}"
Rely primarily on your expertise in Zarma grammar and meaning.
Recognize proper nouns unless contradicted by the glossary.
Use the grammar check and glossary below as supplementary aids.

INPUT DATA:
Grammar check results: {grammar_check}
Glossary information: {glossary_info}

OUTPUT FORMAT:
Provide the analysis in this format:
Is the sentence correct? [Yes/No]
Reason for Incorrectness (if applicable): [Brief reason]
Corrections (if incorrect):
  Option 1: [Corrected sentence with explanation]
  Option 2: [Corrected sentence with explanation]
  Option 3: [Corrected sentence with explanation]

```

Figure 2: Prompt used for RAG-based analysis and evaluation.

C.5 Example Analysis

```

Sentence analyzed: "Demain, a koy Niamey"
Grammar status: Correct (basic syntax, with caveats)

WORD BREAKDOWN:
Demain: Adverb, 'tomorrow' (French loanword)
a: 3rd-person singular pronoun, 'she/he/it'
koy: Verb, 'to go'
Niamey: Proper noun, city name

LINGUISTIC INSIGHT:
Word order: Adheres to Zarma SVO, initial adverbs allowed.
Tense: Lacks future marker "ga", implying habitual / near-future
action.
Context: Suggests "Tomorrow, she/he goes to Niamey"; "Demain" shows
code-switching.

CORRECTNESS ASSESSMENT:
Is the sentence correct? No
Reason: Missing future marker for "tomorrow"; "Demain" is
non-standard.

CORRECTIONS:
Option 1: Suba, a ga koy Niamey
Option 2: Suba, a koy Niamey
Option 3: Demain, a ga koy Niamey

Context sources (RAG retrieval):
Demain: French "demain", Zarma "suba"
a: French "elle", Zarma "a"
koy: French "aller", Zarma "koy"

```

Figure 3: Example of RAG analysis output for a single sentence.

D Generalizability: Adapting InstructLR to Bambara and Fulfulde

To validate the adaptability and scalability of InstructLR across different languages, we applied the framework to two additional West African languages: Bambara and Fulfulde.

Experimental Setup

For these experiments, we maintained the core pipeline structure used in the Zarma implementation. We generated **50,000** instruction-response pairs for both Bambara and Fulfulde using Gemini 2.5 Pro, the same model used for Zarma, with instructions spread randomly across the 20 topics. The objective was to evaluate whether the framework could transfer to other LRLs with minimal modifications.

To assess the raw output quality and better understand the necessity of the automated filtering stage, we implemented a simplified version of the pipeline by excluding the dual-layer quality filtering mechanism. Instead, we provided a random sample of 300 draft instruction-response pairs for each language to native speakers for manual quality assessment.

Evaluation Results

For **Bambara**, the native speaker evaluation revealed that approximately 26% of samples had minor fluency problems. These issues did not significantly impact comprehension but indicated the need for better phrasing. A more significant problem was the detection of hallucinated words in 2% of samples—one instance with a **Hindi** word and another containing a **Russian** word. Despite these issues, the remaining 72% of the samples were considered correct and understandable.

For **Fulfulde**, the evaluation showed a similar pattern, with approximately 17% of samples containing fluency errors and 1% containing hallucinated words. The errors in Fulfulde often related to its complex noun class system—something that our RAG checker could handle.

For both languages, evaluators noted that the content was easily accessible to bilingual speakers. This accessibility stems from the framework’s approach to technical terminology, which remained unchanged or was adapted from French. While this ensures comprehension for bilingual speakers, monolingual speakers might face challenges with these technical concepts.

These scaled experiments with Bambara and Fulfulde demonstrate that the core instruction-response generation component of InstructLR transfers well across linguistically diverse LRLs. The presence of fluency issues and hallucinations underscores the importance of the dual-layer quality filtering approach to produce high-fidelity datasets

at scale.

E Annotator Protocol and Quality Assurance

The integrity of the final datasets relies partially on the quality and consistency of the human validation layer. To ensure a high standard of accuracy, we designed and implemented a structured protocol for annotator recruitment, training, and workflow management. This section provides a detailed account of that process.

E.1 Recruitment and Training

We recruited a team of native speakers for each target language. The primary validation effort for **ZarmaInstruct-50k** was conducted by a team of five annotators. For the initial quality assessments of Bambara and Fulfulde, we worked with two native speakers for each language. All participants are graduate students with a formal background in Computer Science and are fluent in both their native language and French. While none had prior formal experience in linguistic annotation, their technical background facilitated a quick adoption of the structured task requirements.

Before starting the main annotation task, all participants underwent a mandatory 40-minute training session. The session covered:

1. **Project Goals:** An overview of the project’s objective to create high-quality instruction datasets and the role of human validation in correcting the nuanced errors that automated systems miss.
2. **Tooling:** A practical walkthrough of the annotation interface, which was implemented in Google Sheets for its accessibility and real-time collaboration features.
3. **Linguistic Guidelines:** A detailed review of the annotation guidelines (see Section E.3), with a focus on distinguishing between different error types.

Following the training, annotators participated in a calibration phase. During this phase, all annotators independently evaluated a common set of 50 drafts. Afterward, the team convened to discuss their decisions and resolve any disagreements.

E.2 Annotation Workflow and Tooling

The annotation task was managed entirely within a shared Google Sheets environment. Each language had a dedicated workbook, and drafts were assigned to annotators in batches of 200. The sheet was structured with the following columns to create a clear and efficient workflow:

- **draft_id:** A unique identifier for each instruction-response pair.
- **instruction_lrl:** The original, uncorrected instruction in the target LRL, as generated by the LLM. This field was locked.
- **response_lrl:** The original, uncorrected response in the target LRL. This field was locked.
- **rag_status:** The status assigned by the automated checker (e.g., 'top_priority', 'low_priority').
- **is_correct:** A dropdown menu with two options ('Yes', 'No'). Annotators selected 'Yes' if the draft was entirely free of errors.
- **corrected_instruction:** An editable field where the annotator would provide the corrected version of the instruction, if necessary.
- **corrected_response:** An editable field for the corrected version of the response.
- **error_category:** A dropdown menu with predefined error categories (e.g., 'Fluency', 'Suffix Misuse', 'Tense Inconsistency', 'Orthography'). This structured data was essential for our error analysis.
- **comments:** An optional text field for the annotator to leave notes about ambiguous cases or complex corrections.

Annotators were instructed to first assess the draft and set the `is_correct` flag. If they selected 'No', they were then required to provide corrections in the corresponding 'corrected_' fields and select the primary error category.

E.3 Annotation Guidelines

To maintain consistency, all annotators adhered to a defined set of guidelines:

Table 6: **BambaraInstruct-50k Dataset Statistics.**

Metric	Value	% or Average
<i>Instruction Characteristics</i>		
Instructions with 1–10 tokens	1,053	2.11%
Instructions with 11–20 tokens	29,966	59.93%
Instructions with >20 tokens	18,981	37.96%
<i>Response Characteristics</i>		
Responses with <50 tokens	28,346	56.69%
Responses with 50–100 tokens	21,654	43.31%
Instructions with CoT reasoning	12,500	25.00%
<i>Instruction Type Distribution</i>		
Open-ended questions	41,953	83.91%
Definition requests	66	0.13%
Explanation tasks	5,936	11.87%
List generation tasks	2,045	4.09%

Table 7: **FulfuldeInstruct-50k Dataset Statistics.**

Metric	Value	% or Average
<i>Instruction Characteristics</i>		
Instructions with 1–10 tokens	4,390	8.78%
Instructions with 11–20 tokens	31,273	62.55%
Instructions with >20 tokens	14,337	28.67%
<i>Response Characteristics</i>		
Responses with <50 tokens	42,786	85.57%
Responses with 50–100 tokens	7,214	14.43%
Instructions with CoT reasoning	12,500	25.00%
<i>Instruction Type Distribution</i>		
Open-ended questions	39,765	79.53%
Definition requests	219	0.44%
Explanation tasks	7,431	14.86%
List generation tasks	2,585	5.17%

1. **Preserve Semantic Intent:** The primary rule was to correct linguistic errors without altering the core meaning or intent of the original French instruction. The goal was to fix the language, not the content.
2. **Prioritize Fluency and Naturalness:** Corrections should result in text that sounds natural to a native speaker. This often involved rephrasing sentences that were grammatically correct but idiomatically awkward due to literal translation.
3. **Correct All Linguistic Errors:** Annotators were tasked with identifying and fixing all grammatical, orthographic (spelling), and syntactic errors. This included issues with tense, noun-verb agreement, and the misuse of function words or suffixes.

4. **Ensure Consistent Handling of Loanwords:** Annotators followed the same rules provided to the LLM: technical terms from French were to be preserved, and other non-translatable words were to be rendered using phonetic adaptation.

E.4 Common Error Categories and Correction Examples

During the human validation phase, several recurrent error patterns emerged. Table 8 provides illustrative examples of these common errors and the corrections applied by the annotators for the Zarma language.

E.5 Inter-Annotator Agreement (IAA)

To validate the consistency of our annotation process and the clarity of our guidelines, we measured Inter-Annotator Agreement (IAA). We calculated

Table 8: Examples of Common Errors and Applied Corrections in Zarma.

Error Category	Erroneous Draft Example	Corrected Version	Rationale
Suffix Misuse	<i>Ay na hansu di.</i> (I saw dog.)	<i>Ay na hanso di.</i> (I saw the dog.)	The draft was missing the definite article suffix '-o'. The correction adds the suffix to make the noun 'hansu' (dog) definite, which is required by the context.
Tense Inconsistency	<i>Suba, a koy Niamey.</i> (Tomorrow, he/she went to Niamey.)	<i>Suba, a ga koy Niamey.</i> (Tomorrow, he/she will go to Niamey.)	The adverb 'Suba' (tomorrow) establishes a future context, but the verb lacks the future tense marker 'ga'. The correction inserts the marker to ensure grammatical consistency.
Wrong Phrasing (Fluency)	<i>Boro fo kay ga ti alfa go no.</i> (A person who is a teacher is there.)	<i>Alfa fo go no.</i> (A teacher is there.)	The original phrasing is a literal, word-for-word translation (calque) of the French "Une personne qui est un enseignant...". The corrected version is more concise and idiomatically natural in Zarma.
Orthography	<i>Iri ga barma te.</i> (We will do work .)	<i>Iri ga barna te.</i> (We will do work .)	The word for "work" was misspelled. The correction applies the standard orthography for 'barna'.

Krippendorff’s Alpha (α). For the Zarma dataset, a randomly selected sample of 351 drafts was annotated by all five annotators. For Bambara and Fulfulde, a smaller sample of 50 drafts was cross-annotated to validate the initial quality assessment task.

The results, presented in Table 9, show a high level of agreement for the primary Zarma annotation task and agreement for the initial assessments of Bambara and Fulfulde.

The pretty high alpha score for Zarma ($\alpha = 0.793$) indicates that the guidelines were effective and the annotators applied them. An analysis of disagreements revealed two primary sources:

- **Subjectivity in Fluency:** The most frequent source of disagreement arose from the subjective nature of fluency. One annotator might accept a phrasing as adequate, while another would suggest an alternative phrasing.
- **Dialectal Variation:** Minor disagreements occasionally rose from regional variations in vocabulary or preferred sentence structures.

In all cases of disagreement, the final version included in the dataset was determined through a majority vote. If no majority existed, a final decision was made by the lead author in consultation with the annotators.

F Cost Comparison

To quantify the economic efficiency of our framework, we provide a detailed cost comparison for building a **50,000-pair** LRL instruction dataset under three distinct scenarios: *LLM Only* (*No*

QC), *Full Human Correction*, and our proposed *InstructLR (RAG + Human)* pipeline. The analysis, summarized in Figure 4, covers both commercial API models and self-hosted open-source models, factoring in their per-token costs and estimated baseline error rates—the proportion of generated pairs requiring correction before any filtering.

Our cost model is based on the following up-to-date estimates:

- **LLM Costs:** We use an average of 75 tokens per instruction-response pair, totaling approximately 3.75 million tokens for the entire dataset. Commercial API prices are estimated at **\$12/1M tokens for Gemini 2.5 Pro** and **\$10/1M tokens for GPT-4o**. Self-hosted open-source models have a negligible compute cost, estimated at under \$0.01/1M tokens on a single consumer GPU.
- **Human Annotation Cost:** We assume a professional annotator can review and correct a generated pair at a cost of **\$0.40 per pair**. This rate was chosen based on similar study conducted in the past.
- **Baseline Error Rates:** Based on our initial experiments, we use the following error rates for raw generated drafts: Gemini 2.5 Pro (15%), DeepSeek-V3 (25%), GPT-4o (70%), and Llama-3-8B (95%).

The results show cost differences driven primarily by the human labor required. In a **Full Human Correction** scenario, every one of the 50,000 drafts is reviewed. This fixes the human labor cost at

Table 9: **Inter-Annotator Agreement Scores**

Language	Annotation Task	Sample Size	Krippendorff’s Alpha (α)
Zarma	Full Error Correction & Categorization	351	0.793
Bambara	Initial Quality Assessment (Correct/Incorrect)	50	0.821
Fulfulde	Initial Quality Assessment (Correct/Incorrect)	50	0.637

a substantial **\$20,000** (50,000 pairs \times \$0.40/pair) which makes the initial LLM API cost (**\$45** for Gemini) almost irrelevant to the total project budget. This high cost makes large-scale dataset creation “VERY CHALLENGING” for many research teams.

The **InstructLR pipeline** aims to address this challenge. Our dual-layer filtering process reduces the number of pairs requiring human review by approximately 88%, meaning validators only need to inspect the 6,000 pairs flagged as “top priority” or corrected by the RAG system. This slashes the human validation cost from \$20,000 to just **\$2,400** (6,000 pairs \times \$0.40/pair).

This efficiency gain has several implications. For a high-performing commercial model like Gemini 2.5 Pro, InstructLR reduces the total project cost from \$20,045 (Full Correction) to **\$2,445**—a saving of nearly 88%. The framework makes even models with very high error rates economically viable; a self-hosted Llama-3-8B model, despite its 95% error rate, can be used to produce a high-quality dataset for a total cost of approximately \$2,400, as the automated RAG filter handles the vast majority of errors.

These results highlight that the “primary” value of InstructLR lies in its targeted reduction of human labor. By merging scalable LLM generation with an efficient, automated quality filter, our framework makes the creation of large-scale, high-quality instruction datasets for LRLs financially practical.

G Zarma Grammar Rules

We drafted the rules below based on linguistic documentation and observations from multiple sources. The rules are not limited to these ones; however, this constitutes a baseline for future work.

Rule 1: Pronouns — Personal Pronouns

Personal pronouns in Zarma are invariable across nominative, objective, and possessive cases.

- ay — I, me, my
- ni — you, your (singular)
- a (nga) — he, she, it; his, her, its
- iri (ir) — we, us, our
- araj — you (plural), your
- i (ngey, ey) — they, them, their

Rule 2: Pronouns — Demonstrative Pronouns

Demonstrative pronouns indicate specific items; a *din* suffix can be added to nouns for specificity.

- wo — this, that
- wey — these, those

Rule 3: Pronouns — Indefinite Pronouns

Indefinite pronouns refer to non-specific entities.

- boro — someone, one (person)
- hay kulu — everything
- hay fo — something

Rule 4: Nouns — Definite Article

Definite articles are expressed by adding “a” or “o” to the noun based on its ending.

Patterns:

- Ending “a”: add “a” (e.g. zanka \rightarrow zankaa); exceptions: pre-1999 texts may not change.
- Ending “o”: change to “a” or add “a” (e.g. wayboro \rightarrow waybora).

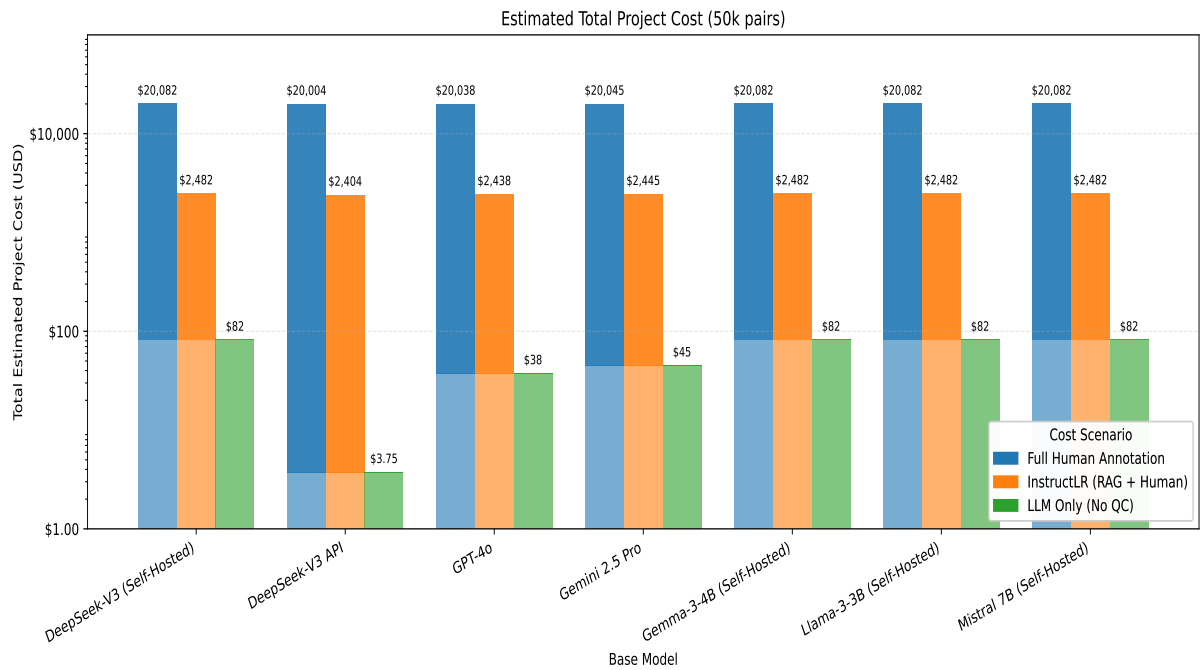


Figure 4: **Estimated total project cost** for producing 50,000 instruction–response pairs under three quality–control scenarios. Each bar shows the combined LLM compute/API cost and any required human annotation.

- Ending “ko”: change to “kwa” (e.g. darbayko → darbaykwa).
- Ending “e, i, u, consonant”: change to “o” or add “o” (e.g. wande → wando).
- Ending “ay”: change “ay” to “a” or add “o” (e.g. farkay → farka or farkayo).

Examples:

- zanka → zankaa — a child → the child
- wayboro → waybora — a woman → the woman
- darbayko → darbaykwa — a fisherman → the fisherman
- hansı → hanson — a dog → the dog
- farkay → farka — a donkey → the donkey

Rule 5: Nouns — Definite Plural

Definite plural is formed by replacing the definite singular vowel with “ey”.

- Replace final vowel with “ey” (e.g. zankaa → zankey).
- zankaa → zankey — the child → the children
- hanson → hansey — the dog → the dogs
- farka → farkey — the donkey → the donkeys

Rule 6: Nouns — Indefinite Article

No explicit indefinite article; “fo” (one) is used to specify “a certain” or “one”.

- Add “fo” after noun for specificity (e.g. musu → musu fo).
- musu — a cat
- musu fo — a (certain) cat, one cat

Rule 7: Nouns — Gender

No grammatical gender; specific words indicate male/female for living beings.

- alboro — man
- wayboro — woman

Rule 8: Verbs — Completed Action (Past Tense)

Verbs without auxiliaries indicate completed actions (past tense).

- Subject + Verb (e.g. ay neera).
- ay neera — I sold
- a neera — he/she sold
- zankaa kani — the child went to bed

Rule 9: Verbs — Uncompleted Action (Future Tense)

Future tense uses the auxiliary “ga” before the verb.

- Subject + ga + Verb (e.g. ay ga neera).
- ay ga neera — I will sell
- i ga neera — they will sell

Rule 10: Verbs — Continuous Aspect

Continuous aspect uses “go no ga” before the verb for ongoing actions.

- Subject + go no ga + Verb (e.g. ay go no ga neera).
- ay go no ga neera — I am selling
- a go no ga neera — he/she is selling

Rule 11: Verbs — Subjunctive

Subjunctive uses “ma” to indicate possible actions.

- Subject + ma + Verb (e.g. ay ma neera).
- ay ma neera — I should sell
- ni ma neera — you should sell

Rule 12: Verbs — Imperative

Imperative uses “ma” or ‘wa” before the verb, or just the verb alone.

- Ma/Wa + Verb or Verb alone (e.g. Ma haŋ or Haŋ).
- Haŋ! — Drink!
- Ma haŋ! — Drink!
- Araŋ ma di! — You (plural) see!

Rule 13: Verbs — To Be

The verb “to be” varies by context: “go”, “ya ... no”, or “ga ti”.

- A go fu — He/she is at home
- Ay ya alfa no — I am a teacher
- Nga ga ti wayboro — She is a woman

Rule 14: Verbs — Irregular Verbs

Some verbs place objects unusually (e.g. direct object before verb without “na”).

- Ay di a — I saw him/her
- A ne ay se — He/she said to me

Rule 15: Adjectives — Qualifying Adjectives

Adjectives follow the noun they modify.

- fu beeri — a big house
- hansi kayna — a small dog

Rule 16: Sentence Structure — Basic Order

Basic sentence order is Subject–Verb–Object (SVO).

- Ay neera bari — I sold a horse

Rule 17: Sentence Structure — Direct Object

Direct object before the verb requires “na” in the past positive.

- Ay na bari neera — I sold a horse

Rule 18: Sentence Structure — Indirect Object

Indirect object is marked with “se” after the object.

- Ay no bari wayboro se — I gave a horse to the woman

Rule 19: Negation — Past Negative

Past negative uses “mana” after the subject.

- Ay mana neera — I did not sell

Rule 20: Negation — Present/Future Negative

Present/future negative uses “si” instead of “ga”.

- Ay si neera — I do not / will not sell

H Topics Selected

In this section, we provide the list of topics—and a short description for each—we used for dataset creation throughout this paper.

I Prompt Templates

In this section, we show all the different prompt templates used in the InstructLR framework.

I.1 Seed Instructions Prompt Template**I.2 Instruction–Response Prompt Template**

We fed the Gemini model with the prompt below to obtain an LRL instruction–response pair from a French input.

J Generated Datasets Snapshots

Table 10: List of the 20 topics used for dataset generation.

Topic	Description
General Knowledge	Includes basic factual information across diverse domains including geography, current events, etc. This category tests very basic knowledge that educated individuals are "expected" to possess.
Biology	Covers living organisms, their structures, functions, growth, evolution, etc.
Economics & Finance	Examines economic principles, financial systems, market mechanisms, etc.
Common Sense Reasoning	Focuses on understanding cause-and-effect relationships in familiar contexts.
History	Explores past events, civilizations, historical figures, their impact on contemporary society, etc.
Mathematics	Involves numerical computations, algebraic manipulations, geometric principles, and mathematical problem-solving.
Computer Science	Includes programming concepts, algorithms, data structures, software engineering, and computational thinking. It covers both theoretical computer science and practical programming applications.
Social Sciences & Psychology	Includes human behavior, mental processes, social interactions, and societal structures.
Adversarial Multi-step Reasoning	Challenges complex problem-solving abilities through multi-layered logical puzzles and sequential reasoning tasks.
Physics	Examines matter, energy, motion, forces, and their interactions in the physical universe.
Engineering	Focuses on the application of scientific and mathematical principles to design and build structures, machines, and systems.
Law & Ethics	Explores legal systems, ethical principles, moral reasoning, and jurisprudence.
Extra-difficult Reasoning	Presents highly challenging logical problems that require advanced cognitive abilities and creative problem-solving approaches.
Chemistry	Studies the composition, properties, and behavior of matter at the atomic and molecular level.
Medicine & Health	Encompasses medical knowledge, healthcare practices, disease prevention, diagnosis, and treatment approaches.
Business & Management	Addresses organizational management, strategic planning, leadership principles, and business operations.
Causal Reasoning	Tests understanding of cause-and-effect relationships, logical inference, and the ability to predict outcomes based on given conditions.
Sports	Covers athletic activities, rules, strategies, and sports-related knowledge including historical achievements and sporting culture.
Sentiment Analysis	Involves identifying and interpreting emotional tones, attitudes, and opinions expressed in text or speech.
Multi-sentence Comprehension	Assesses reading comprehension skills across multiple connected sentences, testing coherence understanding and information synthesis.

SEED INSTRUCTION GENERATION PROMPT

Domaine : {domain}

TASK:
GÉNÉREZ UNE SEULE CONSIGNE OU QUESTION EN FRANÇAIS, REPRÉSENTATIVE DE CE DOMAINE.
VOUS POUVEZ CHOISIR :

- QUESTION À CHOIX MULTIPLES (Options: A)..., B)... etc.)
- QUESTION VRAI/FAUX
- AFFIRMATION À COMPLÉTER
- DEMANDE DE LISTE (ex. : "Donnez x exemples de...")
- TÂCHE OUVERTE (CLASSIFICATION, RÉSUMÉ, EXPLICATION, EXEMPLE, ETC.)
- OU N'IMPORTE QUEL AUTRE STYLE.

CONTRAINTES :

1. RESTEZ EN 1 À 4 PHRASES.
2. NE DEMANDEZ PAS DE DESSIN, DE CHANT, DE GÉNÉRATION D'IMAGE, NI DE RECHERCHE SUR LE WEB.
3. UTILISEZ UN VERBE UNIQUE POUR ÉVITER LA RÉPÉTITION ET MAXIMISER LA DIVERSITÉ.
4. FOURNISSEZ UNE ENTRÉE RÉALISTE (<=150 MOTS).
5. L'ENTRÉE DOIT ÊTRE SPÉCIFIQUE, SUBSTANTIELLE ET FOURNIR UN CONTENU STIMULANT.
6. NE RÉPONDEZ PAS AUX INSTRUCTIONS OU QUESTIONS – LIMITEZ-VOUS JUSTE À L'INSTRUCTION OU À LA QUESTION.

OUTPUT FORMAT (JSON):
RENVOYEZ STRICTEMENT CE JSON :

```
{
  "instruction_fr": "<VOTRE INSTRUCTION>",
  "context_fr": "{domain}"
}
```

Figure 5: Prompt used for generating seed instructions from a specific domain.

LRL INSTRUCTION-RESPONSE GENERATION PROMPT

SYSTEM PREAMBLE:
Vous êtes un assistant IA expert dans la génération de paires instruction-réponse pour des langues à faibles ressources, spécifiquement pour le {target_language}. Votre tâche : (1) générer **instr_lrl**-la version de l'instruction en {target_language}; (2) générer **resp_lrl**-une réponse pertinente et grammaticalement correcte en {target_language}; (3) pour les sujets de raisonnement, générer **CoT_lrl**-une explication des étapes de raisonnement (max 200 mots); pour les autres sujets, **CoT_lrl** doit être "N/A".

CONTRAINTES:

1. LES MOTS TECHNIQUES (SCIENCE, MÉDECINE, ETC.) DOIVENT RESTER INCHANGÉS MAIS UTILISER LEUR VERSION FRANÇAISE.
2. SI UN MOT N'A PAS D'ÉQUIVALENT EN ZARMA, ÉCRIVEZ SA TRANSCRIPTION PHONÉTIQUE EN FRANÇAIS.
3. N'INVENTEZ PAS DE MOTS. SUIVEZ LES DIRECTIVES.
4. PAS DE TRADUCTION MOT À MOT.
5. LES RÉPONSES (**resp_lrl**) NE DOIVENT PAS DÉPASSER 100 MOTS.

USER REQUEST (JSON INPUT):

```
{
  "instruction_fr": "{provided_french_instruction}",
  "context_fr": "{provided_french_context}",
  "specific_guidelines": [
    "La instr_lrl DOIT être uniquement en {target_language}.",
    "Conserver noms propres et emprunts établis, transcrits phonétiquement.",
    "Clarté et grammaire irréprochables."
  ]
}
```

EXPECTED OUTPUT (JSONL):

```
{
  "instr_fr": "...", "instr_lrl": "...", "resp_lrl": "...", "CoT_lrl": "...", "lang": "{code}"
}
```

Figure 6: Prompt configuration for generating Low-Resource Language (LRL) instruction pairs.

Table 11: Snapshot of 20 instruction–response pairs from ZarmaInstruct-50k. **Snapshot of the Generated Drafts.** Columns represent: *instr_fr* = French instruction, *instr_lrl* = Instruction in target LRL, *resp_lrl* = Response in LRL, *CoT_lrl* = Chain of Thoughts in response generation in LRL (N/A for non-reasoning topic), *topic_fr* = Topic name in French, *lang* = Language code.

<i>instr_fr</i>	<i>instr_lrl</i>	<i>resp_lrl</i>	<i>CoT_lrl</i>	<i>topic_fr</i>	<i>lang</i>
Quelle est la capitale du Niger ?	Niamey wo di Niger kebal?	Niamey di Niger gaba kuruso.	N/A	Connaissances générales	dje
Calcule $7 + 5$.	7 nda 5 baani?	7 nda 5 ga baani 12.	N/A	Mathématiques	dje
Explique la loi de la gravitation.	Bay gravite kanun no?	Gravite kanun no: jano hinka deera wusu nda deera.	N/A	Physique	dje
Qu'est-ce qu'une molécule ?	Molekul ya mate no?	Molekul no, kaani zankakwey koyra ga.	N/A	Chimie	dje
Définis la photosynthèse.	Fotosintez ga bay?	Fotosintez no: hanci nda saa kaani ga bi kaba.	N/A	Biologie	dje
Qu'est-ce qu'un algorithme ?	Algorithmme ya mate no?	Algorithmme no, ti hakilay hinka teeri nda namay.	N/A	Informatique	dje
Explique le pont suspendu.	Bay zawo suru di?	Zawo suru no, suuru sila hinka ga hin ka ti zawo fa.	N/A	Ingénierie	dje
Définis l'anémie.	Anemia ya mate no?	Anemia no, doro cee ka si jiri bon.	N/A	Médecine & santé	dje
Que signifie l'inflation ?	Inflation ya hin ka ti?	Inflation no, sarey gooru ka koy boori.	N/A	Économie & finance	dje
Définis le stress.	Stress ya mate no?	Stress no, boro hakilay ka gaama.	N/A	Sciences sociales & psychologie	dje
Qu'est-ce que la loi civile ?	Loi civile ya hin ka ti?	Loi civile no, koyra nda boro farilla hinka.	N/A	Droit & éthique	dje
Qu'est-ce qu'un business plan ?	Business plan ya mate no?	Business plan no, cilogani nda foondiray han.	N/A	Commerce & gestion	dje
Combien de joueurs dans un match de foot ?	Futbol kura ga boro hinka?	Futbol kura ga boro 22.	N/A	Sports	dje
Donne le ton du texte "A ga nafa".	Kurun "A ga nafa" seerey hin?	Kurun no seerey bara kamma, positifu.	N/A	Analyse de sentiment	dje
Si l'eau chauffe, que se passe-t-il ?	So dii, ko moto?	So dii, a ga buburu.	a ga buburu wa; sababaa dii.	Raisonnement causal	dje
Lis ces deux phrases et dis le sujet.	Ay buburu; A koy. Suje di?	Suje di "Ay" nda "A".	N/A	Compréhension multi-phrases	dje
Pourquoi met-on un manteau en hiver ?	Kari wa, ko sabu?	Hima kura, kari ga ke boori.	Fanda kura, kari za daabani.	Raisonnement de sens commun	dje
Résous : $(2 \times 3) + 4$.	2×3 nda 4 baani?	2×3 ga 6; 6 nda 4 ga 10.	multitape: dabari nda daaba.	Raisonnement multi-étape adversarial	dje
Trouve le prochain nombre premier après 29.	29 kuma, numuru kuma surey?	Numuru kuma surey ga 31.	teste divisibilité; 31 si baani.	Raisonnement difficile	dje
En quelle année le Niger fut-il indépendant ?	Niger independansi ci hinka?	Niger independansi ci 1960.	N/A	Histoire	dje

Leveraging CoHere Multilingual Embeddings and Inverted Softmax Retrieval for Automatic Parallel Sentence Alignment in Low-Resource Languages

Auwal Abubakar Khalid
Bayero University, Kano
aka2000078.mcs@buk.edu.ng

Salisu Musa Borodo
Bayero University, Kano
salisuborodo@gmail.com

Amina Imam Abubakar
University of Abuja
amina.imam@uniabuja

Abstract

We present an improved method for automatic parallel sentence alignment in low-resource languages. We used CoHere multilingual embeddings and inverted softmax retrieval. Our technique achieved a higher F1-score of 78.30% on the MAFAND-MT test set, compared to the existing technique’s 54.75%. Precision and recall have shown similar performance. We assessed the quality of the extracted data by demonstrating that it outperforms the existing technique in terms of low-resource translation performance.

1 Introduction

Because of a lack of large parallel datasets, neural machine translation systems trained on low-resource languages produce subpar results (Fernando et al., 2023; Haddow et al., 2022). Parallel corpora continue to be required for effective training of machine translation systems (Althobaiti, 2021; Yousef et al., 2022; Kaufmann, 2012; Paetzold et al., 2017; Resnik and Smith, 2003). Signoroni and Rychlỳ (2023); Haddow et al. (2022) found that neural machine translation is less reliable for language pairs with minimal resources. Even when parallel data is accessible, it is usually of lower quality or obtained from highly specialized sources, such as IT documentation or religious literature (Jaworski et al., 2023; Haddow et al., 2022; Ling et al., 2016). As a result, it cannot be used alone to accurately train general-purpose translation systems. To address this issue, parallel corpora can be mined from the internet (Fernando et al., 2023; Ling et al., 2016; Resnik and Smith, 2003). There are several plausible parallel sentences online, especially on multilingual news and instructional websites (Zhao et al., 2021; Riesa and Marcu, 2012; Makazhanov et al., 2018). Parallel corpora are typically constructed using automated sentence alignment approaches due to time

and resource restrictions (Signoroni and Rychlỳ, 2023; Althobaiti, 2021; Hameed et al., 2016). Automatic sentence alignment is the technique of determining which sentences in a source text match to which sentences in a target text, allowing for the extraction of probable parallel sentences from big corpora (Chousa et al., 2020; Yousef et al., 2022; Brown et al., 1993). Several solutions have been presented. Multilingual sentence embedding-based methods have shown advantage in extending Natural Language Processing (NLP) tasks to a large number of languages, without the need to train a language-specific model (Signoroni and Rychlỳ, 2023; Heffernan et al., 2022; Chousa et al., 2020). Multilingual embeddings provide a universal foundation for sentence alignment that crosses linguistic boundaries (Artetxe and Schwenk, 2019). Using embedding-based methods, sentences in both languages can be represented in a single vector space so that sentences with semantic similarity are adjacent to each other in the vector space (Althobaiti, 2021). A recent study by Abdulmumin et al. (2023) shows that using closed-access CoHere multilingual embeddings resulted in a considerable improvement over earlier state-of-the-art LASER embeddings in parallel sentence alignment for low-resource languages. However, the authors implemented their alignment model using the standard nearest neighbor retrieval method. Although it is a simple and intuitive method for finding similar instances, standard nearest neighbor algorithm can only work well for relatively small datasets. It may suffer from hubness—where certain sentences tend to appear overly similar to many others in an embedding space (Dinu et al., 2015), and may not be effective in capturing complex relationships between sentences in different languages. Therefore, this research aims to propose an improved retrieval method—inverted softmax (Smith et al., 2017)—to enhance the alignment accuracy of the model, and consequently, improve translation quality.

2 Related Works

Research on automatic sentence alignment has evolved from early heuristic-based methods to modern neural approaches, especially in the context of low-resource languages. The methods can be broadly categorized into: (1) length and statistical-based approaches, (2) lexical and dictionary-based methods, (3) hybrid and alignment-tool frameworks, and (4) neural embedding-based approaches.

2.1 Length and Statistical-based Methods

Initial efforts relied heavily on sentence length as a proxy for alignment probability. Church (1993) proposed a character-length-based probabilistic model using dynamic programming. Brown et al. (1991) extended this with token counts and anchor points. Chen (1993); Papageorgiou et al. (1994) incorporated both sentence length and word identity. These methods assumed monotonic alignment and performed well in structured, clean data. Fung and McKeown (1994) introduced DK-Vec, targeting noisy parallel corpora using frequency and position heuristics. These early models laid the foundation for fast and language-independent alignment tools.

2.2 Lexical and Dictionary-based Methods

To address alignment ambiguity, lexical resources became central. McEnery et al. (1994) used approximate string matching, improving results from Kay and Roscheisen (1993). Hunalign (Varga et al., 2007) combined sentence-length statistics with bilingual dictionaries, refining alignments through iterative dictionary induction. Ma (2006) proposed Champollion, which weighted rare words more heavily for alignment but remained dependent on dictionary quality. Chen and Du (2003) tackled one-to-many word alignments in spoken corpora, while Resnik and Smith (2003) used web mining to extract parallel data. Other notable works include Melamed (2001) on idiom-aware alignment, and Semmar and Fluhr (2007) on cross-language information retrieval.

2.3 Hybrid and Tool-based Approaches

Several toolkits emerged combining statistical, linguistic, and rule-based methods. Deng and Byrne (2006) introduced MTTK, a language-independent toolkit for SMT training. JMaxAlign (Kaufmann, 2012) applied maximum entropy classifiers, while MASSAlign (Paetzold et al., 2017) targeted mono-

lingual simplification. Efforts also explored specific domains: Ohmori and Higashida (1999) for Japanese-English collocations, and Volk et al. (2008) for treebank alignment. Some methods incorporated alignment correction and visualization tools (Macdonald, 2001; Cardon and Grabar, 2019). Recently, SpanAlign (Chousa et al., 2020) applied integer linear programming for non-monotonic alignments, and Stodden and Kallmeyer (2022) developed TS-ANNO for simplified corpora.

2.4 Neural Embedding-based Approaches

Recent advances have seen a shift toward embedding based models. VecAlign (Thompson and Koehn, 2019) combined dynamic programming with LASER embeddings (Artetxe and Schwenk, 2019) to align low-resource pairs such as Sinhala-English and Nepali-English. While effective, it suffered from misalignment in distant sentences. Paragraph-level filtering was later introduced to reduce this issue, as applied in Vietnamese-Lao alignment. Fine-tuning multilingual sentence embeddings has further improved performance. Chimoto and Bassett (2022) enhanced zero-shot alignment for Luhya-Swahili by fine-tuning LaBSE on a small Luhya parallel corpus, boosting accuracy from 22% to 85% with cosine similarity filtering. Abdulmumin et al. (2023) demonstrated that Co-Here’s closed-access multilingual embeddings significantly outperformed LASER on Hausa-English. Their method improved downstream MT quality, though it relied on basic nearest-neighbor retrieval, which the authors acknowledged as a limitation. We therefore, propose a better retrieval technique—*inverted softmax* (Smith et al., 2017)—due to its ability to mitigate hubness and capture complex patterns between sentences in different languages

3 Methodology

3.1 Alignment Workflow

Given parallel documents in two languages—source (s) and target (t), the task is to match sentences that are translations of each other. The parallel documents, are the ones in two languages containing similar information. Our sentence alignment workflow is shown in figure 1. Sentences are aligned in three (3) steps:

- i. Sentence Vectorization: The first step is sentence representation. Each sentence in both the source (e.g., Hausa) and target (e.g., English) documents is mapped into a shared vec-

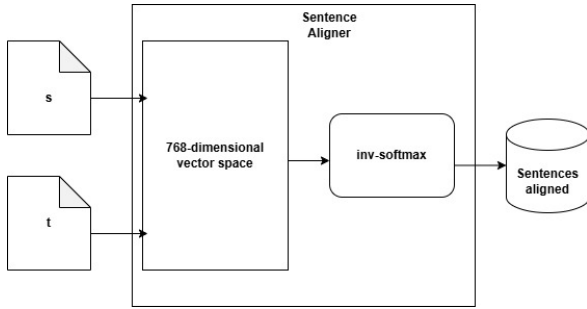


Figure 1: Sentence alignment workflow

tor space that captures their semantic features, using a pretrained multilingual sentence embedding model (CoHere multilingual embedding). This means that a Hausa sentence and its correct English translation should be close together in that space.

- ii. Similarity measure: For each sentence pair, the cosine similarity is computed next. This measures how close two vectors are in direction. Cosine similarity is high (near 1) if sentences are likely translations of each other.
- iii. Translation Retrieval: For each source sentence vector, retrieval algorithm (inverted softmax) searches for the most similar target sentence vector based on cosine similarity. It’s like saying: "For this Hausa sentence, which English sentence is closest in meaning according to their vector representations?"

3.2 CoHere Multilingual Embedding Model

The 768-dimensional CoHere multilingual embedding model was developed to support a number of tasks, including cross-lingual zero-shot content moderation, multilingual semantic search, and customer feedback compilation, in more than 100 languages, including Hausa. This model can only be accessed via an API, which requires authentication with an API key. This key is available for users to generate at their website ¹.

3.3 Retrieval Algorithms

Retrieval algorithms are a key component of automatic parallel sentence alignment systems. They determine how semantically similar sentences across languages are identified in a shared embedding space. In this work, we compare the standard nearest neighbor retrieval with our proposed inverted softmax retrieval, which aims to improve

¹<https://dashboard.cohere.com/api-keys>

alignment quality, especially in low-resource and noisy settings.

3.3.1 Standard Nearest Neighbor Retrieval

Nearest Neighbor (NN) retrieval is one of the simplest and most widely used methods for sentence alignment. For each source sentence embedding \mathbf{x}_i , the most similar target sentence \mathbf{y}_j is selected based on cosine similarity:

$$s(\mathbf{x}_i, \mathbf{y}_j) = \frac{\mathbf{x}_i \cdot \mathbf{y}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{y}_j\|}$$

$$\mathbf{y}_j^* = \arg \max_{\mathbf{y}_j} s(\mathbf{x}_i, \mathbf{y}_j)$$

This approach is efficient and easy to implement, which makes it a popular baseline in multilingual alignment systems. However, it has notable limitations. In high-dimensional embedding spaces, certain sentences—often short or generic ones—tend to appear overly similar to many others, a phenomenon known as *hubness*. As a result, these “hub” sentences are frequently retrieved even when they are not true translations. Moreover, the search is one-directional and does not account for similarity normalization across the dataset.

3.3.2 Inverted Softmax Retrieval

To address these issues, we replace the nearest neighbor search with **Inverted Softmax (ISF)** (Smith et al., 2017) retrieval, a probabilistic method that introduces normalization over source embeddings. This helps counteract the hubness effect and yields more reliable alignments. Given a source embedding \mathbf{x}_i and a target embedding \mathbf{y}_j , the probability that \mathbf{x}_i corresponds to \mathbf{y}_j is computed as:

$$P(\mathbf{x}_i | \mathbf{y}_j) = \frac{\exp(\beta \cdot s(\mathbf{x}_i, \mathbf{y}_j))}{\sum_k \exp(\beta \cdot s(\mathbf{x}_k, \mathbf{y}_j))}$$

where β is a temperature parameter controlling how sharply the similarities are weighted. Unlike the standard approach, ISF conditions on the target sentence by normalizing over all source embeddings. This means that a target sentence similar to many sources (a potential hub) receives lower relative scores, reducing its chance of being incorrectly aligned.

Advantages Over Nearest Neighbor

1. Reduces hubness: ISF penalizes overly generic targets that appear similar to many sources. **2. Improves precision:** It favors more distinctive and

contextually relevant sentence pairs. **3. Remains efficient:** Despite the added normalization, ISF can be implemented efficiently with matrix operations. Overall, the inverted softmax retrieval provides a more balanced and accurate alignment mechanism, particularly useful when working with low-resource or noisy parallel data.

4 Experiment

4.1 Datasets

4.1.1 Crawled Data

We used the 1,000 most recent news stories from the Premium Times News ^{2 3} website that had been crawled in both Hausa and English (Abdulmumin et al., 2023). To preprocess these data, we separated each collected document into a list of sentences using the Natural Language Tool Kit (NLTK) sentence tokenizer. The target and source files were then created by combining these sentences. Following tokenization using the NLTK’s word tokenizer, we removed blank lines and sentences that were either shorter than five words or more than eighty words. Table 1 shows the statistics of the crawled sentences.

4.1.2 MAFAND-MT

We evaluate the proposed and baseline aligners on the MAFAND-MT ⁴ (Adelani et al., 2022) dataset, a multilingual benchmark for African machine translation and alignment tasks. The dataset contains high-quality parallel sentences covering several African–English language pairs, professionally curated and cleaned from news and general-domain sources. In this work, we focus on the *English–Hausa* and *Hausa–English* subsets. Each direction contains parallel text divided into training, development, and test splits, following the official partitioning. The Hausa–English portion consists of several tens of thousands of aligned sentence pairs, with development and test sets typically around 1–2K examples each. This dataset provides a realistic evaluation setting for low-resource alignment due to the moderate corpus size, linguistic diversity, and domain variation.

4.2 Implementation of Sentence Aligners

In accordance with the baseline–nearest neighbor aligner (Abdulmumin et al., 2023), the evaluation

Language	Crawled Sentences	Cleaned Sentences
Hausa	13,916	13,560
English	23,148	22,671

Table 1: Statistics of Monolingual Hausa and English Sentences

script of vecmap⁵ was modified to formulate the source-target sentence aligner. Employing inverted softmax retrieval, the aligner was created by utilizing the CoHere multilingual embedding model to transform both source and target sentences into a 768-dimensional vector. The CoHere embedding API, available for free, imposes a limit of approximately 6,000 sentence conversions to embeddings per minute. Consequently, to address this limitation, the CoHere sentence aligner was designed to pause for 61 seconds after processing a batch of source and target sentences. The batch size was set at 2,000 (or the remaining number of sentences), encompassing both the source and target sentences (totaling 4,000) at each iteration until obtaining embeddings for every sentence. To preserve the generated embeddings for potential future use, we save them to a file and upload it whenever the embedding of a previously converted sentence is required.

4.3 Evaluation

We use the MAFAND-MT (Adelani et al., 2022) datasets, which provide gold-standard target sentences, to compare the performance of the proposed inverted softmax aligner against the existing nearest neighbor aligner. This setup enables the use of precision, recall, and F1-score to objectively measure the quality of the aligned sentence pairs. For the evaluation, we focused on the English–Hausa subset of the MAFAND-MT train, development, and test sets. Furthermore, we utilized the labeled MAFAND-MT training data to train machine translation models in a semi-supervised manner, incorporating the automatically aligned crawled sentences. Specifically, we fine-tuned a publicly available checkpoint of the M2M-100⁶ sequence-to-sequence model on the MAFAND-MT development set. The M2M-100 transformer architecture was designed to enable direct translation across 100 languages without requiring English as an intermediary. Following training, model perfor-

²<https://www.premiumtimesng.com/>

³<https://hausa.premiumtimesng.com/>

⁴<https://github.com/masakhane-io/lafand-mt>

⁵<https://github.com/artetxem/vecmap>

⁶https://huggingface.co/docs/transformers/model_doc/m2m_100

mance was evaluated using the sacreBLEU⁷ metric on the MAFAND-MT test set.

5 Results and Discussion

5.1 Sentence Aligners

As shown in Table 2 the proposed sentence aligner outperforms the existing method across all datasets and evaluation metrics. Notably, it achieves substantial gains in precision (e.g., 74.9% vs. 46.8% on Dev) and recall (e.g., 80.5% vs. 55.4% on Dev), leading to consistently higher F1-scores. These improvements indicate the proposed method’s effectiveness in retrieving more accurate and comprehensive parallel sentence pairs, thereby enhancing overall alignment quality.

Dataset	Existing Aligner			Proposed Aligner		
	P	R	F1	P	R	F1
Mafand-Dev	46.8	55.4	49.0	74.9	80.5	76.7
Mafand-Test	57.7	61.1	54.8	76.6	81.8	78.3
Mafand-Train	37.4	44.8	39.2	67.6	73.8	73.7

Table 2: Performance comparison of the existing and proposed sentence aligners on the Mafand dataset in terms of precision (P), recall (R), and F1 score (%). Bolded values indicate better performance.

5.2 Machine Translation

Table 3 displays the performances of the models that were trained for English to Hausa and Hausa to English translation directions. In both directions, it is evident that the proposed aligner aligned sentences are more advantageous to the translation models than the existing aligner aligned sentences, with an increase of 0.8 bleu points in English-Hausa translation, and 1.4 bleu points in Hausa-English translation.

Training Data	En→Ha	Ha→En
Existing Aligned Data	15.49	12.55
Proposed Aligned Data	16.91	13.44

Table 3: BLEU scores of translation models trained on data generated by the existing and proposed sentence aligners.

Limitations

In this study, we proposed an improved sentence alignment method for low-resource languages, leveraging CoHere’s multilingual embeddings and

inverted softmax retrieval. The proposed aligner consistently outperformed the existing method in precision, recall, and F1-score across all datasets, demonstrating robust and balanced performance. It proved more effective in identifying accurate parallel sentences, which translated to improved BLEU scores in English–Hausa and Hausa–English machine translation tasks. These results highlight the importance of retrieval strategies in sentence alignment quality. Future work will explore extending this technique to other low-resource African languages to support broader multilingual NLP efforts.

References

- Idris Abdulmumin, Auwal Abubakar Khalid, Shamsuddeen Hassan Muhammad, Ibrahim Said Ahmad, Lukman Jibril Aliyu, Babangida Sani, Bala Mairiga Abduljalil, and Sani Ahmad Hassan. 2023. Leveraging closed-access multilingual embedding for automatic sentence alignment in low resource languages. *arXiv preprint arXiv:2311.12179*.
- David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruitter, 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, and 26 others. 2022. *A few thousand translations go a long way! leveraging pre-trained models for african news translation*. *Preprint*, arXiv:2205.02022.
- Maha Jarallah Althobaiti. 2021. A simple yet robust algorithm for automatic extraction of parallel sentences: A case study on arabic-english wikipedia articles. *IEEE Access*, 10:401–420.
- Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the association for computational linguistics*, 7:597–610.
- Peter F Brown, Stephen A Della Pietra, Vincent J Della Pietra, and Robert L Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.
- Peter F Brown, Jennifer C Lai, and Robert L Mercer. 1991. Aligning sentences in parallel corpora. In *29th annual meeting of the association for computational linguistics*, pages 169–176.
- Rémi Cardon and Natalia Grabar. 2019. Parallel sentence retrieval from comparable corpora for biomedical text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 168–177.

⁷<https://huggingface.co/spaces/evaluate-metric/sacrebleu>

- Boxing Chen and Limin Du. 2003. Preparatory work on automatic extraction of bilingual multi-word units from parallel corpora. In *International Journal of Computational Linguistics & Chinese Language Processing*, Volume 8, Number 2, August 2003, pages 77–92.
- Stanley F Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 9–16.
- Everlyn Asiko Chimoto and Bruce A Bassett. 2022. Very low resource sentence alignment: Luhya and swahili. *arXiv preprint arXiv:2211.00046*.
- Katsuki Chousa, Masaaki Nagata, and Masaaki Nishino. 2020. Spanalign: Sentence alignment method based on cross-language span prediction and ilp. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4750–4761.
- Kenneth Church. 1993. Char_align: A program for aligning parallel texts at the character level. In *31st Annual Meeting of the Association for Computational Linguistics*, pages 1–8.
- Yonggang Deng and Bill Byrne. 2006. Mttk: An alignment toolkit for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Demonstrations*, pages 265–268.
- Georgiana Dinu, Angeliki Lazaridou, and Marco Baroni. 2015. [Improving zero-shot learning by mitigating the hubness problem](#). *Preprint*, arXiv:1412.6568.
- Aloka Fernando, Surangika Ranathunga, Dilan Sachintha, Lakmali Piyarathna, and Charith Rajitha. 2023. Exploiting bilingual lexicons to improve multilingual embedding-based document and sentence alignment for low-resource languages. *Knowledge and Information Systems*, 65(2):571–612.
- Pascale Fung and Kathleen McKeown. 1994. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic time warping. *arXiv preprint cmp-lg/9409011*.
- Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.
- Riyafa Abdul Hameed, Nadeeshani Pathirennhelage, Anusha Ihalapathirana, Maryam Ziyad Mohamed, Surangika Ranathunga, Sanath Jayasena, Gihan Dias, and Sandareka Fernando. 2016. Automatic creation of a sentence aligned sinhala-tamil parallel corpus. In *Proceedings of the 6th Workshop on South and Southeast Asian Natural Language Processing (WS-SANLP2016)*, pages 124–132.
- Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. [Bitext mining using distilled sentence representations for low-resource languages](#). *Preprint*, arXiv:2205.12654.
- Rafał Jaworski, Sanja Seljan, and Ivan Dunder. 2023. Four million segments and counting: Building an english-croatian parallel corpus through crowdsourcing using a novel gamification-based platform. *Information*, 14(4):226.
- Max Kaufmann. 2012. Jmaxalign: A maximum entropy parallel sentence alignment tool. In *Proceedings of COLING 2012: Demonstration papers*, pages 277–288.
- Martin Kay and Martin Roscheisen. 1993. Text-translation alignment. *Computational linguistics*, 19(1):121–142.
- Wang Ling, Luis Marujo, Chris Dyer, Alan W Black, and Isabel Trancoso. 2016. Mining parallel corpora from sina weibo and twitter. *Computational linguistics*, 42(2):307–343.
- Xiaoyi Ma. 2006. Champollion: A robust parallel text sentence aligner. In *LREC*, pages 489–492.
- Kirsty Macdonald. 2001. Improving automatic alignment for translation memory creation. In *Proceedings of Translating and the Computer 23*.
- Aibek Makazhanov, Bagdat Myrzakhmetov, and Zhenisbek Assylbekov. 2018. Manual vs automatic bitext extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.
- Anthony M McEnery, Michael P Oakes, and Roger G Garside. 1994. The use of approximate string matching techniques in the alignment of sentences in parallel corpora. In *Proceedings of the Second International Conference on Machine Translation: Ten years on*.
- I Dan Melamed. 2001. *Empirical methods for exploiting parallel texts*. MIT press.
- Kumiko Ohmori and Masanobu Higashida. 1999. Extracting bilingual collocations from non-aligned parallel corpora. In *Proceedings of the 8th Conference on Theoretical and Methodological Issues in Machine Translation of Natural Languages*.
- Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. 2017. Massalign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, System Demonstrations*, pages 1–4.
- Harris Papageorgiou, Lambros Cranias, and Stelios Piperidis. 1994. Automatic alignment in parallel corpora. In *32nd Annual Meeting of the Association for Computational Linguistics*, pages 334–336.
- Philip Resnik and Noah A Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.
- Jason Riesa and Daniel Marcu. 2012. Automatic parallel fragment extraction from noisy data. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational*

- Linguistics: Human Language Technologies*, pages 538–542.
- Nasredine Semmar and Christian Fluhr. 2007. Arabic to french sentence alignment: Exploration of a cross-language information retrieval approach. In *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, pages 73–80.
- Edoardo Signoroni and Pavel Rychlý. 2023. Evaluating sentence alignment methods in a low-resource setting: an english-yorùbá study case. In *Proceedings of the The Sixth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2023)*, pages 123–129.
- Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. 2017. [Offline bilingual word vectors, orthogonal transformations and the inverted softmax](#). *Preprint*, arXiv:1702.03859.
- Regina Stodden and Laura Kallmeyer. 2022. Ts-anno: an annotation tool to build, annotate and evaluate text simplification corpora. In *Proceedings of the 60th annual meeting of the association for computational linguistics: system demonstrations*, pages 145–155.
- Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1342–1348.
- Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. *Amsterdam Studies In The Theory And History Of Linguistic Science Series 4*, 292:247.
- Martin Volk, Torsten Marek, and Yvonne Samuelsson. 2008. Human judgements in parallel treebank alignment.
- Tariq Yousef, Chiara Palladino, David J Wright, and Monica Berti. 2022. Automatic translation alignment for ancient greek and latin. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages*, pages 101–107.
- Shiyu Zhao, Xiaopu Li, Minghui Wu, and Jie Hao. 2021. The mininglamp machine translation system for wmt21. In *Proceedings of the Sixth Conference on Machine Translation*, pages 260–264.

AFRICAPTION: Establishing a New Paradigm for Image Captioning in African Languages

Mardiyyah Oduwole^{1*} Prince Mireku^{1,2*} Fatimo Adebajo^{1†}
Oluwatosin Olajide^{1†} Mahi Aminu Aliyu^{1,3} Jekaterina Novikova^{1‡}

¹ML Collective ²Ashesi University ³Abubakar Tafawa Balewa University
mardiyyah.oduwole@mlcollective.org

Abstract

Multimodal AI research has overwhelmingly focused on high-resource languages, hindering the democratization of advancements in the field. To address this, we present AfriCaption, a comprehensive framework for multilingual image captioning in 20 African languages and our contributions are threefold: (i) a curated dataset built on Flickr8k, featuring semantically aligned captions generated via a context-aware selection and translation process; (ii) a dynamic, context-preserving pipeline that ensures ongoing quality through model ensembling and adaptive substitution; and (iii) the AfriCaption model, a 0.5B parameter vision-to-text architecture that integrates SigLIP and NLLB200 for caption generation across under-represented languages. This unified framework ensures ongoing data quality and establishes the first scalable image-captioning resource for under-represented African languages, laying the groundwork for truly inclusive multimodal AI.

1 Introduction

The digital divide in multimodal AI is starkly evident, with most advancements centered around a selected few Western languages, leaving non-Western languages, especially African languages, underrepresented (Longpre et al., 2024). This under-representation perpetuates a cycle of exclusion, where machine learning systems fail to generalize to global contexts and perform poorly for speakers of low-resource languages, creating a barrier to inclusive AI development.

Datasets and models have both been two strong pillars of machine learning since its inception, where the performance of a good model stems not only from its architecture or training hyperparam-

eters but also from the foundational dataset on which it is trained.

Early benchmark datasets such as ImageNet (Deng et al., 2009) revolutionized computer vision by providing large-scale annotated images, enabling the development of deep learning models. Similarly, MS COCO (Lin et al., 2014), Flickr8k (Hodosh et al., 2013), and Visual Genome (Krishna et al., 2016) provided diverse image-text pairings, facilitating advancements in vision-language tasks like image captioning and visual question-answering. However, these datasets are overwhelmingly monolingual, primarily in English, reflecting an inherent bias in AI research (Geburu et al., 2018). The consequence of this linguistic homogeneity is a failure to generalize AI models across non-Western contexts, limiting their usability and fairness (Bender et al., 2021).

To this end, we introduce AFRICAPTION: an image captioning model and dataset for African languages. AFRICAPTION provides an image-text pair dataset and an image captioning model that, in addition to English, covers 20 African languages, spanning across several language families and regions. To the best of our knowledge, this is the first image captioning model and curated caption corpus of this scale built for African languages. The key contributions of our work include:

- **The AFRICAPTION dataset containing diversified multilingual captions:** We create a corpus of human-readable captions in linguistically diverse African languages, including Igbo, Hausa, Ewe, Yoruba, Luganda, Kinyarwanda, and others spanning Afro-Asiatic, Niger-Congo, and Nilo-Saharan families. AFRICAPTION addresses the lack of coverage for low-resource African languages, creating opportunities to train and evaluate models while ensuring linguistic AI representation (Section 5).

*Equal contribution.

†Equal contribution.

‡Supervisor.

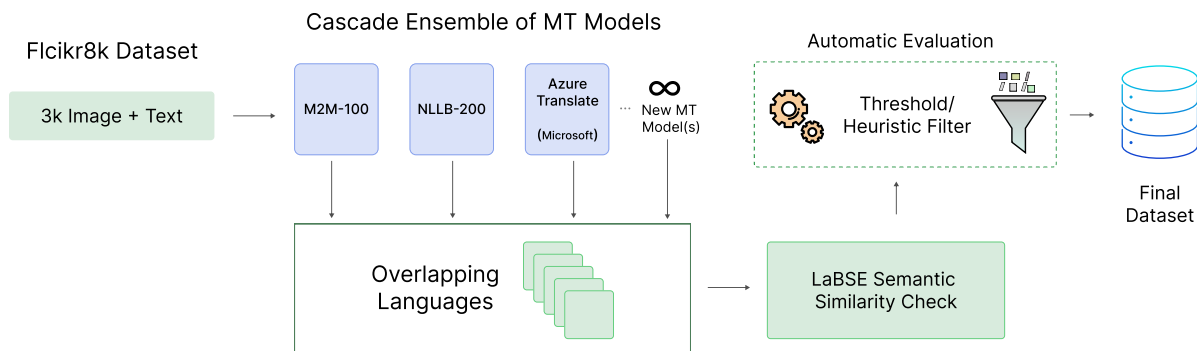


Figure 1: The context-preserving adaptive pipeline, ensuring continuous improvement and high data quality.

- **Context-preserving pipeline:** We present a novel caption translation process to ensure the African language captions remain faithful to the image semantics (Figure 1 and Section 4).
- **The AFRICAPTION image captioning model:** We introduce the first image captioning model designed to generate captions in a wide range of African languages. It is the first to support image captioning for the majority of the 20 African languages covered in our dataset. The model aligns a vision encoder (SIGLIP) with a multilingual text decoder (NLLB) to produce captions across these languages (Section 6).

With this work, we hope to broaden the scope of research and democratise AI, ensuring that cutting-edge technologies benefit a global community rather than just speakers of high-resource languages.

2 Related Work

Recent efforts have sought to address the under-representation of diverse languages in multimodal AI. A prime example is OpenAI’s CLIP model (Radford et al., 2021), which aligns text and images using large-scale datasets primarily in high-resource languages such as English. While CLIP has demonstrated impressive zero-shot learning capabilities, it struggles to generalize across diverse linguistic contexts, particularly for under-represented languages, such as those spoken across Africa. Similarly, multilingual models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and M2M-100 (Fan et al., 2020; Schwenk et al., 2019; El-Kishky et al., 2019) have demonstrated the feasibility of training AI systems across multiple languages, but their reliance on

textual corpora that often exclude low-resource languages results in suboptimal performance for African languages (Kakwani et al., 2020).

One of the more recent efforts to bridge this gap is AViLaMa (Team, 2024), a large open-source text-vision alignment pre-training model specifically targeting African languages. AViLaMa integrates supervision from several African languages, including Swahili, Hausa, Yoruba, Igbo, Zulu, Shona, Arabic, and Amharic, alongside Western languages, such as English, French, and Portuguese.

Despite these advancements, a key limitation remains: many of these datasets and models focus heavily on a small subset of widely spoken African languages, often neglecting lower-resource languages that are equally important for the democratization of AI. Additionally, many multimodal models primarily focus on text-vision alignment without addressing the full spectrum of African linguistic diversity in image captioning and other multimodal settings. Although datasets like Multi30k (Elliott et al., 2016) and the CrissCrossed Captions Dataset (CxC) (Parekh et al., 2020) have expanded multilingual representation, they still lack sufficient African language inclusion.

In our work, we aim to contribute to the inclusion of African languages in the advances in the multimodal domain by introducing AFRICAPTION, the first image captioning model and dataset that broadens the linguistic spectrum and includes 20 African languages. This ensures a wider representation, particularly for under-represented languages. Unlike previous efforts that focus solely on text-vision alignment, our dataset integrates both text and image pairs from the well-known Flickr8k dataset (Hodosh et al., 2013), and our model integrates SigLIP’s (Zhai et al., 2023) vision encoder with NLLB’s decoder (Costa-Jussà et al., 2022), providing a richer multimodal resource. We also prioritize

and linguistic inclusivity by ensuring captions are contextually relevant to each language, fostering nuanced interactions with diverse linguistic groups. We aim to facilitate research in multilingual multimodal AI for low-resource languages and enable models to generalize better across different languages, contributing to a more inclusive global AI landscape.

3 Background

Africa is one of the most linguistically diverse regions in the world. Estimates suggest that the continent is home to between 1,500 and 2,000 distinct languages (Simons and Fennig, 2022).¹ These languages span several major families, including Niger-Congo, Afro-Asiatic, Nilo-Saharan, and Khoisan, and exhibit a wide range of morphological structures. In many African languages, particularly within the Afro-Asiatic and Niger-Congo families, the morphology can be highly complex.

Many African languages remain absent from the corpora despite being spoken by tens of millions and this has led to the challenge of the significant gap in machine learning (ML) resources. For example, while languages like Yoruba, Hausa, and Igbo, which collectively have speaker populations ranging from approximately 20 to 50 million, are included in some ML datasets, the vast majority of Africa’s languages receive little or no attention (Toure, 2025; EqualyzAI, 2025).

3.1 Harms of Misrepresentation

The underrepresentation of African languages in ML datasets and models has significant technical and societal consequences. From a technical perspective, models trained predominantly on high-resource languages fail to generalize to the unique grammatical structures and contexts inherent to many African languages. This can lead to degraded performance, misinterpretation of idiomatic expressions, and ultimately, erroneous outputs when these models are applied in real-world settings. The issue is compounded by the morphological complexity of many African languages, which demands tailored linguistic models that can capture inflectional nuances and tonal variations (Kandybowicz et al., 2018). The harms, however, extend far beyond technical inadequacies. When languages spoken by millions are marginalized in AI research, speakers

¹The Ethnologue reports a similar range, though numbers vary with new surveys.

of these languages are effectively excluded from the benefits of modern technology.

4 Context-Preserving Pipeline

A dedicated and reliable system for obtaining quality data in the context of African NLP is a rarely explored topic. This is a result of some MTs performing better in some languages and poorly in others. To this end, we develop a simple and effective pipeline (Figure 1) that ensures data quality and continual updates through a method of model ensembling and substitution.

We start by using the Flickr8k dataset (Hodosh et al., 2013) as input to several publicly available machine translation (MT) models and build a cascade of ensembles where each model is capable of translating at least one African language. We generate and evaluate similarities between embeddings using language-agnostic BERT (LaBSE)(Feng et al., 2020), compared to other methods like back-translation from the target language to English, eliminates the need for computationally expensive resources.

For languages supported by multiple models, we measure the cosine similarity between their translations and retain the version with the highest score in the final dataset. To ensure ongoing quality, we introduce a dynamic replacement mechanism: when newer models outperform previous ones, that is, yielding higher similarity scores, the corresponding translations are updated accordingly. The novelty of our approach lies in the flexibility of our framework, which allows swapping target language translations when better-performing models become available, ensuring continuous improvement without compromising data integrity.

As a final step to ensure the quality of the dataset, we used heuristic filters to filter out suspicious translations. Inspired by the prior work on unified text-to-text transformer (Raffel et al., 2020) and work on developing heuristic filters through data inspection (Penedo et al., 2025), we devise a mechanism to select close to accurate translations. We implement a method of manually inspecting and eliminating translations that fall below a set threshold. From observation in translated languages, manual human evaluation in a sample revealed translations that adequately describe an image without loss of context, had cosine similarity scores above 0.53. We then used this as the threshold to select quality translations.

Dataset Name	#Samples	#lang	Include African Lang	#African Langs
Multi30K (Elliott et al., 2016)	30,014	2	×	-
Crossmodal-3600 (Thapliyal et al., 2022)	3,600	36	×	-
COCO-CN (Li et al., 2019)	20,342	2	×	-
WIT (Srinivasan et al., 2021)	11.5M	108	✓	unspecified
ArtELingo-28 (Mohamed et al., 2024)	2,000	28	✓	10+
AFRICAPTION (Ours*)	8K	21	✓	20

Table 1: Comparison of multilingual image-text datasets with respect to African language coverage. AFRICAPTION (Ours) is the only dataset to explicitly support 20 African languages, providing broader coverage than existing benchmarks.

For our dataset, the final collection D consists of translations t such that the similarity score $d(t)$ lies in the interval $[0.53, 0.98]$:

$$D = \{t \mid d(t) \in [0.53, 0.98]\}$$

This method allows for data quality assurance even for languages with limited access to human evaluators, which is crucial for creating datasets of under-represented languages with low resources.

5 AFRICAPTION Dataset

5.1 Data Selection and Translation

The Flickr8k dataset, which consists of 8,000 images, each accompanied by five human-generated captions, was chosen particularly for its dense captions upon human review of a few samples.

Caption Selection To compress the dataset to as minimal as possible, a single caption had to be selected among the five to represent a single image. We assume that the best captions will have semantic similarities when compared with each other. We compute a cosine similarity score between all the pairs of captions, and the caption with the highest score is selected in order to avoid potential biases inherent in any singular selection method. In order to ensure preservation of context, which is vital for multilingual tasks, we utilize the pre-trained SentenceBERT model from the sentence transformer (Reimers and Gurevych, 2019) to generate vector representations of the captions.

Translation Process Leveraging the individual translation capability of multiple machine translators, we experimented with SoTA models that support African languages, proven by literature to have a par performance. Our experiment utilized NLLB200 (Team et al., 2022), M2M100 (Fan et al., 2020) and Azure Translate (Microsoft, 2023), of which the first two are publicly available models.

5.2 Quality Assurance

To assess the quality of translations in AFRICAPTION, we adopted a two-pronged approach: (1) an automated similarity evaluation using a back-translation method and (2) a human evaluation to ensure contextual fidelity.

Automatic Evaluation The automatic process leveraged a back-translation strategy using the NLLB200 (Team et al., 2022) MT model. We translated captions from English to target languages and then back to English. The cosine similarity score between the original caption and the back-translated caption was calculated between the embeddings of both the original and translated captions. To preserve context, embeddings were generated using the SentenceBERT model as described in our caption selection process 5.1.

Human Evaluation To complement automatic evaluation, we conducted a human evaluation study on four languages: Yoruba, Igbo, Hausa, and Ewe. We chose these languages based on proximity to communities where these languages are spoken and to cover a mix of high vs. low-resource scenarios. Yoruba and Hausa are widely spoken and have relatively better MT support (we used Azure and NLLB for Yoruba, NLLB for Hausa), whereas Igbo and Ewe are less supported (both used NLLB; Ewe is especially low-resource).

We gathered responses from native speakers of these languages. In total, 102 Yoruba, 38 Igbo, 37 Hausa, and 2 Ewe ratings were collected. We removed a small number of responses that were obviously invalid (e.g., respondents giving all 1’s or all 10’s without variation, which we suspected was not genuine).

5.3 Results

Table 2 shows our effort to evaluate automatic translations using BLEU, which are commonly applied metrics in MTs (Papineni et al., 2002; Popović,

Language	BLEU	Language	BLEU
yor	0.4460	kin	0.5978
amh	0.6307	lua	0.3268
afr	0.3688	kon	0.3496
ibo	0.4945	bem	0.3571
lin	0.2163	dik	0.3714
hau	0.4434	kik	0.2432
cjk	0.4696	ewe	0.3887
lug	0.4468	kam	0.4896
fuv	0.4793	kmb	0.5262
kab	0.8565	dyu	0.5646

Table 2: BLEU scores for the AfriCaption dataset across 20 African languages. Language codes follow ISO 639-3 standards.

2015). Although these metrics capture broad quality trends, we observe considerable variance in translation quality across languages when evaluated using BLEU (some languages score relatively high while others remain lower). A key factor is that many MT systems perform better in the forward direction (Eng \rightarrow target) than in reverse (target \rightarrow Eng). Consequently, our back-translation approach may yield artificially low scores, especially for morphologically complex or extremely low-resource languages (Graham et al., 2019). We therefore complement BLEU with semantic similarity checks and human evaluations for a more robust quality assessment.

Analysis on Human Evaluation We found Yoruba to have the highest quality of translation based on human evaluation, Hausa being the worst, and Igbo and Ewe intermediate (Figure 3, left). The standard deviation of scores was about 2.5–3.0 for all, showing quite a spread of opinions or varying quality across different captions.

Yoruba demonstrated the highest consistency, with an average ICC of 0.68 (moderate agreement). Igbo and Hausa showed lower agreement, with ICCs of 0.52 and 0.41, respectively. Categorical agreement mirrored ICC trends. Yoruba achieved Fleiss’ kappa $\kappa = 0.32$ (moderate), while Igbo and Hausa scored $\kappa = 0.32$ and $\kappa = 0.32$, respectively.

The moderate agreement for Yoruba aligns with its relatively robust machine translation (MT) pipelines (NLLB and Azure Translate) and syntactic simplicity. For instance, the phrase “red-seated swing” translated smoothly as “ìyípadà ìjókòdò pupa” (Yoruba), receiving 78% excellent ratings. In contrast, Hausa’s low agreement correlates with grammatical errors (e.g., “kayaks” mistranslated as “teku,” a general term for “sea”) and limited MT training data (Costa-Jussa et al., 2022). Igbo’s

bimodal scores likely stem from inconsistent handling of idiomatic phrases, such as “taking a swing” translated literally as “ewere swing” (Igbo), which 41% of raters deemed Poor.

Yoruba translation misinterpreted “taking a swing”; a tricky idiom, leading to confusion. Similarly in Hausa, “Three people participate in rock climbing.” received a low 4.4 average; the Hausa translation apparently lost the idea of “rock climbing” (perhaps translating literally in a strange way). On the other hand, Hausa raters gave 8.6 on average to “Women walking down the street.”, indicating that simple captions were handled well. Igbo showed a polarized trend: several captions were rated very high (~ 8.0 – 8.3) but a few were low (~ 4.6 – 5.4). This suggests the Igbo MT sometimes produced excellent results and sometimes failed, perhaps due to inconsistent training data coverage for certain vocabulary. Ewe data is too sparse to draw strong conclusions, but interestingly, the two Ewe evaluators disagreed on many items (one gave much lower scores than the other), reflecting subjectivity or possibly differences in dialect. The human evaluation performed validates that AFRICAPTION machine-translated captions are generally understandable and contextually relevant, though not flawless. They provide a realistic testbed: models trained on or evaluated against these captions will encounter some “noise” or errors as evidenced in our model’s output.

6 AFRICAPTION Model

AFRICAPTION model is a vision-encoder–text-decoder model that integrates a pre-trained vision encoder with a pre-trained sequence-to-sequence language model’s decoder (Figure 2), designed specifically for multilingual image captioning in low-resource African languages. Given an input image and a designated language code, AFRICAPTION generates captions autoregressively, producing text in the specified target language. Our model is capable of generating captions in up to 20 African languages listed in Section 5, thereby addressing a critical gap in image captioning for low-resource languages. AFRICAPTION consists of three components: SigLIP’s Vision Encoder, NLLB Decoder and a linear projector.

6.1 Encoder

For our vision encoder, we use the publicly available multilingual variant of SigLIP’s (Zhai et al.,

	Languages									
	afr	amh	hau	ibo	lug	lin	kin	yor	ckj	dyu
BLEU	71.12	41.03	22.24	39.35	32.74	39.03	18.28	34.85	0.92	0.72
ChrF++	82.32	61.96	42.20	60.85	54.54	59.68	38.21	55.56	15.03	13.31
	dik	ewe	fuv	kam	kab	kmb	kik	kon	lua	bem
BLEU	4.22	1.08	1.31	1.55	0.50	0.44	1.68	0.16	0.76	1.16
ChrF++	20.24	14.57	15.89	17.64	14.26	13.82	17.29	14.77	15.51	16.64

Table 3: Translation quality across languages measured by BLEU and ChrF++ scores. Language codes (e.g., amh for Amharic, afr for Afrikaans, etc.) follow ISO 639-3 standards. Full language definitions are provided in Appendix 6.

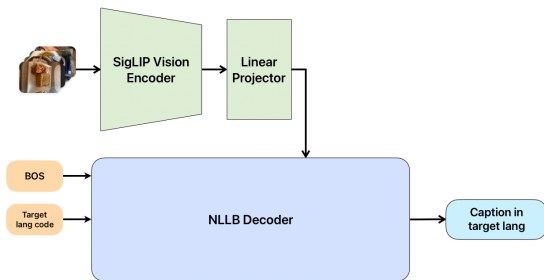


Figure 2: AFRICAPTION model architecture.

2023) image encoder, which is tailored to support multiple languages. This model employs sigmoid loss instead of softmax loss for contrastive pretraining of image-text pairs, demonstrating state-of-the-art performance, particularly given its small size

6.2 Decoder

For our text decoder, we use the publicly available NLLB (Costa-Jussà et al., 2022) checkpoint, which covers 200 of the world’s spoken languages (20 of which AFRICAPTION focuses on). In our setup, the NLLB decoder generates a sequence of wordpiece tokens conditioned on the visual features extracted by the SigLIP encoder. The NLLB decoder produces output sequentially and employs an attention masking mechanism that restricts each generated token to attend only to previously generated tokens, thereby ensuring an autoregressive generation process. NLLB’s tokenizer handles language-specific tokenization.

6.3 Vision-Encoder-Text-Decoder Integration

The image encoder and text decoder are integrated using a modified version of Hugging Face’s VisionEncoderDecoderModel class. The visual features produced by the encoder are projected to match the decoder’s hidden size, ensuring com-

patibility when performing encoder–decoder cross-attention. During training, the model prepares the decoder’s input by shifting the target sequence to the right—ensuring that each output token only attends to preceding tokens, as required in sequence-to-sequence learning. Finally, an `lm_head` linear layer is applied to project the decoder’s hidden states to the size of the vocabulary, and a softmax function produces the probability distribution over the target tokens. This design allows for seamless encoder–decoder cross-attention and end-to-end training, and it is relevant for our task of choosing, image captioning.

6.4 Training

The training of AFRICAPTION follows a two-stage fine-tuning training technique, which we detail in this section.

Stage 0: Selective layer pretraining Firstly, we take the publicly available checkpoints of the pre-trained models off-the-shelf and integrate them using a custom Hugging Face `VisionEncoderTextDecoder` class to include an LM Head at the final layer of the NLLB decoder (Costa-Jussà et al., 2022). We train the last layer of the vision encoder model along with the linear projection layer with the aim of aligning the image and text modalities. SigLIP (Zhai et al., 2023) traditionally uses an encoder language model; however, most language models with African language translation capabilities are either decoder-only transformers or encoder-decoder language models. We opt for the NLLB decoder, as it has decent African language translation capabilities compared to other multilingual language models. We train for 40 epochs, using an lr of $2.0e-5$ and a batch size of 16 on an L4 GPU.

Stage 1: Multimodal Pretraining In this stage, we pretrain the resulting model from Stage 1 for the image captioning task. The goal is to have a model that has acquired image captioning

skills and be able to generate correct image captions in 20 African languages. We do not freeze any layer in our models like we did in the first stage. It is common practice to keep the image encoder frozen during this stage due to findings in LiT [132], reporting multimodal tuning of pretrained image encoders degrading their representations. Studies like CapPa (Tschannen et al., 2023) and (Wan et al., 2024) have shown that captioning tasks can provide valuable signals to image encoders, allowing them to learn spatial and relational understanding capabilities that contrastive models like CLIP or SigLIP typically lack. Hence, we do not freeze the image encoder. We use a slow linear warm-up for our learning rate and an inverse root decay after the warm-up phase, which helps to first stabilize training (via warm-up) and then maintain a slowly decreasing learning rate to allow the model to train its parameters over time. We train for 30 epochs using an lr of 2.0e-5 and a batch size of 16 on an L4 GPU.

6.5 Results and Analysis

The model demonstrated steady improvement across training epochs, as evidenced by the progressive reduction in both training loss and validation loss, indicating an overall improvement in model confidence and generalization.

Model	Lang	Bleu	Cider	Spice
Pangea	amh	0	2.642e-08	2.750e-3
	igb	0.0014	1.127e-07	4.981e-3
AfriCaption	afr	0.8387	8.3207	0.8358
	amh	0.7768	7.6630	0.7906
	bem	0.4813	4.4306	0.4952
	cjk	0.2167	1.7945	0.1977
	dik	0.2521	2.0666	0.2009
	dyu	0.1732	1.3900	0.2288
	ewe	0.2262	1.6527	0.1790
	fuv	0.3552	2.1234	0.2192
	hau	0.8567	8.4716	0.8435
	ibo	0.8506	8.4087	0.8433
	kab	0.1019	0.7688	0.0899
	kam	0.1691	1.2574	0.1493
	kik	0.1844	1.5548	0.1538
	kin	0.7753	7.5940	0.7791
	kmb	0.2407	1.9489	0.1968
	kon	0.4129	3.4451	0.3728
	lin	0.4044	3.6024	0.3807
	lua	0.3017	2.6935	0.3162
	lug	0.5156	5.0911	0.5364
yor	0.8212	7.9930	0.8127	

Table 4: Performance Comparison between Pangea and AFRICAPTION per language.

Table 4 presents a performance comparison between the AFRICAPTION model and Pangea model, a state-of-the-art, open-weight, multilingual multimodal model across BLEU, CIDEr, and SPICE metrics. Table 5 shows our models’ output and it effectively captures the context of the images and generates complete sentences. In some instances, it produces words that are semantically similar to those in the ground truth captions, while in others, it omits one or two words within the caption. For the English translations, we highlight missed or unrelated words in red, indicating that they do not align with the image or the ground truth caption. Words that are contextually similar such as verb tense variations (e.g., a present-tense verb in the ground truth appearing in past tense in the model’s output) are marked in yellow to reflect their near-equivalence in meaning.

7 Discussion

Our results show that the model is capable of generating image captions in a variety of African languages, achieving high-quality outputs in some cases while facing challenges in others. It consistently outperforms the Pangea model across the two overlapping languages; see the detailed table in Table 4. However, performance still varies, revealing broader limitations in existing tools for African language processing. These findings underscore the ongoing need for more robust multilingual AI systems, particularly for low-resource settings.

A key factor contributing to this disparity is the uneven representation of African languages in current “massively multilingual” MT models. Languages like Hausa and Yoruba, which have a relatively stronger digital presence and were likely better represented in training data, yielded better results compared to languages like Ewe or Dinka. This suggests that not all African languages benefit equally from such models, reinforcing the need for more inclusive and balanced training datasets. This raises an important question: *can we bootstrap better translations by leveraging closely related languages?* For example, the strong performance in Luganda, a Bantu language, suggests the potential to improve captions for other Bantu languages like Zulu or Xhosa if extended to those languages. Our dataset provides a benchmark for such explorations, offering a foundation for testing fine-tuned MT models on captioning tasks across diverse African languages.

8 Conclusion

We introduced AFRICAPTION, a family of the multilingual multimodal model that generates image captions in 20 African languages and the dataset that consists of 8k image-text samples in 20 African languages and English. Together, the model and dataset address the problem of exclusion of African languages in the vision-language domain, laying the foundation for broader inclusivity in multimodal AI. AfriCaption serves as a foundation for future research in multilingual image captioning and multimodal learning. We make our dataset and model available publicly on hugging face and we hope this spurs the development of more inclusive AI models that can understand and caption images in the languages spoken by the different communities in Africa.

Moving forward, we plan to adopt a participatory approach, similar to Masakhane, to refine and validate captions. We also advocate incorporating culturally specific imagery and descriptions to ensure models resonate with diverse African contexts. Ultimately, AFRICaption is a pivotal step toward bridging the multimodal resource gap and fostering equitable, multilingual AI systems.

9 Limitations

While AFRICAPTION significantly advances multilingual AI inclusivity, it also highlights systemic gaps in low-resource language research. Translation quality remains uneven, for example, Yoruba outperforms languages like Hausa, Ewe, or Dinka due to richer digital representation, and standard back-translation evaluation metrics (e.g., BLEU) often miss semantic nuances in morphologically complex languages. Furthermore, limited community involvement in human evaluation may overlook subtle, culturally nuanced errors.

Beyond these methodological gaps, this work also lacks cultural awareness. While our dataset and models represent a step toward enabling image captioning in African languages for basic daily conversations, they do not yet capture the deeper cultural context embedded in language use, such as idiomatic expressions, social norms, or culturally specific references. Future iterations of this work would benefit from stronger integration of cultural perspectives, ensuring that captions reflect not only linguistic accuracy but also the lived realities of African communities.

References

- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. [On the dangers of stochastic parrots: Can language models be too big?](#). In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA. Association for Computing Machinery.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). *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, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. [Imagenet: A large-scale hierarchical image database](#). In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Ahmed El-Kishky, Vishrav Chaudhary, Francisco Guzman, and Philipp Koehn. 2019. A massive collection of cross-lingual web-document pairs. *arXiv preprint arXiv:1911.06154*.
- Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. [Multi30k: Multilingual english-german image descriptions](#). *CoRR*, abs/1605.00459.
- EqualyzAI. 2025. [Training language models \(llms\) for low-resource african languages](#).
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. [Beyond english-centric multilingual machine translation](#). *Preprint*, arXiv:2010.11125.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

- Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2018. [Datasheets for datasets](#). *CoRR*, abs/1803.09010.
- Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. [Translationese in machine translation evaluation](#). *Preprint*, arXiv:1906.09833.
- Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.
- Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. [Indicnlp suite: Monolingual corpora, evaluation benchmarks and pre-trained multilingual language models for indian languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4948–4961, Online. Association for Computational Linguistics.
- Jason Kandybowicz, Travis Major, Harold Torrence, and Philip T Duncan. 2018. *African linguistics on the prairie: Selected papers from the 45th Annual Conference on African Linguistics*. Language Science Press.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2016. [Visual genome: Connecting language and vision using crowdsourced dense image annotations](#). *CoRR*, abs/1602.07332.
- Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. [Coco-cn for cross-lingual image tagging, captioning and retrieval](#). *Preprint*, arXiv:1805.08661.
- Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. [Microsoft COCO: common objects in context](#). *CoRR*, abs/1405.0312.
- Shayne Longpre, Nikhil Singh, Manuel Cherep, Kushagra Tiwary, Joanna Materzynska, William Brannon, Robert Mahari, Naana Obeng-Marnu, Manan Dey, Mohammed Hamdy, and 1 others. 2024. Bridging the data provenance gap across text, speech and video. *arXiv preprint arXiv:2412.17847*.
- Microsoft. 2023. [Azure translator](#). Accessed: 2024-03-22.
- Youssef Mohamed, Runjia Li, Ibrahim Said Ahmad, Kilichbek Haydarov, Philip Torr, Kenneth Ward Church, and Mohamed Elhoseiny. 2024. [No culture left behind: Artelingo-28, a benchmark of wikiart with captions in 28 languages](#). *Preprint*, arXiv:2411.03769.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Annual Meeting of the Association for Computational Linguistics*.
- Zarana Parekh, Jason Baldridge, Daniel Cer, Austin Waters, and Yinfei Yang. 2020. [Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO](#). *CoRR*, abs/2004.15020.
- Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, and 1 others. 2025. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. [Learning transferable visual models from natural language supervision](#). *CoRR*, abs/2103.00020.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-bert: Sentence embeddings using siamese bert-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. Ccmatrix: Mining billions of high-quality parallel sentences on the web. *arXiv preprint arXiv:1911.04944*.
- Gary F. Simons and Charles D. Fennig, editors. 2022. *Ethnologue: Languages of the World (25th ed.)*. SIL International, Dallas, TX.
- Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, and 1 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.
- Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. [Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning](#). *Preprint*, arXiv:2103.01913.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.

Sartify LLC Research Team. 2024. Avilama: Learning visual concepts directly from african languages supervision. *To be inserted*.

Ashish V. Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. 2022. [Crossmodal-3600: A massively multilingual multimodal evaluation dataset](#). *Preprint*, arXiv:2205.12522.

Matene Toure. 2025. AI for the world, or just the West? How researchers are tackling Big Tech’s global gaps - zdnet.com. [Accessed 27-03-2025].

Michael Tschannen, Manoj Kumar, Andreas Steiner, Xiaohua Zhai, Neil Houlsby, and Lucas Beyer. 2023. Image captioners are scalable vision learners too. *Advances in Neural Information Processing Systems*, 36:46830–46855.

Bo Wan, Michael Tschannen, Yongqin Xian, Filip Pavetic, Ibrahim M Alabdulmohsin, Xiao Wang, André Susano Pinto, Andreas Steiner, Lucas Beyer, and Xiaohua Zhai. 2024. Locca: Visual pretraining with location-aware captioners. *Advances in Neural Information Processing Systems*, 37:116355–116387.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

A Appendix

Survey on Dataset - Human Evaluation

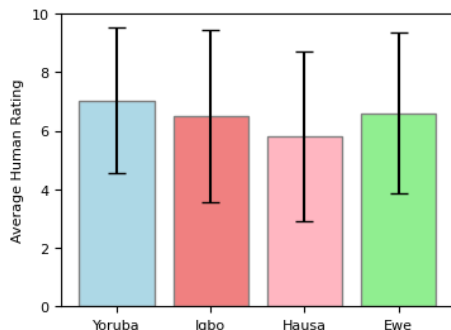


Figure 3: A plot of the **average human score per language** with error bars denoting standard deviation.

We created evaluation surveys where native speakers were presented with the original English

caption and the translated caption in their language. For each caption pair, we asked evaluators to rate the translation’s adequacy on a scale from 1 to 10, with instructions that 1 means “completely wrong translation,” 5 means “understandable gist but with errors,” and 10 means “perfect translation that preserves the full meaning.” We also asked them to flag any catastrophic errors, like when the translation says something entirely different from the original caption.

Perceived Data Quality vs. Average Length: Figure 5 compares the average word count per language caption in our dataset. It shows that our dataset achieves a reasonable balance in caption length across 20 African languages, with the English captions providing a baseline. The consistent average word counts suggest that the translations are neither too brief nor overly verbose, preserving the essential information while ensuring readability. According to previous studies (Singh et al., 2024), balanced caption length is a key feature in preventing model bias in and improving interpretability. This characteristic makes our dataset well-suited for training models that need to generalize across diverse linguistic contexts.

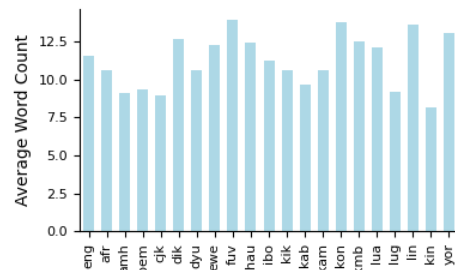


Figure 5: Average word count per language in the AFR-ICaption dataset. The plot highlights variations in caption lengths across different languages.

Lang	Reference	Prediction
yor	Obìnrin kan tó ní àpò èyìn rẹ̀ jókòò sórí àpáta òlá kan, ó sì wo àwọn òkè òlá. (A woman with backpack sits on a large rock and looks down over the mountains.)	Obìnrin kan tó ní àpò èyìn ṣ jókòò lórí àpáta òlá kan tó ó sì wo òkè òkè òlá tó (A woman with a backpack sat on a large rock and looked up at the mountain.)
amh	በህዝብ መካከል አንዲት ልጃገረድ የተንቆጠቆጠች በግ ይዛ! (A girl in a crowd is holding on to a leashed sheep!)	በህዝብ ውስጥ አንዲት ወጣትጃገረድ በንቆጠቆጠውን ለመዛ ተሸ (A young girl in the crowd is holding a fringe)
hau	Mata biyun sun kwana a kan ciyawa tare (The two women slept on the grass together)	Mata biyuun sun kwana a ci ciyawa (The two women slept on [ci] grass)

Table 5: Comparison of the groundtruth translations and our model’s output for 3 African Languages

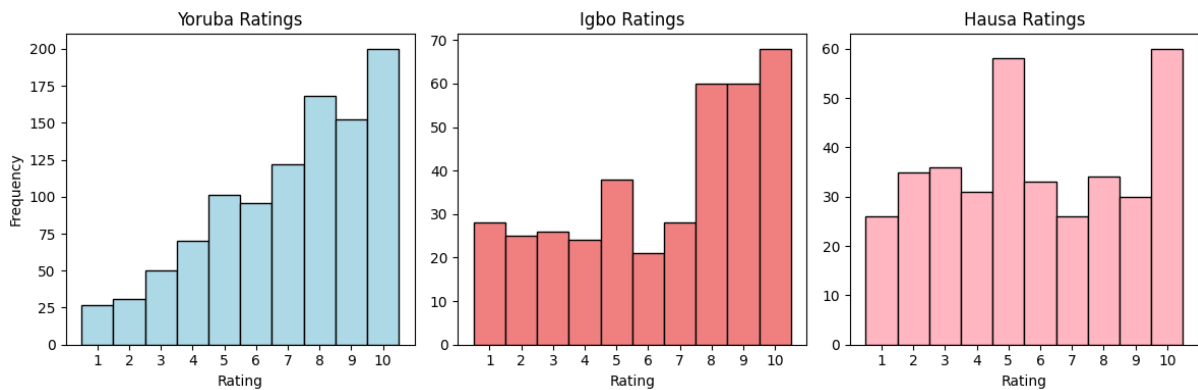


Figure 4: Score distributions for Yoruba, Igbo, and Hausa. We observe that over 50% of Yoruba ratings were 8 or above, and ~20% were perfect 10s. Hausa’s distribution is flatter, with a mode around 10 (16% of scores were 10) but also a substantial portion of low scores (1–4 ratings made up ~35% of Hausa responses, compared to only ~18% for Yoruba). Igbo’s distribution is bimodal – it has a high incidence of 9–10 scores (about one-third of Igbo ratings were 9 or 10, similar to Yoruba) and a noticeable chunk of very low scores (1’s, 2’s, 3’s accounted for ~21% in Igbo, versus ~9% in Yoruba). This bimodality aligns with the earlier observation of Igbo translations being hit-or-miss

ISO 639-3	Language Name	Countries (with Flags)
afr	Afrikaans	South Africa, Namibia
amh	Amharic	Ethiopia
hau	Hausa	Nigeria, Niger, Ghana, Chad, Cameroon
ibo	Igbo	Nigeria
lug	Luganda	Uganda
lin	Lingala	DR Congo, Rep. Congo, Angola, Central African Rep.
kin	Kinyarwanda	Rwanda, DR Congo, Uganda
yor	Yoruba	Nigeria, Benin, Togo
cjk	Chokwe	Angola, DR Congo
dyu	Dyula (Jula)	Burkina Faso, Côte d'Ivoire, Mali
dik	Dinka	South Sudan
ewe	Ewe	Ghana, Togo
fuv	Fulfulde (Fula)	Nigeria, Cameroon, Guinea, Senegal, Mali
kam	Kamba	Kenya
kab	Kabyle	Algeria
kmb	Kimbundu	Angola
kik	Kikuyu	Kenya
kon	Kongo	DR Congo, Rep. Congo, Angola
lua	Luba-Kasai	DR Congo
bem	Bemba	Zambia

Table 6: Languages and their definitions.

	Languages									
	afr	amh	bem	cjk	dik	dyu	ewe	fuv	hau	ibo
No. of Tokens	135336	148517	58640	19805	16042	2021	4734	8126	127429	152722
	kik	kab	kam	kon	kmb	lua	lug	lin	kin	yor
No. of Tokens	12494	2250	6035	15467	2757	16974	66018	22109	120542	172016
No. of Characters	425977	294622	206608	75104	51995	6745	14967	27041	478142	478600
	kik	kab	kam	kon	kmb	lua	lug	lin	kin	yor
No. of Characters	41032	6417	19750	60924	9890	67607	219135	88618	420213	461489
Avg. Length	3.15	1.98	3.52	3.79	3.24	3.34	3.16	3.33	3.75	3.13
	kik	kab	kam	kon	kmb	lua	lug	lin	kin	yor
Avg. Length	3.28	2.85	3.27	3.94	3.59	3.98	3.32	4.01	3.49	2.68

Table 7: Statistical overview of language characteristics. Language codes: afr (Afrikaans), amh (Amharic), bem (Bemba), cjk (Chokwe), dik (Dinka), dyu (Dyula), ewe (Ewe), fuv (Fulfulde), hau (Hausa), ibo (Igbo), kik (Kikuyu), kab (Kabyle), kam (Kamba), kon (Kongo), kmb (Kimbundu), lua (Luba-Katanga), lug (Luganda), lin (Lingala), kin (Kinyarwanda), yor (Yoruba).

Developing an English–Efik Corpus and Machine Translation System for Digitization Inclusion

Offiong Bassey Edet^{1,2}, Mbuotidem Sunday Awak², Emmanuel Oyo-Ita¹,
Benjamin Okon Nyong³, Ita Etim Bassey⁴,

¹University of Cross River River State, ²ML Collective, ³Arthur Jarvis University,

⁴University of Calabar,

offiongbassey99@gmail.com, mbuotidemawak@gmail.com, emmanueloyoita@unicross.edu.ng,

benokon26@gmail.com, basseyita141@gmail.com

Abstract

Low-resource languages serve as invaluable repositories of human history, preserving cultural and intellectual diversity. Despite their significance, they remain largely absent from modern natural language processing systems. While progress has been made for widely spoken African languages such as Swahili, Yoruba, and Amharic, smaller indigenous languages like Efik continue to be underrepresented in machine translation research. This study evaluates the effectiveness of state-of-the-art multilingual neural machine translation models for English–Efik translation, leveraging a small-scale, community-curated parallel corpus of 13,865 sentence pairs. We fine-tuned both the mT5 multilingual model and the NLLB-200 model on this dataset. NLLB-200 outperformed mT5, achieving BLEU scores of 26.64 for English–Efik and 31.21 for Efik–English, with corresponding chrF scores of 51.04 and 47.92, indicating improved fluency and semantic fidelity. Our findings demonstrate the feasibility of developing practical machine translation tools for low-resource languages and highlight the importance of inclusive data practices and culturally grounded evaluation in advancing equitable NLP.

1 Introduction

The rapid advancement in Artificial Intelligence (AI), especially in the field of Natural Language Processing (NLP), has significantly enhanced the ability to translate languages accurately (Vaswani et al., 2017; Fan et al., 2021). However, not all languages have equally benefited from these technological improvements. Many indigenous languages, particularly those classified as “low-resource,” lack sufficient linguistic datasets and digital tools needed for developing effective translation systems (Joshi et al., 2020; Nekoto et al., 2020). One such language is Efik, spoken primarily in the southern part of Cross River State,

Nigeria; particularly Calabar South, Calabar Municipality, Odukpani, Akpabuyo, Bakassi and parts of Akamkpa. Efik is also spoken in some parts of Akwa Ibom State in Nigeria and northwestern Cameroon (Mensah and Dzokoto, 2023; Offiong and Ansa, 2013; Mensah and Mensah, 2014).

Effective translation technologies that bridge English and Efik could therefore have substantial socio-economic and cultural benefits, enhancing communication, fostering inclusion and preserving cultural identity.

Lack of data remains a critical problem in developing neural machine translation systems for African low-resource languages. Most available web data are small and of poor quality (Adelani et al., 2022). Existing systems often fail to translate idiomatic expressions, cultural nuances, and linguistically complex structures accurately, which affects overall translation quality. The primary challenge is the severe scarcity of parallel corpora (sets of sentences paired in English and Efik), which are essential for training AI translation models (Agić and Vulić, 2019). Recent shared tasks such as the AmericasNLP 2025 challenge have shown that MT for Indigenous low-resource languages remains extremely difficult due to severe data scarcity and domain-specific linguistic complexity (Gibert et al., 2025).

Given these challenges, this study evaluates the effectiveness of state-of-the-art multilingual machine translation models, specifically mT5 and NLLB-200, fine-tuned for English-to-Efik translation.

2 Overview of Efik

Efik is the indigenous language of the Efik people, whose major settlements lie in the southern part of Cross River State in Southeastern Nigeria (Mensah and Dzokoto, 2023). Efik is one of the first languages in Nigeria to be written and systemati-

cally studied. Its early prestige led missionaries to develop the orthography, grammar, lexicon, and a substantial body of literature (Offiong and Ansa, 2013).

The language functions as a lingua franca in southern Cross River State and is widely used in radio and television programs, markets, churches, and political activities such as campaigns and rallies. It is also taught in most public primary and secondary schools in the region, as well as in the Linguistics departments of the University of Calabar and the Cross River University of Technology. Several church denominations, including The Apostolic Church Nigeria, Mount Zion Church, and The Presbyterian Church, use Efik during services, often alongside real-time interpretation into English.

Efik was the first Nigerian language to be reduced to writing (Noah, 2004) and the second to receive a full Bible translation after Yoruba (Offiong and Ansa, 2013).

2.1 Decline and Revitalization Efforts

2.1.1 Decline

Efik once featured prominently in national and regional broadcasting, including Radio Nigeria’s news programs in the late 1960s through the early 1980s. Today, however, it is no longer used in national broadcast media. The language was also dropped at one point by the West African Examinations Council (WAEC), despite its long history as a written language. Several factors contribute to its declining use: limited technological support, insufficient institutional encouragement, and decreasing intergenerational transmission. Few new materials have been developed in recent decades compared with earlier periods. In addition, many families now prioritize English over Efik. In numerous households and neighborhoods, English is used in everyday communication, often due to the misconception that fluency in English is a marker of intelligence. As a result, many parents discourage children from learning or speaking Efik, a trend that threatens the vitality of the language in future generations (Offiong and Ansa, 2013).

2.1.2 Revitalization Efforts

Historically, various attempts have been made to expand Efik orthography and vocabulary, though many did not achieve widespread adoption. The long-standing reliance on English loanwords continues to create challenges in developing a robust,

fully standardized writing system (Offiong and Ansa, 2013).

Community tools such as the online Tete Dictionary have emerged to support Efik literacy and vocabulary access for learners (Odey et al., 2025).

3 Related Works

Earlier research by Arikpo and Dickson (2018) focused on the development of an automated English–Efik machine translation system using a rule-based approach rather than modern Neural Machine Translation (NMT). The system relied on a small bilingual dictionary containing 500 English–Efik words and manually defined transfer rules to perform translations. While it produced reasonable results for simple sentences, its performance degraded significantly on complex or composite sentences, highlighting the limitations of rule-based methods in handling diverse syntactic structures, morphological inflections, and idiomatic expressions. The study represents an early attempt to automate translation for Efik but was constrained by the limited dictionary size, absence of robust evaluation metrics, and inability to generalize beyond the manually encoded rules. Despite these limitations, it remains a notable first effort in English–Efik machine translation and underscores the need for larger, data-driven corpora and modern NLP techniques to improve translation quality for low-resource languages.

The JW300 parallel corpus, first collected by Agić and Vulić (2019), provided a large-scale set of English–Efik Bible verse pairs and became foundational for early translation research. Building on this, the Masakhane community initiative (Nekoto et al., 2020) demonstrated the effectiveness of crowdsourced data: volunteers across Africa aligned Efik text, and their Transformer-based machine translation model achieved BLEU scores of 33.48 for English–Efik and 33.68 for Efik–English, demonstrating strong translation performance and contributing open evaluation protocols.

More recently, Kalejaiye et al. (2025) introduced the Ibom NLP dataset, covering several Akwa Ibom languages, including Efik. Their work demonstrates the effectiveness of fine-tuning massively multilingual models, such as M2M-100 and NLLB-200 (NLLB Team et al., 2022), for low-resource Nigerian languages. On English–Efik translation, the models achieved BLEU scores up to 36, while

for Efik–English the best BLEU score reached 34.6, highlighting the potential of cross-lingual transfer even with limited data.

While previous works such as [Nekoto et al. \(2020\)](#) report higher BLEU scores on JW300-based English-Efik data, these results are largely domain-specific (religion-bible). In contrast, our corpus introduces more diverse, natural data across multiple domains, including daily conversation, family and relationships, agriculture, environment, health, and education, emphasizing generalization rather than domain memorization.

4 Methodology

4.1 Dataset Creation

All English sentences were translated by six native Efik speakers and two trained Efik linguists, and independently verified by two additional linguists. All annotators are fluent in both English and Efik. We initially translated 14,000 English sentences into Efik. After removing duplicate entries, incomplete translations, and noisy sentence pairs during data cleaning, the final dataset contains 13,865 high-quality English–Efik parallel pairs. The dataset is stored in UTF-8 encoded JSONL format for efficient preprocessing and reproducibility.

Each sentence was assigned to one primary translator, who produced a single Efik translation. The translation was then independently reviewed by at least two different annotators who did not produce the original translation. Where disagreements occurred, a linguist reviewer performed final adjudication and correction.

4.1.1 Data Collection

English source sentences were created through a combination of controlled large language model (LLM) generation across multiple domains, manually written sentences by Efik–English bilingual linguists and native speakers and example sentences drawn from a printed Efik–English dictionary. The LLM prompts were designed to generate short, domain-diverse sentences covering areas such as daily conversation, food, narratives, religion, and general knowledge. Dictionary examples were manually selected, digitized, and verified to ensure correctness and contemporary usage.

Annotators were instructed to preserve semantic meaning while ensuring natural Efik phrasing rather than literal word-for-word translation. The

Learner’s Dictionary of the Efik Language ([Aye, 1991](#)) was used as a reference resource to resolve lexical ambiguity, verify word meanings, and confirm orthographic conventions, particularly for less frequent or older vocabulary. Each translated sentence was reviewed by at least two independent annotators to ensure grammatical correctness, semantic fidelity, and naturalness.

4.1.2 Domain Coverage

Table 1 shows the domain distribution of the collected English sentences before final cleaning.

The dataset covers a diverse range of domains, with the largest portion coming from daily conversation (7,245 sentences), followed by family and relationships (1,885) and food and agriculture (1,237). Smaller categories include nature and environment, religion, time and number, health, education, and greetings. Sentences that did not clearly fit into any of these categories were grouped under “others” (872), ensuring that all source material was accounted for while maintaining meaningful domain classification.

Domain	Number of Sentences
Daily Conversation	7,245
Family and Relationship	1,885
Food and Agriculture	1,237
Nature and Environment	855
Religion	512
Time and Number	509
Health	475
Education	289
Greetings	121
Others	872

Table 1: Domain distribution of sentences.

4.1.3 Data Extraction

Each parallel pair was stored as a JSON object using the following structure:

```
{ "eng": "my sister is brushing her hair
before going to school.",
  "efi": "eyeneka mi ke asari idet esie mbe-
miso enye aka ufokñwed." }
```

Each line in the JSONL file represents one English–Efik sentence pair, making the dataset easy to preprocess and compatible with standard NLP toolkits such as HuggingFace Transformers, OpenNMT, and Fairseq.

4.2 Data Preprocessing

4.2.1 Data Cleaning

The dataset underwent cleaning steps including removal of duplicates, eliminating extra spaces and standardizing text encoding to UTF-8. Sentences with incomplete or inconsistent translations were discarded.

4.2.2 Dataset Splitting

After cleaning and normalization, the English–Efik parallel corpus containing 13,865 sentence pairs was split into 10,398 for training, 1,387 for validation and 2,080 for testing, as summarized in Table 2.

Split	Number of Sentences
Train	10,398
Dev	1,387
Test	2,080
Total	13,865

Table 2: Dataset Split.

4.3 Dataset Statistics and Analysis

4.3.1 Qualitative Overview

The dataset contains 13,865 English–Efik parallel entries, spanning a wide range of everyday communication.

It includes both sentence-level translations and word or phrase-level entries, particularly for culturally specific terms that require contextual explanation. Lexically, the corpus contains 11,080 unique English word types with 103,734 tokens, and 11,218 unique Efik word types with 100,458 tokens.

Both English and Efik share the same longest sentence in the dataset, consisting of 57 words, while the shortest entries are single-word translations. This distribution reflects a mixture of conversational exchanges, domain-general expressions, and longer narrative or didactic passages.

4.3.2 Sentence Length Distribution

English sentences have an average length of approximately 6 words, with most sentences falling between 2 and 12 words. Efik sentences have an average length of approximately 5 words, with most entries ranging between 2 and 10 words.

In summary, the corpus exhibits a wide variety of sentence lengths, which is important for training a translation model that must handle both short

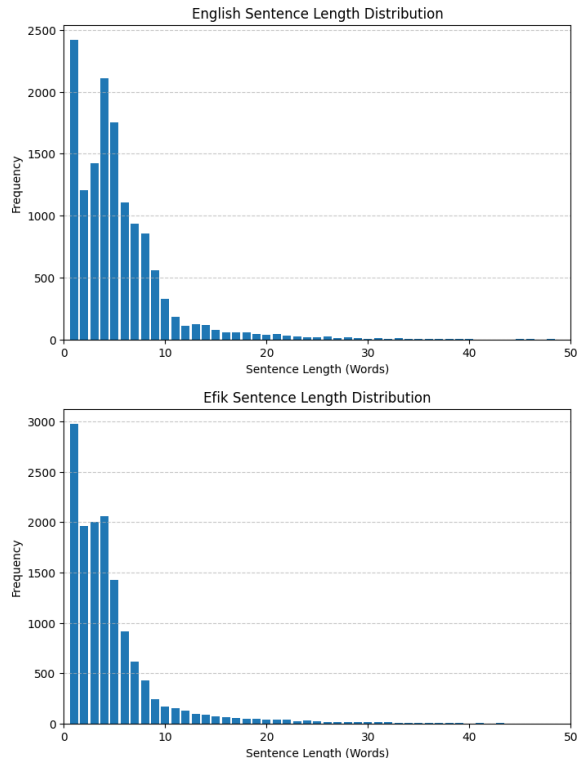


Figure 1: Sentence length distributions for English and Efik.

conversational utterances and longer, structurally complex sentences. Short sentences typically include greetings, responses, exclamations, and simple conversational turns, while longer sentences commonly originate from educational, descriptive, or religious content.

4.3.3 Vocabulary Frequency

English	Count	Efik	Count
house	298	ke	3,457
good	236	enye	2,094
want	220	mi	1,407
thing	211	oro	1,087
money	211	ye	830
one	202	mme	791
like	176	ama	765
food	163	mmọ	762
person	162	edi	738
child	158	owo	698

Table 3: Top 10 most frequent words in English and Efik along with their counts.

From Table 3, it can be seen that certain Efik function words appear with very high frequency compared to English, reflecting the language’s morphological and syntactic patterns. Overall, the

distribution underscores common challenges in low-resource, morphologically rich languages, including sparse lexical diversity, high repetition of grammatical markers, and increased rates of out-of-vocabulary tokens, which motivates the use of subword tokenization and larger, more diverse corpora to improve translation quality.

4.4 Model Training

4.4.1 Training Setup

We adopted a transfer-learning approach by fine-tuning two massively multilingual models: mT5-small (Xue et al., 2021), a text-to-text Transformer pre-trained on 101 languages, and NLLB-200 (NLLB Team et al., 2022), a sequence-to-sequence model covering over 200 languages. Transfer learning has been shown to be more effective for low-resource languages than training models from scratch (Downey et al., 2024).

Both models were selected for their efficiency, generalizability, and suitability for multilingual adaptation. Training was conducted using the Hugging Face Transformers library with structured optimization and hyperparameter settings tuned for low-resource English–Efik translation.

Hyperparameter	mT5	NLLB
Learning rate	1e-4	2e-5
Train batch size per device	8	2
Eval batch size per device	8	2
Gradient accumulation steps	1	2
Number of epochs	20	6
Warmup ratio	-	0.1
Weight decay	0.01	-
FP16	-	True

Table 4: Final hyperparameters used for mT5 and NLLB models.

4.5 Evaluation Methodology

Translation quality was assessed using both automatic metrics and qualitative error analysis.

4.5.1 Automatic Evaluation

We computed BLEU (Papineni et al., 2002) and chrF (Popović, 2015) using the Hugging Face Evaluate library. BLEU measures n-gram overlap, while chrF, a character n-gram-based metric, captures morphological and orthographic patterns, which is particularly important for languages with tone and inflectional complexity.

Table 5 summarizes the BLEU and chrF scores for zero-shot and fine-tuned mT5 and NLLB models on English–Efik translation in both directions. The gray-shaded cells indicate the best-performing results for each metric and direction. Both models perform poorly in zero-shot mode, with mT5 largely unable to generate meaningful translations and NLLB producing semantically weak outputs. Fine-tuning substantially improves performance: mT5-FT achieves moderate gains, while NLLB-FT attains the highest scores across all directions and metrics, demonstrating superior handling of semantic meaning and fluency.

Model	BLEU		chrF	
	en–ef	ef–en	en–ef	ef–en
Zero-shot				
mT5-small	0.01	0.01	2.61	3.15
NLLB-200	0.11	0.22	10.95	15.16
Encoder–Decoder				
mT5-small FT	14.32	18.87	35.55	36.73
NLLB-200 FT	26.64	31.21	51.04	47.92

Table 5: BLEU and chrF scores for zero-shot and fine-tuned mT5 and NLLB models on English–Efik translation.

4.5.2 Qualitative Error Analysis

We conducted a qualitative error analysis on a sample of test sentences for both mT5-FT and NLLB-FT models. Each translation was carefully inspected for substitutions, omissions, and partially correct words. Errors are highlighted in red and omitted content is indicated using italicized red brackets. This analysis provides additional insights into the strengths and weaknesses of each model, complementing the automatic evaluation metrics (Table 5).

Table 6 presents selected examples of English-to-Efik translations. In the first example, the mT5 model omits the repeated phrase *eti eti*, while NLLB correctly preserves it. In the second example, NLLB produces a more fluent and semantically accurate translation than mT5, which fails from the beginning by introducing spurious content and incorrect word choices. In the English-to-Efik examples, while NLLB preserves most of the meaning, it does miss some contextual nuances at the end of the sentence (e.g., translating “fancy”).

Table 7 presents selected Efik-to-English translations. NLLB maintains the overall semantic meaning of the source sentences, even when the transla-

Source (En)	Reference (Ef)	mT5-small FT Output	NLLB-200 FT Output
My mother is very sick and she is in the hospital.	Eka mi ọdọñọ eti eti ndien enye odu ke ufọkibọk.	Eka mi ọsọñọ [<i>eti eti omitted</i>] ndien enye ọdọñọ ke ufọkibọk.	Eka mi ọdọñọ eti eti ndien enye odu ke ufọkibọk.
Are you carrying that umbrella for the rain or just to look fancy?	Ndi akama ufukeyo ke ntak edim mme onyũ akama ke asian?	Ndi afo <i>ebet</i> ufukeyo oro ọno edim mme <i>ńkpọ etiede?</i>	Ndi emen ufukeyo oro ke ntak edim mme <i>ndikut ọkpọsọñ ńkpọ?</i>

Table 6: Qualitative comparison of English to Efik translations produced by mT5 and NLLB.

Source (Ef)	Reference (En)	mT5-small FT Output	NLLB-200 FT Output
Udia usenubọk amakebe, mmọ ewọrọ anwa ndinwam ekpri eyeneka esie ndimen akpasa nya inwañ mbakara kiet nka efe urua.	After breakfast, they went outside to help his younger brother carry a basket of tomatoes to the stall.	<i>The morning went down,</i> they went to help his little sister pick up a basket of <i>yams</i> to the <i>market</i> .	After breakfast, they set off to help his younger brother carry a basket of tomatoes to the stall.
William Otu ekedi akwaowo ke akamba usiakifia aran isọñ ke Lagos.	William Otu was a seasoned manager at a big oil company in Lagos.	William <i>group was a very heavy trader in the mining companies.</i>	William Otu was a senior official in the Lagos oil company.

Table 7: Qualitative comparison of Efik to English translations produced by mT5 and NLLB.

tion is not strictly word-for-word, which is typical and desirable in natural translation. In contrast, mT5 often produces incorrect lexical choices, alters the meaning of domain-specific terms, and fails to correctly translate named entities, such as "Otu".

Overall, NLLB demonstrates higher fidelity, fluency and robust handling of culturally specific content compared to mT5 in Efik-to-English translation.

5 Results and Discussion

Translating between English and Efik remains challenging despite the use of multilingual transfer-learning models. Both mT5 and NLLB perform well on high-frequency words and simple sentences, but they often struggle with idiomatic expressions, morphological inflections, and longer or more complex structures.

Qualitative analysis reveals that mT5 frequently produces incorrect lexical choices, alters domain-specific meanings and sometimes fails to translate named entities correctly. NLLB, in contrast, generally preserves the overall semantic meaning, even when translations are not strictly word-for-word, demonstrating higher fidelity, fluency, and better handling of culturally specific content.

These limitations are partly due to the scarcity of parallel data, the absence of direct Efik equivalents for certain English terms and less commonly used vocabulary. Overall, the results highlight the potential of current multilingual fine-tuning approaches for low-resource languages while emphasizing the

need for additional data and broader domain coverage to improve translation quality and coverage.

6 Conclusion

This study presents fine-tuned multilingual translation systems for English–Efik, using both mT5-small and NLLB-200 models on a manually curated parallel corpus of 13,865 sentence pairs. The results demonstrate that, even with limited data, NLLB produces translations that better preserve semantic meaning, handle culturally specific content, and maintain fluency compared to mT5, which struggles with lexical accuracy and named entities. Both models, however, face challenges with idiomatic expressions, complex sentence structures, and less common or older vocabulary, reflecting the intrinsic difficulties of low-resource language translation.

Our work contributes a community-curated English–Efik dataset and a reproducible pipeline for fine-tuning multilingual models, providing a foundation for future NLP research on underrepresented African languages. The findings highlight the importance of increasing the volume and diversity of parallel data to improve translation fidelity and coverage. Techniques such as back-translation, data augmentation, and crowdsourced annotation could help bridge current gaps in linguistic representation.

Finally, this study highlights the potential of large and specialized multilingual translation models, as demonstrated by mT5 and NLLB-200, for

bidirectional English–Efik machine translation. By combining high-quality parallel data with effective full-parameter fine-tuning, future work can further improve both automatic evaluation metrics and human-perceived translation quality. Overall, this research represents a step toward more inclusive NLP, supporting the development of robust and linguistically informed translation systems for Efik and other low-resource African languages.

7 Limitations

Despite promising results, our study has several limitations. The dataset size (13.8K sentence pairs) restricts exposure to idiomatic expressions and domain-specific language. Orthographic variation, particularly in tone marking and dialectal differences within Efik may reduce translation consistency. While BLEU and chrF (Papineni et al., 2002; Popović, 2015) provide useful quantitative insights, they do not fully capture human judgments of fluency or cultural appropriateness (Kreutzer et al., 2022). Additionally, our human evaluation was limited in scale due to resource constraints. Future work should expand the dataset, apply phonological normalization, and involve community-driven annotation to better capture linguistic diversity and cultural nuances (Joshi et al., 2020).

8 Acknowledgments

We would like to thank Prof. Eyo Mensah, Dr. David Adelani, Dr. Orok Duke, Dr. Enoima Umoh, Gideon George, Abraham Owodunni, Steven Kolawole, Luel Hagos and Saheed Azeez for their invaluable feedback and support throughout this work. We also extend our sincere appreciation to Archibong Archibong, Andrew Nkpanam, Pst. Praize Aye, Mr. Archibong Effiong, Mrs. Edung, Elder Edem Mensah, Mrs. Ansa Edem, Mr. Felix Effiom, Sylvia Edem Mensah, and other native Efik speakers, linguists, and community contributors for their translations and validation efforts. Their linguistic and cultural expertise made this project possible.

We further acknowledge Chief E. U. Aye for his foundational contribution to Efik language documentation through *A Learner’s Dictionary of the Efik Language*, which served as an important reference during dataset development.

We also acknowledge the Africa NLP open-source community, particularly Masakhane, for fostering a collaborative research environment that

enabled reproducible research. Finally, we thank the Machine Learning Collective, a community of machine learning researchers, for their support and guidance.

References

- David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruitter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen R. Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Chinenye Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, and 26 others. 2022. [A few thousand translations go a long way! leveraging pre-trained models for african news translation](#). arXiv preprint. ArXiv:2205.02022.
- Željko Agić and Ivan Vulić. 2019. [Jw300: A wide-coverage parallel corpus for low-resource languages](#). In *Proceedings of the 57th Annual Meeting of the ACL*, Florence, Italy.
- Iwara Arikpo and Iniobong Dickson. 2018. [Development of an automated english-to-local-language translator using natural language processing](#). *International Journal of Scientific and Engineering Research*, 9.
- E. U. Aye. 1991. *A Learner’s Dictionary of the Efik Language*, volume 1. Evans Brothers (Nigeria Publishers) Limited.
- C.M. Downey, Terra Blevins, Dhvani Serai, Dwija Parikh, and Shane Steinert-Threlkeld. 2024. [Targeted multilingual adaptation for low-resource language families](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 15647–15663.
- Angela Fan, Shruti Bhosale, Holger Schwen, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goya, Tom Birch Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2021. [Beyond english-centric multilingual machine translation](#). *JMLR*, pages 1–48.
- Ona de Gibert, Robert Pugh, Ali Marashian, Raúl Vázquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan C. Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno Veliz, Ángel Lino Campos, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. [Findings of the americasnlp 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the americas](#). In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, page 134–152.

- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the nlp world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.
- Oluwadara Kalejaiye, Luel Hagos Beyene, David Ifeoluwa Adelani, Mmekut-Mfon Gabriel Edet, Aniefon Daniel Akpan, Eno-Abasi Urua, and Anietie Andy. 2025. [Ibom nlp: A step toward inclusive natural language processing for nigeria’s minority languages](#). *arXiv preprint arXiv:2511.06531*.
- Julia Kreutzer, Isaac Caswell, Lisa Wang³, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Alahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb¹, Beno^{it} Sagot¹, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. [Quality at a glance: An audit of web-crawled multilingual datasets](#). *Transactions of the Association for Computational Linguistics, Volume 10*, page 50–72.
- Eyo Mensah and Eyamba Mensah. 2014. [The adaptation of english consonants by efik learners of english](#). *English Language Teaching*, 7(3).
- Eyo O. Mensah and Vivian Afi Dzokoto. 2023. [Melting intestines, red hearts, and scattering eyes: exploring embodiment in the efik feeling lexica](#).
- Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Tajudeen Kolawole, Taiwo Fagbohunbe, Solomon Oluwole Akinola, Sham-suddee Hassan Muhammad, Salomon Kabongo, Salomey Osei, Sackey Freshia, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo¹, Orevaoghene Ahia, Musie Meressa, Mofe Adeyemi, Masabata Mokgesi-Seling, Lawrence Okegbemi, and 29 others. 2020. [Participatory research for low-resourced machine translation: A case study in african languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *arXiv preprint arXiv:2207.04672*.
- P. Noah. 2004. Toward a more phonemic efik orthography. In O. Ndimele, editor, *Language and Culture in Nigeria. A Festschrift for Okon Essien*, pages 587–594. Emhai Press, Port Harcourt.
- Elizabeth Akpanke Odey, Emmanuel E. Etta, Louis Ajom Edet, Ikike I. Ufford, Columba Apeh Apeh, James Ajang Aboh, Essien Udoka Edem, Alex Abang Ebu, Eni Eja Alobo, Miebaka Nabiebu, and Gabriel Etim-Ben Inyang. 2025. Efik cultural context in language acquisition: A comparative study of immersion versus traditional methods. *Forum for Linguistic Studies*, 7.
- Offiong Ani Offiong and Stella Ansa. 2013. [The efik language: A historical profile](#). *Research in Humanities and Social Sciences*, 3(6).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, page 311–318.
- Maja Popovi^ć. 2015. [chrF: character n-gram f-score for automatic mt evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, page 392–395.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *arXiv preprint arXiv:1706.03762*.
- Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. [mt5: A massively multilingual pre-trained text-to-text transformer](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 483–498. Association for Computational Linguistics.

Reasoning Beyond Labels: Measuring LLM Sentiment in Low-Resource, Culturally Nuanced Contexts

Millicent Ochieng¹, Anja Thieme¹, Ignatius Ezeani², Risa Ueno¹, Samuel Maina¹, Keshet Ronen³, Javier González¹, Jacki O’Neill¹

¹Microsoft Research, ²Lancaster University, ³University of Washington

Abstract

Sentiment analysis in low-resource, culturally nuanced contexts challenges conventional NLP approaches that assume fixed labels and universal affective expressions. We present a diagnostic framework that treats sentiment as a context-dependent, culturally embedded construct, and evaluate how large language models (LLMs) reason about sentiment in informal, code-mixed WhatsApp messages from Nairobi youth health groups. Using human-annotated data, sentiment-flipped counterfactuals, and rubric-based explanation evaluation, we probe LLM interpretability, robustness, and alignment with human reasoning. Framing our evaluation through a social science measurement lens, we operationalize LLM outputs as an instrument for measuring the abstract concept of sentiment. Our findings reveal significant variation in model reasoning quality, with top-tier LLMs demonstrating greater interpretive stability, while smaller open-weight models in our study show reduced stability under ambiguity or sentiment shifts. This work highlights the need for culturally sensitive, reasoning-aware AI evaluation in complex, real-world communication.

1 Introduction

Sentiment analysis is a prevalent NLP technique used to obtain meaningful information and semantics from text (Onyenwe et al., 2020). It is often conflated with emotion detection (Nandwani and Verma, 2021); or opinion mining such as consumer sentiment (Burnham, 2024). Instead, sentiment analysis primarily determines polarity in the intent behind a written message often characterized

as positive, negative, or neutral (Nandwani and Verma, 2021).

Speech Act Theory by Austin (1975) further highlights that what a reader understands from a message depends on words choices; their individual meanings; ordering; as well as lexical or syntactic variations. Combined, these introduce significant ambiguity on how inferences about a message are drawn (Corvi et al., 2025). Moreover, the social semiotics theory by Halliday (2014) emphasizes that sentiment is not just a linguistic phenomenon; it is also deeply embedded in social and cultural contexts, which highlights how emotions are conveyed and interpreted based on cultural norms and values (Zhang, 2024).

In this paper, we acknowledge that interpreting or measuring sentiment can be difficult—particularly in informal, multilingual, under-resourced, and culturally nuanced communication contexts. Expressions of emotion and attitude are shaped by local language practices, shared cultural knowledge, and interactional context (Matsumoto, 1990; Lindquist, 2021; Fang et al., 2022). In real-world communications such as youth chat, social media, or hyperlocal exchanges among multilingual speakers, language is frequently code-mixed¹, fluid, and shaped by the moment—that is, influenced by who is speaking and who is listening, the topic being discussed, the speaker’s emotional tone, or intentions at that time, and the setting (e.g., online chat vs. in-person talk). Meanings are negotiated, implicit, and frequently ambiguous—making sentiment difficult to interpret, even for humans, espe-

¹The practice where multilingual speakers fluidly shift between languages in conversation

cially when removed from the original platform or context of exchange (O’Neill and Martin, 2003). These complexities do not just complicate classification—they challenge the very *measurement* of sentiment. As argued by Wallach et al. (2025), evaluating GenAI models requires treating such tasks as a social science measurement challenge, where abstract, culturally-contested concepts must be systematically defined and carefully connected to observable indicators.

In our work, we treat **sentiment** not as a fixed label, but as a context-dependent expression of intent. It may be explicit (e.g., “I’m so angry right now”), but more often in our dataset, it appears through muted cues (e.g., “You’re always online”)—subtle, culturally and contextually situated, and open to interpretation. We define **ambiguity** as cases where the intended sentiment is unclear, underspecified, open to multiple readings, or leads to disagreement even among culturally fluent, context-aware annotators—not because the language is misunderstood, but due to differing interpretations of tone or social context (see Table 1).

We use **cultural nuance** to describe how language practices, religious or affective expressions, and shared social knowledge shape how sentiment is conveyed and perceived. In our dataset, such nuance is embedded within: practices of *code-mixing* (e.g., “kama hauko school shindaapo!!”)²; *local shorthand* (e.g., mm for mimi)³; *emoji-only or emoji-enhanced* messages via graphical symbols (e.g., “😞”) or their textual counterpart the *emoticon* (e.g., “:”) (Liu et al., 2021; Yoo and Rayz, 2021); *irony*; and *youth-specific slang* (Sheng)⁴. These elements are often combined to produce rich, but difficult-to-classify, sentiment signals; and these cultural complexities are evident throughout our dataset (see Table 1 and Appendix Table 11)

So far, standard sentiment analysis treats sentiment as a fixed classification problem with a single, context-independent “ground truth” (Mohammad, 2017; Wankhade et al., 2022; Sharma et al., 2024). Recent exploratory work on similar multilingual, code-mixed WhatsApp data has examined LLM sentiment classification and qualitative reasoning (Ochieng et al., 2025), but with-

out a diagnostic framework, robustness testing, or measurement-oriented analysis. Despite advances in LLMs (Brown et al., 2020; Touvron et al., 2023; Jiang et al., 2023; OpenAI, 2023), sentiment evaluation remains label-centric, with metrics like accuracy and F1 obscuring how models reason and whether their decisions align with human interpretation.

We propose a diagnostic approach to sentiment analysis that treats LLMs not only as classifiers but as tools for structuring and probing sentiment in complex, real-world communication. Informed by Wallach et al.’s measurement framework (Wallach et al., 2025) that separates the *conceptualization* of sentiment from its *operationalization*. Our goal is to shift how sentiment is measured in LLMs—from fixed label prediction toward a more interpretive, ambiguity-aware framework. We ask: How do LLMs reason about sentiment in real-world, culturally grounded messages?

To achieve this, we investigate how LLMs reason about sentiment, how they explain their judgments, handle ambiguity, and echo human disagreement. For instance, while a traditional classifier might label “sawa tu 😞”⁵ as neutral, our framework surfaces the emotional nuance by analyzing emoji, tone, and context, revealing how such utterances can signal quiet frustration or withdrawal. We analyze model explanations, confidence scores, and token-level highlights indicating which parts of the message influenced the model’s judgment, across three evaluation settings: messages with annotator agreement (Gold), disagreement (Ambiguous), and **sentiment-flipped counterfactuals**⁶ (Synthetic). These *synthetic* examples are automatically generated by prompting GPT-4 to rewrite real WhatsApp messages in our dataset with their *sentiment flipped* (*positive to negative or vice versa*)—while preserving meaning, cultural tone and informal language. We guide this process using a structured taxonomy of sentiment-bearing components (e.g., negation, emoji, tone, key phrases; see Appendix A.8, Table 12). These counterfactual flips serve as our operationalization of the sentiment concept. Applied in testing whether models respond appropriately to affective changes, we use a dual evaluation protocol with human annotators and LLM-as-a-judge to assess counterfactual plausibility and explanation

²Swahili-English: “If you’re not in school, stay there.” While casual, this often conveys a dismissive stance, reflecting cultural norm that link education to intellectual legitimacy.

³*mimi* means “me” in Swahili

⁴An urban slang spoken by youth in Kenya, blending Swahili, English, and local languages.

⁵*sawa tu* means “just okay,” but can imply resignation or frustration depending on tone and context.

⁶A *counterfactual* is a sentiment-flipped variant of a real message.

Example	Complexities	Annotator 1	Annotator 2
<p>🤔🤔🤔 uyu sasa anachoma manzee “🤔🤔🤔 this guy is now messing up, bro”</p>	<p>Shorthand: “uyu” instead of “huyu” Urban slang (Sheng): “anachoma”, “manzee” Tone: friendly teasing, mockery, or social critique Emoji use: 🤔 Code-mixing: Swahili-Sheng blend Cultural reference: Assumes shared understanding of local slang, social behaviors, and norms</p>	<p>Label: Negative Notes: We see the ridicule and embarrassment from the persona and the audience despite the laugh.</p>	<p>Label: Positive Notes: Expresses criticism with amusement portrayed with laughing emojis.</p>
<p>Nmeacha izea “I’m sorry, I have stopped”</p>	<p>Code-mixing: Swahili-Sheng blend Urban slang (Sheng): “izea” Shorthand: “nmeacha” Tone: flat or understated Ambiguity: lacks strong emotional cues</p>	<p>Label: Neutral Notes: We see a casual apology that doesn’t express strong emotion.</p>	<p>Label: Positive Notes: Speaker is apologetic and remorseful.</p>
<p>U can’t see the future but God can “You can’t see the future but God can”</p>	<p>Shorthand: “U” for “you” Religious expression: appeals to divine foresight Tone: factual or reassuring Cultural context: common in faith-based communication Ambiguity: sentiment depends on interpretation of tone/intention</p>	<p>Label: Neutral Notes: A remark without strong personal emotion.</p>	<p>Label: Positive Notes: Speaker expresses trust in God, offering reassurance.</p>
<p>Hello, guys yani mko tu na mmenyamaza?? “Hello, guys are online and you are quiet?”</p>	<p>Code-mixing: Swahili-English blend Tone: questioning, possibly sarcastic Social cue: expectation of group participation Ambiguity: tone varies between concern and frustration</p>	<p>Label: Negative Notes: We see disappointment and negative shock from the persona on why people are so quiet.</p>	<p>Label: Neutral Notes: Expresses concern and curiosity on the silence of the group.</p>
<p>Yes I eat too much iz it normal “Yes I eat too much is it normal”</p>	<p>Shorthand: “iz” for “is”, informal tone Self-disclosure: reveals possible worry Ambiguity: phrased as a question, unclear tone; genuine concern vs casual comment</p>	<p>Label: Negative Notes: We see worry and distress about too much eating, suggests a negative sentiment.</p>	<p>Label: Neutral Notes: Question seeking clarification.</p>

Table 1: Examples of annotator disagreement illustrating cultural and linguistic complexities.

quality.

This paper makes the following contributions:

- We adapt a social science measurement lens to evaluate model reasoning about language, reframing sentiment analysis as a problem of concept systematization and measurement.
- We introduce a diagnostic framework to analyze how LLMs reason about sentiment in informal, code-mixed, and culturally embedded communication. This involves creating synthetic data using a counterfactual approach based on a taxonomy of sentiment components (e.g., negation, emoji, tone).
- We propose a dual evaluation protocol with human annotators and an LLM-as-a-judge to assess explanation quality and counterfactual plausibility. Through this, we identify reasoning inconsistencies in LLMs, distinguishing between reducible errors and irreducible ambiguity across evaluation settings.

2 Related Work

Sentiment Analysis in Informal and Multilingual Communication: While sentiment analysis has largely focused on English-language data from structured domains such as reviews or news,

real-world communication in informal, multilingual, and code-mixed contexts presents deeper challenges (Choudhary et al., 2018). Prior work on code-mixed sentiment (e.g., Swahili-English, Hindi-English) has highlighted the need for inclusive resources (Zhang et al., 2023; Doğruöz et al., 2023a,b; Kaji and Shah, 2023), yet low-resource, conversational data in health or community settings remains underexplored. Recent exploratory work on multilingual, code-mixed WhatsApp data from Nairobi youth examined LLM-based sentiment classification and qualitative reasoning using standard metrics and manual inspection (Ochieng et al., 2025). Building on this line of work, we move beyond exploratory analysis by introducing a structured diagnostic framework that evaluates how LLMs reason about sentiment across diverse evaluation settings, explicitly models ambiguity, and probes robustness via sentiment-flipped counterfactuals assessed with shared human and LLM-based rubrics.

Evaluating LLM Reasoning: Traditional sentiment evaluation relies on metrics like accuracy and F1, which fail to capture how models reason, especially in ambiguous or culturally situated cases (Lyu et al., 2024). To address this, recent work has explored explanation-based evaluation through token attribution, rationales, and confidence scores (Joshi et al., 2023; Dhaini et al., 2025). Other work

has shown that LLMs like GPT-4 can serve as evaluators, often approximating human ratings in generation tasks (Liu et al., 2023). However, a missing component in this literature is the use of dual evaluation protocols that involve both human and LLM judges applying shared rubrics. Such approaches are particularly valuable in settings with annotator disagreement, where interpretive alignment matters more than single-label accuracy. Our work builds on and extends this direction by systematically comparing model and human evaluations across diverse examples, including ambiguous and counterfactually altered messages.

Counterfactuals and Contrastive Evaluation in NLP: Counterfactuals offer a powerful tool for probing model reasoning by introducing minimal, targeted changes to input data (Yang et al., 2021). In sentiment analysis, these typically flip polarity through shifts in tone, negation, or word choice. While prior work often relied on rule-based or synthetic constructions (Yang et al., 2021), we use GPT-4 to generate sentiment-flipped versions of messages—shifting from positive to negative and vice versa—grounded in a taxonomy of transformation types such as emoji use, phrase substitution, and tone modulation. More broadly, our approach aligns with recent work on problem variation as a diagnostic for reasoning (Xu et al., 2025), which emphasizes the need for systematic, multi-level perturbations, including counterfactuals to reveal model limitations beyond memorization.

3 Evaluation as Measurement: Experimental Setup

3.1 Dataset and Annotation

We build on the WhatsApp Chat Dataset originally collected by Karusala et al. (2021) and annotated by Mondal et al. (2021), which comprises multilingual conversations among young people living with HIV in informal settlements in Nairobi, Kenya. These discussions, drawn from two health-focused WhatsApp groups moderated by a medical facilitator, are informal, context-rich, and code-mixed across English, Swahili, and Sheng. All messages were anonymized, and ethical protocols from the original collection were strictly followed. The dataset is not publicly released due to sensitivity, but researchers may request access for academic use.

For this study, we developed a structured annotation protocol focused on culturally grounded sentiment, interpretive ambiguity, and context-specific

expression. Designed through iterative pilot testing and calibrator discussions (see Appendix A.1). Two trained annotators — Kenyan youth aged 20–24 — labeled each message for sentiment (positive, negative, neutral), provided English translations where needed, and tagged word-level language identifiers. Messages with annotator disagreement were retained for targeted evaluation. From the full dataset of 6,197 messages, we define three evaluation subsets: the **Gold Set** (6,121 messages with full annotator agreement), the **Ambiguous Set** (76 messages with disagreement), and the **Synthetic Set** (sentiment-flipped messages generated from a pool of 1,547 non-neutral messages using GPT-4), see Table 5. No post-processing is applied to normalize emojis, punctuation, or shorthand expressions, as these elements are integral to the communicative and emotional tone of the data.

3.2 Task and Model Setup

We frame sentiment analysis as a multi-class classification task over informal, multilingual WhatsApp messages. Given an input message, the model is prompted to predict a sentiment label (positive, negative, or neutral) and to generate a natural language explanation (max 200 words). The task is performed via in-context learning using few-shot prompting, with manually selected examples from the Gold Set that reflect the natural mix of Swahili, Sheng, and English, including both clear and mildly ambiguous cases (see Table 7 in the Appendix). Pilot comparisons of three prompting strategies (definitions with examples, definitions only, and no definitions) showed that prompts combining definitions and examples yielded the most stable and interpretable outputs. We evaluate a range of LLMs varying in architecture and size, including GPT-4-Turbo and GPT-4-32k (OpenAI, 2023), Gemma-3-27B (Team et al., 2025), LLaMA-3-8B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), OpenChat-3.5 (Wang et al., 2024), and Phi-4 (Gunasekar et al., 2023). We selected models to reflect realistic usage in low-resource settings, contrasting locally deployable lightweight open-weight models with large proprietary models accessed via APIs. Phi-4 was included despite its English-only marketing due to strong pilot performance on code-mixed Swahili–English text. Each model outputs a sentiment label, a natural language explanation, token-level highlights, and a confidence score⁷ scaled

⁷Model-reported confidence where available.

from 0 to 5. Evaluation is conducted across three data partitions: Gold, Ambiguous, and Synthetic.

3.3 Counterfactual Generation Framework

We use sentiment-flipped counterfactuals as a diagnostic stress test, aligned with hypothesis validity testing (Wallach et al., 2025), to evaluate whether models detect and explain controlled shifts in sentiment rather than to construct a new gold-standard dataset. Starting from 1,547 non-neutral messages in the Gold Set, we prompt GPT-4 to generate three sentiment-flipped variants per message, reversing the original polarity (positive to negative or vice versa) while preserving meaning, tone, and conversational style. Some messages require minimal lexical changes, for example, *Napenda wazo lako* (“I like your idea”) → *Sipendi wazo lako* (“I dislike your idea”). Others, however, demand deeper shifts in intent or tone, such as *Sema tuu niache kukuaibisha* (“Just tell me to stop embarrassing you”) → *Sema tu niendeleo kukusifu* (“Just say it, so I continue praising you”). To guide generation across this range of complexity, we developed a taxonomy of sentiment-bearing components, including negation, tone, emoji, and sentiment phrases, which informs the generation prompt (Table 12). Rather than manually selecting outputs, we apply a second GPT-4 prompt that selects the strongest candidate based on plausibility, fluency, and contextual fit (Table 9). This two-step process allows for richer variation at generation time while promoting interpretability and consistency in the resulting counterfactuals. The selected flips constitute the Synthetic Set used for robustness evaluation. Human assessment of a subset of these counterfactuals is conducted later to audit quality and surface limitations (Section 3.4). We discuss the implications of relying on LLM-based generation and filtering in the Limitations Section.

3.4 Human and LLM-as-a-judge Evaluation Protocol

We evaluate model explanations and Synthetic counterfactuals using a structured, rubric-based protocol involving human annotators and GPT-4 as an automated judge. This dual evaluation is designed to assess interpretive quality and counterfactual plausibility, and to compare human and model judgments. For model explanations, two annotators independently rated 480 explanations drawn from the Gold (180), Ambiguous (120), and Synthetic (180) sets. All six models were included where

explanations were available. For the Ambiguous set, only four models (LLaMA-3-8B, GPT-4-Turbo, GPT-4-32k, and Gemma-3-27B) consistently produced usable outputs, reflecting the difficulty of these cases. Explanations were scored on faithfulness, contextual or cultural appropriateness, logical coherence, and clarity or completeness using a binary (0/1) scale. To audit the quality of the Synthetic Set, six annotators evaluated a sample of 50 sentiment-flipped messages on fluency, naturalness, sentiment flip clarity, and meaning preservation, also using a binary (0/1) scale. This human assessment is intended as a diagnostic quality check that surfaces limitations such as semantic drift or stylistic mismatch, rather than as an exhaustive validation of all generated counterfactuals. GPT-4 was prompted to apply the same rubrics using standardized evaluation instructions (Tables 10 and 8), allowing direct comparison between human and LLM judgments. Within the measurement framework of Wallach et al. (2025), this protocol corresponds to the *interrogation* step, enabling analysis of content validity (alignment between explanations and the sentiment concept) and consequential validity (how explanation quality affects interpretation and use). Full rubric definitions and example annotations are provided in Appendix A.7.

4 Results and Analysis

4.1 Overall Model Performance

For predicting sentiment labels as a baseline, we observe that model coverage⁸ varies substantially across settings, especially under counterfactual perturbation revealing a key axis of performance variation (see Table 6). While top-tier models like GPT-4-Turbo and GPT-4-32k, consistently provide labels for all examples (100% coverage), several open-weight models—most notably LLaMA-3-8B—show sharp declines, especially in the Synthetic set, where coverage drops as low as 37.6%. This sharp drop suggests that even fluent, sentiment-flipped rewrites can disrupt model processing, exposing model fragility to subtle language changes in tone, emoji, or phrasing.

We further observe that on the Gold Set, all models achieve strong average F1 scores. The best performance is observed from GPT-4-32k (0.90), Mistral-7B (0.90), and Gemma-3-27B (0.89). Most models maintain balance across sen-

⁸Coverage reflects the percentage of examples for which a model returned a valid sentiment label.

timent classes, but class-specific performance still varies. LLaMA-3-8B underperforms markedly on negative sentiment (0.51), pointing to difficulty detecting more implicit or culturally nuanced negativity. Neutral sentiment is generally the most challenging class, echoing prior findings on underspecified affect and implicit tone. These results establish strong baselines while highlighting gaps in both robustness and class sensitivity that motivate further analysis of model reasoning and explanation quality.

4.2 Reasoning Quality in LLM Explanations

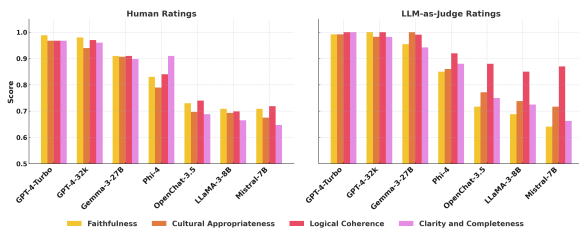


Figure 1: Rubric-based average explanation scores across models.

We evaluated explanation quality of the models reasoning about a message’s sentiment using rubric-based scores from both human annotators and GPT-4-based LLM-as-a-judge systems, see Figure 1. Across all models and dimensions, we observe broad agreement in relative rankings between the two rating sources, though LLM-as-a-judge ratings tend to be more generous overall. GPT-4-32k and GPT-4-Turbo consistently achieve top scores across all rubrics, with near-perfect ratings from both humans and LLMs. Gemma-3-27B also performs well, with high ratings for faithfulness, logical coherence and cultural appropriateness, though with modest drops in clarity. By contrast, Phi-4, OpenChat-3.5, LLaMA-3-8B, and Mistral-7B show significantly lower performance, particularly on faithfulness and clarity—dimensions most sensitive to hallucination and underspecification. Human raters were notably stricter in these areas, especially for open-weight models, revealing that LLM-based evaluations may overestimate explanation quality. Despite these differences in score magnitude, the rubric-level trends are consistent: Logical Coherence is the strongest dimension across most models, while Faithfulness, Cultural Appropriateness and Clarity & Completeness expose key weaknesses in less capable systems. Interestingly, Mistral-7B, which led in clas-

sification F1, ranks lowest in explanation quality by both rating sources, highlighting a persistent disconnect between predictive accuracy and reasoning quality. Conversely, the strongest models (GPT-4 variants and Gemma) exhibit both high classification performance and robust explanatory reasoning. These findings emphasize the importance of explanation-focused evaluation, as high task accuracy alone may mask serious limitations in model understanding and reasoning.

4.3 Probing LLM Robustness to Synthetic Set (Counterfactual Flips)

Criterion	Human	LLM-as-a-judge
Fluency	0.89	1.00
Naturalness	0.68	0.97
Flip Clarity	0.79	0.98
Meaning Preservation	0.78	0.58

Table 2: Average rubric-based scores for synthetic flips.

We categorized each counterfactual by its main transformation and found that flips most commonly altered sentiment-bearing keywords, phrases, tone, and emoji—components central to both explicit and stylistic sentiment signaling, see Figure 3. Less frequent were transformations involving negation, intent framing, or valence modulation, which require more interpretive reasoning. From our analysis, GPT-4 often produced plausible synthetic flips (see examples in Table 14). We assessed the quality of the synthetic flips using rubric-based judgments from both human annotators and LLMs-as-Judges (GPT-4-Turbo and GPT-4-32k), see Table 2. LLM ratings were uniformly high—near-perfect in fluency, naturalness, flip clarity, and slightly lower for meaning preservation. Human annotators, however, were notably stricter, especially in naturalness and meaning preservation, revealing significant gaps in how surface-level and semantic quality are perceived. In particular, humans flagged many cases as semantically incorrect or stylistically unnatural, despite their formal fluency. Manual analysis revealed that *positive-to-negative* flips posed greater challenges. LLMs frequently overcorrected, introducing harsh or exaggerated tone, especially in code-mixed inputs (see Table 15). Conversely, *negative-to-positive* flips tended to be smoother and more culturally appropriate. While human raters penalized positive-to-negative flips for harsh tone or topic drift, LLMs-as-Judges often gave high marks even in such cases—suggesting they were

less sensitive to subtle shifts in meaning or register. While the flipped sentiment was often correct, the model struggled with non-English and code-mixed inputs, frequently normalizing local shorthand, translating content into English, or rewriting messages in Standard Swahili (Kiswahili Sanifu), thereby altering the original language composition (see third example in Table 14).

Model	Eff. F1 (Pre-CF)	Eff. F1 (Post-CF)	Δ Post-Pre
GPT-4-Turbo	0.960	0.980	+0.020
GPT-4-32k	0.970	0.980	+0.010
Phi-4	0.940	0.786	-0.154
Gemma-3-27B	0.940	0.466	-0.474
Mistral-7B	0.892	0.466	-0.425
OpenChat-3.5	0.910	0.441	-0.469
LLaMA-3-8B	0.783	0.349	-0.434

Table 3: Effective F1 before and after counterfactual sentiment flips.

To quantify model robustness under transformation, we compute *Effective F1*—the product of F1 and coverage. As shown in Table 3, both GPT-4-Turbo and GPT-4-32k maintained high or improved post-flip performance (up to 0.980). In contrast, mid-sized and open models suffered significant drops (0.40–0.47), driven by both misclassification and partial outputs. Notably, Phi-4 preserved coverage but underperformed on positive flips, indicating brittle generalization. Beyond label accuracy, explanation quality further reveals this fragility. On the Synthetic Set, only the GPT-4 variants produced consistently faithful, coherent, culturally grounded, and context-sensitive reasoning. Other models often generated fluent but incorrect explanations after sentiment was flipped, with sharp drops in faithfulness and completeness—especially for Mistral-7B, OpenChat-3.5, and LLaMA-3-8B (Table 16).

4.4 How does model confidence and alignment reflect interpretive ambiguity?

Model	Avg. Conf.	Coverage (%)	Eff. Conf.
GPT-4-Turbo	4.639	100.0	4.64
GPT-4-32k	4.440	100.0	4.44
Phi-4	4.711	99.5	4.69
Gemma-3-27B	4.698	47.6	2.24
OpenChat-3.5	4.249	47.4	2.01
Mistral-7B	4.132	47.6	1.97
LLaMA-3-8B	3.981	37.6	1.50

Table 4: Effective Confidence on the Synthetic Set.

To assess confidence calibration, we report average model confidence and coverage across the Gold and Synthetic Sets (Table 17). While most

models maintain high confidence on the Gold Set, only the GPT-4 variants and Phi-4 sustain both high confidence and near-complete coverage on counterfactual inputs. In contrast, models like Gemma-3-27B and OpenChat-3.5 appear overconfident despite skipping over half of the flipped messages. To quantify this further, we compute an Effective Confidence score (confidence \times coverage), reported in Table 4, revealing a sharp drop for open models—underscoring their brittleness under minimal sentiment shifts. Although the Gold Set contains messages with full human agreement, models show only moderate alignment with one another. As shown in Figure 2, the highest agreement is observed between Gemma-3-27B and Phi-4 ($\kappa = 0.73$), and between GPT-4-Turbo and GPT-4-32k ($\kappa = 0.70$). However, other pairings show weaker agreement—such as GPT-4-Turbo and Mistral-7B ($\kappa = 0.48$)—despite similar average F1 scores. This suggests that even on “clear” cases, LLMs diverge in interpretation, reflecting differences in how they weigh tone, cues, and cultural context.

5 Discussion

LLMs-as-Generators: Crafting Cultural Counterfactuals Using GPT-4 to generate sentiment-flipped counterfactuals revealed both the model’s strengths and its limitations. Often, it produced fluent, contextually appropriate flips that successfully reversed sentiment while preserving tone and informal style. However, our diagnostic analysis surfaced key weaknesses. Flips from positive to negative frequently introduced exaggerated emotional intensity, suggesting the model struggles to calibrate negative sentiment in subtle, conversational contexts. Additionally, while GPT-4 provided self-reported labels for the components it modified (e.g., tone, emoji, phrasing), these attributions were often imprecise or inconsistent. These findings underscore both the potential and fragility of using LLMs to generate culturally grounded synthetic data—and highlight the continued need for more iteration in prompt instructions as well as human oversight when precision over tone, meaning, and linguistic structure is essential.

LLMs-as-Judges: Evaluating Counterfactuals

We used GPT-4 as a judge to assess the quality of sentiment-flipped messages—selecting the best rewrite among three generated variants and then scoring the selected flip for fluency, naturalness,

meaning preservation, and successful sentiment reversal. This approach streamlined evaluation and scaled the generation pipeline. In many cases, GPT-4’s selections aligned with the human judgments. However, because these decisions rely entirely on the model’s internal criteria, we observed inconsistencies—especially for non-English messages with culturally layered meaning. For instance, some selected flips introduced subtle shifts in tone or more formal phrasing, reducing cultural fidelity even when sentiment was accurately reversed. In other cases, we observe that GPT-4 successfully produced plausible flips that changed a message’s perceived sentiment, this was achieved in different ways, which do not necessarily reflect the most *minimal* changes to achieve that effect. For example, flipping “Hahaha” (+) could be achieved by “Not funny” (-) or “Ughhh,” (-) or “This is not funny at all” (-). These insights suggests that additional checks should be put into place already at the filtering step to assess if flips are indeed consistent with the tone, phrasing, language composition or cultural meaning of the original message to ensure chosen variants are truly the most faithful transformations. These findings point to the need for human-in-the-loop validation at each stage—particularly when using LLMs to adjudicate nuanced, multilingual language in low-resource settings.

Prediction is not ‘understanding’ Models such as Mistral-7B, Phi-4, and OpenChat-3.5 score competitively on standard metrics, yet generate explanations that often lack coherence, faithfulness, or cultural grounding—especially in cases where sentiment is subtle, indirect, or stylistically embedded, as revealed by human evaluation. These reasoning gaps become even more pronounced under sentiment counterfactuals, with flipped affect lead to sharp performance drops—up to 0.47 F1 for open-weight models—exposing brittle generalization to plausible shifts in tone, emoji, or phrasing. In contrast, GPT-4-Turbo and GPT-4-32k demonstrate greater robustness in both prediction and reasoning, suggesting that scale and stronger instruction tuning support more stable reasoning.

LLMs vary not just in accuracy, but in world-view Agreement scores between models remain low, even on the Gold Set, where human annotators were unanimous. This divergence reflects not just model sensitivity to surface cues, but deeper differences in how LLMs encode sentiment pri-

ors, cultural nuance, and conversational style. That GPT-4-Turbo and Mistral-7B can yield similar F1 yet diverge in label agreement ($\kappa = 0.48$) illustrates that we are not simply comparing better vs. worse models, but different interpretive frameworks. However, we do not understand the models underlying interpretive frameworks, and how well it maps to existing theory, and consistency in reasoning varies significantly across models, especially open-weight models.

Confidence is not calibration While average confidence scores remain high across models, only OpenAI’s models (GPT-4-Turbo and GPT-4-32k) consistently maintain high confidence, full coverage on perturbed data, accurate predictions, and reliable reasoning. In contrast, models such as Phi-4 also exhibits high confidence and broad coverage, but manual inspection reveals frequent reasoning errors, highlighting a gap between confidence and correctness.

Annotation as a site of interpretive complexity Our study highlights the complexities of designing robust annotation protocols for nuanced, real-world data. Annotators frequently encountered edge cases that exposed ambiguity in how sentiment should be labeled, especially when affect was culturally or contextually embedded. This reinforces growing recognition in human-centered NLP that annotation is an interpretive process requiring iteration, theoretical grounding, and thoughtful handling of disagreement.

Sentiment is structured by context Our work challenges simplified views of sentiment as binary or fixed, framing it instead as context-dependent and semantically layered. While our initial definition in the annotation protocol and component taxonomy aimed to capture more nuance, more specification is needed. For example, *context-dependency* emerged as central to interpretation, as seen in our annotation examples. There are many aspects that can shape what context-dependency as an element of sentiment means. As illustrated through our study, context can be informed by: the conversation topic (e.g., health advice); cultural norms (e.g., in Kenya); or religious cues; as well as other interpersonal dynamics (e.g., what the recipient of a message assumes or knows about its writer) that can be harder to capture or specify. Yet, future work will need to expand efforts to further systematize and formalize those components of sen-

timent to be able to achieve more robust evaluation approaches.

6 Conclusion

We reframe sentiment analysis in low-resource, culturally nuanced contexts as a problem of reasoning, not just classification. Using a diagnostic framework grounded in social science measurement, we evaluate how LLMs interpret sentiment in multilingual, code-mixed WhatsApp messages from Nairobi youth health groups. Our findings reveal that while top-tier LLMs demonstrate interpretive robustness, open models often fail under ambiguity and cultural nuance, highlighting deep gaps in reasoning quality. As sentiment increasingly becomes a benchmark task for real-world NLP, our work urges a shift from fixed-label accuracy to context-aware, culturally grounded evaluation. Future sentiment systems must be judged not only by what label they assign, but how and why they reason that way.

Limitations

While our diagnostic framework offers a deeper lens into sentiment reasoning, several limitations remain:

(1) Sentiment itself remains an inherently subjective construct. Our LLM-guided systematization of text components like negation, tone, emojis, keywords and phrase rewordings look reasonable (face validity) and may capture the most salient aspects of the sentiment concept (content validity). However, further research is needed to inspect whether this systematization fully specifies all observable criteria connected to sentiment (substantive validity) (Wallach et al., 2025); as well as how the components may relate to one-another; and whether its operationalization via LLM-as-a-judge is consistent and coherent with the LLMs internal interpretation of these components.

(2) Our counterfactual generation pipeline uses a two-stage prompting process: GPT-4 first generates three flipped variants of a message, then selects the most plausible one for inclusion. While this filtering step improves fluency and contextual fit, it relies entirely on the model’s internal criteria, which we do not independently validate. Future work should investigate how this selection process affects flip quality, what may be lost or altered during filtering, and incorporate human-in-the-loop checks to ensure that selected flips accurately re-

flect the intended sentiment transformation and preserve linguistic and contextual fidelity.

(3) Our study focuses on a single, culturally specific dataset of health-related WhatsApp messages from Nairobi youth. While this setting is intentionally chosen to surface ambiguity and contextual nuance, it limits direct generalization to other populations or sociolinguistic contexts. We view the framework itself as transferable, but its application to other code-mixed languages, age groups, or cultural settings remains an important direction for future research.

Ethical Consideration

This study uses anonymized WhatsApp messages from Nairobi youth health groups, collected with consent under prior research protocols. All data were reviewed to remove identifying information and sensitive content. Our use of LLMs to generate synthetic sentiment data in a code-mixed, culturally grounded setting raises important ethical considerations. Language reflects identity, and synthetic rewrites, especially in informal, multilingual contexts must be handled with care to avoid erasing nuance or reinforcing stereotypes. While we designed prompts to preserve tone and intent, LLMs may still encode harmful biases. We emphasize the importance of cultural sensitivity, context-aware evaluation, and collaboration with local experts to ensure respectful and responsible analysis.

References

- John Langshaw Austin. 1975. *How to do things with words*. Harvard university press.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). *Advances in Neural Information Processing Systems*, 2020-December.
- Michael Burnham. 2024. What is sentiment meant to mean to language models? *Research & Politics*, 11(4):20531680241307941.
- Nurendra Choudhary, Rajat Singh, Ishita Bindlish, and Manish Shrivastava. 2018. [Sentiment analysis of](#)

- code-mixed languages leveraging resource rich languages.
- Emily Corvi, Hannah Washington, Stefanie Reed, Chad Atalla, Alexandra Chouldechova, P Alex Dow, Jean Garcia-Gathright, Nicholas Pangakis, Emily Sheng, Dan Vann, et al. 2025. Taxonomizing representational harms using speech act theory. *arXiv preprint arXiv:2504.00928*.
- Mahdi Dhaini, Kafaite Zahra Hussain, Efstratios Zaradoukas, and Gjergji Kasneci. 2025. Evalxnlp: A framework for benchmarking post-hoc explainability methods on nlp models. *Preprint*, arXiv:2505.01238.
- A. Seza Dođruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2023a. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Annual Meeting of the Association for Computational Linguistics*.
- A. Seza Dođruöz, Sunayana Sitaram, and Zheng-Xin Yong. 2023b. Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. In *Conference on Empirical Methods in Natural Language Processing*.
- Xia Fang, Magdalena Rychlowska, and Jens Lange. 2022. Cross-cultural and inter-group research on emotion perception. *Journal of Cultural Cognitive Science*, 6:1–7.
- Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.
- Michael Alexander Kirkwood Halliday. 2014. Language as social semiotic. *The Discourse Studies Reader. Amsterdam: John Benjamins*, pages 263–272.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. *Mistral 7b*. *Preprint*, arXiv:2310.06825.
- Brihi Joshi, Aaron Chan, Ziyi Liu, Shaoliang Nie, Maziar Sanjabi, Hamed Firooz, and Xiang Ren. 2023. Er-test: Evaluating explanation regularization methods for language models. *Preprint*, arXiv:2205.12542.
- Arshad Kaji and Manan Shah. 2023. Contextual code switching for machine translation using language models.
- Naveena Karusala, David Odhiambo Seeh, Cyrus Mugo, Brandon L Guthrie, Megan Andreas Moreno, Grace C John-Stewart, Irene Inwani, Richard J. Anderson, and Keshet Ronen. 2021. “that courage to encourage”: Participation and aspirations in chat-based peer support for youth living with hiv. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.
- Kristen A. Lindquist. 2021. Language and emotion: Introduction to the special issue. *Affective Science*, 2:91–98.
- Chuchu Liu, Fan Fang, Xu Lin, Tie Cai, Xu Tan, Jianguo Liu, and Xin Lu. 2021. Improving sentiment analysis accuracy with emoji embedding. *Journal of Safety Science and Resilience*, 2(4):246–252.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, 50:657–723.
- David Matsumoto. 1990. Cultural similarities and differences in display rules. *Motivation and Emotion*, 14:195–214.
- Saif M. Mohammad. 2017. Challenges in sentiment analysis. pages 61–83.
- Ishani Mondal, Kalika Bali, Mohit Jain, Monojit Choudhury, Ashish Sharma, Evans Gitau, Jacki O’Neill, Kagonya Awori, and Sarah Gitau. 2021. A linguistic annotation framework to study interactions in multilingual healthcare conversational forums. In *Proceedings of the Joint 15th Linguistic Annotation Workshop (LAW) and 3rd Designing Meaning Representations (DMR) Workshop*, pages 66–77, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Pansy Nandwani and Rupali Verma. 2021. A review on sentiment analysis and emotion detection from text. *Social network analysis and mining*, 11(1):81.
- Millicent Ochieng, Varun Gumma, Sunayana Sitaram, Jindong Wang, Vishrav Chaudhary, Keshet Ronen, Kalika Bali, and Jacki O’Neill. 2025. Beyond metrics: Evaluating LLMs effectiveness in culturally nuanced, low-resource real-world scenarios. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 230–247, Vienna, Austria. Association for Computational Linguistics.
- Jacki O’Neill and David Martin. 2003. Text chat in action. In *Proceedings of the 2003 ACM International Conference on Supporting Group Work, GROUP ’03*,

page 40–49, New York, NY, USA. Association for Computing Machinery.

Ikechukwu Onyenwe, Samuel Nwagbo, Njideka Mbele-dogu, and Ebele Onyedinma. 2020. The impact of political party/candidate on the election results from a sentiment analysis perspective using# anambrade-cides2017 tweets. *Social Network Analysis and Mining*, 10(1):55.

OpenAI. 2023. [Gpt-4 technical report](#).

Neeraj Anand Sharma, A. B.M.Shawkat Ali, and Muhammad Ashad Kabir. 2024. [A review of sentiment analysis: tasks, applications, and deep learning techniques](#). *International Journal of Data Science and Analytics 2024* 19:3, 19:351–388.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenaly, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, Andrés György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Pettrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huiyenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shrivastava, Renjie Wu, Renke Pan, Reza Rokni, Rob

Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Pöder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025. [Gemma 3 technical report](#). *Preprint*, arXiv:2503.19786.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. [Llama: Open and efficient foundation language models](#).

Hanna Wallach, Meera Desai, A Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P Alex Dow, et al. 2025. Position: Evaluating generative ai systems is a social science measurement challenge. *arXiv preprint arXiv:2502.00561*.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024. [Openchat: Advancing open-source language models with mixed-quality data](#). *Preprint*, arXiv:2309.11235.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. [A survey on sentiment analysis methods, applications, and challenges](#). *Artificial Intelligence Review 2022* 55:7, 55:5731–5780.

Xinnuo Xu, Rachel Lawrence, Kshitij Dubey, Atharva Pandey, Fabian Falck, Risa Ueno, Aditya Nori, Rahul Sharma, Amit Sharma, and Javier González. 2025. [Re-imagine: Symbolic benchmark synthesis for reasoning evaluation](#). In *ICLR 2025 - Workshop on Reasoning and Planning for LLMs*.

Linyi Yang, Jiazheng Li, Pdraig Cunningham, Yue Zhang, Barry Smyth, and Ruihai Dong. 2021. [Exploring the efficacy of automatically generated counterfactuals for sentiment analysis](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 306–316, Online. Association for Computational Linguistics.

- Byungkyu Yoo and Julia Taylor Rayz. 2021. Understanding emojis for sentiment analysis. In *The international FLAIRS conference proceedings*, volume 34.
- Junfeng Zhang. 2024. Sentiment and language: A socio-semiotic analysis. *Philosophy Journal*, 3(1):118–127.
- Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023. [Multilingual large language models are not \(yet\) code-switchers](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

A Appendix

A.1 Annotation Protocol

Annotators and Process. This protocol was developed to guide consistent sentiment annotation of informal, multilingual WhatsApp messages exchanged among youth in Nairobi. Two trained annotators—both fluent in English, Swahili, and Sheng—applied the protocol over the course of one month. Annotation covered 6,197 messages drawn from a dataset of informal, code-mixed conversations among youth living with HIV. Work was conducted in Excel. Annotators labeled each message independently, treating it as a standalone utterance while considering cultural context, code-switching, and emoji use. We began with a jointly labeled calibration set of 100 examples, followed by independent annotation with regular meetings to discuss edge cases and resolve ambiguities.

Sentiment Categories. Messages were labeled as **Negative (-1)**, **Neutral (0)**, or **Positive (1)** based on expressed affect. Annotators were instructed to:

- Label as **Negative** if the message expressed frustration, sadness, criticism, or distress (e.g., *"I am tired !!!", "You have a mental problem"*).
- Label as **Neutral** if the message conveyed information, routine conversation, or general greetings without strong sentiment (e.g., *"When are you coming?"*, *"Good morning 🙌"*).
- Label as **Positive** if the message expressed joy, support, pride, or optimism (e.g., *"I'm much happy to interact and share with you guys!"*).

Ambiguity and Cultural Nuance. Annotators flagged ambiguous cases with written justifications. Given the culturally grounded and multilingual nature of the data, particular attention was paid to tone, idioms, emoji use, and context-specific expressions of affect.

A.2 Evaluation Subsets

Subset	Positive	Negative	Neutral	Total
Gold Set	1196	351	4574	6,121
Synthetic Set	351	1196	-	1,547
Ambiguous Set	-	-	-	76

Table 5: Sentiment-wise distribution of messages.

A.3 Overall Model Performance

Model	Pos	Neg	Neu	Avg	Cov. %
Gold Set (annotated Pos/Neg/Neu)					
GPT-4-Turbo	0.98	0.92	0.75	0.88	100.0
GPT-4-32k	0.93	0.90	0.86	0.90	100.0
Gemma-3-27B	0.93	0.96	0.79	0.89	100.0
Phi-4	0.93	0.91	0.80	0.88	100.0
Mistral-7B	0.91	0.88	0.92	0.90	98.9
OpenChat-3.5	0.93	0.77	0.87	0.86	99.9
LLaMA-3-8B	0.94	0.51	0.86	0.77	92.9
Pre-CF (original Pos/Neg examples)					
GPT-4-Turbo	0.98	0.94	—	0.96	100.0
GPT-4-32k	0.99	0.95	—	0.97	100.0
Gemma-3-27B	0.97	0.90	—	0.94	100.0
Phi-4	0.97	0.90	—	0.94	100.0
OpenChat-3.5	0.97	0.85	—	0.91	100.0
Mistral-7B	0.97	0.89	—	0.93	95.9
LLaMA-3-8B	0.97	0.77	—	0.87	90.2
Post-CF (synthetic counterfactuals)					
GPT-4-Turbo	0.97	0.99	—	0.98	100.0
GPT-4-32k	0.97	0.99	—	0.98	100.0
Phi-4	0.67	0.90	—	0.79	99.5
Gemma-3-27B	0.97	0.99	—	0.98	47.6
Mistral-7B	0.97	0.99	—	0.98	47.6
OpenChat-3.5	0.90	0.97	—	0.93	47.4
LLaMA-3-8B	0.91	0.96	—	0.93	37.6

Table 6: F1 scores by sentiment class on the Gold Set, Pre-CF (original positive/negative examples used to generate counterfactuals), and Post-CF (synthetic counterfactuals with flipped sentiment). Coverage rate (%) reflects the proportion of examples for which a model returned a valid sentiment label.

A.4 Prediction Agreement Across Models

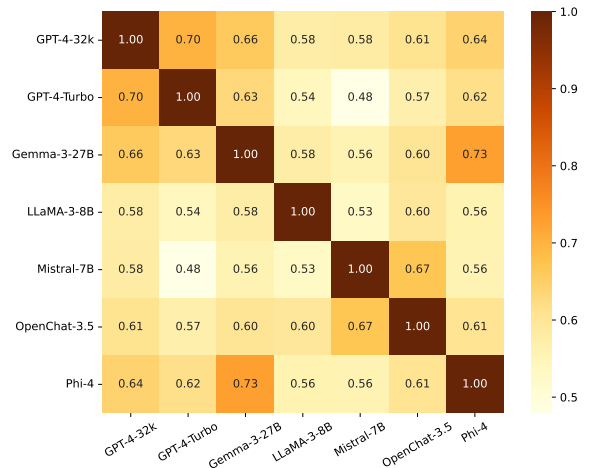


Figure 2: Cohen’s κ agreement between model predictions on the Gold Set. Despite full annotator agreement, models show only moderate pairwise consistency—indicating divergence in their underlying reasoning and sensitivity to sentiment cues.

A.5 Model Prompts

Sentiment Classification + Explanation Prompt
 You are an NLP assistant for sentiment analysis.
 Given a WhatsApp message (QUERY), classify its sentiment as Positive, Negative, or Neutral.
 Provide a justification using extracted keywords and a brief explanation.
 Return your confidence score (0-5). Use JSON output format only.

The prompt includes:

- Sentiment definitions (Positive, Neutral, Negative)
- Examples of clearly and ambiguously labeled messages
- JSON output format with keywords, explanation, label, and confidence

QUERY: "{query}"

Output Format:

```
{
  "justification": {
    "keywords": [ ... ],
    "explanation": "...",
    "sentiment": "...",
    "confidence_score": "..."
  }
}
```

Table 7: Instruction prompt for joint sentiment classification, justification, and confidence scoring.

Counterfactual Evaluation Prompt You are evaluating a synthetic (GPT-4-generated) version of a WhatsApp message. The synthetic message is a sentiment-flipped version of the original.
 Assess the quality of the synthetic message along four criteria using 0 or 1:

1. Fluency - Is the synthetic message grammatically correct and readable?
2. Naturalness - Does it sound plausible for a human to write?
3. Sentiment Flip Clarity - Is the sentiment clearly flipped from the original?
4. Meaning Preservation - Is the core meaning preserved aside from the sentiment?

Original Message: "{original}"
 Synthetic Message: "{flipped}"
 Transformation Type: {transformation}
 GPT-4 Explanation for the Flip: "{explanation}"

Return ONLY this JSON:

```
{
  "fluency": 0 or 1,
  "naturalness": 0 or 1,
  "sentiment_flip_clarity": 0 or 1,
  "meaning_preservation": 0 or 1,
  "annotator_comment": "optional comment (string)"
}
```

Table 8: Prompt used to evaluate quality of synthetic counterfactuals across four rubric dimensions.

(a) Counterfactual Generation Prompt
 You are an NLP assistant helping researchers generate high-quality counterfactual examples for sentiment classification.
 Given a WhatsApp-style message and its sentiment (Positive or Negative), generate 3 distinct versions that flip the sentiment. Only modify necessary components. Preserve fluency and realism. Respect informal tone.
 You may flip sentiment by changing components such as:
 - keywords, phrases, negation, intent framing, tone (e.g., sarcasm), sentiment valence, emojis/icons, code-mixing

Input:
 Original message: "{original_message}"
 Original sentiment: "{original_sentiment}"
 Output Format (JSON List of 3 Objects):

```
{
  "cf_text": "...",
  "components_changed": [...],
  "flip_explanation": "..."
}
```

(b) Counterfactual Filtering Prompt
 You are a sentiment evaluation assistant. Your task is to select the best counterfactual rewrite of a message.

ORIGINAL MESSAGE
 "{original}"
 (Sentiment: {original_sentiment})

COUNTERFACTUAL CANDIDATES

1. "{cf1}"
2. "{cf2}"
3. "{cf3}"

INSTRUCTIONS
 Your goal is to identify which counterfactual most effectively flips the sentiment while remaining realistic and fluent.

- Flip sentiment plausibly
- Sound natural in WhatsApp chat
- Preserve meaning/context where possible

RESPONSE FORMAT (JSON only):

```
{
  "selected_cf": "...",
  "justification": "...",
  "predicted_sentiment": "Positive / Negative"
}
```

Table 9: Combined prompts for generating and selecting counterfactual sentiment flips.

Explanation Evaluation Prompt You are a language model tasked with evaluating the quality of a sentiment explanation. Evaluate the explanation for the following:

1. Faithfulness - Does it reflect the original message and prediction without hallucinating?
2. Contextual Appropriateness - Is it culturally and linguistically aware?
3. Logical Coherence - Is it internally consistent and justified?
4. Clarity and Completeness - Is it clear, specific, and sufficient?

Message:
 "{message}"
 Predicted Sentiment: {prediction}
 Explanation: "{explanation}"

Return ONLY this JSON:

```
{
  "faithfulness": 0 or 1,
  "contextual_appropriateness": 0 or 1,
  "logical_coherence": 0 or 1,
  "clarity_and_completeness": 0 or 1,
  "annotator_comment": "optional comment (string)"
}
```

Table 10: Prompt used to evaluate explanation quality across four rubric dimensions.

A.6 Further Examples from Our WhatsApp Dataset: Cultural Nuance and Annotator Disagreement

Example	Explanation
<i>My friends it was heard to take drugs bt i just take heart</i>	<ul style="list-style-type: none"> • Can be read differently due to <i>situational context (sympathy)</i>. • Shows emotional vulnerability, which may invite empathy or humor depending on setting. • Use of “take heart” is culturally influenced—often heard in African English as a way to express resilience. • The spelling (“heard” instead of “hard”) could be interpreted differently (innocent typo vs. deeper linguistic variation).
<i>Kama hauko School shindaapo</i> “Even you are not in school just stay there”	<ul style="list-style-type: none"> • Can be read differently due to <i>schooling context</i>. • Often used sarcastically or dismissively, especially in online chat. • The phrase can also reflect class-based or knowledge-based exclusion (“If you’re not educated, stay out of this”). • Code-mixing adds a layer of urban youth culture and localized meaning.
<i>He is faithful all the time</i>	<ul style="list-style-type: none"> • Can be read differently due to <i>religion</i>. • Common in Christian communities, especially in African contexts—often part of a call-and-response. • Can express faith during suffering, giving it emotional depth in testimonies or public speeches. • Without context, it may be misread as a general statement about a person rather than a declaration about God.

Table 11: Examples of cultural nuance and their context-dependent interpretations.

A.7 Rubrics for Evaluation

A.7.1 Explanation Evaluation Rubric

Each model-generated explanation was evaluated along four binary (0/1) dimensions:

- **Faithfulness:** Does the explanation accurately reflect the input message and how it informed the model’s sentiment prediction? Explanations that include hallucinated, fabricated, or unrelated content should be scored **0**.
- **Contextual Appropriateness:** Does the explanation show awareness of cultural, social, or linguistic context? If it fails to address relevant tone, code-mixing, or local expressions, assign **0**. Optional comments may highlight cultural or linguistic mismatches.
- **Logical Coherence:** Is the explanation internally consistent and logically connected to the sentiment label? Contradictory or illogical justifications are scored **0**.
- **Clarity and Completeness:** Is the explanation clear, specific, and sufficient to support the sentiment label? Vague or underspecified rationales receive **0**.

Scoring: 1 = Yes; 0 = No

Note: Annotators were asked to leave optional comments when assigning a score of 0, especially for cultural/contextual errors or hallucinations.

A.7.2 Synthetic Data Evaluation Rubric

Each GPT-4-generated counterfactual message was evaluated using the following binary (0/1) criteria:

- **Fluency:** Is the synthetic message grammatically well-formed and fluent?

- **Naturalness:** Does the message sound plausible or likely to have been written by a real user?
- **Sentiment Flip Clarity:** Is the reversal in sentiment (compared to the original message) clear and consistent?
- **Meaning Preservation:** Aside from sentiment, does the core meaning/topic of the original message remain intact? Large semantic shifts receive **0**.

Scoring: 1 = Yes; 0 = No

Note: Annotators were encouraged to flag particularly good or bad examples, especially where tone, fluency, or cultural grounding were notably off.

A.8 Sentiment Transformation Taxonomy

To guide counterfactual generation, we organize sentiment-altering edits into the following transformation types:

Transformation Type	Definition	Example
Negation	Add or remove negation to reverse sentiment.	"I like it" → "I don't like it"
Tone / Intent Shift	Change the tone or implied intent of the message.	"You could do better" → "You're doing great"
Emoji Substitution	Replace emoji's to reflect different sentiment.	"😊" → "😞"
Keyword Substitution	Swap a sentiment-bearing word.	"Useful advice" → "Terrible advice"
Phrase Rewording	Paraphrase to shift sentiment while preserving meaning.	"You always help me" → "You always get in my way"

Table 12: Taxonomy of sentiment-altering transformations used in counterfactual generation.

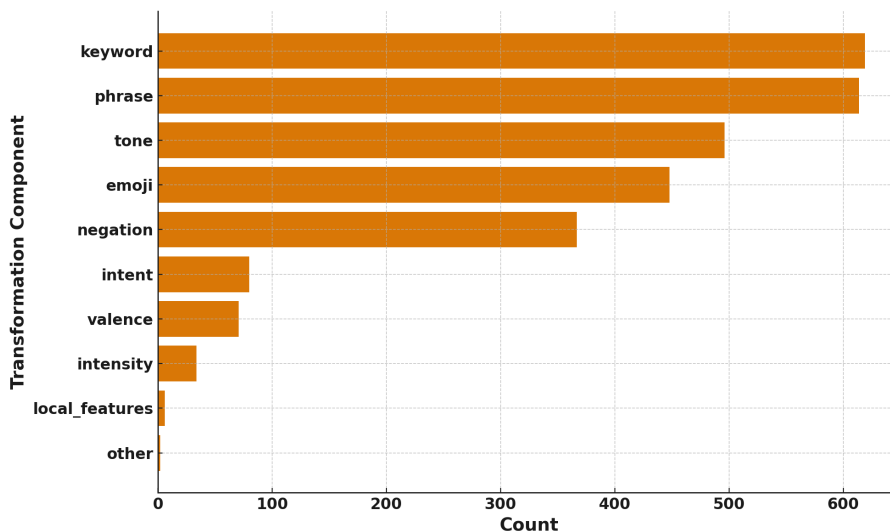


Figure 3: Distribution of transformation components used to generate counterfactual examples. Most sentiment flips involved keyword or phrase substitutions, followed by tone shifts, emoji manipulation, and negation. Less frequent transformations include intent, valence, and intensity adjustments.

A.9 Outputs of Synthetic Experiments

A.9.1 Component Consistency

Original message	Flipped message	LLM identified components	LLM rationale
Hahaha (+)	This is not funny at all. (-)	phrase, valence	Changing ‘😂😂 why?’ to a direct statement expressing displeasure flips the sentiment from positive to negative. Note: hallucinated explanation
Hahaha (+)	not funny at all (-)	keyword, negation	Changing ‘hahaha’ to ‘not funny at all’ adds negation and alters the key phrase to express dissatisfaction, flipping the sentiment to negative.
Ahahhaa (+)	Not funny at all. (-)	keyword, negation	Adding negation and changing the keyword from a positive laughter expression to a negative critique directly flips the sentiment.
Haha (+)	not funny (-)	phrase, negation	Introducing ‘not’ before ‘funny’ adds a negation to the original message, shifting the sentiment from amusement to disapproval.
Hahaaaaa 😊😊😊 (+)	Ughhh 😞😞😞 (-)	emoji, tone	Changing the emoji from 😊 (happy) to 😞 (annoyed) and the exaggerated sound from ‘Hahaaaaa’ to ‘Ughhh’ shifts the sentiment from positive to negative.
hahaha 😂😂😂 (+)	this is not funny at all 😡 (-)	phrase, emoji [no mentioning of negation]	Adding text that explicitly states discontent (‘this is not funny at all’) and using an angry emoji changes the sentiment from amusement to displeasure.

Table 13: Examples from our synthetic experiment showing varied LLM component sensitivity and consistency in handling laughter expressions.

A.9.2 What Worked Well

Original message	Flipped message	LLM identified components	LLM rationale
<i>Pia mi nko poa</i> (+) “Me too, I’m okay.”	<i>Pia mi siko poa kabisa leo</i> (-) “Me too, I’m really not okay today.”	phrase, intensifier	Adding ‘siko’ (not) and ‘kabisa’ (at all) to the original message intensifies the negative sentiment, flipping it from positive. [Reflection: Correct flip with Standard Swahili]
<i>Ata me wananiboo leo yote hawajatext</i> (-) Even me, they’re boring me the whole day they haven’t texted.	<i>Ata me leo nimefurahishwa, wamekuwa wakinitext mara kwa mara.</i> (-) “Even me, today I’ve been made happy, they’ve been texting me again and again.”	keywords, phrase	By modifying the phrase to express happiness and frequent communication, the sentiment changes from negative to positive. [Reflection: Correct flip with Standard Swahili]
<i>I will check on that coz nakohoa sana adi naumwa na kifua</i> (-) “I’ll check on that because I’m coughing a lot - to the point my chest hurts.”	I will check on that coz I’m feeling much better now, the cough and chest pain are gone! (+)	keywords, valence	The switch from describing ongoing symptoms to stating that the symptoms are gone (‘the cough and chest pain are gone’) changes the sentiment from negative to positive by indicating an improvement in health. [Reflection: Correct flip with English Translation]

Table 14: Examples from our synthetic experiment showing successful sentiment flips and clear rationale.

A.9.3 What Worked Less Well

	Example 1	Example 2	Example 3
Original message	<i>but kulinice</i> “But it was nice”	And thank you 5017 for very clear elaboration about the questions at hand. I love how you have handled and answered them to the point (+)	Dia hope you will be coming visiting I enjoyed your company alot your fun to be with (+)
Flipped message	but <i>kulinice</i> was disappointing	And thanks 5017 for very vague explanations about the questions at hand. I hate how you have mishandled and dodged them completely . (–)	Dia, please don’t come over. Honestly, hanging out was more exhausting than enjoyable, you’re not that fun to be with. (–)
LLM identified components	key phrase	keywords, key phrases, tone	negation, tone, key phrases
LLM rationale	Changing ‘kulinice’ to ‘kulinice was disappointing’ directly flips the sentiment by expressing a negative experience.	Changing ‘clear elaboration’ to ‘vague explanations’ and ‘love’ to ‘hate’ directly flips the positive sentiment to negative. Additionally, ‘handled and answered them to the point’ was changed to ‘mishandled and dodged them completely’ to emphasize dissatisfaction.	Introducing a request for absence (‘please don’t come over’) and directly stating negative feelings (‘more exhausting than enjoyable’, ‘not that fun’) shifts the sentiment from positive to negative.
Reflections / Insights	From the LLM’s explanation sw-en codemixed <i>kulinice</i> meaning “it was nice” misunderstood; term “kulinice” was ignored rather than changed because it could not handle the code-mix expression.	Flipped message feels somewhat unnatural: the “thanks” gets in the way of creating weird sarcasm/strong negativity with words like “hate”; hate may be the flip to “love” but it’s rare to have such expressions in feedback about questions.	The flipped message sounds strong/extreme; unnatural.

Table 15: Examples from our synthetic experiment showing less effective sentiment flips.

A.10 Model Explanation Quality, Confidence, and Coverage

Dataset	Model	Faithfulness	Contextual Approp.	Logical Coherence	Clarity & Completeness
Ambiguous Set	GPT-4-32k	1.000	0.950	1.000	0.950
	GPT-4-Turbo	1.000	1.000	1.000	1.000
	Gemma-3-27B	1.000	1.000	1.000	0.975
Gold Set	GPT-4-Turbo	0.983	0.983	1.000	1.000
	GPT-4-32k	1.000	0.983	1.000	0.980
	Gemma-3-27B	0.967	1.000	0.983	0.967
	Phi-4	0.917	0.933	1.000	0.950
	OpenChat-3.5	0.783	0.750	0.900	0.783
	LLaMA-3-8B	0.683	0.683	0.817	0.733
	Mistral-7B	0.617	0.650	0.850	0.650
Synthetic Set	GPT-4-32k	1.000	1.000	1.000	1.000
	GPT-4-Turbo	1.000	1.000	1.000	1.000
	Gemma-3-27B	0.906	1.000	1.000	0.875
	Phi-4	0.750	0.750	0.800	0.775
	LLaMA-3-8B	0.700	0.900	0.950	0.700
	Mistral-7B	0.688	0.844	0.906	0.688
OpenChat-3.5	0.594	0.812	0.844	0.688	

Table 16: Explanation quality scores by dataset and model across four dimensions.

Model	Gold Conf.	Gold Cov. %	Synth. Conf.	Synth. Cov. %	Eff. Conf.
GPT-4-Turbo	4.174	100.0	4.639	100.0	4.64
GPT-4-32k	4.283	100.0	4.440	100.0	4.44
Phi-4	4.464	100.0	4.711	99.5	4.69
Gemma-3-27B	4.265	100.0	4.698	47.6	2.24
OpenChat-3.5	4.204	99.9	4.249	47.4	2.01
Mistral-7B	4.237	98.9	4.132	47.6	1.97
LLaMA-3-8B	4.311	92.9	3.981	37.6	1.50

Table 17: Average model confidence and coverage across Gold and Synthetic Sets.

ÒWE-Voice: An Evaluation of Monolingual and Multilingual ASR Model Using Yoruba Proverb Speech Dataset

Daud Abolade

Masakhane

NITHUB, University of Lagos

NKANDA

aboladedawud@gmail.com

Abstract

Given the advancement of various Artificial Intelligence (AI) technologies in the 21st century, Automatic Speech Recognition (ASR) plays a vital role in human and machine interaction and serves as an interface for a wide range of applications. The development of these high-performing, robust and useful technologies continue to gain more attention on high-resource languages due to high availability of language data, market profitability dominance and access to funding and research initiatives compared to the marginalised low-resource languages. Despite efforts to develop ASR systems for African languages, there are still numerous challenges due to limited speech datasets, tonal complexity and dialectal variation. In this study, we curated a domain-specific speech dataset for one of the oral Yoruba literatures, proverbs, which are highly culturally inclined. We used the Yoruba recording app that was developed for Iroyin-speech project to record 6 hours of Yoruba proverb sentences. The NCAIR1/Yoruba-ASR model which was finetuned on Open AI Whisper Small and Massively Multilingual Speech, a multilingual speech model featuring low-resource languages including Yoruba language was evaluated with the recorded Yoruba proverbs. Evaluation was conducted based on Word Error Rate (WER) and Tone Error Rate (TER). Our result shows that current ASR systems that support

Yoruba does not capture cultural nuances. These findings highlight an urgent need to curate more robust speech datasets that are culturally embedded for low resource languages and in this case particularly, Yoruba language in order to build technological tools that preserve African culture, language and identity.

1 Introduction

Prior to the widespread adoption of literacy in Africa, knowledge systems were transmitted primarily through oral traditions. Among these traditions, proverbs constitute one of the most significant forms of cultural knowledge. Proverbs, alongside with folktales, praise poetry and the Ifa oral corpus is beyond mere creative expressions. It serves as a cognitive tool for teaching, solving problems, reasoning and preserving the language and cultural identity. Orality is African heritage, through societal observations proverbs are born. In Yoruba culture, proverbs hold a special place because they convey values, ideas, logic, expressions, and experiences with figurative language which make it difficult for non-native speakers and native speakers that are not well grounded in the language to decode its meaning (Olusanya et al., 2025). Their meaning often depends on logical reasoning which makes it challenging for both humans and machines.

With the rapid expansion of speech technologies across the globe, the absence of accurate and culturally inclined automatic speech recognition (ASR) systems for African languages presents a major barrier to digital preservation of indigenous knowledge thereby contributing to the digital divide. One of the

major solutions is to create culturally inclined datasets and incorporate them into models for more inclusivity. While there have been series of development in building general purpose speech synthesis and speech recognition system which has resulted to the birth of digital products such as Siri, Alexa, Google’s Gemini, and Microsoft Cortana for high-resource languages, African languages still struggle to get to this level due to limited amount of quality speech corpora. It must be mentioned that many African oral traditions are at risk due to rapid urbanization, linguistic shift, and diminishing transfer of indigenous knowledge across generations. Recent advances in multilingual speech models such as Whisper, MMS and SeamlessM4T show promising opportunities for many low resource languages (Radford et al., 2022; Pratap et al., 2023; Communication et al., 2023). However, their performance on culturally nuanced dataset remains unexplored and African oral literature remains challenging for ASR systems because it often deviates from everyday conversational patterns, contains uncommon lexical items and relies heavily on tonal accuracy to preserve semantic meaning. There have been efforts towards the curation of speech dataset to train, evaluate and fine-tune speech synthesis and speech recognition models for low-resource languages, (Junczyk, 2024; Emezue et al., 2025; Oliveira et al., 2023) yet, these datasets often neglect African oral literature content which is the bedrock of Africa’s rich culture.

To address this challenge, this study presents Òwe-voice, a 6 hours Yoruba proverb speech datasets as well as an experimentation of two state-of-the-art ASR models, evaluated on a culturally inclined speech dataset of 1,250 recorded Yoruba proverbs. We evaluate NCAIR1/Yoruba-ASR which is a Yoruba specific model and MMS-1b-all using WER and TER.¹

2 Yoruba language

The Yoruba language is spoken in 10 states in southwestern Nigeria and in some communi-

ties in the republics of Benin and Togo with over 40 million native speakers. A language that belongs to the Niger-Congo family, it has about 20 dialects and it is one of the national languages of Nigeria also spoken in other countries like Ghana, Côte d’Ivoire, Sierra Leone, Cuba and Brazil (Owolabi, 2006) which makes it one of the prominent and most widely spoken African languages in the world. The language has 25 letters of the Latin alphabet including additional letters containing subdots, such as (e., gb, ş, and o.). Yoruba is a tonal language, meaning that it has three distinctive tone levels-high, mid, and low that are decisive in word distinction. High and Low tones are marked with acute (´) and grave (`) diacritics respectively, while Mid tone is typically unmarked in standard orthography. Accurate pronunciation depends greatly on the tonal marks and subdots. Yoruba is a culturally rich language which has its own special way of preserving and passing oral knowledge before colonization.

3 Related works

Several research has highlighted the need for a large amount of both textual and speech dataset to build ASR models. Earlier work has focused on creating general purpose speech dataset across the three major Nigerian languages (Igbo, Hausa and Yoruba). (Ogunremi et al., 2024) Created about 42 hours of speech data recorded by 80 volunteers, and 6 hours of validated recordings of news and creative writing domains. (Meyer et al., 2022) A religious domain dataset of 86 hours open speech dataset for ten languages spoken in Sub-Saharan Africa where they trained the VITS end-to-end speech synthesis model. (van Niek-erk et al., 2015) this dataset was claimed to be curated for speech recognition research, about 33 diverse speaks both male and female gender. The current largest speech corpus for Hausa, Igbo and Yoruba is (Emezue et al., 2025), where a dataset creation process known as data farming was implemented to curate 1839 hours of speech recording on several domains. (Ahia et al., 2024) introduced a parallel text and speech corpus of standard Yoruba and its dialects to perform a machine translation,

¹<https://github.com/Holuwasege/OWE-Voice-Evaluation>

automatic speech recognition, and speech-to-text translation task.

While all these studies have greatly explored curation of speech dataset with different modalities, substantial amounts of speech data to train ASR models for Nigerian languages and other low resource African languages are still limited.

4 ÒWE-Voice Dataset

4.1 Textual Data Preparation

The proverb text used in the development of the Òwe-Voice, Yoruba Proverb Speech Dataset was obtained from two different sources. Our primary textual source was Òwe-Yor, a Yoruba text classification dataset containing both proverbial and non-proverbial sentences (Olusanya et al., 2025). We filtered the dataset to isolate only texts containing proverbs. To extend the coverage of the dataset beyond corpus-based collections, we conducted community fieldwork in Lagos and Ogun State. We engaged both the elders and young native speakers who exhibit strong cultural knowledge and oral tradition competence. Through the short interview session that was conducted, we collected additional 70 proverbs. The fieldwork was adopted to capture proverbs that do not appear in existing proverbial datasets. Manual cleaning and orthographical verification, following the standard Yoruba orthography with the accurate tone marks were conducted on the collected data.

4.2 Speech Data Recording

To create high-quality audio datasets of ÒWE-Voice, we followed the process used for creating IroyinSpeech. Yoruba native speakers who are linguistics students at University of Lagos volunteered and they were tested to confirm and verify their competence in reading Yoruba language. A total of 12 voice talents were engaged, representing both male female speakers to ensure acoustics diversity. The recording was conducted in a controlled environment using an Audio-Technical AT2020USB-X microphone and a quiet studio setup to minimize background noise. The Yorùbá Voice SpeechRecorder that was developed for Iroyin

Source	Number of Proverbs
Òwe-Yor Corpus	4,930
Fieldwork Collection	70
Total	5000

Table 1: Dataset curated via Òwe Corpus and Fieldwork Collection.



Figure 1: An illustration showing the environment setup and how Òwe-Voice was curated.

speech (Orife et al., 2022) The app processes a text file, typically containing 250 sentences, by sequentially displaying each line for voice input. Functionality includes: recording, playback, file management and deletion, in the case of multi-take recordings. To standardize the recording protocol we adopted some modality which are:

- 1) Each speaker received a batch of 250 proverbs.
- 2) Sentences were displayed with the correct tone mark and its standard orthography.
- 3) All audio was recorded at 48 kHz, 16-bit WAV format.
- 4) The voice talent was guided by a prompter who was both a linguist and technically sound in operating the Yorùbá Voice SpeechRecorder.

The final dataset contain 6hrs Yoruba proverb audio samples, each paired with its transcription.

5 Experiment and Result

In this study, we conducted a zero-shot evaluation of the following existing ASR models on Òwe-Voice dataset. We sampled 25% of the utterances across all speakers for evaluation, ensuring that each speaker was represented in the evaluation split. The remaining 75% of the data was reserved for potential fine-tuning

Models	WER	TER
NCAIR1/Yoruba-ASR	72.45	27.83
MMS-1b-all	95.42	66.75

Table 2: Result of the models evaluated on 25% Òwe-Voice dataset

experiments.

NCAIR1/Yoruba-ASR-v1.0 (Awarri Technologies & National Centre for Artificial Intelligence and Robotics (NCAIR), 2025): This is a monolingual automatic speech recognition (ASR) model finetuned on the Whisper Small architecture, specifically for Yoruba language which is expected to capture linguistics patterns and orthographical conventions.

MMS (Pratap et al., 2023): This is Meta’s open-source 1B parameter wav2-vec2 architecture (Baevski et al., 2020) model, supporting 1162 languages, including Yoruba language.

Comparing NCAIR1/Yoruba-ASR with MMS-1b-all allows us to investigate the current state of this model’s output on Yoruba proverb speech dataset as part of the indigenous oral knowledge of the Yoruba people.

5.1 Word Error Rate

Word Error Rate (WER) was used as the primary metrics to quantify transcription accuracy at the lexical level. The NCAIR1/Yoruba-ASR model achieved a 72.45% WER which performs better than the multilingual speech model. Although the high error rate shown in this result indicates that the model still struggles with proverbial expressions even though the model was fine-tuned specifically on Yoruba language. On the other hand, MMS produced a much higher WER of 95.42%. This result confirms that multilingual speech models, despite its large training coverage, lack sufficient representation of Yoruba orthography. The model misrecognized common Yoruba lexical items and produced high deletion and substitutions errors. Overall, the WER analysis shows that both models struggle significantly with transcribing Yoruba proverb speech, with the Yoruba specific model offering only partial improvements while the multilingual model largely fails to generalize. These results highlight the difficulty of ASR for low-resource

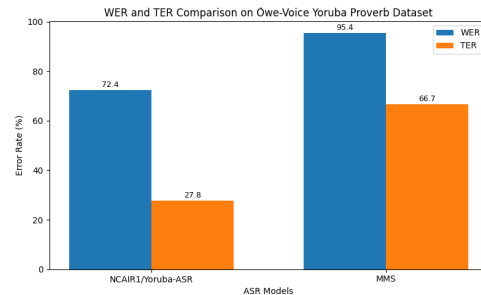


Figure 2: Visualization of the WER and TER score

tonal languages.

5.2 Tone Error Rate

Going beyond lexical investigation, we evaluated tonal accuracy using the TER, a metrics that compares predicted tone sequences to reference tone sequences extracted from the predicted transcript of the model and the reference transcript curated alongside with the speech dataset. TER provides insight into the model’s ability to preserve Yoruba prosodic distinctions which is a crucial phenomenon because tonal differences often signal different meanings in proverbial expressions. The NCAIR1/Yoruba-ASR model revealed a lower TER score than MMS, showing that the model captures Yoruba tonal patterns better than the MMS model. However, its overall tone accuracy remained challenging. Many tone errors occurred even when the segmental transcription was correct, indicating that the model can predict the right word but fail to assign the correct tone. The MMS model displayed a significantly higher TER, demonstrating very poor tonal generalization. This is expected because multilingual training typically does not emphasize tonal information, and MMS does not explicitly model Yoruba tones. In conclusion, the TER results reveal that tonal errors remain a major bottleneck for Yoruba language in ASR model performance.

6 Beyond Experiments

Yoruba proverb dataset by participating in a Speech Hackathon where we developed a Yoruba Proverb Text-to-Speech web application aimed at evaluating the performance of the Spitch TTS model, a commercial speech technology model built by a language technol-

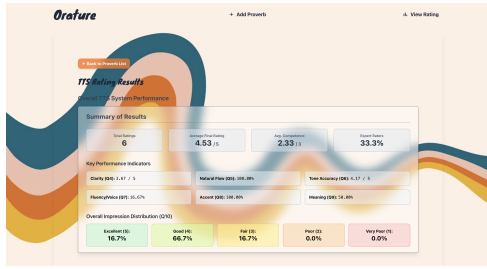


Figure 3

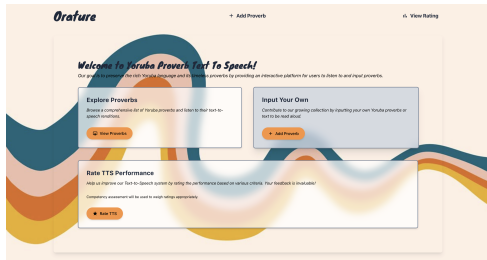


Figure 4: Landing page of Yoruba Proverb Text-to-Speech web application

ogy company in Nigeria. The web app was designed to output spoken Yoruba proverbs directly from the text and to provide an interface through which users could interact with the model.

The primary motivation behind this hackathon was twofold. First, it offered a practical opportunity to assess how well an industry grade TTS model can handle the prosodic complexities of Yoruba language, especially the proverbial expressions, since we are having language tech start ups coming up in Nigeria. Second, it aligned with our mission of preserving Yoruba oral knowledge through technology.²

7 Conclusion and Future work

This work introduced Òwe-Voice, a Yoruba proverb speech dataset and the dataset was used to evaluate the performance of both monolingual and multilingual ASR models on culturally rich, low-resource linguistic material. Yoruba proverbs are structurally complex which makes it challenging for ASR models trained on general speech. Due to lack of enough resources and access to GPU we couldn't fine-tune the models. Our evaluation demonstrated that the monolingual

²<https://orature.vercel.app/>

NCAIR1/Yoruba-ASR model outperformed the MMS model, having a lower Word Error Rate (WER) and Tone Error Rate (TER) score. Building on this work, We plan to expand the proverb dataset to other African languages and also curate large hours of speech datasets to train existing contemporary ASR systems, including large language model-based speech systems such as GPT-4o/GPT-5-style multi-modal models, Qwen3-ASR, Whisper variants, and commercial ASR APIs. We also aim to build other datasets that target African oral knowledge such as praise poetry, folklores and restructure that Yoruba proverb web app. Òwe-Voice can also be explored in other wide range of tasks such as speech translation.

Limitations

There are several limitations in this study. Firstly, due to constraints in computational resource, we were unable to fine-tune the evaluated models before evaluation. Secondly, the scope of the dataset was limited due to the challenges faced in gathering the dataset. Although there was a focus on finding culturally grounded Yoruba proverbs, time, logistics, and budget limitations made it impossible to gather a large amount of speech data. This constraints also prevented the expansion of the dataset to other African languages that have a rich oral literature.

Thus, even though the Òwe-Voice dataset is an important resource that allows one to critically assess ASR models on Yoruba cultural nuances, the corpus is yet to adequately document the African oral knowledge system at large.

References

Orevaoghene Ahia, Anuoluwapo Aremu, Diana Abagyan, Hila Gonen, David Ifeoluwa Adelani, Daud Abolade, Noah A. Smith, and Yulia Tsvetkov. 2024. *Voices unheard: Nlp resources and models for yorùbá regional dialects*. *Preprint*, arXiv:2406.19564.

Awari Technologies & National Centre for Artificial Intelligence and Robotics (NCAIR). 2025. *Yoruba-asr v1.0: Automatic speech*

- recognition for yoruba language. Hugging Face Model Hub. Accessed: 2025-12-09.
- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. [wav2vec 2.0: A framework for self-supervised learning of speech representations](#). *Preprint*, arXiv:2006.11477.
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenhaler, Paul-Ambroise Duquenne, Brian Ellis, Hady El-sahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Iliia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023. [Seamless: Multilingual expressive and streaming speech translation](#). *Preprint*, arXiv:2312.05187.
- Chris Emezue, NaijaVoices Community, Busayo Awobade, Abraham Owodunni, Handel Emezue, Gloria Monica To-bechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, and Chris Pal. 2025. [The najjavavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages](#). *Preprint*, arXiv:2505.20564.
- Michał Junczyk. 2024. [Framework for curating speech datasets and evaluating asr systems: A case study for polish](#). *Preprint*, arXiv:2408.00005.
- Josh Meyer, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky, Iro-ro Orife, Colin Leong, Perez Ogayo, Chris Emezue, Jonathan Mukiibi, Salomey Osei, Apelete Agbolo, Victor Akinode, Bernard Opoku, Samuel Olanrewaju, Jesujoba Alabi, and Shamsuddeen Muhammad. 2022. [Biblets: a large, high-fidelity, multilingual, and uniquely african speech corpus](#). *Preprint*, arXiv:2207.03546.
- Tolulope Ogunremi, Kola Tubosun, Anuoluwapo Aremu, Iro-ro Orife, and David Ifeoluwa Adelani. 2024. [Ìròyìnspeech: A multi-purpose yorùbá speech corpus](#). *Preprint*, arXiv:2307.16071.
- Frederico S. Oliveira, Edresson Casanova, Arnaldo Cândido Júnior, Anderson S. Soares, and Arlindo R. Galvão Filho. 2023. [Cmltts a multilingual dataset for speech synthesis in low-resource languages](#). *Preprint*, arXiv:2306.10097.
- Joy Olusanya, NLP Masakhane, and Daud Abolade. 2025. [Owe-yor: Leveraging transformer based models for yoruba proverb classification](#). In *Proceedings of the Conference*.
- Iro-ro Orife, Aremu Anuoluwapo, Kólá Túbòsún, David Ifeoluwa Adelani, and Tolúlopé Ógúnrẹmí. 2022. [Yorùbá voice speech recorder](#).
- K. Owolabi. 2006. [Yoruba](#). In Keith Brown, editor, *Encyclopedia of Language & Linguistics (Second Edition)*, second edition edition, pages 735–738. Elsevier, Oxford.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. [Scaling speech technology to 1,000+ languages](#). *Preprint*, arXiv:2305.13516.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *Preprint*, arXiv:2212.04356.
- Daniel van Niekerk, Etienne Barnard, Oluwapelumi Giwa, and Azeez Sosimi. 2015. [Lagos-NWU yoruba speech corpus](#). SADIaR Language Resource Repository, License: Creative Commons Attribution 2.5 South Africa License: <http://creativecommons.org/licenses/by/2.5/za/legalcode>.

Language Choice in Nigerian Social Media Hate Speech

Nneoma C. Udeze

Northwestern University
nneomaudeze2027@u.northwestern.edu

Rob Voigt

University of California, Davis
robvoigt@ucdavis.edu

Abstract

Language choice in multilingual societies is rarely arbitrary. In Nigerian, English, Nigerian Pidgin (NP) and indigenous languages are strategically deployed in online discourse, yet little is known about how they function in hostile contexts. Here we conduct the first systematic analysis of NP in online hate speech on two platforms, Twitter and Instagram. Using a linguistically enriched annotation scheme, we label each post for class, targeted group, language variety, and hate type. Our results show that NP is disproportionately used in offensive and hateful discourse, particularly against Hausa, women, and LGBTQ+ groups, and that insults are the dominant hate strategy. Cross-domain evaluation further reveals that classifiers trained on Twitter systematically overpredict hate on Instagram, highlighting challenges of domain transfer. These findings underscore NP’s role as a linguistic resource for hostility and its sociolinguistic salience in amplifying stereotypes and affect. For NLP, the work demonstrates the need for NP-specific resources, sensitivity to figurative strategies, and domain adaptation across platforms. By bridging sociolinguistics and computational modeling, this study contributes new evidence on how language choice shapes online hate speech in a multilingual African context.

1 Introduction

The choice of language in multilingual contexts such as Nigeria is rarely arbitrary. It reflects social alignments, ideological positions, and strategic rhetorical intent. Among the languages spoken in Nigeria, Nigerian Pidgin (NP), an English-based pidgin/creole, is the most widely spoken language despite lacking an official political government status (Faraclas, 2021; Soneye, 2019). Although NP has long been studied for its roles in informality, solidarity, and identity expression, especially in digital and youth culture (Osoba, 2014; Agantiem

and Alagbe, 2023; Usoro and Nsit, 2024; Nweke et al., 2024; Adegbiya, 2004; Oluyinka Adebayo, 2023), its role in conflict-driven discourse remains underexplored. It is deeply embedded in everyday life and plays a role as a unifying force in Nigeria’s multicultural society, functioning as both a street and home language, valued for humor and storytelling. Hate speech adds a layer of complexity to its sociolinguistic analysis. Existing research often focuses on hateful content or political context, but pays limited attention to the specific language varieties through which hate is articulated. In Nigeria’s highly multilingual digital landscape, users appear to alternate between English, NP, and indigenous languages in ways that can be interpreted as framing hostility, intensifying emotion, or encoding in-group messages. Previous datasets (Muhammad et al., 2025) indicate that multilingual hate expression is common and moderation of large-scale and targeted hate speech remains limited due to scarcity of high-quality data in local languages and the exclusion of local communities from data collection, annotation, and moderation efforts.

This study therefore examines the highly charged context of online hate speech to ask whether language choice matters, particularly the use of Nigerian Pidgin. This question is significant because language choice in a multilingual society like Nigeria is never neutral; it reflects social identities, power relations, and ideological alignments. Nigerian Pidgin, in particular, functions as both a unifying lingua franca and a marker of authenticity, solidarity, and informality. Its use in online hate speech therefore offers a lens into how speakers draw on linguistic resources to frame hostility, express stance, and index in-group belonging.

Although computational models have begun to incorporate NP in hate detection systems (Adegoke et al., 2024), these approaches may overlook the discursive and socio-ideological functions of NP in expressing toxicity. As a result, the role of NP

in framing hostility, reinforcing stereotypes, and drawing in-group boundaries remains poorly understood. To address this gap, we present these research questions:

1. Is Nigerian Pidgin used more frequently in offensive and hate content than in neutral content within the multilingual Nigerian social media context?
2. Are there specific social groups that are more frequently targeted with Nigerian Pidgin in online discourse?
3. What are the lexical and multi-word patterns associated with various hate types (insults, dehumanization, demonization, incitement to violence) and which groups are most frequently targeted with specific hate types?

To achieve this, we examine two datasets of online discourse: a re-annotated sample from the NaijaHate Twitter Corpus (Tonneau et al., 2024) and a novel dataset of Instagram comments scrapped from Instablog9ja (from July-Sept. 2024). We combine computational classification with manual linguistic annotation to identify languages used, isolate languages expressing toxicity, and categorize hate speech types. Our contributions are as follows:

1. A cross-platform analysis of Nigerian Pidgin in online hate speech: We provide the first systematic investigation of NP across Twitter and Instagram, examining how language choice varies by class (neutral, offensive, hateful) and targeted group.
2. A linguistically enriched annotation framework for hate speech: We develop and apply an annotation scheme that captures not only class labels and targeted groups, but also hate types (insults, dehumanization, demonization, and incitement to violence) and language choice, enabling both quantitative modeling and qualitative analysis.
3. Empirical and computational insights into multilingual hate detection: Through cross-domain error analysis and regression modeling, we show how NP disproportionately encodes hostility, and we highlight the implications for building culturally informed hate-speech detection systems.

Systematically analyzing the role of Nigerian Pidgin in constructing hostility across platforms is critical because hate speech classifiers trained on a single platform or on majority-language data (e.g. English) may fail when applied to other platforms or to local language varieties. This paper addresses this gap by providing a linguistically enriched, cross-platform analysis of NP in hate speech, combining computational modeling with fine-grained annotation of language choice, hate type, and targeted groups. By doing so, we contribute new empirical evidence to sociolinguistics and NLP, demonstrating how language choice functions as a strategic mechanism in online hostility.

2 Background of Study

Research on language choice in multilingual societies shows that speakers switch among codes to express identity, solidarity, authority, or informality (Ifukor, 2011; Oduma and Gomwalk, 1986; Igboanusi, 2008a). Language switching in digital forums and commercial discourse often serve rhetorical goals, reflecting the strategic use of the language that best resonates with target audiences (Ifukor, 2011; Dalamu, 2017; Doğruöz et al., 2021). English often indexes authority, while Nigerian Pidgin indexes informality and solidarity, particularly in informal interactions (Oluyinka Adebayo, 2023; Balogun, 2013; Taiwo, 2010). It is often preferred for its neutrality and broad intelligibility in diverse multilingual public settings such as markets, prisons, or interethnic gatherings. On social media, NP appears frequently for strong opinions, identity marking, and rhetorical effects (Taiwo, 2010; Oluyinka Adebayo, 2023).

Despite this rich body of research, little is known about how NP functions in hostile discourse. Most sociolinguistic studies focus on uses of NP in positive or neutral contexts (Akande and Salami, 2010; Osoba, 2014; Balogun, 2013), while computational research has only recently begun to include NP in hate speech detection (Adegoke et al., 2024). Hate speech introduces an added layer of complexity to the study of language choice. While much of the existing literature concentrates on the semantic content or ideological thrust of hateful messages (AYENI, 2018, 2024; Adepoju and Kalu, 2022; Ononye and Nwachukwu, 2019), less attention has been paid to the linguistic strategies (like language choice and hate types) through which such content is framed and delivered.

Hate speech detection and characterization has been investigated extensively in high-resource contexts (Zannettou et al., 2020; ElSherief et al., 2018; Warner and Hirschberg, 2012) and, increasingly, in multilingual and low-resource contexts with specialized datasets and models (Vargas et al., 2022; Geleta et al., 2023; Muhammad et al., 2025; Ayele et al., 2023). For this study, we adopt the definition of hate speech proposed by the United Nations and used by Tonneau et al. (2024), as “any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor” (United Nations, 2019). Hate speech as noted by Waseem et al. (2017) can manifest in both explicit and implicit forms (Talat et al., 2017). Explicit hate speech typically includes clearly offensive language, slurs, or threats, making it more easily recognizable through automated detection methods (Talat et al., 2017). Implicit hate speech, on the other hand, is often expressed through sarcasm, irony, or coded language, and requires greater contextual awareness to identify, posing significant challenges for both human annotators and machine learning models. Papcunová et al. (2023) further contribute to the operationalizations of hate speech by proposing a structured set of indicators (Papcunová et al., 2023). These include expressions that promote violence, deny human rights, invoke negative stereotypes, employ ad hominem attacks, or manipulate historical facts. Such indicators offer a measurable framework for annotation and detection.

Within Nigerian-focused work, several datasets and computational systems have emerged (Ilevbare et al., 2024; Adegoke et al., 2024; Aliyu et al., 2022; Tonneau et al., 2024; Asogwa et al., 2022; Nkemdilim and Somtochukwu, 2024), but most prioritize model evaluation and performance rather than linguistic profiling of hate expressions. Recent studies show that abusive content is often more prevalent in Yoruba, Nigerian Pidgin, and code-switched messages than in Standard English in political contexts (Ilevbare et al., 2024). However, these studies may often neglect the nuanced interplay between linguistic form, hate type, and language variety. Therefore, our work bridges sociolinguistics and NLP by directly testing the prevalence of NP in hate contexts, profiling hate expressions at the lexical and phrasal level, and mapping

their group-specific targets, offering the contributions mentioned above.

3 Data Collection and Annotation

3.1 Datasets

This study is based on two social media datasets: Twitter and Instagram. The primary dataset is the NaijaHate Twitter Corpus (Tonneau et al., 2024), a curated collection of approximately 36,000 tweets collected between July 2021 and July 2023, annotated for tweet class (neutral, offensive, hateful) and target groups. From the NaijaHate Twitter corpus, we drew a stratified sample of 6,000 tweets (2,000 per class) to support a balanced comparative analysis. To compare cross-domain behavior, we scraped Instablog9ja, a high-traffic Nigerian Instagram account, using Instaloader (Instaloader Developers, 2024) and collected 35,000 public comments between July and September 2024. The NaijaXLM-T classifier, an existing XLM-R pretrained model finetuned on the NaijaHate Twitter corpus, initially classified the scraped Instagram comments.¹ From the classification set, we selected a balanced sample of 1,500 comments (500 per class) for manual re-annotation and analysis.² We used the NaijaHate-trained classifier to propose language labels and an initial class label for Instagram. All further analysis reported here rely on manual re-annotation to ensure consistency of labels across platforms.

3.2 Annotation scheme

To capture linguistic and ideological nuances, we used an expanded annotation schema based on Bahador (2020) (Bahador, 2020) (figure 1). Rather than a binary distinction, we annotate class (neutral, offensive, hateful), language variety (English, Nigerian Pidgin, Yoruba, Igbo, Hausa), target groups (Women, Igbo, Hausa, Yoruba, Christians, Muslims, LGBTQ+, Northerners, Southerners, Fulani, Herdsmen, Other), and hate type. Hate types include:

- **Insults:** Negative group characterizations (e.g. 'stupid', 'lazy').
- **Dehumanization:** Equating a group to subhuman entities (e.g. 'rats', 'pigs').

¹The NaijaHate dataset and the NaijaXLM-T classifier can be accessed here <https://github.com/worldbank/NaijaHate>

²The annotated Instagram dataset can be accessed here <https://github.com/Nneoma-Udeze/Language-Choice-In-Nigerian-Social-Media-Hate-Speech>

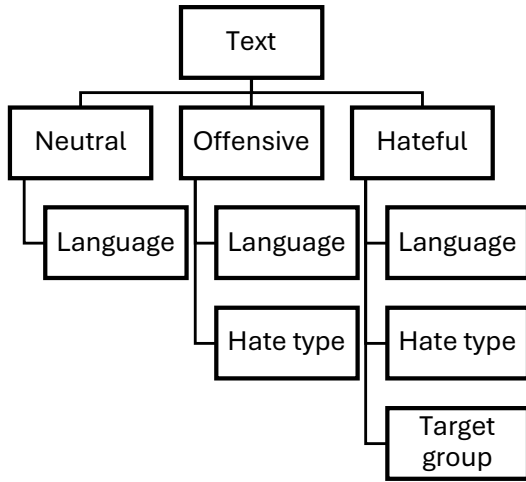


Figure 1: Hierarchical Annotation Framework for Classes, Languages, hate Types, and Target Groups

- **Demonization:** Portraying group as an existential threat (e.g. 'monsters', 'diseases').
- **Incitement to violence:** Calls for harm against a group.

Both target group and hate type annotations are multi-label; a single comment can target many groups and use multiple hate strategies.

3.3 Annotators

The annotations were conducted by the lead author and a graduate student in linguistics. The NaijaHate sample was re-annotated for speech class, targeted group, language(s) used, and hate type. A 20% overlapping sample (1,200 tweets) was used to compute inter-annotator agreement for the speech class alone. The agreement was high (Cohen's Kappa = 0.88), and most disagreements occurred between offensive and hateful classes, reflecting their inherent ambiguity.

4 NaijaHate Cross-Domain Performance

We evaluated the NaijaHate Twitter-trained classifier on the manually re-annotated Instagram dataset to test domain adaptability. This revealed an imbalanced class distribution in the Instagram sample, underscoring the discrepancy between model predictions and human annotations. The model achieved an overall accuracy of 63.3% when evaluated against human annotations (N = 1,396). Figure 2 shows a confusion matrix that compares model predictions and human labels.

We observe a clear bias toward hate classification, and the model exhibits exceptionally high

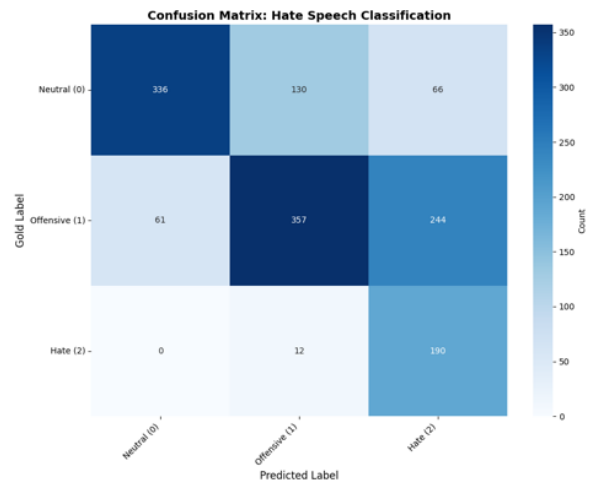


Figure 2: Confusion matrix evaluating NaijaHate model performance against human annotations

recall for the hate class (94.1%) at the cost of very low precision (38.0%). This bias is problematic for automated moderation systems as it risks inflating rates of false positives, potentially leading to unjustified censorship of neutral or offensive content. In contrast, the model shows more conservative behavior for neutral content, achieving high precision (84.6%) but moderate recall (63.2%), indicating that it correctly identifies neutral content when predicted, but misses a substantial portion of neutral instances.

We initially hypothesized that the poor performance of the NaijaHate model on Instagram data could be attributed to the prevalence of Nigerian Pidgin, a language not seen in large proportions on the Twitter-based training corpus but representing a substantially larger proportion of the Instagram discourse.

However, contrary to this expectation, the classifications in the Nigerian Pidgin comments demonstrated a better classification performance across all metrics. Comments containing only NP achieved the highest hate class F1 score (0.67), substantially outperforming English-only content (F1 = 0.57) despite English being the primary training language. The English-NP code-mix yielded an F1 score of 0.61.

Our findings point to fundamental differences between platforms in the manifestation of hate speech rather than language-specific classification challenges. The poor performance of the model on English content suggests that Twitter English and Instagram English could represent different communicative domains with different rhetorical

strategies for expressing toxicity. The superior performance on the NP content implies that Nigerian Pidgin employs more direct rhetorical strategies that align better with the model’s learned patterns, while the English discourse on Instagram may utilize more sophisticated linguistic strategies that diverge from the Twitter training patterns. This reveals the need for domain-sensitive calibration or strategies to prevent harmful overgeneralization.

5 Analysis and Discussion of Language Choice in Hate Speech

Analysis of language use within Instagram and Twitter datasets revealed distinct distributions across the three classes (neutral, offensive, and hateful). English and Nigerian Pidgin dominate both neutral and harmful content; indigenous languages (Yoruba, Igbo and Hausa) occur less frequently. Figure 3 & 4 shows the distribution of language by comment and tweet class in the Instagram and twitter datasets.

Among indigenous languages, Yoruba appears the most frequently, and many affective terms in NP are borrowed from Yoruba. When such loanwords occur embedded within a Pidgin syntactic frame, they are labeled as Nigerian Pidgin; when they appear in isolation, they are annotated as Yoruba, illustrating fluid boundaries and annotation challenges. This overlap introduces annotation ambiguity that has direct implications for downstream modeling. In supervised classification, such hybrid usage effectively introduces label noise, as lexical items associated with Yoruba may appear in instances labeled as Nigerian Pidgin and vice versa. As a result, models relying on surface-level lexical features or language identification assumptions may conflate language membership with affective intensity, potentially inflating or obscuring language-specific effects. These challenges suggest that computational models applied to Nigerian social media should move beyond strictly discrete language categories and instead accommodate mixed-language representations, for example, through subword modeling, contextual embeddings, or multi-label language tagging. Explicitly acknowledging and modeling language overlap is therefore crucial for both predictive performance and the interpretability of sociolinguistic findings.

Given the data, we structure our analysis around three central research questions that aim to address key gaps in existing studies on hate speech in Nige-

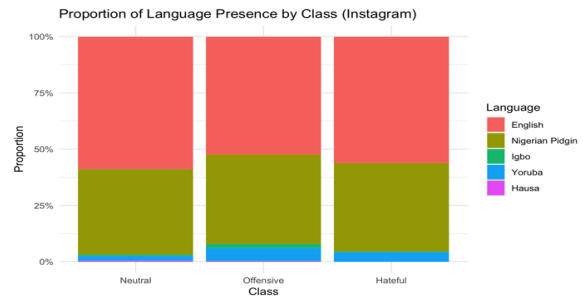


Figure 3: Language Distribution by Comment Class in Instagram Data

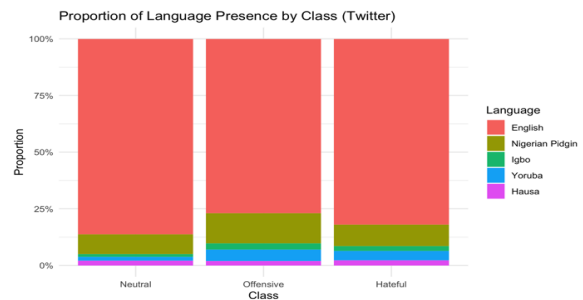


Figure 4: Language Distribution by Comment Class in Twitter Data

ria. The findings of these analyses offer a nuanced understanding of how NP functions within online hate speech and contributes new insights into the sociolinguistic dynamics of digital discourse in Nigeria.

5.1 RQ1: Is Nigerian Pidgin more frequent in offensive and hateful comments than in neutral content?

In figures 5 & 6, we notice that Nigerian Pidgin (both used alone and mixed) is used more in the Instagram data than in the Twitter data. Two regression models were fit on 1) the full dataset to determine the relationship of NP with toxic comments and tweets, and 2) a subset of the full dataset to text if an effect was mostly driven by single language use only (NP) or code-mixed language (NP mixed with other languages)

To answer RQ1, a regression model was fitted using a generalized linear model (GLM) framework in R to predict the probability that a comment is written in Nigerian Pidgin as a function of class, controlling for platform and code-mixing status. Two models were estimated: 1) on the full dataset to assess language-class relationships; 2) on a subset examining single-language NP vs NP in code-mixed contexts.

Statistical testing revealed a significant effect

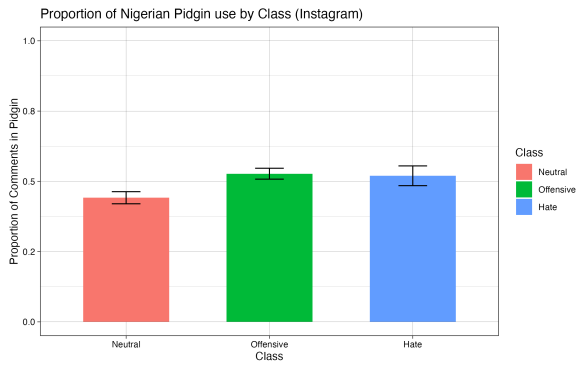


Figure 5: Distribution of Nigerian Pidgin Use across Neutral, Offensive, and Hateful Comments in Instagram Data

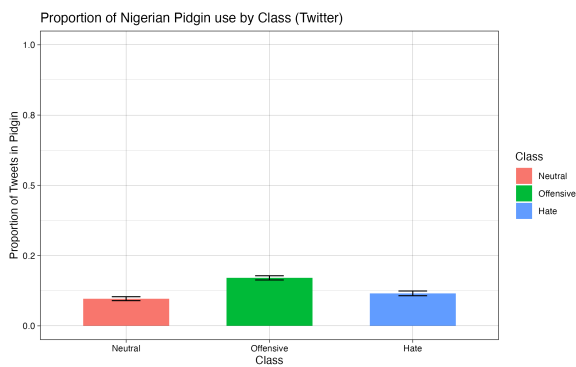


Figure 6: Distribution of Nigerian Pidgin Use across Neutral, Offensive, and Hateful Comments in Twitter Data

for the offensive class compared to neutral ($p < 0.005$) in both Instagram and Twitter, indicating offensive content is more likely to be written in NP than neutral content. For the hateful class, effects were marginal on Instagram ($b = -0.234$, $p = 0.059$) and non-significant on Twitter ($b = -2.234$, $p = 0.083$), suggesting trends but limited power given fewer hateful instances and potential moderation effects prior to data collection. These patterns should be interpreted in light of pronounced class imbalance, as hateful content constituted a substantially smaller proportion of the dataset relative to offensive and neutral comments. This imbalance reduces statistical power by inflating standard errors for class-specific estimates, making it more difficult to detect statistically significant effects even when underlying trends are present. As a result, the coefficients associated with the hateful class are inherently less stable and more sensitive to sampling variability. For content expressed solely in NP versus code-mixed with English or other languages, neither offensive nor hateful classes showed sig-

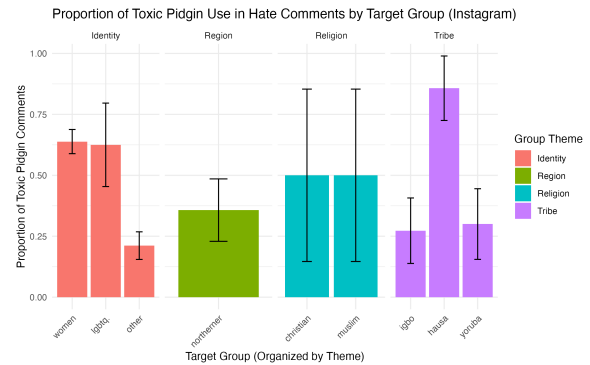


Figure 7: Distribution of Nigerian Pidgin in Hate Speech by Target Group in Instagram Data

nificant effects driving the association of NP in these classes for both Instagram and Twitter ($p > 0.05$). This suggests that the observed association between NP and harmful speech may not be driven by one subset of NP usage alone. Instead, the effect appears to emerge from the combined influence of both monolingual NP comments and code-mixed comments.

From a sociolinguistic perspective, this is consistent with previous work showing that NP functions as an informal and affective code used to express humor, critique, and solidarity (Oluwinka Adebayo, 2023). However, in hostile contexts, these same pragmatic features make NP a natural vehicle for insults and ridicule. The fact that NP is more predictive of offensive than neutral discourse highlights its strategic use in emotionally charged interactions. The trend observed in the hateful class warrants further investigation with a more balanced dataset, as it may reveal whether NP also systematically functions as a language of hate speech, beyond offensive expression.

5.2 RQ2: Which groups are disproportionately targeted with Nigerian Pidgin?

Figure 7 & 8 shows the proportion of hateful comments and tweets (which is the percentage of comments and tweets in NP compared to other languages seen in the dataset) expressed in Nigerian Pidgin across targeted groups, organized by thematic categories (identity, region, religion, and tribe).

NP is used disproportionately when targeting certain groups, especially Hausa, women, and LGBTQ+ people on Instagram. By contrast, non-Nigerians ('other') received the least hate expressed in NP. This suggests that NP is strategically

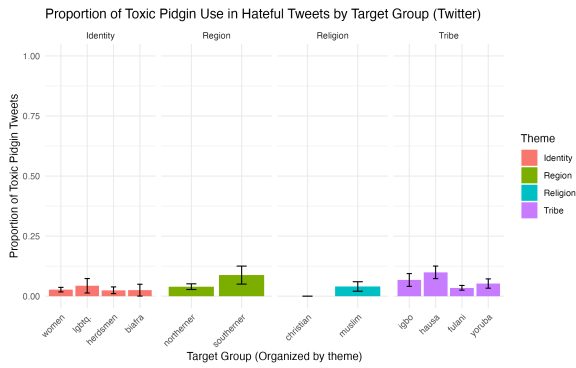


Figure 8: Distribution of Nigerian Pidgin in Hate Speech by Target Group in Twitter Data

mobilized in online hostility where the speaker seeks resonance within an in-group audience. On Twitter, the overall proportions are smaller for all groups. Across the two platforms, the Hausa and LGBTQ+ groups are the dominant NP targets. Thus, the selective use of NP demonstrates that language choice in hate speech is sensitive to the audience.

The Hausa case is particularly notable. As the group most frequently targeted on both social media platforms in NP, their prominence suggests that the language is strategically mobilized in relation to the Hausa group, potentially reinforcing their marginalization. Similarly, NP’s recurrent use in hate directed at women and LGBTQ+ groups highlights its affective force in amplifying ridicule and hostility within these group audiences. Nigerian Pidgin was least frequently employed in hostile comments directed at non-Nigerians. Instead, English predominated in these cases. This suggests that NP is primarily reserved for intra-national targeting, where it achieves stronger affective resonance within the local speech community. This supports Osoba’s (2015: 138) claim that NP can be regarded as a highly sensitive informal language where “nastier, sharper, more basal, and more naturally unobtrusive conceptions and inclinations towards a brutally lower level of emotion laden acquaintance can be observed to feature more prominently” than any other language captured in the dataset. Thus, the selective use of NP demonstrates that language choice in hate speech is sensitive to the audience. Speakers appear to choose NP when addressing co-nationals, amplifying hostility through a language that maximizes shared cultural meaning, while turning to English in addressing outsiders. These findings show that hate detec-

tion cannot be based solely on generic NP lexicons. Models must account for target-specific usage. Bias-aware training and group-sensitive annotation schemes are needed to mitigate misclassification and under-detection of harm.

5.3 RQ3: What rhetorical strategies appear in NP hate speech?

Insults are the dominant hate strategy across platforms and languages. The frequency analysis of the hateful class revealed that Nigerian Pidgin (NP) terms such as “mumu” (fool) and “ashawo” (prostitute) were among the 20 most frequently used words. These elements, documented in online resources such as Naija Lingo, exemplify how NP functions as a lexical resource for emotionally charged expressions. Their prominence underscores the role of NP as a key medium through which hostility and contempt are linguistically encoded in Nigerian digital spaces. In contrast, a frequency analysis of the hateful class in the Twitter data showed that English terms such as “stupid” and “useless” were among the 35 most frequent words. This pattern is not surprising, given that only 8% of the Twitter dataset is written in Nigerian Pidgin. Notably, across both datasets, women emerged as the group most frequently targeted with insults.

The dependence on NP for insults must be understood within Nigeria’s multilingual ecology. As an English-based creole shaped by indigenous languages, NP combines broad intelligibility with an affective, informal register (Obi, 2014; Osoba et al., 2015). Previous studies show that speakers often choose NP to express humor, solidarity, or critique (Affia, 2025; Oluyinka Adebayo, 2023). However, in hostile contexts, this same versatility allows NP to be strategically mobilized to intensify insult and ridicule, particularly against marginalized groups. Negative attitudes toward NP, considering it an inferior language and associated with low education, vulgarity, and lack of prestige (Osoba et al., 2018; Akande and Salami, 2010; Igboanusi, 2008b; Mann, 1996; Nwoda, 2023) further reinforce its role as a linguistic resource to position targets as socially inferior.

Beyond insults, dehumanizing language appeared mostly against the Muslim group on Instagram and women on Twitter. Mendelsohn & Budak (2025) highlight how metaphors such as natural disasters (e.g. ‘floods of immigrants’ or ‘infestations’, operate at the level of discourse to re-

inforce exclusionary ideologies and justify hostile policies toward migrants. In the Nigerian context, while the specific metaphors differ, the logic of dehumanization remains consistent. These framings are not merely rhetorical; they have material consequences, as dehumanizing discourse lowers thresholds for violence and legitimizes social or physical harm (Haslam, 2006; Haslam and Stratemeyer, 2016). By showing that Nigerian online discourse deploys dehumanization in ways like how western immigration discourse deploys metaphors, we provide further evidence that metaphor is a mechanism through which hostility is linguistically constructed. It also emphasizes the urgent need for hate speech detection models to account for metaphorical and figurative language, which may encode hostility more subtly than direct insults but can be equally potent in legitimizing violence.

Demonizing language was used against several groups, with women being the most frequent targets on Instagram, reflecting broader discourses that portray women as the source of societal decline. However, it was used the most to target the Fulani group on Twitter. Notably, incitement to violence was directed solely against women on Instagram, where examples explicitly encourage harm, attack, and even death. These patterns indicate that women occupy a uniquely vulnerable position in Nigerian online hate discourse, targeted not only with contempt but also with explicit calls for violence. On Twitter, it was used across targets but in very low proportions.

6 Discussion

Taken together, our results show NP plays a systematic rhetorical role in hostile online discourse. NP functions as a resource for expressing affective meaning in ways English often does not. Its strategic deployment reflects broader sociolinguistic patterns of language choice in multilingual societies, where codes carry ideological and affective meanings (Igboanusi, 2008a; Ifukor, 2011). NP, often dismissed as “low status,” is mobilized to reinforce social hierarchies and intensify hostility against vulnerable groups. This is evident in the dataset, where NP is disproportionately used in hateful expressions targeting specific groups. For NLP, these findings underscore the necessity of incorporating sociolinguistic insight into computational models. This is because hate speech expression in Nigerian social media is deeply shaped by local linguistic

practices and sociocultural meanings that current NLP models fail to capture. The cross-domain overprediction by the NaijaHate classifier reveals practical consequences: models trained on platform-specific data (Twitter) can misclassify content in other environments (Instagram), especially where language use and conversational norms differ. This has implications for moderation systems and for dataset curation: domain adaptation and platform-aware annotations are necessary. Computationally, the use of NP (and NP’s overlap with Yoruba loanwords) suggests model architectures need to be sensitive to code-mixed tokens, loanword handling, and figurative devices. Multilingual embeddings, metaphor detection modules, and bias-aware metrics can help, as can human-in-the-loop annotation practices that incorporates local linguistic expertise.

7 Conclusion

This paper provided the first systematic corpus analysis of Nigerian pidgin in online hate speech across Twitter and Instagram. By combining manual annotation with error analysis of a classifier trained on the NaijaHate Twitter corpus, we show that NP is disproportionately used in offensive contexts and is an important medium for insults and other figurative hate strategies, particularly targeting Hausa, women and LGBTQ+ groups. Our findings suggest that NP is not merely an informal code, but a linguistic resource strategically deployed to reinforce stereotypes and intensify hostility. The sociolinguistic associations of NP with low prestige make it especially potent in positioning groups as inferior, while its broad intelligibility ensures affective resonance in multiethnic online settings. Cross-domain evaluation also reveals that Twitter-trained classifiers over-predict hate on Instagram, underscoring the need for domain adaptation.

For NLP, these results underscore three key implications. First, hate detection in Nigerian and other multilingual contexts requires NP-specific resources, including lexicons, embeddings, and annotated corpora. Second, models must be sensitive to figurative and metaphorical strategies, which encode hostility more subtly than direct slurs. Third, domain adaptation across platforms is essential, as language use varies significantly between Twitter and Instagram. By foregrounding both sociolinguistic and computational perspectives, this work bridges a critical gap and provides a foundation

for building more robust, culturally informed hate detection systems.

8 Future Work

Future research should expand the size and balance of annotated datasets, particularly for the hateful class, to improve statistical power and classifier robustness. They should also expand demographic annotation to contextualize language choice. Extending analysis beyond Instagram and Twitter to platforms such as TikTok or WhatsApp would test the portability of models across even more diverse digital environments. Incorporating multilingual embeddings, metaphor detection modules, and bias-aware evaluation metrics will be crucial for capturing the subtle, context-dependent ways in which hostility is encoded. Finally, collaboration with linguists, sociologists, and local communities can help ensure that computational approaches remain sensitive to the cultural and political dynamics of language choice in Nigeria and other multilingual societies.

Limitations

This work has limitations. Datasets lack user demographics, limiting sociolinguistic contextualization. It remains unclear whether NP's emotional impact is carried by specific words or by full-language usage in code-mixed vs single-language contexts; experimental work is needed because the findings here are largely descriptive. Resource constraints meant annotators were not expert specialists, which may affect label quality. Distinguishing offensive vs hateful categories remains challenging.

References

Efurosibina E Adegbija. 2004. *Multilingualism: A Nigerian case study*. Africa World Press.

Folake Oluwatoyin Adegoke, Bashir Tenuche, and Eneh Agozie. 2024. Development of pidgin english hate speech classification system for social media. *American Journal of Information Science and Technology*, 8(2):34–44.

B Adepoju and T Kalu. 2022. An ideological analysis of hate speech comments on selected online new reports.

Precious Isaac Affia. 2025. Nigerian pidgin: The identity of a nigerian away from home.

AA Agantiem and Joseph O Alagbe. 2023. A comparative study of nigerian english pidgin usage among

students in federal university of lafia. *Ansu Journal of Language and Literary Studies*, 2(3).

- Akinmade T Akande and L Oladipo Salami. 2010. Use and attitudes towards nigerian pidgin english among nigerian university students. In *Marginal Dialects: Scotland, Ireland and Beyond*. Aberdeen: Forum for Research on the Languages of Scotland and Ireland, pages 1–79.
- Saminu Mohammad Aliyu, Gregory Maksha Wajiga, Muhammad Murtala, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, and Ibrahim Said Ahmad. 2022. Herdphobia: A dataset for hate speech against fulani in nigeria. *arXiv preprint arXiv:2211.15262*.
- Doris Chinedu Asogwa, Chiamaka Ijeoma Chukwunke, CC Ngene, and GN Anigbogu. 2022. Hate speech classification using svm and naive bayes. *arXiv preprint arXiv:2204.07057*.
- Abinew Ali Ayele, Skadi Dinter, Seid Muhie Yimam, and Chris Biemann. 2023. Multilingual racial hate speech detection using transfer learning. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 41–48.
- ABIODUN COMFORT AYENI. 2018. Socio-pragmatic analysis of online hate speeches in nigeria.
- ABIODUN COMFORT AYENI. 2024. *Pragmatic Analysis of Hate Rhetoric on Selected Social Media Platforms in Nigeria*. Ph.D. thesis, Doctoral dissertation, Federal University Lokoja.
- Babak Bahador. 2020. Classifying and identifying the intensity of hate speech. *Social Science Research Council*. <https://items.ssrc.org/disinformation-democracy-and-conflictprevention/classifying-and-identifying-the-intensity-of-hate-speech>.
- Temitope Abiodun Balogun. 2013. In defense of nigerian pidgin. *Journal of languages and culture*, 4(5):90–98.
- Taofeek Olaiwola Dalamu. 2017. *A discourse analysis of language choice in MTN® and Etisalat® advertisements in Nigeria*. University of Lagos (Nigeria).
- A Seza Dođruöz, Sunayana Sitaram, Barbara Bullock, and Almeida Jacqueline Toribio. 2021. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666.
- Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

- Nicholas Faraclas. 2021. Naija: A language of the future. *Current trends in Nigerian Pidgin English: A sociolinguistic perspective*, pages 9–38.
- Raisa Romanov Geleta, Klaus Eckelt, Emilia Parada-Cabaleiro, and Markus Schedl. 2023. Exploring intensities of hate speech on social media: A case study on explaining multilingual models with xai. In *Proceedings of the 4th Conference on Language, Data and Knowledge*, pages 532–537.
- Nick Haslam. 2006. Dehumanization: An integrative review. *Personality and social psychology review*, 10(3):252–264.
- Nick Haslam and Michelle Stratemeyer. 2016. Recent research on dehumanization. *Current Opinion in Psychology*, 11:25–29.
- Presley Ifukor. 2011. Linguistic marketing in “... a marketplace of ideas”: language choice and intertextuality in a nigerian virtual community. *Pragmatics and Society*, 2(1):110–147.
- Herbert Igboanusi. 2008a. Changing trends in language choice in nigeria. *Sociolinguistic Studies*, 2(2):251–269.
- Herbert Igboanusi. 2008b. Empowering nigerian pidgin: A challenge for status planning? *World Englishes*, 27(1):68–82.
- Comfort Ilevbare, Jesujoba Alabi, David Ifeoluwa Adelani, Firdous Bakare, Oluwatoyin Abiola, and Oluwaseyi Adeyemo. 2024. Ekohate: Abusive language and hate speech detection for code-switched political discussions on nigerian twitter. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 28–37.
- Instaloader Developers. 2024. Instaloader: Download instagram photos and metadata. [urlhttps://instaloader.github.io/](https://instaloader.github.io/).
- Charles C Mann. 1996. 10 anglo-nigerian pidgin in nigerian education: A survey of policy. *Language, Education, and Society in a Changing World*, page 93.
- Shamsuddeen Hassan Muhammad, Idris Abdulmu-min, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Sa’id Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwunke, and 1 others. 2025. Afrihate: A multilingual collection of hate speech and abusive language datasets for african languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1854–1871.
- Mbeledogu Njideka Nkemdilim and Ike-Okonkwo Mishael Somtochukwu. 2024. Navigating the dark web of hate: Supervised machine learning paradigm and nlp for detecting online hate speeches. *International Journal of Advanced Engineering Research and Science*, 11:3.
- Ogochukwu C Nweke, Koryoe Anim-Wright, and Bukola Towuru. 2024. The influence of pidgin english in nigerian media and its societal impact: A systematic literature review. *Journal of Humanities, Arts and Social Science*, 8(12).
- Chinazor Nwoda. 2023. Attitudes regarding the use of nigerian pidgin english among nigerian students at coventry university. *SCHOOL OF LANGUAGES, CULTURES AND LINGUISTICS SOAS UNIVERSITY OF LONDON*, page 46.
- Edith Ifeyinwa Obi. 2014. Language attitude and nigerian pidgin. *AFRREV IJAH: An International Journal of Arts and Humanities*, 3(4):34–46.
- A Oduma and V Gomwalk. 1986. Towards a typology of variation in nigerian english: A critique of some existing frameworks of analysis. In *17th NESAC Conference. University of Lagos*.
- Mosunmola Oluyinka Adebayo. 2023. Exploring the meaning of pidgin english on social media: A sociolinguistic analysis of nigerian pidgin hashtags as adapted speech. , 4(2):68–87.
- Chuka Fred Ononye and Nkechinyere Juliana Nwachukwu. 2019. Metalinguistic evaluators and pragmatic strategies in selected hate-inducing speeches in nigeria. *Indonesian Journal of Applied Linguistics*, 9(1):48–57.
- Joseph Babasola Osoba. 2014. The use of nigerian pidgin in media adverts. *International Journal of English Linguistics*, 4(2):26.
- Joseph Babasola Osoba and 1 others. 2015. Analysis of discourse in nigerian pidgin. *Journal of universal language*, 16(1):131–159.
- Joseph Babasola Osoba and 1 others. 2018. Power in nigerian pidgin (np) discourse. *Journal of Universal Language*, 19(1):1–32.
- Jana Papcunová, Marcel Martončík, Denisa Fedáková, Michal Kentoš, Miroslava Bozogánová, Ivan Srba, Robert Moro, Matúš Pikuliak, Marián Šimko, and Matúš Adamkovič. 2023. Hate speech operationalization: a preliminary examination of hate speech indicators and their structure. *Complex & intelligent systems*, 9(3):2827–2842.
- Taiwo Soneye. 2019. David jowitt: Nigerian english.
- Rotimi Taiwo. 2010. The dynamics of language mixing in nigerian digital communication. In *Handbook of research on discourse behavior and digital communication: Language structures and social interaction*, pages 180–190. IGI Global Scientific Publishing.
- Zeerak Talat, Thomas Davidson, Dana Warmesley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the first workshop on abusive language online*, pages 78–84.

- Manuel Tonneau, Pedro Quinta De Castro, Karim Lasri, Ibrahim Farouq, Lakshmi Subramanian, Victor Orozco-Olvera, and Samuel Fraiberger. 2024. Naijahate: Evaluating hate speech detection on nigerian twitter using representative data. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9020–9040.
- United Nations. 2019. United Nations Strategy and Plan of Action on Hate Speech. <https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf>. Accessed on [Date you accessed the document].
- Donatus George Usoro and Afaha Nsit. 2024. The status of nigerian pidgin in nigeria’s linguistic landscape: Another examination. *Erudite Compendiums in Education*, pages 89–100.
- Francielle Vargas, Isabelle Carvalho, Fabiana Rodrigues de Góes, Thiago Pardo, and Fabrício Benevenuto. 2022. Hatebr: A large expert annotated corpus of brazilian instagram comments for offensive language and hate speech detection. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7174–7183.
- William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the second workshop on language in social media*, pages 19–26.
- Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and characterizing hate speech on news websites. In *Proceedings of the 12th ACM conference on web science*, pages 125–134.

Appendix: Annotated Hate Speech Examples

Platform	Targeted group	Insults	Dehumanization	Demonization	Incitement to violence
Twitter	Women	women are actually the greatest hypocrites of all time, all these trolls coz it was not the other way round? A creature that always wants to be on the receiving end... Enjoy your family...	Women are Scum	na only God go save us from all this girls hand, they are all blood suckers (bed bug)	
	Lgbtq+	play good football and stay off of that madness,homosexuals are insane people animals don't gay not to talk of humans	Gay baboon doesn't know about getting consent?	Gay Terrorist	
	Herdsmen	You all really thought you had the monopoly of crime. The only reason I think herdsmen and some other northerners are doing all this crap is because they are uneducated and brainwashed, but "uneducated" comes first for a reason.	We demand aâ€¦ @USER Yet one foolish talkative will come here 2 bark like a curse dog dat they don't suppose 2 chase or killed Fulani cows, but when killer herdsmen were massacring, maiming and raping our people she lost her voice.	Killer herdsmen are heartless terrorists, they killed and they are killing people in North. The same herder-terrorists (might have also) disguised as UGM to kill the police. LINK	
	Biafra	The moment you ask questions, they resort to insults, abuse and name-calling. @USER Swears nothing theyâ€™d will make foolish tribe who thinks everyone is against them like theyâ€™re the best thing in the world win. Biafraud biafools now theyâ€™re Nigerians suddenly confused souls These Biafrans just want to waste a part of Nigeria	Crowd funding a terrorist organization, u are a magnetic fool Igbos are smart nd industrious people , but IPOB are the pigs and idiotic morons among them	Accept it ! You ipobians , are terrorist . How many people did members of IPOB kill before President Buhari ordered an armed invasion of the South East?	
	Northerner	Northern Nigeria and their lies. They formed BH and they are Reaping the result. My heart bleeds for the innocent ppl loosing their lives.	I stand with southern governors. No to open grazing! Northern extremists have continuously destroyed investments in alcoholic beverage belonging to southerners with impunity but want the South to consult them before Open Grazing can be banned in the South. Away with parasites! LINK	Northerners are the biggest, greediest, erratically selfish and nepotistic thieves & corrupt people in Nigeria...	They are terrorists supported by northern leadersâ€¦. The truth is the north is ok with themâ€¦.. let them burn! But should not cross down to the south. They should kill their own people.
	Southerner	Should worry his southern Christian morons.	Some people are dogs They are blaming	You are very stupid, an undesirable idiot, Fulani	

Appendix: Annotated Hate Speech Examples

			northerners for the deed of SA people Northerners are innocent but southerners are looking for trouble look at how heartless southerners people are blaming us	are not the course of any conflict the southerners are because of their greedy and selfishness. You want dominate the country so you start hating the Fulani after that the northern, we know all I'm with @USER Nigeria will never change, cause of these Narcissistic Southerners especially Igbo people they hate us, they are happy with how this #COVID__19 spreading in kano.	
	Christian	If there are two christians and they do not understand each other,then one of them must be a bastard in christ jesus.		I'm an atheist born of Christians but at the same time I'm in love with a Muslim. I actually see Christians as the biggest problem but muslim fanatics the biggest mistake the world has	
	Muslim	Some of you Muslims are THE most judgemental people ever. You say the most disgusting things to your fellow brothers and sister	If this is what Islam teaches, then, Mohammad and Allah are scum and shit	Islamic terrorists would have overrun Igbo land if not the presence of ESN and the so called IPOB. Those animals are afraid of these two - IPOB and ESN. Don't ever attribute evil deed to these two, but the government mercenaries	useless fool say islamic terrorism, the world has to kill islam b4 islam kills the world. we know you added ojukwu to ur name to look an Igbo but u r just a pig.
	Igbo	Igbo people are lazy,jobless, deceitful? Nah sorry we dont fit into that category...	Igbos are cannibals	idiot, don't change the rhetoric.These are simply Igbo kidnappers.Most of your people hide under the guise of Fulani herdsmen to kidnap people for a living, and fools like you would crucify and persecute the Fulani herders for your evils	STOP SPREADING HATE MSGS! By the way I am IGBO! @USER Igbos re d scum of d Earth. Jeopardizing d future of Nigerians..kill them all"
	Hausa	A Hausa man is an ungrateful human. Hypocrites, fanatics,pretenders and evil doers	Na Hausa you be I can't blame you Come to southern you will see what's happening	I agree with you on this, we need to find a way of repatriating these abokis in the south back to their	Shoot me 4 saying dis bt I tink Hausas especially d muslim ones shld b exterminated "Le

Appendix: Annotated Hate Speech Examples

			Mumu man You go be the next person that sars go kill even your family #animal #EndSarsNow	north. The ploy to take over Nigeria is so real.	struggle to trend. Bitch have a seat on thorns
	Fulani	Idiots! All of you are idiots! All this talk is to kill more Igbo people on the east without investigating anything. Buhari has never sounded this way about Terrorists ravaging Nigeria, murderous fulani herdsmen God will wash you and	Leave the deaf Fulani baboon	Are you happy that fulani herdsmen are killing, raping women and destroying people's farms and property? Do fulani have any land in the south west or south east? Can an ibo/yoruba man come to the north with pigs or goats to destroy the farmland	Sometimes, Fayose is my guy. Other times I just cringe at his actions. But on this murderous Fulani bastards, I support him! KILL THEM ALL!
	Yoruba	God!!! Yoruba's are so daft! I never see	They are mostly yoruba pigs	Yorubas are the problem of this country, you can never see hausa man to treat his brother like this coz of a peaceful protest... Animal in uniform, you guys will soon get sense by force Yorubas are mugus nah, we've sold our right to the Hausa people so they are the one controlling us. Why can't they just share the materials in Lagos and distribute to every other state? Wicked souls ðŸ˜ðŸ˜ðŸ˜ðŸ˜ LINK RT @USER: Yorobber criminals & thieves ve been defeated. Many of them are having heart attack right now as I am writing.	Yoruba dirty goat like u should slaughtered by boko haram. Idiat fool
Instagram	Women	women na real wa for una anything for money some of you would do totally ashamed	everything about this gender na cheating cheating as if we have no life aside cheating men doesnt own women any form or loyalty i brought you into my life	women are more deadly when it comes to power this your analysis is not a reason	girls too many for this world sef let them continue make feminist reduce
	Lgbtq+	thats why you decided to be a gay you are very use less		na gay people cause em	
	Northerner	northern nigeria has to be the shame of africa		anything wey no be religious violence no concern northerners	
	Christian	sebi bible says if dem slap u u turn the other cheek look at christians embarrassing themselves las las everybody just dey play christ like kor			
	Muslim	mumu muslims under bridge na muslims full there dey suffer	you see this when a man or group of men start forcefully controlling women they themselves are weak very weak thats why they look for the weak beings to		

Appendix: Annotated Hate Speech Examples

			control and treat any how animals in veils always causing problems in the name of islam		
	Igbo	some brainless ibo sha the price of rice in ibadan must be different from imo state abi always the victim		later now they will say igbo are not tribalistic igbos are the must tribalistic and the haters of Nigeria we dont even want the protest so allow people to complain as united nigerians but the protest is not a good thing but its the last option	
	Hausa	this hausa people for this video so and sense are far apart cause wtf are you singing for someone who doesnt even care about your existance someone who dislikes the black		hausa na the problem of this country sense dey always pain them	
	Yoruba	yoruba president the worst president ever		yorubas people are the one destroying nigeria and yoruba leaders are all control by usa and nato countries	

The Token Tax: Systematic Bias in Multilingual Tokenization

Jessica M. Lundin¹, Ada Zhang², Nihal Karim², Hamza Louzan²,
Victor Wei², David Ifeoluwa Adelani³, Cody Carroll^{2,4}

¹Institute for Disease Modeling, Gates Foundation,

²Data Institute, University of San Francisco,

³Mila - Quebec AI Institute, McGill University, and Canada CIFAR AI Chair,

⁴Department of Mathematics and Statistics, University of San Francisco

Correspondence: jessica.lundin@gatesfoundation.org

Abstract

Tokenization inefficiency is associated with structural disadvantages for morphologically complex, low-resource languages, inflating compute resources and reducing accuracy. We evaluate 10 large language models (LLMs) on AfriMMLU (5 subjects; 16 African languages) and show that token fertility reliably predicts accuracy. Higher fertility consistently predicts lower accuracy across all models and subjects. We further find that reasoning models (e.g., DeepSeek, o1) consistently outperform non-reasoning peers across high- and low-resource languages in the AfriMMLU dataset, narrowing accuracy gaps observed in prior generations. In terms of economics, a doubling in tokens results in quadrupled training cost and time, underscoring the "token tax" faced by many languages. These results motivate morphologically aware tokenization, fair pricing, and multilingual benchmarks for equitable natural language processing (NLP).

1 Introduction

Tokenization serves as the foundational layer of modern NLP systems, yet it is associated with systematic inequalities that disproportionately affect morphologically complex and low-resource languages. Prior work decisively establishes tokenization as a source of computational and economic inequality (Ahia et al., 2023), with quantified impacts ranging from inflated token counts to substantial BLEU point performance degradation (Petrov et al., 2023; Ali et al., 2024). Because transformer attention scales quadratically with sequence length, even modest increases in token counts can meaningfully raise compute requirements and reduce effective context capacity (Keles et al., 2022). As a result, morphologically complex languages with high fertility values suffer from compound disadvantages that are difficult to fully overcome within widely deployed transformer architectures (Sreedhar et al.,

2023). Recent work on adaptive and flexible tokenization strategies (e.g., MAGNET; FlexiTokens) suggests promising directions for mitigating these inefficiencies, but these approaches are not yet standard in large-scale multilingual LLM training and remain mostly unevaluated on low-resource African languages (Ahia et al., 2024; Owodunni et al., 2025).

These disparities particularly affect morphologically rich languages, where agglutinative and fusional morphology leads to systematic tokenization inefficiency. The technical disparities translate directly into economic exclusion through what we term the "token tax", prohibitive training and inference costs measured in dollars and tons of CO₂, and systematic underrepresentation in model capabilities that affects billions of speakers worldwide.

A reasonable cost to train a small-medium model or a large frontier model is easily \$1M (1 month) to \$100M (~3 months) with primarily English tokens (see Appendix C for derivation of estimates from publicly reported petaFLOP-day requirements and standard cloud compute pricing). If we instead train on a language with 2× or 5× more tokens for the same content, the transformer’s quadratic $O(n^2)$ compute costs result in a 4× or 25× increase in energy consumption, dollar cost, training time, and CO₂ emissions relative to English. In this example, the cost becomes \$4-25M (4 months-2 years) or \$400M-2.5B (1-6 years), respectively.

Our contributions herein are as follows:

- We extend prior fertility and accuracy analysis to 10 models and 16 languages, confirming fertility as a reliable predictor of MCQA accuracy.
- We conduct the first large-scale comparison of tokenization effects for reasoning vs. non-reasoning LLMs on AfriMMLU, showing that reasoning capabilities substantially reduce but do not eliminate tokenization bias.

- We quantify the economic impact of tokenization inefficiency, demonstrating how the "token tax" creates barriers to multilingual NLP development.
- We release public datasets containing: (i) model results from AfriMMLU benchmark including reasoning models, (ii) comprehensive tokenization metrics.

2 Related Work

The impact of tokenization on multilingual model performance has received increasing attention. [Petrov et al. \(2023\)](#) demonstrated that tokenizers disadvantage non-English languages, with text length exceeding 15 times for some language pairs. [Rust et al. \(2021\)](#) showed that tokenizer quality significantly impacts downstream performance, finding that morphologically complex languages suffer from both vocabulary underrepresentation and sub-optimal segmentation strategies. [Ali et al. \(2024\)](#) conducted extensive ablation experiments showing that tokenizer choice can impact downstream performance and increase training costs by 68%. Their analysis is for European languages that have different fertility distributions than African languages, the focus of this work.

[Joshi et al. \(2020\)](#) categorized languages into six resource levels, showing that 88% of world languages fall into the lowest category. These systemic inequalities correlate strongly with tokenization efficiency, suggesting that addressing tokenization could substantially improve language technology equity.

Recent benchmarks have begun addressing the evaluation gap for African languages ([Adelani et al., 2025](#); [Singh et al., 2025](#); [Alhanai et al., 2024](#); [Adebara et al., 2025](#); [Beyene et al., 2025](#); [Team et al., 2022](#)). While these benchmarks evaluate various aspects of multilingual performance, from culturally relevant knowledge questions to speech tasks, they primarily focus on documenting performance gaps without systematically analyzing the underlying tokenization disparities. Our work complements these efforts by explicitly connecting tokenization efficiency to performance degradation, revealing that much of the observed performance gap can be attributed to systematic tokenization bias rather than model capability limitations.

While the computational complexity of transformers is well-established ([Vaswani et al., 2017](#)), the economic implications of tokenization inefficiency

remain underexplored. Our work bridges this gap by quantifying both performance degradation and financial costs, demonstrating how technical choices create economic barriers to equitable NLP development.

3 Experimental Setup

3.1 Dataset and Languages

AfriMMLU ([Adelani et al., 2025](#)) comprises 9,000 multiple-choice questions across 5 subjects: elementary mathematics, global facts, high school geography, high school macroeconomics, and international law. The benchmark covers 16 African languages spanning multiple language families including Niger-Congo (e.g., Swahili, Yoruba, Zulu), Afro-Asiatic (e.g., Amharic, Hausa), and Nilo-Saharan, providing substantial typological diversity. Each subject contains between 1,500 and 2,000 questions, professionally translated by native speakers with quality verification.

3.2 Model Selection

We evaluate 10 models spanning three categories to enable comparison, including general LLMs (Llama 3.1 405B, Gemini 1.5 Pro, Claude Sonnet 3.5, DeepSeek V3, GPT-4o, Qwen 2.5 32B) and multilingual-focused ones (Aya 23 35B, Pixtral 12B). We also include reasoning models (DeepSeek R1, OpenAI o1) to test whether enhanced reasoning capabilities mitigate tokenization-related performance decline. If reasoning models show reduced gaps despite similar tokenization inefficiency, this would suggest that architectural advances can partially mitigate tokenizer limitations.

3.3 Metrics and Analysis

Our primary metric is Fertility ($F = T / W$), the average number of tokens per word, quantifying tokenization efficiency. Lower fertility indicates more efficient tokenization for a given language. Our analysis also includes zero-shot accuracy of MCQA tasks using the prompt template detailed in the Appendix.

Performance gaps are calculated as the difference between English and French accuracy and the mean accuracy across all 16 African languages for each model-subject combination. English and French here provide baselines near optimal tokenization efficiency. The "Random" baseline of 25% represents the probability of guessing a cor-

rect response by chance for multiple choice questions with 4 options.

3.4 Statistical Methodology

The pipeline for analysis is as follows:

- 1: **for** each language ℓ in AfriMMLU **do**
- 2: **for** each model m **do**
- 3: Tokenize all questions using m 's tokenizer
- 4: Calculate fertility $F_{\ell,m} = \frac{\text{tokens}}{\text{words}}$
- 5: Evaluate zero-shot accuracy on all subjects
- 6: Store results for regression analysis
- 7: **end for**
- 8: **end for**
- 9: Fit linear models: Accuracy \sim Fertility for each model-subject pair
- 10: Calculate regression statistics: slopes, standard errors, t-values, p-values
- 11: Calculate Pearson correlation coefficients (ρ), R^2 , and adjusted R^2
- 12: Apply Benjamini-Hochberg FDR correction for multiple comparisons

We conduct linear regression analysis for each model-subject combination, treating fertility as the predictor and accuracy as the outcome variable. For each regression, we report the intercept, slope (effect of fertility on accuracy), standard error, t -statistic, and p -value. We calculate Pearson correlation coefficients (ρ) to quantify the strength and direction of the linear relationship, and coefficients of determination (R^2) to measure the proportion of variance in accuracy explained by fertility. Adjusted R^2 values correct for sample size and number of predictors. Statistical significance is assessed at $p < 0.05$ with Benjamini-Hochberg False Discovery Rate (FDR) correction to account for multiple comparisons across 50 model-subject pairs. We consider $R^2 \geq 0.25$ as large effects and $|\rho| \geq 0.50$ as strong correlations following standard conventions.

4 Results

4.1 Performance Gaps Across Languages

Figure 1 shows substantial performance disparities across languages. African languages trail English by 30 percentage points and French by 24 percentage points on average across model results. This gap varies by subject, with Geography and Economics showing the largest disparities and Elementary Mathematics and Global Facts showing

relatively smaller gaps.

Reasoning models DeepSeek R1 and o1 reduce performance gaps across subjects. These models outperform non-reasoning counterparts while maintaining strong English performance, suggesting that enhanced reasoning capabilities provide particular benefits for low-resource settings.

4.2 Tokenization as Performance Predictor

Figure 2 demonstrates the strong negative relationship between fertility and accuracy for Llama 3.1 405B as an example model (results for all models are displayed in Figure 3). Across all 10 models and five subjects, higher fertility consistently predicts lower accuracy. Regression analyses quantify this relationship with slopes ranging from -0.08 to -0.18, meaning each additional token per word reduces accuracy by 8 to 18 percentage points on average.

We note that fertility is not independent of language resource level, that is languages underrepresented in training corpora tend to have higher fertility due to fewer subword merges during tokenizer training. The observed fertility accuracy relationship is likely reflecting both the direct tokenization effects and indirect effects mediated by training data availability.

Table 1 reports detailed regression results. Several effects are both large and statistically significant after FDR correction, including Llama-3.1-405B on Microeconomics (slope = -0.185 , $p = 0.002$) and Qwen-2.5-32B on Geography (slope = -0.155 , $p = 0.006$). Between 20 to 50% of variance in accuracy is explained by variation in fertility, with particularly strong effects in technical subjects requiring precise terminology.

5 Analysis

5.1 Economic Impact of Token Inflation

The tokenization inefficiencies documented above translate directly to economic barriers. Because transformer training scales quadratically with sequence length, a 2 \times increase in fertility produces a 4 \times increase in training time and cost. Table 2 quantifies these impacts for the Llama model family.

Inference costs show similar patterns. Generating 1M English-equivalent tokens costs \$5-20 with GPT-4o, but \$10-40 for a language with twice the token fertility. Latency doubles correspondingly, creating user experience degradation alongside cost inflation.

Model	Total	Math	Facts	Geog	Econ	Law
Baseline Performance (English Language)						
o1-preview-2024-09-12	91	99	75	91	97	91
DeepSeek-R1	90	100	65	95	99	91
gemini-15-Pro-002	88	93	69	91	96	92
gpt-4o-2024-08-06	89	95	68	92	97	91
DeepSeek-V3-0324	88	96	67	92	98	89
Llama-3.1-405B	86	85	66	92	97	89
claude-3-5-sonnet-202410	75	63	66	87	81	78
Qwen2.5-32B	79	78	55	86	88	89
phi-4	77	66	45	92	95	89
Pixtral-12B-2409	62	41	43	78	73	76
aya-23-35B	57	42	43	67	69	66
Random	25	25	25	25	25	25
Average Performance (all African Languages)						
o1-preview-2024-09-12	76	88	67	72	75	76
DeepSeek-R1	67	88	56	60	63	68
gemini-15-Pro-002	62	78	55	56	55	68
gpt-4o-2024-08-06	62	80	48	58	58	67
DeepSeek-V3-0324	54	76	49	47	45	54
Llama-3.1-405B	51	69	45	42	41	59
claude-3-5-sonnet-202410	50	59	46	48	43	52
Qwen2.5-32B	39	53	34	27	31	50
phi-4	37	44	36	29	29	48
Pixtral-12B-2409	33	35	31	30	31	39
aya-23-35B	24	26	20	19	24	28
Random	25	25	25	25	25	25
Performance Gap (English - African Languages)						
o1-preview-2024-09-12	15	11	8	19	22	15
DeepSeek-R1	23	12	9	35	36	23
gemini-15-Pro-002	26	15	14	35	42	24
gpt-4o-2024-08-06	26	15	20	34	39	24
DeepSeek-V3-0324	34	20	18	45	53	35
Llama-3.1-405B	35	16	21	50	57	30
claude-3-5-sonnet-202410	25	4	20	39	39	27
Qwen2.5-32B	40	25	22	59	57	39
phi-4	40	22	9	63	66	41
Pixtral-12B-2409	29	6	12	48	42	37
aya-23-35B	34	16	23	48	45	38

(a) Accuracy Aggregation (English)

Model	Total	Math	Facts	Geog	Econ	Law
Baseline Performance (French Language)						
o1-preview-2024-09-12	89	93	75	90	95	91
DeepSeek-R1	87	96	68	85	97	91
gemini-15-Pro-002	85	86	65	89	94	91
gpt-4o-2024-08-06	83	88	60	86	90	91
DeepSeek-V3-0324	83	89	63	86	90	88
Llama-3.1-405B	81	82	63	85	90	84
claude-3-5-sonnet-202410	71	70	53	87	80	66
Qwen2.5-32B	73	70	54	78	80	82
phi-4	73	65	47	83	82	86
Pixtral-12B-2409	55	37	41	66	58	71
aya-23-35B	44	35	28	41	51	65
Random	25	25	25	25	25	25
Average Performance (all African Languages)						
o1-preview-2024-09-12	76	88	67	72	75	76
DeepSeek-R1	67	88	56	60	63	68
gemini-15-Pro-002	62	78	55	56	55	68
gpt-4o-2024-08-06	62	80	48	58	58	67
DeepSeek-V3-0324	54	76	49	47	45	54
Llama-3.1-405B	51	69	45	42	41	59
claude-3-5-sonnet-202410	50	59	46	48	43	52
Qwen2.5-32B	39	53	34	27	31	50
phi-4	37	44	36	29	29	48
Pixtral-12B-2409	33	35	31	30	31	39
aya-23-35B	24	26	20	19	24	28
Random	25	25	25	25	25	25
Performance Gap (French - African Languages)						
o1-preview-2024-09-12	13	5	8	18	20	15
DeepSeek-R1	20	8	12	25	34	23
gemini-15-Pro-002	23	8	10	33	40	23
gpt-4o-2024-08-06	21	8	12	28	32	24
DeepSeek-V3-0324	29	13	14	39	45	34
Llama-3.1-405B	30	13	18	43	50	25
claude-3-5-sonnet-202410	22	11	7	39	38	15
Qwen2.5-32B	34	17	21	51	49	32
phi-4	36	21	11	54	53	38
Pixtral-12B-2409	22	2	10	36	27	32
aya-23-35B	20	9	6	22	27	37

(b) Accuracy Aggregation (French)

Figure 1: Baseline performance shows English (a) and French (b) accuracy (in percentage points). The mean accuracy across all 16 African languages is shown in the middle charts of (a) and (b). The bottom charts of (a) and (b) show performance gaps between the African languages and higher-resource languages, though reasoning-oriented models narrow this gap.

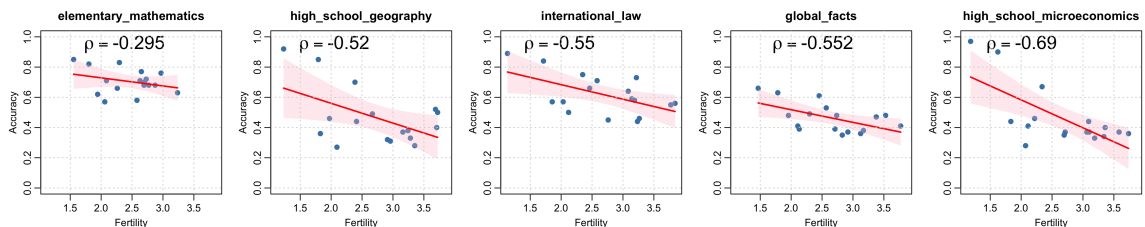


Figure 2: Fertility and accuracy for Llama 3.1 405B across subjects. Strong negative correlations (ρ) demonstrate systematic performance degradation with tokenization inefficiency. Fertility is not independent of training data representation; fertility captures tokenization inefficiency that covaries with performance, but does not isolate causal effects independent of pretraining data availability or quality.

Subject	Model	Intercept	Slope	Std. Error	t-value	P-value	ρ	R^2	Adj. R^2
Elementary Math	Sonnet3.5	0.652	-0.018	0.029	-0.609	0.552	-0.155	0.024	-0.041
	Aya23 35B	0.480	-0.079	0.019	-4.058	0.001*	-0.723	0.523	0.492
	DeepSeek R1	1.045	-0.066	0.045	-1.475	0.161	-0.356	0.127	0.068
	DeepSeek V3	0.884	-0.044	0.048	-0.922	0.371	-0.232	0.054	-0.009
	Gemini 1.5 Pro	0.907	-0.045	0.042	-1.078	0.298	-0.268	0.072	0.010
	Llama3.1 405B	0.836	-0.054	0.045	-1.195	0.251	-0.295	0.087	0.026
	Phi4	0.773	-0.125	0.034	-3.641	0.002*	-0.685	0.469	0.434
	GPT-4o	1.002	-0.089	0.057	-1.571	0.137	-0.376	0.141	0.084
	Pixtral 12B	0.417	-0.024	0.014	-1.717	0.106	-0.405	0.164	0.109
	Qwen2.5 32B	0.857	-0.113	0.037	-3.012	0.009*	-0.614	0.377	0.335
Global Facts	Sonnet3.5	0.508	-0.011	0.027	-0.390	0.702	-0.100	0.011	-0.056
	Aya23 35B	0.335	-0.044	0.023	-1.930	0.073	-0.446	0.199	0.146
	DeepSeek R1	0.574	-0.002	0.038	-0.061	0.952	-0.016	0.000	-0.066
	DeepSeek V3	0.619	-0.045	0.034	-1.308	0.211	-0.320	0.102	0.042
	Gemini 1.5 Pro	0.585	-0.011	0.051	-0.222	0.827	-0.057	0.003	-0.063
	Llama3.1 405B	0.685	-0.084	0.033	-2.564	0.022	-0.552	0.305	0.258
	Phi4	0.408	-0.015	0.017	-0.845	0.411	-0.213	0.045	-0.018
	GPT-4o	0.638	-0.063	0.054	-1.169	0.261	-0.289	0.084	0.022
	Pixtral 12B	0.428	-0.038	0.018	-2.068	0.056	-0.471	0.222	0.170
	Qwen2.5 32B	0.505	-0.052	0.024	-2.171	0.046	-0.489	0.239	0.188
High School Geography	Sonnet3.5	0.781	-0.080	0.045	-1.779	0.096	-0.417	0.174	0.119
	Aya23 35B	0.475	-0.097	0.038	-2.512	0.024	-0.544	0.296	0.249
	DeepSeek R1	0.847	-0.082	0.056	-1.466	0.163	-0.354	0.125	0.067
	DeepSeek V3	0.843	-0.124	0.053	-2.331	0.034	-0.516	0.266	0.217
	Gemini 1.5 Pro	0.750	-0.065	0.070	-0.937	0.363	-0.235	0.055	-0.008
	Llama3.1 405B	0.822	-0.131	0.055	-2.359	0.032	-0.520	0.271	0.222
	Phi4	0.808	-0.162	0.048	-3.343	0.004*	-0.653	0.427	0.389
	GPT-4o	0.952	-0.151	0.068	-2.211	0.043	-0.496	0.246	0.195
	Pixtral 12B	0.688	-0.121	0.035	-3.414	0.004*	-0.661	0.437	0.400
	Qwen2.5 32B	0.755	-0.155	0.049	-3.190	0.006*	-0.636	0.404	0.365
High School Microeconomics	Sonnet3.5	0.750	-0.096	0.042	-2.307	0.036	-0.512	0.262	0.213
	Aya23 35B	0.549	-0.105	0.038	-2.775	0.014	-0.582	0.339	0.295
	DeepSeek R1	0.888	-0.088	0.074	-1.194	0.251	-0.295	0.087	0.026
	DeepSeek V3	0.906	-0.157	0.049	-3.179	0.006*	-0.634	0.403	0.363
	Gemini 1.5 Pro	0.883	-0.129	0.067	-1.920	0.074	-0.444	0.197	0.144
	Llama3.1 405B	0.953	-0.185	0.050	-3.691	0.002*	-0.690	0.476	0.441
	Phi4	0.858	-0.184	0.053	-3.479	0.003*	-0.668	0.447	0.410
	GPT-4o	0.942	-0.150	0.084	-1.779	0.096	-0.417	0.174	0.119
	Pixtral 12B	0.622	-0.105	0.033	-3.179	0.006*	-0.635	0.403	0.363
	Qwen2.5 32B	0.779	-0.154	0.048	-3.196	0.006*	-0.636	0.405	0.365
International Law	Sonnet3.5	0.645	-0.040	0.028	-1.426	0.174	-0.346	0.119	0.061
	Aya23 35B	0.578	-0.101	0.042	-2.403	0.030	-0.527	0.278	0.230
	DeepSeek R1	0.813	-0.043	0.043	-1.010	0.329	-0.252	0.064	0.001
	DeepSeek V3	0.771	-0.073	0.045	-1.617	0.127	-0.385	0.148	0.092
	Gemini 1.5 Pro	0.796	-0.039	0.052	-0.758	0.460	-0.192	0.037	-0.027
	Llama3.1 405B	0.876	-0.096	0.038	-2.548	0.022	-0.550	0.302	0.256
	Phi4	0.804	-0.101	0.041	-2.452	0.027	-0.535	0.286	0.238
	GPT-4o	0.889	-0.085	0.072	-1.175	0.258	-0.290	0.084	0.023
	Pixtral 12B	0.686	-0.095	0.033	-2.859	0.012*	-0.594	0.353	0.310
	Qwen2.5 32B	0.787	-0.092	0.040	-2.297	0.036	-0.510	0.260	0.211

Table 1: Fertility and Accuracy by Model and Subject. Results from linear models regressing accuracy on fertility across 16 languages for each model-subject combination. The table reports intercepts, slopes (negative values indicate higher fertility correlates with lower accuracy), standard errors, t -statistics, and p -values for each regression. Pearson correlation coefficients (ρ) quantify the strength of the linear relationship. R^2 values show the proportion of variance in accuracy explained by fertility, and adjusted R^2 values correct for sample size. Bold p -values indicate statistical significance ($p < 0.05$). Asterisks (*) indicate results that remain significant after Benjamini-Hochberg FDR correction ($FDR < 0.05$). Bold ρ values indicate strong correlations ($|\rho| \geq 0.50$), and bold R^2 values indicate large effects ($R^2 \geq 0.25$). Regressions for o1 are not included because OpenAI has not released details on the tokenizer for this model.

These cost projections assume equivalent training objectives across languages. In practice, the relationship between fertility and cost interacts with

data availability: high-fertility languages lack training data to realize theoretical training costs, which is an additional and compounded barrier to devel-

opment.

5.2 Linguistic Factors Driving Fertility

The fertility disparities correlate strongly with morphological typology. Agglutinative languages in our dataset (Swahili: $F = 2.8$, Zulu: $F = 2.6$) consistently show higher fertility than more analytic languages. Fusional languages like Amharic occupy intermediate positions. This pattern suggests that current BPE-based tokenizers, predominantly trained on English and European language text, systematically fail to capture morphological processes. Affixation patterns that would constitute single semantic units are split across tokens, fragmenting the meaning and increasing the sequence length.

For example, Swahili verb forms that encode subject, tense, object, and mood in a single word are often split into multiple tokens, while equivalent English constructions using auxiliary verbs and pronouns may use fewer tokens despite containing more words. This fundamental mismatch between tokenizer training data and target language morphology drives the observed inefficiencies.

6 Conclusion

This study demonstrates that tokenization inefficiency predicts systematic disadvantages for morphologically complex languages. Across 10 large language models and 16 African languages, fertility explains up to 50% of the variance in model accuracy. In the strongest cases, each additional token per word is associated with accuracy drops of up to 18 percentage points.

While reasoning models like DeepSeek and o1 narrow accuracy gaps—improving African language performance by 8 to 12 points on average—substantial disparities remain. The economic implications are severe: doubling fertility quadruples training costs, creating a "token tax" that turns linguistic diversity into computational liability.

Addressing these inequities requires coordinated intervention: technical advances in morphologically-aware tokenization, economic reforms to pricing structures, and expanded multilingual evaluation infrastructure. Without such efforts, billions of speakers will remain excluded from the benefits of language technology, perpetuating digital divides along linguistic lines.

7 Limitations

Our analysis establishes correlation between fertility and accuracy but cannot establish causation. Token fertility correlates with training data availability, where languages with less representation in pre-training corpora receive fewer subword merges during tokenizer training, and have consequent higher fertility. The performance degradation observed may reflect tokenizer inefficiency, scant data, or a combined effect. Also, while we report aggregate performance for such models where possible, fine-grained fertility analyses cannot be conducted without tokenizer transparency.

In terms of access to model components, we could not analyze OpenAI's o1 model tokenizer as implementation details remain proprietary, limiting our ability to fully characterize reasoning models' tokenization strategies. Our analysis focuses on African languages which limits generalization to other language families. It's possible that similar patterns would be present for morphologically complex, low-resource languages, this is untested. While we analyze 16 African languages, this represents less than 1% of Africa's 2,000+ languages. Our findings may not generalize to languages with different morphological properties or non-Latin scripts. We evaluate only multiple-choice question answering, MCQA tasks. Performance degradation patterns may differ for generative tasks where token inflation affects both input and output, potentially compounding the effects we observe, but this remains speculation. Our cost calculations use publicly available pricing and may not reflect negotiated rates or future hardware improvements. We focus on direct computational costs without quantifying broader environmental impacts.

By quantifying performance gaps, we risk reinforcing perceptions that some languages are lesser for NLP applications. We emphasize that these disparities reflect technological limitations, not inherent language properties. Documenting the "token tax" could discourage investment in low-resource language technologies if stakeholders focus solely on costs rather than equity. This document is intended to motivate solutions. To address these risks, we suggest developing inclusive tokenization standards through community participation; advocating for subsidized compute resources for low-resource language research; and creating evaluation metrics that explicitly penalize tokenization inefficiency.

References

- Ife Adebara, Hawau Olamide Toyin, Nahom Tesfu Ghebremichael, AbdelRahim A. Elmadany, and Muhammad Abdul-Mageed. 2025. [Where are we? evaluating LLM performance on African languages](#). In [Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics \(Volume 1: Long Papers\)](#), pages 32704–32731, Vienna, Austria. Association for Computational Linguistics.
- David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Ijeoma Chukwunkeke, Happy Buzaaba, Blessing Kudzaishe Sibanda, Godson Koffi Kalipe, Jonathan Mukiibi, Salomon Kabongo Kabenamualu, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, and 8 others. 2025. [IrokoBench: A new benchmark for African languages in the age of large language models](#). In [Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies \(Volume 1: Long Papers\)](#), pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.
- Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Valentin Hofmann, Tomasz Limisiewicz, Yulia Tsvetkov, and Noah A Smith. 2024. [Magnet: Improving the multilingual fairness of language models with adaptive gradient-based tokenization](#). [Advances in Neural Information Processing Systems](#), 37:47790–47814.
- Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. 2023. [Do all languages cost the same? tokenization in the era of commercial language models](#). In [Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing](#), pages 9904–9923, Singapore. Association for Computational Linguistics.
- Tuka Alhanai, Adam Kasumovic, Mohammad Ghassemi, Aven Zitzelberger, Jessica Lundin, and Guillaume Chabot-Couture. 2024. [Bridging the gap: Enhancing llm performance for low-resource african languages with new benchmarks, fine-tuning, and cultural adjustments](#). [Preprint](#), arXiv:2412.12417.
- Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveiling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, Charvi Jain, Alexander Weber, Lena Jurkschat, Hammam Abdelwahab, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Samuel Weinbach, Rafet Sifa, and 2 others. 2024. [Tokenizer choice for LLM training: Negligible or crucial?](#) In [Findings of the Association for Computational Linguistics: NAACL 2024](#), pages 3907–3924, Mexico City, Mexico. Association for Computational Linguistics.
- Luel Hagos Beyene, Vivek Verma, Min Ma, Jesujoba O. Alabi, Fabian David Schmidt, Joyce Nakatumba-Nabende, and David Ifeoluwa Adelani. 2025. [msteb: Massively multilingual evaluation of llms on speech and text tasks](#). [Preprint](#), arXiv:2506.08400.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In [Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics](#).
- Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. 2022. [On the computational complexity of self-attention](#). [Preprint](#), arXiv:2209.04881.
- Abraham Toluwase Owodunni, Orevaoghene Ahia, and Sachin Kumar. 2025. [Flexitokens: Flexible tokenization for evolving language models](#). [arXiv preprint arXiv:2507.12720](#).
- Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. 2023. [Language model tokenizers introduce unfairness between languages](#). [Preprint](#), arXiv:2305.15425.
- Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. [How good is your tokenizer? on the monolingual performance of multilingual language models](#). In [Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing \(Volume 1: Long Papers\)](#), pages 3118–3135, Online. Association for Computational Linguistics.
- Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Sebastian Ruder, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, and 5 others. 2025. [Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation](#). [Preprint](#), arXiv:2412.03304.
- Makeesh Narsimhan Sreedhar, Xiangpeng Wan, Yu Cheng, and Junjie Hu. 2023. [Local byte fusion for neural machine translation](#). [Preprint](#), arXiv:2205.11490.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). [Preprint](#), arXiv:2207.04672.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). [Preprint](#), arXiv:1706.03762.

Appendix

A Regression Results by Subject

Table 1 and Figure 3 show results of accuracy-on-fertility regressions for the 10 models over 5 subjects.

B Fertility

Fertility measures the average number of tokens required to represent a word in a corpus:

$$F = \frac{T}{W}$$

for T and W the token and word counts. Higher F inflates sequence length, affecting a model’s ability to learn long-range dependencies and compute costs (Ali et al., 2024).

C Inference

C.1 Training and Inference Cost Comparison: English vs. Language X

This is a thought exercise in training and inference costs for LLMs applied to English and Language X. The analysis assumes the same model architecture and tokenizer across languages, with cost differences due to tokenization inefficiencies and quadratic $O(n^2)$ training scaling (Vaswani et al., 2017) of transformer models.

We assume Language X has a fixed **2× increase** in tokens across tokenizers (although there are variations not included here). We assume English has 1 000 000 tokens (baseline) and Language X: approximately 2 000 000 tokens for equivalent content. There is a $2^2 = 4x$ increase in training cost.

In addition to cost, token inflation impacts time. Transformer models scale quadratically in sequence length. With a $2\times$ token increase, Language X requires $4\times$ more compute. This means training that takes 90 days for English would take ~ 360 days for Language X on the same hardware. For inference time, decoding scales approximately linearly with token count. A prompt completion that takes 2 seconds in English may take about 4 seconds in Language X.

These multipliers apply whether the additional tokens appear in the input (prompt) or output (completion), and they exacerbate cost disparities for low-resource languages.

Using published petaFLOP-day figures and assuming a compute cost of \$240 per petaFLOP-day, Table 3 displays order-of-magnitude estimates of compute and cost for training, while Table 4 displays order-of-magnitude estimates for inference.

C.2 Prompt

```
You must only reply with 'Final Answer: X'
where X is A, B, C, or D.
Do NOT add explanations, reasoning,
or extra text.
Question: <question text>
Choices:
A. <option 1>
B. <option 2>
C. <option 3>
D. <option 4>
Your response must be strictly formatted as:
Final Answer: X
```

D Reproducibility: Code and Data

Raw LLM outputs, tokens, fertility and parity measures for each model, and scripts to replicate our analysis are available at:

Model	English	2× Fertility	5× Fertility
Llama 2 70B	\$5M	\$20M	\$125M
Llama 3 70B	\$24M	\$96M	\$600M
Llama 3.1 405B	\$105M	\$420M	\$2.6B

Table 2: Training costs scale quadratically with fertility

Model	petaFLOP-days	English \$	Language X (\$4×)
LLaMA 2 (69B)	21 000	5 M	20 M
LLaMA 3 (70B)	100 000	24 M	96 M
LLaMA 3.1 (405B)	440 000	105 M	420 M

Table 3: Training compute and cost estimates for LLaMA models (USD).

Provider	Model (type)	English \$	Language X (~2×)
OpenAI	GPT-4o	5 / 20	10 / 40
OpenAI	o4-mini*	4 / 16	8 / 32
Google	Gemini 2.5 Flash	0.30 / 2.50	0.60 / 5.00
Google	Gemini 2.5 Pro*	1.25 / 10	2.50 / 20
Anthropic	Claude 4 Sonnet	3 / 15	6 / 30
Anthropic	Claude 4 Opus*	15 / 75	30 / 150

Table 4: Inference cost per 1M English-equivalent tokens (USD) including reasoning models. The costs are shown for input/output.

https://github.com/jessicalundin/multilingual_token_tax.

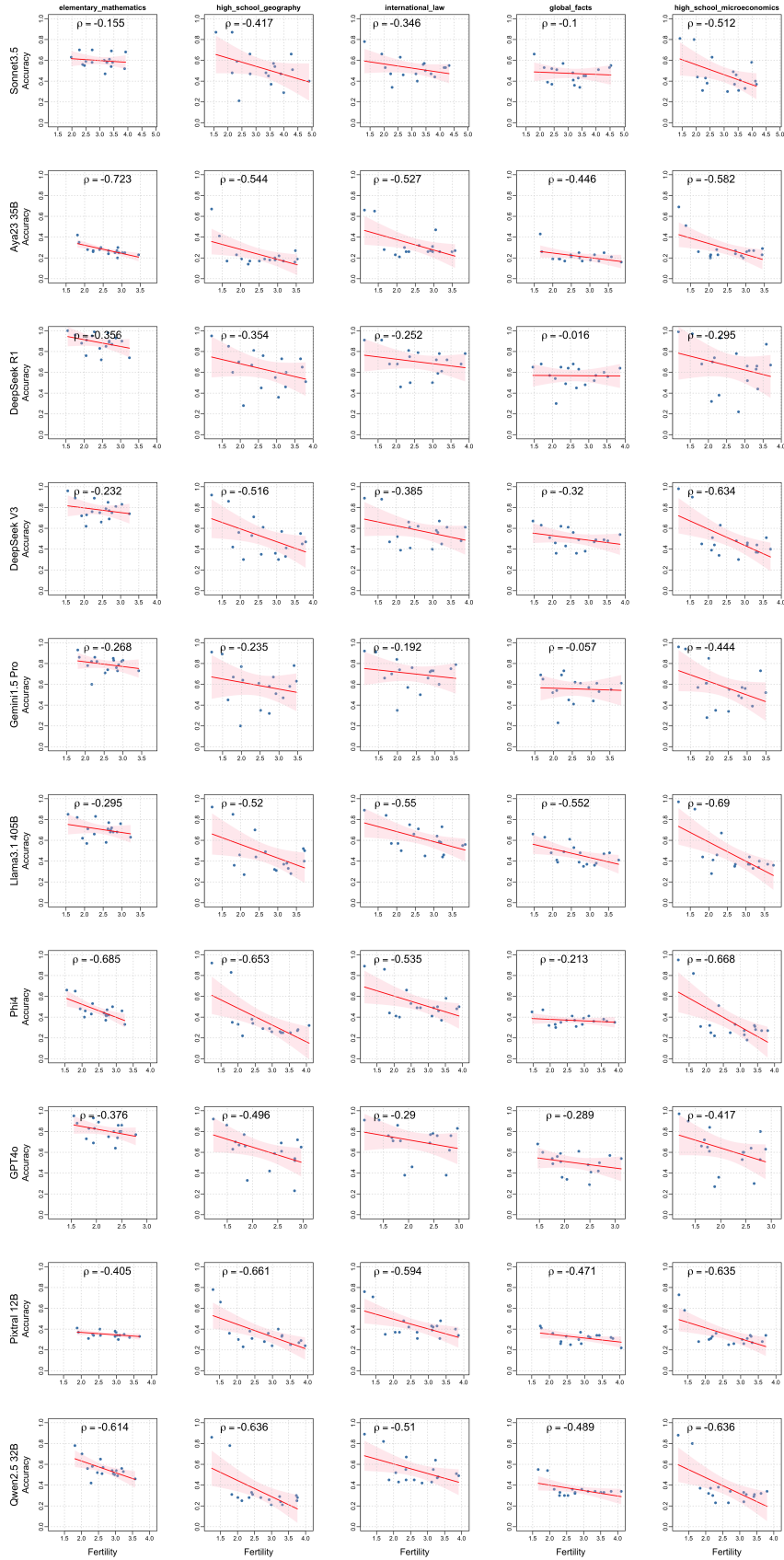


Figure 3: Fertility and accuracy trade-offs for the 10 models across five MMLU subjects. Note that these associations are correlational: fertility captures tokenization inefficiency that covaries with performance, but does not isolate causal effects independent of pretraining data availability or quality.

EduNaija AI Tutor: A Multi-Agent Retrieval-Augmented Generation System for Nigerian Curriculum Education

Israel Odeajo
WorldQuant University
isrealodeajo@gmail.com

Edifon Jimmy
University of Benin
edifonemmanuel114@gmail.com

Abstract

Equitable access to quality education remains a critical challenge in Nigeria, where millions of students prepare annually for standardized examinations (WAEC, NECO, JAMB) with limited access to personalized tutoring. This research presents **EduNaija AI Tutor**, a multi-agent Retrieval-Augmented Generation (RAG) system designed to democratize educational support aligned with Nigerian curricula. The system integrates conversational AI with document-based question answering, automated assessment generation, and multilingual support for English, Yoruba, Hausa, and Igbo. Using LangChain for agent orchestration and FAISS for vector retrieval, the system enables students to interact with educational content through natural language queries while maintaining cultural relevance. Preliminary evaluation demonstrates the system’s capability to provide curriculum-aligned explanations and generate practice assessments.

1 Introduction

Despite rising enrollment in secondary and tertiary education, access to quality educational support in Nigeria remains highly unequal. Each year, over two million candidates register for the Joint Admissions and Matriculation Board (JAMB) examination (Joint Admissions and Matriculation Board, 2023; West African Examinations Council, 2022). However, structured tutoring is disproportionately concentrated in urban centers (UNESCO, 2023).

Private tutoring typically costs between ₦5,000 and ₦50,000 per month, presenting a significant financial barrier (World Bank, 2022). While Large Language Models (LLMs) offer scalable solutions, generic models often lack strict adherence to local syllabi (Adelani et al., 2022). We introduce **EduNaija AI Tutor**, a localized systems framework engineered to solve the lack of curriculum-aligned, multilingual tutoring for Nigerian national exams.

2 Background and Related Work

2.1 AI in Education

Intelligent Tutoring Systems (ITS) have evolved from rule-based systems to LLM-powered tutors (Woolf, 2010). While tools like Khanmigo (Khan Academy, 2023) utilize LLMs, they are predominantly designed for Western curricula, leaving a gap in systems aligned with African examination structures like WAEC.

2.2 Retrieval-Augmented Generation (RAG)

LLMs often suffer from hallucination. RAG addresses this by grounding responses in authoritative documents (Lewis et al., 2020). While applied in scientific domains (Liu et al., 2025), its application to African national curricula remains underexplored.

3 Methodology

3.1 System Architecture

The system employs a hierarchical multi-agent architecture. The main orchestrator acts as the central hub, classifying user intent and routing queries to specialized agents.

3.2 Data Processing Pipeline

To ensure relevance, the system allows users to upload specific study materials. Text is extracted via PyMuPDF and split using RecursiveCharacterTextSplitter (size=1024, overlap=100). Chunks are embedded using OpenAI’s text-embedding-ada-002 and stored in FAISS.

4 Evaluation and Results

4.1 Evaluation Setup

We recruited $N = 50$ professional evaluators across three departments to assess the system (Table 1).

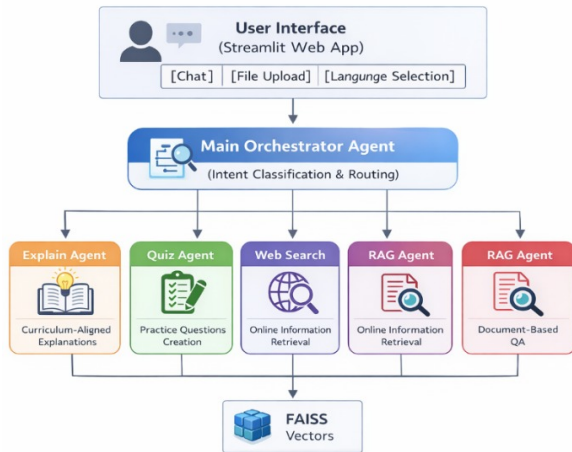


Figure 1: EduNaija Multi-Agent Architecture routing intent to specialized tools.

Department	Count	Key Subjects Evaluated
Science	20	Mathematics, Physics, Chemistry, Biology
Commercial	15	Economics, Accounting, Commerce
Art	15	Lit-in-English, Government, Christian Religious Studies
Total	50	

Table 1: Demographics of the 50 evaluators across major academic departments.

4.2 Multilingual Performance

Table 2 summarizes the aggregate scores based on language accuracy and syllabus alignment.

Criteria	Yoruba	Igbo	Hausa
Language Accuracy	8.0	7.0	7.0
Edu. Info Accuracy	8.0	6.0	8.0
Doc. Handling	7.0	6.0	8.0
Output Structure	8.0	6.0	9.0
Normalized (%)	77.5%	62.5%	80.0%

Table 2: Expert evaluation summary showing performance across languages.

5 Conclusion

EduNaija demonstrates that Multi-Agent RAG systems can effectively bridge the educational gap in Nigeria. Future work will focus on fine-tuning open-source models on native African educational corpora to reduce dependency on translation.

Limitations

A primary limitation is the reliance on translation. Deep nuances in Yoruba or Igbo may be lost during the translation to English for reasoning.

Ethical Considerations

There is a risk that students may rely solely on the AI. We explicitly label the tool as a supplement, not a replacement for schooling.

References

- David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Hassan Muhammad, Peter Nabende, and Cheikh M. B. Dione. 2022. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4488–4508.
- Joint Admissions and Matriculation Board. 2023. Annual registration statistics and examination report. Technical report, JAMB, Abuja, Nigeria.
- Khan Academy. 2023. [Khanmigo: AI-powered teaching assistant and tutor](#). Online resource.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.
- S. Liu and 1 others. 2025. Improving large language model applications in biomedicine with retrieval-augmented generation: A systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association (JAMIA)*.
- UNESCO. 2023. Global education monitoring report: Technology in education. Technical report, UNESCO, Paris, France.
- West African Examinations Council. 2022. Waec syllabus and regulations for west african senior school certificate examination. Technical report, WAEC, Lagos, Nigeria.
- Beverly Park Woolf. 2010. *Building Intelligent Interactive Tutors*. Morgan Kaufmann.
- World Bank. 2022. Nigeria poverty assessment 2022: A better future for all nigerians. Technical report, The World Bank Group.

A System Interface

The EduNaija interface is designed for accessibility across various devices. As shown in Figure 2, students can interact via a chat-based interface that supports document uploads and language selection.

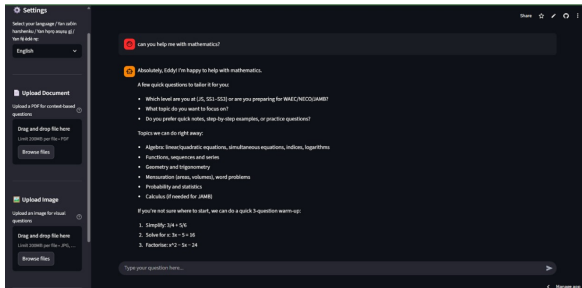


Figure 2: Main chat interface showing language selection and document upload panel.

B Reproducibility

To facilitate future research, we have made our work open-source. The complete source code and implementation notebooks are available at: <https://github.com/Eddy-Emmanuel/EduNaija-LLM>.

Synthetic Data Generation Pipeline for Low-Resource Swahili Sentiment Analysis: Multi-LLM Judging with Human Validation

Samuel Gyamfi¹, Alfred Malengo Kondoro^{1,2}, Yankı Öztürk¹
Richard H. Schreiber¹, Vadim Borisov¹

¹tabularis.ai ²Hanyang University
samuel.gyamfi@tum.de

Abstract

Despite serving over 100 million speakers as a vital African lingua franca, Swahili remains critically under-resourced for Natural Language Processing, hindering technological progress across East Africa. We present a scalable solution: a controllable synthetic data generation pipeline that produces culturally grounded Swahili text for sentiment analysis, validated through automated LLM judges. To ensure reliability, we conduct targeted human evaluation with a native Swahili speaker on a stratified sample, achieving 80.95% agreement between generated sentiment labels and human ground truth, with strong agreement on judge quality assessments. This demonstrates that LLM-based generation and quality assessment can transfer effectively to low-resource languages. We release a dataset and provide a reproducible pipeline in tandem, providing ample knowledge and working material for NLP researchers in low-resource contexts. Our dataset and the full reproducible generation pipeline are publicly available in the [Swahili Sentiment Dataset repository](#) and the [GitHub repository](#).

1 Introduction

Advances in Natural Language Processing (NLP) have not been distributed evenly across the world’s languages, leaving many “low-resource” languages without sufficient data for modern model development (Joshi et al., 2020; Nekoto et al., 2020; Hedderich et al., 2021). Swahili (Kiswahili), a major lingua franca of East and Central Africa and an official African Union language, remains one of these under-served languages despite its more than 100 million speakers (Orife et al., 2020; Ndimbo et al., 2025). Data scarcity limits the development of high-quality NLP tools for Swahili and risks widening the digital divide. The language’s agglutinative morphology and complex noun-class system further amplify the challenge, since such linguistic

structures are often underrepresented in corpora dominated by English and other high-resource languages (Gutman and Avanzati, 2013; Arnett and Bergen, 2024).

Synthetic data (sometimes artificial data (Borisov and Schreiber, 2024)) generation using foundation models, particularly Large Language Models (LLMs) (Bommasani et al., 2022), offers a scalable alternative to manual annotation (Liu et al., 2024; Nadas et al., 2025). However, producing high-quality, diverse, and culturally grounded synthetic text for low-resource languages requires careful control to ensure linguistic fidelity (Kirk et al., 2024; Li et al., 2023).

This work introduces a controllable synthetic data pipeline for Swahili that prompts LLMs to generate culturally grounded text, employs LLM-as-a-judge (Zheng et al., 2023) for multi-dimensional quality scoring, and filters aggressively to retain only high-quality samples. We validate this approach through rigorous human evaluation, achieving 80.95% agreement between generated sentiment labels and native speaker ground truth. On the AfriSenti benchmark (Muhammad et al., 2023), models fine-tuned with our synthetic data show consistent macro-F1 gains over zero-shot baselines, demonstrating that judged synthetic supervision can reliably transfer sentiment capability to a low-resource setting.

Our main contributions are as follows:

1. A controllable generation pipeline for Swahili that incorporates sentiment, domain, aspect, tone, and cultural relevance constraints.
2. An automated LLM-based judging mechanism validated against human ground truth.
3. Empirical evidence that judged synthetic data improves Swahili sentiment classification performance on AfriSenti.

2 Related Work

Prior work on low-resource languages addresses data scarcity through transfer learning, cross-lingual and semi-supervised methods, and distantly supervised labeling, often trading scalability for noise (Hedderich et al., 2021). Recent work increasingly favors data augmentation, including back-translation and LLM-based synthetic data generation, as scalable alternatives to manual annotation.

Traditional Data Augmentation. Early approaches relied on rule-based transformations such as synonym replacement and token-level perturbations (Feng et al., 2021), which are easy to implement but offer limited diversity and can introduce grammatical errors (Wei and Zou, 2019). Related work has also explored simpler augmentation techniques, such as contextual augmentation for low-resource English-Swahili MT, yielding moderate gains (Gitau and Marivate, 2023).

Sentiment Analysis in Low-Resource Settings. Sentiment analysis is a widely studied NLP task for modeling opinions expressed in text (Medhat et al., 2014; Wankhade et al., 2022), but approaches developed for resource rich languages often transfer poorly to LRLs (Joshi et al., 2020).

For Swahili, the scarcity of large, high-quality labeled corpora comparable to English benchmarks, such as IMDb or Yelp, remains a primary bottleneck (Tunga and David, 2025). This challenge is compounded by the language’s agglutinative morphology, which embeds sentiment-bearing morphemes within complex verb forms and complicates standard tokenization and feature extraction (Arnett and Bergen, 2024; Mathayo and Kondoro, 2025), as well as by cultural nuances expressed through idioms, proverbs, and context-dependent phrasing that multilingual models trained on Western-centric data often fail to capture (Muhammad et al., 2023). In response, recent efforts have introduced dedicated benchmarks such as AfriSenti (Muhammad et al., 2023) and explored synthetic data generation to alleviate labeled data scarcity (Sundarreson and Kumarapathirage, 2024), a direction this work further advances.

Generative AI Models for Synthetic Data. Earlier generative approaches such as GANs and VAEs were explored for text generation (Goodfellow et al., 2014; Kingma and Welling, 2022), but they offered limited controllability for task-specific data synthesis (Liu et al., 2024).

The emergence of large language models has

substantially advanced synthetic data generation (Nadas et al., 2025; Davidson et al., 2025; Sundarreson and Kumarapathirage, 2024), enabling fluent and contextually grounded text generation through prompting (Brown et al., 2020). Prior work leverages LLMs for synthetic data creation across tasks such as text classification, instruction following, question answering, and information extraction (Li et al., 2023; Tan et al., 2024; IR-LLM Community, 2023), using techniques including zero- and few-shot prompting, self-instruction pipelines (Wang et al., 2023), and retrieval-augmented generation (Lewis et al., 2020). Recent studies further show that fine-tuned teacher LLMs can generate large synthetic corpora for training smaller student models, a paradigm that is particularly effective in low-resource settings (Kaddour and Liu, 2024).

Synthetic Data for Low-Resource Languages. Although high-quality human-annotated datasets exist for some low-resource languages, including Swahili (Tunga and David, 2025; Zawuya et al., 2025), they are often limited in scale, making synthetic data generation a central strategy for resource expansion (Doshi and Bhattacharyya, 2024). Back-translation remains a foundational approach in low-resource neural machine translation (Li et al., 2020; Bojar et al., 2016), while more recent work explores LLMs for synthetic data generation. However, LLM performance frequently declines in LRL settings due to pre-training data imbalance (Wang et al., 2024; Qin et al., 2025), motivating adaptation strategies such as multilingual fine-tuning (Moskvoretskii et al., 2024), cross-lingual transfer (Latouche et al., 2024), and specialized prompting (Deshpande et al., 2024). Related studies also investigate machine translation-based synthetic data (translationese) for language-model pre-training in Indic languages (Doshi et al., 2024). For Swahili specifically, retrieval-augmented generation has been applied to conversational AI, demonstrating the potential of combining LLMs with external knowledge sources in this low-resource context (Ndimbo et al., 2025).

Handling Morphological Complexity. Morphologically rich, agglutinative languages such as Swahili pose well-known challenges for NLP due to dense sequences of inflectional and derivational morphemes. Standard subword tokenization methods, like byte-pair encoding (BPE), frequently misalign with morpheme boundaries of such languages, introducing sparsity and obscuring grammatical structure (Arnett and Bergen, 2024). Prior work

demonstrates that morphology-aware approaches to subword segmentation, such as deploying the Prefix-Root-Postfix-Encoding (PRPE) algorithm for machine translation (Chen and Fazio, 2021), can substantially benefit low-resource settings.

Linguistic and computational studies of Swahili highlight a highly structured, multi-slot verbal morphology that contributes to segmentation errors in neural models (Wahome et al., 2023; Mathayo and Kondoro, 2025), motivating rule-based approaches such as SwaRegex, a finite-state lexical transducer that achieves high-accuracy in verb segmentation (Muthee et al., 2022). Recent work further shows that Swahili morphology is highly productive in digital contexts, with novel derivations and reduplication patterns emerging under code-mixing and social media use (Makulilo, 2025; Gabriel et al., 2018). Although our approach does not perform explicit morphological segmentation, these findings motivate the evaluation of the morphological plausibility of generated Swahili text.

Quality Control and Evaluation. A critical aspect of synthetic data generation is assessing its quality and utility. Automated metrics like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) are widely used but often show limited correlation with human judgment (Schmidová et al., 2024; Pillutla et al., 2021). Evaluating downstream task performance remains the gold standard (Guo et al., 2024). Recent work explores using LLMs themselves as evaluators or “judges” (Zheng et al., 2023), similar to the approach adopted in our pipeline. Ensuring diversity and avoiding distribution shift are also key concerns (Nadas et al., 2025). The subjectivity of tasks can also impact the effectiveness of synthetic data (Li et al., 2023).

Our work builds upon these insights by employing a conditional LLM prompting strategy specifically designed for Swahili, incorporating culturally relevant prompts, and utilizing a multi-faceted LLM-based judging mechanism for quality control, and then evaluating the efficacy of this pipeline on sentiment classification tasks. We contribute a large-scale dataset generated using this pipeline and evaluate its quality based on automated judgments, providing insights into the effectiveness of and challenges facing this approach for Swahili.

3 Synthetic Data Generation Pipeline

In this section, we present our synthetic data generation pipeline, which consists of five steps.

3.1 Step 1: Criteria Definition

To ensure diversity without relying on seed samples, we define a comprehensive set of controllable generation parameters spanning sentiment, structure, quality, and semantic context. **Sentiment** is modeled on a 5-point scale with intermediate values (e.g., 1.5, 4.5) to capture nuance, with a higher sampling weight assigned to *neutral* to reflect natural distributions. **Target Length** ranges from *micro* (10–25 words) to *extensive* (250–300 words), acknowledging that LLMs do not always adhere strictly to word-count constraints. To mimic real-world noise, we introduce a **Desired Quality** parameter ranging from *abysmal* to *exceptional*, explicitly prompting for imperfections such as typos. Semantic variation is controlled through curated sets of **Domains**, **Aspects**, and **Tones**: domains cover over 160 categories relevant to East African digital experiences (e.g., mobile money, politics, hospitality, technology), aspects span over 180 evaluation dimensions (e.g., quality, customer service, cultural relevance), and tones include more than 50 nuanced emotional states such as *sarcastic* or *disappointed but hopeful*. This fine-grained control enables the generation of highly structured and diverse synthetic data for robust downstream fine-tuning; the full list of criteria is provided in Appendix E.

3.2 Step 2: Prompt Construction

Our pipeline utilized two distinct prompt templates: one for generating samples and another for evaluating these samples. These were provided to the generator models and the judge models, respectively, and were dynamically filled with randomized criteria during the generation process. We explicitly evaluated the use of a generator prompt written in the target language to promote closer stylistic adherence. Additional details regarding the prompt design and structure are provided in Appendix B.

3.3 Step 3: Synthetic Text Generation

To increase linguistic and stylistic diversity in the generated data, we employ two distinct generator models: **Llama 3 70B** (AI@Meta, 2024) and **Gemini Flash 2.0** (Google DeepMind). The models are used in an interleaved manner, allowing com-

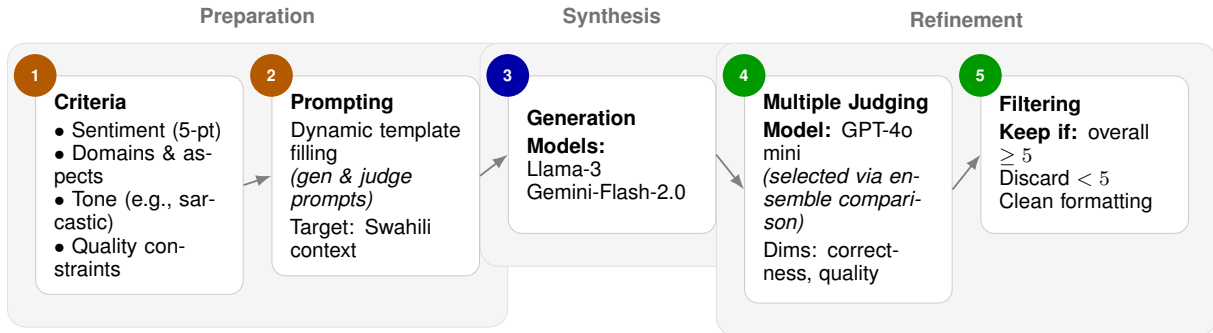


Figure 1: The proposed pipeline for robust synthetic data generation.

plementary generation behaviors to surface across domains, tones, and sentiment configurations. Both models were selected based on evidence of strong multilingual capabilities and the likelihood of exposure to substantial Swahili data during pre-training, as suggested by model scale, provider documentation, and reported multilingual performance. The pipeline itself is agnostic to the total number of generated samples and can be scaled to arbitrarily large datasets as computational resources permit.

3.4 Step 4: Automated Judging with Multi-Model Ensemble

After generating all 50,000 synthetic Swahili samples using our two generator models, we implemented a comprehensive automated evaluation strategy to assess sample quality and determine the most reliable judge for large-scale filtering.

Judge Model Selection. We evaluated three distinct LLM judges to score the generated samples: **GPT-4o mini** (OpenAI, 2024), **Claude 4.5 Haiku** (Anthropic, 2025), and **Grok 4.1 Fast** (xAI, 2025). These models were selected based on their performance on public leaderboards, cost-effectiveness for large-scale evaluation, and recent state-of-the-art releases. Each judge evaluated samples across five dimensions: Language Correctness (0–5), Cultural Relevance (0–5), Sentiment Alignment (0–5), Instruction Following (0–5), and Overall Quality (0–10).

Ensemble Evaluation. All three judges independently scored the complete set of 50,000 generated samples. We then conducted ensemble experiments, averaging scores across the three judges and comparing ensemble performance against individual judge performance. We observed strong Pearson correlations between judges, particularly between GPT-4o mini and the ensemble average.

Human Validation Study. To validate the reliability

of our automated judges, we conducted targeted human evaluation on a stratified random subset of 126 samples. A native Swahili speaker with expertise in East African linguistics reviewed each sample, providing ground-truth sentiment labels and independent quality judgments across the same five evaluation dimensions used by the automated judges. Comparing human annotations against automated judge scores revealed strong agreement across all judges, with performance metrics detailed in Appendix C.

Final Judge Selection. Based on these validation results, we selected **GPT-4o mini** (OpenAI, 2024) as the sole judge for our final filtering pipeline. This decision prioritizes cost-effectiveness and reproducibility: by demonstrating that a single, affordable model can serve as a reliable evaluator when properly validated, we provide a practical blueprint for researchers in low-resource settings who may lack budgets for expensive multi-model ensembles. GPT-4o mini’s performance was nearly identical to the full ensemble while offering significantly lower inference costs, making it the optimal choice for scaling to our complete dataset.

3.5 Step 5: Automated Filtering and Dataset Finalization

Using GPT-4o mini as the selected judge, we filtered the complete dataset of 50,000 generated samples based on their Overall Quality scores. We applied a threshold of Overall Quality ≥ 5 (out of 10), automatically removing any sample that fell below this standard. This threshold was chosen to balance dataset scale with quality assurance, retaining samples that met minimum acceptable standards while removing only egregiously poor outputs.

This filtering process retained 47,980 samples. As a final preprocessing step, we cleaned the retained samples by removing any formatting arti-

facts (such as brackets or placeholders used to guide generation), ensuring the text was production-ready for downstream training.

4 Analysis of Synthetic Data

4.1 Intrinsic Quality and Generation Analysis

From an initial set of 50,000 generated samples, 47,980 were retained after filtering for an Overall Quality score ≥ 5 (out of 10), as assigned by the GPT-4o mini judge.



Figure 2: Pearson correlation heatmap between the five judgment metrics: Language Correctness (LC), Cultural Relevance (CR), Sentiment Alignment (SA), Instruction Following (IF), Overall Quality (OQ).

Judge Score Distributions: The selected judge, GPT-4o mini, assigned scores across five dimensions. The distribution of these scores for the initial 50,000 samples is shown in Table 1. The heatmap in Figure 2 indicates positive correlations between all judgment metrics, with Overall Quality strongly correlating with Instruction Following ($r=0.85$) and moderately with Language Correctness ($r=0.69$) and Sentiment Alignment ($r=0.68$).

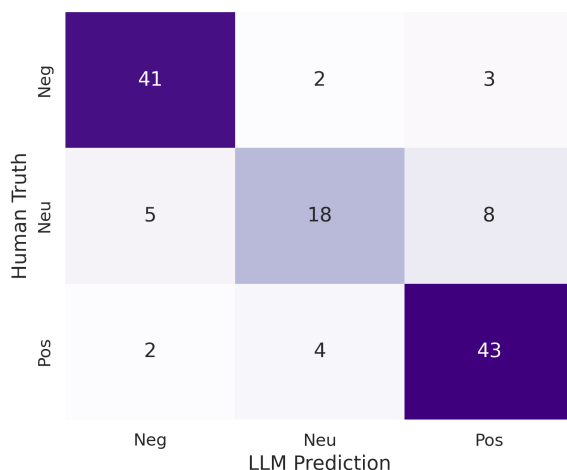
Generator Comparison: Gemini Flash 2.0 generally outperformed **Llama3-70B** in judged quality, receiving higher average scores across all dimensions (Table 1). Notably, **Gemini Flash 2.0** achieved an average Overall Quality of 7.81 compared to 6.55 for **Llama3-70B**. These differences are consistent across all evaluation dimensions.

4.2 Human Verification and Morphological Analysis

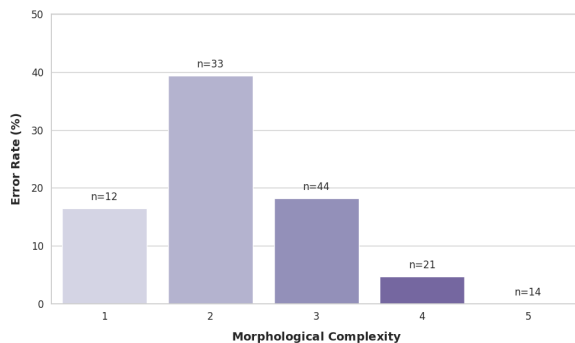
To validate the reliability of our automated generation and judging pipeline, we conducted targeted

human evaluation on a stratified random subset of 126 generated samples (described in Section 3.4). A native Swahili speaker with expertise in East African linguistics reviewed each sample, providing ground-truth sentiment labels and independent quality judgments.

Generator Sentiment Accuracy: As illustrated in Figure 3a, we observed **80.95% accuracy** (Cohen’s Kappa of 0.706) between the sentiment labels assigned by our generator LLMs (Llama 3 70B and Gemini Flash 2.0) and human ground-truth sentiment. The confusion matrix shows that the vast majority of samples were correctly tagged, with only 2 false negative instances and 3 false positive instances, confirming that our generators reliably produce sentiment-aligned text.



(a) Confusion Matrix: Human vs. LLM sentiment tagging



(b) Error Rate by Morphological Complexity

Figure 3: Human validation results showing strong sentiment agreement and the impact of morphological complexity on error rates.

Morphological Complexity Analysis: Swahili is an agglutinative language in which sentiment cues are often encoded within complex verbal morphology (e.g., *si-ta-ku-pend-a*; NEG-FUT-2SG-love-FV, "I will not love you"). To examine whether

Generator Model	Lang. Correct.	Cult. Relevance	Sent. Align.	Instr. Follow.	Overall Qual.
Gemini Flash 2.0	4.17 ± 0.49	4.55 ± 0.67	4.19 ± 1.30	4.02 ± 0.81	7.81 ± 1.37
Llama3-70B	3.81 ± 0.53	3.45 ± 0.93	3.70 ± 1.39	3.25 ± 0.80	6.55 ± 1.26

Table 1: Average judged scores and their standard deviation by generator model (Pre-filtering, N=50,000). The Gemini model show superior performance on all criteria.

morphological complexity influences error patterns, each sample was assigned a *Morphological Complexity Score* (1–5) based on the number of realized suffix slots within verbs (subject agreement, tense/aspect, negation, object marking, verbal extensions). Scores were assigned through manual inspection by the native speaker following consistent slot-based criteria.

As shown in Figure 3b, the highest error rate (approximately 39%) occurred at moderate complexity levels (score 2), while samples with highly complex morphology (score 5) showed no sentiment errors in this subset. Qualitative analysis indicates that misclassifications arise less from morphological form itself and more from pragmatic factors such as hedging, politeness markers, and evaluative ambiguity. Highly inflected but structurally canonical verb forms align more closely with training data patterns, yielding more stable sentiment judgments. These results suggest that conversational-level morphology combined with pragmatic nuance poses greater challenges than morphologically dense but structurally regular constructions, while confirming the pipeline’s robustness to Swahili’s morphological richness.

5 Experimental Validation of Synthetic Data Pipeline

In addition to evaluating three pre-trained multilingual sentiment models¹, we trained and evaluated a simple DistilBERT base model² (Sanh et al., 2019) on our generated data. Their performance was assessed on the Swahili test set (‘swa’) of the AfriSenti benchmark (Muhammad et al., 2023), comparing their zero-shot performance against performance after fine-tuning solely on our generated data. The metric of choice was macro F1 as a way to better handle class imbalances, as well as providing a fair assessment of performance. Fine-

¹tabularisai/multilingual-sentiment-analysis (tabularisai et al., 2025), lxyuan/distilbert-base-multilingual-cased-sentiments-student, and nlptown/bert-base-multilingual-uncased-sentiment

²distilbert/distilbert-base-cased

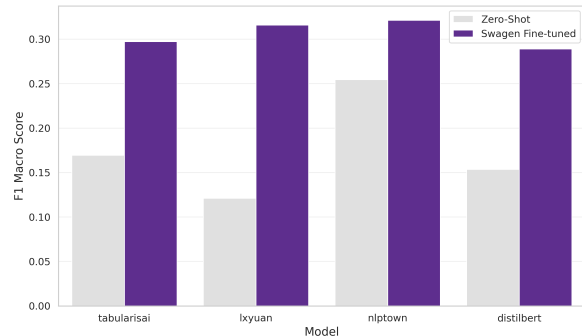


Figure 4: Macro F1 scores on AfriSenti-Swahili: Zero-Shot vs. Fine-tuned on Synthetic Data.

tuning used the following hyperparameters: learning rate of 5×10^{-5} , 5 epochs, AdamW optimizer (Loshchilov and Hutter, 2019) with the HuggingFace Transformers library (Wolf et al., 2020), the rest of the hyperparameters are default to the model or selected package.

5.1 Results

Fine-tuning on our synthetic dataset substantially improved the performance of all three multilingual sentiment classification models, as well as the trained distilbert-base-uncased model on the AfriSenti-Swahili test set compared to their zero-shot capabilities (Figure 4, Table 2). All models benefited from fine-tuning with our synthetic data. Even smaller models like our Tabularis multilingual-sentiment-analysis model and the distilbert base model singularly trained for swahili sentiment analysis show competitive performance with much larger models. Models like the nlptown/bert-base-multilingual-uncased-sentiment trained on a massive corpus of data still benefited from training.

6 Discussion and Future Work

Our synthetic data generation pipeline demonstrates a viable strategy for augmenting data for Swahili. The notable improvement in sentiment analysis performance post-fine-tuning on this LLM-generated data underscores its value for adapting existing multilingual models. According to auto-

Model	Zero-Shot Macro F1	Synthetic Data Fine-tuned Macro F1
tabularisai	0.1695	0.2972
lxyuan	0.1211	0.3158
nlptown	0.2545	0.3213
swahili distilbert base	0.1536	0.2891

Table 2: Sentiment Model Performance (Macro F1) on AfriSenti-Swahili Test Set

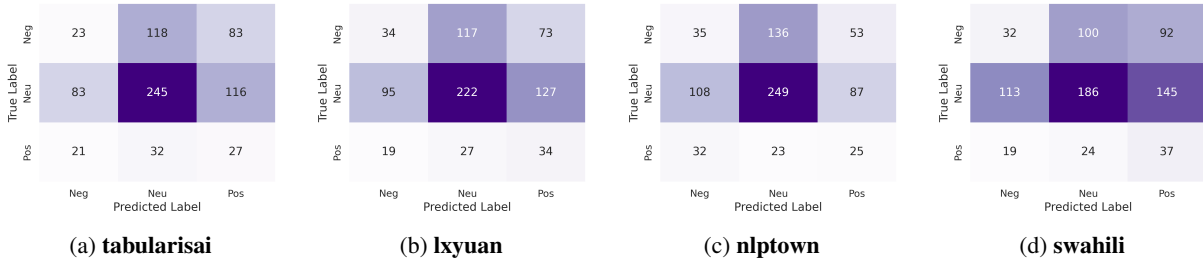


Figure 5: Confusion matrices for models fine-tuned with synthetic data, evaluated on the AfriSenti-Swahili test set.

mated judgments, **Gemini Flash 2.0** was a more effective generator than **Llama3-70B** for this task.

A recurring failure mode in low-resource language data augmentation is the proliferation of “translationese,” text that is grammatically correct but culturally hollow. By explicitly prompting for and judging *Cultural Relevance*, such as the use of *methali* (proverbs) and local entities, our pipeline moves beyond mere translation. The high agreement between the human evaluator and the LLM judge confirms that modern models, when prompted correctly, can distinguish between generic translation and culturally resonant text. The performance gains on AfriSenti, a benchmark known for difficult and colloquial samples, suggest that cultural grounding is a critical factor in transfer learning success for African languages.

Limitations include the inherent differences between synthetic and real-world data, contradictions in generated samples (sometimes asking for contradictory tones and sentiments as a result of randomized criteria insertion), and the intrinsic challenges of the AfriSenti benchmark. Additionally, while we validated our automated judge against a human baseline, we utilized the raw judgment scores for filtering. As noted by recent statistical frameworks for LLM evaluations (Lee et al., 2025), naive estimates from imperfect judges can introduce bias, potentially overestimating performance at low accuracy levels or underestimating it at high levels. Future iterations of this pipeline would benefit from applying bias-adjusted estimators and constructing confidence intervals that explicitly account for the judge’s sensitivity and specificity.

The meaning of the term “low resource language” is not unanimously agreed upon among researchers of language technologies (Cieri et al., 2016). Differences in available information and the goals of such research inevitably lead to some obscurity in what can be understood from the term: *low density*, *less commonly taught*, *under-resourced*, *less computerized*, and *less privileged* are all among possible connotations (Singh, 2008). We interpret a low-resource language as one that is unable to directly benefit from state-of-the-art statistical methods directly due to its scarcity of structured data.

Although an understudied language in NLP tasks, Swahili enjoys the privileges of official language status in several countries, a written literary tradition, and well documented dialects (Miachina, 1981). We recognize that the methods described in this paper presume state-of-the-art LLMs’ ability to readily generate text in the target language and might not be reproducible for the vast majority of the world’s languages.

Future work should prioritize expanded human evaluation of synthetic data, specifically addressing semantic cohesion. Our native speaker expert noted that instances with generated proverbs, while linguistically accurate, were incongruent with the context in which they appeared. This misalignment likely arises from the stochastic combination of disparate generation criteria (e.g., requesting a proverb within a technical error report). Further investigation is warranted to refine prompt constraints, ensuring that cultural markers are applied only where they are contextually appropriate.

7 Conclusion

This work demonstrates that carefully controlled synthetic data generation, when paired with automated LLM-based quality assessment and targeted human verification, provides a viable and scalable pathway for advancing sentiment analysis in morphologically complex, low-resource languages such as Swahili. By integrating fine-grained generation constraints with multi-dimensional judging and aggressive filtering, our pipeline produces linguistically coherent and culturally grounded synthetic text that transfers effectively to downstream sentiment classification tasks. Empirical validation shows strong agreement between automated judges and native-speaker annotations, supporting the reliability of LLM-as-a-judge approaches beyond high-resource settings. Fine-tuning multilingual models on the resulting dataset yields consistent macro F1 improvements over zero-shot baselines on the AfriSenti-Swahili benchmark, confirming that high-quality synthetic supervision can meaningfully bridge labeled data gaps. Beyond Swahili, our findings suggest that combining controllable generation, automated evaluation, and selective human oversight offers a reproducible blueprint for building NLP resources in under-resourced languages where manual annotation is costly or infeasible.

References

- AI@Meta. 2024. [Llama 3 model card](#).
- Anthropic. 2025. Claude haiku 4.5. <https://www.anthropic.com/news/claude-haiku-4-5>. Released: 2025-10-15.
- Catherine Arnett and Benjamin K. Bergen. 2024. [Why do language models perform worse for morphologically complex languages?](#)
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Aurélie N. Névéol, Mariana Neves, Martin Popel, Matt Post, Raphael Rubino, Carolina Scarton, Lucia Specia, Marco Turchi, and 2 others. 2016. [Findings of the 2016 conference on machine translation \(WMT16\)](#). In *Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 131–198, Berlin, Germany. Association for Computational Linguistics.
- Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 95 others. 2022. [On the opportunities and risks of foundation models](#). *Preprint*, arXiv:2108.07258.
- Vadim Borisov and Richard H Schreiber. 2024. Open artificial knowledge. *arXiv preprint arXiv:2407.14371*.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- William Chen and Brett Fazio. 2021. [Morphologically-guided segmentation for translation of agglutinative low-resource languages](#). In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 20–31, Virtual. Association for Machine Translation in the Americas.
- Christopher Cieri, Mike Maxwell, Stephanie Strassel, and Jennifer Tracey. 2016. [Selection criteria for low resource language programs](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 4543–4549, Portorož, Slovenia. European Language Resources Association (ELRA).
- Tim Davidson, Benoit Seguin, Enrico Bacis, Cesar Magalhaes, and Hamza Harkous. 2025. [Orchestrating synthetic data with reasoning](#). In *SynthData @ ICLR2025*.
- Tejas Deshpande, Nidhi Kowtal, and Raviraj Joshi. 2024. [Chain-of-translation prompting \(cotr\): A novel prompting technique for low resource languages](#). *Preprint*, arXiv:2409.04512.
- Meet Doshi and Pushpak Bhattacharyya. 2024. [Synthetic data for multilingual nlp: A survey](#). CFILT, IIT Bombay.
- Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. [Pretraining language models using translationese](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5843–5862, Miami, Florida, USA. Association for Computational Linguistics.

- Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Edward Hovy. 2021. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*.
- B. Gabriel, Bisamaza Emilien, and Ndayishimiye Jean Léonard. 2018. Morphological doubling theory to two bantu languages reduplication: A comparative perspective of kinyarwanda and swahili. *International Journal of English and Literature*, 3:31–40.
- Catherine Gitau and Vukosi Marivate. 2023. Textual augmentation techniques applied to low resource machine translation: Case of swahili.
- Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. *Preprint*, arXiv:1406.2661.
- Google DeepMind. Gemini flash. <https://deepmind.google/models/gemini/flash/>.
- Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. APPLS: Evaluating evaluation metrics for plain language summarization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9194–9211, Miami, Florida, USA. Association for Computational Linguistics.
- Alejandro Gutman and Beatriz Avanzati. 2013. Swahili. The Language Gulper.
- Michael A. Hedderich, Lukas Lange, Heike Adel, Jan-nik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.
- IR-LLM Community. 2023. Awesome information retrieval in the age of large language model. GitHub repository.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Jean Kaddour and Qi Liu. 2024. Synthetic data generation in low-resource settings via fine-tuning of large language models. *Preprint*, arXiv:2310.01119.
- Diederik P Kingma and Max Welling. 2022. Auto-encoding variational bayes. *Preprint*, arXiv:1312.6114.
- Hannah Rose Kirk, Jatinder Singh, and Bertie Vidgen. 2024. Transparency in the wild: Navigating transparency in a deployed ai system to broaden need-finding approaches. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*, Rio de Janeiro, Brazil. ACM.
- Gaetan Lopez Latouche, Marc-André Carbonneau, and Benjamin Swanson. 2024. Zero-shot cross-lingual transfer for synthetic data generation in grammatical error detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3002–3016, Miami, Florida, USA. Association for Computational Linguistics.
- Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy yong Sohn, and Kangwook Lee. 2025. How to correctly report llm-as-a-judge evaluations. *Preprint*, arXiv:2511.21140.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.
- Hongliang Li, Yang Gao, Jingbo Zhao, and Philip S. Yu Zhou. 2020. Revisiting back-translation for low-resource machine translation between chinese and vietnamese. *IEEE Access*, 8:115032–115041.
- Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. 2024. Best practices and lessons learned on synthetic data. *Preprint*, arXiv:2404.07503.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- Prisca Boniphace Makulilo. 2025. Morphological productivity and lexical innovation in swahili: Digital communication and language transformation in social media spaces. *Language, Technology, and Social Media*.
- Irene Masiringi Mathayo and Alfred Malengo Kondoro. 2025. Unveiling swahili verb conjugations: A comprehensive dataset for low-resource nlp. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval, NLPPIR '24*, page 149–156, New York, NY, USA. Association for Computing Machinery.

- Walaa Medhat, Ahmed Hassan, and Hoda Korashy. 2014. [Sentiment analysis algorithms and applications: A survey](#). *Ain Shams Engineering Journal*, 5(4):1093–1113.
- E.N. Miachina. 1981. The swahili language: a descriptive grammar.
- Viktor Moskvoretskii, Nazarii Tupitsa, Chris Biemann, Samuel Horváth, Eduard Gorbunov, and Irina Nikishina. 2024. [Low-resource machine translation through the lens of personalized federated learning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8806–8825, Miami, Florida, USA. Association for Computational Linguistics.
- Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Seid Muhie Yimam, David Ifeoluwa Adelan, Ibrahim Sa'id Ahmad, Nedjma Ousidhoum, Abinew Ali Ayele, Saif M. Mohammad, Meriem Beloucif, and Sebastian Ruder. 2023. [AfriSenti: A Twitter sentiment analysis benchmark for african languages](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13968–13981, Singapore. Association for Computational Linguistics.
- Mutwiri George Muthee, Mutua Makau, and Omamo Amos. 2022. [Swaregex: a lexical transducer for the morphological segmentation of swahili verbs](#). *African Journal of Science, Technology and Social Sciences*.
- Mihai Nadas, Laura Diosan, and Andreea Tomescu. 2025. [Synthetic data generation using large language models: Advances in text and code](#).
- Edmund V. Ndimbo, Qin Luo, Gimo C. Fernando, Xu Yang, and Bang Wang. 2025. [Leveraging retrieval-augmented generation for swahili language conversation systems](#). *Applied Sciences*, 15(2):524.
- Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohunge, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Seling, Lawrence Okegbemi, and Laura Martinus. 2020. [Participatory research for low-resourced machine translation: A case study in african languages](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.
- OpenAI. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Iroko Fred O. 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, and 7 others. 2020. [MASAKHANE – machine translation for africa](#). In *Proceedings of the AfricaNLP Workshop 2020 co-located with ICLR 2020*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. [MAUVE: Measuring the gap between neural text and human text using divergence frontiers](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 25154–25167. Curran Associates, Inc.
- Libo Qin, Qiguang Chen, Yuhang Zhou, Zhi Chen, Yinghui Li, Lizi Liao, Min Li, Wanxiang Che, and Philip S. Yu. 2025. [A survey of multilingual large language models](#). *Patterns (N Y)*, 6(1):101118.
- Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *NeurIPS EMC² Workshop*.
- Patrícia Schmidtová, Saad Mahamood, Simone Balloccu, Ondřej Dušek, Albert Gatt, Dimitra Gkatzia, David M Howcroft, Ondřej Plátek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. *arXiv preprint arXiv:2408.09169*.
- Anil Kumar Singh. 2008. [Natural language processing for less privileged languages: Where do we come from? where are we going?](#) In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.
- Pushpika Sundarreson and Sapna Kumarapathirage. 2024. [Sentigen: Synthetic data generator for sentiment analysis](#). *Journal of Computing Theories and Applications*, 1(4):461–477.
- tabularisai, Samuel Gyamfi, Vadim Borisov, and Richard H. Schreiber. 2025. [multilingual-sentiment-analysis \(revision 69afb83\)](#). <https://huggingface.co/tabularisai/multilingual-sentiment-analysis>.
- Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. [Large language models for data annotation and synthesis: A survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957, Miami, Florida, USA. Association for Computational Linguistics.

Mahadia Tunga and Davis David. 2025. [Introducing a swahili social media sentiment analysis dataset for the telecom industry](#). *Language Resources and Evaluation*.

Maina Wahome, Agus Subiyanto, and Oktiva Herry Chandra. 2023. [An analysis of swahili verbal inflection and derivational morphemes: An item and arrangement approach](#). *Journal of Languages, Linguistics and Literary Studies*.

Daniel Wang, Tim Bakkenes, and Anton Johansson. 2024. [Fine tuning methods for low-resource language](#). OpenReview Submission.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. [Self-instruct: Aligning language models with self-generated instructions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni. 2022. [A survey on sentiment analysis methods, applications, and challenges](#). *Artificial Intelligence Review*, 55(7):5731–5780.

Jason Wei and Kai Zou. 2019. [EDA: Easy data augmentation techniques for boosting performance on text classification tasks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. [Transformers: State-of-the-art natural language processing](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

xAI. 2025. Grok 4.1 fast. <https://x.ai/blog/grok-4-1-fast>. Released: 2025-11-19.

Chaddy Anthony Zawuya, Alfred Malengo Kondoro, Diana Rwegasira, and Juma H. Lungo. 2025. [Maneno yetu: Dynamic corpus construction and pretraining for swahili nlp](#). In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM '25*, page 6564–6569, New York, NY, USA. Association for Computing Machinery.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhonghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang,

Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). Preprint, arXiv:2306.05685.

A Judge Model Selection Rationale

The selection of the "Judge" models for the initial ensemble experiments (described in Section 3.4) was based on a combination of performance capability, cost, and availability at the time of the study.

1. **GPT-4o mini**: This model was selected primarily due to its availability, speed, and cost-effectiveness. As a model from OpenAI (OpenAI, 2024), we hypothesized it had been trained on a massive and diverse corpus of internet data, likely including a non-trivial amount of Swahili text, making it a strong candidate for linguistic validation. Its performance in our initial pilot showed high correlation with larger, more expensive models, making it the ideal choice for scaling the judging process to the full dataset.
2. **Grok 4.1 Fast**: This model (xAI, 2025) was included in our initial ensemble based on its high ranking on OpenRouter and competitive pricing. We aimed to diversify the "judge panel" by including a model from xAI to mitigate potential biases inherent in using only one provider's model architecture.
3. **Claude 4.5 Haiku**: Chosen as a representative of the Anthropic model family (Anthropic, 2025), Claude 4.5 Haiku was selected due to its recent release as a state-of-the-art (SOTA) lightweight model. It offered a compelling balance of high-quality instruction following and lower inference costs compared to the larger Opus or Sonnet models.

B Prompting

This section details the prompt templates used for generating Swahili text with specific characteristics and for subsequently evaluating the generated text. These templates were designed to guide the language models effectively.

B.1 Generation Prompt Template

The following template was used to instruct the generator language model (e.g., Llama3-70B, Gemini-Flash) to produce Swahili text. The placeholders (e.g., `criteria['sentiment']`) are dynamically filled based on the specific criteria for each desired sample.

Swahili Text Generation Prompt Template

Tafadhali tengeneza maandishi ya Kiswahili yanayokidhi vigezo vifuatavyo kwa umakini mkubwa:

1. **Mwelekeo wa Hisia (Sentiment):** `{criteria['sentiment']}`
2. **Kikoa/Mada (Domain):** `{criteria['domain']}`
3. **Kipengele Maalum (Aspect):** `{criteria['aspect']}`
4. **Mtindo wa Lugha (Tone):** `{criteria['tone']}`
5. **Urefu Unaolengwa (Target Length):** `{criteria['target_length']}`
6. **Uhora Unaotarajiwa (Desired Quality):** `{criteria['desired_quality']}`

Maagizo Muhimu Zaidi:

- **Utamaduni wa Afrika Mashariki:** Jumuisha methali, misemo, au marejeleo yanayofahamika katika utamaduni wa Kiswahili/Afrika Mashariki inapowezekana na kwa njia ya asili. Lenga muktadha wa kitamaduni unaoeleweka.
- **Lugha Bora:** Tumia Kiswahili sanifu na sahihi kisarufi. Zingatia matumizi ya visawe na msamiati unaofaa.
- **Upekee wa Kienyeji:** Ongeza maelezo au vidokezo vinavyoonyesha uelewa wa mazingira ya Afrika Mashariki (k.m., majina ya maeneo, bidhaa za kawaida, hali za kijamii).
- **Fuata Maagizo:** Hakikisha maandishi yanayotokana yanaakisi kwa uaminifu vigezo vyote vilivyotolewa (hisia, kikoa, kipengele, mtindo, urefu, ubora).

Muundo wa Majibu: Tafadhali toa maandishi yaliyotengenezwa PEKEE, bila maelezo ya ziada, utangulizi, au hitimisho lako mwenyewe. Weka maandishi yote ndani ya mabano ya mraba '[]'.

Mfano wa muundo unaotarajiwa:

[<Weka maandishi yako ya Kiswahili yaliyotengenezwa hapa>]

B.2 Judging Prompt Template

To evaluate the generated Swahili text, the following prompt template was provided to the judging model. The `{criteria_str}` placeholder is replaced with a JSON string detailing the original generation criteria, `{text}` is replaced with the Swahili text to be evaluated, and `{criteria['sentiment']}` is replaced with the target sentiment for specific alignment checking.

Swahili Text Evaluation Prompt Template

You are an expert evaluator of Swahili text, focusing on linguistic quality, cultural relevance, and adherence to instructions. Evaluate the following Swahili text based on the provided criteria.

Criteria: {criteria_str}

Swahili Text to Evaluate: {text}

Evaluation Tasks: Please provide scores based on the following dimensions:

- **Language_Correctness:** (Scale 0-5) How grammatically correct, fluent, and natural is the Swahili used? (0=Very Poor, 5=Excellent)
- **Cultural_Relevance:** (Scale 0-5) How well does the text incorporate East African/Swahili cultural context, proverbs, or nuances appropriately? (0=Not relevant/Inappropriate, 5=Highly relevant and natural)
- **Sentiment_Alignment:** (Scale 0-5) How accurately does the text reflect the target sentiment ('{criteria['sentiment']}')? (0=Completely misaligned, 5=Perfectly aligned)
- **Instruction_Following:** (Scale 0-5) How well does the text adhere to all other specified criteria (Domain, Aspect, Tone, Length, Quality)? (0=Ignores most criteria, 5=Follows all criteria closely)
- **Overall_Quality:** (Scale 0-10) Considering all aspects, what is the overall quality of this generated text as a representative Swahili sample? (0=Very low, 10=Outstanding)

Output Format: Provide ONLY a JSON object containing the scores. Do not include any explanations or surrounding text.

Example Output:

```
{
  "Language_Correctness": 4,
  "Cultural_Relevance": 3,
  "Sentiment_Alignment": 5,
  "Instruction_Following": 4,
  "Overall_Quality": 8
}
```

Your JSON Output:

C Ensemble Experiments

In this section, we present the quantitative results of our inter-judge ensemble experiments. These visualizations illustrate the level of agreement between the selected LLM judges and their individual vs. collective performance when compared to human-annotated ground truth.

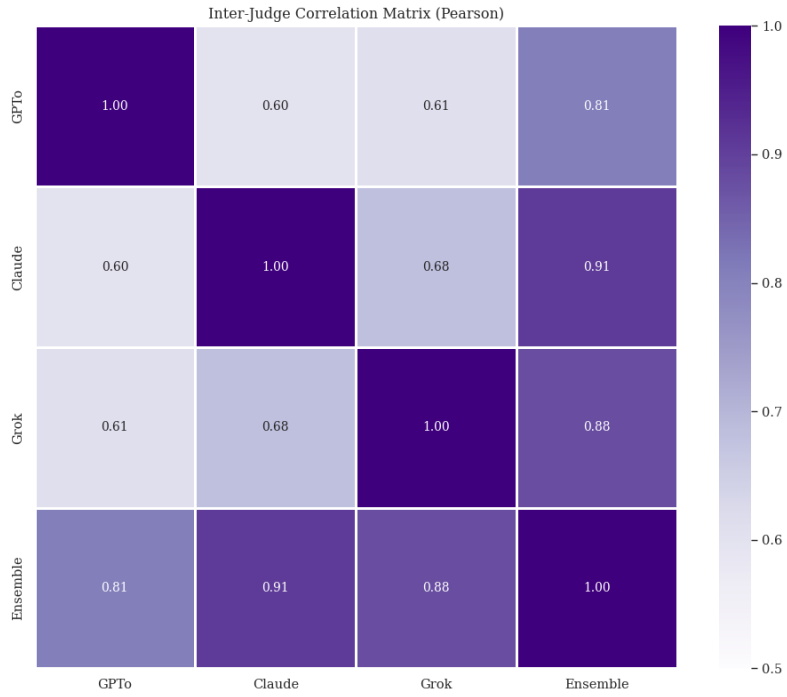


Figure 6: Correlation Heatmap of Judgment Scores: This matrix shows the Pearson correlation between the different evaluation dimensions across the judge ensemble.

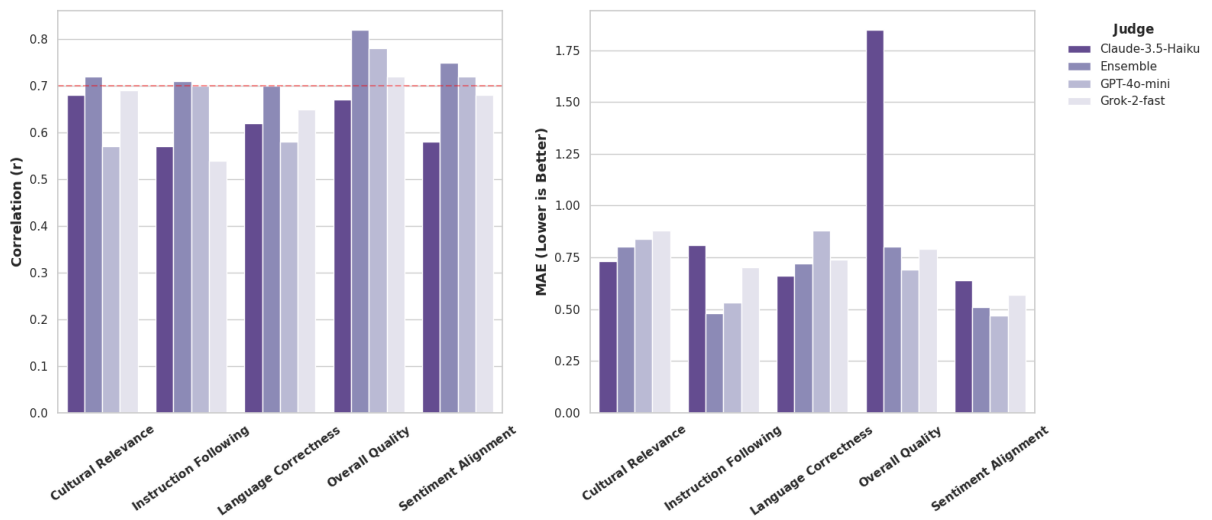


Figure 7: Judge Performance Metrics: Comparison of individual judge models and the final ensemble against human ground truth for Pearson Correlation (left) and Mean Absolute Error (right).

D Generated Samples

This section presents illustrative examples of synthetic data generated during our experiments. We showcase samples that received low overall quality scores, highlighting common failure modes, followed by samples that were judged to be of high quality, demonstrating successful generation according to the specified criteria.

D.1 Examples of Low-Quality Samples (Overall_Quality < 5)

The following samples were rated with an Overall_Quality score below 5, indicating significant deviations from the desired output or failures in adhering to instructions.

Sample ID: swahili_13932

- **Generator Model:** Llama3-70B
 - **Overall Quality:** 2.0
 - **Text:**

Tafadhali niwaonyeshe ukarimu wao katika kituo cha fitness cha [Jina la Kituo
 - **Criteria:**
 - sentiment: '3 - Neutral'
 - domain: 'Gym or Fitness Center Review'
 - aspect: 'Security Protocols'
 - tone: 'enthusiastic'
 - target_length: 'medium-long (125-175 words)'
 - desired_quality: 'Exceptional'
 - language: 'Swahili'
 - **Judgment:**
 - Language_Correctness: 2.0
 - Cultural_Relevance: 1.0
 - Sentiment_Alignment: 2.0
 - Instruction_Following: 1.0
 - Overall_Quality: 2.0
-

Sample ID: swahili_17332

- **Generator Model:** Gemini-Flash
- **Overall Quality:** 3.0
- **Text:**

Huyu mkufunzi binafsi ni janga! Anadai anajua sheria, lakini ni uongo mtupu! Kwani yeye ni nani, mjanja asiyejua lolote? "Mganga hajigangi," lakini huyu anatuuzia sumu! Anatuambia anafuata kanuni, lakini hati zake ni kama "kitumbua kimeingia mchanga" - hakuna kinachoeleweka. Anatuibia pesa zetu kwa visingizio vya kijinga, halafu anatuambia "subira yavuta heri." Heri gani, wakati tunazidi kuumia na upuuzi wake? Hata leseni hana! Anafanya kazi kinyume cha sheria kama "panya akimbia bila shati." Nataka pesa zangu zirudishwe! Hana adabu wala heshima! Atajuta!
- **Criteria:**
 - sentiment: '5 - Extremely Positive'
 - domain: 'Personal Trainer Evaluation'
 - aspect: 'Legal Compliance'
 - tone: 'angry'
 - target_length: 'medium (100-150 words)'
 - desired_quality: 'Very Poor'
 - language: 'Swahili'
- **Judgment:**
 - Language_Correctness: 4.0

- Cultural_Relevance: 5.0
 - Sentiment_Alignment: 0.0
 - Instruction_Following: 2.0
 - Overall_Quality: 3.0
-

Sample ID: swahili_7573

- **Generator Model:** Llama3-70B
- **Overall Quality:** 2.0
- **Text:**

Nilikuwa na matumaini makubwa juu ya huduma ya bima ya [jina la kampuni

- **Criteria:**
 - sentiment: '2 - Negative'
 - domain: 'Insurance Service Feedback'
 - aspect: 'Overall Experience'
 - tone: 'surprised'
 - target_length: 'extensive (250-300 words)'
 - desired_quality: 'Abysmal'
 - language: 'Swahili'
 - **Judgment:**
 - Language_Correctness: 2.0
 - Cultural_Relevance: 1.0
 - Sentiment_Alignment: 2.0
 - Instruction_Following: 1.0
 - Overall_Quality: 2.0
-

Sample ID: swahili_10177

- **Generator Model:** Gemini-Flash
- **Overall Quality:** 3.0
- **Text:**

Hebu niseme ukweli, sijawahi kukutana na mfumo mbovu kama huu. Wanaita "kitega uchumi," lakini kwangu mimi, ni shimo la kutupia pesa. Lakini, tuseme, nilikuwa najaribu kutafuta hisa za kampuni ya sukari ya Mumias, nikidhani labda kuna mtu bado anaamini katika muujiza. Kazi ya utafutaji? Usiniambie! Ni kama kutafuta sindano kwenye fungu la nyasi, au kama vile wasemavyo, "Kila ndege huruka kwa bawa lake," lakini hapa, mbawa zimevunjika.

Nimejaribu majina mengi, herufi, hata alama za uandishi, lakini hakuna chochote! Inarudi tupu. Kweli, kama vile msemu unavyosema, "Mficha ugonjwa, kifo humuumbua." Hivyo ndivyo mfumo huu ulivyo: unajaribu kuficha udhaifu wake, lakini udhaifu wenyewe unajitokeza waziwazi. Sijui kama mtu yeyote anafanya kazi hapa, lakini kama wanafanya, wanalipwa bure. Bora niendeleo na biashara yangu ya kuuza vitumbua, angalau huko najua ninachokifanya.

- **Criteria:**
 - sentiment: '5 - Extremely Positive'
 - domain: 'Investment Platform Review'
 - aspect: 'Search Functionality'
 - tone: 'pessimistic'
 - target_length: 'medium-long (125-175 words)'
 - desired_quality: 'Very Poor'
 - language: 'Swahili'
- **Judgment:**

- Language_Correctness: 4.0
 - Cultural_Relevance: 5.0
 - Sentiment_Alignment: 0.0
 - Instruction_Following: 2.0
 - Overall_Quality: 3.0
-

Sample ID: swahili_26557

- **Generator Model:** Gemini-Flash
- **Overall Quality:** 3.0
- **Text:**

Hii simu janja inanikera! Wanadai ina uwezo wa uchanganuzi wa kutabiri? Uongo mtupu! Ni kama kumpa nyani shilingi! Nimepoteza pesa zangu bure! Wanasema inatabiri hali ya hewa? Hata mzee wa kienyeji anajua vizuri zaidi! Hawana haya! "Mganga hajigangi!" Nimeghadhabika sana! Watu wanadanganywa kila siku. Hii ni wizi mtupu! Nataka pesa zangu zirudishwe! Hii ni hasira tupu!

- **Criteria:**
 - sentiment: '4.5 - Very Positive'
 - domain: 'Tech Gadget Evaluation'
 - aspect: 'Predictive Analytics'
 - tone: 'angry'
 - target_length: 'short (50-100 words)'
 - desired_quality: 'Exceptional'
 - language: 'Swahili'
- **Judgment:**
 - Language_Correctness: 4.0
 - Cultural_Relevance: 5.0
 - Sentiment_Alignment: 0.0
 - Instruction_Following: 2.0
 - Overall_Quality: 3.0

D.2 Examples of High-Quality Samples (Overall_Quality \geq 5)

In contrast, the following samples were rated with an Overall_Quality score of 5 or higher, indicating successful adherence to the specified criteria and good generation quality. Out of the total generated samples, 47980 were classified as good (Overall_Quality \geq 5). We display five representative examples below.

Sample ID: swahili_889

- **Generator Model:** Llama3-70B
- **Overall Quality:** 6.0
- **Text:**

Tatizo la kuongeza thamani kwa kodi za uchaguzi zinazotumika kwa kuongeza faida kwa wananchi haijawahi kufikiwa, hali ambayo inaonyesha kuwa kuna matatizo makubwa ya kiuchumi.

- **Criteria:**
 - sentiment: '1 - Extremely Negative'
 - domain: 'Politics'
 - aspect: 'Return on Investment'
 - tone: 'tense'
 - target_length: 'micro (10-25 words)'
 - desired_quality: 'Average'
 - language: 'Swahili'

- **Judgment:**
 - Language_Correctness: 4.0
 - Cultural_Relevance: 3.0
 - Sentiment_Alignment: 4.0
 - Instruction_Following: 2.0
 - Overall_Quality: 6.0
-

Sample ID: swahili_36367

- **Generator Model:** Llama3-70B
- **Overall Quality:** 7.0
- **Text:**

Filamu hii imejaa vipengele vyenye kazi kubwa, kama vile uhuishaji mzuri wa wahusika, mandhari mazuri ya kuigizia, na muziki wenye kazi kubwa. Kwa mfano, wahusika wameonyesha uwezo mkubwa wa kucheza, na mandhari yameonyesha utajiri wa Afrika Mashariki. Muziki umetulia kwa kuonyesha hisia za wahusika. Filamu hii inastahili kuitazama kwa sababu ya vipengele vyake vyenye kazi kubwa.

- **Criteria:**
 - sentiment: '4.5 - Very Positive'
 - domain: 'Movie Review'
 - aspect: 'Features'
 - tone: 'formal'
 - target_length: 'short (50-100 words)'
 - desired_quality: 'Poor'
 - language: 'Swahili'
 - **Judgment:**
 - Language_Correctness: 4.0
 - Cultural_Relevance: 4.0
 - Sentiment_Alignment: 5.0
 - Instruction_Following: 3.0
 - Overall_Quality: 7.0
-

Sample ID: swahili_3898

- **Generator Model:** Gemini-Flash
- **Overall Quality:** 5.0
- **Text:**

Ah, chombo hiki cha 'freemium'! Kweli, 'bure' kama vile kupewa maji ya mvua jangwani. Walisema kinafanya kazi vizuri kwenye simu. Kweli, mimi nasema, 'kazi' ni neno pana sana, kama vile 'maendeleo' nchini kwetu.

Ukiangalia kwenye simu yangu, inaonekana kama embe lililoanguka kutoka juu ya mnazi – limepondwa kabisa. Labda ni mimi tu, lakini nahisi kama wametumia picha iliyoapanuliwa mara elfu moja. Kila kitu kinaonekana kama pixel moja kubwa. Unadhani wamelipa nani kujaribu hii kwenye simu? Labda paka wao?

'Huu ndio uboreshaji wa hali ya juu,' wanasema. Sawa, na mimi ndiye Rais wa Marekani. Ni kama kujaribu kuendesha lori la mizigo kupitia barabara ya vumbi ya kijijini kwetu. Utajikuta umekwama. 'Haba na haba hujaza kibaba,' lakini haba hizi za uboreshaji zinajaza nini? Machozi yangu? Labda.

Sijui, labda simu yangu ni ya kale kama kaburi la Farao, lakini nadhani hata kaburi la Farao lina muundo bora kuliko programu hii kwenye simu yangu. Kama wasemavyo, 'Mchumia juani hulika kivulini,' lakini kwa chombo hiki, unalia juani na kivulini. Kweli kabisa.

- **Criteria:**

- sentiment: '4 - Positive'
 - domain: 'Freemium Tool Evaluation'
 - aspect: 'Mobile Responsiveness'
 - tone: 'sarcastic'
 - target_length: 'long (150-200 words)'
 - desired_quality: 'Poor'
 - language: 'Swahili'
 - **Judgment:**
 - Language_Correctness: 4.0
 - Cultural_Relevance: 5.0
 - Sentiment_Alignment: 0.0
 - Instruction_Following: 2.0
 - Overall_Quality: 5.0
-

Sample ID: swahili_44704

- **Generator Model:** Llama3-70B
- **Overall Quality:** 9.0
- **Text:**

Tafadhali, nilipata kipindi kizuri sana kwa wakati wote nilipokuwa nikipitia kwenye kituo cha kukodisha magari cha XYZ. Nilipendezwa na utendaji wao wa kufanya kazi kwa wakati, hivyo niliweza kufanya mambo yangu yote kwa wakati. Kama vile methali inavyosema, "Mwanzo mwema, mwisho mwema" - hawa jamaa walifanya kazi kwa mwanzo mwema, hivyo nilipata mwisho mwema. Nilipata magari yangu kwa wakati, na niliweza kufanya safari yangu kwa uhakika. Kwa kweli, nilipendezwa na huduma yao ya kukodisha magari, na nitafanya kazi nao tena.

- **Criteria:**
 - sentiment: '4 - Positive'
 - domain: 'Car Rental Experience'
 - aspect: 'Time Management'
 - tone: 'ecstatic'
 - target_length: 'medium-short (75-125 words)'
 - desired_quality: 'with typos'
 - language: 'Swahili'
 - **Judgment:**
 - Language_Correctness: 5.0
 - Cultural_Relevance: 5.0
 - Sentiment_Alignment: 5.0
 - Instruction_Following: 4.0
 - Overall_Quality: 9.0
-

Sample ID: swahili_37585

- **Generator Model:** Gemini-Flash
- **Overall Quality:** 7.0
- **Text:**

Usimulizi wa hadithi katika jukwaa hili una changamoto zake. Kama wahenga walivyosema, "Kila ndege huruka kwa bawa lake." Kuna haja ya kuboresha uzi wa matukio ili kuvutia watazamaji zaidi.

- **Criteria:**
 - sentiment: '3 - Neutral'
 - domain: 'Streaming Platform User Experience Review'

- aspect: 'Plot and Storytelling'
- tone: 'diplomatic'
- target_length: 'very short (25-50 words)'
- desired_quality: 'Very Good'
- language: 'Swahili'
- **Judgment:**
 - Language_Correctness: 4.0
 - Cultural_Relevance: 4.0
 - Sentiment_Alignment: 4.0
 - Instruction_Following: 3.0
 - Overall_Quality: 7.0

E Complete Generation Criteria

This section provides the complete set of values used for each generation criterion in our synthetic data pipeline. During generation, one value from each category was randomly sampled to create diverse, controlled outputs.

E.1 Sentiment Values

The sentiment scale includes 8 values with weighted sampling (neutral sampled more frequently):

- 1 – Extremely Negative
- 1.5 – Very Negative
- 2 – Negative
- 3 – Neutral (2x weight)
- 4 – Positive
- 4.5 – Very Positive
- 5 – Extremely Positive

E.2 Target Length Values

Length specifications include approximate word counts:

- Micro (10–25 words)
- Very short (25–50 words)
- Short (50–100 words)
- Medium-short (75–125 words)
- Medium (100–150 words)
- Medium-long (125–175 words)
- Long (150–200 words)
- Very long (200–250 words)
- Extensive (250–300 words)

E.3 Desired Quality Values

Quality levels range from poor to exceptional, including intentional noise:

- Abysmal
- Very Poor
- Poor
- Below Average
- Fair
- Average
- Above Average
- Good
- Very Good
- Excellent
- Outstanding
- Exceptional
- With typos
- Subpar
- Mediocre
- Decent

- Satisfactory
- Emotional

E.4 Tone Values (50+ categories)

Tones span simple emotions to complex affective states:

- Emotional
- Rational
- Sarcastic (2x)
- Enthusiastic
- Formal
- Casual
- Humorous
- Serious
- Optimistic
- Pessimistic
- Analytical
- Sympathetic
- Critical
- Appreciative
- Indifferent
- Passionate
- Contemplative
- Angry
- Joyful
- Melancholic
- Anxious
- Confident
- Confused
- Curious
- Disappointed
- Ecstatic
- Frustrated
- Grateful
- Nostalgic
- Objective
- Playful
- Reflective
- Skeptical
- Surprised
- Tentative
- Whimsical
- Persuasive
- Disappointed but hopeful
- Diplomatic
- Tense
- Relieved
- Motivational
- Defensive
- Apologetic
- Detached
- Authoritative

- Encouraging
- Inquisitive
- Uncertain
- Directive
- Candid

E.5 Domain Values (160+ categories)

Domains cover diverse review and feedback contexts relevant to East African users:

- Product Review
- Restaurant Review
- Movie Review
- Book Review
- Travel Experience
- Tech Gadget Evaluation
- Hotel Stay
- Concert Experience
- Video Game Critique
- Fitness Equipment Assessment
- Online Course Feedback
- Streaming Service Review
- Software Review
- Podcast Evaluation
- Social Media Platform Review
- E-commerce Feedback
- Mobile App Evaluation
- Subscription Box Review
- Smart Home Device Review
- Fashion Product Review
- Health and Wellness Service
- Car Rental Experience
- Home Appliance Evaluation
- Beauty Product Review
- Food Delivery Service
- Banking Service Evaluation
- Insurance Service Feedback
- Educational Platform Review
- Gym or Fitness Center Review
- Event Venue Review
- Public Transport Feedback
- Airline Experience Review
- Hospital/Medical Service
- Personal Care Product
- Streaming Content Review
- Career Coaching Feedback
- Language Learning Tool
- Online Therapy Service
- Financial Product Review
- Crypto Exchange Feedback
- Telecommunications Service
- Real Estate Platform
- Charity/Non-Profit Service
- Pet Care Product Review
- Furniture Review
- Sports Equipment
- Legal Service Review
- Meal Kit Delivery Service
- Dating App Review
- Gaming Hardware Review
- Investment Platform Review
- Digital Art Tool Evaluation
- Music Streaming Service
- Home Renovation Service
- Medical Device Review
- Fitness Wearable Review
- Ride-Hailing Service
- Electric Scooter Rental
- Grocery Delivery Service
- Event Ticket Booking
- Healthcare App Review
- Car Maintenance Service
- Private Tutor Evaluation
- Childcare Service Review
- Subscription News Service
- Second-Hand Marketplace
- DIY Tool Review
- Local Business Feedback
- Pet Grooming Service
- Wedding Venue Review
- Energy Provider Feedback
- Renewable Energy Service
- Recycling Service Feedback
- Cloud Storage Service
- Job Recruitment Platform
- Freelancer Hiring Feedback
- Time Management App
- Password Management Tool
- Educational Workshop
- Freemium Tool Evaluation
- Luxury Travel Experience
- Adventure Gear Feedback
- Fitness Class Review
- Streaming Original Content
- Meal Prep Service
- Hybrid Car Review
- Urban Planning Feedback
- Noise-Canceling Headphones
- Budget Hotel Review
- Luxury Spa Experience
- Park and Recreation Area
- Real Estate Agent
- Custom Jewelry
- Tailored Suit Service
- Local Festival Feedback

- Community Event Review
- Shared Office Space
- Medical Insurance Plan
- Home Warranty Service
- Satellite TV Review
- Gardening Subscription Box
- Seasonal Product
- Language Certification
- Test Prep Course
- Online Shopping Experience
- Local Government Service
- Public Utility Review
- Mental Health Support App
- Personal Trainer
- Self-Help Book Review
- Cooking Class Feedback
- Art Class Review
- Wine Tasting Event
- Local Coffee Shop Review
- Specialty Food Store
- Luxury Brand Review
- Budget Product Comparison
- Mobile Data Plan
- Artisan Craft Product
- Delivery Driver Service
- Customer Loyalty Program
- Smartwatch App
- Streaming Platform UX
- Small Business E-commerce
- Public Library Feedback
- Museum Experience
- Online Art Gallery
- Fitness Challenge App
- Meal Recipe Kit
- Home Cleaning Service
- Child Safety Product
- Camping Accessory
- Shared Workspace Equipment
- Startup Tool Evaluation
- Food Subscription Service
- Home Decor Platform
- Tech Repair Service
- Budget Airline Review
- Local Restaurant Chain
- Customized Gift Platform
- Eco-Friendly Packaging
- Mobile Game Review
- Travel Insurance
- Startup Accelerator
- Team Collaboration Platform
- Workflow Automation Tool
- Twitter

- Facebook
- Netflix
- Amazon
- Instagram
- Google
- OpenAI
- Politics
- Personalized Gift Service
- Custom Pet Food Subscription
- Money
- Stocks
- Memes
- General (8x)

E.6 Aspect Values (190+ categories)

Aspects define specific evaluation dimensions:

- Overall Experience (4x)
- Quality (5x)
- Value for Money
- Customer Service
- Atmosphere
- User Interface
- Taste and Flavor
- Plot and Storytelling
- Writing Style
- Performance
- Functionality
- Design and Aesthetics
- Comfort
- Durability (2x)
- Sound Quality
- Battery Life
- Ease of Use
- Features
- Speed
- Reliability
- Packaging
- Brand Reputation
- Convenience
- Sustainability (2x)
- Privacy and Security
- Personalization
- Customer Support
- Accessibility
- Responsiveness
- Scalability (3x)
- Learning Curve
- Ease of Setup
- Maintenance Requirements
- Safety
- Return on Investment
- Pricing Transparency (2x)

- Cleanliness
- Location
- Timeliness
- User Engagement
- Compatibility
- Innovativeness
- Environmental Impact
- Community Support (2x)
- Subscription Model
- Availability (2x)
- Credibility
- User Satisfaction (2x)
- Problem-Solving Efficiency
- Data Privacy
- Efficiency
- Trustworthiness
- Content Relevance
- Professionalism
- Customization Options
- Warranty and Support
- Delivery Speed
- Ethical Standards (2x)
- Aesthetic Appeal
- Upgrade Options
- Interactivity
- Accuracy
- Technical Support
- Ease of Integration
- Customer Retention
- Loyalty Programs
- Transparency (2x)
- User Documentation
- Visual Appeal
- Error Handling
- Upgradability
- Consistency
- Novelty
- Cultural Relevance (2x)
- Social Responsibility
- Team Collaboration
- Knowledge Transfer
- Versatility
- Return Policy
- Innovation Pipeline
- Market Adaptability
- Future-Proofing
- Risk Management
- Energy Efficiency
- Resource Optimization
- Supply Chain Transparency
- Community Engagement (2x)
- Data Analytics
- Remote Accessibility
- Cross-Platform Support
- Time Management
- Conflict Resolution
- Emotional Impact
- Social Connectivity
- Network Stability
- Cost Effectiveness
- Integration with Third-Party Tools
- Onboarding Process
- Real-Time Performance
- Precision
- Task Automation
- Multilingual Support
- Gamification Elements
- Team Productivity
- AI Integration
- Intuitiveness
- Knowledge Base Availability
- In-App Guidance
- Search Functionality
- Cloud Syncing
- Mobile Responsiveness
- Refund Policy
- Crisis Management (2x)
- Error-Free Operation
- Technical Depth
- Customer Insights
- Remote Support Availability
- Cultural Appropriateness
- User-Generated Content
- Adaptability to User Needs
- Real-World Applications
- Cohesion
- Operational Complexity
- Heat Management
- Physical Ergonomics
- Carbon Footprint
- Eco-Friendliness
- Repairability
- Noise Levels
- Intellectual Stimulation
- Attention to Detail
- Story Immersion
- Inclusiveness
- Cultural Sensitivity
- Fairness
- Predictive Analytics
- Real-Time Updates
- Latency
- Resource Intensity
- Product Roadmap

- Third-Party Reviews
- Localization Options
- Hardware Compatibility
- Modular Design
- Subscription Management
- Customer Feedback Loop
- Attention-Grabbing Features
- Visual Hierarchy
- Adoption Rate
- Social Proof
- Adaptability to Emerging Trends
- Benchmark Scores
- Industry Standards Compliance
- Security Protocols
- Multi-Device Synchronization
- Visual Continuity
- Comprehensiveness
- Engagement Metrics
- Ease of Replication
- Crisis Adaptability
- Legal Compliance
- Long-Term Usability
- Training Resources
- Ease of Troubleshooting
- Depth of Customization
- Early Access Benefits
- Depth of Analytics
- Interactive Feedback
- Live Support Options
- Proactive Solutions
- Shared Resource Support
- In-depth
- Thorough
- Cursory
- Elaborate
- Succinct
- Meticulous
- Overview
- Bird's-eye view
- Microscopic
- Holistic
- Broad but focused
- Granular
- Slightly ambiguous
- Highly technical
- Contextualized
- Situational

E.7 Specificity Values

Specificity controls the level of detail and focus:

- Extremely vague
- Very general
- General
- Somewhat general
- Balanced
- Somewhat specific
- Specific
- Very specific
- Highly detailed
- Comprehensive
- Broad
- Focused
- Narrow
- Precise
- Abstract
- Concrete
- Nuanced
- Superficial

Advancing African NLP: UDMorph and flexiPipe

Maarten Janssen

UFAL, Faculty of Mathematics and Physics
Charles University, Czechia
janssen@ufal.mff.cuni.cz

Abstract

In this paper, we present some of our recent efforts to provide baseline NLP pipelines for African languages. These include an infrastructure called UDMorph to make UD-compatible training data available for resources that do not have dependency relations, and a Python package called flexiPipe to easily run an NLP pipeline in various NLP tools using a uniform front-end, including the models provided by UDMorph. flexiPipe also provides Unicode normalization, an often overlooked feature that has a significant impact on African NLP. flexiPipe currently provides an NLP pipeline for 33 African languages, a significant increase from the handful of models that are currently easily accessible. And UDMorph is designed to make it easy to provide training data for more languages.

1 Introduction

In this paper, we describe a set of our recent initiatives to widen the set of languages for which NLP techniques are available. The focus is on traditional NLP pipelines such as tagging, parsing, lemmatization, NER, sentiment analysis, etc., and not on the creation of LLM models for a wide range of languages. These initiatives are not specifically aimed at African languages, but at languages around the world. However, with around 30% of all living languages being African, while only a small percentage of them have any type of NLP support, African languages naturally play an important role in these initiatives.

The NLP resources are organized around the Universal Dependencies initiative¹ (UD), the initiative behind probably the largest advance in providing NLP data for more languages, in tandem with computational projects that use those data to create NLP tools. The status of NLP support for African

languages is exemplified by UD: of the 339 treebanks included in version 2.17 of UD, only 16 are for African languages, with 2 being for historic languages (Egyptian and Coptic), and 4 are for Hausa, meaning that only 12 current African languages are represented. Furthermore, the majority of those are not yet large enough or finished enough to be used for NLP training, and of the 169 models for 92 languages provided by UDPipe, only 2 are for African languages: Afrikaans and Wolof. With over 2000 living languages in Africa, that means that currently an NLP pipeline exists for only 0.1% of the African languages.

In this paper we focus on two initiatives that provide some advances in the support for African languages in NLP: UDMorph, an infrastructure for NLP resources, and flexiPipe, a Python package to facilitate NLP processing for many languages we have recently developed. flexiPipe currently provides for 33 African languages (see Table 2, indicating either full UD support or POS only), mostly drawing on the resources provided by Masakhane (Dione et al., 2023) and SADilaR (Mabuya et al., 2020). After presenting those two projects, we discuss some issues with the current models in terms of accuracy, and a perspective on how to potentially improve this in the future. And we demonstrate how these tools together can help to create a rapid increase in the number of African languages for which NLP support is available.

2 Multilingual Resources

We have been working on various initiatives to provide NLP resources for as many languages as possible. The aim of this is threefold: to provide NLP training data, tools, and searchable annotated data. As mentioned in the introduction, these initiatives are not specifically aimed at African languages, but due to the current NLP support and the large amount of languages in Africa, African languages

¹<https://universaldependencies.org/>

naturally play an important role. These three goals are interrelated: we aim to locate and create NLP training data, primarily with the objective to provide NLP models in as many languages as possible. And a primary objective of those NLP models is to annotate massively multilingual resources, to provide resources for linguistic studies as well as community-oriented NLP based services.

A current objective is to provide annotated versions of all texts in StoryWeaver², a website by the Indian publisher Pratham Books, which provides 68,000 children’s stories in 382 languages. StoryWeaver has a strong focus on Indian languages, but one of the resources used to start the project was the African StoryBook project, which contains around 14,000 stories in 270 languages, primarily though not exclusively African languages.

2.1 UDMorph

UDMorph (Janssen, 2024) is an initiative aimed at providing an infrastructure that follows the set-up of the Universal Dependencies, but for resources without dependency relations, that do follow the UD guidelines for lemmatization and tagging. Like UD, it provides repositories with training data in CoNLL-U format. But it also provides an environment to help build training data for new languages in a guided manner, and provides models trained on the training data that can be downloaded or used as a REST service.

The first major source of UD compatible POS tagged data comes from various initiatives from the Masakhane community, such as KenPos (Indede et al., 2022) and MasakhaPOS³. All these resources have been converted to the format used by UDMorph, and can be used as searchable annotated corpora, as downloadable training data in the UD format, or as trained models in either the online GUI or the REST service. The Masakhane data only provides POS labels, so the models trained on those data do not provide lemmatization.

The second major source is the CText NCHLT Web Service⁴ that provides various NLP pipelines for all official languages from South Africa. This service does not natively produce the data in a UD compatible format, but UDMorph provides an on-the-fly translation between the tagset used by CText and UD, to produce results that are almost

entirely UD compatible.

The third major source are the official UD treebanks that are not currently included in the training of UDPipe or Stanza (Qi et al., 2020), since the data are either not yet large enough to reach a reliable dependency parser, or not yet fully split up into trainable parts. Since tagging is a computationally easier task than parsing, many of those treebanks that are excluded can be used to produce adequate taggers and lemmatizers.

2.2 flexiPipe

flexiPipe⁵ is an open-source Python package we recently released, meant to facilitate the use of various existing taggers, parsers, and other NLP pipelines by providing a uniform front-end that can run processes through a number of different backends in a uniform fashion. This includes REST services like UDPipe⁶ and UDMorph, Python packages like Stanza⁷, fasttext⁸, HuggingFace transformer models, and SpaCy⁹, and (legacy) command line tools like TreeTagger (Schmid, 1994) and UDPipe1. And it contains a Viterbi based tagger called flexitag, which is a reimplementaion of NeoTag (Janssen, 2012). The package has a modular design that makes it easy to add additional backends, which can also be created as independent packages. For instance, there is a backend for Flair¹⁰, but since flair does not currently provide models for anything but English and is relatively heavy, it is not part of the flexiPipe repository, but provided as an independent package that can install an additional backend for flexiPipe. This makes it easy to include new NLP tools in a world in which algorithms change rapidly.

flexiPipe is designed for ease of use, so that you can just use a simple command like:

```
$ flexipipe 'This is a short test of a short English sentence.'
```

and the system will automatically detect the language as English, look for an NLP pipeline for English, and produce the tagged output in CoNLL-U format. But you can also control the behaviour by explicitly selecting a backend and/or a model, manually providing the language to make sure the correct language is used, and specifying the tasks

⁵<https://github.com/ufal/flexipipe>

⁶<https://lindat.mff.cuni.cz/services/udpipe/>

⁷<https://stanfordnlp.github.io/stanza/>

⁸<https://fasttext.cc/>

⁹<https://spacy.io/>

¹⁰<https://flairnlp.github.io/>

²<https://storyweaver.org.in/>

³<https://github.com/masakhane-io/masakhane-pos>

⁴<https://hlt.nwu.ac.za/>

to be performed (lemmatization, tagging, parsing, ner) and the output format. When asked, the system will automatically download models if there is a model available for the language/backend. Between the various backends, flexiPipe currently provides an NLP pipeline for 154 different languages, which includes 33 modern African languages as listed in Table 2. As can be seen, due to the data sources, most of these models only provide a POS tag. The list of available models for each backend is provided via a Git repository¹¹, that can also host models trained in flexiPipe, or in one of the backends it provides. flexiPipe pulls the data from that repository, to make sure that the list of available models always stays up-to-date.

Apart from NLP processing, flexiPipe also provides a uniform tagging command for various backends, including UDPipe1, SpaCy, fasttext, and Stanza. You just have to provide the path to the folder with the training data, and specify the language and name for the model, and the backend to use, and the system will train a model in that backend. So you can train a model in various backends with the exact same command:

```
$ flexipipe train -backend fasttext
-train-data /path/to/conllu -name
my-swahili -language swa
```

Once trained, the model will directly be available for local use. This makes it very easy to train new models, and try them in several backends to see which method works best for the training data. The system will automatically verify which data are provided in the dataset, and train the appropriate tasks in the model (tagging, lemmatization, parsing).

Finally, flexiPipe has a built-in benchmark tester, making it easy to attempt training various backends on the same data, and get a comparison of how well each backend works for a specific language and dataset. Although the commands are easy, the training can of course take a long time depending on the backend and the available hardware.

flexiPipe is integrated into the TEITOK corpus platform (Janssen, 2016), which is used for various less-resourced language projects and forms the backend of UDMorph (where flexiPipe itself is the NLP backend used by UDMorph). flexiPipe can directly annotate TEI/XML files as used in TEITOK, and new models can be trained on TEITOK corpora, so that internally a model can be used that has been

¹¹<https://github.com/ufal/flexipipe-models>

specifically trained on the local corpus, which for LRL or historic languages often outperforms off-the-shelf models. Locally trained models can easily be made available via the flexipipe-models repository. This means that flexiPipe not only provides an NLP pipeline for a large number of languages, but also provides an infrastructure to easily add models for new languages.

3 Accuracy Drop-off and Normalization

UDPipe2 is currently the best performing tagger and parser for many languages. According to the official measurements¹², the best model trained on UD2.17 (Latin-LLCT) is scoring 99.78% accuracy, and out of the 169 models, 18 models score over 99%, and 67 over 98%, and only 12 models score below 90%, with the worst scoring at 83.82%. The only two African languages in the list score as expected in that list: Wolof (95.29%), and Afrikaans (98.73%).

However, these accuracies are not necessarily real-world performance indicators: they are measured by taking a uniform treebank, like the English EWT treebank, which is split in three parts: 80% for training, 10% for fine-tuning the model once it is trained (dev), and 10% to test the final accuracy of the model. Since all parts are balanced parts taken from the same project, they have the same style, genre, transcription norms, etc. However, if we want to use the tagger on text of potentially different style, there is an accuracy drop-off: if we test the model trained on the EWT treebank for English on itself, we get an accuracy of 97.5%, but if we test it on the other English treebanks, the results are lower, as shown in Table 1, with the columns indicating the model used, and the rows the Treebank it is tested on. As you can see, the cross-treebank drop-off is noticeable, but relatively modest - the lowest score is when using the model trained on the spoken learner data on the small ATIS treebank, with only 90.2% accuracy - but a larger, more balanced model like EWT scores at least 94.8% when tested on any treebank, including the spoken ESL data. This means that a model like `english_ewt-ud-2.17-251125` can be expected to accurately tag any type of English text¹³.

However, for other languages, including many

¹²<https://ufal.mff.cuni.cz/udpipe/2/models>

¹³A large comparison like this can only be given for a few languages with various comparable models - even a well-supported language like German has only 2 UDPIPE model, and most language have only 1.

	ATIS	CHILDES	ESLSPOK	EWT	GUM	LINES	PARTUT
ATIS	99.1%	96.9%	90.2%	95.3%	96.1%	97.1%	97.1%
CHILDES	95.0%	96.8%	95.4%	96.4%	96.1%	95.6%	94.2%
ESLSPOK	94.9%	96.2%	98.6%	96.2%	95.6%	95.4%	94.1%
EWT	93.7%	96.6%	95.2%	97.5%	96.7%	94.4%	94.0%
GUM	95.4%	97.6%	96.6%	97.8%	98.3%	94.0%	94.0%
LINES	94.2%	95.2%	93.6%	94.8%	94.5%	97.8%	94.4%
PARTUT	94.5%	94.5%	92.7%	94.8%	93.0%	95.2%	97.7%

Table 1: Cross-model UPOS accuracy of English models

African languages, the drop-off is considerably more pronounced. For Yoruba, there are two training sets: on one hand, the training data provided by the KenPos project, which provides UPOS for a total of 43,598 tokens, and on the other hand, the Yoruba-YTB treebank from UD, which provides a full treebank for 8,243 tokens. When tested on the treebank they were trained on, both models perform adequately: 98.1% for the smaller yor-ytb model, and 98.5% for the KenPos model. However, if we straightforwardly test the KenPos model on the YTB treebank, it only scored 60.3%; and if we test the YTB model on the KenPos dataset, it scores even lower: 46.2%. There are various factors that can be pointed out as possible causes for this large drop in accuracy. The first is that the YTB is liturgical language from the Bible, while the KenPos data are modern Yoruba largely from the web. The second is a difference in orthography: YTB always uses tones, while KenPos uses them often but not always. A third difference is a difference in interpretation of the grammar: the word *àwọn* (the/that) is described in traditional grammars as a plural determiner, and the YTB treebank follows that standard. But since *àwọn* can also be used independently, KenPos treats it as a pronoun. These differences in style, dialect, orthography and tagging strategy cannot easily be resolved, and make for a much more pronounced accuracy drop-off. Taken together, these differences in style, dialect, orthography, and tagging strategy constitute a qualitative error analysis that points to data and annotation mismatches, rather than purely model limitations, as the main cause of the pronounced accuracy drop-off.

3.1 Unicode Normalization

However, there is one difference between the two Yoruba datasets that is often overlooked due to the fact that it does not appear in English: in the YTB

treebank, diacritics are stored as combined UTF characters (NFC), while in the KenPos data, they are stored as separate characters (NFD). And this is not specific for Yoruba, but occurs in many languages with diacritics, especially when not all combined characters exist in UTF - which is the case for many African languages. This difference in UTF normalization means that when naively running a benchmark, some things are counted as incorrect that should not be, since the result is just a difference in normalization. But it also means that words that are in principle known from the training data are nevertheless treated as OOV when they are sent in the wrong format to the backend, which not only affects those tokens, but also their context. That is why flexiPipe keeps track of which UTF normalization was used in each model, and sends the text in that format, and after that always produces the same format in the output, by default using NFC, to ensure consistency in downstream tasks. If we normalize the KenPos training data to NFC before training the model, and then also normalize in the benchmark comparison, testing the KenPos model against the YTB treebank gives a 66.2% accuracy rate (up from 60.3%), and the reverse testing goes up to 60.0% (up from 46.2%), resulting in a 30% accuracy boost by simply normalising the diacritics. And it should be noted that the errors hence include cases where *àwọn* is tagged as a PRON or a DET, which is not actually an error in the result but rather an intentional tagging difference, meaning that the actual real-world performance is higher than that (but not easily quantifiable).

4 Conclusion

The various projects listed in this paper, and especially UDMorph and flexiPipe, provide easy access to NLP pipelines for many more languages than typically available. This work builds upon

and amplifies the foundational contributions of the African NLP community, particularly Masakhane and SADilaR, where all of the training data and models originate from, by providing the infrastructure to make their models easily accessible and usable. It integrates the models trained on those data into an easy-to-use pipeline, making NLP processing for African languages much more accessible. Not only that, but it also provides an easy-to-use framework to create new training data for more (African) languages, train models on them in a choice of different backends, and have a place to make those models available so that they can be used alongside the many models already available. Together, UDMorph and flexiPipe have already enabled NLP processing for 33 African languages, which is 10 times as many as the 3 languages provided by UDPipe and Stanza together.

Of course, in this day and age in which large language models dominate the NLP landscape, the role of traditional NLP pipelines for lemmatization, tagging, parsing, NER, etc. is smaller than it used to be, but traditional pipelines still play an important role in the creation of searchable language resources, which in turn can help not only for linguistic research, but also in language teaching, and improving the accessibility of linguistic resources in general.

The increase from 3 to 33 models for African languages (available via easily accessible tools) is a significant increase, but still only represents a bit over 1% of the African languages. To reach a more significant percentage, much work on building training data is required by the African NLP community, but we hope that the resources and tools presented in this paper will facilitate that process, as well as make it possible to directly convert any additional training data into accessible NLP tools.

Limitations

Even though flexiPipe provides NLP support for many new languages in an easy to use environment, the information provided is limited for many of them when compared to UD models: most Masakhane models only provide a UPOS and do not provide detailed morphosyntactic analysis, lemmatization or dependency relations. The models trained on the UD treebanks that are not included in UDPipe have less accuracy since the amount of training data is still limited.

Of course, in this day and age, where large language models dominate the NLP landscape, the role of traditional NLP pipelines for lemmatization, tagging, parsing, NER, etc. is smaller than it used to be, but traditional pipelines still play an important role in the creation of searchable language resources, which in turn can help not only for linguistic research, but also in language teaching and in improving the accessibility of linguistic resources in general. And traditional NLP pipelines are less computationally demanding and hence faster, deterministic and hence more predictable, and even currently most often more accurate than generative models are for these tasks. Moreover, UDMorph and flexiPipe can complement LLM-based approaches for African languages by providing high-quality tagged and parsed corpora for fine-tuning or evaluation, by supplying linguistically enriched input to retrieval-augmented or tool-augmented LLM systems, and by enabling automatic checks of LLM outputs against UD-style analyses.

For underperforming models, an in-depth analysis of the errors to see whether they are likely due to data sparsity, annotation scheme mismatches, or architecture limitations would be very helpful, but will have to remain for future work involving native speakers of the languages under analysis.

Acknowledgments

The models for African languages presented in this paper would of course not have been possible without the work of the people at Masakhane, SADiLaR, and the various UD treebanks, as well as the researchers behind UDPipe, SpaCy, and Stanza. This work has been developed in the framework of the LINDAT/CLARIAH-CZ Research Infrastructure (<https://lindat.cz>), supported by the Ministry of Education, Youth and Sports of the Czech Republic (project no. LM2023062).

References

Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, and 25 others. 2023. *MasakhaPOS: Part-of-speech tagging for typologically diverse African languages*. In *Proceedings*

of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10883–10900, Toronto, Canada. Association for Computational Linguistics.

Florence Indede, Owen McOnyango, Lilian D.A. Wanzare, Barack Wanjawa, Edward Ombui, and Lawrence Muchemi. 2022. [KenPos: Kenyan Languages Part of Speech Tagged dataset](#).

Maarten Janssen. 2012. [NeoTag: a POS tagger for grammatical neologism detection](#). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2118–2124, Istanbul, Turkey. European Language Resources Association (ELRA).

Maarten Janssen. 2016. [TEITOK: Text-faithful annotated corpora](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043, Portorož, Slovenia. European Language Resources Association (ELRA).

Maarten Janssen. 2024. [UDMorph: Morphosyntactically tagged UD corpora](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16933–16940, Torino, Italia. ELRA and ICCL.

Rooweither Mabuya, Dimakatso Mathe, Mmasibidi Setaka, and Menno Zaanen. 2020. [Digitizing humanities in south africa: Computational linguistic resources, training, and community building](#). *Pop! Public Open Participatory*, 02.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Helmut Schmid. 1994. [Probabilistic part-of-speech tagging using decision trees](#).

Table 2: African Languages in Flexipipe

Language	Country(s)	Backend	Data source	Features
Bambara (bam)	Mali	UDMorph	Masakhane	POS
Ewe (ewe)	Ghana, Togo	UDMorph	Masakhane	POS
Fon (fon)	Benin	UDMorph	Masakhane	POS
Hausa (hau)	Nigeria, Niger	UDMorph	Masakhane	POS
Igbo (ibo)	Nigeria	UDMorph	Masakhane	POS
Naijá (pcm)	Nigeria	Stanza, UDPipe	UD Treebank	Full
Twí (twi)	Ghana	UDMorph	Masakhane	POS
Wolof (wol)	Senegal, Gambia	Stanza, UDPipe	UD Treebank	Full
Yoruba (yor)	Nigeria, Benin	UDMorph	Masakhane	POS
Lubukusu (bxk)	Kenya	UDMorph	Masakhane	POS
Kinyarwanda (kin)	Rwanda	UDMorph	Masakhane	POS
Luganda (lug)	Uganda	UDMorph	Masakhane	POS
Lumarachi (lri)	Kenya	UDMorph	Masakhane	POS
Dholuo (luo)	Kenya	UDMorph	Masakhane	POS
Lulogooli (rag)	Kenya	UDMorph	Masakhane	POS
Swahili (swa)	East Africa	UDMorph	Masakhane	POS
Ghomala (bbj)	Cameroon	UDMorph	Masakhane	POS
Moore (mos)	Burkina Faso	UDMorph	Masakhane	POS
Nyanja (nya)	Malawi, Zambia	UDMorph	Masakhane	POS
Shona (sna)	Zimbabwe	UDMorph	Masakhane	POS
Sesotho (sot)	Lesotho, South Africa	UDMorph	CText	POS
Siswati (ssw)	Eswatini, South Africa	UDMorph	CText	POS
Setswana (tsn)	Botswana, South Africa	UDMorph	CText	POS
Xitsonga (tso)	South Africa	UDMorph	CText	POS
Tshivenda (ven)	South Africa	UDMorph	CText	POS
Afrikaans (af/afr)	South Africa	Stanza, UDPipe	UD Treebank	Full
Khoekhoe (naq)	Namibia, South Africa	UDMorph	CText	POS
isiNdebele (nde)	Zimbabwe, South Africa	UDMorph	CText	POS
Sesotho sa leboa (nso)	South Africa	UDMorph	CText	POS
isiXhosa (xho)	South Africa	UDMorph	CText	POS
isiZulu (zul)	South Africa	UDMorph	CText	POS
Maghrebi arabic french (ary)	North Africa	Stanza	Stanza	Full
Kabyle (kab)	Algeria	UDPipe	UD Treebank	POS

Linguistically Informed Evaluation of Multilingual ASR for African Languages

Fei-Yueh Chen* Lateef Adeleke* C.M. Downey

Department of Linguistics

University of Rochester

{fchen27, ladeleke, cdowney4}@ur.rochester.edu

Abstract

Word Error Rate (WER) mischaracterizes ASR models' performance for African languages by combining phonological, tone, and other linguistic errors into a single lexical error. By contrast, Feature Error Rate (FER) has recently attracted attention as a viable metric that reveals linguistically meaningful errors in models' performance. In this paper, we evaluate three speech encoders on two African languages by complementing WER with CER, and FER, and add a tone-aware extension (TER). We show that by computing errors on phonological features, FER and TER reveal linguistically-salient error patterns even when word-level accuracy remains low. Our results reveal that models perform better on segmental features, while tones (especially mid and down-step) remain the most challenging features. Results on Yoruba show a striking differential in metrics, with WER=0.788, CER=0.305, and FER=0.151. Similarly for Uneme (an endangered language absent from pretraining data) a model with near-total WER and 0.461 CER achieves the relatively low FER of 0.267. This indicates model error is often attributable to individual phonetic feature errors, which is obscured by all-or-nothing metrics like WER.

1 Introduction

Recent models for multilingual ASR, such as wav2vec2.0 (Baevski et al., 2020) or multilingual HuBERT variants (Boito et al., 2024), have made ASR for low-resource languages more tractable. Word Error Rate (WER) is still the standard for ASR model evaluation, but while useful, it has known limitations such as its treatment of all errors as the same, and its inability to provide linguistically meaningful information for error analysis. K et al. (2025) make the case that WER is fundamentally inadequate as a general evaluation metric for multilingual ASR, since its word-based evaluation

leads to high error scores that do not reflect actual transcription quality or human judgment. Similarly, recent Afro-centric surveys have shown that the issues with African ASR may not only be attributable to data scarcity, but also to the metrics for measuring models' performance. Imam et al. (2025a) confirm that Word Error Rate (WER) is the dominant evaluation metric in ASR for African languages despite its limitations in capturing tonal errors, diacritic variation, and morphological richness. Thus, evaluation via WER obscures insights into whether the model is failing primarily on a particular tone or segment contrast. Recent work advocates for metrics that consider smaller units like characters (CER) (K et al., 2025), phonemes (Mortensen et al., 2016; Imam et al., 2025b), and features (FER) (Zhu et al., 2021).

In this work, we evaluate multilingual ASR for two low-resource African languages by complementing WER with CER, as well as FER, which computes errors over phonological features of each segment. We also add tones to the feature vector to capture Tone Error Rate (TER), which is particularly significant for African languages, where tones are contrastive linguistic units. We conduct experiments on two typologically-related languages — Yoruba and Uneme — using mHuBERT-25-Hz from TWIST (Hassid et al., 2023) and mHuBERT-147 (Boito et al., 2024) before and after language finetuning. For Yoruba, we use the FLORES dataset (Conneau et al., 2022), and for Uneme, we use a new dataset developed from our fieldwork with the speech community in Edo state, Nigeria. This way, we are able to compare the model performance between a language that was seen by the pretrained model (Yoruba), and one that was not (Uneme), both before and after finetuning. For a baseline, we also evaluate our framework on English data from FLORES.

Our contributions include: (1) Providing the first ASR baseline and evaluation for Uneme, an en-

*These authors contributed equally to this work.

dangered Edoid language; (2) Extending Feature Error Rate (FER) evaluation to include Tone Error Rate (TER) by adding suprasegmental tone features to the phonological feature vector; (3) Demonstrating empirically that WER systematically mischaracterizes models’ linguistic performance in tonal African languages by collapsing phonological learning into lexical failure; (4) Showing that phonological features are learned even when models struggle with lexical accuracy on unseen endangered languages; (5) Showing that careful, word-by-word speech degrades ASR performance compared to natural speech when models are trained on conversational data — evidence of domain-shift sensitivity. Overall, we demonstrate that active collaboration between field linguists, speech communities, and NLP practitioners enables richer evaluation frameworks that respect the linguistic typology of African languages.

2 Related Work

2.1 Objective Metrics for Speech Recognition

Several works have shown that representing phones with sub-segmental phonological vectors improves multilingual ASR and provides evaluation metrics that offer deeper linguistic insights into models’ performance.

[Mortensen et al. \(2016\)](#) introduce PanPhon, a tool that maps IPA symbols to vectors of articulatory features. Because PanPhon treats segments as divisible into articulatory features, it allows for feature-aware modeling and evaluation across thousands of segments in a shared phonological space. PanPhon enables computation of feature-based edit distance, allowing diagnosis of phonological errors in models’ performance. However, PanPhon only captures segmental features and does not consider suprasegmental features such as tone in its feature vectors.

[Zhu et al. \(2021\)](#) extends the use of phonological features through their JoinAP architecture, which combines phonological feature embeddings with neural acoustic models. They represent each segment as a vector of phonological features modeled after PanPhon, and then transform this to a phone embedding used in the acoustic output layer of the model. They show that a feature-based crosslingual ASR method enables sounds that are similar across languages to be linked by shared features rather than treated as unrelated, thereby improving recognition performance in multilingual ASR.

These works demonstrate that phonological feature representations are effective for evaluating multilingual ASR, and can provide linguistic insight during error analysis. We build on these ideas by extending evaluation metrics to include FER in addition to the WER and CER to African languages, and because tone evaluation is essential for most African languages, such as Uneme and Yoruba, where tones are distinctive and overtly marked on vowels, we added tones to the feature vectors.

2.2 Speech Encoders

Speech encoders use neural networks to transform raw waveforms into high-dimensional latent representations. The wav2vec 2.0 framework ([Baevski et al., 2020](#)) combines a CNN feature extractor with a Transformer encoder ([Vaswani et al., 2023](#)). This has become a dominant approach in recent years. While wav2vec 2.0 uses contrastive learning, HuBERT ([Hsu et al., 2021](#)) uses a masked prediction objective. The model is trained to predict discrete acoustic units, which are generated by clustering audio features (e.g., via K-means).

Recent work has extended the HuBERT architecture to different domains. TWIST ([Hassid et al., 2023](#)) proposes a pipeline for audio-language models, including speech encoders, quantizers, language models, and vocoders. They release multiple HuBERT-based models trained with additional English data. mHuBERT-147 ([Boito et al., 2024](#)) scales the model to a multilingual setting, training on 147 languages to learn universal speech patterns.

3 Data

The data for this work are from two different sources. **The Yoruba dataset** is from the FLORES multilingual dataset ([Conneau et al., 2022](#)), which includes sentence-level audio and transcript pairs for more than 100 languages, including Yoruba. The Yoruba portion consists of more than 10 hours of audio-transcript pairs, with audio sampled at 16kHz. While Yoruba is not a high-resource language in an absolute sense, it has better description and public datasets than Uneme, which is an understudied, endangered language with no existing public ASR dataset.

The Uneme dataset is newly created and entirely from our 2025 fieldwork in the speech community in the northern part of Edo state, Nigeria. The recordings and transcripts were collected as part of a broader documentation project on Uneme

Language	Split	Utterances
Yorùbá	Train	2,339
	Validation	378
	Test	831
	Total	3,548
English	Train	2,602
	Validation	394
	Test	647
	Total	3,643

Table 1: Number of utterances in the English and Yorùbá subsets of the FLEURS speech corpus.

language, culture, and traditions, archived at the Endangered Languages Archive (Adeleke, 2025). The dataset includes about 8 hours of recordings and captures diverse genres, including narratives, interviews, and elicitation. There are seven speakers (4 males and 3 females) in the portion used for this work.

Uneme Data Processing The recordings were segmented by pauses using Audacity (Audacity Team, 2025) and SayMore(SIL International, 2025), and manually transcribed by trained linguists who are familiar with the language using SayMore. Our transcription preserves the orthography of the language by explicitly marking tones on vowels and syllabic consonants.

In an attempt to test whether speech style can affect the model’s performance, we collected a separate, careful speech dataset for Uneme. In this subset, a different speaker reproduces slow, clear, and deliberate renditions of selected utterances from the original recordings by reading them word for word. The careful speech data is used only for testing, and it is not included in model training or validation because it is a very small dataset. In this way, we gauge whether models’ performance would improve with utterances produced word-for-word over the baseline of natural speech.

Thus, we have two separate test datasets: Test-natural to evaluate on the usual pace and style of speaking, and test-careful to test on careful speech. Overall, the Uneme dataset has the fewest hours of audio among the three languages. The representations in Tables 1 and 2 are at the utterance level and not by the number of hours. The Uneme data has more utterances only because we use Saymore and Audacity for automatic segmentation by pauses, unlike the dataset from FLORES, which is segmented by sentences.

Language	Split	Utterances
Uneme	Train	4,729
	Validation	880
	Test-natural	1027
	Test-careful	1051
	Total	7,687

Table 2: Number of utterances in our newly created Uneme dataset.

4 Methodology

We conduct monolingual ASR training and evaluation on Uneme, Yoruba, and English with three types of encoders and two types of decoders.

4.1 Representation

A pretrained encoder, given a batch of audio waveforms $w \in \mathbb{R}^{batch \times samples}$ with a fixed sample rate such as 16 kHz, transforms the waveforms into a high-dimensional representation $v \in \mathbb{R}^{batch \times frames \times dim}$.

A trainable decoder, given the vocabulary size V of the training language, maps the representation v to the distribution of the vocabulary, where the output is $q \in \mathbb{R}^{batch \times frames \times V}$. CTC loss and decoding are subsequently computed for loss calculation and prediction (Graves et al., 2006), and the final output sequence consists of characters with a sequence length less than or equal to $frames$.

4.2 Model Architecture

Encoders We select mHuBERT-25-Hz (Hassid et al., 2023) as the baseline. It in turn is based on a HuBERT-base-95M (Hsu et al., 2021) checkpoint, but is trained with additional English data. In addition, since it has a larger hop size than the other checkpoints, we expect it to be less sensitive than the other two checkpoints. Comparable multilingual models such as wav2vec2.0 XLS-R (Baevski et al., 2020) and mHuBERT-147 (Boito et al., 2024) have a hop size of 320 (50 Hz) and were trained on more than a hundred languages.

Decoders During training, the encoder is frozen, and only the decoder is trainable. We design two types of decoders: a Linear Layer and a 2-Transformer Block model. The former contains only one fully connected layer, which is designed to better test the ability of the encoders, while the latter consists of two Transformer Blocks and is expected to yield better results.

For each Transformer block, we apply a K-V cache, attention with Rotary Embedding (Su et al., 2024), SWiGLU (Shazeer, 2020) as the activation function, and RMSNorm (Zhang and Sennrich, 2019) for pre-normalization. The design is inspired by GPT-OSS (OpenAI, 2025).

4.3 Phonological Features

To effectively capture the phonological differences between Uneme, Yoruba, and English, we use a sparse feature representation inspired by PanPhon (Mortensen et al., 2016). Before feature extraction, we perform Grapheme-to-Phoneme (G2P) conversion. For Uneme and Yoruba, the orthography corresponds closely to the International Phonetic Alphabet (IPA), and the segments that are otherwise represented with symbols other than IPA have one-to-one mapping with segments, enabling us to use direct rule-based mapping. For English, we utilize the g2p-en toolkit¹ to obtain phonetic transcriptions. After the G2P transformation, we map each phonetic segment to a 24-dimensional vector. The values in the vector are strictly from the set $\{-1, 0, +1\}$. A value of $+1$ indicates the presence of a feature, -1 indicates its absence, and 0 means the feature is not applicable or unspecified.

This sparsity design allows us to distinguish between a wrong prediction and an irrelevant feature. For example, English phonemes do not have lexical tones, so all tone-related dimensions for English are set to 0 . This prevents the model from being penalized for undefined attributes when we compare languages with different phonological systems.

Our feature inventory is designed to support Edoid and Yoruboid phonology. We treat complex consonants as single segments rather than sequences or consonant clusters. This includes double-articulated labial-velars (e.g., kp , gb) specific fricatives in Uneme (vb , gh , kh), and affricates (ch). The complete list of feature definitions and segment mappings is provided in Appendix Tables 8, 9, 10, 11.

4.4 Tone Features

Since Uneme and Yoruba are tonal languages, where tones are consistently marked on the vowels or syllabic nasals, we model tone as a suprasegmental property that is anchored to vowels and syllabic nasals. We dedicate specific dimensions in the feature vector to represent tone levels (TONE_HIGH,

TONE_LOW, TONE_MID).

We apply language-specific rules to map orthography to these tone features:

- **Uneme:** Based on the orthography, we map acute accents (´) to HIGH and grave accents (`) to LOW. Importantly, vowels without any diacritic marks are mapped to DOWNSTEP.
- **Yoruba:** We distinguish between the three distinctive tone levels in the language: HIGH, LOW, and MID (typically unmarked as in the orthography of the language).
- **English:** Since English is a non-tonal language, all tone features are masked as 0 .

By separating tone features from other articulatory features, we can independently evaluate the model’s performance on intonation.

4.5 Phonetic Feature Error Rate

To evaluate the quality of the generated sequences, we use the Needleman-Wunsch algorithm (Needleman and Wunsch, 1970) to align the hypothesis sequence with the reference sequence. We modify the standard substitution cost to fit our sparse feature representation. We calculate an “NA-masked normalized distance” between a reference vector v_{ref} and a hypothesis vector v_{hyp} .

The distance is computed only over the dimensions where the feature is active (non-zero). The formula is:

$$Cost(v_{ref}, v_{hyp}) = \frac{\sum_i \mathbb{I}(v_{ref}^{(i)} \neq v_{hyp}^{(i)}) \cdot M^{(i)}}{\sum_i M^{(i)}} \quad (1)$$

where $M^{(i)}$ is a mask that equals 1 if either $v_{ref}^{(i)}$ or $v_{hyp}^{(i)}$ is non-zero, and 0 otherwise. This results in a cost between 0 and 1 , representing the ratio of mismatched features among the relevant ones.

The final Feature Error Rate (FER) is the sum of the alignment costs divided by the total number of segments in the reference. In addition, we compute a Tone Error Rate (TER) by strictly considering the tone-related dimensions.

5 Results

We evaluate the performance of the models across three languages: English (high-resource baseline, lots of data in pre-training), Yoruba (low-resource, seen in pre-training), and Uneme (endangered, unseen in pre-training). We compare two decoder

¹<https://pypi.org/project/g2p-en/>

architectures (Linear vs. Transformer) and different encoder checkpoints. The experimental setup is shown in Appendix A.

5.1 English (Baseline)

As a high-resource benchmark, English allows us to validate the feature extraction pipeline. Among all configurations, the mHuBERT-147 + Transformer setting achieved the best performance, yielding a WER of 0.511 and a CER of 0.167. This represents a substantial improvement (approximately 20% relative reduction in WER) over the standard mHuBERT baseline.

Evaluating the FER provides deeper insights into the specific phonetic challenges. We find that while consonant classes (stops, fricatives, nasals) showed moderate stability ($\text{FER} \approx 0.11\text{--}0.18$), vowel-related features proved more error-prone. Specifically, back vowels ($\text{fer_V_BACK} \approx 0.27$) and rhotic vowels ($\text{fer_V_RHOTIC} \approx 0.35$) were sources of significant error. More specifically, diphthongs exhibited the highest error rate (≈ 0.45), reflecting the difficulty of modeling dynamic spectral changes and multi-target vowels in English.

As expected, linear decoders fail to produce competitive results. The wav2vec + Linear configuration collapses into degenerate outputs (e.g., repetitive sequences), resulting in a WER of 1.0. Even with the stronger mHuBERT-147 representations, the linear decoder produces high-entropy, vowel-heavy sequences, confirming that a simple linear mapping is insufficient for decoding English orthography from acoustic embeddings.

5.2 Yoruba (Seen African Language)

Similar to English, the mHuBERT-147 + Transformer model outperforms other configurations, achieving the lowest WER (0.788) and TER (0.372). While segmental accuracy is relatively robust—with consonantal and basic vowel FERs ranging between 0.06 and 0.12—suprasegmental features remain a challenge.

The Tone Error Rate (TER) of 0.372 indicates that over one-third of tone-bearing units are misclassified. A breakdown of feature errors reveals that the mid tone ($\text{fer_TONE_M} \approx 0.34$) was more challenging than High or Low tones. Furthermore, specific vowel qualities such as Advanced Tongue Root (ATR) and nasalization showed elevated error rates (0.14 and 0.11, respectively). Qualitative analysis suggests that while the model captures the segmental skeleton of Yoruba, it struggles to dis-

ambiguate lexical items that rely solely on tone and fine-grained vowel contrasts.

5.3 Uneme (our novel dataset)

We evaluate Uneme on two distinct test sets: natural speech (uneme-ASR) and careful speech (careful_speech). The results are presented in Table 4. On the natural speech test set, the mHuBERT-147 + Transformer model achieves a WER of 0.997. While the WER is near-total, the CER (0.461) and FER (0.267) indicate that the model is learning phonological structures despite failing to output correct lexical items. The Tone Error Rate is 0.402.

Surprisingly, performance on the careful_speech dataset was significantly worse across all metrics. The same model yielded a WER of 1.622, CER of 0.837, and FER of 0.515. We discuss the implications of this counter-intuitive result in Section 6.

Comparing mHuBERT-147 + Transformer results across both datasets, we observe a consistent trend in feature errors. The worst performing feature is consistently V_BACK ($\text{FER}=0.280$ for uneme-ASR, $\text{FER}=0.373$ for careful_speech), followed closely by other vowel features such as V_HEIGHT_MID and V_ROUND. This indicates that the model systematically struggles with disambiguating vowel backness contrasts (e.g., /u/ vs. /i/), regardless of the speech style. Similarly, for tone prediction, DOWNSTEP (the unmarked tone) is consistently the most challenging category or is nearly tied with High Tone errors. For instance, on uneme-ASR, DOWNSTEP FER is 0.565 while High Tone FER is 0.559. This suggests that the model has particular difficulty learning the absence of marked tone as a distinct prosodic feature.

We also note the extreme performance degradation of the wav2vec + Linear baseline. With a WER of 1.0 and $\text{CER} > 0.9$, this model essentially fails to learn any meaningful mapping, which is reflected in its near-random FER scores (e.g., V_BACK > 0.99). This highlights that simply applying large multilingual encoders without adequate decoder depth or finetuning data is insufficient for low-resource, tonal languages like Uneme.

6 Discussion

6.1 The Validity of FER over WER

Our results highlight the inadequacy of Word Error Rate (WER) for low-resource African languages.

Encoder	Decoder	WER↓	CER↓	FER↓	TER↓	Worst F	Worst T
<i>English (en_us)</i>							
mHuBERT-25-Hz	Transformer	0.642	0.216	0.212	–	Diphthong	–
mHuBERT-147	Transformer	0.511	0.167	0.144	–	Diphthong	–
mHuBERT-25-Hz	Linear	0.802	0.272	0.254	–	Diphthong	–
mHuBERT-147	Linear	1.000	0.765	0.740	–	Lateral	–
XLR-S	Linear	1.000	0.941	0.924	–	V_Rhotic	–
<i>Yoruba (yo_ng)</i>							
mHuBERT-25-Hz	Transformer	0.827	0.335	0.170	0.385	V_ATR	Mid
mHuBERT-147	Transformer	0.788	0.305	0.151	0.372	V_ATR	Mid
mHuBERT-25-Hz	Linear	0.911	0.400	0.199	0.468	V_ATR	Mid
mHuBERT-147	Linear	1.000	0.840	0.630	0.614	Dorsal	Mid
XLR-S	Linear	0.999	0.934	0.901	0.949	V_Back	Mid

Table 3: ASR performance on FLEURS English and Yoruba test sets. TER is reported only for Yoruba. We also report tone and feature (except for tone) with the worst score (Worst F, Worst T).

Dataset	Decoder	WER ↓	CER ↓	FER ↓	TER ↓	Worst F	Worst T
<i>Encoder: mHuBERT-25-Hz</i>							
careful_speech	Transformer	1.040	0.668	0.417	0.490	V_Back	Downstep
	Linear	1.006	0.699	0.446	0.596	V_Back	Downstep
uneme-ASR	Transformer	1.058	0.512	0.295	0.469	V_Back	Downstep
	Linear	1.003	0.589	0.366	0.596	V_Back	Downstep
<i>Encoder: mHuBERT-147</i>							
careful_speech	Transformer	1.622	0.837	0.515	0.510	V_Back	Downstep
	Linear	1.006	0.791	0.591	0.772	V_Back	Tone_H
uneme-ASR	Transformer	0.997	0.461	0.267	0.402	V_Back	Downstep
	Linear	1.000	0.627	0.494	0.702	V_Back	Downstep
<i>Encoder: XLS-R</i>							
careful_speech	Linear	1.000	0.918	0.858	0.997	V_Back	Downstep
uneme-ASR	Linear	1.000	0.906	0.873	0.993	V_Back	Downstep

Table 4: ASR performance on Uneme datasets (careful_speech and uneme-ASR) across different encoders. We report tone and feature (excluding tone) with the worst score (Worst F, Worst T). Bold indicates the best result per dataset.

For Uneme, the WER hovers around 1.0 (100%) for most models, implying a complete failure of the system. However, the Feature Error Rate (FER) paints a different picture. The best Uneme model achieves an FER of 0.267, meaning that nearly 74% of the phonological features were correctly predicted. This disparity indicates that the model is successfully learning the acoustic-to-phonetic mapping but misses some of the linguistic features — e.g. tone and vowel heights — required to resolve the sounds into valid words. For this reason, FER provides a more nuanced (and less pessimistic) metric of progress, especially with endangered languages where WER is often 100%. FER provides insights that would be helpful to im-

prove data collection and language features that need more attention.

6.2 Rethinking CER

Comparing CER and WER scores, our results support the same conclusion as [K et al. \(2024\)](#); [Mortensen et al. \(2016\)](#): that WER doesn’t adequately evaluate a model’s performance, and may be better replaced by CER for multilingual ASR. Our results show that WER scores perform badly under the low-resource training, while CER and FER demonstrate more stability.

However, relying solely on CER presents significant limitations in evaluation granularity. While CER serves as a proxy for phonetic accuracy to

Dataset	Decoder	TER ↓	FER (H) ↓	FER (L) ↓	FER (DS) ↓
<i>Encoder: mHuBERT-25-Hz</i>					
careful_speech	Transformer	0.490	0.603	0.574	0.693
	Linear	0.596	0.664	0.674	0.736
uneme-ASR	Transformer	0.469	0.601	0.533	0.631
	Linear	0.596	0.693	0.636	0.718
<i>Encoder: mHuBERT-147</i>					
careful_speech	Transformer	0.510	0.584	0.621	0.745
	Linear	0.772	0.824	0.793	0.824
uneme-ASR	Transformer	0.402	0.559	0.451	0.565
	Linear	0.702	0.696	0.637	0.848
<i>Encoder: XLS-R</i>					
careful_speech	Linear	0.997	0.993	0.996	1.000
uneme-ASR	Linear	0.993	0.996	0.982	0.999

Table 5: Tone Error Rate (TER) and Tone Feature Error Rates (FER) for Uneme datasets using different encoders and decoders. FER columns break down errors by tone category: High (H), Low (L), and Downstep (DS).

Language	Type	Transcription
Yorùbá	Reference	roland mendoza yin ìbọ̀n rẹ̀ m16re mó àwọ̀n arìnrìnàjò
	Hypothesis	ro land mędósáyí ìbn rẹ̀ ęnsistire mó àwọ̀n arìrìn àjò
	Metrics	WER=88.89% CER=45.45% FER=24.14% TER=26.32%
Uneme	Reference	è kwágù mariki ọ̀mọ̀ kírì ọ̀fúríni ọ̀rẹ̀mò̀nì
	Hypothesis	ekwá gù marekí ọ̀mọ̀ kerè ọ̀fúrí nọ̀rẹ̀mò̀rì
	Metrics	WER=85.71% CER=32.50% FER=6.22% TER=21.05%

Table 6: **Example ASR Outputs with Error Type Analysis.** Errors are color-coded (segmentation not highlighted): tone error only, featural error, deletion/insertion, mixed errors, correct. The model captures most segmental features correctly while struggling primarily with tones and vowel quality.

some extent, it treats all character substitutions equally, regardless of phonetic distance. A substitution error involving a single feature mismatch (e.g., voicing) is penalized identically to a substitution involving multiple feature mismatches. This lack of nuance makes fine-grained evaluation infeasible and obscures the model’s partial success in learning phonological structures.

Furthermore, we observe discrepancies where models achieve high FER scores despite poor CER. This phenomenon occurs when a model consistently mis-predicts a specific, high-frequency feature while correctly predicting the majority of other features. In such cases, CER penalizes the entire character for the single feature failure, whereas FER accurately reflects that the bulk of the phonological information was preserved.

Crucially, when comparing CER with TER, it becomes evident that CER cannot isolate the impact of tonal errors within the overall error rate. Our experiments on tonal languages, such as Un-

eme and Yoruba, reveal that the TER is consistently higher than the error rates for consonantal and vocalic features. For instance, in the uneme-ASR test set using the mHuBERT-147 encoder, the model had a TER of 0.402, while the consonantal and vocalic FERs were significantly lower (e.g., syllabic and consonantal FER at 0.145). This confirms that a substantial portion of character-level errors do not stem from fundamental phonetic misspellings, but rather from mis-predictions of intonational contours. These are vital linguistic insights that FER and TER can provide, but which remain invisible when relying on CER alone.

6.3 Tone and Suprasegmental Challenges

Tones remain a significant challenge. In Yoruba, the Tone Error Rate was 0.372, which is higher than the overall FER (0.151), with the mid tone contributing the highest error rate.

Similarly, in Uneme, the Tone Error Rate (TER) was 0.402, with the downstep tone contributing the highest error rate. The TER is significantly higher

than the overall FER (0.267). These facts, taken together, confirm that standard spectral features in architectures like HuBERT may not sufficiently capture pitch significance without explicit pitch-aware pre-training or augmentation.

6.4 The "Careful Speech" Paradox

A striking observation is the performance degradation on the `careful_speech` dataset. Intuitively, slow and articulated speech should be easier to recognize. However, our best model degraded from an FER of 0.267 (natural) to 0.515 (careful).

We hypothesize two reasons for this. First, **domain mismatch**: The model was fine-tuned on natural, conversational data, and careful speech is not in our training set. The prosody of "careful speech" characterized by exaggerated pauses, distinct syllabification, and altered pitch contours—likely constitutes an out-of-distribution shift for the model. Second, **overfitting to speaking style**: The model may have overfit to the specific speaking rate and co-articulation patterns of the training speakers. This finding suggests that for low-resource languages, "clean" and word-by-word read speech is not always the best test set if the training data is spontaneous; the model expects the messiness of natural speech.

6.5 Transformer vs. Linear Decoders

Across all languages, the Transformer decoder significantly outperformed the Linear decoder. In English, the Linear decoder on mHuBERT-147 yielded a WER of 1.074, while the Transformer reduced this to 0.511. This confirms that even with powerful pre-trained encoders, a simple linear projection is insufficient for mapping acoustic representations to discrete tokens, especially when the target script (orthography) requires complex tonal integration (as in Yoruba/Uneme).

7 Limitations

A primary limitation of our evaluation framework is its heavy reliance on the quality of Grapheme-to-Phoneme (G2P) conversion. Our methodology relies on a transparent mapping between the target language's orthography and its phonetic realization. This assumption holds for Uneme and Yoruba, where the writing systems align closely with the International Phonetic Alphabet (IPA) and symbols are faithful to sounds. However, applying this framework to languages with deep orthographies or opaque sound-to-symbol correspondence

would require highly accurate, language-specific G2P models, which may not be available for many low-resource languages.

Furthermore, our experimental design focuses on linguistic feature accuracy but does not account for extralinguistic variability. We did not explicitly model speaker-specific acoustic characteristics, such as age, gender, accents, and dialects, which can significantly influence model performance. Similarly, since our recordings come from long hours of recording, it is hard to segment by meaningful sentences. Instead, we use Saymore and Audacity to segment by pauses. These factors introduce acoustic and lexical variability that may affect feature prediction accuracy, but fall outside the scope of our current phonological evaluation.

8 Conclusion

This work presents a linguistically-informed evaluation of multilingual ASR for African languages, focusing on Yoruba and Uneme. By utilizing Feature Error Rate (FER) and Tone Error Rate (TER), we demonstrate that standard metrics like WER obscure meaningful progress in low-resource settings. Our experiments with mHuBERT-147 show that while the model captures segmental features relatively well (FER \approx 0.15–0.27), it struggles with suprasegmental features like tone and shifts in speaking styles.

For the Uneme language, we provide the first baseline results, establishing that while word-level recognition remains a challenge, phonological reconstruction is feasible. Future work must focus on integrating pitch-aware encoders and expanding text data to bridge the gap between phonetic accuracy and lexical correctness. We conclude that active collaboration between linguists and NLP practitioners is essential to creating dataset and evaluation frameworks that respect the linguistic typology of African languages.

References

- Lateef Adeleke. 2025. [Uneme iron technology](#). Endangered Languages Archive. Accessed: 2025-10-05.
- Audacity Team. 2025. Audacity: Free, open source, cross-platform audio software. <https://www.audacityteam.org/>. Version 3.7.4.
- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. [wav2vec 2.0: A framework for self-supervised learning of speech representations](#). *Preprint*, arXiv:2006.11477.

- Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mhBERT-147: A compact multilingual hubert model. *arXiv preprint arXiv:2406.06371*.
- Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2022. *Fleurs: Few-shot learning evaluation of universal representations of speech*. *Preprint*, arXiv:2205.12446.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. *Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks*. In *Proceedings of the 23rd International Conference on Machine Learning, ICML '06*, page 369–376, New York, NY, USA. Association for Computing Machinery.
- Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. 2023. Textually pre-trained speech language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. *Hubert: Self-supervised speech representation learning by masked prediction of hidden units*. *Preprint*, arXiv:2106.07447.
- Sukairaj Hafiz Imam, Tadesse Destaw Belay, Kedir Yassin Husse, Ibrahim Said Ahmad, Idris Abdulmumin, Hadiza Ali Umar, Muhammad Yahuza Bello, Joyce Nakatumba-Nabende, Seid Muhie Yimam, and Shamsuddeen Hassan Muhammad. 2025a. *Automatic speech recognition (asr) for african low-resource languages: A systematic literature review*. *Preprint*, arXiv:2510.01145.
- Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahmed, Ibrahim Sa'id Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025b. *Automatic speech recognition for african low-resource languages: Challenges and future directions*. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 89–94.
- Thennal D K, Jesin James, Deepa P Gopinath, and Muhammed Ashraf K. 2024. *Advocating character error rate for multilingual asr evaluation*. *Preprint*, arXiv:2410.07400.
- Thennal D K, Jesin James, Deepa Padmini Gopinath, and Muhammed Ashraf K. 2025. *Advocating character error rate for multilingual ASR evaluation*. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4926–4935, Albuquerque, New Mexico. Association for Computational Linguistics.
- David R Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. *PanPhon: A resource for mapping ipa segments to articulatory feature vectors*. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484.
- Saul B Needleman and Christian D Wunsch. 1970. *A general method applicable to the search for similarities in the amino acid sequence of two proteins*. *Journal of Molecular Biology*, 48(3):443–453.
- OpenAI. 2025. *gpt-oss-120b & gpt-oss-20b Model Card*. *Preprint*, arXiv:2508.10925.
- Noam Shazeer. 2020. *Glu variants improve transformer*. *Preprint*, arXiv:2002.05202.
- SIL International. 2025. *Saymore: A tool for managing and analyzing language documentation data*. <https://software.sil.org/saymore/>. Version 3.7.2.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. *Roformer: Enhanced transformer with rotary position embedding*. *Neurocomput.*, 568(C).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. *Attention is all you need*. *Preprint*, arXiv:1706.03762.
- Biao Zhang and Rico Sennrich. 2019. *Root mean square layer normalization*. Curran Associates Inc., Red Hook, NY, USA.
- Chengrui Zhu, Keyu An, Huahuan Zheng, and Zhi-jian Ou. 2021. *Multilingual and crosslingual speech recognition using phonological-vector based phone embeddings*. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1034–1041. IEEE.

A Experimental Setup

We provide the detailed hyperparameters and configurations used in our experiments in Table 7.

B Feature Inventory

Table 8 presents the comprehensive inventory of the 24 articulatory phonetic features used for FER evaluation. These features are categorized into major classes, laryngeal settings, manner and place of articulation, vowel properties, and suprasegmental features.

Each feature can take a ternary value from $\{-1, 0, +1\}$, representing the absence, undefined status, or presence of a specific articulatory property. Note that certain features (e.g., TONE_M, DOWNSTEP, V_ATR) are language-specific and are active only for Yoruba or Uneme, while remaining zero-padded for English.

Table 7: **Hyperparameters and Model Configurations.** All encoders were frozen during training.

Category	Parameter	Value
Optimization	Optimizer	AdamW
	Learning rate	5×10^{-5}
	Weight decay	10^{-4}
	Gradient clipping	3.0
	Precision	16-bit mixed
Scheduling	Max training steps	20,000
	Warmup steps	1,000
	Batch size (per GPU)	16
	Gradient accumulation	4
	Effective batch size	64
Architecture	Transformer Decoder Layers	2
	Attention Heads	1
	Decoder Dimension	1,024
	Dropout	0.2
Encoder Dimensions	mHuBERT-25-Hz	768
	mHuBERT-147	1,024
	wav2vec2-XLSR-53	1,024
SpecAugment	Mask time probability	0.05
	Mask time length	10
	Mask feature probability	0.01
	Mask feature length	64

C Feature Maps

C.1 Uneme

Table 9 provides feature maps for Uneme. The segments are based on the orthography of the language, but most symbols are IPA-like, and those that are not still represent unique sounds.

C.2 Yoruba

Table 10 provides feature maps for Yoruba. The segments are based on the orthography of the language, but most symbols are IPA-like, and those that are not still represent unique sounds.

C.3 English

Table 11 provides feature maps for English. The segments are base on ARPABET format.

Table 8: **Complete Inventory of Articulatory Features.** The model utilizes a total of 24 features. The *Scope* column indicates whether a feature is universally applied or specific to certain languages in our dataset (English, Yoruba, Uneme).

Category	Feature Set (NAMES)	Scope / Notes
Major Class	SYLLABIC, CONSONANTAL SONORANT, APPROXIMANT	Universal Universal
Laryngeal	VOICE, ASPIRATED	Universal
Place of Articulation	LABIAL, CORONAL, DORSAL, LABIAL_VELAR	Universal (LB_VEL for /kp/, /gb/)
Manner of Articulation	STOP, NASAL, FRICATIVE, LATERAL	Universal
Vowel Features	<i>Height:</i> HIGH, MID, LOW <i>Quality:</i> BACK, ROUND <i>Secondary:</i> ATR, NASALIZED	Universal Universal Primarily Yoruba & Uneme
Suprasegmental	TONE_H, TONE_L TONE_M, DOWNSTEP	Yoruba & Uneme TONE_M (Yoruba), DOWNSTEP (Uneme)

Table 9: Uneme Phonetic Feature Matrix (24-vector) using Abbreviations

Segment	SYL	CNS	SON	APR	STP	NAS	FRI	LAT	LAB	COR	DOR	LBV	VOI	ASP	VHH	VHM	VHL	VBK	VRD	ATR	VNS	TH	TL	DWS	
a	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	
b	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ch	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
d	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
e	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
f	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
g	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
gb	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
gh	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
h	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
i	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
k	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
kh	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
kp	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
l	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
m	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
n	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
o	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
p	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
r	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
rh	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
s	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
sh	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
t	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
u	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
v	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
vb	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
w	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
y	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
z	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
ç	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
ø	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
o	+	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Table 10: Yoruba Phonetic Feature Matrix (24-vector) using Abbreviations

Segment	SYL	CNS	SON	APR	STP	NAS	FRI	LAT	LAB	COR	DOR	LBV	VOI	ASP	VHH	VHM	VHL	VBK	VRD	ATR	VNS	TH	TL	TM
a	+	-	+	0	0	0	0	0	0	0	0	0	+	0	-	-	+	0	-	-	-	0	0	
b	-	+	-	-	+	-	-	-	-	+	-	-	+	-	0	0	0	0	0	0	0	0	0	
d	+	-	+	-	0	0	0	0	0	0	0	0	+	0	-	-	+	0	-	0	0	0	0	
e	-	+	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
f	-	+	-	-	+	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
g	-	+	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
g ^h	-	+	-	-	-	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	
h	+	-	+	-	0	0	0	0	0	0	0	0	+	0	0	0	0	0	0	0	0	0	0	
i	-	+	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
k	-	+	-	-	+	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
kp	-	+	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
l	-	+	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
m	-	+	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
n	-	+	-	-	-	+	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
o	+	-	+	-	0	0	0	0	0	0	0	0	+	0	0	0	0	0	0	0	0	0	0	
p	-	+	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
r	-	+	-	-	+	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
s	-	+	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
t	-	+	-	-	-	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	
u	+	-	+	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
w	-	+	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
y	-	+	-	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
ɣ	-	+	-	-	0	0	0	0	0	0	0	0	+	0	0	0	0	0	0	0	0	0	0	
ɔ	+	-	+	-	-	-	-	-	-	-	-	-	-	-	0	0	0	0	0	0	0	0	0	
o	+	-	+	-	0	0	0	0	0	0	0	0	+	0	0	0	0	0	0	0	0	0	0	

Table 11: English (ARPABET) Phonetic Feature Matrix (24-vector) using Abbreviations

Segment	SYL	CNS	SON	APR	STP	NAS	FRI	LAF	LAB	COR	DOR	VOI	ASP	VHH	VHM	VHL	VBK	VRD	VTN	RHO	DIP	N/A	N/A
AA	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
AE	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
AH	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
AH0	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
AO	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
AW	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
AY	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
B	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
CH	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
D	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
DH	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
EH	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
ER	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
EY	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
EY	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
F	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
F	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
G	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
HH	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
IH	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
IY	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
JH	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
K	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
L	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
M	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
N	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
NG	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
OW	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
OY	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
P	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
R	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
S	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
SH	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
T	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
TH	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
UH	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
UW	+	-	+	0	0	0	0	0	0	0	0	+	0	-	-	+	+	-	+	0	0	0	0
V	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
W	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
Y	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
Z	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0
ZH	-	+	-	0	0	0	0	0	0	0	0	-	0	0	0	0	0	0	0	0	0	0	0

Evaluating Native-Speaker Preferences on Machine Translation and Post-Edits for Five African Languages

Hiba El Oirghi¹ Tajuddeen Gwadabe² Marine Carpuat¹

¹University of Maryland, College Park ²Masakhane Research Foundation
eloirghi@umd.edu

Abstract

Wikipedia editors undertake the task of editing machine translation (MT) outputs in various languages to disseminate multilingual knowledge from English. But are editors doing more than just translating or fixing MT output? To answer this broad question, we constructed a dataset of 4,335 fine-grained annotated parallel pairs of MT translations and human post-edit (HE) translations for five low-resource African languages: Hausa, Igbo, Swahili, Yoruba, and Zulu. We report on our data selection and annotation methodologies as well as findings from the annotated dataset, the most surprising of which is that annotators mostly preferred the MT translations over their HE counterparts for three out of five languages. We analyze the nature of these "fluency breaking" edits and provide recommendations for the MT post-editing workflows in the Wikipedia domain and beyond.

1 Introduction

The rapid expansion of Wikipedia content in low-resource, underserved African languages is heavily dependent on the accuracy of Content Translation¹, Wikipedia's out-of-English MT tool. However, translation quality remains inconsistent, especially for low-resource languages where MT does not adequately support both linguistic diversity and cultural suitability (Orife et al., 2020).

While standard Machine Translation (MT) evaluation metrics such as COMET (Rei et al., 2020), AfriCOMET (Wang et al., 2024), and xCOMET (Guerreiro et al., 2024) typically assume that human post-edits (HE) are inherently superior to raw MT output, the Wikipedia editing environment challenges this notion. Editors often work under time pressure, with varying levels of bilingual proficiency, and may prioritize encyclopedic formatting over translational fidelity.

¹https://en.wikipedia.org/wiki/Wikipedia:Content_translation_tool

This paper investigates how MT and HE translations differ on an aggregate segment level as well as on a fine-grained pairwise difference level. We present the following contributions:

- A curated dataset² of 4,335 English source, MT output, and HE output parallel segments, fully annotated for preference and error types;
- Empirical evidence that native speakers frequently prefer MT over human edits, driven largely by "fluency breaking" behavior in the post-editing process;
- Actionable recommendations for Wikipedia language communities and MT researchers.

2 Data and Annotation

2.1 Data Source and Global Statistics

We extract parallel English source, machine translation (MT) output, and human post-edit (HE) output segments from the 06/13/2025 Wikipedia data dumps³ for the five following out-of-English Language Pairs (LPs): Hausa (eng-hau), Igbo (eng-ibo), Swahili (eng-swa), Yoruba (eng-yor), and Zulu (eng-zul).

Table 1 provides a statistical overview of this initial dataset. While Hausa and Igbo represent the largest corpora, a more telling metric is the Levenshtein character-level edit distance (Levenshtein, 1965) between their MT and HE pairs. Most importantly, all five languages have a null or near-zero average segment-level difference between the AfriCOMET (Wang et al., 2024) Quality Estimation scores of their MT and HE pairs. This surprising observation—that a quality estimation metric detects little to no quality differences after human editing—is the primary motivation for our deeper analysis.

²<https://github.com/hibaeloirghi/Wiki-Data>

³<https://dumps.wikimedia.org/other/contenttranslation/>

LP	Edit Dist. (MT vs. HE)	AfriCOMET (MT)	AfriCOMET (HE)	AfriCOMET Diff	Token Diff
eng-hau (n = 177,387)	40.39	0.65	0.66	0.01	19.34
eng-ibo (n = 200,161)	89.35	0.59	0.58	0.00	38.71
eng-swa (n = 10,100)	93.50	0.72	0.72	0.00	-7.41
eng-yor (n = 9,988)	91.97	0.56	0.58	0.02	24.45
eng-zul (n = 7,319)	40.77	0.67	0.66	-0.01	-3.01

Table 1: Global statistics comparing machine translation (MT) and human-edited (HE) Wikipedia segments across five African languages. n denotes the number of parallel segments found in the 06/13/2025 Wikipedia dump. Global AfriCOMET (Wang et al., 2024) Difference refers to the difference between the HE and MT AfriCOMET scores. Global Edit Distance represents the Levenshtein edit distance (Levenshtein, 1965) between the MT and HE pairs. Global Token Difference refers to the word-count difference between the MT and HE pairs.

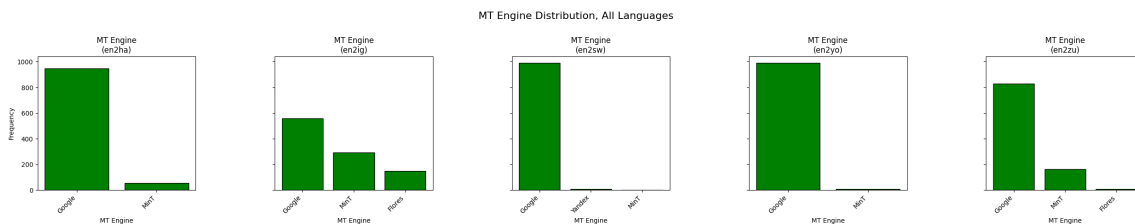


Figure 1: Distribution of MT engines used for out-of-English MT in the selected 5,000 segments for all five African languages.

2.2 Curation of the Annotation Dataset

From the large corpora described in §2.1, we curated a subset of 1,000 representative source-MT-HE triplets per language for annotation and analysis. To ensure the selected segments were informative and suitable for a detailed comparison, we applied the following filtering criteria:

- **Bounded Quality Difference:** The quality score differences between HE and MT must be between -0.5 and 0.5, as measured by AfriCOMET QE. We exclude segments with zero differences to avoid pairs with minimal variation.
- **Meaningful Edit Distance:** The character-level Levenshtein distance must be between 5 and 40 to filter out both cases with insignificant changes (e.g., punctuation edits) and complete re-translations which are hard to compare directly.
- **Sufficient Source Length:** The English source segments must contain at least 20 words to ensure segments are substantial enough to annotate.
- **High Absolute Quality:** The AfriCOMET QE score for both HE and MT must be greater than 0.5 to focus the analysis on higher quality translations.

- **Data Cleaning:** We remove duplicate segments and segments with excessive special characters (e.g. !]>*), which often signal lower quality segments.

We observe a diverse mix of MT engines in the source data (including Google Translate, NLLB, and potentially others), as illustrated in Figure 1.

2.3 Annotation Protocol

We used a customized version of the Appraise tool (Federmann, 2018; Kocmi et al., 2024)⁴ to deploy our annotation protocol⁵. We recruited three native speakers per language through Masakhane⁶ to annotate the same segments. Crucially, the identities of the candidates (MT vs. HE) were masked on the annotation interface. Annotators were asked to select the better translation and justify their choice using a fine-grained span-level mapping including labels for *Fluency*, *Adequacy*, and *Explicitation*. Appendix A shows a collage of screenshots of our annotation interface.

Inter-annotator Agreement Table 2 details overall high Fleiss’ κ inter-annotator agreement (IAA) scores for the aggregate preference task. IAA is

⁴<https://github.com/AppraiseDev/Appraise>

⁵<https://github.com/hibaeloirghi/Appraise-wiki>

⁶<https://www.masakhane.io/>

LP	Fleiss' κ	N Segments
eng-hau	0.83	404
eng-ibo	0.72	1000
eng-swa	0.37	1000
eng-yor	0.57	835
eng-zul	0.70	1000

Table 2: Fleiss' κ (Fleiss, 1971) inter-annotator agreement scores for aggregate HE vs. MT preference across languages, considering only segments with exactly three annotators.

substantial for Hausa (0.83), Igbo (0.72), and Zulu (0.70) and moderate for Yoruba (0.57). However, IAA is lower for Swahili (0.37), indicating that judging overall translation quality was a more subjective task in this language context.

3 Findings

Our analysis of the annotated data reveals a complex and often counterintuitive relationship between MT and HE, challenging the assumption that post-editing Wikipedia content targets improvements against the original MT (see Figure 2). In Zulu, for instance, MT was preferred in 86% of cases, and in Igbo, 62%. In Yoruba, MT is preferred 27% of the time and in barely more than 4% of cases in Hausa. This variety suggests that the assumption of post-edit supremacy does not hold for the Wikipedia post-editing domain. Figures 3 and 4 summarize the distribution of factors cited for translation superiority and inferiority, respectively, across all five languages in our study. Appendix B contains a detailed key of the advantages and disadvantages selected by the annotators. We summarize findings and implications of those below.

The preference trends described above hold steady when we restrict our analysis to examples with a clear majority preference by filtering out segments where at least two out of three annotators spotted no difference between the candidate translations (selected "NoDiff" meaning no real difference in fluency or meaning between translation candidates as detailed in Appendix B). As shown in Figure 6 this filtering had a significant impact on the number of remaining segments, where a large number of post-edits are assessed as having no effect on quality.

3.1 Why is a translation better?

Our analysis reveals that **fluency** is overwhelmingly the dominant driver of annotator preference.

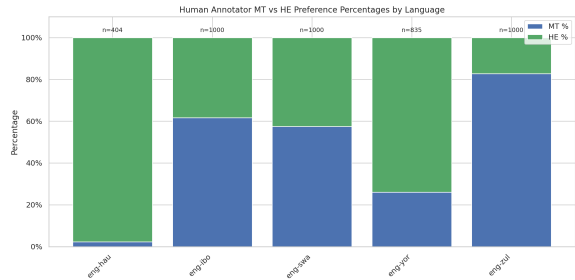


Figure 2: Human annotator MT vs. HE preference percentages. For Igbo, Swahili, and Zulu, annotators mostly prefer MT.

In all languages, the most frequently cited factors for preferring one translation are "Fluency_Natural" and "Fluency_Grammar_Spelling." Interestingly, machine translation (MT) outputs often outscore human post-edits on fluency grounds in Zulu, Swahili, and Igbo. In contrast, human-edited output tends to surpass MT in Yoruba and Hausa for fluency.

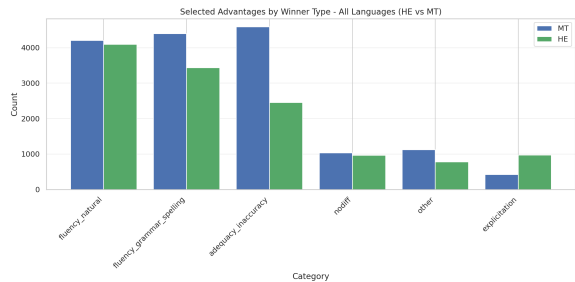


Figure 3: Distribution of justifications for the preferred translation. Fluency dominates the decision-making process.

Adequacy plays a complex role. Contrary to common assumptions, MT is also frequently cited as more adequate than the human post-edits—in all languages except Hausa. This suggests that in some cases, human interventions introduce errors or omit important information that MT maintains. Explicitation (the addition of beneficial context or clarifications) emerges as a notable factor, especially in Hausa, Swahili, and Yoruba. Finally, segments with "NoDiff" (no real difference in fluency or meaning between candidates as detailed in Appendix B) form a substantial subset in most languages, except for Igbo, where differences are more often perceived.

These patterns, visualized in Figure 3, underscore that *language-specific strategies may be necessary to improve translation workflows*. They also suggest that some post-edits may be superficial or

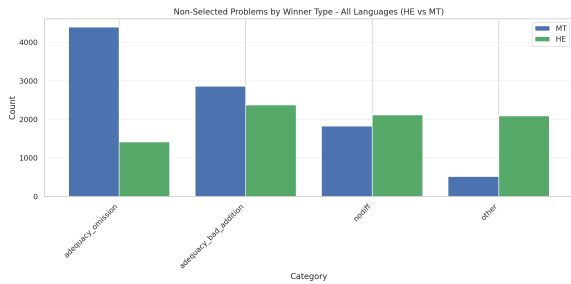


Figure 4: Distribution of factors cited by annotators for why a translation is judged worse than its counterpart. Post-edits often suffer from omissions or bad additions.

even detrimental to translation quality, particularly when editors make changes that reduce fluency or adequacy.

3.2 Why is a translation worse?

When annotators identify why a translation is inferior, a distinct language-dependent error profile emerges (Figure 4). Human editors are most frequently identified as omitting information present in the source, a pattern that is striking in Igbo, Swahili, and Zulu. In addition to omissions, human editors also sometimes add or elaborate information not present in the source text.

By contrast, the primary weaknesses of MT translations are less often tied to adequacy errors; more frequently, annotators cite general quality issues (“Other” or “NoDiff”). This indicates that, while MT typically remains closer to the original content, its output sometimes lacks the refinements or contextual adaptations made by human editors, though such adaptations may not always improve fidelity.

Collectively, these results suggest that MT is generally more **faithful to the source**, while humans are more likely to commit major errors per established MQM categories, particularly omissions or problematic additions. This raises important questions: Do human editors sometimes aim for goals beyond faithful translation, such as localization, summarization, or re-writing? If so, should such edits be distinguished in Wikipedia workflows (e.g., with a tag different from the standard “HE” marker), and how should downstream systems recognize and support these distinctions?

4 Recommendations and Implications

Our analysis strongly suggests that “post-editing” MT content on Wikipedia is not a monolithic task of error correction. Editors often engage in more

complex activities that go beyond faithful translation, such as adding or removing content.

While these adaptations can be valuable, our analysis shows they are also risky as they can degrade fluency. We offer the following actionable insights for the Wikipedia and the African NLP communities:

- **Distinguishing Editorial Roles:** Post-editing workflows could benefit from distinguishing between different types of post-edits. The workflow could be enhanced by allowing editors to tag their intent, for example, distinguishing a “Faithful Correction” from a “Cultural Adaptation”. This distinction is critical for quality control; otherwise, valuable adaptations may be incorrectly flagged as translation errors, and the true intention behind the editor’s work is lost.
- **Develop Language Specific Strategies:** The clear differences in editing patterns and fluency outcomes between languages (as shown in Figures 7 and 8) suggest that a one-size-fits-all approach to quality control is suboptimal. Workflows need to be adapted to the specific needs of each language community.
- **Implement Fluency Checks:** This study reveals that human edits can, paradoxically, decrease the fluency of machine-translated text, an issue mainly seen in Igbo, Swahili, and Zulu. To address this, the workflow could integrate lightweight automated checks that flag potentially awkward or ungrammatical sentences in post-edited content.

5 Conclusion

Our paper investigated native-speaker human preference between machine translation and human post-editing in the Wikipedia domain for five African languages: Hausa, Igbo, Swahili, Yoruba, and Zulu. We discovered that native speaker preferences are mainly driven by perceived fluency, and human edits can have varying effects on fluency of MT outputs. These findings highlight the need for better editor training and more domain-specific, human-preference-aligned automated metrics to support the creation of Wikipedia articles in low-resource Wikipedia.

Limitations

Our study focuses on five African languages; results may not generalize to other low-resource languages with different editing communities. Additionally, our analysis relies on the judgments of three annotators per language. While we filtered for majority agreement, subjective preferences regarding "naturalness" can vary by dialect and region. Finally, we did not explicitly model the intent of the editors (e.g., distinguishing between vandalism, partial edits, and genuine corrections), which adds noise to the "Human Edit" class.

Ethical Considerations

The annotators we hired are all native speakers of the five relevant African languages. We envision a participatory approach to MT evaluation and aim for our work to help expand the field's interest and understanding of MT for low-resource languages, with the ultimate hope that this will benefit the language communities of the five languages examined in this work: Hausa, Igbo, Swahili, Yoruba, and Zulu.

Acknowledgements

The authors are grateful to Eleftheria Briakou for her contributions, and to Vilém Zouhar and Tom Kocmi who answered our questions on the Appraise open source code.

References

- Christian Federmann. 2018. [Appraise evaluation framework for machine translation](#). In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 86–88, Santa Fe, New Mexico. Association for Computational Linguistics.
- Joseph L. Fleiss. 1971. [Measuring nominal scale agreement among many raters](#). *Psychological Bulletin*, 76:378–382.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xCOMET: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. [Error span annotation: A balanced approach for human evaluation of machine translation](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.
- Vladimir I. Levenshtein. 1965. [Binary codes capable of correcting deletions, insertions, and reversals](#). *Soviet physics. Doklady*, 10:707–710.
- Arle Lommel, Aljoscha Burchardt, Maja Popović, Kim Harris, Eleftherios Avramidis, and Hans Uszkoreit. 2014. [Using a new analytic measure for the annotation and analysis of MT errors on real data](#). In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation*, pages 165–172, Dubrovnik, Croatia. European Association for Machine Translation.
- 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, Oreaoghene 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. 2020. [Masakhane – machine translation for africa](#). *Preprint*, arXiv:2003.11529.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgo, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Hassan Ayinde, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Njoroge Kiragu, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Ochieng', Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, Yao Lu, and Pontus Stenertorp. 2024. [AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Mexico City, Mexico. Association for Computational Linguistics.

A Appendix: Annotation Screenshots

B Appendix: Annotation Mapping Categories

B.1 Aggregate Annotation Mapping

Why is the selected translation better? (check all that apply)

- **Fluency_Natural:** Reads more naturally in the target language, regardless of meaning preservation.
- **Fluency_Grammar_Spelling:** Contains fewer grammatical or spelling errors (typos, punctuation mistakes, etc.), regardless of meaning preservation.
- **Adequacy_Inaccuracy:** Preserves factual information correctly (dates, numbers, proper nouns, etc.) compared to the unselected translation.
- **Explicitation:** The selected translation adds helpful context or clarifications beyond what is present in the source text. The non-selected translation does not.
- **NoDiff:** No real difference in fluency or meaning.
- **Other:** Other (please specify)

What problems does the non-selected translation have? (check all that apply)

- **Adequacy_Bad_Addition:** Inaccurately introduces words/phrases not in the source text.
- **Adequacy_Omission:** Omits information present in the source text.
- **NoDiff:** No real difference in fluency or meaning.
- **Other:** Other (please specify)

Which of the two candidate translations is adequate for a Wikipedia entry, even if it is not a perfect translation? (check all that apply)

- **Wiki_Style_selected:** The selected translation.
- **Wiki_Style_non_selected:** The non-selected translation.
- **Neither.**

B.2 Span Annotation Mapping

Why did you select this span? Please select one or more options below to explain your choice and share any additional thoughts.

- **Fluency_natural:** The selected span reads more naturally in the target language.
- **Fluency_grammar_spelling:** The selected span contains fewer grammatical or spelling errors (typos, punctuation mistakes, etc.).
- **Adequacy_inaccuracy:** The selected span preserves factual information correctly (dates, numbers, proper nouns, etc.) compared to the non-selected span.
- **Adequacy_untranslated:** The non-selected span is partially or fully untranslated.
- **Explicitation:** The selected span adds helpful context or clarifications beyond what is present in the source text. The non-selected span does not.
- **NoDiff:** No real difference in fluency or meaning.
- **Other:** Other (please specify)

C Appendix: Filtered Dataset

Osibona had previously worked as a shoe salesman, and had developed properties at Albion Drive, Hackney, London, in Atlanta, Georgia, and near Johannesburg, South Africa.[4] He was an evangelist and a member of the Celestial Church of Christ.[4] Osibona was educated at Mayflower School, Ikenne, and then took an HND in business and finance, reportedly at Croydon University[note 2] in the UK.[4]

— Source text

Which of the two candidate texts below most accurately and fluently convey the original meaning of the source text above in the target language? Simply put: which candidate translation do you prefer?

Osibona ti **gege** bi oniṣowo **bata tele**, o si ti ni idagbasoke awọn ohun-ini ni Albion Drive, Hackney, London, ni Atlanta, ati nitosi Johannesburg, South Africa. [1] Ó jẹ aḣhinrere àti ọmọ egbẹ́ kan ti lẹ́ọ Celestial ti Kristi . [1] Osibona ti kọ ẹ̀kọ̀ ni Ile-iwe Mayflower, Ikenne, ati lẹhinna gba HND ni iṣowo ati iṣuna, ti a sọ ni Croydon University [note 2] ni UK. [1]

I prefer this translation

Osibona ti **ṣiṣẹ́ tẹlẹ** bi oniṣowo **bata**, o si ti ni idagbasoke awọn ohun-ini ni Albion Drive, Hackney, London, ni Atlanta, **Georgia**, ati nitosi Johannesburg, South Africa. [1] Ó jẹ aḣhinrere àti ọmọ egbẹ́ kan ti lẹ́ọ Celestial ti Kristi . [1] Osibona ti kọ ẹ̀kọ̀ ni Ile-iwe Mayflower, Ikenne, ati lẹhinna gba HND ni iṣowo ati iṣuna, ti a sọ ni Croydon University [note 2] ni UK. [1]

I prefer this translation

(a) MT vs. HE Preference Annotation

Why is the selected translation better? (check all that apply)

- Reads more naturally in the target language, regardless of meaning preservation.
- Contains fewer grammatical or spelling errors (typos, punctuation mistakes, etc.), regardless of meaning preservation.
- Preserves factual information correctly (dates, numbers, proper nouns, etc.) compared to the non-selected translation.
- The selected translation adds helpful context or clarifications beyond what is present in the source text. The non-selected translation does not.
- No real difference in fluency or meaning.
- Other (please specify)

What problems does the non-selected translation have? (check all that apply)

- Inaccurately introduces words/phrases not in the source text.
- Omits information present in the source text.
- No real difference in fluency or meaning.
- Other (please specify)

Are the candidates appropriate content for a Wikipedia article? (check all that apply)

- The selected translation.
- The non-selected translation.
- Neither.

(b) Aggregate MT vs. HE annotation

Span Annotation

For each highlighted difference between the two candidate translations above, please select the option you prefer.

Difference 1

- gege** (from the first candidate translation)
- ṣiṣẹ́ tẹlẹ** (from the second candidate translation)
- No meaningful difference

Why did you select this span? Please select one or more options below to explain your choice and share any additional thoughts.

- The selected span reads more naturally in the target language.
- The selected span contains fewer grammatical or spelling errors (typos, punctuation mistakes, etc.)
- The selected span preserves factual information correctly (dates, numbers, proper nouns, etc.) compared to the non-selected span.
- The non-selected span is partially or fully untranslated
- The selected span adds helpful context or clarifications beyond what is present in the source text. The non-selected span does not.
- No real difference in fluency or meaning.
- Other (please specify)

(c) Span-level annotation

Figure 5: Screenshots from one sample Yoruba segment annotation page of the annotation interface. Screenshot (a) shows the. Screenshot (b) shows two simplified MQM-style (Lommel et al., 2014) questions about the advantages and disadvantages of the candidate translations in (a). Screenshot (c) shows pairwise difference-level simplified MQM-style annotation.

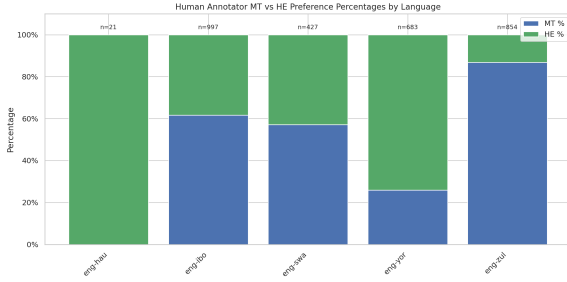


Figure 6: Human annotator MT vs. HE preference percentages. For Igbo, Swahili, and Zulu, annotators mostly prefer MT. Segments where at least two out of a total of three annotators marked that they see no difference between the two aggregate candidate translations are dropped.

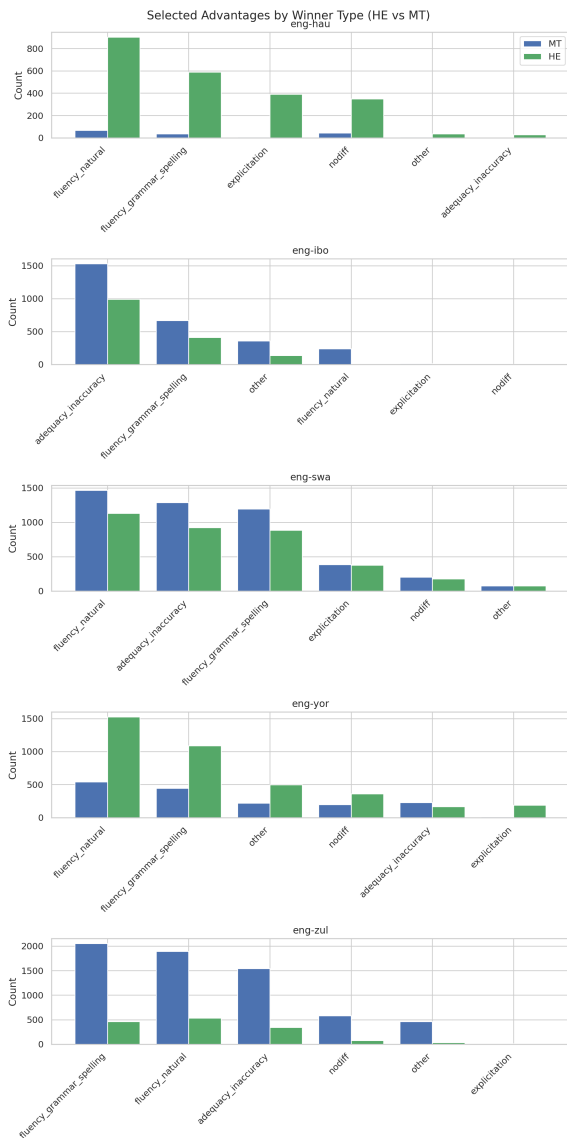


Figure 7: Distribution of justifications for the preferred translation for each language. Fluency dominates the decision-making process.

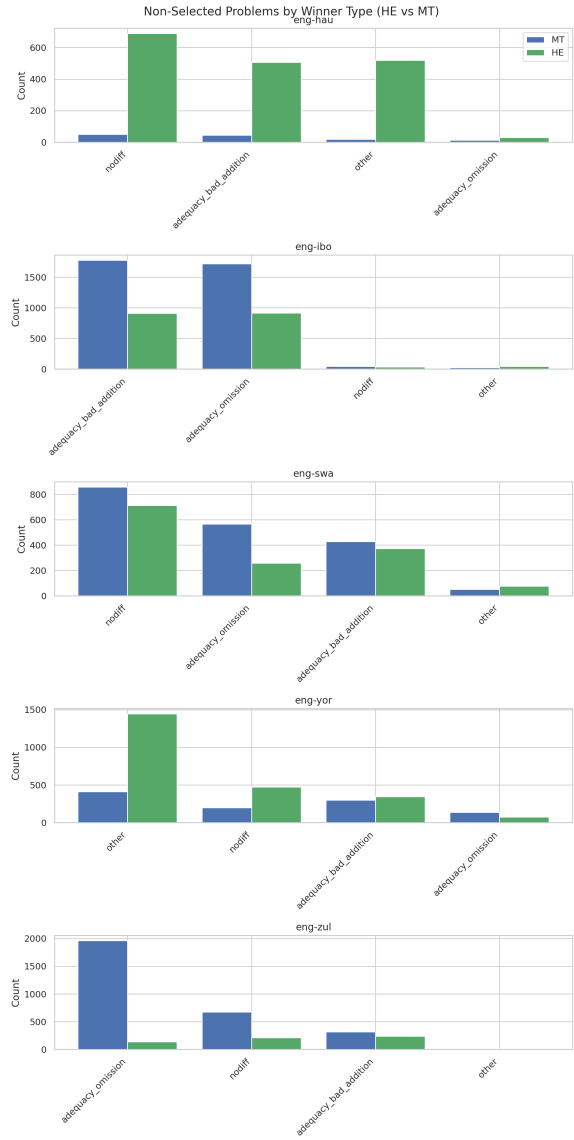


Figure 8: Distribution of factors cited by annotators for why a translation is judged worse than its counterpart for each language. Post-edits often suffer from omissions or bad additions.

Building a Conversational AI Assistant for African Travel Services with LLMs and RAG

Grace Kevine Tadaha Ngoufo¹
Shamsuddeen Hassan Muhammad², Kevin Jeff Fogang Fokoa¹

¹AIMS Cameroon, ²Imperial College London

Correspondence: kevine.tadaha@aims-cameroon.org, s.muhammad@imperial.ac.uk

Abstract

Travel agencies in many African countries face increasing pressure to handle large volumes of customer inquiries with limited staff or, either non-existent or outdated rule-based chatbots. To address this challenge, we develop a conversational virtual assistant powered by a Large Language Model (LLM) and enhanced with a Retrieval-Augmented Generation (RAG) pipeline. The system combines LLM reasoning, company-specific knowledge retrieval, and real-time API (Application Programming Interface) integration to deliver accurate, context-aware responses through WhatsApp, the region's most widely used communication platform. A dedicated web interface enables staff to upload and update internal documents, ensuring that the assistant remains aligned with changing service information. Demonstrations show that the proposed solution improves response speed, enhances user experience, and reduces operational burden.

1 Introduction

The rapid evolution of artificial intelligence worldwide has led to the emergence of Large Language Models (LLMs), which demonstrate exceptional abilities in natural language understanding, coherent text generation, and task automation. These models are becoming essential tools in various domains, offering new opportunities for intelligent decision-making and human-machine interaction. However, despite this global progress, the effective adoption of artificial intelligence (AI) technologies in many African countries remains limited. Several sectors continue to show reluctance toward integrating and trusting AI systems, even though such tools could significantly simplify operational workflows and contribute to economic growth and digital transformation across the continent (Azaroual, 2024).

In African contexts, and across varying levels of digital innovation and automation adoption, AI

tools show an adoption rate of 41%, primarily focused on analytics, fraud detection and chatbots King et al., 2025. Due to restricted access to advanced technologies and infrastructure, as well as regulatory and policy gaps such as outdated regulations and insufficient data King et al., 2025, chatbots struggle to support natural and flexible conversations, often failing to deliver a satisfactory user experience. As a result, their adoption remains low, and organizations face increasing pressure to handle customer requests manually leading to service delays and operational inefficiencies.

These challenges are particularly visible in the travel industry. In this study, we focus on a travel agency offering a broad range of services, including flight booking, visa assistance, tourism packages, car rentals, and accommodation arrangements. Due to the large volume of client inquiries and limited human resources, the company often struggles to respond promptly and accurately to customer needs. This situation highlights the need for an adaptive and intelligent virtual assistant capable of managing diverse customer queries in real time while reducing the workload on agency staff.

In this paper, we address these challenges through two main system components. (1) We transform an existing rule-based chat-bot into a fully conversational AI assistant powered by a Retrieval-Augmented Generation (RAG) architecture, enabling the system to produce accurate, context-aware, and up-to-date responses grounded in the agency's verified documents. (2) We develop a dynamic and secure web platform that allows authorized staff to directly upload and update information (text or PDF form) in the chat-bot's knowledge database, ensuring that the system remains up to date, reliable, and aligned with changes in travel policies or service offerings.

In general, this work provides a practical demonstration of how LLM-based conversational systems can be effectively integrated into African service

environments, offering a foundation for future extensions such as multilingual support and deployment in other sectors.

2 Related Work

Conversational agents have been extensively studied in artificial intelligence, progressing from early rule-based systems to modern neural and LLM-driven architectures. Traditional dialog systems relied on hand-crafted rules and slot-filling approaches (Weizenbaum, 1966; McTear, 2002), which limited their flexibility and naturalness. The emergence of sequence-to-sequence models and transformer-based architectures (Vaswani et al., 2017) enabled more fluent responses, but these systems still required large task-specific datasets and often struggled to generalize beyond their training domain.

Recent advances in Large Language Models (LLMs), such as GPT-3 (Brown et al., 2020), PaLM (Chowdhery et al., 1), Llama-2/3 (Touvron et al., 2023) and Mistral (Jiang et al., 2024), have significantly improved conversational quality by leveraging large-scale pretraining. These models possess strong zero-shot and few-shot capabilities, enabling them to support open-domain and task-oriented dialogue with minimal supervision. However, LLMs are also prone to hallucinations (Maynez et al., 2020) and often produce incorrect or unverifiable information, making them unreliable for domains requiring precise factual knowledge such as travel regulations, visa requirements, and service policies.

To mitigate hallucinations, Retrieval-Augmented Generation (RAG) approaches combine LLMs with external knowledge sources (Lewis et al., 2020; Izacard and Grave, 2021). Retrieval-based augmentation has proven effective for grounding model outputs, improving factual accuracy, and ensuring up-to-date responses in dynamic domains, including customer service and information retrieval pipelines.

Within the African Natural Language Processing (NLPs) research landscape, significant progress has been made in addressing the scarcity of digital resources, datasets, and language technologies for African languages. Initiatives such as Masakhane (Nekoto et al., 2020), the AfriBERTa model (Ogueji et al., 2021), and the MasakhaNER 2.0 Africa-centric transfer learning or Named Entity Recognition (Adelani et al., 2022) efforts have contributed to multilingual NLP resources for low-

resource African languages. Nevertheless, conversational AI applications tailored to African service industries remain under-explored. Existing chatbots deployed in African contexts often rely on rule-based or template-driven designs (Marone and Mbengue, 2025), leading to rigid interactions and limited scalability.

Despite the growing availability of African NLP resources, very few studies focus on LLM-driven conversational assistants for industry-specific workflows, such as travel agencies, tourism or customer support. Previous work on domain-specific assistants focuses mainly on general-purpose RAG pipelines (Gao et al., 2023) or enterprise knowledge systems, but does not address the unique infrastructure, data availability, or adoption challenges faced in African markets.

Our work contributes to filling this gap by demonstrating a practical use case of LLM-driven, RAG-based conversational assistance in an African travel agency context. Unlike previous rule-based systems used locally, we build a system capable of natural conversation, grounded retrieval, and dynamic knowledge updates through a dedicated web platform.

3 System Architecture

This section presents the architecture of the proposed conversational chat-bot, designed to improve the existing system with a more adaptive and scalable approach. The system integrates an LLM for natural and fluid conversation, a RAG pipeline using a vector database and an embedding model to provide to the bot precision on company-specific knowledge, external API (Application Programming Interface) for real-time information such as flight options, and Meta Webhook integration to enable conversations via WhatsApp.

3.1 Data Sources and Knowledge Database

To ensure the delivery of accurate and reliable information over time, our chat-bot relies on three primary data sources.

General Knowledge of the LLM One of the core sources of information used by our chat-bot is the general knowledge embedded in our LLM. Large language models are trained on vast and diverse corpora, including publicly available web documents, code, images, audio, video, and more; giving them broad world knowledge, linguistic patterns, and strong reasoning capabilities (Google

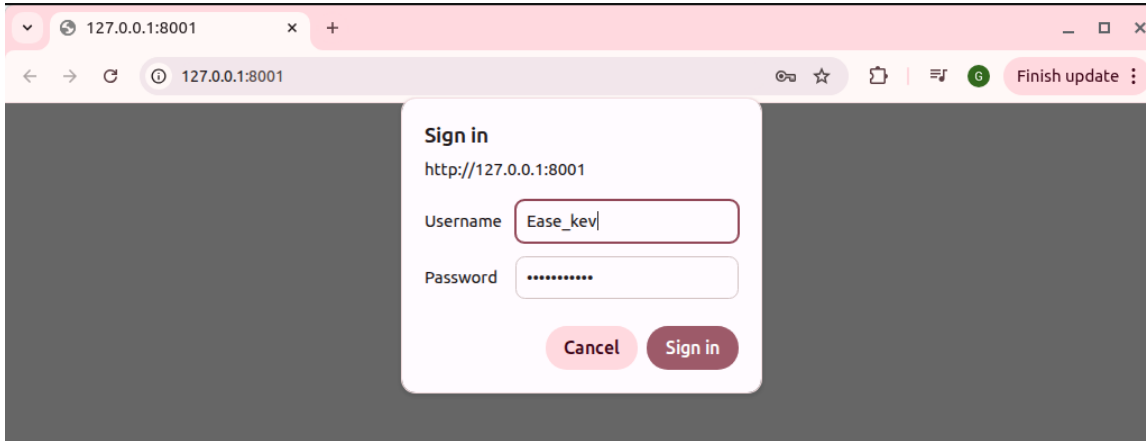


Figure 1: **Initial Login Interface:** The administrator accesses the system by entering authentication credentials, including a username and password, in order to securely log in and manage system functionalities.

DeepMind, 2024a). This general-purpose knowledge forms the foundational layer of our system and enables the model to provide coherent, contextually relevant responses even when domain-specific information is not available externally.

In this study, we use Gemini-2.5-flash, accessed through its official API, as the primary LLM powering the chat-bot. Gemini-2.5-flash is part of the Gemini family, developed by Google Deep-mind. It is designed for efficient, real-time inference, making it well-suited for interactive applications such as chat-bots (Google DeepMind, 2024b). The use of langchain community provide to the system an internal memory which help to keep the history of the conversation per user (Community, 2023).

To adapt the general knowledge of the LLM to the specific context of our application, namely the company domain, internal data, and user needs, we rely on prompt engineering. Prompt engineering uses carefully designed instructions, context templates, and query formulations to guide the LLM toward producing outputs aligned with the desired domain and style. This method has been widely studied and shown to significantly enhance the performance of LLMs across many tasks without changing their internal weights (Li et al., 2023).

External API (Application Programming Interface) An API is a set of rules and specifications that enables different software systems to communicate and exchange data and functionalities (Postman, 2024). In our work, we use several APIs provided by the company, particularly for retrieving IATA (International Air Transport Association) airport codes, fetching real-time flight availability based on the destination and flight type (round

trip or one-way), and other operational information. The integration of these APIs with the LLM is achieved through the Model Context Protocol (MCP), an architecture primarily composed of a client–server interaction model that facilitates secure and efficient communication between the chatbot and external services (Protocol, 2023).

Specific Knowledge of the Company Knowing that LLMs only provide general knowledge, it is essential to supply them with company-specific information such as organizational activities, available products, and up-to-date service details. To address this need, we developed a dedicated web interface for staff members (Figure 2). Access to this platform is protected by an authentication mechanism (Figure 1), ensuring that only verified employees can upload, update, or manage internal company documents.

Through this secure interface, employees can upload or update internal documents, including text files and PDFs, containing information relevant to the company’s services. Once uploaded, the documents are automatically processed, embedded, and stored in a vector database. This controlled access helps protect sensitive company information and prevents unauthorized disclosure or manipulation of internal data, while ensuring that the chatbot consistently provides accurate and up-to-date responses during user interactions.

In addition, the system supports document management functionalities, including the deletion of outdated or irrelevant documents and the computation of statistics to track the volume of documents available in the database. These features further contribute to data governance and privacy by allowing

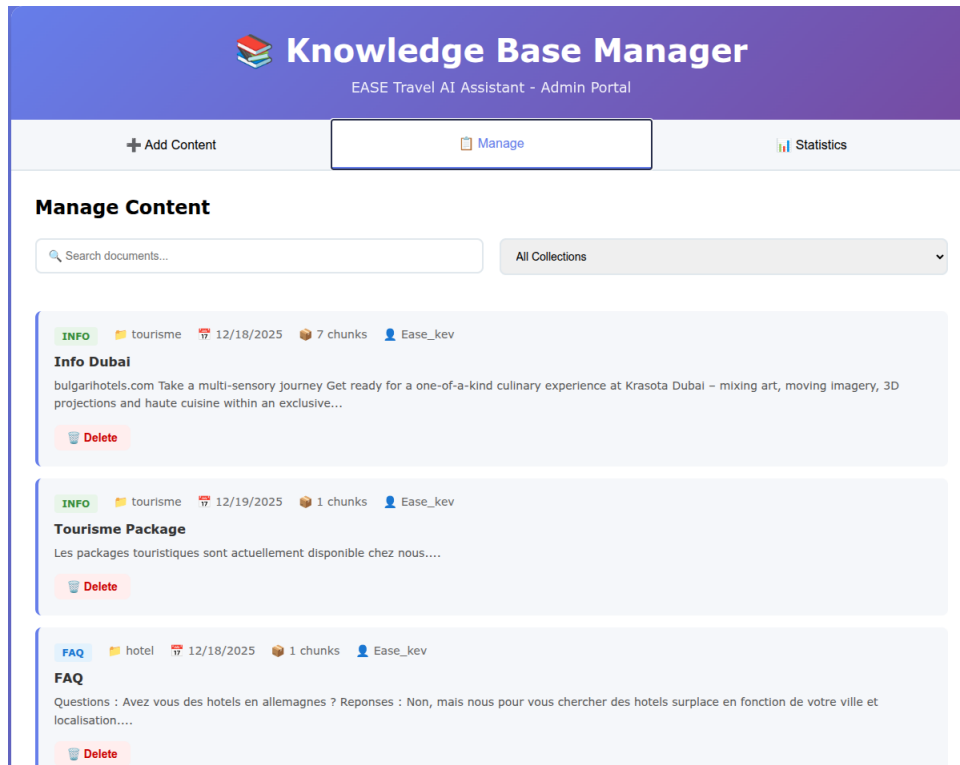


Figure 2: **Web Knowledge Base Manager:** Web-based interface for data management, allowing authorized staff to upload content, manage existing data, and view vector database statistics.

administrators to maintain only relevant and current information within the system.

To support this retrieval process, we rely on a vector database, a system that stores data in the form of high-dimensional numerical vectors. This structure enables efficient similarity search, allowing the chat-bot to retrieve the most relevant information based on semantic meaning rather than simple keyword matching. Several vector databases exist (e.g., Pinecone, Weaviate, Milvus), each with distinct capabilities. In this study, we employ Qdrant, an open-source vector database optimized for high-performance vector similarity search and scalable AI applications (Team, 2023). Qdrant was selected because it offers strong support for semantic search, delivers fast query performance, and remains free and user-friendly, making it well-adapted to our deployment constraints.

Before being stored in Qdrant, textual data must be transformed into vectors using an embedding model. This model converts text into numerical representations that preserve semantic meaning, allowing the system to compare and retrieve documents based on their conceptual similarity. For this purpose, we use the BGE-M3 embedding model, which supports dense and multi-vector retrieval as well as multilingual processing (Chen et al., 2024; Xiao et al., 2024). Our choice is motivated by its balance of accuracy and computational efficiency:

the model is free, lightweight, and capable of generating rich embeddings that capture nuanced semantic relationships. These characteristics are essential for ensuring high-quality semantic search and, ultimately, improving the chat-bot’s ability to deliver reliable, context-aware answers.

3.2 System Implementation

The full system implementation is designed around a hybrid architecture combining generation, retrieval, and real-time API interaction. Figure 3 shows an overview of the workflow, and the main components are described below.

User’s Interface. To enable real-world deployment and effectively meet the operational needs of the travel agency, the chat-bot has been integrated with WhatsApp, which is the most widely used communication platform in the region. This choice ensures that users can interact with the system through a familiar and accessible interface, minimizing barriers to adoption.

The integration follows a structured workflow: When a user sends a message to the agency’s WhatsApp Business number, the message is first forwarded by the Meta Developer Webhook to the backend server. The backend then processes the query, leveraging the chat-bot’s natural language understanding and retrieval capabilities to generate

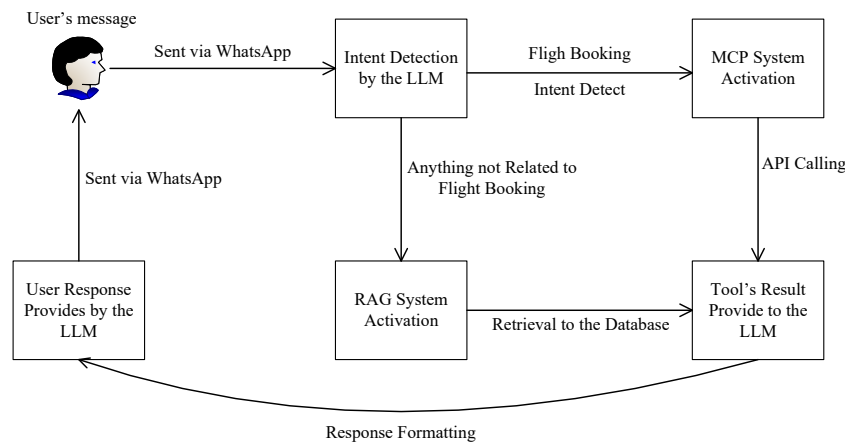


Figure 3: **System Architecture:** Our hybrid conversational assistant workflow combining intent-based routing, Multi-Component Processing (MCP), and Retrieval-Augmented Generation (RAG). User messages received via WhatsApp are first classified by the LLM; flight booking intents trigger API calls, while general travel queries activate the RAG system to retrieve knowledge from the database.

an appropriate response. Once generated, the response is transmitted back to the user through the WhatsApp Cloud API, completing the interaction cycle. This architecture not only guarantees real-time communication but also ensures scalability, reliability, and ease of maintenance. By leveraging a widely adopted messaging platform, the system provides a seamless and intuitive user experience while supporting the agency’s operational objectives and enhancing customer engagement.

Backend Service Once a user message is captured and forwarded to the backend server, it is first processed by the Gemini model, which is responsible for interpreting the message and performing intent detection. If the query is general in nature, for instance, greetings or requests for broad, non-specific information, the model responds directly using its pre-trained knowledge. However, if the query pertains to the company’s services, the system activates one of two specialized orchestration pathways.

For flight-ticket booking intents, the Model Context Protocol (MCP) system is triggered. In this case, the external flight-booking API is not invoked immediately. Instead, the Gemini model initiates a clarification dialogue to gather all essential booking parameters from the user, including destination, number of travellers, travel dates, and flight type. Once the information is complete, the MCP system retrieves the corresponding results. These results are then passed back through the Gemini model, which reformats, verifies, or restructures the output as needed before producing the final user-ready response.

For all other service-related intents not involving flight booking, the system activates a Retrieval-Augmented Generation (RAG) pipeline. In this workflow, the embedding of the user query is compared against stored document embeddings within the company’s vector database. The most relevant information is retrieved (result from the RAG retrieval) and supplied to the Gemini model alongside the original query. The model then integrates this domain-specific context with its generative capabilities to produce a precise, consistent, and contextually grounded response.

The entire orchestration is governed by carefully engineered prompts that direct the model’s behaviour across different scenarios. This design is further strengthened by few-shot prompting strategies, which provide the model with concrete examples of expected behaviour, thereby enhancing its ability to generalize across diverse queries while ensuring accuracy, consistency, and alignment with the company’s operational requirements.

4 Usability and Analysis

Travel companies often face a high volume of customer inquiries while lacking sufficient customer service staff, leading to long waiting times and customer dissatisfaction. Existing procedural chatbots are limited by rigid, step-by-step interactions that can be slow and cumbersome. To address this issue, we propose a conversational chatbot that allows customers to interact naturally via WhatsApp and receive immediate, relevant responses without human intervention. The system was developed and deployed in collaboration with a real travel

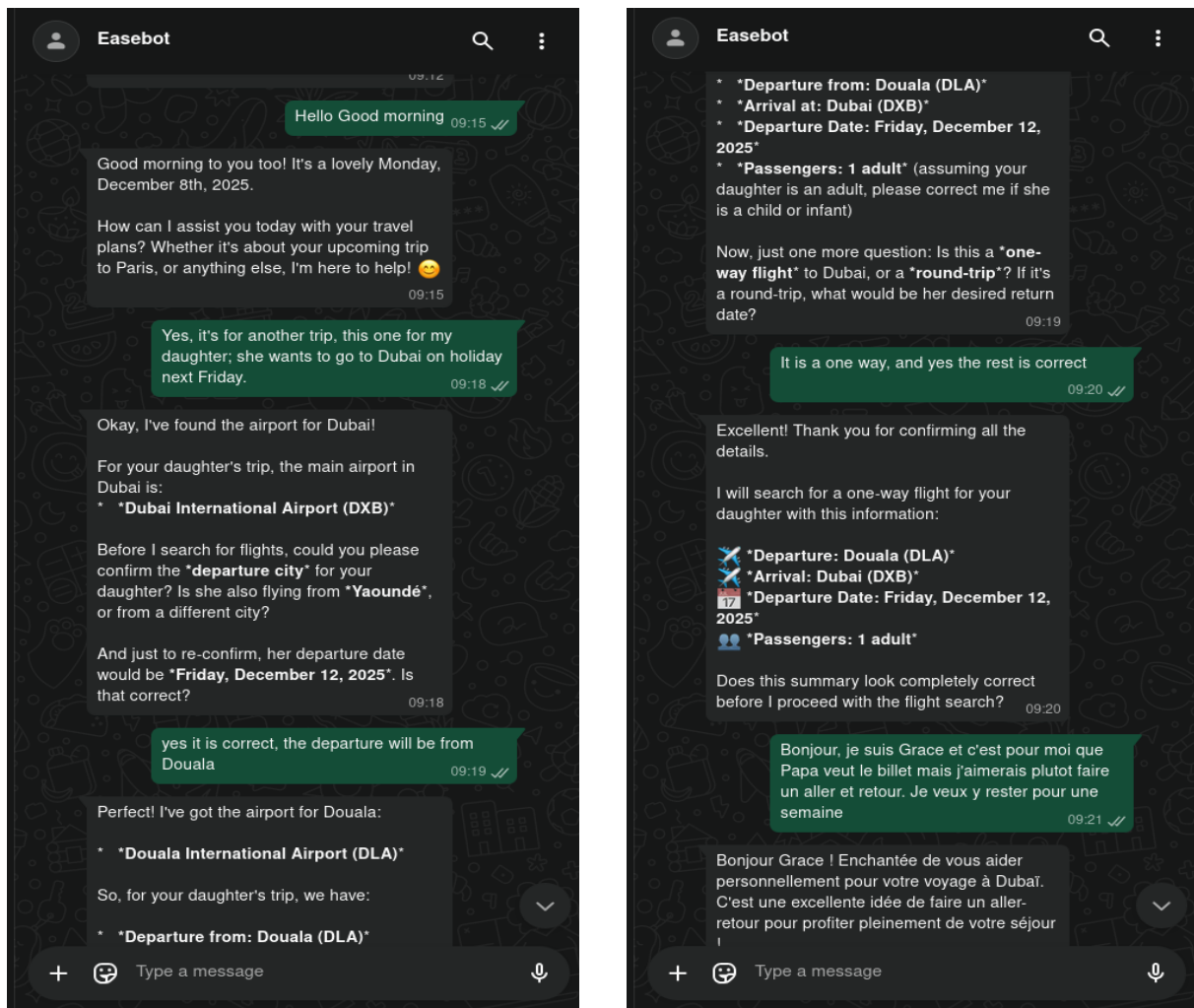


Figure 4: **Dialogue State Management:** Example showing the assistant’s handling of complex conversational dynamics. The system manages a context shift (father to daughter), corrects initial assumptions (one-way to round-trip), confirms multiple parameters through clarification questions, and seamlessly switches languages (English to French) while maintaining contextual coherence.

agency experiencing these challenges and was designed to handle all customer inquiries related to the agency’s services, including bookings, pricing, schedules, policies, and general support. This real-world deployment demonstrates the practical applicability of the solution and its potential to improve response times, enhance customer service quality, and reduce operational burden for travel agencies.

Figures 4 and 5 illustrates a typical interaction with the proposed chat-bot. When a user restarts a session, the system retains the context of previous conversations and offers the option to resume where the dialogue left off. After the required information is gathered, the chat-bot initiates a flight search only once the user explicitly confirms that the collected details are correct. In addition, the system seamlessly adapts to the user’s language primarily (English and French), which are among

the most widely spoken languages in many African countries. The response time is also remarkably fast, giving the impression of an almost real-time conversation with a human agent.

The example presented in this paper highlights the chat-bot’s ability to remain aligned with the intended domain whether the query concerns flight booking or any other service offered by the company, while delivering a genuinely conversational and context aware user experience.

Conclusion and Future Work

In this work, we present a conversational chat-bot powered by a large language model (LLM), equipped with a Retrieval-Augmented Generation (RAG) system, and designed for a travel agency offering a wide range of services. The virtual as-

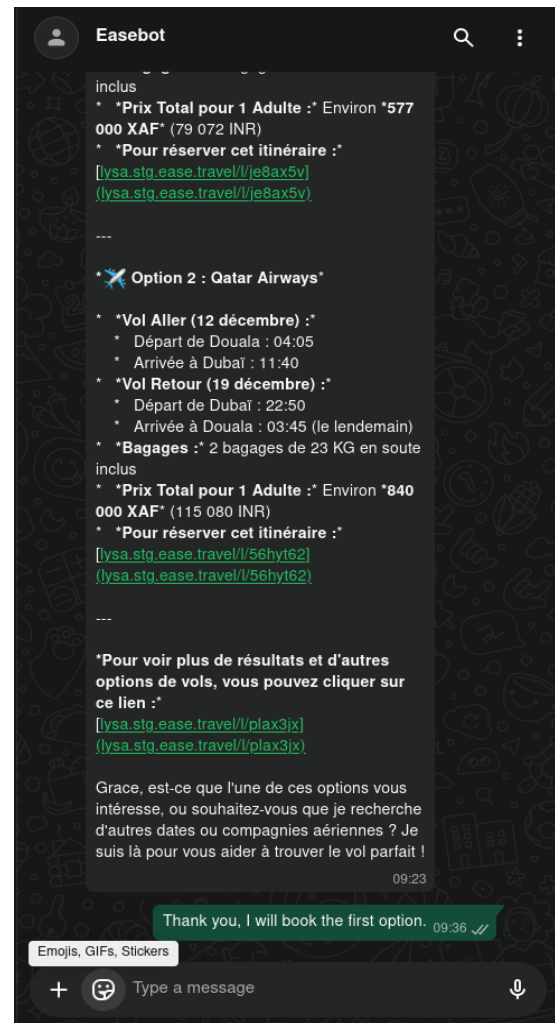
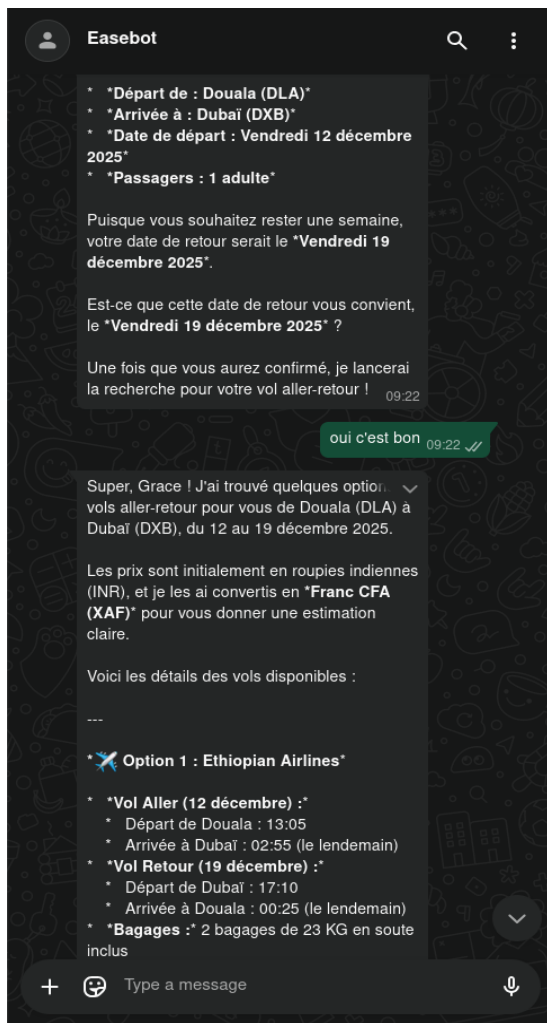


Figure 5: **Example Conversation Flow:** An illustrative WhatsApp exchange demonstrating the assistant’s flight booking capability. The system extracts travel parameters (origin, destination, dates), proposes a return date, presents multiple flight options with converted pricing (INR to XAF), and provides booking links.

sistant demonstrates the ability to understand user queries, provide relevant guidance, and support clients throughout the interaction, while maintaining a smooth conversational flow thanks to its low latency and user friendly tone. For future work, we will focus on two main directions. First, we plan to develop a voice-enabled version of the chatbot. This will include integrating speech-to-text and text-to-speech models, enabling users to interact through WhatsApp voice notes. Second, we aim to integrate African languages into the system to increase accessibility and cultural relevance. We will begin with widely spoken languages such as Swahili and Hausa, and gradually expand to others, ensuring that the chatbot becomes more inclusive and better aligned with the linguistic diversity of the continent. Such a feature would significantly improve accessibility, particularly for users with

low literacy levels or those who naturally prefer voice communication. By supporting spoken interaction, the chatbot would become more intuitive and adaptable to real-world user behaviors across Africa.

References

- David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen H Muhammad, Peter Nabende, and 1 others. 2022. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508.
- Fahd Azaroual. 2024. Artificial intelligence in africa: Challenges and opportunities. *Policy Brief. PB*, pages 23–24.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie

- Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.](#) *arXiv preprint arXiv:2401.XXXXX*.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 1. others. 2022. palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.
- LangChain Community. 2023. Langchain: Building applications with llms. <https://docs.langchain.com/oss/python/langchain/>.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).
- Google DeepMind. 2024a. Gemini 2 flash model card. <https://modelcards.withgoogle.com/assets/documents/gemini-2-flash.pdf>.
- Google DeepMind. 2024b. Gemini flash. <https://deepmind.google/technologies/gemini/flash/>.
- Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pages 874–880.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Stephen King, Toni Morrison, Ernest Hemingway, F Scott Fitzgerald, and Mark Twain. 2025. Digital innovation and automation in african enterprises.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Y Li, J Chen, and X Wang. 2023. Prompt engineering for large language models: A survey. *arXiv preprint*.
- Reine Marie Ndéla Marone and Moustapha Mbengue. 2025. Chatbots and artificial intelligence to support digital university libraries in africa: Opportunities and challenges. *Digital Libraries Across Continents*, pages 72–92.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.
- Michael F McTear. 2002. Spoken dialogue technology: enabling the conversational user interface. *ACM Computing Surveys (CSUR)*, 34(1):90–169.
- Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohunbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.
- Keneth Ogueji and 1 others. 2021. Afriberta: African multilingual language model. In *NeurIPS Workshop on AfricanNLP*.
- Postman. 2024. What is an api? a beginner’s guide to apis. <https://www.postman.com/what-is-an-api/>.
- Model Context Protocol. 2023. What is the model context protocol (mcp)? <https://modelcontextprotocol.io/docs/getting-started/intro>.
- Qdrant Team. 2023. Qdrant: High-performance vector database for scalable ai applications. <https://qdrant.tech>.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Joseph Weizenbaum. 1966. Eliza—a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45.
- Jun Xiao, Sheng Zhang, Zhuosheng Zhang, Fei Huang, and Heng Ji. 2024. Bge-m3: Multi-functionality, multi-linguality, and multi-granularity embedding model. https://bge-model.com/bge/bge_m3.html.

Morphologically-informed Somali Lemmatization Corpus built with a Web-based Crowdsourcing Platform

Abdifatah A. Gedi¹ Shafie A. Mohamed¹ Yusuf A. Yusuf¹
Muhidin A. Mohamed^{1,2} Fuad M. Hassan³ Houssein A. Assoweh⁴

¹Jamhuriya University of Science & Technology, Mogadishu, Somalia; ²Aston University, UK
³Somali National University, Mogadishu, Somalia; ⁴Université de Djibouti, Djibouti

Correspondence: Gedi@just.edu.so

Abstract

Lemmatization, which reduces words to their root forms, plays a key role in tasks such as information retrieval, text indexing, and machine learning-based language models. However, a key research challenge for low-resourced languages such as the Somali is the lack of human-annotated lemmatization datasets and reliable ground truth to underpin accurate morphological analysis and training relevant NLP models. To address this problem, we developed the first large-scale, purpose-built Somali lemmatization lexicon, coupled with a crowdsourcing platform for ongoing expansion. The system leverages Somali’s agglutinative and derivational morphology, encompassing over 5,584 root words and 78,629 derivative forms, each annotated with part-of-speech tags. For data validation purpose, we have devised a pilot lexicon-based lemmatizer integrated with rule-based logic to handle out-of-vocabulary terms. Evaluation on a 294-document corpus covering news articles, social media posts, and short messages shows lemmatization accuracies of 51.27% for full articles, 44.14% for excerpts, and 59.51% for short texts such as tweets. These results demonstrate that combining lexical resources, POS tagging, and rule-based strategies provides a robust and scalable framework for addressing morphological complexity in Somali and other low-resource languages.

1 Introduction

Lemmatization is a foundational step in Natural Language Processing (NLP) which supports tasks such as information retrieval, text classification, and machine translation by reducing words to their canonical forms. It is a morphological process that converts inflected words to their base forms, also known as lemmas. For nouns, this corresponds to the singular form; for verbs, the infinitive; and for adjectives or adverbs, the positive form. Essentially, lemmatization normalizes different morphological

variants of a word by mapping them to the same underlying lemma, allowing them to be analyzed as a single term or concept. By reducing the number of distinct terms, lemmatization simplifies text and benefits downstream processing tasks. For example, in information retrieval systems, lemmatization can improve recall, as queries and documents that are morphologically normalized are more likely to match (Liu et al., 2012).

For many low-resource African languages, lemmatization is particularly challenging due to sparse annotated data and intricate morphological patterns (Adelani, 2025). Recent research has made progress through approaches such as multilingual pre-training, morphological segmentation, and limited supervised datasets, but overall performance remains inconsistent. This paper examines the state of the lemmatization task and resource for the Somali language.

Prior work on Somali lemmatization has been limited, characterized by minimal data and no integration of part-of-speech (POS) information, despite its importance for handling Somali’s complex morphology (Mohamed and Mohamed, 2023). In this paper, we expand the dataset, introduce an annotation tool to streamline labeling, and propose a hybrid approach combining rule-based methods, root derivation, lookup strategies, and POS tagging. Our evaluation demonstrates significant improvements over earlier methods, establishing a more robust foundation for future Somali NLP research in low-resource conditions.

In this study, we address the problem of lemmatization for the Somali language, aiming to develop a method for normalizing words derived from the same root. Our focus is primarily on the “MAXAA TIRI,” – the principal written dialect of Somali – which was previously explored in our previous initial study (Mohamed and Mohamed, 2023). Building on that work, we have significantly expanded the datasets in terms of the root words, their derivative

forms, and the addition of a purposeful annotation tool to facilitate the creation of high-quality linguistic resources supporting crowdsourcing and future related NLP research. The pilot lemmatizer built on the developed dataset identifies and extracts meaningful root forms from inflected variants, employing a hybrid approach that integrates lookup method with rule-based processing and providing a robust foundation for further computational processing of Somali word normalization.

The main contributions of this work are as follows:

1. First, we constructed an expanded Somali lemmatization lexicon that integrates Somali morphological rules and covers a wider range of root forms and their inflections.
2. Second, we designed and implemented an annotation tool to enable effective collaboration across the annotation team.
3. Third, we created a Somali word lemmatization algorithm built on the expanded lexicon incorporated with rule-based method
4. Fourth, we tested the lemmatizer on a Somali corpus of various lengths and domains to evaluate its performance.

2 Related Work

Text lemmatization is a fundamental NLP task, which is considered a solved research problem for high-resource languages such as English, French, and Chinese (Bergmanis and Goldwater, 2018; Manjavacas et al., 2019). However, it remains a significant research challenge for under-resourced languages like the Somali (Miletić and Siewert, 2023; Mohamed and Mohamed, 2023). Although significant progress has been made for various NLP tasks with the emergence of neural networks and transformer-based models, the development of manually annotated lemmatization datasets, such as root-derived word pairs enriched with part-of-speech tags, remains a well-established and indispensable approach for languages with complex morphology and limited digital resources, such as Somali (Sahala et al., 2023; Stanković et al., 2016; Gordin et al., 2025).

Linked with the above, several related studies have contributed to the resource development of core text normalization NLP steps including word lemmatization. For example, the recent work of

Mathayo and Kondoro (2024) on Swahili, which is a low-resourced language, introduced a large verb conjugation dataset to address its agglutinative morphology. Covering over 319,000 verb forms, this dataset facilitates essential NLP pipeline steps including lemmatization, and morphological analysis, making it a valuable resource for advancing NLP in low-resource Bantu languages. Moreover, KinyaBERT (Nzeyimana and Rubungo, 2022) demonstrates that integrating explicit morphological structure into transformer models improves performance over subword-only methods for Kinyarwanda. While multilingual neural parsing studies on Bambara, Wolof, and Yoruba (Dione, 2021) show that neural transfer approaches still rely on structured annotations such as lemmas and morphological features.

Despite the growing interest in NLP for low-resource languages, Somali remains notably under-represented in the literature. Few studies have addressed core NLP tasks or the development of language resources for the language. For instance, Mohammed (2020) investigated part-of-speech (POS) tagging using statistical and machine learning methods, achieving an accuracy of 87.51% through ten-fold cross-validation. Additionally, (Badel et al., 2023) develop an annotated corpus – a dataset consisting of 2,335 documents sourced from prominent online platforms, including Hiiraan Online, Dhacdo.net, and collections of Somali poetry – for Somali language information retrieval. Also, Nimaan et al. (2006) explored automatic speech transcription for Somali language, constructing a 10-hour audio corpus and reporting a word error rate (WER) of approximately 21%. Recently, Mohamed et al. (2025) developed two Somali datasets for fake news detection and toxicity classification sourced from the social media and labelled by human annotators. Their work has also introduced the first monolingual BERT-based Somali language model, named SomBERTa, which outperformed compared multilingual models like AfriBERTa and AfroXLMR in fake news and toxicity classification, achieving the highest average accuracy of 87.99% and highlighting promising directions for Somali NLP research. Other research studies on low-resourced languages have utilized multilingual LLMs covering Somali NLP tasks such as machine translation (Wang et al., 2024; Adelani et al., 2022) and text classification (Adelani et al., 2023; Alabi et al., 2022).

The current study follows our previous work

(Mohamed and Mohamed, 2023) which pioneered the development of a Somali lemmatization resource and has specifically addressed lemmatization for the Somali language. While this research builds on that initial study (which was released only as a non-archived preprint), it makes several significant extensions, including the expansion of the dataset from 1247 root words to 5584 words, the addition of morphological information such as part-of-speech (POS) tags, and most importantly, the development of a tailored web-based data annotation tool to facilitate community crowdsourcing and further corpus development.

3 Methodology

This section outlines the methodology employed in developing the data management platform, and the comprehensive Somali lexical database for word lemmatization paired with associated PoS tags. The research realization pipeline includes six key phases: system development, data collection, annotation, lexical database construction, crowdsourced validation, and pilot testing the constructed data in the form of lemmatization evaluation.

3.1 Lemmatization platform and database

A custom web-based annotation platform was developed to facilitate the data collection, annotation, and expansion of the dataset beyond this research via community crowdsourcing¹. Technically speaking, and drawing inspiration from similar platforms and lexical resources (Habash and Dorr, 2003), we implemented a relational database architecture in which each root word serves as a lexical anchor linked to its derived forms and their PoS tags. All data were stored using a structured SQL schema, allowing for efficient retrieval, expansion, and linkage between morphological variants. This system also ensures data quality by enforcing structural consistency, preventing duplicate entries, and standardizing POS categories while allowing annotators to systematically record derivational variants alongside their PoS tags, thereby producing reliable and high quality lexical resource (Figure 1).

This collaborative lemmatization system formalizes the division of tasks among contributors, combining linguistic expertise with scalable annotation practices. By embedding validation and verification at multiple points in the pipeline, the resulting resource achieves a balance between morpholog-

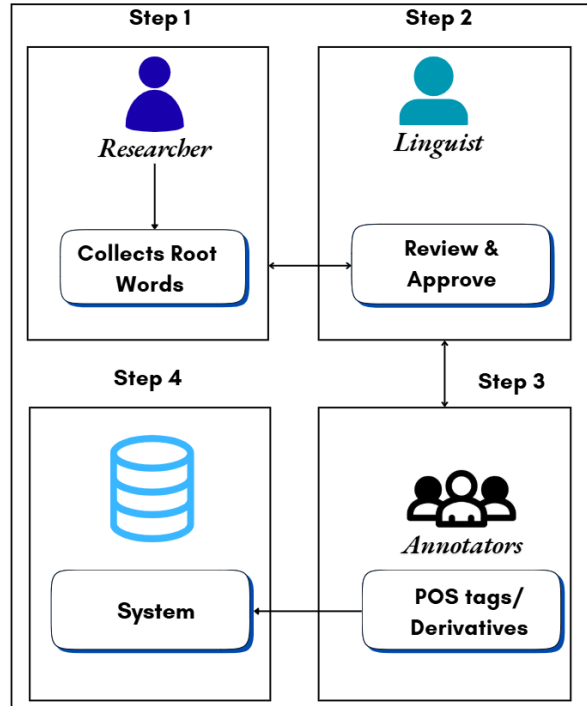


Figure 1: Data annotation and expert approval process

ical precision and usability for downstream NLP tasks such as lemmatization, POS tagging, and machine translation.

All collected and annotated data, including the compilation of root words, the derived forms, and their PoS tags were iteratively verified, with Somali language experts from the *Intergovernmental Academy for Somali Language* who conducted cross-checks of the data quality, derivational logic, the POS assignments, and overall data consistency. Discrepancies were occasionally identified during these verifications and resolved, which strengthened both the reliability and linguistic validity of the data set.

3.1.1 Core platform features

The developed platform consists of all necessary features from uploading word base forms and annotating them with their derived forms and PoS tags, to data quality moderation and statistical monitoring through tailed dashboard. The root word uploading feature (Figure 7 appendix A) serves as a central component of the system and database, enabling insertion, modification, and deletion of root words. Integrated validation mechanisms prevent duplication and enforce conformity with Somali morphological rules. Once approved, root words form the foundation for derivative generation and POS annotation, ensuring accuracy and scalability

¹<http://annot.just.edu.so/>

in the lexical resource.

The annotators then start enriching approved root words with derivative forms and assign the corresponding part-of-speech (POS) tag. As shown in (Figure 8 *appendix B*), the interface displays only validated roots, ensuring quality control. Users specify the word type, POS category, root word, and all related derivatives. This structured workflow maintains annotation consistency, prevents duplicates, and links each derivative to its validated root, creating linguistically robust lexical entries that support downstream NLP tasks such as lemmatization and morphological analysis.

The platform also includes a publicly accessible search interface and supports querying root and derived forms, filtering by POS tags, and exporting data for external analysis. For example, it enables users to explore lexical data by entering full or partial terms. The system retrieves root forms, POS tags, and derived forms, including their morphological structures. For example, a search query such as “abuu” returns all words that contain this substring, whether as a root or derived word (Figure 9). This transparent and intuitive interface facilitates open access to the lexical database, enabling linguists, NLP researchers, and language learners to analyze and explore Somali morphology effectively. The system architecture was designed for scalability, allowing integration with additional NLP tools in the future.

Finally, the system is designed to support community crowdsourcing to future expansion of the developed lemmatization resource, enhance scalability and maintain data quality (Figure 2). This would require new users and data annotators to register and undergo approval before contributing to the data expansion. And as with current data, all future submissions need to be reviewed by expert moderators through a cross-verification process to ensure consistency in POS tagging and derivation logic.

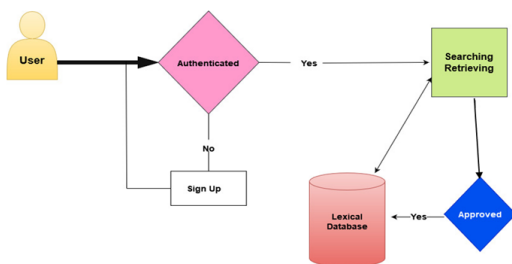


Figure 2: Platform’s crowd-sourcing feature/process

3.2 Dataset Summary

Overall, we have compiled a corpus consisting of 5,584 root words from which a total of 78,629 PoS tagged derived forms were generated. As summarized in Table 1, the nouns constitute 56.73% (44,603 words) of the corpus, while the verbs make up 43.27% (34,026 words). Furthermore, Table 2 shows fine-grained breakdown distributions for the noun and verb subcategories.

Table 1: Dataset distribution by main part of speech

Part of Speech (Qaybta hadalka)	Count	Percentage
MAGAC (Noun)	44,603	56.73%
FAL (Verb)	34,026	43.27%

4 Pilot Evaluation

In this study, we extend the approach proposed in Mohamed and Mohamed (2023), which introduced a two-stage Somali language lemmatization framework (Figure 3). Briefly, that framework began with the construction of a lexicon by manually compiling and pairing root and derived words based on defined linguistic criteria. In the second stage, written morphological rules were applied to lemmatize words that are not found in the lexicon. This combination of lexicon-based and rule-based methods were employed to pilot test and validate the usability of the lemmatization corpus created in the current work.

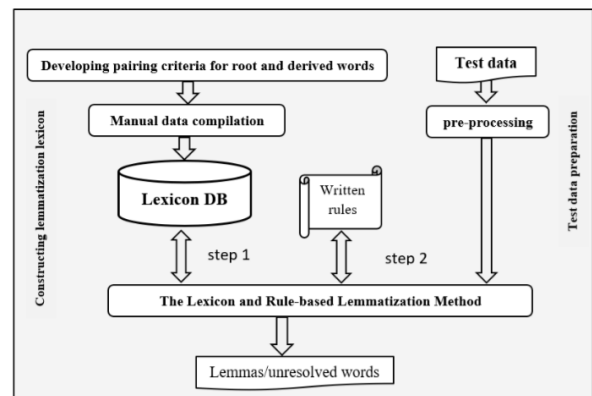


Figure 3: Lexicon and rule-based lemmatization method

Although the present evaluation focuses on rule-based methods, the developed resource is designed to support future neural and LLM-based Somali NLP systems. Large language models typically rely on subword tokenization and often struggle with morphologically rich languages when explicit morphological supervision is absent. The curated

Table 2: Distribution of Grammatical Patterns in the Somali Lexical Dataset

Part of Speech (Qaybta hadalka)	Count	Percentage
MAGAC+TILMAAN+LAHAANSHO (Noun cloalesced with possive adjective)	22,955	29.19%
FAL TAAGAN (Progressive verb)	15,115	19.22%
FAL TAGAY (Verb in past form)	15,108	19.21%
MAGAC+TILMAAN+TUSMO (Noun cluster)	6,954	8.84%
MAGAC+QODOB (Defined noun (noun with a definite article))	6,898	8.77%
MAGAC (Noun)	5,587	7.11%
MAGAC+TILMAAN+WEYDIIMO (Noun coalesced with interrogative adjective)	2,209	2.81%
FAL AMAR (Imperetive verb)	1,907	2.43%
FAL MAXADANE (Nonfinite verb)	1,896	2.41%

root-derivative pairs and POS annotations in this dataset can be used to generate supervised training data, construct evaluation benchmarks, guide morphological post-processing, and support hybrid pipelines that combine neural modeling with linguistic constraints. Thus, this work provides foundational linguistic infrastructure that complements and enables future data-driven approaches.

In particular, to evaluate the robustness of the constructed lexicon, 294 Somali text documents of varying lengths were collected, which was a diverse dataset spanning multiple high-frequency public discourse domains of Somali texts sourced from a variety of digital platforms, including social media posts, BBC-Somali, and other reputable online news outlets. The corpus was carefully categorized into 8 thematic domains that reflect the most prominent areas of public discourse in Somali society (Table 3).

Table 3: Distribution of the test dataset by domain

Category	Count	Percentage (%)
Caafimaad (Health)	32	10.88%
Ciyaaro (Sports)	45	15.31%
Diin (Religion)	34	11.56%
Waxbarasho (Education)	20	6.80%
Ganacsi (Business)	68	23.13%
Siyaasad (Political)	62	21.09%
Madadaalo (Entertainment)	14	4.76%
Tiknolojiyo (Technology)	19	6.46%

Beyond topical diversity, the dataset was further classified according to text size to capture structural and contextual variations across different communication settings. Specifically, the texts were grouped into three categories: small, medium, and long texts (full-length news articles). This stratification is essential for evaluating the lemmatization algorithm, as short texts often exhibit high lexical variability and limited context, whereas long texts present more coherent discourse structures and richer morphological forms.

Quantitative analysis of the dataset revealed that

small texts contained an average of 86.56 tokens, medium texts averaged 257.17 tokens, and long texts averaged 563.79 tokens. This clear distinction in token length validates the effectiveness of the size-based categorization and provides a robust foundation for assessing model performance across heterogeneous Somali text types. By incorporating this size-based categorization, the dataset provides a robust foundation for assessing model performance across heterogeneous Somali text types (Figure 4).

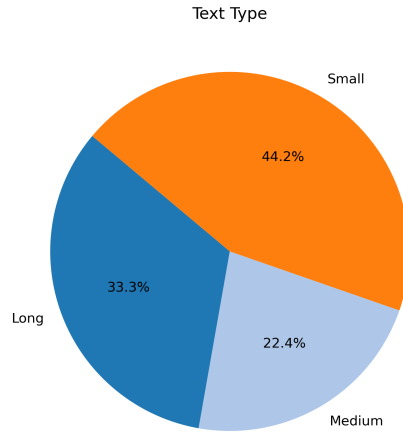


Figure 4: Document token length distributions

Following preprocessing, the corpus was normalized by removing punctuation, numerical tokens, and common stop words, leaving only cleaned textual content for analysis. This step was essential to ensure that morphological processing operated exclusively on linguistically meaningful units. The cleaned dataset was subsequently tested using our rule-based lemmatization system, which integrates lexical resources with Somali-specific morphological rules.

To contextualize the performance of the pro-

posed lexicon-and-rule lemmatization framework, we compare it against two simple baselines. The first baseline is lexicon-only lookup, where tokens are matched directly against the lexical database without applying morphological rules; unmatched tokens remain unresolved. The second baseline is identity mapping, where each token is returned unchanged as its own lemma. These baselines provide lower-bound references that help quantify the contribution of morphological rules beyond dictionary coverage.

The evaluation revealed that the system successfully lemmatized and annotated over 51 percent of the tokens, reducing them to their canonical root forms and simultaneously assigning appropriate part-of-speech (POS) tags. This dual outcome is critical for downstream NLP tasks, as it not only standardizes lexical variation but also provides syntactic and grammatical information that can improve applications such as text classification, machine translation, and information retrieval.

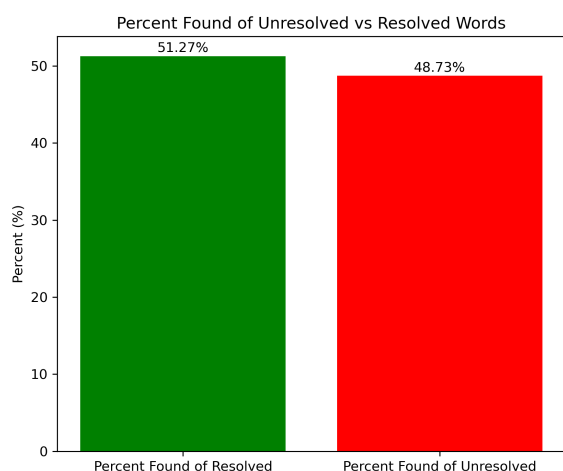


Figure 5: Proportion of tokens successfully lemmatized versus unresolved tokens

Moreover, in addition to domain-level variation, we examined how text size influenced lemmatization outcomes. The results show a clear relationship between text length and system performance. Short texts achieved the highest resolution rate, with approximately 60 percent of tokens assigned a lemma and an associated POS label derived from the lexicon. In contrast, medium-length texts achieved 47% resolution, while long texts recorded the lowest performance.

This trend can be explained by the structural characteristics of different text sizes. Short texts, such as social media posts and user comments, tend

to contain fewer tokens and simpler morphological constructions, making them more amenable to rule-based lemmatization. Medium-length texts, while offering richer context, often include greater lexical diversity and more complex derivational structures, which pose challenges to rule-based systems. Long texts, such as full-length news articles and analytical reports, are particularly difficult due to their higher frequency of compounding, derivation, and rare vocabulary, which increases the number of unresolved tokens.

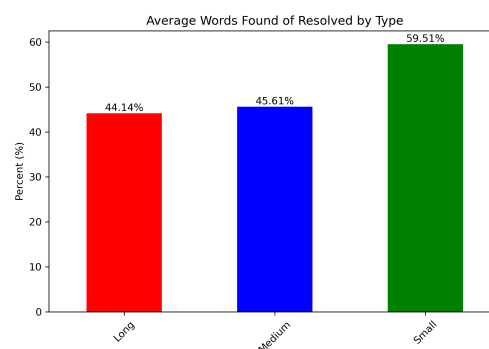


Figure 6: Lemmatization accuracy (percentage of tokens successfully reduced to root forms) across text length categories

To further assess the robustness of the rule-based lemmatizer, we evaluated its performance across different topical domains in the corpus. The results indicate notable variation in resolution rates depending on the domain. Specifically, the average proportion of tokens successfully lemmatized and assigned POS tags was highest in Business (72.38 percent), followed by Education (49.38 percent) and Politics (46.14 percent), while the remaining text categories resolved 45 percent.

This distribution suggests that domains such as business benefited from higher lexical regularity and stronger overlap with the curated lexicon, resulting in a significantly higher resolution rate. In contrast, categories such as technology, health, and sports exhibited lower resolution.

Table 4: % of lemmatised words by text domain

Category	Words Resolved (%)
Caafimaad (Health)	43.01%
Ciyaaro (Sports)	43.99%
Diin (Religion)	44.59%
Waxbarasho (Education)	49.38%
Ganacsi (Business)	72.38%
Siyaasad (Political)	46.14%
Madadaalo (Entertainment)	44.67%
Tiknoolojiyo (Technology)	42.38%

5 Discussion

This study demonstrates the application of a linguistically informed and empirically grounded methodology to build a large-scale lemmatization resource for the Somali language with a focus on the MAXAA TIRI written dialect. Our data collection and analysis framework was designed to capture linguistic authenticity, morphological diversity, and domain coverage, all of which are key requirements for evaluating NLP systems in low-resource contexts.

We compiled a total of 5,584 linguistically verified root words sourced from reputable corpus-based platforms and thoroughly validated by Somali linguists to reduce noise and ensure accuracy. These root words were expanded to 78,629 POS-tagged derivatives, creating a comprehensive resource. The resulting lexical database captures the agglutinative and derivational properties of the Somali language which provides a strong foundation for downstream applications, including lemmatization, POS tagging, information retrieval, and classification.

To preserve structural consistency and data integrity, the study introduces a web-based annotation and management platform utilizing relational database architecture with real-time validation to prevent duplication. Controlled crowdsourcing, subjected to expert validation, ensured linguistic precision. Such hybrid annotation strategies increasingly reflect best practices in low-resource NLP, where fully manual annotation is impractical and fully automated methods risk propagating noise. Empirical evaluations across 294 documents spanning multiple textual domains and length categories demonstrated clear performance patterns. Short texts, particularly social media content, achieved the highest lemmatization accuracy (approximately 60%), attributable to simpler syntactic structures and reduced morphological complexity. In contrast, longer news articles demonstrated lower accuracy due to increased lexical variability and complex derivational patterns.

Domain-specific analysis further confirmed that rule-based systems depend heavily on the level of structuredness in the text. Some written business documents reached 72.38% accuracy, benefiting from standardized and repetitive terminology such as ‘heshiis’ (agreement), ‘maalgashi’ (investment), ‘deyn’ (loan), ‘dakhliga’ (revenue) and ‘shirkad’ (company), which map cleanly to the constructed

lexicon. In contrast, domains such as technology and health showed weaker performance possibly because they contain:

- Borrowed terminology, e.g., ‘kombiyuutar’ (computer), ‘antibaayootik’ (antibiotic), etc.
- Code-switching, e.g., ‘waxaan update-gareeyay system-ka’ (I updated the system), etc.
- Orthographic inconsistency, e.g., ‘cafimaad/ caafimaad’ (health), etc.

These findings illustrate that dataset diversity, not only in linguistic structure but also in domain and text length, is essential for robust system evaluation. These features challenge deterministic methods by increasing out-of-vocabulary rates and morphological ambiguity. Additionally, the observed performance variation further highlights the limitations of rule-based systems in handling dialectal variation, code-switching, and orthographic inconsistency.

Overall, the study presents a pioneering lemmatization lexicon and a web-based annotation tool that resulted in a linguistically validated Somali NLP resource with its pilot empirical assessment on diverse test datasets showing promising performance. The study’s findings also motivate the integration of curated lexical resources with statistical and neural modeling techniques to achieve greater generalization in morphologically complex, low-resource languages.

6 Conclusion

This study developed a scalable and linguistically grounded infrastructure for the Somali NLP resources, moving beyond proof-of-concept to create a reusable lemmatization lexicon. The work delivered a large expert-validated database linking 5,584 roots to over 78,629 POS-tagged derivatives, supported by a web-based annotation and crowdsourcing platform that enables sustainable future data expansion. Empirical testing across 294 documents covering multiple domains and text lengths demonstrated that dictionary and rule-based approach built on the developed dataset can lemmatise more than 51% of tokens overall. More precisely, the performance varied across domains, with the highest accuracy observed in business texts (72%). The system also performed better on shorter and less complex content, such as social media comments.

Future research related to this work will focus on the following directions:

- Expanding the lexical database through controlled crowdsourcing and community-driven annotation to increase domain coverage and scalability.
- Addressing dialectal variation, irregular morphology, orthographic inconsistency, and borrowed terminology through hybrid linguistic–neural approaches.
- Integrating machine learning and deep learning techniques to complement the rule-based system to enable context-sensitive lemmatization.
- Leveraging cross-lingual transfer learning to other morphologically rich languages such as Amharic and Arabic.

Overall, this paper establishes both a novel lexical corpus for Somali and a replicable methodology for other under-resourced languages, demonstrating how linguistic expertise, community participation, and computational methods can converge to build sustainable NLP ecosystems for under-represented languages.

Data Availability

Datasets are made publicly available for the research community on this GitHub page ².

References

- David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreuzer, Xiaoyu Shen, Machel Reid, Dana Ruitter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajudeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, and 26 others. 2022. [A few thousand translations go a long way! leveraging pre-trained models for African news translation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.
- David Ifeoluwa Adelani. 2025. Natural language processing for african languages. *arXiv preprint arXiv:2507.00297*.
- ²<https://github.com/ShafieAbdi/Somali-Lemmatization-Crowdsourcing->
- David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure FP Dossou, Akintunde Oladipo, Doreen Nixdorf, and 1 others. 2023. Masakhanews: News topic classification for african languages. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159.
- 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. *arXiv preprint arXiv:2204.06487*.
- Abdisalam Badel, Ting Zhong, Wenxin Tai, and Fan Zhou. 2023. Somali information retrieval corpus: Bridging the gap between query translation and dedicated language resources. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 7463–7469.
- Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with lematus. In *16th annual conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1391–1400. Association for Computational Linguistics (ACL).
- Cheikh Bamba Dione. 2021. Multilingual dependency parsing for low-resource african languages: Case studies on bambara, wolof, and yoruba. *Proceedings of the International Conference on Parsing Technologies (IWPT)*.
- Shai Gordin, Aleksy Sahala, Shahar Spencer, and Stav Klein. 2025. Evacun 2025 shared task: Lemmatization and token prediction in akkadian and sumerian using llms. In *Proceedings of the Second Workshop on Ancient Language Processing*, pages 242–250.
- Nizar Habash and Bonnie Dorr. 2003. A categorial variation database for english. Technical report.
- Haibin Liu, Tom Christiansen, William A Baumgartner Jr, and Karin Verspoor. 2012. Biolemmatizer: a lemmatization tool for morphological processing of biomedical text. *Journal of biomedical semantics*, 3(1):3.
- Enrique Manjavacas, Ákos Kádár, and Mike Kestemont. 2019. Improving lemmatization of non-standard languages with joint learning. *arXiv preprint arXiv:1903.06939*.
- Irene Masiringi Mathayo and Alfred Malengo Kondoro. 2024. Unveiling swahili verb conjugations: A comprehensive dataset for low-resource nlp. In *Proceedings of the 2024 8th International Conference on Natural Language Processing and Information Retrieval*, pages 149–156.

Aleksandra Miletić and Janine Siewert. 2023. Lemmatization experiments on two low-resourced languages: Low saxon and occitan. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 163–173.

Muhidin A Mohamed, Shuab D Ahmed, Yahye A Isse, Hanad M Mohamed, Fuad M Hassan, and Houssein A Assowe. 2025. Detection of somali-written fake news and toxic messages on the social media using transformer-based language models. *arXiv preprint arXiv:2503.18117*.

Shafie Abdi Mohamed and Muhidin Abdullahi Mohamed. 2023. Lexicon and rule-based word lemmatization approach for the somali language. *arXiv preprint arXiv:2308.01785*.

Siraj Mohammed. 2020. Using machine learning to build pos tagger for under-resourced language: the case of somali. *International Journal of Information Technology*, 12(3):717–729.

Abdillahi Nimaan, Pascal Nocéra, and Jean-François Bonastre. 2006. Towards automatic transcription of somali language. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. Kinyabert: a morphology-aware kinyarwanda language model. *Proceedings of the Association for Computational Linguistics (ACL)*.

Aleksi Sahala, Tero Alstola, Jonathan Valk, and Krister Lindén. 2023. Lemmatizing and pos-tagging akkadian with babylemmatizer and dictionary-based post-correction. In *CLARIN Annual Conference*, pages 111–119. CLARIN ERIC.

Ranka Stanković, Cvetana Krstev, Ivan Obradović, Biljana Lazić, and Aleksandra Trtovac. 2016. Rule-based automatic multi-word term extraction and lemmatization. In *LREC*, pages 507–514.

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

A Appendix

Appendix A: Uploading approved root Words

B Appendix

Appendix B: Annotating root Words

C Appendix

Appendix C: Searching words

D Appendix

Appendix D: Home Page Interface

Search The Root word

Annotated: 5000

#	Root Word	Action
1	abuur	<input type="button" value="edit"/> <input type="button" value="delete"/>
2	adeeg	<input type="button" value="edit"/> <input type="button" value="delete"/>
3	adeegso	<input type="button" value="edit"/> <input type="button" value="delete"/>

Figure 7: Uploading approved root words to the platform

Word Type *

Part of Speech *

Root Word *

Enter the Derivative_Word Words *

Figure 8: Data Annotation

Derived Words	Part of Speech	Root Word
abuuray	FAL TAGAY	abuur
abuurtay	FAL TAGAY	abuur
abuurnay	FAL TAGAY	abuur
abuureen	FAL TAGAY	abuur
abuuraynay	FAL TAGAY	abuur
abuuraysay	FAL TAGAY	abuur

Figure 9: Annotation tool: word search interface

★ Somali Lexical Database
Home Article Features Cat Var Search Contact Login [Signup](#)

Welcome to the Somali Lexical Database

A comprehensive resource for exploring and understanding the richness of the Somali language.

The **Somali Lexical Database** is a digital initiative to preserve and document the Somali language. It was developed in partnership with universities and language experts to support research, education, and technology. This project aims to:

- Create a central, searchable repository of Somali root, derived words and their POS Tags.
- Enable research and development of Somali language tools (e.g., NLP).

This project is made possible by Jamhuriya University NLP research group.



Jamhuriya University



Somali National University



Universite de Djibouti

Figure 10: Annotation Tool: Home Page Interface

Kunnafonidilaw ka Cadeau: an ASR dataset of present-day Bambara

Yacouba Diarra, Panga Azazia Kamaté,
Nouhoum Souleymane Coulibaly, Michael Leventhal

RobotsMali AI4D Lab, Bamako, Mali; Correspondence: research@robotsmali.org

Abstract

We present Kunkado, a 160-hour Bambara ASR dataset compiled from Malian radio archives to capture present-day spontaneous speech across a wide range of topics. It includes code-switching, disfluencies, background noise, and overlapping speakers that practical ASR systems encounter in real-world use. We finetuned Parakeet-based models on a 33.47-hour human-reviewed subset and apply pragmatic transcript normalization to reduce variability in number formatting, tags, and code-switching annotations. Evaluated on two real-world test sets, finetuning with Kunkado reduces WER from 44.47% to 37.12% on one and from 36.07% to 32.33% on the other. In human evaluation, the resulting model also outperforms a comparable system with the same architecture trained on 98 hours of cleaner, less realistic speech. We release the data and models to support robust ASR for predominantly-oral languages.

1 Introduction

Data availability for automatic speech recognition (ASR) for the Bambara language has increased significantly this year. For about three years, Jeli-ASR (Diarra et al., 2022), a corpus of 30 hours of transcribed griot narrations, had been the only open ASR dataset for Bambara, but in late 2025, a team at RobotsMali AI4D Lab released a 612 hour dataset as part of the African Next Voices (ANV) project (Diarra et al., 2025a), scaling the amount of open data available by a factor of 20x. In both cases, the audio was recorded during the project in a relatively controlled environment with consistent quality control prior to transcription.

Cost and the challenges of field collection have led to many initiatives aiming to increase speech data for low resource languages (LRLs) to either align data that has already been recorded such as Bible readings (Black, 2019; Pratap et al., 2023) or

employ synthetic generation techniques (DeRenzi et al., 2025). The former requires a pretrained acoustic model to produce alignments between acoustic features and the corresponding phonemes in a long transcript (Tsoukala et al., 2023) or an existing aligned set to train a speech synthesizer to generate phones from the transcript and find those phones in the audio (Black, 2019). Synthetic speech data can be useful for robustness training but its value is significantly lower than that of real speech. It also increases the risk of propagating the biases in the generating distribution in the trained models (Rosenberg et al., 2019; Moslem, 2024).

Gender and age are usually the primary concerns with respect to the representativeness of the dataset. Naturalness in spontaneous speech is often neglected even though it is an equally important factor in creating a dataset that truly represents speech in all its dimensions. Many projects curate data following guidelines that hinder the capture of spontaneous speech, seeking to minimize or prohibit code-switching, background noise, slang, and ungrammatical constructions. Such guidelines reduce the usability of the models for real-world deployment scenarios (Diarra et al., 2025a), as these phenomena reflect the realities of daily interactions and the linguistic evolution of many low resource languages. Day-to-day speech in many African languages feature high rates of inter-sentential and intra-sentential language shifts both between African languages and with high-resource colonial languages. Code-switching may enable models to use the high-resource language constantly appearing in sentences and conversations to improve accuracy on the LRL.

Among LRLs, predominantly-oral languages (POL) constitute a large subset where speech, to the almost complete exclusion of writing, has been and remains the dominant means of knowledge transmission. Bambara along with most African languages, is a POL. For many of those languages,

the only readily-available body of natural communication is radio and television broadcasts. This resource includes background and foreground music, various types of noise, and phone calls in which the audience jumps into the conversation, bringing a variety of accents and dialects (Doubouya et al., 2021). This abundant source of data is rarely exploited due to the many challenges in transcribing such unpredictable conversations and because it implies renouncing control over topics and audio quality.

In this paper, we introduce **Kunkado**, a 160 hour transcribed ASR dataset compiled from radio archives. The title of this paper, *Kunnafonidilaw ka cadeau* can be translated as "Media's Gift". It is also a good example of the everyday code-switching in Bambara, the word *cadeau* being a direct borrowing from French. The entire dataset was automatically transcribed, with 25% corrected by humans. We report on human evaluation performed on a model trained with the reviewed subset of Kunkado, comparing this result to that of the same model trained with a much larger quantity of curated data. (see Section 4). In the next section, we share insights on handling code-switching and numbers and how trade-offs can be made between consistency and simplicity in transcription.

2 Characteristics of Kunkado

Audio collection and segmentation: We obtained approximately 300 hours of broadcast recordings from 4 Malian radio stations. After segmentation, we only retained segments between 600 milliseconds and 45 seconds of duration. We have released, on Hugging Face¹, 118,925 segments totaling 161.15 hours. Approximately 94% of segments are less than 15 seconds in duration. The mean duration of segments is 4.9 seconds.

For audio segmentation, we employed a simple energy threshold-based method implemented via the `split_on_silence` function from the `pydub` library (Robert, 2018). This function performs segmentation by analyzing the root mean square (RMS) energy of the audio and splitting the signal where the energy drops below a predefined absolute loudness threshold. Specifically, we used the parameters `min_silence_len = 600` milliseconds and `silence_thresh = -35` dBFS. This configuration ensured that the audio stream was only split

¹The full dataset is released under CC BY-SA 4.0 at [RobotsMali/kunkado](https://huggingface.co/RobotsMali/kunkado) on HF

when the signal's loudness dropped below -35 dB relative to the maximum possible volume for a minimum duration of 600 milliseconds.

While this kind of silence proxy segmentation is much faster than modern voice-activity-detection based methods and does not discard any part of the original recording—i.e., it simply finds endpoints to split on based on the given parameters, yielding $segments_duration = original_duration$, it also results in much rougher segmentation and speech cut-offs. We modelled those cut-offs during transcription (see Table 1).

Noise estimation: We calculate signal-to-noise ratio (SNR) as an estimate of the level of noise/non-speech signal in the segmented dataset, using the same implementation and classification thresholds defined by Diarra et al.. 69.2% of the segments fall above the High SNR category (>15 dB). Although they feature a considerable amount of acoustic non-speech information, this relatively good SNR measurement is explained by the fact that the volume of those events is significantly lower than that of speech, as broadcasts are recorded with professional equipment. The measurements demonstrate that radio data is a quality source to create speech datasets (Doubouya et al., 2021). Figure 1 shows the density distribution of SNR values in the dataset.

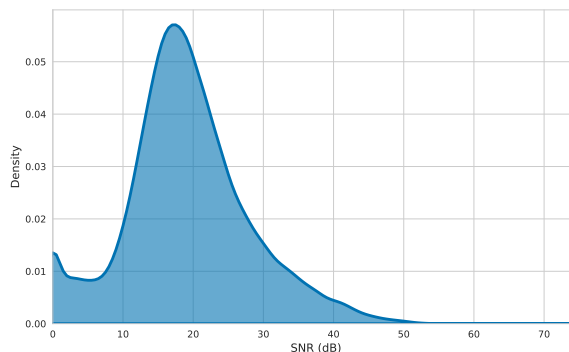


Figure 1: Density Distribution of Signal-to-Noise Ratio values in Kunkado. This figure includes both subsets

Transcription: We have re-engineered transcription, aiming at better matching the requirements of the task to the skill level of annotators that were available to us and speeding up the process. We used the process described in Diarra et al., redefining segment annotation as a first review task where the objective is to correct an automatic transcription generated

Tag	Meaning
<BRUITS>	generic noise
<INCOMPRÉHENSIBLE>	fully inaudible speech
<CHEVAUCHEMENT>	speaker overlap
<RIRES>	laughter
<MUSIQUE>	music / jingle (no lyrics)
<TOUX>	cough
<INVOCATION>	prayers, quranic excerpts
<ECHO>	echo artifact
<APPLAUDISSEMENTS>	applause
<CRIS>	screams
<PLEURES>	crying
__phrase__	double underscores delimit code-switched words and sentences
...	used to mark speech cut-offs and hesitations

Table 1: Tags/Annotations for acoustic and linguistic events captured in Kunkado

by RobotsMali/soloni-114m-tdt-ctc-v0 Bambara ASR model and to ensure it conforms to our guidelines.

Our guidelines direct annotators to transcribe numbers with Hindu-Arabic numerals, in contrast to orthodox ASR practice where numbers are written out as words. Writing numerals speeds up transcription, reduces potential ambiguity in parsing the quantity, simplifies downstream processing, and is more easily validated by human reviewers. We also have created data for comparison of number formatting approaches in end-to-end ASR—which remains a challenge even in high-resource languages (Huber and Waibel, 2025). Annotators were instructed to use 13 tags to capture as much acoustic and linguistic information as possible. Code-switching to French was written out using the French orthography, ignoring the existence of Bambara-ized spellings commonly, but inconsistently, used for many French words and expressions. Code-switching to Arabic also followed this principle, though transliterated to the Latin alphabet. A transliteration standard was not enforced due to the specialized knowledge required and the complexity that this would have added to the annotation task. While French is often woven into Bambara in a wide variety of ways, Arabic is generally limited to a small set of formulaic Islamic expressions, though these are used frequently. Table 1 presents the complete list of tags and their significations. Other rules on proper nouns, acronyms and spelled-out words are similar to those used in the ANV Bambara project (Diarra et al., 2025a). Although annotators were not required to follow

a single standard Bambara orthography, they used the Bamadaba dictionary (Vydrin, 2022) as their primary reference. The annotators used the same data annotation interface described in (Diarra et al., 2025a).

Despite added complexity with respect to tagging and code-switching, our team of seven annotators were able to correct 39.3 hours of segments in roughly 1260 hours of human labor, yielding a 32x ratio, a bit faster than the 36x reference datapoint reported in the transcription cost analysis study by (Diarra et al., 2025b) which used the same model for generating the automated transcriptions. We speculate that more flexible orthography and the use of numerals rather than spelled-out numbers contributed to this 4x difference.

3 ASR Experiments

We finetuned multiple Bambara ASR models, previously trained on Jeli-ASR (Diarra et al., 2022), from RobotsMali’s baseline ASR experiments. These models are based on NVIDIA’s Parakeet family of monolingual English ASR models (Rekesh et al., 2023). We evaluated all the models on a 5 hour test set taken from the Kunkado data, and Nyana-Eval, a small, stratified human evaluation dataset with only 45 entries of 3 minutes total duration (Diarra et al., 2025a). We report in Section 4 the automatic and human evaluation results for all the RobotsMali soloni models, except the v3 version for which we only report the WER gains since it was not part of the human evaluation².

²We still release all the other models with the corresponding WER scores. hf.co/RobotsMali/models

Model	WER (%) ↓		CER (%) ↓	
	Kunkado Test	Nyana-Eval	Kunkado Test	Nyana-Eval
soloni (jeli-asr)				
Unfinetuned (v0)	46.91	40.75	30.56	24.71
Finetuned (v1)	39.13	39.44	20.98	20.5
soloni (afvoices)				
Unfinetuned (v2)	44.47	36.07	29.61	20.24
Finetuned (v3)	37.12	32.33	21.17	16.72

Table 2: ASR Evaluation results: We apply the same normalization steps as explained in Section 3 and remove the tags from both the reference and the prediction before calculating the WER and CER. The values in bold highlight the best performance per metric for each benchmark.

Experimental setup: We used 4 NVIDIA A100 GPUs with 80GB VRAM each for these experiments. We finetuned soloni-114m-tdt-ctc-v0 and soloni-114m-tdt-ctc-v2 in this study. The two models have a hybrid architecture with a Fast-Conformer encoder (Rekesh et al., 2023) and two jointly trained decoders: an autoregressive TDT decoder, Token-and-Duration Transducer (Xu et al., 2023; Graves, 2012) and a convolutional decoder trained with a Connectionist Temporal Classification (CTC) loss function (Graves et al., 2006). We will refer to the models as soloni-v0 and soloni-v2. The finetuned versions of these models are identified on Hugging Face as v1 and v3, respectively.

Training Data: We finetuned the two models on the training set of the human-reviewed subset of Kunkado (33.47 hours). We simplified training for the ASR model by removing symbols and diacritics, and making number and Bambara-specific normalizations using our `bambara-normalizer` Python package. We then removed code-switching markers (the double underscores) and all punctuation. We kept only a reduced set of acoustic event tags (e.g., overlaps, paralinguistic vocalizations such as laughter, and music), which are modeled during training but ignored during evaluation. We applied these normalization steps after noticing that, with only just over 30 hours of training data, the models were struggling with the human-annotation variability present in the reference transcripts. These models rarely improved beyond 70% WER due to inconsistencies in numbers, tags and code-switching. We concluded that it would require much more data for an end-to-end ASR model to learn the original task. Table 3 contains 3 examples of how this normalization simplifies the task.

Training configurations: We trained soloni-v0 for 100k steps on 2 GPUs, with batch size 40, using the AdamW optimizer and Noam scheduler with a learning rate scaling factor of 0.03 and a 10% warmup ratio (Vaswani et al., 2017). We froze the encoder for the last 7,000 steps, training only the 5M combined parameters of the two decoders.

For soloni-v2, trained on much more data with the ANV dataset, we simply trained all the 114M parameters for 13k steps on 4 GPUs, with batch size of 64 and LR scaling factor of 1 and 3,000 warmup steps. Both models were trained with bf16 float precision.

4 Evaluation & Results

We evaluated the Word Error Rate (WER) and Character Error Rate (CER) of the resulting models, and we also report the findings of the human evaluation conducted during the Bambara ANV project (Diarra et al., 2025a, Tall, 2025). Table 2 presents the WERs and CERs of the two models before and after our finetuning experiments; the terms in parentheses represent the dataset on which the unfinetuned versions were trained and their version IDs on Hugging Face. Since these models have two decoders, we only report the best scores; detailed per-decoder metrics can be found in their respective model cards. Both finetuned versions reduce WER on the Kunkado test set. Soloni-v3, our latest finetuned model, achieves the best results on both benchmarks³.

In the detailed comparative analysis report by Tall on several RobotsMali ASR models, we found

³Note that, as one third of Nyana-Eval (15 examples) comes from the Kunkado test set, there is a small intersection between the two test sets

Original	Normalized
Bamananw ko ten ko maa jugu t'i ba sinamuso ye, nka n'a ni ba be kele la, <CHEVAUCHEMENT>__ voilà__ o kɔni ka di ye __ donc__ jamana...	bamananw ko ten ko maa jugu t'i ba sinamuso ye nka n'a ni ba be kele la <tag> voilà o kɔni ka di ye donc jamana
<MUSIQUE>an b'an sinsin ni Alahutala tɔɔ barikama ye.	<MUSIQUE> an b'an sinsin ni Alahutala tɔɔ barikama ye
εε... nɛɛjuru sira 76 64 10 10... __76 64 10 10__	εε nɛɛjuru sira bi wolonwula ni wɔɔɔ bi wɔɔɔ ni naani tan tan soixante-seize soixante-quatre dix dix

Table 3: Sample Kunkado transcripts pre and post normalization

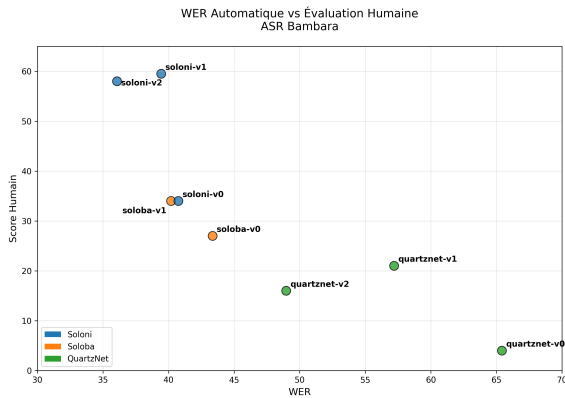


Figure 2: WER vs human evaluation. Figure from (Tall, 2025)

that soloni-v1 achieved the highest human evaluation score on Nyana-Eval (59.5 out of 135), outperforming soloni-v2, which was trained with an additional 98 hours from the African Next Voices Bambara dataset. Tall evaluated the model outputs on a 0–3 scale, where 0 indicated barely understandable transcriptions with multiple errors and 3 indicated semantically and lexically accurate transcriptions. Scores were assigned using criteria such as semantic and lexical fidelity, handling of code-switching and disfluencies, and accurate recognition of proper names. According to Tall, soloni-v1 shows clear gains in handling speech disfluencies, proper names, code-switching, and noisy or overlapping speech—phenomena that are common in natural Bambara conversations (Tall, 2025). The v3 models are not included in that report because they were finetuned from v2 after the report was published; however, the substantial WER improvement in soloni-v3 (32.33% vs. 36.07%) suggests that a proportional improvement in human scores is plausible. Figure 2 (from the comparative analysis report) ranks the models by both human scores and WER. All of these models, along with their v3

variants, are available on RobotsMali’s Hugging Face profile.

5 Conclusion

In this work, we introduce Kunkado, a 160-hour Bambara ASR dataset compiled from Malian radio archives, "Media’s Gift" to LRL Bambara NLP, and designed to better reflect present-day, naturally occurring speech. By shifting from curated source materials and controlled recording conditions toward broadcast content containing code-switching, disfluencies, background noise, and overlapping speech, we address a major source of domain mismatch that often yields models that perform well on curated data using standard metrics but poorly on real-world data and applications. Our experiments show that finetuning on 33.47 hours of human-reviewed Kunkado data yields substantial gains, and the best-performing configuration (soloni-v3) improves WER on both Kunkado-sourced test data and Nyana-Eval relative to its unfinetuned counterpart. These results support the broader conclusion that, for Bambara, representativeness and linguistic realism can matter as much as (or more than) raw hours when the goal is real-world usability. At the same time, the linguistic richness of spontaneous speech increases annotation and modeling difficulty, motivating pragmatic transcription guidelines and normalization choices, as well as continued investment in human review. We release Kunkado and the associated models to encourage research, community-driven quality standards led by native speakers, and future work on code-switching-aware evaluation and data collection and training from large-scale resources such as radio broadcasts for low-resource, predominantly-oral languages.

Limitations

The design of the Kunkado corpus intentionally prioritized the linguistic authenticity of the Bambara language. Specifically, the data reflects the most common and contemporary register of spoken Bambara, characterized by features inherent to natural conversational settings (e.g., podcasts, TV shows, and debates). These features include extensive code-switching, prevalent slang, and frequent spontaneous-speech disfluencies—exactly the features that the ground rules for the African Next Voices project required us to exclude from the dataset. (Diarra et al., 2025a)

Our ASR experiments demonstrated that these added linguistic complexities necessitate significantly greater volumes of data and advanced model engineering to achieve robust performance. Although radio broadcasts, in the West African region, represent a virtually inexhaustible, daily-generated source of data, resource constraints limited our labeling efforts to only ≈ 40 hours within the scope of this project, relying solely on limited internal funding.

The compelling results from the human evaluation indicate that continued annotation and development on more authentic data could substantially accelerate the deployment of high-fidelity, real-world speech applications for Bambara speakers. Moving forward, we advocate for a community-driven dataset design strategy where quality standards are organically defined with robust participation by native speakers.

Acknowledgments

We would like to extend our gratitude to Radio Benkouma, Mouso TV, ORTM and Radio Sahel FM for graciously sharing their archives with us and allowing us to release the data with an open source license.

References

- Alan W Black. 2019. [Cmu wilderness multilingual speech dataset](#). In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975.
- Brian DeRenzi, Anna Dixon, Mohamed Aymane Farhi, and Christian Resch. 2025. [Synthetic voice data for automatic speech recognition in african languages](#). *Preprint*, arXiv:2507.17578.
- Sebastien Diarra, Michael Leventhal, and Allahsera Auguste Tapo. 2022. [Robotsmali griots speech dataset](#), and asr. <https://github.com/robotsmali-ai/jeli-asr/>.
- Yacouba Diarra, Nouhoum Souleymane Coulibaly, Panga Azazia Kamaté, Madani Amadou Tall, Emmanuel Élisé Koné, Aymane Dembélé, and Michael Leventhal. 2025a. [Dealing with the hard facts of low-resource african nlp](#). *Preprint*, arXiv:2511.18557.
- Yacouba Diarra, Nouhoum Souleymane Coulibaly, and Michael Leventhal. 2025b. [Cost analysis of human-corrected transcription for predominately oral languages](#). *Preprint*, arXiv:2510.12781.
- Moussa Doumbouya, Lisa Einstein, and Chris Piech. 2021. Using radio archives for low-resource speech recognition: Towards an intelligent virtual assistant for illiterate users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35.
- Alex Graves. 2012. [Sequence transduction with recurrent neural networks](#). *Preprint*, arXiv:1211.3711.
- Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. [Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks](#). In *ICML 2006*, volume 2006, pages 369–376.
- Christian Huber and Alexander Waibel. 2025. [Handling numeric expressions in automatic speech recognition](#). *Preprint*, arXiv:2408.00004.
- Yasmin Moslem. 2024. [Leveraging synthetic audio data for end-to-end low-resource speech translation](#). *Preprint*, arXiv:2406.17363.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. [Scaling speech technology to 1,000+ languages](#). *Preprint*, arXiv:2305.13516.
- Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. [Fast conformer with linearly scalable attention for efficient speech recognition](#). *Preprint*, arXiv:2305.05084.
- James Robert. 2018. [Pydub: Manipulate audio with a simple and easy high level interface](#).
- Andrew Rosenberg, Yu Zhang, Bhuvana Ramabhadran, Ye Jia, Pedro J. Moreno, Yonghui Wu, and Zelin Wu. 2019. [Speech recognition with augmented synthesized speech](#). *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 996–1002.
- Madani Amadou Tall. 2025. [Analyse comparative humaine des modèles asr bambara de robotsmali](#).

- Chara Tsoukala, Kosmas Kritsis, Ioannis Douros, Nikolaos Kokkas, Vasileios Arampatzakis, Vasileios Sevetlidis, Stella Markantonatou, and George Pavlidis. 2023. [Asr pipeline for low-resourced languages: A case study on pomak](#). pages 40–45.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). *Preprint*, arXiv:1706.03762.
- Valentin Feodosievich Vydrin. 2022. [Vers un dictionnaire orthographique bambara](#). *Mandenkan : Bulletin Semestriel d'Études Linguistiques Mandé*, (68):59–82.
- Hainan Xu, Fei Jia, Somshubra Majumdar, He Huang, Shinji Watanabe, and Boris Ginsburg. 2023. [Efficient sequence transduction by jointly predicting tokens and durations](#). *Preprint*, arXiv:2304.06795.

Full Fine-Tuning vs. Parameter-Efficient Adaptation for Low-Resource African ASR: A Controlled Study with Whisper-Small

Sukairaj Hafiz Imam^{1,2}, Muhammad Yahuza Bello¹, Hadiza Ali Umar¹,
Tadesse Destaw Belay³, Idris Abdulmumin⁴, Seid Muhie Yimam⁵,
Shamsuddeen Hassan Muhammad^{1,6}

¹Bayero University Kano, ²Northwest University Kano, ³Instituto Politécnico Nacional,

⁴University of Pretoria, ⁵University of Hamburg, ⁶Imperial College London

Correspondence: sukhimam00@gmail.com

Abstract

Automatic speech recognition (ASR) for African low-resource languages (LRLs) is often limited by scarce labelled data and the high cost of adapting large foundation models. This study evaluates whether parameter-efficient fine-tuning (PEFT) can serve as a practical alternative to full fine-tuning (FFT) for adapting Whisper-Small with limited labelled speech and constrained compute. We used a 10-hour subset of NaijaVoices covering Hausa, Yorùbá, and Igbo, and we compared FFT with several PEFT strategies under a fixed evaluation protocol. DoRA attains a 22.0% macro-average WER, closely aligning with the 22.1% achieved by FFT while updating only 4M parameters rather than 240M, and this difference remains within run-to-run variation across random seeds. Yorùbá consistently yields the lowest word error rates, whereas Igbo remains the most challenging, indicating that PEFT can deliver near FFT accuracy with substantially lower training and storage requirements for low-resource African ASR.

1 Introduction

African ASR has advanced rapidly, driven by large self-supervised and multilingual foundation models such as wav2vec 2.0, HuBERT, MMS, and Whisper (Imam et al., 2025). While these models achieve strong performance for many high-resource languages (Yang et al., 2025; Palivela et al., 2025), recognition accuracy for African LRLs often remains lower, typically reflected in higher word error rates and more limited standardised evaluation.

Recent progress in African ASR has been driven by multilingual pretraining, targeted adaptation, and the release of more diverse speech corpora such as AfriSpeech-200 by Afonja et al. (2024), ÌròyìnSpeech by Ogunremi et al. (2024), and NaijaVoices developed by Emezue et al. (2025). However, adapting large encoder-decoder models remains challenging in practice due to the computa-

tional cost of FFT and the linguistic complexity of many African languages, including tone, diacritics, and frequent code-switching.

PEFT addresses this limitation by freezing the backbone and updating only a small number of newly introduced adapter parameters, rather than the entire model (Ali et al., 2025). Methods such as LoRA (Kwok et al., 2024), DoRA (Joseph and Baby, 2024), AdaLoRA (Kwok et al., 2024), and IA³ (Wang et al., 2024) have shown that lightweight adaptations can approach FFT performance while requiring far fewer trainable parameters. Despite growing interest, their relative behaviour for multilingual African ASR under severely low-resource supervision remains under-explored.

In this work, we evaluate whether PEFT can effectively adapt Whisper-Small to Hausa, Yoruba, and Igbo under extremely low-resource conditions. Using NaijaVoices, we construct a stratified 10-hour subset (approximately 3.3 hours per language), apply standard audio and text preprocessing, and train a set of controlled adaptation strategies (FFT and PEFT). To enable fair comparison across PEFT methods variants, we fix the adapter rank to $r = 32$ and scaling to $\alpha = 64$, and we standardise optimisation and decoding across all runs. We report per-language WER and macro-average WER on held-out test data, emphasising the accuracy-parameter trade-off when adapting large ASR models in compute and data-constrained African language settings.

This paper contributes in three ways: (i) First, we present a controlled setup for adapting Whisper-Small on NaijaVoices using only 10 hours of transcribed speech. (ii) benchmark FFT against common PEFT methods across Hausa, Yorùbá, and Igbo, and (iii) we analyse the trade-off between recognition accuracy and the number of trainable parameters to show when PEFT is a practical alternative to FFT for low-resource African ASR.

2 Related work

Recent progress in African ASR has been supported by both methodological advances and the release of more diverse speech corpora. Notable datasets include AfriSpeech-200, a 200-hour African-accented English corpus covering 120 accents (Afonja et al., 2024), as well as language-specific resources such as ÌròyìnSpeech, a 42-hour Yorùbá corpus spanning news and creative speech (Ogunremi et al., 2024), and NaijaVoices, a large Nigerian speech dataset with varied accents and recording contexts (Emezue et al., 2025). Collectively, these efforts highlight the central role of dataset scale, linguistic coverage, and transcription quality in determining the effectiveness of pre-trained ASR models for African languages.

Alongside dataset expansion, improvements have been driven by self-supervised learning (SSL) and multilingual training strategies. Mdhaftar et al. (2024) compared several SSL encoders for ASR and spoken language understanding on the Tunisian dialect, reporting the lowest WER among the evaluated models with w2v-BERT 2.0. Similarly, Abdou Mohamed et al. (2024) examined multilingual ASR across multiple African languages by contrasting joint training, language-dependent training with language identification, and language-independent tokenisation, showing that multilingual training and careful handling of diacritics can improve recognition for tonal scripts. Fine-tuning remains particularly important in domain- and code-switch-sensitive scenarios. For example, Babatunde et al. (2025) developed a Yorùbá–English code-switching ASR system by fine-tuning monolingual and multilingual models, while Chevtchenko et al. (2025) fine-tuned Wav2Vec 2.0, HuBERT, and Whisper for Xhosa child reading assessment.

Despite these advances, most prior work in African ASR still relies on FFT, which updates all model parameters, by increasing training memory requirements and producing saved model weights that typically require access to modern GPUs. These costs become more restrictive when only limited transcribed speech is available, as in datasets such as NaijaVoices. PEFT offers a practical alternative by freezing the backbone and training only lightweight task-specific modules, reducing the number of trainable parameters and the size of the saved model weights while supporting iterative training under constrained compute (Kwok et al., 2024; Wang et al., 2024; Joseph and Baby, 2024).

3 Methodology

Figure 1 summarises the experimental pipeline designed to separate and quantify the impact of PEFT under extreme data scarcity. The pipeline consists of four stages: (i) low-resource data curation and preprocessing, (ii) backbone model instantiation, (iii) standardised comparative evaluation of FFT and PEFT strategies with defined parameter updates, and (iv) standardised evaluation under a fixed decoding and scoring metrics. To ensure a standardised and consistent evaluation across methods, we fixed the training set size (10 hours), text normalisation, decoding configuration, and evaluation metrics for all experiments.

3.1 Data Curation and Preprocessing

We use the NaijaVoices corpus and focus on Hausa, Yorùbá, and Igbo, widely spoken Nigerian languages with diverse phonological and orthographic properties, including tone and diacritics, which enables a controlled comparison within a single dataset. To reflect a realistic low-resource setting while keeping the languages comparable, we sample a stratified random 10-hour subset that is approximately balanced across the three languages (about 3.3 hours per language). This limits the influence of language imbalance on the results, so differences in WER are less likely to be driven by unequal training data across languages. We then partition this 10-hour subset into 60/20/20 train/val/test splits using a speaker-disjoint split (by speaker_id) to prevent speaker leakage across splits. All audio is resampled to 16 kHz to match the Whisper front-end, and each segment is padded or truncated to Whisper’s 30-second context window (Simic and Bocklet, 2024). Transcripts are normalised to a consistent orthography while preserving language-specific diacritics to maintain lexical distinctions and avoid evaluation errors from overly aggressive normalisation.

3.2 Base Model

We adopt Whisper-Small as the encoder-decoder ASR base model. It transforms an input waveform into log-Mel features and generates a token sequence through autoregressive decoding. Although Whisper offers strong multilingual transfer, performance on underrepresented languages can lag behind that of high-resource languages (Pratama and Amrullah, 2024). We therefore evaluate targeted adaptation using FFT and PEFT under iden-

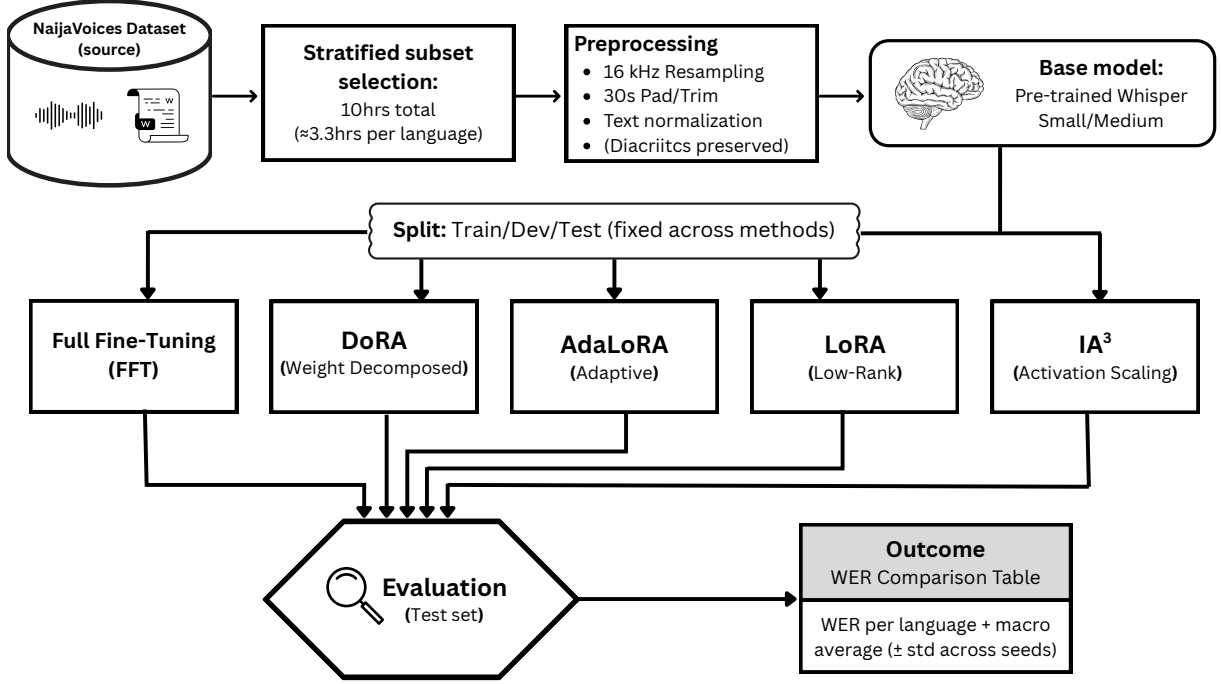


Figure 1: The pipeline shows a low-resource ASR pipeline: a stratified 10-hour subset from NaijaVoices (Hausa/Yoruba/Igbo) is preprocessed (16 kHz, 30s pad/trim, diacritics preserved), used to adapt a pre-trained Whisper model with FFT and PEFT methods (DoRA, AdaLoRA, LoRA, IA³), and then evaluated on the fixed held-out test split from the 10-hour subset (60/20/20).

tical experimental conditions. Whisper-Small is chosen as a practical compromise between multilingual capability and computational cost, enabling controlled comparisons and multi-seed replication under limited compute.

3.3 Fine-Tuning Strategies

As illustrated in Figure 1, our pipeline keeps the dataset, preprocessing steps, and evaluation protocol fixed and then applies five adaptation strategies to the same pre-trained Whisper base model. Let $W \in \mathbb{R}^{d \times k}$ denote a representative model weight matrix, where d is the input feature dimension and k is the output dimension, and let $W_0 \in \mathbb{R}^{d \times k}$ denote the corresponding pre-trained weight matrix from the Whisper checkpoint. This notation provides a unified view of the adaptation branches by expressing each method as a particular way of updating W relative to W_0 .

FFT (FFT): As a high-capacity baseline, FFT updates all model parameters (Simic and Bocklet, 2024):

$$W = W_0 + \Delta W, \quad (1)$$

where $\Delta W \in \mathbb{R}^{d \times k}$ is unconstrained. While offering maximum flexibility, FFT is computationally

and memory-intensive and is more prone to overfitting in low-data settings.

LoRA (Low-Rank Adaptation): LoRA freezes the pre-trained weights W_0 and learns a low-rank update (Kwok et al., 2024):

$$\Delta W = BA, \quad (2)$$

where $B \in \mathbb{R}^{d \times r}$ and $A \in \mathbb{R}^{r \times k}$ are the *trainable* low-rank factors. In particular, B maps an r -dimensional latent adaptation space to the d -dimensional output space, while A projects the k -dimensional input into the same r -dimensional space. The rank r controls the adapter capacity, with $r \ll \min(d, k)$.

The layer output for an input vector x becomes

$$h = (W_0 + BA)x, \quad (3)$$

so only A and B are updated during training, constraining adaptation to a low-dimensional subspace.

DoRA (Weight-Decomposed Low-Rank Adaptation): DoRA decouples the weight magnitude from its direction to better approximate the behaviour of FFT at PEFT-level size (Joseph and Baby, 2024). Specifically, it represents weights

via a magnitude vector m and a directional matrix V :

$$W = m \frac{V}{\|V\|_c}, \quad (4)$$

where $\|\cdot\|_c$ denotes the column-wise norm and \odot is the Hadamard product. During adaptation, DoRA applies a LoRA update to the direction while learning the magnitude explicitly:

$$W' = m \odot \frac{V_0 + BA}{\|V_0 + BA\|_c}. \quad (5)$$

This formulation allows independent control of feature scaling (via m) and directional shifts (via $V_0 + BA$), which can be beneficial in low-resource settings where overly flexible updates may overfit.

AdaLoRA: AdaLoRA extends LoRA by dynamically allocating rank capacity across layers during training, concentrating parameters in layers that contribute more to the task while maintaining a fixed overall adaptation size (Kwok et al., 2024; Ali et al., 2025).

IA³: IA³ modulates activations using lightweight learned vectors (Kwok et al., 2024). A common instantiation scales the projected activations as

$$h = l \odot (W_0 x), \quad (6)$$

where l is trainable and W_0 remains frozen. IA³ is highly parameter-efficient but can be less expressive than low-rank weight updates.

3.4 Standardised Evaluation Protocol

All results are reported on a held-out test split constructed from the same 10-hour subset to ensure a fair and directly comparable evaluation across methods. We do not use the original NaijaVoices test set because our goal is not to benchmark performance on the full corpus, but to compare adaptation strategies under a fixed and tightly controlled amount of transcribed speech. Evaluating on the full test set would move the assessment outside this controlled setting and make it harder to attribute performance differences to the adaptation method rather than to differences in data conditions.

4 Results and Discussion

4.1 Evaluation Setup

All experiments use the same Whisper-Small checkpoint as the base model and follow the data curation and preprocessing procedure described in Section 3.1.

Training was performed on a single NVIDIA A100 GPU using the AdamW optimiser with a linear learning-rate schedule. For PEFT methods (LoRA, AdaLoRA, and DoRA), we fixed the adapter configuration to rank $r = 32$ and scaling $\alpha = 64$ to enable a controlled, capacity-matched comparison. IA³ does not rely on low-rank adapters and was configured under its standard formulation. During inference, we applied greedy decoding (temperature $T = 0$) to produce deterministic hypotheses.

Multi-seed replication. Because the DoRA-FFT margin is very small (0.1 WER), we additionally quantify run-to-run variability by retraining only the two strongest configurations (FFT and DoRA) across multiple random seeds (42, 43, 44), while keeping the data split, preprocessing, and all hyperparameters fixed. We report mean \pm standard deviation WER on the same held-out test split for these replications.

4.2 Main ASR Performance and Language-Wise Trends

Table 1 summarises the comparative performance of full and parameter-efficient adaptation strategies on the held-out test split from our stratified 10-hour subset. The results reveal a stable ordering across methods: **DoRA** achieves the best macro-average WER, followed closely by FFT (FFT), then AdaLoRA, LoRA, and finally IA³. DoRA attains an average WER of **22.0%**, effectively matching the FFT baseline of 22.1% despite updating only a small fraction of the model parameters. Although the absolute gap between DoRA and FFT (0.1 WER) is narrow and may fall within typical experimental variance, the key implication is methodological. To validate this, we replicate FFT and DoRA across three random seeds and report mean \pm std WER in Table 2. The replicated results support the same interpretation as the single-run comparison: DoRA is statistically comparable to FFT under the 10-hour regime. Under a strict 10-hour constraint, a PEFT approach can achieve parity with FFT, supporting PEFT as a practical alternative when compute and memory budgets are constrained or when FFT is prone to instability.

Language-specific analysis further highlights systematic patterns. **Yoruba** consistently yields the lowest WER across methods (20.1%-23.6%), suggesting stronger transfer from the backbone’s pretraining distribution or more favourable acous-

Method	Params	Hausa	Yoruba	Igbo	Avg.	Rel.
<i>Baseline</i>						
FFT (FFT)	~240M	22.4	20.1	23.8	22.1	1.00×
<i>Parameter-Efficient Methods</i>						
DoRA	~4M	22.1	20.4	23.5	22.0	0.99×
AdaLoRA	~4M	22.6	20.9	24.1	22.5	1.02×
LoRA	~4M	23.4	21.7	24.9	23.3	1.05×
IA ³	<1M	25.3	23.6	27.2	25.4	1.15×

Table 1: WER comparison on the held-out test split from our stratified 10-hour NaijaVoices subset, reported per language (Hausa, Yoruba, Igbo) and as a macro-average across languages for all fine-tuning strategies.

Method (3 seeds)	Hausa	Yoruba	Igbo	Avg.
FFT	22.4 ± 0.18	20.1 ± 0.12	23.8 ± 0.21	22.1 ± 0.14
DoRA	22.1 ± 0.16	20.4 ± 0.10	23.5 ± 0.19	22.0 ± 0.13

Table 2: Seed sensitivity on the held-out test split (same 10-hour subset): mean±std WER across three random seeds (42, 43, 44) for the two top configurations.

tic/phonological alignment under our subset. In contrast, **Igbo** remains the most challenging language (23.5%-27.2%), reflecting the difficulty of learning robust mappings from limited exposure in a tone-sensitive setting where pitch can alter lexical meaning; under such constraints, residual tonal confusions may contribute to substitution and deletion errors even when orthography is preserved. Hausa generally falls between these two extremes. Importantly, DoRA improves upon standard LoRA by **1.3** absolute WER on the macro-average (22.0% vs. 23.3%), indicating that the gains are not merely due to PEFT itself, but to DoRA’s specific architectural refinement. Figure 2 visualises the per-language WER profiles in Table 1, highlighting the consistent trend that Yoruba is easiest and Igbo is most challenging under the 10-hour constraint. At the same time, DoRA remains competitive across all three languages.

4.3 Parameter Efficiency and Practical Implications

Beyond accuracy, a key motivation for PEFT is reducing the number of trainable parameters and the size of task-specific updates. FFT (FFT) updates essentially all Whisper-Small parameters, which typically increases training-time memory requirements because gradients and optimiser states must be stored for all parameters. In contrast, DoRA trains only a small set of additional parameters while achieving closeness to FFT in macro-average WER. A smaller set of trainable parameters also

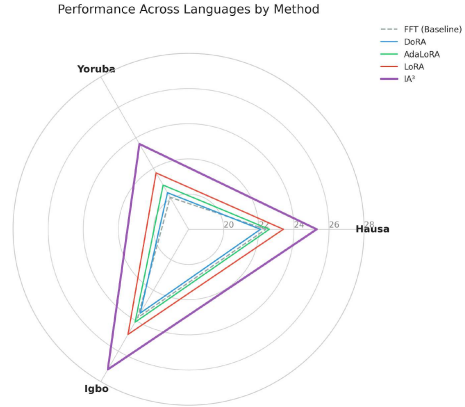


Figure 2: Radar plot of WER (↓) across Hausa, Yoruba, and Igbo. DoRA exhibits near-baseline performance relative to FFT across languages, AdaLoRA and LoRA yield higher WER, and IA³ performs worst; Yoruba attains the lowest WER while Igbo remains most challenging.

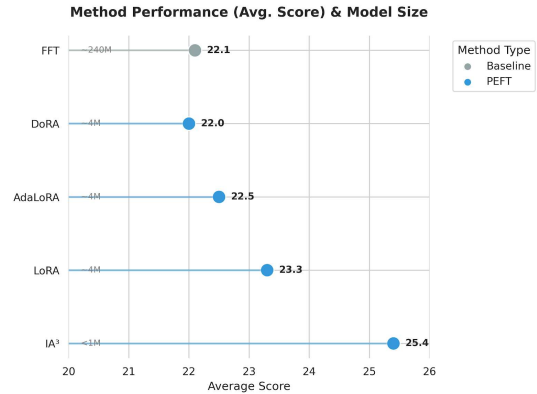


Figure 3: Average WER (↓) plotted against trainable parameters. DoRA achieves FFT-level accuracy (22.0 vs. 22.1) while training ~4M parameters rather than ~240M, whereas IA³ is the most parameter-efficient (<1M) but incurs the largest accuracy degradation.

implies a smaller adaptation footprint, which can reduce the storage required to save and share adapted models and can make iteration more feasible under constrained compute. Figure 3 summarises the accuracy-efficiency trade-off observed in our results.

4.4 Ablation: Contribution of DoRA’s Weight Decomposition

To isolate the contribution of DoRA’s magnitude-direction decomposition, we compare DoRA directly against standard LoRA under matched adapter capacity ($r = 32$, $\alpha = 64$). We treat this comparison as an **architectural ablation**, because LoRA performs only a low-rank directional

update. In contrast, DoRA additionally learns an explicit magnitude component that decouples feature scaling from directional shifts. As shown in Table 1, reverting from DoRA to LoRA increases macro-average WER from 22.0% to 23.3% (+1.3 absolute). This degradation suggests that in extremely low-resource constraints, adapting only via a low-rank directional subspace may be insufficient; the ability to explicitly rescale feature magnitudes helps align pre-trained Whisper representations with the target-language domain, improving stability and generalisation. In this sense, the magnitude component in DoRA is not redundant—it appears to be a critical mechanism for effective adaptation when labelled data is scarce.

4.5 Limitations and Future Work

Despite the encouraging results, several limitations remain. First, the performance gap between DoRA and full fine-tuning is small, so we explicitly quantify run-to-run variability by replicating the two strongest configurations across multiple random seeds (Table 2). However, the remaining PEFT baselines are reported from single runs; broader replication and formal statistical testing across all methods remain important directions for future work.

Second, the study evaluates only three languages, which are linguistically distinct and are drawn from a single national context. Extending the analysis to languages from other African regions and families would help assess whether the observed PEFT–FFT parity generalises across a wider range of linguistic and geographic conditions.

Third, our experiments focus on a single backbone model, Whisper-Small. Future work should examine whether similar parameter-efficiency trends hold for other architectures to clarify the extent to which the findings are model-specific.

Fourth, all evaluations are conducted in-domain on a fixed subset of NaijaVoices. While this controlled setup enables a clean comparison of adaptation strategies, additional experiments under domain shift would better reflect real-world African ASR deployment.

Finally, our analysis considers a single extreme low-resource regime (10 hours of transcribed speech). Exploring multiple data budgets (e.g., 2h, 5h, 20h) would provide insight into when PEFT is most advantageous and how its benefits evolve as more labelled data becomes available.

5 Conclusion

This study evaluated FFT and parameter-efficient adaptation of Whisper-Small for Hausa, Yoruba, and Igbo using a stratified 10-hour NaijaVoices subset, with results reported on the held-out test split (60/20/20 train/val/test). DoRA achieved macro-average performance indistinguishable from FFT (22.0% WER), matching FFT (22.1%) while updating roughly 4M parameters instead of 240M. Multi-seed replication of the two best configurations further supports this parity conclusion under the 10-hour setting. These results show that, in extremely low-resource constraints, structured low-rank adaptation can preserve Whisper’s multilingual priors while delivering FFT-level recognition accuracy with substantially lower training and storage overhead. Treating DoRA versus LoRA as an architectural ablation, the added magnitude–direction decomposition improved average WER by 1.3 points, suggesting that explicit feature rescaling strengthens adaptation stability beyond directional low-rank updates alone. Practically, this makes PEFT a compelling option for African ASR development, where iterative training and deployment are often constrained. While additional robustness evaluation is needed, the findings establish a strong baseline for efficient adaptation and motivate broader studies across domains, dialects, and larger language coverage.

References

- Naira Abdou Mohamed, Anass Allak, Kamel Gaanoun, Imade Benelallam, Zakarya Erraji, and Abdessalam Bahafid. 2024. Multilingual speech recognition initiative for african languages. *International Journal of Data Science and Analytics*, pages 1–16.
- Tejumade Afonja, Tobi Olatunji, Sewade Ogun, and 1 others. 2024. Performant asr models for medical entities in accented speech. *arXiv preprint arXiv:2406.12387*.
- Mohamed Nabih Ali, Daniele Falavigna, and Alessio Brutti. 2025. *Efl-peft: A communication efficient federated learning framework using peft sparsification for asr*. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Oreoluwa Boluwatife Babatunde, Victor Tolulope Olufemi, Emmanuel Bolarinwa, Kausar Yetunde Moshood, and Chris Chinenye Emezue. 2025. Beyond monolingual limits: Fine-tuning monolingual asr for yoruba-english code-switching. In *Proceedings of the 7th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 18–25.

- Sergio Chevtchenko, Nikhil Navas, Rafaella Vale, Franco Ubaudi, Sipumelele Lucwaba, Cally Ardington, Soheil Afshar, Mark Antoniou, and Saeed Afshar. 2025. *An end-to-end approach for child reading assessment in the Xhosa language*, pages 106–119. Lecture Notes in Computer Science. Springer. International Conference on Artificial Intelligence in Education, AIED ; Conference date: 22-07-2025 Through 26-07-2025.
- Chris Emezue, NaijaVoices Community, Busayo Awobade, Abraham Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, and 1 others. 2025. The naijavoices dataset: Cultivating large-scale, high-quality, culturally-rich speech data for african languages. *arXiv preprint arXiv:2505.20564*.
- Sukairaj Hafiz Imam, Tadesse Destaw Belay, Kedir Yassin Husse, Ibrahim Said Ahmad, Idris Abdulmumin, Hadiza Ali Umar, Muhammad Yahuza Bello, Joyce Nakatumba-Nabende, Seid Muhie Yimam, and Shamsuddeen Hassan Muhammad. 2025. Automatic speech recognition (asr) for african low-resource languages: A systematic literature review. *arXiv preprint arXiv:2510.01145*.
- George Joseph and Arun Baby. 2024. *Speaker Personalization for Automatic Speech Recognition using Weight-Decomposed Low-Rank Adaptation*. In *Interspeech 2024*, pages 2875–2879.
- Chin Yuen Kwok, Sheng Li, Jia Qi Yip, and Eng Siong Chng. 2024. Low-resource language adaptation with ensemble of peft approaches. *IEEE*, pages 1–6.
- Salima Mdhaffar, Haroun Elleuch, Fethi Bougares, and Yannick Estève. 2024. *Performance analysis of speech encoders for low-resource SLU and ASR in Tunisian dialect*. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 130–139. Association for Computational Linguistics.
- Tolulope Ogunremi, Kola Tubosun, Anuoluwapo Aremu, Iroko Orife, and David Ifeoluwa Adelani. 2024. *ÌròyìnSpeech: A multi-purpose Yorùbá speech corpus*. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9296–9303, Torino, Italia. ELRA and ICCL.
- Hemant Palivela, Meera Narvekar, David Asirvatham, Shashi Bhusan, Vinay Rishiwal, and Udit Agarwal. 2025. Code-switching asr for low-resource indic languages: A hindi-marathi case study. *IEEE Access*.
- Riefkyanov Surya Adia Pratama and Agit Amrullah. 2024. Analysis of whisper automatic speech recognition performance on low resource language. *Jurnal Pilar Nusa Mandiri*, 20(1):1–8.
- Christopher Simic and Tobias Bocklet. 2024. Self-supervised adaptive av fusion module for pre-trained asr models. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12787–12791. IEEE.
- Luping Wang, Sheng Chen, Linnan Jiang, Shu Pan, Runze Cai, Sen Yang, and Fei Yang. 2024. Parameter-efficient fine-tuning in large models: A survey of methodologies. *arXiv preprint arXiv:2410.19878*.
- Yifan Yang, Zhesu Song, Jianheng Zhuo, Mingyu Cui, Jinpeng Li, Bo Yang, Yexing Du, Ziyang Ma, Xunying Liu, Ziyuan Wang, and 1 others. 2025. Gigaspeech 2: An evolving, large-scale and multi-domain asr corpus for low-resource languages with automated crawling, transcription and refinement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2673–2686.

Real-Time Spoken Instruction Following and Translation in Ugandan Languages

Benjamin Akera, Tim Wenjie Hu, Patrick Walukagga, Evelyn Nafula Ouma, Gilbert Yiga, Engineer Bainomugisha, Ernest Tonny Mwebaze and John Quinn

Sunbird AI, Kampala, Uganda

Corresponding Authors: (info, bakera)@sunbird.ai

Abstract

Many languages are predominantly spoken rather than written, and to bring the benefits of LLMs to speakers of these languages, it is essential that models cater to the voice modality. The typical approach is to cascade ASR, LLM and TTS models together, though this results in systems with high latency, making them unsuitable for natural, real-time interaction. We describe results on taking the encoder part of a Whisper-based model trained to recognise ten languages common in Uganda, and using the Ultravox architecture to project its output directly to the input embedding space of a text model based on Qwen 3 32B, also trained to have comprehension of those languages. The result is a speech LLM with high accuracy and very low latency. For most spoken prompts, we can begin streaming a text response within as low as 50 ms, and a speech audio response within around one second, making real-time spoken interaction with an LLM possible for the first time in these languages. The model is available open source on [Hugging Face](#).

1 Introduction

Speech LLMs are of particular significance for languages which are primarily spoken rather than written. When a practical application calls for speech input to an LLM in a low-resource language, the most common method is to separately train a speech recognition (ASR) model and optionally a machine translation (MT) model, then to simply chain them together such that text is input to the LLM. The drawbacks of this approach are that (1) latency is high, as the ASR and MT models must in turn complete generating tokens before the LLM can begin; (2) errors are amplified along the cascade of models, often leading to nonsensical output in the case of low-resource languages.

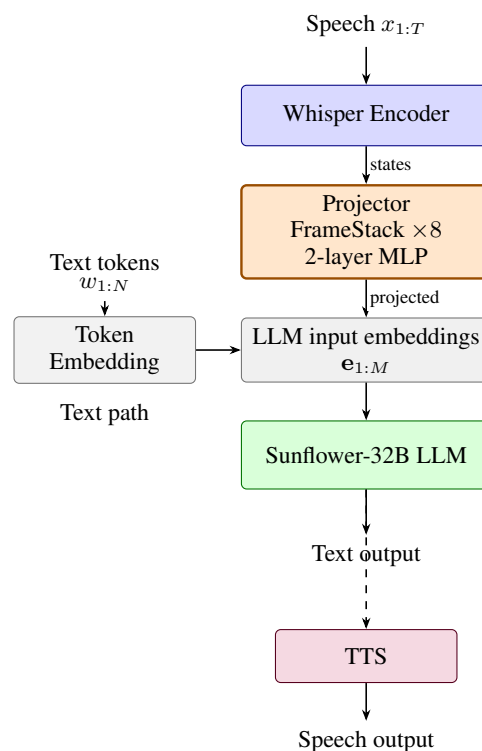


Figure 1: Ultravox-style speech LLM architecture: Whisper encoder states are mapped by a trainable projector into the text model’s input embedding space, so that speech and text prompts can be concatenated. TTS is optional.

In this work, we apply recent findings in coupling ASR and LLM models at a deeper level than text tokens to a low resource setting, specifically Ugandan and East African languages. The Ultravox architecture (Fixie AI, 2024) is a linear multimodal projector that takes a high-dimensional embedding from the encoder part of an ASR model and maps it to the input dimension of an LLM. The resulting fused speech LLM model is trained so that the output is the same regardless of whether prompts are

input as text or read out as speech.

Starting by training both an ASR model and multilingual text LLM for our languages of interest (Luganda, Runyankore, Acholi, Ateso and Lugbara) we then apply this multimodal projector approach to create a single speech model that can respond to audio prompts directly, without needing a separate ASR step. We demonstrate that this results in a model which is very low latency and has high accuracy for speech translation and instruction following.

The contributions of this paper are:

- We describe the development of the first speech LLM model, to our knowledge, capable of real-time translation and responses to spoken instructions in some African languages.
- Rather than adding an Ultravox adapter to an off-the-shelf LLM such as Gemma or Llama, we demonstrate how this can be applied to a new multilingual LLM and ASR models separately trained to have comprehension of several new languages.
- We extend the training pipeline for Ultravox, specifically on the use of Group Relative Policy Optimisation for post-training the LLM.

Furthermore, we make the model publicly available, which we hope will make progress towards important speech applications e.g. in education, healthcare and real-time translation.

2 Related work

Speech Language Models: Recent work has moved beyond cascaded ASR-LLM pipelines towards end to end speech understanding. Qwen Audio (Chu et al., 2023) and Qwen2-Audio (Chu et al., 2024) extend the Qwen LLM family with audio encoders demonstrating strong performance on speech recognition, translation and audio reasoning tasks across several languages. Moshi introduces a fully duplex spoken dialogue system capable of simultaneous listening and speaking with latencies under 200ms (Défossez et al., 2024). Most relevant to our work, Ultravox (Fixie AI, 2024) projects Whisper encoder outputs directly into an LLM’s embedding space via a lightweight multimodal projector, avoiding explicit transcription while preserving the flexibility to use any open-weight LLM

backbone. We adopt this architecture for its modularity: the audio encoder and LLM can be trained independently and then aligned through the projector alone.

Low Resource and African Language ASR:

While large scale models like Whisper (Radford et al., 2023) and MMS (Pratap et al., 2024) include some African languages, coverage remains sparse and performance lags significantly behind high resource languages. Recent efforts have produced targeted datasets such as SALT (Owomugisha et al., 2023; Nakatumba-Nabende et al., 2024), Commonvoice (Ardila et al., 2020), Flores and the NLLB Project (Pratap et al., 2024) enabling fine-tuned models with substantially improved accuracy. However, these advances in ASR have not yet been coupled with LLMs capable of understanding and generating text in these languages. To our knowledge, no prior work has applied speech-native LLM architectures to regional African languages. Existing Speech LLMs focus predominantly on high resource languages or rely on translation to English as an intermediate step. We bridge independently trained components: a Whisper encoder fine-tuned on 10 Ugandan languages and an LLM with native comprehension of those languages, enabling direct spoken interaction without cascaded errors or translation bottlenecks. Our system achieves accuracy comparable with the cascade approach, but with extremely low latency of down to 50ms to first token.

3 Model

Our architecture (Figure 1) follows the Ultravox design (Fixie AI, 2024), comprising three components: (1) an audio encoder that converts speech to continuous representations, (2) a multimodal projector that maps these representations into the LLM’s embedding space, and (3) a large language model that generates text conditioned on the projected audio. Following recent taxonomy (Arora et al., 2025), our system can be characterized as a *speech-aware language model*: an LLM augmented with a speech encoder that processes audio directly, rather than relying on cascaded ASR.

Audio encoder. We use the encoder from a Whisper Large V3 model (Radford et al., 2023) fine-tuned on ten Ugandan languages. Given input audio, the encoder produces a sequence of hidden states $\mathbf{h} \in \mathbf{R}^{T \times 1280}$, where T depends on audio duration. The decoder part of the Whisper model is

discarded; the larger multilingual LLM that we add below becomes the new decoder. The encoder takes 16 KHz audio, used to compute an 80-dimensional log Mel spectrogram on 25 ms windows with a stride of 10 ms. The Whisper encoder layers down-sample the time dimension by a factor of 2, so that each frame of the resulting encoding represents 20 ms of audio.

Multimodal projector. The projector bridges the audio and text representation spaces. Following Ultravox, we first apply a stacking operation that concatenates 8 adjacent frames, reducing sequence length to 6.25 frames per second while increasing dimensionality to $1280 \times 8 = 10240$. This is critical for efficiency: raw encoder outputs would otherwise produce prohibitively long sequences for the LLM’s attention. The stacked features pass through a two-layer MLP with SwiGLU activation and RMS normalization, outputting embeddings matched to the LLM’s hidden dimension.

The projector is trained to minimize cross-entropy loss on transcription tokens. Additionally, we employ KL divergence to encourage projected audio embeddings to match the text embeddings of equivalent transcriptions, ensuring modality alignment. The aim of this training is that the LLM should produce similar output logits for a given input prompt wording, regardless of whether those words are in text or spoken form.

Language model. We use Sunflower-32B (Aker et al., 2025), based on Qwen 3 32B and adapted for Ugandan languages through continued pretraining and instruction tuning on a corpus of Ugandan language data. During inference, projected audio embeddings replace placeholder tokens in the input sequence, and the LLM generates text autoregressively.

Text-to-speech. Text-to-speech output is optional, but preferred to enable fully spoken interaction with the model. We use a version of SparkTTS fine-tuned to produce speech in Ugandan English, Luganda, Acholi, Lugbara, Runyankole and Ateso, and then apply inference optimisations to stream back audio sentence-by-sentence with latency of a few hundred milliseconds.

4 Training

Each component is trained independently before projector alignment.

4.1 Audio Encoder

The Whisper Large V3 model was fine-tuned on the SALT dataset (Owomugisha et al., 2023), which provides transcribed speech across ten languages common in Uganda: English, Luganda, Acholi, Lugbara, Ateso, Runyankole, Lumasaba, Lusoga, Swahili, and Kinyarwanda. Training incorporated language-specific tokens, enabling language-aware recognition without explicit identification as a pre-processing step. The resulting model achieves 1.8% WER on English, 14.2% on Luganda, and 11.1% on Kinyarwanda. A variant trained on approximately 1,400 hours of Kinyarwanda speech achieved 7.1% WER, placing first in the Digital Umuganda ASR competition.¹

4.2 Language Model

Sunflower-32B (Aker et al., 2025) builds on Qwen 3 through continued pretraining on approximately one billion characters of Ugandan language text, including digitized books, transcripts, and parallel corpora across 31 languages. The model then undergoes supervised fine-tuning with LoRA for instruction-following, emphasizing translation tasks, followed by reinforcement learning via Direct Preference Optimization to reduce hallucinations. On translation benchmarks, Sunflower-32B achieves state-of-the-art performance in 24 of 31 Ugandan languages, outperforming GPT-4o and Gemini 2.5 Pro on both $xx \rightarrow \text{eng}$ and $\text{eng} \rightarrow xx$ directions.

4.3 Projector Alignment

With the encoder and LLM frozen, we train only the projector on paired speech-transcription data from SALT. Training follows a chat-based format where each example is structured as:

```
System: You are an ASR assistant that transcribes speech
User: Transcribe in {lang}: <|audio|>
Assistant: {transcription}
```

We modify the Ultravox training recipe to use a specific prompt for translation (“Translate to [language]: [text to be translated]”) which matches the Sunflower LLM instruction-tuning templates. This ensures the LLM recognizes the task framing from its prior training, reducing the burden on the projector to learn both modality alignment and task specification simultaneously.

¹<https://www.kaggle.com/competitions/kinyarwanda-automatic-speech-recognition-track-b>

The `<|audio|>` placeholder is replaced by projected audio embeddings during the forward pass. We use language-aware prompts, inserting the detected language name (e.g., “Luganda”, “Acholi”) to provide the model with explicit language context.

Loss is computed only on the assistant response tokens; system and user tokens are masked. This focuses learning on the transcription task rather than prompt reconstruction. Audio samples are filtered to durations between 0.5 and 25 seconds.

We train for 2 epochs with batch size 4 and gradient accumulation over 8 steps (effective batch size 32). We use AdamW with learning rate 2×10^{-5} , cosine decay schedule, and 500 warmup steps.

Optionally, we apply LoRA to the encoder and LLM for joint fine-tuning: rank 32 for the audio encoder attention layers, and rank 64 for the LLM attention and feed-forward projections. This allows limited adaptation of the frozen backbones while keeping computational costs manageable.

4.4 LLM Post-training

The instruction-tuned LLM (Section 4.2) occasionally responds conversationally to prompts rather than following the literal task instruction. For example, given “Transcribe: Tell me a story about a goat”, the model may generate a story rather than echoing the input text. This behaviour, while inherent for question answering, degrades performance for transcription and translation tasks where verbatim output is required. We address this through two-stage post-training. First, we perform supervised fine-tuning on examples pairing prompts with the expected verbatim output. This establishes the basic transcription behavior. Second, we apply Group Relative Policy Optimization (GRPO) (Guo et al., 2025) with a reward function based on the negative edit distance between the model output and expected transcription. For each prompt, we generate multiple completions and reward those with lower Levenshtein distance to the target, penalizing deviations up to a maximum of 100 characters. This reinforcement learning phase strengthens the model’s adherence to literal transcription even when the input resembles a question or instruction.

5 Results

We evaluate on transcription (WER) and speech translation (BLEU) across the SALT test sets, covering 10 language pairs for translation (5 $xx \rightarrow \text{eng}$

Table 1: Speech translation quality (BLEU \uparrow). *Ours*: Ultravox model. *Cascaded*: Whisper transcription followed by Sunflower translation. *Text-only*: Sunflower translating ground-truth text (oracle upper bound).

Direction	Ours	Cascaded	Text-only
lug \rightarrow eng	38.2	38.3	42.1
eng \rightarrow lug	30.9	31.3	34.3
nyn \rightarrow eng	22.8	24.0	26.0
ach \rightarrow eng	19.9	24.0	24.0
lgg \rightarrow eng	18.4	17.0	23.0
teo \rightarrow eng	18.2	17.5	26.0
eng \rightarrow nyn	17.4	18.0	18.5
eng \rightarrow lgg	18.3	17.5	20.0
eng \rightarrow ach	16.6	16.8	17.5
eng \rightarrow teo	15.0	16.0	17.0
Average	21.6	22.0	24.8

and 5 $\text{eng} \rightarrow xx$ directions) and 8 languages for transcription.

Speech translation. Table 1 compares our Ultravox model against two baselines: (1) a cascaded system that transcribes with Whisper then translates with Sunflower, and (2) an oracle that translates from ground-truth text. Our model matches the cascaded baseline across most language pairs (Table 1), achieving 38.2 BLEU on Luganda \rightarrow English versus 38.3 for the cascaded system. We can therefore match the translation accuracy of a cascaded system, but with very low latency, enabling real-time speech translation in these languages for the first time.

Instruction following. A key advantage of our architecture is that the model naturally generalizes beyond the transcription and translation tasks seen during projector training. When we omit any text instruction and prompt the model with speech audio alone, it defaults to conversational assistant behavior, enabling free-form voice interaction. Table 2 shows an example where the model answers a factual question in Luganda. This emergent capability arises because we train only on response tokens, allowing the LLM’s instruction-following abilities to transfer directly to the speech modality.

Transcription. Table 3 compares our model against the full Whisper encoder-decoder model trained on identical data. For pure transcription, our model underperforms Whisper, as expected: we discard Whisper’s decoder (optimized for transcription) in favor of a general-purpose LLM that enables broader capabilities. When we look at the pattern of errors, the higher resulting word error

Table 2: Example model response when given an audio prompt in Luganda. When prompted with audio and no text instruction, the model defaults to responding as a conversational assistant. The response is generated directly from the audio, without any separate ASR step.


Input audio:

<p>"Yingini y'emmotoka ekola etya?" (How does a car engine work?)</p>
Response:
<p>Yingini y'emmotoka ekola ng'efulumya amaanyi okuva mu mafuta oba amasannyalaze, nga gakyusa amafuta oba amasannyalaze okugafuula amaanyi agasobozesa emmotoka okutambula... (The engine of a car works by converting energy from fuel or electricity into mechanical energy, which is then used to power the wheels of the car...)</p>

Table 3: Transcription quality (WER ↓, median %). The Whisper values are from a standard encoder-decoder model trained on the same data. Our model trades transcription accuracy for flexibility to support different tasks and lower latency.

Language	Ours	Ours (GRPO)	Whisper
English	0.0	0.0	0.0
Luganda	22.6	16.7	5.0
Kinyarwanda	27.8	28.2	1.0
Acholi	41.0	36.9	16.8
Runyankole	43.5	37.5	19.2
Lugbara	55.5	45.8	15.8
Lusoga	58.3	50.0	28.6
Ateso	62.5	58.6	28.6

rates have two causes. First, the model is prone to replying conversationally when asked to transcribe speech containing a question or something that sounds like an instruction. Second, the model often paraphrases the meaning of what was spoken in the text, rather than a word-for-word verbatim transcription. Initial experiments we have carried out with Group Relative Policy Optimisation indicate that this gap can be reduced, which will be the focus of future work.

Latency. Table 4 reports inference latency measured on an A100-80GB GPU using vLLM. Our model achieves a median time-to-first-token (TTFT) of 55ms after warm-up, enabling respon-

Table 4: Latency on A100-80GB. Time-to-first-token (TTFT) measures server-side delay from request receipt to first generated token. The cascaded system cannot produce output until Whisper completes, making true TTFT undefined. Excludes cold-start (~8s).

System	TTFT	Streamable
Ours (p50)	55 ms	✓
Ours (p90)	61 ms	✓
Cascaded	>1 s	×

sive streaming output. The first request after cold-start incurs ~8s latency due to model loading and KV cache initialization; subsequent requests benefit from prefix caching (hit rate >85%).

The cascaded baseline cannot stream output until Whisper completes full transcription. Whisper’s encoder-decoder architecture processes the entire audio context non-causally, precluding incremental output (Radford et al., 2023). This architectural constraint results in latencies exceeding one second for typical utterances, an order-of-magnitude slower than our end-to-end model. This difference is critical for real-time applications such as live translation and voice assistants.

6 Limitations

Several limitations warrant discussion. First, transcription accuracy lags behind the standalone Whisper model, reflecting the trade-off between latency and fidelity when discarding the decoder. Potential mitigations include adding another speech encoder, e.g. based on Wav2Vec2, and concatenating the input embeddings together for a richer representation of the input audio. Second, the model occasionally paraphrases utterances or responds conversationally to questions rather than transcribing verbatim. These behaviors are inherited from the instruction-tuned LLM and can be partially mitigated through reinforcement learning, as shown in Table 3. Finally, our evaluation focuses on translation and transcription; the model’s capability for more complex spoken instructions (e.g., multi-turn dialogue, reasoning over audio) remains unexplored.

Efficiency considerations. Our model requires an A100-80GB GPU for inference, which limits accessibility in resource-constrained settings. Several paths exist for improving efficiency:

1. *Quantization* via GPTQ or AWQ could reduce memory requirements to 20GB (4bit quantization) with minimal quality loss.

2. *Smaller backbones*: The modular architecture allows independent scaling of each component. Using Whisper Small or Medium alongside a 7B–14B LLM would reduce memory requirements from 80GB to under 20GB, with expected trade-offs in transcription accuracy for lower-resource languages.
3. *Distillation*: The speech projector could potentially be transferred to a smaller student LLM. We leave systematic evaluation of these efficiency trade-offs to future work, but note that our architecture’s modularity makes such adaptations straightforward.

7 Conclusion

We presented the first speech-native large language model for Ugandan languages, combining a Whisper encoder fine-tuned on SALT with the Sunflower-32B language model via the Ultravox multimodal projector. Our model responds to free-form speech instructions and matches cascaded Whisper→LLM translation quality while reducing time-to-first-token from over one second to 55 milliseconds, enabling real-time streaming applications for the first time in these languages. Our evaluation focuses on single-turn transcription and translation; multi-turn dialogue, spoken question answering, and constrained instruction-following remain to be systematically evaluated, pending the development of purpose-built benchmarks for Ugandan languages. Nonetheless, qualitative evidence suggests that the LLM’s conversational capabilities transfer effectively to the speech modality.

Future work will extend coverage to additional African languages, improve transcription fidelity through reinforcement learning, and reduce TTS latency to enable full-duplex conversation. We release our models to support further research on speech-native LLMs for low-resource languages.

Acknowledgments

We acknowledge the linguists, educators, and translators from Makerere University whose expertise made this dataset possible. We also thank the Ultravox team for their great open source work, and the anonymous reviewers whose comments improved the paper.

References

- Benjamin Akera, Evelyn Nafula Ouma, Gilbert Yiga, Patrick Walukagga, Phionah Natukunda, Trevor Saaka, Solomon Nsumba, Lilian Teddy Nabukeera, Joel Muhanguzi, Imran Sekalala, and 1 others. 2025. Sunflower: A new approach to expanding coverage of african languages in large language models. *arXiv preprint arXiv:2510.07203*.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the twelfth language resources and evaluation conference*, pages 4218–4222.
- Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, and 1 others. 2024. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.
- Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.
- Fixie AI. 2024. Ultravox: A fast multimodal llm for real-time voice. <https://github.com/fixie-ai/ultravox>. Accessed: 2025-12-19.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shitong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.
- Joyce Nakatumba-Nabende, Claire Babirye, Peter Nabende, Jeremy Francis Tusubira, Jonathan Mukibi, Eric Peter Wairagala, Chodrine Mutebi, Tobias Saul Bateesa, Alvin Nahabwe, Hewitt Tusiime, and 1 others. 2024. Building text and speech benchmark datasets and models for low-resourced east african languages: experiences and lessons. *Applied AI Letters*, 5(2):e92.
- Isaac Owomugisha, Benjamin Akera, Ernest Tonny Mwebaze, and John Quinn. 2023. Multilingual model and data resources for text-to-speech in ugandan languages. In *4th Workshop on African Natural Language Processing*.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, and 1 others. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

SALT-31: A Machine Translation Benchmark Dataset for 31 Ugandan Languages

Solomon Nsumba^{1,2}, Benjamin Akera¹, Evelyn Nafula Ouma¹, Medadi Ssentanda²
Deo Kawalya², Engineer Bainomugisha^{1,2}, Ernest Tonny Mwebaze¹, John Quinn¹

¹Sunbird AI, Kampala, Uganda, ²Makerere University, Kampala, Uganda

Correspondence: snsumba@sunbird.ai

Abstract

We present the SALT-31 benchmark dataset for evaluation of machine translation models covering 31 Ugandan languages. Unlike sentence-level evaluation sets, SALT-31 is constructed from short, scenario-driven mini-dialogues designed to preserve discourse context, pragmatics, and culturally grounded communication patterns common in everyday Ugandan settings. The dataset contains 100 English sentences organized into 20 typical communication scenarios, each represented as a five-sentence mini-sequence. It can therefore be used to evaluate both sentence-level and paragraph level machine translation, and includes nearly every language spoken in a country with high linguistic diversity. It is available at hf.co/datasets/Sunbird/salt-31.

1 Introduction

Evaluation datasets play a central role in advancing machine translation for low-resource languages, particularly in multilingual and linguistically diverse regions such as Africa. Prior work has highlighted that progress in African NLP is often constrained not only by model capacity, but also by the availability of representative, high-quality evaluation resources (Wilhelmina et al., 2020).

In many existing benchmarks, including widely used multilingual evaluation suites, sentences are presented in isolation. While effective for measuring lexical or syntactic adequacy, such sentence-level evaluation often fails to capture discourse phenomena such as coreference, turn-taking, pragmatic intent, and culturally grounded meaning (Bawden et al., 2018; Voita et al., 2018; Goyal et al., 2022).

This limitation is especially pronounced for Ugandan and other African languages, which are predominantly used in spoken, conversational, and community-centered contexts. Recent studies have emphasized the need for context-aware and locally

grounded evaluation protocols to better reflect real-world language use in African settings (Joshi et al., 2020).

The SALT-31 Evaluation Dataset was developed to address this gap by introducing a context-aware evaluation resource grounded in realistic Ugandan communication scenarios. Building upon the Sunbird African Language Technology (SALT) initiative (Akera et al., 2022; SunbirdAI, 2023), SALT-31 emphasizes evaluation quality, linguistic diversity, and contextual coherence rather than large-scale parallel corpora. The dataset targets 31 Ugandan languages, spanning multiple language families, and additionally includes a small number of closely related regional languages. Many of these languages remain severely underrepresented in standard machine translation benchmarks. For several of them, this dataset represents the first time that any publicly available parallel text beyond bib-

Table 1: Communication scenarios in SALT-31, covering formal, informal and everyday modes of communication

ID	Scenario Domain
1	Banking transaction
2	Educational instruction
3	Conversational greetings
4	Narrative/fiction
5	Medical consultation
6	News report
7	Encyclopedic information
8	Market/shopping
9	Emergency response
10	Practical technical guide
11	Health/prenatal advice
12	Official announcement
13	Government speech
14	Agricultural extension
15	Opinion poll response
16	Public transport
17	Family conversation
18	Nutrition guidance
19	Food security
20	Civic office dialogue

Table 2: Summary of the SALT-31 Evaluation Dataset

Property	Value
Number of scenarios	20
Sentences per scenario	5
Total English sentences	100
Target languages	31
Domains	Health, market, family, school, transport, daily life
Unit of evaluation	Five-sentence mini-sequence

Table 3: Generated English text for scenario 10 (practical guide for mechanic or builder)

First, make sure the foundation trench is at least two feet deep and well compacted before laying any bricks.
 Mix the cement, sand, and aggregate in a 1:2:4 ratio for strong concrete, especially for the pillars.
 Use a spirit level to keep the walls straight and check alignment after every few layers of bricks.
 Keep the site clean and water the curing concrete regularly for at least seven days to prevent cracks.
 Finally, make sure all electrical and plumbing points are marked clearly before plastering the walls.

lical translations has been made available.

While multilingual benchmark datasets such as FLORES provide broad coverage of languages, they are typically sentence-level and therefore do not directly probe discourse consistency across turns. Furthermore, they do not capture types of communication relevant to the local context. SALT-31 complements such benchmarks by focusing on short, coherent five-sentence sequences that enable assessment of context-sensitive phenomena (e.g., role consistency, anaphora, pragmatic intent) in realistic Ugandan communication settings.

This paper describes the design principles, data generation pipeline, translation workflow, and release of the SALT-31 Evaluation Dataset. We also summarize its application in evaluating the Sunflower multilingual language model (Akeru et al., 2025), which is currently deployed in a production setting.

2 Data and Methods

An overview of the SALT-31 Evaluation Dataset, including its size, scenario structure, language coverage, and release format, is summarized in Table 2.

2.1 Scenario and Sequence Generation

The English source data was generated using a structured prompt that instructed large language models (LLMs) to produce five-sentence mini-sequences for predefined scenarios. A total of

20 scenarios were defined, each yielding one five-sentence sequence, resulting in 100 English sentences. LLMs were used only to generate controlled English seed text; final sequences were manually reviewed and selected, and all translations were produced by native speakers.

An example scenario is patient-doctor communication in a medical setting, where dialogue captures symptoms, follow-up questions, and clinical reasoning. Other scenarios include everyday greetings, small talk, market negotiations, school interactions, and community discussions. An example for scenario 10, practical instructions for a builder, are shown in Table 3.

To encourage stylistic diversity and reduce model-specific artifacts, multiple LLMs were used to generate candidate sequences, including GPT-4.5, GPT-4o, DeepSeek R1, LLaMA 3.3 70B, Mistral Large, Gemini 2 Flash, and Claude Sonnet. For each scenario, outputs were reviewed and a single sequence was selected based on contextual appropriateness, cultural grounding, and linguistic naturalness for Uganda.

2.2 Data Structuring

Once finalized, the 20 mini-sequences were compiled into a structured spreadsheet, with each row corresponding to a single English sentence and metadata linking it to its scenario and sequence position. This intermediate format facilitated systematic translation and quality control.

2.3 Community-Driven Translation

Translation was carried out in collaboration with experts in mother-tongue education, early literacy development and language policy. Native speakers translated the English sentences into 31 target languages spanning three major families: Bantu (17 languages), Nilotic (11 languages), and Central Sudanic (3 languages). Representative languages from each family are shown in Table 4; the complete language list is provided in Appendix B.

Translations emphasized meaning preservation, natural phrasing, and cultural appropriateness rather than literal word-for-word mapping.

To ensure gold-standard quality, all translations underwent an independent verification process conducted by the Department of Linguistics, English Language Studies and Communication Skills. Teams of trained linguists and native speakers reviewed each translation for semantic fidelity, grammatical correctness, naturalness of expression, and

cultural appropriateness. Discrepancies were resolved through consensus review, with reference to the original English source and the intended communicative context of each scenario. This multi-stage verification process provided an additional layer of quality assurance beyond initial translation, increasing confidence in the reliability of SALT-31 as a gold-standard evaluation dataset.

3 Results and Discussion

SALT-31 evaluation covers 20 distinct communication domains (Table 1), ranging from health-care consultations to agricultural extension contexts identified through community consultation as critical for Ugandan language use. This breadth enables assessment of MT systems across diverse registers, from formal government speeches to informal market negotiations.

Table 4 illustrates the linguistic diversity captured across these scenarios by showing a single English sentence from Scenario 10 (practical construction guide) translated into representative languages from three major families: Bantu (e.g., Luganda, Lusoga), Nilotic (e.g., Acholi, Lango), and Central Sudanic (e.g., Lugbara, Ma'di). These translations reveal systematic structural differences in word order, morphological complexity, and noun class agreement; features that pose distinct challenges for MT systems.

3.1 Baseline Model Performance

We evaluated proprietary models (GPT-4o, Gemini 2.5 Pro, Grok-3) and open-weight alternatives (Sunflower-14B/32B, DeepSeek-Chat, NLLB-1.3B) using chrF as the primary metric, which better captures morphological similarity in agglutinative languages than BLEU (Popović, 2015).

Table 5 presents average performance across all 31 languages. Sunflower models achieve the highest scores in both translation directions, with Sunflower-32B excelling at local-to-English translation (chrF=0.435) and Sunflower-14B performing best for English-to-local (chrF=0.366). Notably, these regionally specialized models outperform substantially larger general-purpose systems. GPT-4o achieves only 0.354 chrF for $xx \rightarrow en$ and 0.235 for $en \rightarrow xx$, despite being substantially larger in scale.

This performance gap reflects fundamental challenges in massively multilingual training. Large-scale models trained on hundreds of languages

must allocate limited parameter capacity across diverse distributions, often resulting in reduced performance on low-resource languages compared to focused regional approaches (Conneau et al., 2020; Arivazhagan et al., 2019). Our results suggest that regional specialization concentrating model capacity on linguistically related languages within a coherent geographic area can yield superior performance for underrepresented languages.

3.2 Performance by Language Family

Performance varies systematically across language families (Table 6). While most models achieve reasonable scores on Bantu languages (mean chrF 0.30–0.41), performance degrades for Nilotic and Central Sudanic languages. GPT-4o achieves 0.323 chrF on Bantu languages but drops to 0.139 for Nilotic and 0.094 for Central Sudanic languages. This 3–4× performance gap exposes significant inequities in current MT systems.

In contrast, Sunflower models maintain more consistent performance across families, with only modest degradation from Bantu (0.413) to Nilotic (0.312) to Central Sudanic (0.299). This consistency validates the effectiveness of training on regionally coherent language groups where structural similarities enable cross-lingual transfer even for extremely low-resource languages.

3.3 Context-Aware Evaluation Findings

The mini-sequence structure enabled qualitative analysis of discourse-level phenomena. Manual inspection of model outputs by native speakers revealed three recurring failure patterns across the mini-sequences:

Coreference errors: Models frequently failed to maintain consistent pronominal reference across sentences. In Scenario 5 (medical consultation), references to "the patient" were inconsistently translated across the five-sentence sequence, sometimes incorrectly switching gender or number.

Register inconsistency: In formal scenarios (e.g., Scenario 13, government speeches), models produced mixed registers, inappropriately combining colloquial and formal constructions within the same sequence.

Cultural misalignment: Technical terms and culturally specific concepts were often mistranslated. In Scenario 14 (agricultural advice), references to traditional farming practices were sometimes rendered with urban or formal terminology inappropriate for rural extension contexts.

Table 4: Example sentence from Scenario 10 translated across three language families, illustrating typological diversity in word order, morphology, and agreement systems

Code	Language	Translation (Scenario 10: Construction Guide)
eng	English	First, make sure the foundation trench is at least two feet deep and well compacted before laying any bricks.
<i>Bantu Family</i>		
lug	Luganda	Ekisooka, kakasa nti omusingi gukka waakiri fuuti bbiri era nga guggumizibbwa bulungi nga tonnassaayo mataffaali.
xog	Lusoga	okusooka, kakasa nti olukonko lw'omusingi lughanvu okuswika fuuti eibiri era nga lwidhiziibwa bukalamu nga okali kutandiika kuzimba ku matafali.
nyn	Runyankole	Ekyokubanza, reeba ngu omusingye gutimbirwe kuhisya fuuti ibiri ahansi kandi gwijwize kurungi, otakatandikire kwombeka amatafaari.
ttj	Rutooro	Ekyokubanza, rora ngu omusingi guhikire fuuti ibiri hansi, gusokiire kurungi otakataireho amatafaali.
<i>Nilotic Family</i>		
ach	Acholi	Me acel, nen ni tut pa bur me te ot olo to romo tyen aryo dok kitoro maber ma pud pe iketo matapwali mo iye.
laj	Lango	Me acel, nen ni bur me pandecon tye ame tuttere romo tyelo aryo dang ocwiny aber ame pwod pe iketo birikkoro keken.
teo	Ateso	Kigeari, kowany ebe idulu aipany na ibokarit ijo bala ipuutin iare ido kibamakina kokwap kojokan eroko ijo inapakina amatapaalin adio kere.
<i>Central Sudanic Family</i>		
lgg	Lugbara	Okoria, I'ba kini 'bile 'bani gale fawundasoni e'dozu abi sizuri ma aliniri ma ovu fiti iri azini 'bama omi eri kililiru denga 'ba 'bani mutufali eyi vaa kuru.
mhi	Ma'di	Atijoa ru rii, kole'a nyi ba fa'undesoni a'a bu ni kolu fiti eri guru vua ure okpo oca matafali bi re ku.

Table 5: Average chrF scores across 31 languages. Specialized regional models outperform general-purpose systems despite smaller size

Model	xx→en	en→xx
<i>Regionally Specialized</i>		
Sunflower-32B	0.435	0.357
Sunflower-14B	0.419	0.366
<i>General Purpose</i>		
Gemini 2.5 Pro	0.408	0.301
GPT-4o	0.354	0.235
Grok-3	0.347	0.247
DeepSeek-Chat	0.308	0.237

These failures underscore SALT-31's value as a diagnostic tool. While aggregate metrics provide useful quality estimates, discourse-level evaluation reveals systematic weaknesses in handling context, a critical requirement for deploying MT systems in real-world Ugandan communication settings.

4 Conclusions

We have introduced the SALT-31 Evaluation Dataset, a context-aware machine translation benchmark covering 31 Ugandan languages across Bantu, Nilotic, and Central Sudanic language families. Unlike conventional sentence-level evaluation sets, SALT-31 structures translation tasks as five-sentence sequences drawn from 20 realistic

Table 6: Average chrF (en→xx) by language family. General-purpose models exhibit 2-3× performance degradation for non-Bantu languages, while Sunflower maintains more consistent quality across families

Model	Bantu	Nilotic	C. Sud.
Sunflower-14B	0.413	0.312	0.299
Sunflower-32B	0.406	0.299	0.295
Gemini 2.5 Pro	0.369	0.230	0.173
GPT-4o	0.323	0.139	0.094
Grok-3	0.299	0.190	0.161
DeepSeek-Chat	0.295	0.172	0.148

communication scenarios that reflect authentic language use in Uganda. The option to group sentences together as mini-documents allows evaluation of models with respect to coreference resolution, register consistency, and cultural grounding phenomena that sentence-level benchmarks cannot capture. Because the dataset is focused on the contexts in which these languages are typically used, evaluation metrics help as diagnostic indicators of models' ability to handle real-world practical applications – even though the small size of the dataset means that more thorough assessment, and especially human evaluation by native speakers, would typically be needed to assess readiness for deployment in a production setting.

Evaluation of state-of-the-art MT systems on

SALT-31 reveals substantial performance gaps. While regionally specialized models like Sunflower achieve robust performance across all groups, general-purpose systems underperform despite their scale. We observed a drastic performance degradation for all models when carrying out machine translation on the lower-resourced of these languages, for which very little training data is available.

Future work will expand SALT-31 to cover code-switching patterns, dialectal variation, and speech modalities, enabling evaluation of speech-to-text systems critical for Uganda where many languages remain primarily oral. SALT-31 is released openly on Hugging Face to support reproducible evaluation and comparative benchmarking.

Limitations

SALT-31 is small (100 English sentences) and is designed for controlled, context-aware evaluation and diagnostic usage, rather than being a dataset which would support fine-grained evaluation or model training. Although scenarios were curated to reflect realistic Ugandan communication, they cannot cover all sociolinguistic registers, dialectal variation, or code-switching patterns present across the 31 languages. Future expansions will increase scenario diversity, incorporate longer contexts, and include additional evaluation protocols involving human judgments.

Acknowledgments

We acknowledge the linguists, educators, and translators from Makerere University whose expertise made this dataset possible. We also thank the Sunbird AI team for supporting dataset preparation, hosting, and evaluation.

References

- Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Nagayi, Claire Babirye, Isaac Owomugisha, Solomon Nsumba, Joyce Nakatumba-Nabende, Engineer Bainomugisha, Ernest Mwebaze, and John Quinn. 2022. 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, and 1 others. 2025. Sunflower: A new approach to expanding coverage of african languages in large language models. *arXiv preprint arXiv:2510.07203*.
- Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, and 1 others. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.
- Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8440–8451.
- 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.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. *arXiv preprint arXiv:2004.09095*.
- Maja Popović. 2015. chrF: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.
- SunbirdAI. 2023. Salt: Sunbird african language technology. <https://github.com/SunbirdAI/salt>.
- Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. *arXiv preprint arXiv:1805.10163*.
- Nekoto Wilhelmina, Marivate Vukosi, Matsila Tshinondiwa, Fasubaa Timi, Fagbohunge Taiwo, Akinola Solomon Oluwole, Muhammad Shamsuddeen, Kabenamualu Salomon Kabongo, Osei Salomey, Sackey Freshia, and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.

5 Appendices

A Prompt Used for Scenario Generation

We used the following prompt to generate English mini-sequences for all scenarios:

You are helping design a multilingual evaluation set for machine translation for use in Uganda. For each scenario, create a short dialogue or paragraph consisting of exactly 5 sentences, capturing natural, contextually realistic communication that would make sense in a Ugandan context. Use plain, accessible language, and keep the tone appropriate to the scenario (e.g., professional for business, conversational for daily life, etc.).

B Complete Language List

Table 7 presents the complete list of 31 target languages, organized by language family.

#	Code	Language	#	Code	Language
<i>Bantu (17 languages)</i>					
1	cgg	Rukiga	10	nyn	Runyankole
2	gwr	Lugwere	11	nyo	Runyoro
3	kin	Kinyarwanda	12	rub	Lugungu
4	koo	Rukonjo	13	ruc	Ruruuli
5	lsm	Samia	14	rwm	Kwamba
6	lug	Luganda	15	swa	Swahili
7	myx	Lumasaba	16	tlj	Lubwisi
8	nuj	Lunyole	17	ttj	Rutooro
9	xog	Lusoga			
<i>Nilotic (9 languages)</i>					
18	ach	Acholi	23	kpz	Kupsabiny
19	adh	Dhopadhola	24	laj	Lango
20	alz	Alur	25	pok	Pokot
21	kdi	Kumam	26	teo	Ateso
22	kdj	Karamojong			
<i>Central Sudanic (5 languages)</i>					
27	bfa	Bari	30	luc	Aringa
28	keo	Kakwa	31	mhi	Ma'di
29	lgg	Lugbara			

Table 7: Complete list of 31 target languages in SALT-31, organized by language family.

Sample-Size Scaling of the African Languages NLI Evaluation

Anuj Tiwari¹, Oluwapelumi Ogunremu², Terry Oko-odion³, Jesujuwon Egbewale⁴,
Hannah Nwokocha⁵

Noida Institute of Engineering and Technology¹, ML Collective^{1,2,3,4,5}
aj11anuj123@gmail.com, ogunremuoluwapelumi@gmail.com, terryokoodion@gmail.com,
egbewalejesujuwon7@gmail.com, hannahsopuruchi@gmail.com

Abstract

African languages have very little labelled data, and it is unclear if augmenting the quantity of annotation data reliably enhances downstream performance. The study is a systematic sample-size scaling study of natural language inference (NLI) on 16 African languages based on the AfriXNLI benchmark. Under controlled conditions, two multilingual transformer models with roughly 0.6B parameters XLM-R Large fine-tuned on XNLI and AfroXLM-R Large are tested on sample sizes of between 50 and 500 labeled examples and average their results across random subsampling runs. As opposed to the usual belief of monotonic increase with increased data, we find a strongly language-sensitive and often non-monotonic scaling behavior. Some languages show early saturation or decrease in performance with sample size as well as high variance in low resource regimes. These results indicate that the volume of data is not enough to guarantee stable profits to African NLI, creating the necessity of language-sensitive datasets creation and stronger multilingual modelling strategies.

1 Introduction

The latest advancements in NLP have been fuelled by massive pretraining and access to large amounts of labeled data. The advances have however benefited more high resource languages unfairly and many African languages are still underexamined in training and evaluation standards. Therefore, the key problem of multilingual and low-resource NLP is to learn the performance of the models in relation to the amount of labeled data available. One of the most widely used assumptions in machine learning is that as more and more data is labelled, the better the performance will be downstream. Although this assumption is usually true in high-resource environments, it has not yet been carefully studied in the case of the low-resource languages especially the African languages with various typological and

morphological characteristics. Practically, annotation is expensive, and expansion of dataset without a clear indication of the benefit can be inefficient or even counterproductive.

This paper provides an analysis of the behavior of natural language inference (NLI) with respect to the amount of labeled data in African language models by using the AfriXNLI benchmark. Rather than proposing new models or datasets, our objective is to empirically characterize scaling behavior, performance stability, and variance across languages and models under controlled experimental conditions. In particular, we answer the following research questions:

- **RQ1:** Does larger labeled data positively affect NLI performance when using African languages in AfriXNLI?
- **RQ2:** How does scaling behavior vary across languages and models?
- **RQ3:** To what extent are observed trends stable under random subsampling?

By answering these questions, we hope to offer empirical recommendations to dataset construction and evaluation practice in African NLP, as well as requirements on expectations of data scaling in low-resource semantic reasoning problems.

2 Related Work

Multilingual and African NLP Benchmarks
The current research has contributed to the development of African-language NLP by extending the current standards and developing new assessment tools. AfriXNLI is a human-translated version of the XNLI benchmark of various African languages, allowing to evaluate the human natural language inference in low-resource conditions (Community, 2024) in a unified way.

MasakhaNER offers a named entity recognition system on a large scale on ten African languages, and this project proves that community-driven sets construction are effective in African NLP (Adelani et al., 2021). AfroLID presents neural language identification toolkit, spanning 517 languages in Africa, and greatly increasing the coverage of languages compared to the previous multi-lingual systems (Adebara et al., 2022). These combined efforts spell out long-term development towards determining assessment materials of the African languages in the context of multilingual NLP.

Scaling Data Laws and Efficiency In high-resource settings, language models exhibit predictable scaling behavior. (Kaplan et al., 2020) demonstrate that language modeling loss reduces according to power-law dependencies on both model size and dataset size. (Hoffmann et al., 2022) also show that compute-optimal training depends in proportionately more data, as the Chinchilla model can outperform much larger models trained on smaller data. (Muennighoff et al., 2023) however establish that performance improvements reduce quickly in data-constrained environments, and more compute or repeated data produces only small increases in performance. These results cast some doubts on the fact that classical scaling laws can be directly applied to low-resource and multilingual settings.

Data Scaling in Low-Resource NLP (Eiselen and Gaustad, 2023) look into the impact of the training data size on performance in African languages with particular attention to the morphologically diverse languages of South Africa. They demonstrate that although small data sets can be used to obtain useful models, languages with complex, conjunctive morphology need considerably more data to give similar performance. The importance of linguistic structure in relation to data efficiency is brought out in this work. Nonetheless, they only tested embedding-based models and problems like part-of-speech tagging and they pose the open question of behavior with data scaling in semantic reasoning problems and contemporary fine-tuned pretrained language models.

African Languages: Scarcity of Data and Benchmarking Systemic under-representation (Hussen et al., 2025) report that today only a tiny share of the 2000+ languages of Africa have been trained on modern large language models,

and that the field of African languages has been far under-represented 15 compared to its representation across the world. (Adebara and Abdul-Mageed, 2022) attributes such scarcity to the fact that African languages are structurally unsupported by the current large language model development, and are significantly underrepresented relative to their global distribution. The latest benchmark projects like AfroBench (Ojo et al., 2025) and IrokoBench (Adelani et al., 2025) extend assessment to African languages and task categories, like reasoning and natural language understanding. Even with this extended coverage, it is evident that these benchmarks always indicate significant performance differences between African and high-resource languages, and that there are still continued issues in modeling and evaluation.

Multilingual Representation Models Multilingual encoders such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020), which have training based on pretraining on multiple languages, are commonly used as multilingual NLP baselines. (Conneau et al., 2020) demonstrate that multilingual pretraining significantly improves the cross-lingual test of XLM-R, especially on low-resource languages. Although these models can be shown to be effective in zero-shot transfer, their response to an incremental scale of data of a single African language has not been thoroughly explored. Our study supplements this literature by giving an empirical examination of sample-size scaling action of African-languages NLI.

3 Experimental Setup

3.1 Dataset

We use the AfriXNLI benchmark, which consists of the sentence pairs of NLI translated to various African languages. We use 16 languages in our experiments which adopt a variety of language families, scripts, and typological properties which are represented in AfriXNLI. All tests are performed on the test splits. Simulating the various labeled data regimes, we adjust the number of test examples to be evaluated by randomly subsampling, but we do not adjust the model parameters.

3.2 Models

To allow us to compare pretraining strategies, we assess two multilingual transformer models from similar architectures with about 0.6 billion parameters. The first model, XLM-R Large fine-

tuned on XNLI (Davison, 2020) is a powerful task-aligned multilingual baseline constructed on the XLM-R model (Conneau et al., 2020). The second model, AfroXLM-R Large (Alabi et al., 2022) is an African-based form of XLM-R trained with more focus on African-language-based data.

With the choice of similar scale and architecture model, we factor out the influence of pretraining data composition and language coverage and reduce the impact of model size.

3.3 Evaluation

For each language-model pair, we evaluate performance at sample sizes ranging from 50 to 500 examples. To control variance due to the selection of the data, we run several random subsampling runs to calculate the mean and standard deviation between the runs of a given sample size.

The most common metric that we report is accuracy, however, we also report precision and F1-score. Such an assessment plan allows us to differentiate between systematic scaling effects and those caused by sampling.

4 Results

4.1 Evaluation Variance between Sample Sizes

We initially analyse the patterns of the evaluation variance with the change in size of the sample. The standard deviation of the accuracy is reported in Figure 1 in the aggregate form over all the languages and models. The maximum variance is in low-resource samples (50-100 examples) and is sharply decreasing with increase in the sample size that reaches a point of an average of 300 samples after which the variance is constant.

This tendency shows that small sets of evaluation produce very unstable performance estimates which are highly dependent on the specific sets of samples that one is analyzing. With an increase in sample size, the variance decreases implying that the bigger the evaluation sets, the more true model performance is likely to have been estimated.

Figure 2 also further modifies this effect by model. Both XLM-R Large and AfroXLM-R Large have the following qualitative trend: large variance at the beginning of the sample size, and this is followed by a sharp rise in the value of the sample size. Although the absolute levels of different variance are slightly different, the overall trend is maintained in all models which implies that the

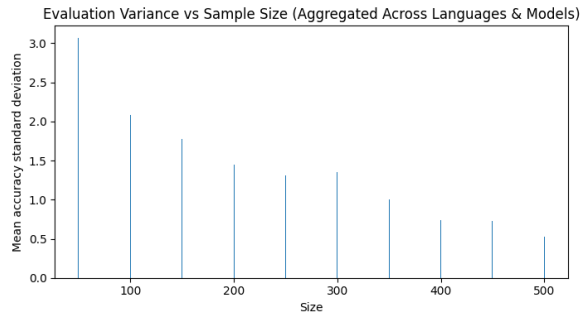


Figure 1: Evaluation variance (standard deviation of accuracy) as a function of sample size, aggregated across all languages and models. The variance decreases very rapidly with the size of the sample meaning that evaluation regimes with low resources are unstable.

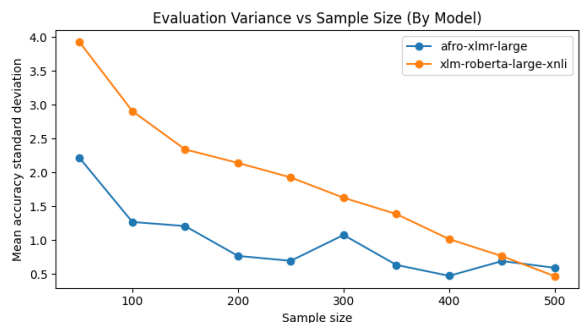


Figure 2: Comparison of evaluation variance with sample size, separately by XLM-R Large and AfroXLM-R Large. Both models show high variance in low-resource regimes and stabilize with larger evaluation sets.

instability of evaluation is not particular to a given pretraining strategy.

4.2 Trends of Global Scaling across Languages and Models

As the sample size increases, the variance reduces, however, accuracy may not necessarily increase monotonically. Figure 3 provides a heatmap of scaling slopes of every language model pair, indicating whether the performance improves, stays constant or reduces with increased sample size.

The heatmap indicates that there is significant heterogeneity in languages. There are some nearly zero or slightly positive slopes, that is, weak gains or early saturation, and also negative slopes, that is, systematic degradation in performance with increasing evaluation sets becoming larger and more representative. Such patterns exist in both models, indicating that the scaling behaviour is highly language-specific, as opposed to being model-driven.

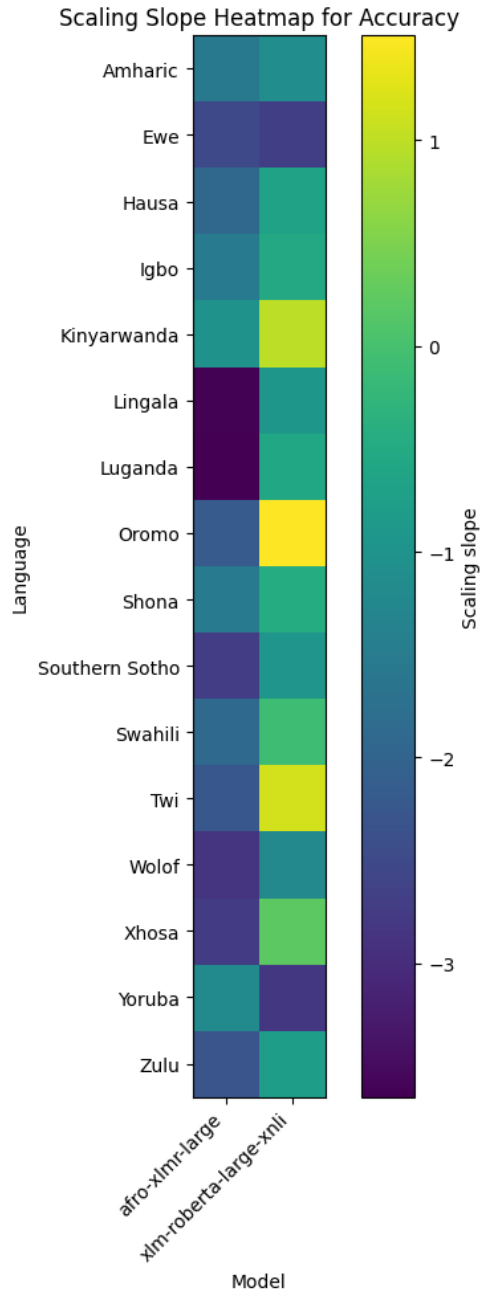


Figure 3: Scaling slope heatmap on accuracy between language-model pairs. Positive slopes mean that the performance improves as the sample size increases and negative slope means that the performance deteriorates.

4.3 Scaling Behavior Dependent on Language: Yoruba vs Kinyarwanda

To illustrate these trends concretely, we analyze scaling behavior for Yoruba and Kinyarwanda under each model. Figure 4 shows results for XLM-R Large. Yoruba exhibits pronounced small-sample optimism, with relatively high accuracy at 50 examples followed by a consistent decline as sample size increases. This monotonic degradation suggests

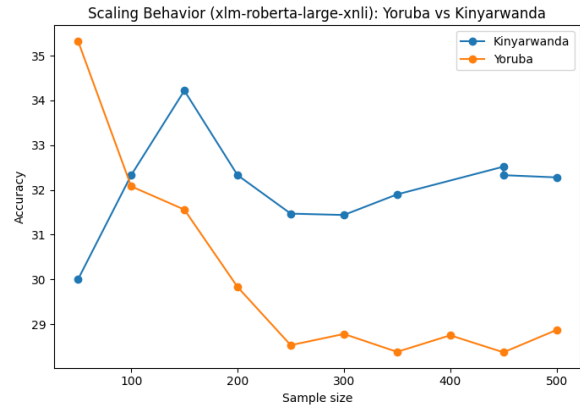


Figure 4: Yoruba and Kinyarwanda evaluation scaling behaviour with XLM-R Large. Yoruba experiences monotonic deterioration as the sample size increases and Kinyarwanda experiences initial improvement and afterwards saturation.

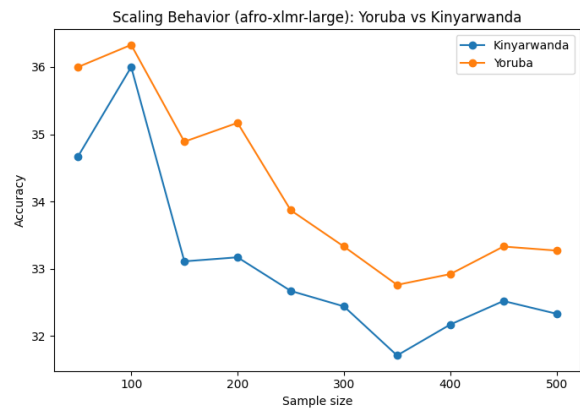


Figure 5: Yoruba and Kinyarwanda AfroXLM-R Large evaluation scaling behavior. The non-monotonic tendencies that are specific to language prevail within models.

that small evaluation subsets overestimate performance, masking systematic errors that emerge with broader coverage.

Conversely, there is a slight rise in performance of Kinyarwanda up to around 150 examples after which it starts to decrease and stabilize. At bigger sample sizes, variance collapses, meaning that it can no longer be measured with its actual performance level by the smaller subsets.

The same comparison is made in Figure 5 on AfroXLM-R Large. The qualitative trends are similar: Yoruba demonstrates again decreasing accuracy with the growing sample size at the same time Kinyarwanda demonstrates initial gains and then stabilities. The fact that such trends are maintained in the models supports the conclusion that the scaling behavior is more language-specific than model-specific.

5 Discussion

Our findings dispute the widely held belief that more uniformly augmented data is always better to use when applying African-language natural language inference. In languages, non-monotonic scaling behavior in evaluation accuracy is observed, where information other than data quantity influences it, e.g. the distribution of labels, ambiguity in translations and representativeness of evaluation subsets contribute significantly.

Notably, we come up with conclusions about AfriXNLI only and these need not be generalized as being applicable in all African languages. However, the similarity in trends observed at random subsampling runs and between two different multilingual models indicates that it is systematic, and not incidental.

Evaluation bias and small-sample optimism

Among the main conclusions that this research made is that small sample evaluation is systematically biased in overestimating the performance of the models. The figure below (Figures 1 and 2) indicated that at sample sizes smaller than 200 evaluation variance is high resulting in unstable and optimistic bias accuracy estimates. With increase in sample size, both variance and accuracy collapse and this is not due to decreasing model quality but rather due to harder instances, neutral cases, and translation ambiguities displayed by larger evaluation sets that are underrepresented in small samples. This point is crucial: we are examining reliability of evaluation and not learning curves, and the apparent decline in performance indicates the lower estimation bias and not the alterations of the model behavior.

Language-specific non-monotonic scaling Figure 3, the scaling slope heatmap, indicates that there is a lot of heterogeneity between languages. Although in some languages, the slopes are weakly positive or near-zero, in others, the slopes are negative, which means that the performance of a language deteriorates as sets of evaluations increase. These patterns are present in both models and it is revealed that scaling behavior is more language dependent than model dependent. The differences between Yoruba and Kinyarwanda case studies, Figures 4 and 5, illustrate this contrast quite well: Yoruba has high small-sample optimism and decreases in a monotonic way, whereas Kinyarwanda has small gains at the start and stagnates. These

variations indicate that an individual set evaluation size might provide inaccurate results when used across languages when used consistently.

Pretraining strategies Model effects Even though AfroXLM-R Large can achieve significantly better accuracy and reduced variance at very small sample sizes than XLM-R Large fine-tuned on XNLI, both models have similar qualitative scaling behaviors across languages. Africa-centric pretraining improves initial stability but does not eliminate non-monotonic scaling or language specific evaluation bias. This implies that data composition pretraining is not enough to consider the heterogeneity of African-language NLI assessment and that the choice of the model cannot influence not only stability in the evaluation but also absolute performance.

Benchmarking and evaluation implications

These findings have direct methodological implications. Single scores on small test sets of accuracy can significantly exaggerate the ability of models to perform under low resource conditions. The larger sets of evaluation decrease variance and bias but can indicate lower real performance and therefore makes comparing studies more difficult. We also suggest that African language benchmarks should report variance between subsamples, should not over rely on small held-out sets, and should take into account language specific evaluation sizes instead of fixed-size test sets.

Overall, we find that the amount of data is not sufficient to ensure credible assessment on the part of African NLI. Rather, meaningful benchmarking in a low resource multilingual setting requires representative sampling, meticulous dataset construction, and stability analysis to contribute to it.

Evaluation stability and saturation In order to measure stability in evaluation, we approximate a saturation value of each language-model combination, which is the minimum size of the evaluation where the mean accuracy is varied by at most $\pm 0.5\%$. The smallest sample size at which additional increments in evaluation data do not produce significant performance variations is called the n^* .

$$n^* = \min \left\{ n \mid \max_{m>n} |A(m) - A(n)| \leq \epsilon \right\}.$$

In case there is no $n \leq 500$, we declare saturation point as > 500 . We are not training, but only testing. Here saturation refers to the amount of

Table 1: The saturation points are estimated at a sample size of ">500" at which the average error levels off within $\pm 0.5\%$ between the languages and models.

Language	xlm-roberta-large-xnli	afro-xlmr-large
Amharic	450	250
Ewe	250	400
Hausa	400	400
Igbo	300	250
Kinyarwanda	450	400
Lingala	300	350
Luganda	350	450
Oromo	400	400
Shona	200	400
Southern Sotho	300	350
Swahili	300	450
Twi	450	400
Wolof	>500	450
Xhosa	250	400
Yoruba	250	400
Zulu	300	400

evaluation data one only needs to achieve the stabilization of the estimated performance.

Table 1 provides the summary of these saturation points in all languages and models. We find very great language to language disparity. In some languages (e.g., Shona, Yoruba, Xhosa) as few as 200 or 250 evaluation samples can give the required stable performance estimates, whereas in others (e.g., Swahili, Kinyarwanda, Oromo) it can take 400 or 450 samples. It is worth noting that Wolof underestimates performance even in 500 samples of XLM-R Large, and this implies that Wolof remains unstable in performance estimation.

These variations are much the same across models indicating that language and dataset specific factors are the primary causes of saturation behaviour as opposed to model architecture itself. Overall, these findings indicate that model performance can be significantly misestimated using fixed-size evaluation benchmarks in the case of African languages, and that the required volume of evaluation data to make reliable estimates can differ significantly across languages.

6 Limitations

The limitations of our study are as follows:

- **Dataset scope** All experiments are conducted on AfriXNLI; thus, observed trends may reflect dataset-specific properties such as translation artifacts or label distribution biases.
- **Evaluation vs learning** We get the evaluation behavior as opposed to the dynamics of

learning. Models do not optimize on successively large training sets hence results reflect the stability and bias of the performance estimates, not the improvement in performance with more training data.

- **Model scale** The scale of experiments is restricted to 2 models of multilingual size 0.6b. Relationships between data scaling and scales of models are not investigated.

Regardless of these constraints, the fact that trends were similar among languages, models, and random subsampling runs implies that we have been able to capture systematic elements of evaluation reliability to African NLI.

7 Future Work

This analysis can be developed in several ways in future work:

- **Broader tasks and datasets** The generalizability of the found evaluated scaling behaviour of the study should be tested by extending the study to other African NLP benchmark tasks, like sentiment analysis or named entity recognition.
- **Linguistic and dataset effects** Adding linguistic metadata, label distributions and tokenization statistics can be useful to explain language behavioral specifics of saturation and non-monotonic scaling.
- **Learning dynamics** Learning scaling behavior: Fine-tuning, but not evaluation, would help illuminate the effect of the addition of labeled data on real model learning with African languages.

8 Conclusion

Here we provide a detailed study of scale behavior in terms of sample-size on the AfriXNLI benchmark using African languages. Our results based on controlled evaluation in 16 languages and two multilingual models and a series of random subsampling runs demonstrate that growth in evaluation data does not come at uniform or monotonic benefits. Rather, scaling behavior is very language-specific, usually non monotonic, and hugely influenced by evaluation variation under low resource limitations.

We show that the performance estimates of small evaluation subsets are often optimistically biased, whereas the estimates of larger subsets are indicative of latent difficulty and their estimates are more stable. This demonstrates the difference between the evaluation reliability and model learning as one of the key issues in the African NLP.

On our results, we suggest: (i) do not use single-point assessment on very small test sets and report the mean \pm standard deviation on more than one subsample, (ii) supplement aggregate measures with per-class measures and (iii) use at least 300 evaluation samples unless otherwise, and consider results below this scale as high-noise measures. In a broader sense, our paper warns on naive beliefs about the reliability of increased information as a means to have dependable evaluation and the need to have practices that are evaluation conscious in benchmarking of African languages.

References

- Ife Adebara and Muhammad Abdul-Mageed. 2022. [Towards afrocentric NLP for African languages: Where we are and where we can go](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841. Association for Computational Linguistics.
- Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. [AfroLID: A neural language identification tool for African languages](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981. Association for Computational Linguistics.
- David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, and 1 others. 2021. [MasakhaNER: Named entity recognition for African languages](#). *Transactions of the Association for Computational Linguistics*, pages 1116–1131.
- David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, and 1 others. 2025. [IrokoBench: A new benchmark for African languages in the age of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757. Association for Computational Linguistics.
- Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and 1 others. 2022. [Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning](#). In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics.
- Masakhane NLP Community. 2024. [AfriXnli: Dataset](#).
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, and 1 others. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- Joe Davison. 2020. [Xlm-roberta large fine-tuned on xnli](#). Hugging Face model.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Roald Eiselen and Tanja Gaustad. 2023. [Deep learning and low-resource languages: How much data is enough? a case study of three linguistically distinct south african languages](#). In *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 42–53. Association for Computational Linguistics.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, and 1 others. 2022. [Training compute-optimal large language models](#). In *Proceedings of NeurIPS*.
- Kedir Yassin Hussen, Walelign Tewabe Sewunetie, Abinew Ali Ayele, Sukairaj Hafiz Imam, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. [The state of large language models for african languages: Progress and challenges](#). *ArXiv*, abs/2506.02280.
- Jared Kaplan, Sam McCandlish, T. J. Henighan, and 1 others. 2020. [Scaling laws for neural language models](#). *ArXiv*, abs/2001.08361.
- Niklas Muennighoff, Sébastien Candido, Jonathan Vamalké, and 1 others. 2023. Scaling data-constrained language models. <https://sllnk.com/wyS64>.
- Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, and 1 others. 2025. [AfroBench: How good are large language models on African languages?](#) In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095. Association for Computational Linguistics.

A Appendix - Full Results for models

Table 2: Evaluation of Swahili with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	33.33 \pm 1.89	0.339 \pm 0.009	0.331 \pm 0.014
100	34.08 \pm 1.41	0.341 \pm 0.014	0.340 \pm 0.014
150	32.89 \pm 0.31	0.332 \pm 0.002	0.330 \pm 0.008
200	33.33 \pm 0.85	0.335 \pm 0.004	0.334 \pm 0.006
250	33.87 \pm 1.61	0.341 \pm 0.015	0.340 \pm 0.016
300	32.89 \pm 1.59	0.331 \pm 0.015	0.330 \pm 0.015
350	32.95 \pm 1.52	0.332 \pm 0.015	0.331 \pm 0.015
400	33.25 \pm 1.14	0.335 \pm 0.011	0.334 \pm 0.011
450	33.19 \pm 0.93	0.333 \pm 0.010	0.333 \pm 0.009
500	33.13 \pm 0.41	0.331 \pm 0.004	0.331 \pm 0.004

Table 3: Evaluation of Lingala with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	34.00 \pm 0.00	0.346 \pm 0.027	0.323 \pm 0.010
100	29.67 \pm 1.70	0.293 \pm 0.026	0.281 \pm 0.021
150	31.33 \pm 0.94	0.308 \pm 0.028	0.294 \pm 0.014
200	30.83 \pm 1.84	0.301 \pm 0.030	0.284 \pm 0.022
250	30.67 \pm 1.91	0.302 \pm 0.029	0.283 \pm 0.022
300	31.89 \pm 1.66	0.314 \pm 0.018	0.296 \pm 0.018
350	31.71 \pm 1.46	0.313 \pm 0.015	0.294 \pm 0.015
400	31.67 \pm 1.36	0.309 \pm 0.015	0.293 \pm 0.014
450	31.93 \pm 0.90	0.311 \pm 0.008	0.295 \pm 0.008
500	31.87 \pm 0.82	0.312 \pm 0.008	0.294 \pm 0.008

Table 4: Evaluation of Igbo with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	31.33 \pm 1.89	0.424 \pm 0.008	0.314 \pm 0.026
100	32.00 \pm 3.56	0.456 \pm 0.011	0.314 \pm 0.040
150	32.44 \pm 2.06	0.458 \pm 0.036	0.315 \pm 0.024
200	31.33 \pm 1.93	0.441 \pm 0.020	0.306 \pm 0.020
250	30.67 \pm 1.80	0.424 \pm 0.009	0.298 \pm 0.019
300	29.89 \pm 1.50	0.406 \pm 0.015	0.299 \pm 0.018
350	29.62 \pm 1.28	0.408 \pm 0.013	0.299 \pm 0.014
400	30.08 \pm 0.82	0.403 \pm 0.009	0.296 \pm 0.010
450	29.85 \pm 0.46	0.406 \pm 0.005	0.293 \pm 0.005
500	30.08 \pm 0.33	0.409 \pm 0.005	0.294 \pm 0.003

Table 5: Evaluation of Hausa with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	32.67 \pm 2.49	0.317 \pm 0.032	0.316 \pm 0.026
100	32.67 \pm 1.79	0.335 \pm 0.007	0.327 \pm 0.011
150	31.78 \pm 1.37	0.317 \pm 0.017	0.316 \pm 0.015
200	32.80 \pm 1.08	0.314 \pm 0.006	0.316 \pm 0.008
250	32.53 \pm 2.93	0.325 \pm 0.026	0.325 \pm 0.027
300	32.56 \pm 3.22	0.326 \pm 0.029	0.325 \pm 0.031
350	32.10 \pm 1.94	0.321 \pm 0.021	0.321 \pm 0.020
400	31.42 \pm 1.84	0.317 \pm 0.016	0.315 \pm 0.017
450	31.19 \pm 1.18	0.317 \pm 0.012	0.314 \pm 0.012
500	31.13 \pm 0.38	0.318 \pm 0.006	0.314 \pm 0.005

Table 6: Evaluation of Yoruba with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	35.33 \pm 2.49	0.439 \pm 0.060	0.342 \pm 0.035
100	32.08 \pm 2.83	0.421 \pm 0.077	0.296 \pm 0.026
150	31.56 \pm 2.27	0.396 \pm 0.046	0.291 \pm 0.020
200	29.83 \pm 1.43	0.391 \pm 0.020	0.283 \pm 0.021
250	28.53 \pm 1.24	0.367 \pm 0.010	0.268 \pm 0.020
300	28.78 \pm 0.68	0.359 \pm 0.007	0.271 \pm 0.013
350	28.38 \pm 0.36	0.353 \pm 0.005	0.268 \pm 0.008
400	28.75 \pm 0.89	0.369 \pm 0.017	0.272 \pm 0.011
450	28.37 \pm 1.06	0.358 \pm 0.018	0.269 \pm 0.011
500	28.87 \pm 0.52	0.358 \pm 0.011	0.273 \pm 0.005

Table 7: Evaluation of Kinyarwanda with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	30.00 \pm 4.32	0.455 \pm 0.064	0.305 \pm 0.031
100	32.33 \pm 3.09	0.430 \pm 0.040	0.325 \pm 0.023
150	34.22 \pm 5.45	0.440 \pm 0.040	0.337 \pm 0.050
200	32.33 \pm 4.50	0.420 \pm 0.041	0.318 \pm 0.043
250	31.47 \pm 3.09	0.412 \pm 0.019	0.311 \pm 0.030
300	31.44 \pm 1.75	0.402 \pm 0.012	0.309 \pm 0.018
350	31.90 \pm 1.75	0.406 \pm 0.015	0.312 \pm 0.016
400	32.33 \pm 1.45	0.415 \pm 0.011	0.319 \pm 0.015
450	32.52 \pm 0.38	0.412 \pm 0.004	0.320 \pm 0.003
500	32.28 \pm 0.16	0.403 \pm 0.001	0.315 \pm 0.003

Table 8: Evaluation of Zulu with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	32.08 \pm 4.32	0.337 \pm 0.057	0.319 \pm 0.048
100	31.33 \pm 2.49	0.349 \pm 0.026	0.323 \pm 0.029
150	33.33 \pm 0.54	0.358 \pm 0.011	0.342 \pm 0.009
200	31.17 \pm 2.09	0.359 \pm 0.012	0.327 \pm 0.016
250	30.88 \pm 1.96	0.351 \pm 0.014	0.326 \pm 0.017
300	30.56 \pm 2.01	0.341 \pm 0.012	0.328 \pm 0.017
350	30.67 \pm 2.38	0.340 \pm 0.018	0.321 \pm 0.021
400	30.80 \pm 1.47	0.338 \pm 0.010	0.316 \pm 0.013
450	30.07 \pm 0.93	0.339 \pm 0.006	0.317 \pm 0.008
500	30.33 \pm 0.77	0.341 \pm 0.004	0.319 \pm 0.006

Table 9: Evaluation of Amharic with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	33.33 \pm 6.18	0.333 \pm 0.058	0.332 \pm 0.059
100	33.67 \pm 4.11	0.333 \pm 0.043	0.334 \pm 0.041
150	32.08 \pm 3.49	0.332 \pm 0.022	0.325 \pm 0.028
200	31.58 \pm 2.27	0.333 \pm 0.020	0.323 \pm 0.021
250	32.13 \pm 1.36	0.337 \pm 0.086	0.329 \pm 0.010
300	31.67 \pm 0.82	0.327 \pm 0.094	0.321 \pm 0.006
350	31.85 \pm 0.59	0.327 \pm 0.092	0.318 \pm 0.006
400	31.88 \pm 0.82	0.329 \pm 0.009	0.319 \pm 0.008
450	30.96 \pm 0.91	0.329 \pm 0.010	0.319 \pm 0.010
500	30.73 \pm 0.52	0.327 \pm 0.006	0.317 \pm 0.004

Table 10: Evaluation of Southern sotho with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	32.00 \pm 3.27	0.420 \pm 0.040	0.308 \pm 0.036
100	28.33 \pm 4.11	0.397 \pm 0.022	0.268 \pm 0.027
150	29.78 \pm 4.43	0.383 \pm 0.093	0.278 \pm 0.036
200	30.08 \pm 4.14	0.399 \pm 0.027	0.285 \pm 0.037
250	29.73 \pm 2.54	0.376 \pm 0.021	0.281 \pm 0.023
300	29.44 \pm 2.20	0.355 \pm 0.028	0.277 \pm 0.021
350	29.14 \pm 1.42	0.354 \pm 0.028	0.275 \pm 0.013
400	29.33 \pm 1.25	0.357 \pm 0.020	0.279 \pm 0.011
450	29.48 \pm 0.46	0.356 \pm 0.011	0.280 \pm 0.004
500	29.80 \pm 0.49	0.357 \pm 0.010	0.282 \pm 0.004

Table 11: Evaluation of Oromo with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	24.00 \pm 12.96	0.269 \pm 0.118	0.246 \pm 0.126
100	27.67 \pm 4.64	0.333 \pm 0.047	0.295 \pm 0.046
150	28.44 \pm 2.57	0.347 \pm 0.022	0.305 \pm 0.023
200	28.00 \pm 2.04	0.341 \pm 0.026	0.300 \pm 0.022
250	28.80 \pm 1.31	0.346 \pm 0.011	0.308 \pm 0.013
300	28.89 \pm 0.83	0.342 \pm 0.019	0.306 \pm 0.011
350	28.76 \pm 0.97	0.342 \pm 0.019	0.306 \pm 0.012
400	27.67 \pm 0.82	0.337 \pm 0.013	0.297 \pm 0.010
450	28.15 \pm 0.90	0.339 \pm 0.011	0.301 \pm 0.010
500	27.47 \pm 0.34	0.333 \pm 0.005	0.294 \pm 0.004

Table 12: Evaluation of Twi with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	25.33 \pm 4.71	0.276 \pm 0.034	0.214 \pm 0.034
100	27.00 \pm 3.74	0.319 \pm 0.064	0.225 \pm 0.033
150	27.78 \pm 3.94	0.349 \pm 0.049	0.237 \pm 0.040
200	27.00 \pm 2.83	0.332 \pm 0.053	0.231 \pm 0.029
250	26.00 \pm 2.47	0.316 \pm 0.045	0.224 \pm 0.023
300	26.78 \pm 1.50	0.302 \pm 0.031	0.232 \pm 0.015
350	27.33 \pm 1.66	0.322 \pm 0.017	0.240 \pm 0.015
400	27.08 \pm 1.01	0.324 \pm 0.022	0.237 \pm 0.008
450	27.93 \pm 0.21	0.334 \pm 0.003	0.244 \pm 0.001
500	28.00 \pm 0.43	0.328 \pm 0.014	0.245 \pm 0.005

Table 13: Evaluation of Shona with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	26.00 \pm 5.66	0.301 \pm 0.122	0.240 \pm 0.064
100	23.00 \pm 4.55	0.256 \pm 0.104	0.208 \pm 0.050
150	25.33 \pm 3.81	0.319 \pm 0.063	0.237 \pm 0.043
200	24.50 \pm 2.27	0.315 \pm 0.055	0.233 \pm 0.029
250	24.53 \pm 2.00	0.311 \pm 0.032	0.233 \pm 0.024
300	24.67 \pm 1.52	0.307 \pm 0.027	0.234 \pm 0.018
350	24.18 \pm 0.94	0.298 \pm 0.013	0.228 \pm 0.011
400	24.42 \pm 1.30	0.314 \pm 0.020	0.233 \pm 0.014
450	24.81 \pm 1.34	0.311 \pm 0.014	0.235 \pm 0.012
500	24.93 \pm 0.77	0.310 \pm 0.009	0.235 \pm 0.008

Table 14: Evaluation of Xhosa with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	29.33 \pm 2.49	0.369 \pm 0.043	0.316 \pm 0.031
100	29.33 \pm 2.87	0.346 \pm 0.027	0.311 \pm 0.027
150	29.78 \pm 3.62	0.354 \pm 0.019	0.317 \pm 0.031
200	30.33 \pm 3.09	0.357 \pm 0.017	0.324 \pm 0.027
250	29.60 \pm 2.36	0.354 \pm 0.024	0.319 \pm 0.024
300	29.33 \pm 2.87	0.348 \pm 0.028	0.315 \pm 0.029
350	29.62 \pm 2.58	0.345 \pm 0.027	0.316 \pm 0.026
400	29.92 \pm 1.12	0.349 \pm 0.013	0.320 \pm 0.012
450	30.00 \pm 0.54	0.353 \pm 0.005	0.321 \pm 0.003
500	29.80 \pm 0.43	0.349 \pm 0.007	0.319 \pm 0.004

Table 15: Evaluation of Wolof with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	33.33 \pm 5.25	0.396 \pm 0.058	0.306 \pm 0.024
100	32.67 \pm 2.62	0.398 \pm 0.040	0.311 \pm 0.020
150	33.33 \pm 0.54	0.392 \pm 0.015	0.321 \pm 0.008
200	32.67 \pm 1.70	0.399 \pm 0.021	0.317 \pm 0.018
250	31.68 \pm 1.50	0.389 \pm 0.030	0.305 \pm 0.016
300	32.80 \pm 1.66	0.375 \pm 0.034	0.306 \pm 0.017
350	31.24 \pm 1.28	0.373 \pm 0.023	0.301 \pm 0.013
400	31.58 \pm 0.47	0.383 \pm 0.008	0.304 \pm 0.002
450	31.26 \pm 0.38	0.379 \pm 0.004	0.301 \pm 0.003
500	30.53 \pm 0.57	0.375 \pm 0.007	0.294 \pm 0.006

Table 16: Evaluation of Luganda with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	32.80 \pm 1.63	0.470 \pm 0.037	0.324 \pm 0.020
100	30.08 \pm 0.82	0.420 \pm 0.020	0.294 \pm 0.004
150	32.22 \pm 0.83	0.428 \pm 0.023	0.315 \pm 0.005
200	32.58 \pm 0.71	0.421 \pm 0.013	0.325 \pm 0.003
250	31.68 \pm 0.86	0.402 \pm 0.009	0.314 \pm 0.009
300	32.22 \pm 0.79	0.395 \pm 0.007	0.318 \pm 0.007
350	31.24 \pm 0.49	0.384 \pm 0.007	0.318 \pm 0.005
400	31.25 \pm 0.35	0.388 \pm 0.006	0.311 \pm 0.002
450	31.63 \pm 0.28	0.395 \pm 0.003	0.315 \pm 0.004
500	31.53 \pm 0.34	0.393 \pm 0.002	0.313 \pm 0.004

Table 17: Evaluation of Ewe with xlm-roberta-large-xnli. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	36.00 \pm 3.27	0.424 \pm 0.154	0.318 \pm 0.049
100	33.00 \pm 2.16	0.468 \pm 0.042	0.284 \pm 0.018
150	32.22 \pm 1.26	0.417 \pm 0.012	0.278 \pm 0.010
200	30.08 \pm 1.47	0.390 \pm 0.010	0.260 \pm 0.011
250	29.73 \pm 1.86	0.385 \pm 0.025	0.255 \pm 0.015
300	30.22 \pm 1.40	0.388 \pm 0.020	0.260 \pm 0.011
350	29.52 \pm 1.52	0.392 \pm 0.027	0.259 \pm 0.016
400	30.00 \pm 0.54	0.400 \pm 0.012	0.265 \pm 0.005
450	29.85 \pm 0.64	0.394 \pm 0.013	0.261 \pm 0.004
500	29.80 \pm 0.16	0.393 \pm 0.009	0.257 \pm 0.003

Table 18: Evaluation of Swahili with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	37.33 \pm 0.94	0.184 \pm 0.067	0.240 \pm 0.057
100	35.33 \pm 0.94	0.169 \pm 0.052	0.223 \pm 0.042
150	33.33 \pm 0.94	0.210 \pm 0.072	0.206 \pm 0.040
200	34.67 \pm 0.62	0.222 \pm 0.082	0.219 \pm 0.044
250	33.07 \pm 0.19	0.202 \pm 0.067	0.204 \pm 0.045
300	32.89 \pm 0.68	0.210 \pm 0.074	0.202 \pm 0.044
350	32.19 \pm 0.13	0.204 \pm 0.077	0.197 \pm 0.045
400	32.42 \pm 0.62	0.204 \pm 0.077	0.198 \pm 0.041
450	32.96 \pm 0.55	0.207 \pm 0.074	0.201 \pm 0.038
500	33.00 \pm 0.57	0.201 \pm 0.068	0.201 \pm 0.038

Table 19: Evaluation of Lingala with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	41.33 \pm 3.40	0.246 \pm 0.073	0.294 \pm 0.060
100	37.67 \pm 1.25	0.217 \pm 0.061	0.257 \pm 0.050
150	37.33 \pm 2.72	0.216 \pm 0.072	0.253 \pm 0.065
200	35.83 \pm 1.43	0.202 \pm 0.061	0.238 \pm 0.051
250	33.87 \pm 0.68	0.187 \pm 0.054	0.218 \pm 0.039
300	33.56 \pm 0.42	0.187 \pm 0.053	0.215 \pm 0.035
350	32.86 \pm 0.62	0.183 \pm 0.057	0.209 \pm 0.042
400	32.42 \pm 0.24	0.179 \pm 0.053	0.207 \pm 0.038
450	32.96 \pm 0.55	0.207 \pm 0.074	0.201 \pm 0.038
500	33.00 \pm 0.57	0.201 \pm 0.068	0.201 \pm 0.038

Table 20: Evaluation of Igbo with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	36.67 \pm 0.94	0.164 \pm 0.039	0.215 \pm 0.023
100	35.67 \pm 1.25	0.157 \pm 0.034	0.208 \pm 0.018
150	34.67 \pm 0.54	0.271 \pm 0.148	0.201 \pm 0.031
200	34.50 \pm 0.71	0.262 \pm 0.137	0.201 \pm 0.025
250	32.93 \pm 0.19	0.254 \pm 0.135	0.187 \pm 0.025
300	32.89 \pm 0.87	0.261 \pm 0.135	0.186 \pm 0.029
350	32.48 \pm 0.49	0.259 \pm 0.139	0.184 \pm 0.031
400	32.75 \pm 0.41	0.216 \pm 0.084	0.187 \pm 0.032
450	32.96 \pm 0.42	0.217 \pm 0.080	0.189 \pm 0.027
500	33.13 \pm 0.19	0.217 \pm 0.082	0.189 \pm 0.027

Table 21: Evaluation of Hausa with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	36.67 \pm 0.94	0.238 \pm 0.064	0.236 \pm 0.030
100	35.67 \pm 0.47	0.206 \pm 0.054	0.223 \pm 0.032
150	33.56 \pm 0.31	0.192 \pm 0.054	0.208 \pm 0.033
200	33.67 \pm 0.24	0.179 \pm 0.050	0.201 \pm 0.034
250	32.40 \pm 0.86	0.171 \pm 0.051	0.188 \pm 0.037
300	31.09 \pm 1.50	0.163 \pm 0.046	0.185 \pm 0.037
350	31.43 \pm 0.62	0.157 \pm 0.050	0.188 \pm 0.037
400	31.83 \pm 0.31	0.163 \pm 0.048	0.185 \pm 0.034
450	32.15 \pm 1.00	0.169 \pm 0.042	0.188 \pm 0.028
500	32.27 \pm 0.84	0.166 \pm 0.040	0.188 \pm 0.028

Table 22: Evaluation of Zulu with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	38.00 \pm 1.63	0.148 \pm 0.017	0.213 \pm 0.020
100	37.00 \pm 0.82	0.192 \pm 0.083	0.205 \pm 0.015
150	35.11 \pm 1.57	0.196 \pm 0.114	0.189 \pm 0.023
200	34.17 \pm 0.24	0.174 \pm 0.081	0.181 \pm 0.009
250	32.88 \pm 0.86	0.166 \pm 0.085	0.167 \pm 0.012
300	32.22 \pm 1.23	0.138 \pm 0.047	0.162 \pm 0.012
350	31.81 \pm 0.49	0.136 \pm 0.050	0.158 \pm 0.008
400	32.33 \pm 0.12	0.138 \pm 0.048	0.162 \pm 0.005
450	32.59 \pm 0.69	0.159 \pm 0.058	0.165 \pm 0.004
500	32.73 \pm 0.50	0.149 \pm 0.057	0.165 \pm 0.003

Table 23: Evaluation of Yoruba with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	36.00 \pm 4.32	0.220 \pm 0.069	0.239 \pm 0.024
100	36.33 \pm 1.25	0.210 \pm 0.057	0.244 \pm 0.041
150	34.89 \pm 0.63	0.208 \pm 0.066	0.228 \pm 0.049
200	35.17 \pm 1.03	0.201 \pm 0.061	0.233 \pm 0.051
250	33.87 \pm 0.94	0.196 \pm 0.061	0.221 \pm 0.050
300	33.33 \pm 0.47	0.193 \pm 0.056	0.216 \pm 0.042
350	32.76 \pm 0.71	0.190 \pm 0.062	0.214 \pm 0.048
400	32.92 \pm 0.51	0.188 \pm 0.059	0.214 \pm 0.046
450	33.33 \pm 0.48	0.188 \pm 0.053	0.217 \pm 0.041
500	33.27 \pm 0.34	0.191 \pm 0.057	0.216 \pm 0.038

Table 24: Evaluation of Kinyarwanda with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	34.67 \pm 3.40	0.166 \pm 0.041	0.219 \pm 0.027
100	36.00 \pm 1.63	0.280 \pm 0.141	0.231 \pm 0.038
150	33.11 \pm 1.26	0.189 \pm 0.052	0.205 \pm 0.038
200	33.17 \pm 1.55	0.185 \pm 0.049	0.206 \pm 0.033
250	32.67 \pm 0.50	0.182 \pm 0.051	0.200 \pm 0.046
300	32.44 \pm 1.50	0.172 \pm 0.045	0.197 \pm 0.049
350	31.71 \pm 0.62	0.168 \pm 0.048	0.191 \pm 0.049
400	32.17 \pm 0.12	0.165 \pm 0.047	0.194 \pm 0.046
450	32.52 \pm 0.73	0.166 \pm 0.041	0.196 \pm 0.041
500	32.33 \pm 0.82	0.162 \pm 0.039	0.194 \pm 0.038

Table 25: Evaluation of Amharic with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	36.67 \pm 0.94	0.167 \pm 0.043	0.223 \pm 0.033
100	37.33 \pm 1.89	0.186 \pm 0.076	0.229 \pm 0.050
150	34.67 \pm 1.44	0.163 \pm 0.068	0.203 \pm 0.045
200	34.33 \pm 0.85	0.161 \pm 0.064	0.208 \pm 0.038
250	32.80 \pm 1.18	0.149 \pm 0.062	0.183 \pm 0.035
300	32.67 \pm 1.66	0.152 \pm 0.069	0.183 \pm 0.038
350	32.57 \pm 1.68	0.181 \pm 0.064	0.183 \pm 0.041
400	33.08 \pm 1.24	0.192 \pm 0.065	0.188 \pm 0.037
450	33.04 \pm 0.73	0.184 \pm 0.053	0.188 \pm 0.029
500	33.07 \pm 0.47	0.182 \pm 0.052	0.187 \pm 0.027

Table 26: Evaluation of Southern sotho with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	39.33 \pm 3.40	0.202 \pm 0.092	0.238 \pm 0.054
100	37.08 \pm 0.82	0.174 \pm 0.058	0.211 \pm 0.022
150	34.67 \pm 0.54	0.264 \pm 0.150	0.191 \pm 0.017
200	34.50 \pm 0.71	0.261 \pm 0.137	0.192 \pm 0.014
250	33.33 \pm 0.50	0.226 \pm 0.089	0.182 \pm 0.018
300	33.08 \pm 0.94	0.226 \pm 0.087	0.179 \pm 0.019
350	32.57 \pm 0.81	0.208 \pm 0.075	0.174 \pm 0.022
400	32.67 \pm 0.42	0.184 \pm 0.058	0.177 \pm 0.018
450	32.89 \pm 0.48	0.193 \pm 0.057	0.179 \pm 0.009
500	33.07 \pm 0.25	0.192 \pm 0.057	0.180 \pm 0.011

Table 27: Evaluation of Oromo with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	38.00 \pm 3.27	0.186 \pm 0.038	0.230 \pm 0.016
100	37.33 \pm 1.25	0.287 \pm 0.143	0.228 \pm 0.030
150	35.56 \pm 1.13	0.231 \pm 0.082	0.212 \pm 0.034
200	34.83 \pm 0.62	0.216 \pm 0.073	0.204 \pm 0.030
250	33.33 \pm 0.50	0.204 \pm 0.073	0.190 \pm 0.022
300	33.08 \pm 0.54	0.186 \pm 0.052	0.187 \pm 0.021
350	32.19 \pm 0.49	0.177 \pm 0.053	0.179 \pm 0.023
400	32.50 \pm 0.35	0.179 \pm 0.053	0.182 \pm 0.025
450	32.81 \pm 0.52	0.192 \pm 0.058	0.185 \pm 0.023
500	33.00 \pm 0.33	0.193 \pm 0.059	0.186 \pm 0.025

Table 28: Evaluation of Twi with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	37.33 \pm 0.94	0.179 \pm 0.058	0.236 \pm 0.049
100	35.33 \pm 1.25	0.194 \pm 0.048	0.238 \pm 0.034
150	33.11 \pm 0.83	0.188 \pm 0.053	0.213 \pm 0.033
200	33.33 \pm 1.03	0.186 \pm 0.051	0.213 \pm 0.032
250	31.73 \pm 1.00	0.175 \pm 0.047	0.199 \pm 0.030
300	31.67 \pm 1.25	0.177 \pm 0.046	0.197 \pm 0.028
350	31.05 \pm 0.75	0.171 \pm 0.049	0.198 \pm 0.031
400	31.83 \pm 0.62	0.177 \pm 0.052	0.196 \pm 0.033
450	31.85 \pm 1.21	0.173 \pm 0.044	0.195 \pm 0.025
500	32.13 \pm 1.09	0.175 \pm 0.046	0.196 \pm 0.026

Table 29: Evaluation of Shona with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	36.00 \pm 1.63	0.174 \pm 0.053	0.229 \pm 0.042
100	36.00 \pm 0.00	0.173 \pm 0.058	0.228 \pm 0.050
150	33.78 \pm 0.31	0.156 \pm 0.058	0.208 \pm 0.051
200	34.00 \pm 0.00	0.155 \pm 0.055	0.208 \pm 0.048
250	32.67 \pm 0.75	0.146 \pm 0.057	0.196 \pm 0.052
300	31.67 \pm 1.44	0.139 \pm 0.049	0.187 \pm 0.044
350	31.52 \pm 0.49	0.138 \pm 0.054	0.186 \pm 0.049
400	32.50 \pm 0.54	0.144 \pm 0.057	0.194 \pm 0.052
450	32.67 \pm 0.65	0.144 \pm 0.051	0.195 \pm 0.045
500	32.53 \pm 0.74	0.180 \pm 0.048	0.194 \pm 0.049

Table 30: Evaluation of Xhosa with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	39.33 \pm 2.49	0.258 \pm 0.149	0.238 \pm 0.028
100	37.67 \pm 0.94	0.358 \pm 0.156	0.214 \pm 0.013
150	35.33 \pm 1.44	0.347 \pm 0.164	0.198 \pm 0.016
200	35.00 \pm 0.41	0.306 \pm 0.193	0.189 \pm 0.009
250	33.47 \pm 1.05	0.307 \pm 0.153	0.174 \pm 0.014
300	32.89 \pm 1.40	0.277 \pm 0.146	0.169 \pm 0.016
350	32.38 \pm 0.71	0.260 \pm 0.137	0.163 \pm 0.018
400	32.83 \pm 0.31	0.263 \pm 0.135	0.167 \pm 0.006
450	32.89 \pm 0.63	0.219 \pm 0.093	0.167 \pm 0.007
500	33.00 \pm 0.43	0.219 \pm 0.092	0.168 \pm 0.005

Table 31: Evaluation of Wolof with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	39.33 \pm 1.89	0.236 \pm 0.067	0.241 \pm 0.031
100	38.08 \pm 1.63	0.223 \pm 0.076	0.226 \pm 0.030
150	34.67 \pm 1.44	0.172 \pm 0.055	0.198 \pm 0.028
200	34.17 \pm 0.24	0.171 \pm 0.040	0.194 \pm 0.015
250	32.67 \pm 0.50	0.166 \pm 0.041	0.183 \pm 0.013
300	32.11 \pm 1.10	0.159 \pm 0.034	0.177 \pm 0.006
350	31.81 \pm 0.13	0.160 \pm 0.041	0.174 \pm 0.014
400	32.17 \pm 0.31	0.160 \pm 0.040	0.177 \pm 0.015
450	32.44 \pm 1.01	0.163 \pm 0.040	0.179 \pm 0.013
500	32.73 \pm 0.81	0.170 \pm 0.044	0.182 \pm 0.015

Table 32: Evaluation of Luganda with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	41.33 \pm 3.40	0.298 \pm 0.140	0.282 \pm 0.052
100	35.67 \pm 2.05	0.219 \pm 0.072	0.235 \pm 0.031
150	34.22 \pm 2.45	0.210 \pm 0.075	0.222 \pm 0.035
200	33.33 \pm 1.31	0.199 \pm 0.065	0.213 \pm 0.030
250	32.67 \pm 1.05	0.196 \pm 0.067	0.205 \pm 0.030
300	32.08 \pm 1.36	0.184 \pm 0.052	0.198 \pm 0.026
350	31.43 \pm 0.62	0.176 \pm 0.053	0.192 \pm 0.032
400	31.67 \pm 0.82	0.176 \pm 0.052	0.194 \pm 0.031
450	32.44 \pm 0.79	0.181 \pm 0.049	0.201 \pm 0.033
500	32.93 \pm 0.66	0.183 \pm 0.051	0.205 \pm 0.038

Table 33: Evaluation of Ewe with afro-xlmr-large. Measures are provided of means \pm standard deviation.

Size	Accuracy (%)	Precision	F1
50	39.33 \pm 1.89	0.233 \pm 0.070	0.260 \pm 0.042
100	38.00 \pm 2.83	0.226 \pm 0.071	0.252 \pm 0.051
150	36.44 \pm 1.75	0.224 \pm 0.078	0.241 \pm 0.050
200	35.33 \pm 1.25	0.203 \pm 0.063	0.227 \pm 0.041
250	33.20 \pm 0.33	0.188 \pm 0.055	0.207 \pm 0.029
300	32.78 \pm 0.83	0.185 \pm 0.051	0.208 \pm 0.022
350	32.67 \pm 0.75	0.192 \pm 0.064	0.208 \pm 0.031
400	33.08 \pm 0.59	0.192 \pm 0.063	0.203 \pm 0.033
450	33.26 \pm 0.58	0.193 \pm 0.057	0.204 \pm 0.027
500	33.53 \pm 0.82	0.195 \pm 0.059	0.207 \pm 0.030

Evaluating Yoruba Text-to-Speech Systems for Accessible Computer-Based Testing in Visually Impaired Learners

Kausar Moshood^{1,2}, Victor Olufemi^{1,3}, Oreoluwa Babatunde¹,
Emmanuel Bolarinwa¹, and Williams Oluwademilade¹

¹LyngualLabs

²Oregon State University

³Carnegie Mellon University

{kausar,victor,oreoluwa,emmanuel,williams}@lynguallabs.org

Abstract

Text-to-Speech (TTS) technology offers potential to improve exam accessibility for visually impaired learners, but existing systems often underperform in underrepresented languages like Yoruba. This study evaluates current Yoruba TTS models in delivering standardized exam content to five visually impaired students through a web-based interface. Before testing, four Yoruba TTS systems were compared; only Facebook’s mms-tts-yor and YarnGPT produced intelligible Yoruba speech. Students experienced exam questions delivered by human voice, Braille, and TTS. All preferred Braille for clarity and independence, some valued human narration, while TTS was least favored due to robotic and unclear output. These results reveal a significant gap between TTS capabilities and the needs of users in low-resource languages. The paper highlights the urgency of developing tone-aware, user-centered TTS solutions to ensure equitable access to digital education for visually impaired speakers of underrepresented languages.

1 Introduction

Computer-Based Testing (CBT) has become a widely adopted format in modern educational assessments, thanks to benefits like scalability, automation, and reduced logistical costs. For example, most exams today from school evaluations to professional certifications use computer-based delivery (Patel et al., 2021). This approach streamlines administration for large candidate numbers while cutting out printing and distribution expenses (Chukwuma-Nosike and Chukwuma, 2023). However, this digital transition also presents serious accessibility challenges for visually impaired learners. Studies have shown that blind or low-vision students often face barriers with standard CBT interfaces, limiting their ability to independently access test content (Patel et al., 2021). These difficulties are especially pronounced in underrepresented

language contexts, where assistive tools may not support the local language script or speech (Tubosun, 2023).

Assistive technologies such as screen readers and Braille displays have facilitated access to digital education for many blind users. Screen readers, for instance, convert on-screen text to speech or Braille and are a primary interface for blind computer users (American Foundation for the Blind, 2025). Yet the effectiveness of these tools varies significantly depending on the language and infrastructure involved. Many African languages, including Yoruba, Igbo, and Hausa, have historically been marginalized in technology, resulting in limited support in mainstream screen readers and voice assistants (Tubosun, 2023). Indeed, until recently, no major smart assistant could operate in any indigenous African language, forcing users to rely on English for voice feedback (Tubosun, 2023). This gap means that visually impaired students in such language contexts are at a higher risk of exclusion when exams move to digital formats.

Text-to-Speech (TTS) systems offer a promising path for making CBTs more accessible to blind and low-vision learners. These systems convert written content into spoken output in real time, enabling visually impaired students to independently interact with computer-based questions and answers (Dhaliwal and Sharma, 2024). TTS technology has been hailed for “revolutionizing the world by enabling disabled people to access information and achieve independence” (Dhaliwal and Sharma, 2024). In an exam setting, a reliable TTS voice could allow a blind test-taker to listen to questions and select answers without needing a human reader or proctor, and without the delays of Braille translation. However, most state-of-the-art TTS models today are developed and optimized for high-resource languages like English and Chinese, leaving a significant performance gap when they are applied to low-resource languages with tonal and complex

structures such as Yoruba (Ogunremi et al., 2024).

Yoruba is a major Nigerian language spoken by over 40 million people across West Africa (Eberhard et al., 2019). It is also one of the few indigenous African languages offered as a subject in standardized national exams like those administered by the West African Examinations Council (WAEC) for example, there is a dedicated Yoruba language paper in the WAEC senior secondary exams (waec). Despite its large speaker base and official status in education, little is known about how well existing Yoruba TTS systems can perform in high-stakes CBT environments for visually impaired learners. To date, there has been virtually no published research on applying Yoruba TTS for exam delivery, which indicates a clear knowledge gap that this study aims to fill.

In this study, we investigate the feasibility and effectiveness of using AI-driven Yoruba TTS systems for administering CBT-style exam questions to visually impaired secondary school students. By engaging five blind students in an experimental setup, we evaluate their experience using two publicly available Yoruba TTS models, comparing the results with traditional Braille and human narration modes. The study provides both technical insights and human-centered evidence to inform the development of inclusive, linguistically aware educational technologies. Ultimately, our goal is to highlight whether current TTS technology is up to the task of delivering Yoruba exam content accessibly, and what improvements are needed to ensure equitable access to digital assessments for visually impaired learners.

1.1 Research Questions

This study carried out an experiment on the use of TTS systems to assist visually impaired students during computer-based tests (CBTs). The aim was to understand how effective current Yoruba TTS models are when used in real exam conditions. The following research questions guided the study:

- **Clarity of TTS Output:** Can existing Yoruba TTS systems read out standardized exam questions clearly and correctly for blind students (i.e., with sufficient intelligibility and accuracy)?
- **Comprehension Compared to Braille/Human Voice:** Do visually impaired students comprehend Yoruba TTS-delivered

questions as well as they do with Braille or human narration during test scenarios?

- **Student Experience and Preference:** How do students feel about using TTS in a CBT setting, particularly regarding comfort, focus, and sense of independence?

2 Background and Related Work

2.1 TTS Systems and Accessibility

TTS systems play a central role in accessibility, particularly for blind and visually impaired users. These systems are widely used in screen readers, digital assistants, and educational tools to provide real-time auditory access to written content (American Foundation for the Blind, 2025; Rella, 2023). In educational settings, TTS enables learners with visual impairments to independently engage with computer-based content, including assessments, without relying on human assistance (Dhaliwal and Sharma, 2024).

A typical TTS pipeline includes text normalization, linguistic analysis, acoustic modeling, and waveform generation (Rella, 2023). Advances in deep learning have led to the development of end-to-end neural models like Tacotron 2, FastSpeech, and VITS, which have demonstrated near-human speech quality in high-resource languages such as English and Mandarin (Shen et al., 2018; Ren et al., 2019; Kim et al., 2021). These models have significantly improved voice quality in screen readers and accessibility tools, providing smoother and more natural reading experiences.

However, deploying TTS in high-stakes scenarios like exams requires more than fluency. The system must produce highly intelligible, accurate, and well-pronounced speech, especially in tonal languages where pitch can change word meaning (Ogunremi et al., 2024). For languages like Yoruba, which require tone sensitivity and language-specific modeling, this remains a major challenge.

2.2 Challenges in TTS for Underrepresented Languages

Developing TTS systems for underrepresented languages such as Yoruba presents significant challenges due to limited linguistic resources. Yoruba is a tonal language where pitch accents (e.g., á, à, a) influence meaning. When TTS systems fail to model tone accurately, they risk generating unnatural or misleading speech (Ogunremi et al., 2024;

Tubosun, 2023). The lack of large high-quality Yoruba datasets, particularly diacritically marked text further complicates model training and tone learning.

While multilingual efforts like Meta’s Massively Multilingual Speech (MMS) project and Mozilla’s Common Voice have introduced Yoruba into their data pools (Pratap et al., 2023; Ardila et al., 2020), the models built on these datasets still struggle with tonal fluency and pronunciation. Early Yoruba TTS systems based on festival or rule-based engines achieved syntactic correctness but lacked expressive prosody and were rarely tested in applied contexts like exam delivery (Gutkin et al., 2020). Additionally, many existing corpora are domain-limited, such as religious readings, and do not reflect diverse use cases like standardized assessments.

As a result, current Yoruba TTS tools are often intelligible at the sentence level but unreliable in scenarios that demand tonal precision, such as multiple-choice exams where slight tonal differences can change answer meanings.

2.3 Speech vs. Braille in Educational Testing

Braille remains a trusted method for blind students during standardized testing. It offers silent, tactile interaction with exam content, supporting independence, concentration, and accuracy (Willings, 2017). However, Braille literacy rates remain low globally due to limited access to training and materials. In Nigeria and similar contexts, many visually impaired students lack the infrastructure or instruction required to become fluent Braille users (National Federation of the Blind, 2009).

Moreover, digital Braille solutions such as refreshable Braille displays are costly and scarce in low-resource environments (Perkins School for the Blind, 2024). Paper Braille, while helpful, requires advance preparation and lacks flexibility for real-time or adaptive CBT systems. This limits its practicality for digital-first education systems and national e-assessment platforms.

TTS systems offer an alternative that can scale across devices without special hardware. When properly designed, they can render questions in local languages on standard laptops or mobile phones. However, TTS for exams must be accurate, fast, and linguistically tuned. In tonal languages like Yoruba, incorrect pronunciation or flat prosody can compromise understanding and fairness. Despite TTS’s potential, few studies have directly compared its performance with Braille in timed, exam-

like scenarios, particularly in African languages. This study addresses that gap using real-world testing with Yoruba-speaking students.

3 Methodology

3.1 Study Design

This study used a three-phase approach to compare the effectiveness of human voice, Braille, and TTS in delivering standardized exam content to visually impaired students. The same set of WAEC Yoruba questions was presented to each participant using all three formats. The responses and experiences of the students across these three modes formed the basis of the findings.

3.2 Participants

All five students (three males and two females) who participated in this experiment were visually impaired, fluent in Yoruba language, and between the ages of 15 and 18. They were selected from an educational institution that supports students with special needs and each of them had prior experience using Braille but limited or no experience with TTS systems.

3.3 Question Selection

Ten multiple-choice questions were selected from the 2024 WAEC Yoruba exam paper. The questions were standardized and covered common themes such as comprehension, tone-sensitive vocabulary, and culturally rooted expressions. Each question had four answer options (A–D) and was stored in a structured CSV file for use during the TTS phase.

3.4 Phase One: Human Voice Delivery

In the first phase, each question was read aloud to the students by a fluent Yoruba speaker in a quiet room. The speaker maintained a consistent pace and tone to minimize variability while students listened carefully and selected their answers verbally. This phase served as a familiar benchmark, as human narration is often used in assisted testing environments.

3.5 Phase Two: Braille Delivery

The second phase involved presenting the same 10 questions to each student in Braille format. Each student was given enough time to read the questions and options independently and then respond verbally. This phase reflected current best practices in inclusive exam administration and provided a direct comparison point for evaluating TTS usability.

3.6 Phase Three: TTS Web Interface Delivery

The third phase involved presenting the same WAEC Yoruba questions to participants using a custom-built web interface. This approach tested whether modern Yoruba TTS models could serve as a viable delivery method for computer-based tests (CBTs) designed for visually impaired students.

3.6.1 Model Evaluation

Four Yoruba TTS models were evaluated to determine their suitability for delivering exam content to visually impaired learners. The models tested were: Facebook’s MMS-TTS-Yor, YarnGPT, Tacotron 2 (Google)(Shen et al., 2018), YorubaTTS (Túbsún and Olúòkun, 2017). Each model was assessed using *Mean Opinion Score (MOS)*, a standard subjective metric used in speech quality testing (ITU-T, 1996). Each model was rated by four native Yoruba speakers using a 5-point Likert scale, where 1 = Bad(very unnatural), 2= Poor (unnatural), 3=Fair (somewhat unnatural), 4=Good (mostly natural), 5=Excellent (very natural).

The MOS for each model was calculated using the formula:

$$MOS = \frac{1}{N} \sum_{i=1}^N r_i$$

where r_i represents the rating assigned by the i^{th} evaluator, and N is the total number of raters. This method, standardized by the International Telecommunication Union, ensures a consistent and interpretable quality metric across speech systems (ITU-T, 1996).

Models with a MOS score of 3.0 or higher were considered acceptable for inclusion in the live exam testing phase. The final selection was based on both quantitative MOS results and qualitative listener feedback on pronunciation, fluency, and tone handling.

3.6.2 Model Selection

Of the four TTS models evaluated, two were selected for further testing based on their intelligibility, tone accuracy, and integration ease for Yoruba exam content.

- **Facebook MMS-TTS-Yor** is a neural TTS model released by Meta AI as part of the *Massively Multilingual Speech* project. It was trained on public Yorùbá Bible recordings and produced moderately natural, intelligible

speech suitable for exam content (Pratap et al., 2023).

- **YarnGPT** is a lightweight, open-source Yorùbá TTS model hosted on Hugging Face (Azeez, 2025). Though less natural than MMS-TTS-Yor, it remained intelligible and was easy to integrate into the test interface.

3.6.3 System Architecture and Web Deployment

To make the models usable in a controlled testing environment, we built a web interface using Streamlit, a Python-based framework for building interactive data applications.

- **Frontend Interface:** Each question was presented one at a time along with audio playback controls. Yoruba ordinal numbering (e.g., Keta, Kewàà) was used to guide question progression.
- **TTS Processing:** For each question, the text was passed into the Facebook mms-tts-yor model via Hugging Face’s AutoTokenizer and AutoModelForTextToWaveform classes. The resulting waveform was saved using the soundfile library and played directly in the browser.

The system is deployed on Streamlit Community Cloud (<https://yoruba-cbt-tts.streamlit.app/>), and the source code is available at https://github.com/Moshood-Kausar/Yoruba_CBT_TTS.

4 Results and Findings

This section presents the feedback from five visually impaired students who participated in the experiment. Each student engaged with the same set of Yoruba WAEC questions through three delivery methods: human voice, Braille, and Text-to-Speech (TTS) via a web-based interface. After the sessions, they were asked to identify which method they preferred and why.

4.1 Delivery Method Preferences

- All five students chose Braille as their top preference. It gave them a sense of control, quietness, and independence during the test.
- Three students found the human voice method helpful and clear but also pointed out that depending on someone else was less ideal. The remaining two preferred working alone.

Table 1: Mean Opinion Scores (MOS) of four tested Yoruba TTS models.

Yoruba TTS Model	MOS (1–5)	Remarks
MMS-TTS-Yor	3.7	Best overall; moderately natural
YarnGPT	3.2	Moderate
YorubaTTS	2.5	Partially intelligible
Tacotron 2	2.1	Poor clarity and hard to follow

Table 2: Preferences expressed by five students after using all three delivery methods.

Delivery Method	Preferred (n=5)	Key Feedback
Human Voice	3	Easy to understand but made students feel dependent. Two preferred not to rely on others.
Braille	5	Most comfortable and familiar. Allowed for full independence, no distractions, and no background noise.
TTS System (Facebook mms-tts-yor and YarnGPT)	1	Liked the idea of using technology but found the speech robotic and harder to understand.

- Only one student liked the idea of using TTS, mainly for its modern approach, but raised concerns about poor voice quality, mispronunciation, and tonal inaccuracies.

5 Discussion

The current generation of Text-to-Speech (TTS) models shows promising capabilities in high-resource languages, but remains largely inadequate for underrepresented languages like Yoruba, especially in educational contexts such as Computer-Based Testing (CBT) for visually impaired learners. While the use of TTS offers a scalable and digital alternative to traditional methods, this study reveals critical shortcomings in its ability to support comprehension, independence, and user comfort in real exam settings.

5.1 Gaps in TTS for Accessibility

The results indicate that existing TTS systems, particularly mms-tts-yor and YarnGPT, still struggle with producing clear and natural speech. All five participants preferred Braille over TTS, citing better independence, fewer distractions, and stronger understanding. The main issues with TTS included flat, robotic delivery, mispronunciation, and inconsistent tone handling.

Yoruba’s tonal structure makes it more complex than many high-resource languages, a single word

can have different meanings based on tone alone. Unfortunately, the TTS models tested were not trained with sufficient tonal context or high-quality Yoruba data, which made them unreliable in a CBT scenario where clarity is critical. This gap mirrors broader concerns about the digital marginalization of African languages in AI systems. Despite the rise of multilingual models, training data imbalances remain a major barrier to inclusive performance.

5.2 Technology vs. User Experience

From a usability perspective, the TTS web interface was functional, accessible, and easy to navigate. However, good design alone did not make up for the weak audio quality. While one student appreciated the idea of using technology for independent testing, the others found the synthesized speech difficult to follow. These findings reinforce the idea that accessibility is not just about adding speech output, it must be accurate, culturally aware, and linguistically appropriate. Visually impaired learners need tools that help them feel in control and confident, especially in exam settings. At present, the TTS systems tested fall short of delivering that experience.

6 Conclusion

This study examined how effective current Text-to-Speech (TTS) models are for delivering computer-based exam content in an underrepresented language, Yoruba, for visually impaired students. Using real WAEC Yoruba questions and feedback from five blind students, we compared the experience of listening to TTS-generated speech with human narration and Braille.

Our findings show that:

- **Braille was the most preferred method**, offering clarity, ease of navigation, and a strong sense of independence throughout the test.
- Human voice was partially helpful, but introduced dependency.
- TTS models, though promising, were not yet usable for CBT due to poor tone handling and robotic delivery.

The performance limitations of even the best available models (like Facebook's 'mms-tts-yor') reveal a clear accessibility gap in speech technology for low-resource languages. While tools like Braille remain reliable, they are not scalable in digital-first environments, making TTS a critical area for future development.

6.1 Future Work

There is still a significant gap between what current TTS systems can do and what visually impaired students actually need, especially in languages that are underrepresented in AI development. Moving forward, one key step is the creation of high-quality, open-source Yoruba speech datasets that reflect the natural rhythm, tone, and variation of the language. This would make it possible to train models that not only produce intelligible speech but also capture the linguistic richness that matters in real exam settings.

It will also be important to design and evaluate these systems with direct input from users. Many TTS models today are built and tested using automated metrics, but this study shows that actual student experience tells a very different story. Working closely with educators, accessibility centers, and Yoruba speakers can help ensure that future tools are truly usable in classrooms and testing environments. As these systems improve, there will also be a need to think about how they fit into educational policies and standardized testing frameworks. With the right support, TTS technology

can become a reliable, inclusive tool not just for Yoruba, but for many languages that have been left behind in the digital space.

References

- American Foundation for the Blind. 2025. [Screen readers](#). Accessed: 2025-02-22.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. [Common Voice: A massively-multilingual speech corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Saheed Azeez. 2025. YarnGPT: Nigerian-accented English text-to-speech model.
- Chika Chukwuma-Nosike and Peace Chukwuma. 2023. Computer-based test (CBT) evaluation innovation: Prospects in curriculum implementation in Nigeria. *Journal of Agriculture and Food Sciences*, 21(2). Special Issue: Man, Environmental Safety and Sustainability the Role of Research, Chapter 24, pp. 331–343.
- Manpreet Kaur Dhaliwal and Rohini Sharma. 2024. [Improving accessibility and independence for blind/visually impaired persons based on speech synthesis technology](#). *International Journal of Computer Applications*, 186(28):12–20.
- David Eberhard, Gary Simons, and Chuck Fennig. 2019. *Ethnologue: Languages of the World*, 22nd edition. SIL International.
- Alexander Gutkin, Isin Demirsahin, Oddur Kjartansson, Clara E. Rivera, and Kólá Túboşún. 2020. [Developing an open-source corpus of Yoruba speech](#). In *Proc. of Interspeech 2020*, pages 404–408, Shanghai, China.
- ITU-T. 1996. Method for subjective determination of transmission quality. Technical Report Recommendation P.800, International Telecommunication Union - Telecommunication Standardization Sector. Retrieved from <https://www.itu.int/rec/T-REC-P.800-199608-I/en>.
- Jaehyeon Kim, Jungil Kong, and Juhee Son. 2021. [Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech](#). *Preprint*, arXiv:2106.06103.
- National Federation of the Blind. 2009. *The Braille Literacy Crisis in America: Facing the Truth, Reversing the Trend, Empowering the Blind*. NFB Jernigan Institute, Baltimore, MD.
- Tolulope Ogunremi, Kola Tubosun, Anuoluwapo Aremu, Iroro Orife, and David Ifeoluwa Adelani. 2024. [‘Ir’oy’inSpeech: A multi-purpose Yorubá speech corpus](#). *Preprint*, arXiv:2307.16071.

- Pawan Kumar Patel, Amey Karkare, and Gaurav Raheja. 2021. Inclusive accommodations for persons with visual impairments in computer-based tests. *Studies in Health Technology and Informatics*, 282:219–237.
- Perkins School for the Blind. 2024. An overview of Braille devices. <https://www.perkins.org/resource/overview-braille-devices/>.
- Vineel Pratap, Adithya Tjandra, Bowen Shi, Jing Huang, Qiantong Xu, Aravindh Krishnaswamy Babu, and Michael Auli. 2023. Scaling speech technology to 1,000+ languages. *Preprint*, arXiv:2305.13516.
- Sirisha Rella. 2023. Exploring unique applications of text-to-speech technology.
- Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2019. FastSpeech: Fast, robust and controllable text to speech. *Preprint*, arXiv:1905.09263.
- Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural TTS synthesis by conditioning WaveNet on MEL spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.
- Kola Tubosun. 2023. Making machines speak Yorùbá.
- Klá Túbsún and Adéday Olúòkun. 2017. Yorùbá text-to-speech system. <https://www.ttsyoruba.com>. Accessed: 2025-02-22.
- waec. The west african examinations council. Accessed: 2025-02-22.
- Carmen Willings. 2017. Educational assessments for students who are blind or visually impaired.

Power Asymmetries, Bias, and AI: A Reflection of Society on Low-Resourced Languages – African Languages as a Case Study

Simbiat Ajao

Masakhane

University of Lagos

simbiatajao18@gmail.com

Abstract

In recent times, artificial intelligence (AI) systems have become the primary intermediary to information access, services, and opportunities. Currently, there are growing concerns as to how existing social inequalities are reproduced and amplified through AI. This is significantly evident in language technologies, where a small number of dominant languages or what we'll refer to as big languages and cultural contexts shape the training, design, and evaluation of models. This paper examines the intersections of power asymmetries, linguistic bias, and cultural representation in AI, with a major focus on African languages and communities. We argue that current Natural Language Processing (NLP) systems reflect a high level of global imbalances in the availability of data, infrastructure, and decision making power, often marginalizing low-resourced languages and cultural peculiarities. It is important we know that how these data are structured is a great determinant in what their outcome will be. With reference to examples from speech recognition, machine translation, and large language models, we highlight the social and cultural consequences of linguistic exclusion, including reduced accessibility, misinterpretation, and digital invisibility. Finally, we identify and discuss pathways toward more equitable language technologies, emphasizing community-led data practices, interdisciplinary collaboration, and context-aware evaluation frameworks. By foregrounding language as both a technical and political concern, this work advocates for African-centered approaches to NLP that promote fairness, accountability, and linguistic justice in AI development (Bender et al., 2021b) (Masakhane Community, 2020–2025).

Keywords: Power asymmetries, Bias, AI, Low-resourced, African natural language processing.

1 Introduction

African Languages continually tops the charts as to the intensity of work being done in Artificial Intelligence (AI) to drive inclusivity, particularly with the various approaches for culturally embedded or nuanced data/models. Natural Language Processing (NLP) systems increasingly shape how we, as humans, communicate, access information, and participate in digital life. From the birth of voice assistants and machine translation to automated content moderation and large language models, language technologies are no longer secondary; they are infrastructural. However, despite their global deployment, these systems are not built on equal linguistic ground.

In the context of NLP, these inequalities are especially visible in the uneven representation of the world's languages. Although a small number of dominant languages, especially English, receive disproportionate attention in data collection, model training, and evaluation, the vast majority of the over 7,000 languages in the world remain underrepresented or completely excluded from modern language technologies (Joshi et al., 2020).

The majority of contemporary NLP models are trained on data drawn disproportionately from a small number of dominant languages, particularly English. As a result, African languages, despite representing immense linguistic diversity and millions of speakers, remain underrepresented, underperforming, or entirely absent in many AI systems. This imbalance is not merely technical; it reflects deeper asymmetries of power, knowledge production, and cultural authority.

Colonial language policies, limited research infrastructure, and the concentration of AI development in the Global North have all contributed to the marginalization of African languages in digital spaces (UNESCO, 2021)

This paper argues that language bias in AI is inseparable from global power structures. We position African NLP not only as a technical challenge of low-resource languages, but as a site where questions of equity, representation, and justice must be confronted. By examining how power operates through data, models, and evaluation practices, we aim to reframe linguistic inclusion as a core requirement of ethical AI.

2 Related Work

Research on bias and inequality in artificial intelligence has expanded significantly over the past decade, with scholars documenting how AI systems reproduce and amplify existing social hierarchies. Foundational work by (Noble, 2018) demonstrates how search and ranking systems encode racial and gendered bias, while (Benjamin, 2019) situates such systems within broader structures of power, arguing that technological harms are inseparable from social context. These perspectives provide an essential grounding for understanding bias in NLP beyond purely technical metrics.

Within the NLP community, several studies have examined linguistic bias and uneven language representation. (Joshi et al., 2020) provide a comprehensive analysis of linguistic diversity in NLP, showing how benchmark datasets and research attention overwhelmingly favor a small number of high-resource languages. Their work highlights how structural research practices, not linguistic properties, drive disparities in model performance across languages. Similarly, (Bender et al., 2021b) critiques the prevailing modeling assumptions in NLP, arguing that many architectures implicitly treat English as a linguistic norm, thus marginalizing typologically diverse languages. Research focusing specifically on African contexts further underscores the role of power asymmetries. (Birhane, 2020) frames the marginalization of African languages and communities as a form of algorithmic colonization, highlighting how data extraction and model deployment often occur without a meaningful local agency. This perspective aligns with broader critiques of data colonialism articulated by (Couldry and Mejias, 2020), who argue that data-driven systems replicate historical patterns of resource extraction and control. Empirical studies have also documented the real-world consequences of lin-

guistic bias. (Koenecke et al., 2020) show that commercial speech recognition systems perform significantly worse for non-dominant accents and dialects, illustrating how linguistic marginalization translates into measurable performance gaps. More recently, African-led NLP initiatives have begun to challenge dominant research paradigms. The Masakhane research community promotes collaborative and community-driven approaches to NLP for African languages, emphasizing capacity building and equitable knowledge production. Complementary efforts such as Lanfrica focus on surfacing and organizing resources for African languages, addressing visibility gaps in the research ecosystem. While existing work has made significant contributions to understanding bias, language diversity, and African NLP, much of the literature treats these issues in isolation. This paper builds on and connects these strands by explicitly foregrounding power as a unifying analytical lens. By situating linguistic bias within historical and institutional asymmetries, we extend prior work and argue for a more holistic, justice-oriented approach to NLP for African languages.

3 Power Asymmetries in the AI and NLP Ecosystem

Power asymmetry in AI refers to the unequal distribution of control, influence, and benefits in the development of language technologies (Benjamin, 2019); (Birhane, 2020). AI development is characterized by a pronounced concentration of resources, expertise, and decision-making authority.

English operates as the default language of AI research, reinforcing Anglophone norms in datasets, benchmarks, and evaluation practices. Such dynamics reflect broader patterns of unequal knowledge production, where communities most affected by AI systems often have limited influence over their design and deployment (Benjamin, 2019). In the context of African languages, this results in technologies that poorly reflect local linguistic practices, including dialectal variation, tone, and code-switching.

These asymmetries are not new but are rooted in longer histories of how African knowledge, languages, and perspectives have been systematically marginalized. As (Badawi, 2024) argues, dominant global narratives have often spoken *about* Africa rather than *from* Africa, privileging external interpretations over indigenous voices and

epistemologies. This historical pattern is reproduced in contemporary AI systems, where African languages and communicative practices are frequently underrepresented or framed as peripheral. Situating NLP within this broader historical context highlights that linguistic exclusion in AI is not merely a technical oversight, but part of a continuing struggle over whose knowledge is valued, preserved, and amplified in global technological systems.

3.1 Language as a Site of Power in AI Systems

Language is both a tool of communication and a repository of culture, worldview, and identity. In AI, it becomes a site where power is exercised. When AI systems privilege certain languages, they also privilege the worldviews embedded in those languages. In NLP, dominant modeling choices such as tokenization schemes, pretrained embeddings, and benchmark datasets are optimized primarily for Indo-European languages. African languages, many of which are tonal, agglutinative, or exhibit complex morphology, are poorly served by these assumptions (Bender et al., 2021b).

/ In the Yorùbá corpora sampled, this fragmentation typically results in 23% more tokens per sentence compared to English sentences expressing equivalent semantic content. Standard subword tokenizers, such as **BPE** and **WordPiece** disproportionately fragment Yorùbá text because they ignore morphemic structure and tone-bearing units, producing subword sequences that do not align with linguistically meaningful representations. This fragmentation collapses tone-dependent distinctions as represented in table 1, into overlapping token patterns, weakening lexical representations, and increasing semantic ambiguity in downstream tasks including machine translation and named entity recognition. At the same time, over-fragmentation inflates token counts for semantically equivalent content relative to English, leading to higher training and inference costs in transformer-based models due to the quadratic complexity of self-attention and more rapid exhaustion of fixed context windows. Consequently, token-level evaluation metrics and fixed-context benchmarks systematically penalize Yorùbá, making models appear less efficient or low-performing when the observed disparities primarily arise from tokenizer/language mismatch rather than inherent linguistic complexity.

As a result, African languages are often treated as deviations from a presumed linguistic norm rather than as central objects of study. This reinforces a hierarchy in which some languages are considered "standard" technologically, while others are framed as difficult, noisy, or peripheral.

4 Bias Across the NLP Pipeline

AI research and innovation are structurally biased toward actors with the greatest computational, financial, and institutional resources, leading to unequal influence over whose objectives are prioritized and who benefits the most from AI technologies (Ahmed and Wahed, 2020)

4.1 Data Bias

Training data for NLP systems are heavily skewed toward web-based sources dominated by English and other high-resource languages. Even when data on African languages exist, it is often fragmented, inconsistently annotated, or stripped of sociolinguistic context. These limitations directly affect model performance and robustness.

4.2 Model Bias

Models trained on skewed datasets inherit their biases. Previous work has shown that NLP systems exhibit reduced accuracy for non-dominant dialects and accents, leading to systematic performance gaps (Koenecke et al., 2020). For African users, this often results in unreliable speech recognition and culturally inappropriate text generation.

4.3 Evaluation Bias

Standard evaluation metrics prioritize surface-level accuracy while neglecting cultural appropriateness, pragmatic meaning, and code-switching features central to everyday language use across Africa. This misalignment reinforces the invisibility of African communicative practices in NLP benchmarks.

4.4 Algorithmic bias

Models amplify the patterns present in the data, often reinforcing stereotypes. Example: Speech recognition systems perform poorly for African-American Vernacular English or Nigerian English because the training data are skewed. AI bias is not accidental; it results from systemic imbalances in data collection, design, and validation.

Sentence	Tokenizer	Tokens Produced	Impact
I bought a black cap (English)	WordPiece	I / bought / a / black / cap	These are mainly whole words or sub-words with meaning
Mo ra fià dúdú kan	WordPiece	Mo / ra / fi / là / dú / dú / kan	Diacritics force arbitrary splits, breaking lexical units

Words	Gloss	BPE / WordPiece Output	Linguistic Implication
mlúàbí	person of good character	/ m / lú / à / bí	Morphemes fragmented; cultural concept split into ambiguated pieces
w	hand	/ w /	Loses distinction from owo (money) when diacritics are removed or mismatched

Table 1: Comparative Analysis: Subword Tokenization Effects in Yorùbá

5 Linguistic and Sociocultural Context of African Languages

Africa is the continent with the longest human history in the world (Badawi, 2024). In this book, Badawi went on to say that history is not only about the past, it also informs our present and shapes our future, which is why the drive for a digitally and infrastructurally inclusive AI system constantly remains valid. African languages represent one of the richest and most diverse linguistic ecologies worldwide, comprising more than 2,000 languages that span multiple language families, including Niger-Congo, Afroasiatic, Nilo-Saharan, and Khoisan (Guthrie, 1967); (Heine and Nurse, 2000). This diversity is characterized not only by the number of languages, but also by the extensive variation in morphology, syntax, phonology, and pragmatics. Almost, if not all African languages are tonal, morphologically rich, and rely heavily on contextual meaning, posing distinct challenges for standard NLP pipelines developed primarily for a different specific language class or family.

5.1 Sociocultural Embeddedness of Meaning

Meaning in African languages is often shaped by cultural norms, social roles, politeness strategies, and shared communal knowledge rather than by literal semantic content alone (Bamgbose, 1991). Pragmatic features such as honorifics, kinship terms, indirectness, and metaphor are central to interpretation, but are often flattened or lost in surface-level NLP representations. Literal translations may therefore misrepresent intent, especially

in culturally dense expressions such as idioms or proverbs (Bender and Koller, 2020).

5.2 Power, Representation, and Digital Marginalization

Language technologies developed without local participation risk misrepresenting speakers or excluding entire communities. Minority languages, rural dialects, and non-elite speakers are often absent from datasets, leading to AI systems that encode urban, male, and elite speech as normative. This exclusion has tangible consequences, particularly when AI systems are deployed in sensitive domains such as education, healthcare, or civic participation.

6 Empirical Case Studies and Illustrative Evidence from African NLP

6.1 Community-Driven Machine Translation: The Masakhane Paradigm

Languages and Tasks: Yorùbá-English, Hausa-English, Kiswahili-English machine translation (MT) Recent work within the Masakhane community provides a concrete illustration of how community-centered research practices improve NLP outcomes for African languages. Rather than relying exclusively on web-scraped or legacy parallel corpora often dominated by religious or colonial texts, Masakhane emphasizes participatory data creation involving native speakers, linguists, and translators. Empirically, MT systems trained on community-curated datasets demonstrate measurable improvements over baselines trained on

noisier corpora. For example, Yorùbá-English systems that preserve diacritics and enforce orthographic consistency outperform non-diacritized baselines, as tone-sensitive distinctions reduce semantic ambiguity. Similarly, Hausa MT models that explicitly account for Boko and Ajami orthographies generalize better across domains than models trained on homogenized text.

This case study empirically supports the argument that data quality, linguistic fidelity, and community participation are critical determinants of model performance in low-resource settings ((Orife et al., 2020; Adelani and Alabi, 2021)

6.2 Orthography and Diacritics in Yorùbá NLP

Language and Tasks: Yorùbá - POS tagging, named entity recognition (NER), machine translation, and ASR preprocessing

Yorùbá is a tonal language in which diacritics encode essential lexical and grammatical distinctions. However, many publicly available datasets omit tone marks and underdots, collapsing distinct lexical elements into identical surface forms. This practice introduces systematic ambiguity into downstream tasks. Empirical experiments across multiple Yorùbá NLP tasks show that models trained on fully diacritized text consistently outperform those trained on stripped text. In POS tagging and NER, diacritics-aware models achieve higher accuracy and more interpretable error patterns. In MT, diacritics preservation reduces mistranslations arising from homographs that differ only in tone or vowel quality. This case highlights how seemingly minor preprocessing decisions can materially affect model performance and reinforces the need for language-specific data handling strategies rather than language-agnostic pipelines ((Alabi et al., 2020; Adelani and Alabi, 2021).

6.3 Named Entity Recognition in Morphologically Rich African Languages

Amharic (Morphological Complexity) In Amharic NER, person and location names frequently appear with attached prefixes, suffixes, and case markers. For example, the surface form (*b-Addis Ababa*, in Addis Ababa) combines a preposition (, in) with the name of the location . Standard whitespace tokenization treats this as a single token, causing baseline NER systems to miss or mislabel the entity boundary. (Gashaw

et al., 2020) show that segmenting functional morphemes from named entities either through rule-based morphological preprocessing or morphology-aware tokenization leads to substantial gains in NER performance. When combined with gazetteers of Ethiopian place and personal names, F1 scores improve markedly compared to models trained on raw text, demonstrating that the error source lies in tokenization and annotation mismatch rather than model capacity.

Hausa (Honorifics and Titles) In Hausa, personal names are commonly preceded by titles and honorifics such as *Alhaji*, *Mallam*, *Dr.*, or *Sarki*. In sentences like *Alhaji Musa ya isa Abuja* (Alhaji Musa arrived in Abuja), baseline NER models often label *Alhaji Musa* as a single entity or misclassify *Alhaji* as part of the name. Empirical annotation studies reported by (Adelani and Alabi, 2021) demonstrate that explicitly separating honorifics from name spans in annotation guidelines reduces entity boundary errors and improves consistency across annotators. Models trained on such linguistically informed annotations achieve higher precision in PERSON entities, particularly in news and administrative text.

Igbo (Compounding, Prefixation, and Semantic Heads) Igbo exhibits productive nominal compounding and derivational morphology, which frequently affects named entities. Institutional and location names often include common nouns such as *l* (house), *obodo* (town/city) or *mahadum* (university) followed by a proper name, e.g., *l Akwkw Nnamdi Azikiwe* (Nnamdi Azikiwe Library) or *Obodo Owerri* (Owerri city). In standard NER pipelines, these constructions pose two challenges. First, whitespace-based tokenization fragments semantically unified entities, causing systems to label only the proper-name head (*Nnamdi Azikiwe*, *Owerri*) while excluding the institutional or locative marker. Second, English-centric annotation schemes often misclassify the common-noun component (*l*, *obodo*) as non-entity context rather than part of the named entity span. Empirical annotation studies on Igbo NER reported by (Adelani and Alabi, 2021) show that the consistency of the entity span improves when annotation guidelines explicitly treat such nounproper-name compounds as single named entities. Models trained on these revised annotations achieve higher recall for ORGANIZATION and LOCATION entities, particularly in news and educational-domain corpora, where such constructions are frequent.

Across Amharic, Hausa, and Igbo, these examples demonstrate that NER errors are not primarily due to low data volume but to misaligned linguistic assumptions embedded in standard NER pipelines. Incorporating morphology-aware tokenization, culturally grounded annotation schemes, and locally relevant gazetteers yields consistent performance gains, validating the need for language-specific NER design in African NLP.

7 Implications for African NLP Research

Understanding African languages requires moving beyond a purely technical framing of low-resource status toward a sociotechnical perspective that recognizes historical marginalization, data extraction, and uneven power relations. Ethical and effective African NLP should therefore integrate linguistic expertise, community inclusivity, and culturally grounded evaluation frameworks, ensuring that AI systems support linguistic diversity rather than erode it.

8 Re-framing African Languages Beyond the Low-Resource

Describing African languages as low-resource risks naturalizing their marginalization. These languages are not inherently under-resourced; rather, they have been systematically under-supported due to historical and political factors. Re-framing the problem shifts attention from perceived linguistic deficiency to structural inequality.

Crucially, African languages are not inherently under-resourced. Many are spoken by millions of people and have rich oral traditions, established writing systems, and long-standing histories of literacy and scholarship. Their exclusion from mainstream NLP pipelines is better understood as the result of structural neglect rather than linguistic deficiency. Colonial language policies privileged European languages in education, governance, and publishing, shaping which languages were standardized, archived, and later digitized ((UNESCO, 2021)). These historical choices continue to influence the availability and research priorities of contemporary data. The low-resource label also narrows the scope of technical innovation by framing African languages primarily as problems to be solved through data augmentation or transfer learning. While such techniques are valuable, they risk reinforcing a deficit-

oriented narrative in which African languages are treated as beneficiaries of models trained on dominant languages, rather than as central objects of inquiry in their own right. As (Bender et al., 2021b) argues, modeling choices in NLP often embed assumptions about linguistic structure that align poorly with typologically diverse languages, resulting in systems that perform inadequately despite technical sophistication. An alternative framing emphasizes *resource redistribution* rather than resource scarcity. From this perspective, the central challenge is not the absence of linguistic data, but the lack of sustained investment in community-driven data creation, annotation, and governance. African-led initiatives such as the Masakhane demonstrate how collaborative research models can shift this balance by centering local expertise, shared ownership, and contextual knowledge. These efforts challenge extractive research practices and highlight the importance of building long-term capacity alongside technical outputs. Re-framing African languages beyond the low-resource paradigm also has implications for evaluation. Standard benchmarks often fail to capture key features of African language use, including code-switching, dialectal variation, and culturally grounded meanings. Treating these features as noise rather than signal further marginalizes African communicative practices. More inclusive evaluation frameworks, informed by sociolinguistics and community input, are necessary to ensure that NLP systems meaningfully serve their intended users. Ultimately, moving beyond the low-resource label requires a shift in both language and practice. It calls for recognizing African languages as sites of knowledge, identity, and agency, and for designing NLP systems that reflect this reality. Such a shift aligns with broader calls for linguistic justice in AI, where fairness is understood not only as parity in performance metrics, but as equitable participation in the creation and governance of language technologies (Benjamin, 2019); (Birhane, 2020).

9 Roles of Key Stakeholders in Advancing African NLP

The development of NLP systems for African languages is shaped by the interaction of multiple stakeholders, each contributing distinct forms of expertise, resources, and constraints. Prior work suggests that progress in African NLP is most con-

sistent when responsibilities are distributed across policy institutions, research communities, industry actors, and international funders, rather than concentrated within a single sector (Joshi et al., 2020)(Bird, 2020).

9.1 African Governments: Policy Frameworks and Public Resources

African governments influence NLP research primarily through language policy, public infrastructure, and access to state-held textual resources. Government recognition of indigenous languages in education, administration, and media contributes to the availability and standardization of written materials, which in turn affects data availability for NLP (UNESCO, 2021).

9.2 Local Research Communities: Linguistic Expertise and Task Design

Universities, research institutes, and community-led initiatives contribute linguistic knowledge, annotation expertise, and contextual understanding that are difficult to obtain through large scale automated approaches alone. Empirical evidence from community-driven projects such as Masakhane shows that datasets developed with direct involvement from native speakers and local researchers exhibit higher linguistic fidelity and clearer task definitions (Orife et al., 2020)(Adelani and Alabi, 2021). For languages such as Yorùbá and Hausa, local researchers are particularly well positioned to identify issues related to tone marking, honorific usage, orthographic variation, and code-switching phenomena that directly affect modeling and evaluation. Their role is therefore central to ensuring that research assumptions align with actual language use, rather than inferred abstractions.

9.3 Industry: Engineering Capacity and Deployment Experience

Industry actors contribute engineering expertise, computational infrastructure, and experience with large-scale deployment. These capacities are especially relevant for training and serving models that operate under real-world latency, memory, and cost constraints. In African language contexts, industry-supported systems such as speech recognition or conversational agents for Kiswahili or Hausa have demonstrated the feasibility of deploying NLP beyond research settings.

At the same time, prior studies note that industry-driven development tends to prioritize scalability and reuse, which can benefit from collaboration with linguistically informed research teams to ensure that language-specific characteristics are adequately represented (Bender et al., 2021a). Viewed in this way, industry participation complements academic and community research by translating prototypes into usable systems while benefiting from external linguistic validation.

9.4 International Funding: Enabling Long-Term Research Capacity

International funding organizations play an enabling role by supporting training, access to infrastructure, and sustained research programs, particularly in contexts where local funding for computational research is limited. Beyond individual projects, funding mechanisms influence research practices through requirements related to openness, collaboration, and capacity development (Birhane, 2023).

Evidence from African NLP initiatives suggests that funding models emphasizing local leadership, open resources, and long-term partnerships are associated with broader participation and more reusable research outputs (Adelani and Alabi, 2021). In this sense, international funders act less as directors of research agendas and more as facilitators of stable research ecosystems.

Taken together, the literature indicates that African NLP research benefits from complementary stakeholder roles: governments provide policy and public resources, local research communities contribute linguistic grounding, industry supports scaling and deployment, and international funding enable continuity and capacity building. Rather than attributing challenges to any single group, existing evidence points to coordination and alignment among these actors as a key factor in reducing structural bottlenecks related to data availability, modeling assumptions, and evaluation relevance (Joshi et al., 2020)(UNESCO, 2021).

10 Conclusion

This paper has argued that linguistic bias in AI is fundamentally shaped by power asymmetries in global AI development. African languages are marginalized not because of inherent technical limitations, but because of historical, institu-

tional, and political inequalities embedded in the NLP ecosystem. Building fair and inclusive language technologies requires redistributing power over data, models, and evaluation to include the communities whose languages are being modeled. An ethical future for AI depends on recognizing language as a site of justice, identity, and agency. A truly intelligent AI must be multilingual, culturally grounded, and accountable to all its users.

References

- David Ifeoluwa Adelani and Jesujoba Oluwadara Alabi. 2021. Deep learning approaches for low-resource african languages. In *Proceedings of the 3rd Workshop on African Natural Language Processing (AfricaNLP)*. Association for Computational Linguistics.
- Nur Ahmed and Muntasir Wahed. 2020. [The democratization of ai: Deep learning and the compute divide in artificial intelligence research](#). Preprint, arXiv:2010.15581.
- Jesujoba Oluwadara Alabi, David Ifeoluwa Adelani, and 1 others. 2020. Massively multilingual neural machine translation in low-resource settings. In *Proceedings of the 2nd Workshop on African Natural Language Processing (AfricaNLP)*. Association for Computational Linguistics.
- Zeinab Badawi. 2024. *An African History of Africa: From the Dawn of Humanity to Independence*. Penguin Books, London.
- Ayo Bamgbose. 1991. [Language and the nation : the language question in sub-saharan africa](#).
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021a. [On the dangers of stochastic parrots: Can language models be too big?](#) In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, pages 610–623, New York, NY, USA. Association for Computing Machinery.
- Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021b. [On the dangers of stochastic parrots: Can language models be too big?](#) . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, page 610623, New York, NY, USA. Association for Computing Machinery.
- Emily M. Bender and Alexander Koller. 2020. [Climbing towards NLU: On meaning, form, and understanding in the age of data](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.
- Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity.
- Steven Bird. 2020. [Decolonising speech and language technology](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Abeba Birhane. 2020. [Algorithmic colonization of africa](#). *SCRIPTed*, 17(2):389–409.
- Abeba Birhane. 2023. [Algorithmic colonization of africa](#). In *Imagining AI: How the World Sees Intelligent Machines*. Oxford University Press.
- Nick Couldry and Ulises A Mejias. 2020. [The costs of connection: How data are colonizing human life and appropriating it for capitalism](#). *Social Forces*, 99(1):e6–e6.
- Ibrahim Gashaw, Solomon Teferra Abate, and Mengistu Tachbelie. 2020. Named entity recognition for amharic using deep learning. In *Proceedings of the 2nd Workshop on African Natural Language Processing (AfricaNLP)*, Seattle, USA. Association for Computational Linguistics.
- Malcolm Guthrie. 1967. [Comparative bantu: An introduction to the comparative linguistics and prehistory of the bantu languages](#).
- B. Heine and D. Nurse. 2000. *African Languages: An Introduction*. African Languages: An Introduction. Cambridge University Press.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. [Racial disparities in automated speech recognition](#). *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.
- Masakhane Community. 2020–2025. [Masakhane: Machine translation for africa](#). Online; accessed 2025-12-22.
- S.U. Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.
- Iroro Orife, David Ifeoluwa Adelani, Jade Abbott, and 1 others. 2020. [Masakhane: Machine translation for africa](#). In *Proceedings of the 2nd Workshop on African Natural Language Processing (AfricaNLP)*. Association for Computational Linguistics.
- UNESCO. 2021. *Recommendation on the Ethics of Artificial Intelligence*. UNESCO.

Sudanese-Flores: Extending Flores+ to Sudanese Arabic Dialect

Hadia Mohmmedosman Ahmed Samil

Mila - Quebec AI Institute, Montreal, Canada

hadia.samil2@mila.quebec

David Ifeoluwa Adelani

Mila - Quebec AI Institute, McGill University & Canada CIFAR AI Chair

david.adelani@mila.quebec

Abstract

In this work, we introduce Sudanese-Flores, an extension of the popular Flores+ machine translation (MT) benchmark to the Sudanese Arabic dialect. We translate both the DEV and DEVTEST splits of the Modern Standard Arabic dataset into the corresponding Sudanese dialect, resulting in a total of 2,009 sentences. While the dialect was recently introduced in Google Translate, there are no available benchmark in this dialect despite it being spoken by over 40 million people. Our evaluation on GPT-4.1, Gemini 2.5 Flash, and open-weight models (LLaMA-3.3 70B and Qwen-3 80B) showed that while the performance from English to Arabic is impressive (more than 23 BLEU), they struggle on Sudanese dialect (less than 11 BLEU) in zero-shot settings. In the few-shot scenario, we achieved only a slight improvement in performance.

1 Introduction

Sudanese Arabic is a widely spoken variety in Sudan, yet it remains underrepresented in natural language processing (NLP) resources, limiting research on machine translation and multilingual modeling for this language. Despite the growing interest in low resource languages and Arabic dialects, most NLP benchmarks focus on Modern Standard Arabic (MSA) or other widely spoken dialects, leaving Sudanese Arabic largely unexplored. This gap motivates the need for parallel datasets that enable systematic evaluation of machine translation models for Sudanese Arabic.

In this work, we introduce Sudanese-Flores, an extension of the Flores+ evaluation benchmark that builds on Flores-200 (NLLB-Team et al., 2022) with Sudanese Arabic translations for all sentences in the original DEV & DEVTEST dataset. Our new data is primarily translated from Modern Standard Arabic (MSA), with additional verification from the original English split. The Sudanese-Flores

enables translation between Sudanese Arabic and all other >200 languages in FLORES including the six directions: MSA \leftrightarrow Sudanese Arabic, Sudanese Arabic \leftrightarrow English and MSA \leftrightarrow English. Our new dataset preserves the diversity of domains and sentence types present in FLORES+, ensuring a broad linguistic and cultural representation.

We perform zero-shot MT experiments using proprietary models (GPT-4.1, Gemini 2.5 Flash) as well as open-weight models (LLaMA-3.3, Qwen-3) to balance performance evaluation and reproducibility. Translation quality is evaluated using BLEU and ChrF metrics, highlighting both strengths and weaknesses across different directions. Our experiments show that LLMs struggle to translate from English to Sudanese Arabic compared to MSA, with over -13.0 BLEU drop in performance. Our results highlight the need to expand machine translation capabilities to more low-resource languages and dialects. By releasing Sudanese-Flores, we provide a benchmark that enables reproducible research, supports the development of inclusive multilingual NLP models, and contributes to expanding NLP coverage to underrepresented languages and dialects.

2 Related Work

Research on Arabic NLP has traditionally focused on MSA, often overlooking dialectal varieties. While more African Arabic dialects are represented in large-scale datasets such as Flores-200 (NLLB-Team et al., 2022) and SIB-200 (Adelani et al., 2024) (Tunisian Arabic, Moroccan Arabic, and Egyptian Arabic), Sudanese Arabic is not covered. In general, Sudanese Arabic is particularly underrepresented in NLP research such as MT, sentiment analysis & speech recognition. Existing dialectal corpora that include Sudanese Arabic (Jarrar et al., 2022) are relatively small and primarily designed for morphological

English	MSA	Sudanese Arabic
Such children are called “feral” or wild. Some feral children have been confined by people (usually their own parents); in some cases this child abandonment was due to the parents’ rejection of a child’s severe intellectual or physical impairment.	يَمَّ حبس بعض الأطفال الصّالين من قبل أشخاص (عادةً والديهم)؛ في بعض الحالات، كان هذا التخلي عن الطفل بسبب رفض الوالدين لإعاقته العقلية أو الجسدية الشديدة، ويطلق على هؤلاء الأطفال اسم «الوحشي» أو البري.	حبسوا شوية شفيع ضالين من ناس (عادةً والديهم)؛ في بعض الحالات، كان سبب التخلي هو رفض الوالدين لأعاقه الشافع العقلية أو الجسدية الشديدة، وبسموا الشفع ديل «الوحشي» أو البري
Lion prides act much like packs of wolves or dogs, animals surprisingly similar to lions (but not other big cats) in behavior, and also very deadly to their prey.	قطيع الأسود يعمل مثل قطيع الذئاب أو الكلاب، والمفاجئ أن الحيوانات مشابهة للأسد (لكن ليس القطط الكبيرة الأخرى) في السلوك ومميتة جداً لفراسمها.	قطيع الاسود يشتغل زي قطيع الذئاب او الكلاب، و المفاجئ انو الحيوانات بتشبه الاسد (لكن ما الكدائيس الكبيرة الثانية) في السلوك و مميتة خالص لفراسمها.
It warns that no one can guarantee that any course of action in Iraq at this point will stop sectarian warfare, growing violence, or a slide toward chaos.	يحذر من أنه لا أحد يمكنه ضمان أن أي مسار عمل في هذه اللحظة في العراق سيؤدي إلى وقف الحرب الطائفية أو العنف المتزايد أو الانحدار نحو الفوضى.	يحذر من انو مافي زول بقدر يضمن انو أي مسار عمل في اللحظة دي في العراق حيايدي لي وقف الحرب الطائفية أو العنف المتزايد او الانحدار ناحية الفوضى.

Table 1: **Example sentences from Sudanese-FLORES.** In the Sudanese Arabic column, **dialect-specific words are highlighted in bold** to show the linguistic differences from MSA. These words reflect local lexical choices (e.g., **الشفيع** ، **الكدائيس** ، **زول**) and demonstrate authentic Sudanese Arabic expressions captured in the dataset. To the best of our knowledge, these words are used only in Sudan.

analysis rather than MT, highlighting the need for larger, translation-focused datasets. Similarly, other large-scale pre-training data covering over 400 languages, such as FineWeb2 (Penedo et al., 2025) and MADLAD-400 (Kudugunta et al., 2023), do not include the Sudanese dialect.

3 Dataset and Experimental Setup

3.1 Dataset creation

Sudanese-Flores was created by a single native Sudanese speaker, who is trilingual (speaks English, Arabic and Sudanese dialect). The translations represent the Khartoum/Central Sudanese dialect variety, which is widely understood across Sudan and neighboring regions. The data set contains 1,012 sentences in the DEV set and 997 sentences in the DEVTEST set. Although she is fluent in the three languages, for ease of translation, she translated Arabic into Sudanese since they are closely related. In cases where the MSA text was ambiguous, the English source was consulted to ensure the meaning was preserved. This translation took between three to four weeks, with feedback from other native speakers during the process. [Table 1](#) provides examples that highlight some differences between the MSA and Sudanese Arabic dialect.

Additionally, 100 sentences were randomly selected for verification by a second native speaker, who confirmed that each sentence preserved the

original meaning and used natural Sudanese Arabic. We note that there is no standardized way to write Sudanese Arabic and some letters have the same pronunciation. These differences in writing do not cause any issues in understanding, and readers can clearly interpret the meaning of all sentences.

3.2 Experiments Setup

We prompted two leading LLMs in both zero-shot and few-shot (5-shots): GPT-4.1 and Gemini 2.5 Flash. In addition, we also evaluate open-weight large language models, including Llama-3.3 and Qwen-3, to improve reproducibility and accessibility. We make use of a single prompt. The same prompt template is used for all models and all translation directions. The prompt we used is very simple, obtained from AfroBench paper (Ojo et al., 2025):

```
You are a translation expert. Translate
the following {{source_lang}}
sentences to {{target_lang}}
```

```
{{source_lang}} sentence: {{source_text
}}
{{target_lang}} sentence:
```

For few shots, we added five examples in the DEV set into the prompt. We note that while few-shot prompting can improve performance in some cases, it can also negatively affect results, particularly for dialectal translation, suggesting that few-shot prompting does not consistently generalize across

Models	Setting	Arabic - English		English - Arabic		MSA - Sudanese	
		arb-eng	apd-eng	eng-arb	eng-apd	apd-arb	arb-apd
GPT-4.1	0-shot	39.9 / <u>68.1</u>	32.3 / 63.0	27.9 / 59.7	10.4 / 46.6	54.4 / 78.7	24.2 / <u>64.1</u>
	5-shots	<u>41.7 / 69.0</u>	36.5 / 65.8	<u>24.8 / 58.6</u>	12.0 / 47.9	46.9 / 75.2	24.1 / 63.1
Gemini 2.5 Flash	0-shot	24.4 / 60.1	21.5 / 57.2	23.5 / 56.4	9.7 / 46.2	47.0 / 74.3	<u>27.0 / 66.4</u>
	5-shots	40.3 / 67.9	<u>35.1 / 64.3</u>	24.7 / 58.0	<u>11.9 / 47.3</u>	<u>49.2 / 76.2</u>	25.2 / 63.2
Llama-3.3 70B	0-shot	42.6 / 66.5	32.9 / 59.9	23.5 / 51.7	8.1 / 37.9	44.1 / 69.0	17.7 / 48.8
	5-shots	43.2 / 66.2	36.0 / 61.2	23.6 / 51.6	10.7 / 40.3	48.1 / 71.6	29.3 / 61.2
Qwen-3-Next-80B-A3B	0-shot	40.7 / 64.5	33.5 / 59.1	24.1 / 52.4	8.7 / 39.2	41.2 / 66.8	16.9 / 50.1
	5-shots	40.6 / 64.3	34.7 / 59.6	24.1 / 52.4	8.2 / 38.9	45.2 / 69.4	18.8 / 53.0

Table 2: **MT translation BLEU / ChrF++ performance on Sudanese-Flores** across six translation directions: Arabic dialects \rightarrow English (arb-eng & apd-eng), English \rightarrow Arabic dialects (eng-arb & eng-apd), and MSA \leftrightarrow Sudanese Arabic (apd-arb & arb-apd). The results reported on DEVTEST set. The best result are in **Bold** and the second best are underlined.

translation directions for Sudanese Arabic.

We evaluate six translation directions to comprehensively assess model performance: MSA \leftrightarrow English, Sudanese Arabic \leftrightarrow English, and MSA \leftrightarrow Sudanese Arabic. The language directions reported in Table 1 are ordered by source–target pairs and include all combinations of arb, apd, and eng.

3.3 Evaluation Metrics

We measure the translation quality using BLEU (SacreBLEU) (Post, 2018) and ChrF++ (Popović, 2017).

4 Results and Discussion

Table 2 shows the result of GPT-4.1 and Gemini 2.5 Flash as well as open-weight models (Llama-3.3 and Qwen-3) on Sudanese-Flores in a zero-shot and few-shot settings across the six translation directions. Results are reported on the DEVTEST split using BLEU and ChrF metrics.

GPT-4.1 generally outperforms Gemini 2.5 Flash In zero-shot settings, we find GPT-4.1 is able to translate better than Gemini-2.5 Flash in almost all directions except arb–apd. In the apd–eng direction, the performance gap in BLEU points is more than +10, however, in eng–apd, it is lower. This finding is similar to the observation of AfroBench-Lite (Ojo et al., 2025)

Proprietary LLMs are consistently better than Open LLMs In almost all settings, we find the proprietary LLMs such as GPT-4.1 and Gemini 2.5 Flash having better performance than the open LLMs such as Llama 3.3 70B and Qwen 3 80B.

However, with few shots, there is significant boost in performance for Llama 3.3 70B with at least +2.0 increase for direction into Sudanese dialect, especially for the arb–apd direction. On the other hand, Qwen 3 80B gains after few shots are not consistent and often smaller.

Translation from English into Sudanese dialect is challenging We find a big gap in performance between translating from eng–arb and eng–apd with almost -17 BLEU point drop with GPT-4.1 and -13 points with Gemini-2.5 Flash in zero-shot settings. A similar degradation is observed for open-weight models, indicating that translation into Sudanese Arabic remains challenging across different model families. Surprisingly, few-shot did not seem to help much although there are some slight improvements.

Translation from Sudanese dialect to English seems easier Given the similarity of MSA and Sudanese dialect, it is unsurprising that LLMs are able to easily translate into English. Although, the performance is lower than from MSA, it is noticeably high (i.e. over 32 BLEU with GPT-4.1 and over 21 BLEU with Gemini). This trend is consistent across models, though performance remains lower than translation from MSA.

Few-shot is not so helpful, and may even hurt performance The easiest direction we find is from MSA to Sudanese, and vice versa. However, with few-shots, sometimes it helps in the direction of apd–arb while in some other cases it severely hurt the performance when generating Sudanese. This further suggests that few-shot prompting does

Source → Target	Example Source	Model Output	Reference	Type of Error	Observations
Sudanese → Arabic	المصطلح دا مشتق من المعرفة القديمة ببق الفراش	المصطلح مشتق من المعرفة القديمة ببق الفراش	هذا المصطلح مشتق من المعرفة القديمة ببق الفراش	MSA hallucination	Insertion of standard determiner; meaning preserved
Sudanese → English	كيرتس كوبر قام باكتشاف أكبر عدد أولي معروف	Curtis Cooper... discovered the largest known prime number to date...	Curtis Cooper... has discovered the largest known prime number to date...	Tense mismatch	Present perfect reduced to past tense
Arabic → English	قام كيرتس كوبر باكتشاف أكبر عدد أولي معروف	Curtis Cooper... discovered the largest known prime number to date...	Curtis Cooper... discovered the largest known prime number to date...	Minimal	Stylistic variation only
English → Arabic	There, they found the body of Saroja Balasubramanian...	وهناك وجدوا جثة ساروجا بالاسوبرامانيان	وهناك وجدوا جثة ساروجا بالاسوبرامانيان	Punctuation / style	Comma insertion; meaning preserved
English → Sudanese	The find also grants insight into the evolution of feathers in birds.	ال اكتشاف ده كمان بيعطي نظرة على تطور الريش في الطيور	ال اكتشاف بدى برضو نظرة على تطور الريش في الطيور	Dialectal variation	Alternative Sudanese lexical realization
Arabic → Sudanese	جميع سكان مدينة الفاتيك من الروم الكاثوليك	كل سكان مدينة الفاتيك من الروم الكاثوليك	جميع سكان مدينة الفاتيك من الروم الكاثوليك	Minor	Dropped period; meaning preserved

Table 3: **Qualitative error analysis across translation directions.** Each entry shows the source sentence, model output, reference, type of error, and observations describing the deviation. Dialect-specific or stylistic deviations are highlighted in the observations column; Sudanese-specific words are bolded in the model output.

not consistently generalize across translation directions for Sudanese Arabic. We leave further investigation of prompting strategies for future work.

5 Qualitative Error Analysis

To better understand model behavior on dialectal data, we perform a qualitative error analysis with representative examples shown in Table 3. The model exhibits occasional **MSA hallucination** when translating from Sudanese Arabic, inserting standard forms not present in the source. For Sudanese outputs, errors mainly involve **dialectal realization**, including inconsistent lexical choices and weakening of Sudanese-specific idioms or morphological markers, while core meaning is preserved. Overall, errors are largely stylistic rather than semantic and are most pronounced in Sudanese-related directions due to dialectal variability and limited training data.

6 Conclusion

In this paper, we introduce, Sudanese-Flores, a new benchmark that extends Flores-200 to Sudanese Arabic dialect. Our evaluations clearly shows the need for such a benchmark, since current LLMs struggle to generate Sudanese dialect, especially when translating from English which is also an official language in Sudan.

7 Acknowledgement

This research was supported by Mila-Quebec AI Institute P2v5 grant. We are grateful for the support of Siva Reddy throughout the project.

Limitations

While Sudanese-Flores offers a valuable benchmark for Sudanese Arabic MT, it has some inherent constraints. The dataset was primarily created by a single native speaker, which may limit dialectal variation, though we included verification from a second native speaker to ensure naturalness. The dataset size (2,009 sentences) is modest, reflecting the resource-scarce nature of this dialect, but it is sufficient for benchmarking and evaluation. Our focus on the Khartoum/Central Sudanese dialect ensures broad intelligibility across Sudan, though regional variations may exist. Finally, the orthography of Sudanese Arabic is not fully standardized, which may affect surface-level evaluation but does not hinder semantic understanding.

Ethical Considerations

We prioritize the linguistic and cultural integrity of Sudanese Arabic. Translations were performed by a native speaker and verified by a second native speaker to ensure accuracy and naturalness. The dataset contains no personally identifiable information or sensitive content. Users should be aware that MT models may reproduce biases present in

their training data; careful evaluation is advised when applying models in real-world contexts involving this dialect.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191. Association for Computational Linguistics.

References

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

Mustafa Jarrar, Fadi A. Zaraket, Tymaa Hammouda, Daanish Masood Alavi, and Martin Wählisch. 2022. [Lîsan: Yemeni, iraqi, libyan, and sudanese arabic dialect corpora with morphological annotations](#). *Preprint*, arXiv:2212.06468. Accessed: 2025-12-20.

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. [MADLAD-400: A multilingual and document-level large audited dataset](#). In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

NLLB-Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. [AfroBench: How good are large language models on African languages?](#) In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

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](#). In *Second Conference on Language Modeling*.

Maja Popović. 2017. chrF++: Towards character and word n-gram F-score for automatic mt evaluation. In *Proceedings of the Second Conference on Machine Translation*, pages 568–575. Association for Computational Linguistics.

Where Are We at with Automatic Speech Recognition for the Bambara Language?

Seydou Diallo^{1,4,5}, Yacouba Diarra^{1,2}, Mamadou K. KEITA^{1,3},
Panga Azazia Kamaté^{1,2}, Adam Bouno Kampo¹, Aboubacar Ouattara⁴

¹MALIBA-AI ²RobotsMali AI4D Lab ³Rochester Institute of Technology ⁴DJELIA

⁵Dakar American University of Science and Technology

Abstract

This paper introduces the first standardized benchmark for evaluating Automatic Speech Recognition (ASR) in the Bambara language, utilizing one hour of professionally recorded Malian constitutional text. Designed as a controlled reference set under near-optimal acoustic and linguistic conditions, the benchmark was used to evaluate 37 models, ranging from Bambara-trained systems to large-scale commercial models. Our findings reveal that current ASR performance remains significantly below deployment standards in a narrow formal domain; the top-performing system in terms of Word Error Rate (WER) achieved 46.76% and the best Character Error Rate (CER) of 13.00% was set by another model, while several prominent multilingual models exceeded 100% WER. These results suggest that multilingual pre-training and model scaling alone are insufficient for underrepresented languages. Furthermore, because this dataset represents a best-case scenario of the most simplified and formal form of spoken Bambara, these figures are yet to be tested against practical, real-world settings. We provide the benchmark and an accompanying public leaderboard to facilitate transparent evaluation and future research in Bambara speech technology.

1 Introduction

Automatic Speech Recognition (ASR) for Bambara has seen growing interest in the past three years. Since the 2022 release of Jeli-ASR (Diarra et al., 2022), the first open ASR dataset for the language, numerous models and datasets have emerged from both research labs and community initiatives. However, this rapid growth raises concerns about quality and usability, concerns that cannot be addressed without standardized evaluation.

Quality, when it comes to low resource African languages, is the object of strong debates among

the African NLP community due to the variety of dialects, writing systems, and standards (Hussen et al., 2025), but also the complexity of the contact phenomenon between African languages and western languages, namely code switching.

As the Word Error Rate (WER) is only relevant when we have already defined and assessed the quality of the evaluation set, whatever quality means for one, some researchers recommend defaulting to human evaluation by native speakers (Lau et al., 2025; Tall, 2025). However, this process is time consuming and expensive, furthermore edit distance metrics like WER or Character Error Rate (CER) remain insightful on a curated and standardized benchmark.

However, no such benchmark existed for evaluating Bambara ASR models, most openly released models¹ report values for WER and CER on internal test sets. To address this issue and offer a *reference test set*, we publish the first Bambara ASR benchmark and leaderboard backed with experts validated transcriptions.

As more data collection initiatives for African languages emerge, often with strict rules to capture simplified language and context, such as no slang, no code-switching, no background noise etc, we have designed this first benchmark to represent an equally "pure" version of the Bambara language. Relatively poor evaluation results of models trained on more modern and accessible Bambara (see section 3) raise questions about the representativeness and usability of simplified language for real-world applications where natural data often include noise, informal terms, and code-switching. Therefore, we anticipate that this benchmark will be among the most difficult test sets for current Bambara ASR systems, covering a specialized and highly formal domain, and we argue for its

¹hf.co/facebook; facebookresearch/omnilingual-asr;
hf.co/asr-africa; hf.co/MALIBA-AI; hf.co/RobotsMali;
hf.co/djelias

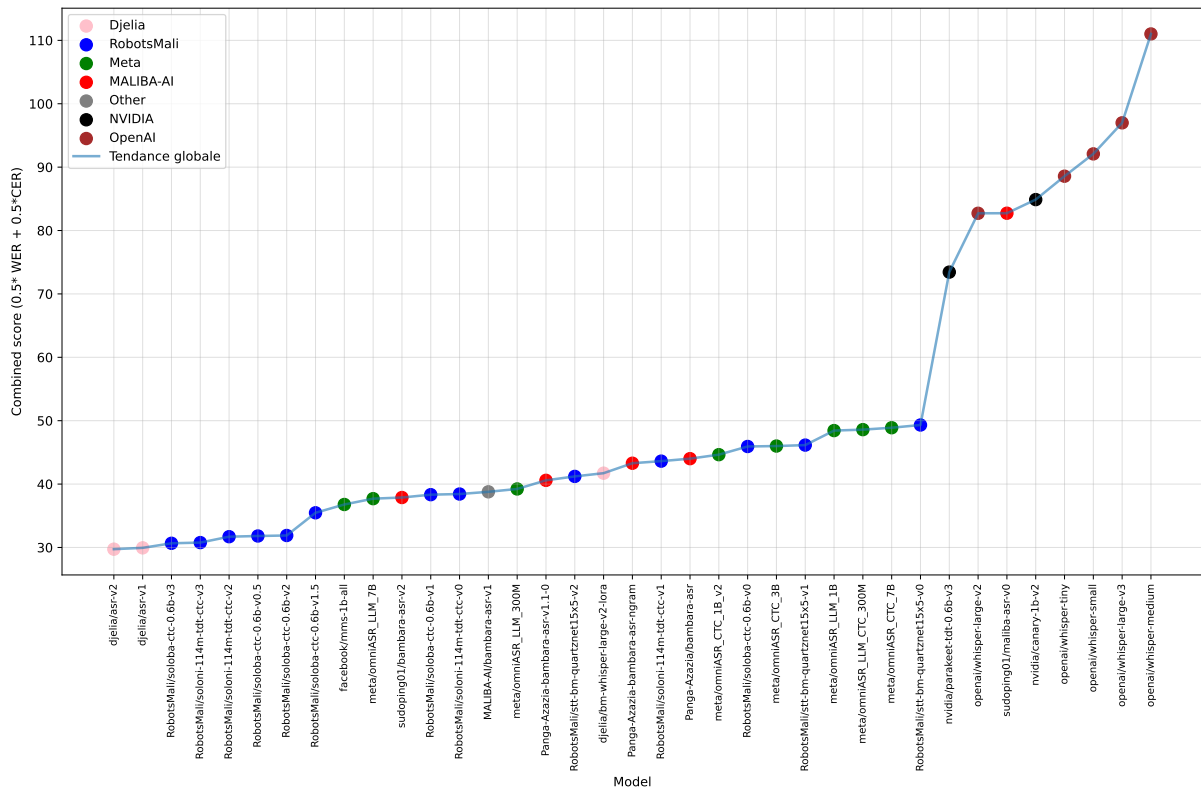


Figure 1: Models combined performance on Bambara Benchmark. Lower is better.

interpretation as a reference test set for *pure Bambara*.

2 Characteristics of the Benchmark

This first version of the evaluation set consists of a 1 hour recording of a professionally translated version of the Malian constitution, translated and recorded by the Direction Nationale de l’Education Non Formelle et des Langues Nationales (DNENF-LN)² under studio conditions, featuring one unique adult male voice.

With the premier legal text of Mali as topic, the dataset features a highly formal and diverse vocabulary that unpacks many aspects of the organization of Malian society, laws, institutions, rights and responsibilities, all written in the Bambara latin script using standard orthography and **without code switching**. The dataset also has an important representation of numbers, as the constitution contains 191 articles as of July 2023, 160 of which are clearly spelled out in the recording, specifically in ordinal forms.

We ran manual segmentation and audio-text

²DNENF-LN is the government founded organization in charge of literacy training and official documents translation in all the 13 national languages of Mali: <https://dnenfn.ml/>

alignment using the Audacity software (Audacity Team, 2024). Then we performed a final quality assurance step wherein the aligned utterances were reviewed to correct divergences resulting from the corpus’ read speech (READ) nature, specifically addressing instances where the speaker paraphrased or interpreted the text rather than providing a literal recital. This process resulted in 500 variable-length audio utterances ranging from 600 ms to 46 seconds, with a mean duration of 7.57 seconds. With this variability the benchmark aims to test models’ capabilities on both short and long form transcription.

We calculated Signal-to-noise Ratio (SNR) as an estimate of the acoustic purity of the benchmark (a higher value is best). We used the same Voice-activity-detection based implementation and classification thresholds as Diarra et al. but we calculated SNR on the segmented utterances, because the original recording features transition music and longer silences that would hinder the accuracy of the estimation as any non-speech segment is considered for estimating the Noise Power in this implementation (Diarra et al., 2025; Vondrasek and Pollák, 2005). Note that we still kept 8 of these silent/music segments in the final benchmark to

test the robustness of the models, especially the tendency to "hallucinate" tokens as the silence becomes lengthy when there is in fact no speech. Table 1 shows the SNR distribution the 492 remaining speech segments.

SNR Category	Threshold (dB)	Recordings
Medium SNR	[5, 15)	5
High SNR	[15, 25)	109
Very high SNR	≥ 25	378
Total Audios		492

Table 1: Distribution of Audio utterances by Signal-to-noise Ratio Category.

We note that 99% of the utterances are classified as relatively noise-free. This is an important point for interpreting our results: **this benchmark represents near-optimal acoustic conditions**³. Any production deployment would face significantly more challenging audio quality, so these results should be interpreted with caution given the specialized and nature of the reference test set and its acoustic purity.

3 Leaderboard and Results of Open Bambara ASR Models

We evaluated 37 publicly available ASR models on our benchmark, including monolingual ASR models, multilingual models with Bambara support, and large-scale commercial ASR systems. Table 2 presents the complete leaderboard ranked by a weighted average score of WER and CER (50% WER + 50% CER). This equal weighting reflects a neutral stance that does not privilege either word-level or character-level accuracy, treating both as equally informative for assessing transcription quality. We acknowledge that optimal weighting may depend on downstream application requirements for instance, applications sensitive to semantic accuracy may prioritize WER, while those tolerant of word boundary errors may favor CER. To address this, our public leaderboard allows users to adjust these weights according to their specific needs, and we report sensitivity analysis under alternative weightings in Table 5. All evaluations were conducted using normalized text

³In future versions, we will collect data in various domains under different recording conditions, trying to maximize diversity and real world representativeness instead of purity

(lowercase, no punctuation & consecutive whitespace) to ensure fair comparison between models.

3.1 Assessment

The main finding of this evaluation is that current Bambara ASR systems do not yet meet the commonly accepted production-readiness thresholds in the narrow domain represented in our test set. Under our combined evaluation metric, the highest-ranked model attains a Word Error Rate of 47.50%, indicating that nearly half of all words are incorrectly transcribed.

For context, production-grade ASR systems for well-resourced languages typically achieve Word Error Rates in the 5–15% range (Nahabwe et al., 2025). Current Bambara ASR performance therefore remains approximately 30–40 percentage points below these levels, suggesting a substantial gap that will require significant advances in data, modeling, and evaluation to close.

Real-world Bambara speech introduces additional challenges: phone-quality or ambient recordings, multiple speakers with varying accents and dialects, ubiquitous French code-switching, informal vocabulary, variable recording equipment, and background noise. Therefore, this benchmark gives little insight into the performance of these models with truly naturalistic speech.

3.2 Model-Specific Findings

We find that specialized fine-tunes from Djelia and RobotsMali substantially outperform their base version (parakeet, whisper) and all the other models from large multilingual initiatives.

Multilingual models exhibit high error rates. All evaluated OpenAI Whisper variants exhibit WER exceeding 100%, indicating that models generate more tokens than present in the reference audio, a hallucination phenomenon. This pattern is consistent across model sizes: whisper-tiny (112.72%), whisper-small (109.97%), whisper-medium (123.18%), whisper-large-v2 (106.84%), and whisper-large-v3 (121.06%). NVIDIA’s Parakeet-tdt-0.6b-v3 (100.06% WER) and Canary-1b-v2 (111.64% WER) show similar behavior.

These results are consistent with findings that off-the-shelf multilingual ASR models require language-specific adaptation to perform well in underrepresented languages (Nahabwe et al., 2025). It is important to note that, while multilingual, the base versions of Whisper and Canary, along with

Rank	Model	WER (%)↓	CER (%)↓	Combined (%)↓	License
1	djelia/asr-v2	47.50	13.56	29.73	Proprietary
2	djelia/asr-v1	48.56	13.00	29.94	Proprietary
3	RobotsMali/soloba-ctc-0.6b-v3	46.76	16.02	30.66	Open Source
4	RobotsMali/soloni-114m-tdt-ctc-v3	48.32	14.81	30.77	Open Source
5	RobotsMali/soloni-114m-tdt-ctc-v2	49.42	15.58	31.70	Open Source
6	RobotsMali/soloba-ctc-0.6b-v0.5	49.93	15.33	31.81	Open Source
7	RobotsMali/soloba-ctc-0.6b-v2	48.06	17.19	31.89	Open Source
8	RobotsMali/soloba-ctc-0.6b-v1.5	52.56	19.93	35.47	Open Source
9	facebook/mms-1b-all	61.06	14.71	36.78	Open Source
10	meta/omniASR_LLM_7B	62.57	15.08	37.70	Open Source
11	sudoping01/bambara-asr-v2	60.33	17.46	37.88	Open Source
12	RobotsMali/soloba-ctc-0.6b-v1	57.59	20.81	38.33	Open Source
13	RobotsMali/soloni-114m-tdt-ctc-v0	55.79	22.65	38.43	Open Source
14	MALIBA-AI/bambara-asr-v1	61.74	17.90	38.78	Open Source
15	meta/omniASR_LLM_300M	63.32	17.32	39.23	Open Source
16	Panga-Azazia/bambara-asr-v1.1-0	60.39	22.60	40.59	Proprietary
17	RobotsMali/stt-bm-quartznet15x5-v2	65.66	18.98	41.21	Open Source
18	djelia/bm-whisper-large-v2-lora	59.17	25.85	41.72	Proprietary
19	Panga-Azazia/bambara-asr-ngram	69.13	19.80	43.29	Open Source
20	RobotsMali/soloni-114m-tdt-ctc-v1	61.14	27.69	43.62	Open Source
21	Panga-Azazia/bambara-asr	70.00	20.39	44.01	Open Source
22	meta/omniASR_CTC_1B_v2	69.62	21.93	44.64	Open Source
23	RobotsMali/soloba-ctc-0.6b-v0	62.93	30.48	45.93	Open Source
24	meta/omniASR_CTC_3B	72.62	21.80	46.00	Open Source
25	RobotsMali/stt-bm-quartznet15x5-v1	72.98	21.75	46.15	Open Source
26	meta/omniASR_LLM_1B	78.31	21.29	48.44	Open Source
27	meta/omniASR_LLM_CTC_300M	76.87	22.87	48.59	Open Source
28	meta/omniASR_CTC_7B	74.65	25.47	48.89	Open Source
29	RobotsMali/stt-bm-quartznet15x5-v0	75.82	25.23	49.32	Open Source
30	vidia/parakeet-tdt-0.6b-v3	100.06	49.24	73.44	Open Source
31	openai/whisper-large-v2	106.84	60.80	82.72	Open Source
32	sudoping01/maliba-asr-v0	94.86	71.72	82.73	Open Source
33	vidia/canary-1b-v2	111.64	60.55	84.88	Open Source
34	openai/whisper-tiny	112.72	66.61	88.57	Open Source
35	openai/whisper-small	109.97	75.84	92.09	Open Source
36	openai/whisper-large-v3	121.06	75.10	96.99	Open Source
37	openai/whisper-medium	123.18	99.95	111.01	Open Source

Table 2: Bambara ASR Benchmark Leaderboard. Combined Score = $0.5 \times \text{WER} + 0.5 \times \text{CER}$. Lower scores indicate better performance.

Nvidia’s monolingual Parakeet models, included in this study, did not include Bambara in respective their training sets. However, evaluating them allowed us to rule out the hypothesis that massive multilingualism may translate to better performance on unseen, underrepresented African languages like Bambara through transfer learning. On the other end, remarkably better performance from Meta’s Omnilingual ASR and MMS models shows that even a negligible amount of Bambara data in the training set can drastically change these figures.

Model scale does not compensate for data scarcity. Meta’s omniASR family provides insight into scaling effects. The 7B parameter CTC model (74.65% WER) performs worse than the 300M LLM variant (63.32% WER), and both lag behind the 114M parameter monolingual soloni models (48.32% WER).

Character-level accuracy exceeds word-level accuracy. CER results are notably better than WER across all models, with the best achieving 13.00% (djelia/asr-v1). This suggests that models capture phonetic patterns more successfully than word boundaries and vocabulary, a pattern consistent with the challenges of morphologically rich languages where compound words and agglutination are frequent.

3.3 Qualitative Error Analysis

To better illustrate model failure modes, we present representative examples from our evaluation.

Hallucination in multilingual models. Table 3 shows severe hallucination in Whisper models, where the output contains scripts entirely unrelated to Bambara.

3.2. Benchmarking studies indicate that competitive ASR performance generally requires substantial volumes of labeled data (Nahabwe et al., 2025).

Domain mismatch. Most available Bambara speech datasets consist of over-simplified spontaneous speech with limited vocabulary, recorded under controlled conditions (Diarra et al., 2025; Diarra et al., 2022). This creates distribution mismatch when models encounter highly formal or inversely very informal registers, specialized vocabulary, or challenging acoustic conditions (Tall, 2025). Our benchmark also exposes this gap through its legal/constitutional domain.

Orthographic and dialectal variation. Standardizing written Bambara is a recent research (Konta and Vydrin, 2014; Vydrin, 2022), despite the creation of a dedicated institution—the Académie Malienne des Langues (AMALAN)—the most recent orthography is not universally adopted, and dialectal variation across regions introduces additional complexity (Imam et al., 2025). Additionally, Bambara text available on the internet often features inconsistencies, old and mixed standards, models trained on one variant may struggle with others, fragmenting an already limited data pool.

Morphological complexity. Bambara’s agglutinative morphology makes word boundary detection inherently challenging. The gap between CER and WER across models reflects this difficulty phonetic patterns are captured more successfully than word structure.

4.3 Implications for Research and Development

Our findings have several implications:

Standardized benchmarking supports progress. The field benefits from rigorous evaluation against common benchmarks. We encourage researchers to report results on standardized test sets in addition to internal evaluations.

Data collection should prioritize diversity. Current data collection efforts, while valuable, may not adequately prepare models for real-world deployment. Future efforts should consider naturalistic speech, code-switching, dialectal variation, and varied acoustic conditions.

Architecture research may be needed. The consistent underperformance of scaled multilingual models suggests that existing architectures may not be optimally suited to low-resource scenarios. Research into architectures designed for data-scarce settings may prove valuable.

Multilingual transfer has limits. The poor performance of Whisper and similar systems demonstrates that multilingual pre-training does not automatically transfer to underrepresented languages. The dominance of RobotsMali’s monolingual models suggests that, for Bambara and similar languages, targeted development appears more effective than relying on transfer from massive multilingual training.

4.4 Directions for Progress

Despite current limitations, our results suggest promising directions:

The success of smaller, Bambara-specific models (114M–600M parameters) over massive multilingual systems indicates that focused development yields better results than scale alone. The narrowing gap between proprietary and open-source solutions suggests that community-driven development can produce competitive systems. The reasonable CER performance (13–15% for top models) indicates that phonetic modeling is more tractable than word-level transcription, suggesting that improvements in language modeling and vocabulary handling through post-processing could yield significant gains.

Closing the gap to production readiness will require sustained investment in data collection, architecture research, and evaluation infrastructure at scales that do not currently exist for Bambara and similar languages.

5 Conclusion

We present the first standardized benchmark for evaluating Bambara Automatic Speech Recognition systems and provide an empirical answer to the question posed in our title: **current Bambara ASR systems are not yet ready for production deployment.**

Our evaluation of 37 ASR models on a one-hour, studio-quality benchmark reveals that:

- The best-performing model on our benchmark is **djelja/asr-v2**, achieving a Combined Score of 29.73 (WER 47.50%, CER 13.56%) under ideal conditions.

- No evaluated system reaches the 5–15% WER range typical of production-ready ASR systems.
- All OpenAI Whisper variants and commercial multilingual systems (not trained on Bambara) exhibit catastrophic failure, with WER exceeding 100%, worse than how a randomly initialized model would perform. Suggesting that transfer learning fails where similarity between the target language and training languages stops.

These results should inform expectations for Bambara ASR deployment. Current systems may be suitable for research and development purposes, but deployment in production applications where users depend on accurate transcription should be approached with caution.

The benchmark and leaderboard are publicly available to support continued development and enable rigorous comparison of future systems. We hope this resource contributes to honest assessment of progress and motivates the sustained investment necessary to achieve production-ready Bambara ASR.

6 Limitations

This benchmark has several limitations:

Simplified evaluation conditions. Our benchmark represents near-ideal acoustic conditions: studio recording, professional speaker, high SNR, standardized orthography. Although we do speculate that the metrics reported here likely represent upper bounds on real-world performance, this assertion may not hold if some of the models that we evaluate have been trained on more naturalistic data. In other terms, the inverse assertion that models trained on natural data may experience more struggle on this benchmark may also be a valid interpretation.

Single speaker and domain. The current version features recordings from a single adult male speaker reading constitutional text. This limits assessment of speaker and domain variability, though it also provides a consistent and controlled evaluation environment.

Limited size. One hour of audio is a minimal benchmark. However, consistent patterns across 37 models suggest findings would generalize to larger evaluations.

Metric limitations. WER and CER may not optimally capture transcription quality for morphologically rich languages. Future work could explore morpheme-level metrics or semantic similarity measures.

Normalization sensitivity. Our evaluation applied minimal text normalization (lowercase, punctuation removal, whitespace normalization) to ensure fair comparison. However, Bambara orthography permits substantial valid variation that our normalization does not fully address. Contractions such as *b'a* versus *bε a*, or the ambiguous *k'a* which can legitimately expand to *ka a*, *kε a*, or *ko a* depending on grammatical context, represent equivalent transcriptions that would be penalized as errors under standard WER computation. Similarly, compound word segmentation (*yεɛmahɔɾɔnya* versus *yεɛma hɔɾɔnya*) and legacy orthographic variants (*è/ε*, *ny/ɲ*) introduce scoring artifacts unrelated to recognition accuracy. A more sophisticated normalization framework that accounts for these linguistic equivalences could yield different and potentially more meaningful error rates. Future work should investigate normalization strategies that distinguish genuine recognition errors from valid or outdated orthographic variation.

Code-switching. Real Bambara speech frequently incorporates French, particularly in urban context but also formal settings, quite frequently. However, this first benchmark does not inform on a model ability to handle code-switching as this feature is deliberately absent from the data.

We view this benchmark as a foundation for continued development, with future versions incorporating speaker diversity, domain variation, naturalistic speech, and code-switching.

Data and Code Availability

The benchmark dataset, evaluation code, and public leaderboard are available to support reproducibility and future research:

Benchmark Dataset :

- <https://huggingface.co/datasets/MALIBA-AI/bambara-asr-benchmark>

Public Leaderboard :

- <https://huggingface.co/spaces/MALIBA-AI/bambara-asr-leaderboard>

- <https://github.com/MALIBA-AI/bambara-asr-leaderboard>

We encourage researchers to submit their model results to the leaderboard and to report performance on this benchmark in future publications.

References

- Audacity Team. 2024. *Audacity(r)*. Free software distributed under the terms of the GNU General Public License (GPL). Accessed: February, 2025.
- Sebastien Diarra, Michael Leventhal, and Allahsera Auguste Tapo. 2022. Robotsmali griots speech dataset, and asr. <https://github.com/robotsmali-ai/jeli-asr/>.
- Yacouba Diarra, Nouhoum Souleymane Coulibaly, Panga Azazia Kamaté, Madani Amadou Tall, Emmanuel Élisé Koné, Aymane Dembélé, and Michael Leventhal. 2025. *Dealing with the hard facts of low-resource african nlp*. *Preprint*, arXiv:2511.18557.
- Kedir Yassin Hussen, Walelign Tewabe Sewunetie, Abinew Ali Ayele, Sukairaj Hafiz Imam, Shamsuddeen Hassan Muhammad, and Seid Muhie Yimam. 2025. *The state of large language models for african languages: Progress and challenges*. *Preprint*, arXiv:2506.02280.
- Sukairaj Hafiz Imam and 1 others. 2025. Automatic speech recognition (asr) for african low-resource languages: A systematic literature review. *arXiv preprint arXiv:2510.01145*.
- Mamadou Konta and Valentin Vydrin. 2014. Propositions pour l’orthographe du bamanankan. *Mandenkan*, (52):3–38.
- Mingfei Lau, Qian Chen, Yeming Fang, Tingting Xu, Tongzhou Chen, and Pavel Golik. 2025. *Data quality issues in multilingual speech datasets: The need for sociolinguistic awareness and proactive language planning*. *Preprint*, arXiv:2506.17525.
- Alvin Nahabwe and 1 others. 2025. Benchmarking automatic speech recognition models for african languages. *arXiv preprint arXiv:2512.10968*.
- Madani Amadou Tall. 2025. *Analyse comparative humaine des modèles asr bambara de robotsmali*.
- Martin Vondrasek and Petr Pollák. 2005. Methods for speech snr estimation: Evaluation tool and analysis of vad dependency. *Radioengineering*, 14.
- Valentin Feodosievich Vydrin. 2022. *Vers un dictionnaire orthographique bambara*. *Mandenkan : Bulletin Semestriel d’Études Linguistiques Mandé*, (68):59–82.

Enhancing Automatic Speech Recognition Models for Maternal and Reproductive Health: Fine-Tuning and Real-World Evaluation in Wolof

Ertony Basilwango¹, Yann Le Beux¹, Oche David Ankele¹, Pierre Herve Berdys¹

Dhananjay Balakrishnan^{1,2},

¹ YUX Design, Senegal

²Stanford University, USA

{yann, pierre, erton, oche}@yux.design

dhananj@stanford.edu

Abstract

Automatic Speech Recognition (ASR) systems perform well for high-resource languages, but most African languages, including Wolof, remain underrepresented, particularly in maternal and reproductive healthcare. This work proposes a domain-specific approach to improving Wolof ASR under low-resource conditions, addressing limited annotated data, orthographic variability, and code-switching. We curated a dataset of 750 validated Wolof utterances covering 250 maternal health keywords and applied data augmentation to increase acoustic diversity. Pretrained models, including wav2vec 2.0 and Whisper, were benchmarked to select candidates for fine-tuning. Using parameter-efficient Low-Rank Adaptation (LoRA), a Whisper model was adapted to the maternal health domain. Evaluation using Word Error Rate (WER), Character Error Rate (CER), and Keyword Error Rate (KER), which measures medically critical term transcription accuracy, shows substantial gains, reducing WER from 46.5% to 23.2% and KER from 17% to 11%. Community-based evaluation on 1,340 real-world utterances reveals a moderate degradation, with WER increasing by 35%. These results demonstrate that lightweight domain adaptation with small, high-quality data can significantly improve ASR for low-resource healthcare applications. This work introduces one of the first Wolof ASR datasets for healthcare and presents a practical framework for developing reliable speech recognition tools in underrepresented languages, improving access to healthcare information and services

Keywords: Automatic Speech Recognition, Low-Resource Languages, Wolof, Maternal Health, Data Augmentation, Domain Adaptation, Real World ASR Evaluation.

1 Context problem statement

Oral communication plays a central role in daily life across African societies, yet speech technolo-

gies have not evolved at the same pace for local languages (Caubrière and Gauthier, 2024). Existing transcription tools offer limited or no support for the majority of African languages, restricting the development of research, digital services, and domain-specific applications in health, finance, education, and governance (Imam et al., 2025). Although the African continent is home to more than 2,000 languages, only a very small fraction has any form of automatic speech recognition (ASR) resources, and even fewer have systems that perform reliably in real-world contexts. The few available corpora often contain government speeches or religious texts, which do not reflect spontaneous, conversational speech, and therefore hinder downstream performance in other domains (Mak et al., 2024).

Wolof poses significant challenges for ASR due to its rich phonology (17 vowels and 45 consonants) (Cissé and Sadat, 2023) and high orthographic variability shaped by regional usage, French code-switching, and the lack of a standardized writing system, leading to transcription and pronunciation inconsistencies that hinder both acoustic and language modeling (Bourdeau, 2024), (Cissé and Sadat, 2023). In Wolof, such variations are further amplified by flexible orthographic conventions and morphophonological alternations, making normalization and error handling essential preprocessing steps for reliable ASR (Aliou, 2010).

This technological gap disproportionately affects maternal and reproductive health, a domain where accurate documentation and communication are critical for patient safety, continuity of care, and public health monitoring. The lack of reliable ASR systems for local languages, increasing the risk of information loss and clinical error, further marginalizes already underserved populations and prevents healthcare systems from leveraging AI-driven efficiencies.

2 Related works

2.1 ASR for Low-Resource African Languages

Automatic speech recognition (ASR) for African languages remains challenging due to limited data, high linguistic diversity, and the lack of standardized writing systems. Most African languages do not have enough labeled speech data, making it hard to build reliable ASR systems (Hedderich et al., 2021). Available datasets often come from formal sources, such as government speeches or broadcast news, which do not reflect everyday, conversational speech (Gauthier et al., 2016). Recent initiatives like AfriSpeech-200 aim to improve coverage across African accents (Olatunji et al., 2023), but they still do not fully capture domain-specific contexts such as healthcare, where vocabulary, pronunciation, and discourse differ. Evaluating ASR in low-resource settings is also difficult because there are no standard benchmarks or consistent annotation practices. Common metrics, like Word Error Rate (WER) and Character Error Rate (CER), are widely used to measure transcription accuracy (Jurafsky and Martin, 2009). However, these metrics are very sensitive to differences in spelling and writing style, which is especially challenging for African languages with evolving orthographies.

2.2 Wolof ASR and Linguistic Challenges

Wolof, spoken by 10–12 million people in Senegal, The Gambia, and Mauritania, is still underrepresented in speech technology. The early datasets (Gauthier et al., 2016);(Diop, 2015);(Aliou, 2010) are limited in size and domain. The language shows a wide spelling variation, code-switching with French, and complex sound changes, which make ASR challenging (Aliou, 2010), (Bourdeau, 2024). Recent datasets, such as Kallaama (Gauthier et al., 2024), cover agriculture, but domain mismatch remains, highlighting the need for specialized corpora and preprocessing, especially for sensitive areas such as healthcare.

2.3 Domain Adaptation and Fine-Tuning for ASR

To address data scarcity, researchers have adapted pretrained ASR models to low-resource languages using limited labeled data. Self-supervised models like wav2vec 2.0 (Baevski et al., 2020), Massively Multilingual Speech(MMS)(Pratap et al., 2023) and Whisper (Radford et al., 2022) learn

robust speech representations that transfer well. Parameter-efficient fine-tuning methods such as Low-Rank Adaptation (LoRA) (Hu et al., 2021) allow adaptation with lower computational cost, which is important in African contexts with limited GPU access. Prior work shows that lightweight, language-specific adaptation can outperform multilingual models trained on larger but less relevant datasets (Caubrière and Gauthier, 2024). These approaches are especially useful for domain-specific tasks, such as maternal health, where collecting labeled data is challenging.

2.4 Data Augmentation and Robustness

Data augmentation is commonly used to improve ASR in low-resource settings. Methods like speed perturbation, noise injection, pitch shifting, and volume changes can reduce Word Error Rate (WER) (Alex et al., 2023); (Ko et al., 2015), especially when collecting real-world data is difficult, such as in healthcare. However, models trained only on augmented data may still struggle in real-world conditions (Flynn and Ragni, 2024). Using imperfect or noisy speech can also help; carefully selected community-recorded data can improve acoustic modeling when clean data is scarce. (Badenhorst and de Wet, 2019).

2.5 Real-World and Human-Centered Evaluation of ASR

Standard ASR evaluation on clean benchmark datasets often fails to reflect real-world performance. Models tested on a single dataset can lose 35–50% WER when deployed in different domains or acoustic conditions (Likhomanenko et al., 2021); (Shah et al., 2024); In healthcare, this is critical: even models with low overall WER can misrecognize medical terms, risking patient safety (Afonja et al., 2024). Fine-tuning on domain-specific, accented clinical speech improves recognition of medical entities, highlighting the need for domain-aware and human-centered evaluation. Moreover, speech quality metrics, like DNSMOS (Reddy et al., 2021), do not always predict ASR accuracy, emphasizing the importance of direct evaluation on real-world data.

3 Study Area, Data, and Methods

3.1 Data collection

In collaboration with domain experts and community partners, we created a Wolof dataset of 750

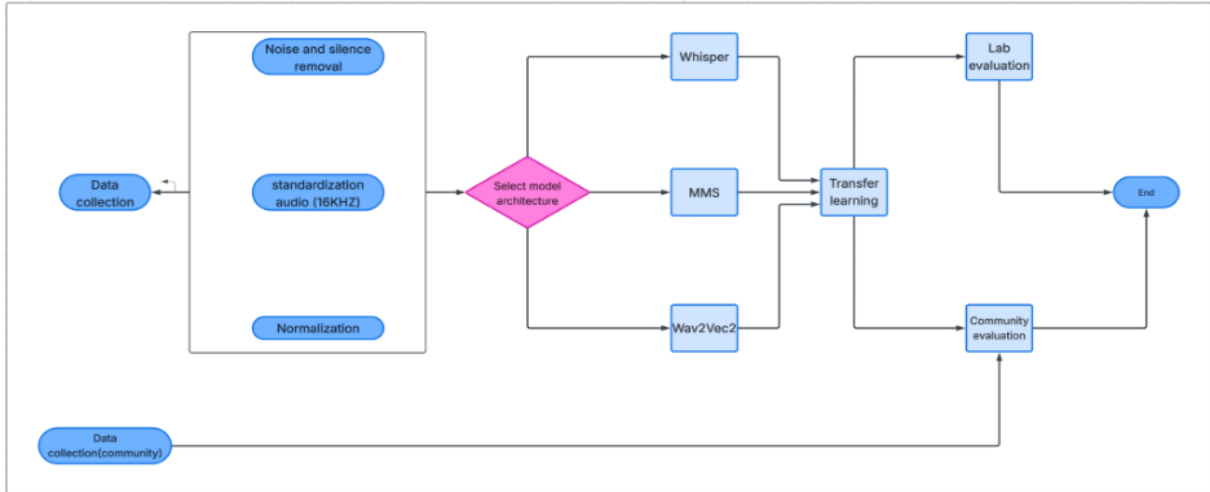


Figure 1: High-Level Workflow

utterances (2 hours) based on 250 maternal and reproductive health keywords spanning antenatal care, delivery, complications, infections, and family planning. Phrases were translated and culturally validated, then recorded via the Looka platform, a digital data collection system that facilitates remote recording and survey participation, by 20 regionally diverse speakers with balanced gender representation. To reduce sensitivity to individual speakers or specific keywords, we prioritized speaker diversity, applied data augmentation, and performed utterance-level splits while maintaining speaker diversity across train and test sets. While some sensitivity is inevitable in low-resource settings, this design mitigates overfitting to particular speakers or lexical items. Each iteration was reviewed by linguists, health professionals, and social scientists to identify and correct potential biases.

3.2 Ethics, Privacy, and Deployment Constraints

Given the sensitive nature of maternal and reproductive health data, all participants provided informed consent. Audio files were anonymized and securely stored, and risk mitigation measures were implemented to ensure participation posed no harm. These steps make the dataset ethically suitable for research in healthcare contexts.

3.3 Pre-processing

- Normalization

Wolof shows significant lexical and orthographic variation, where multiple forms can express the same meaning depending on region, speaker, or

Mispellings	Correct wolof
dadial	dajale
guinaw	guinnaw
Mousiba	Musiba
Infection	infekcion
gnakk	ñakk
thiosane	cosaan

Table 1: Wolof Spelling Normalization Examples

transcription conventions (Afonja et al., 2024). To address this, we applied a multi-stage normalization pipeline, including spelling standardization, letter-case normalization, canonical representation of numbers, and removal of non-essential punctuation (Rahimi and Homayounpour, 2022).

- Data augmentation

To improve the robustness of ASR in low-resource settings, we applied standard data augmentation techniques known to reduce WER (Ko et al., 2015; Bagchi et al., 2020). Using speed perturbation ($\pm 10\%$), pitch shifting (± 2 semitones), volume adjustment (± 3 dB), and additive background noise (SNR 20–30 dB) in the original data set resulted in approximately 10 hours of data in total. This approach complements naturally noisy speech, which is beneficial in low-resource ASR settings (Badenhorst and de Wet, 2019).

3.4 Benchmarking of ASR Models on Maternal and Reproductive Health Data

Before fine-tuning, open-source models were benchmarked on 120 cleaned samples from our maternal health dataset using Word Error Rate

Models	WER	CER	Latency(s)
Alwaly/whisper	0.464	0.172	1.394
facebook/mms	0.526	0.181	0.55
bilalfaye/wav2vec2	0.533	0.186	0.55
CAYTU/whisper	0.544	0.228	1.626
cibfaye/whisper	0.546	0.224	0.463

Table 2: Benchmarking of ASR models on maternal and reproductive health Wolof data.

(WER) and Character Error Rate (CER) (Rahimi and Homayounpour, 2022). Models with WER below 50% were selected for domain-specific adaptation, as they provide a sufficient baseline for learning from limited in-domain data. The threshold therefore serves as a computationally efficient screening step rather than an optimal boundary. This pre-selection follows established ASR transfer learning practice, where models with reasonable initial performance adapt more effectively than poorly performing ones (Baeviski et al., 2020); (Hu et al., 2021)

- Model Notation

These models are referred to by their abbreviated names in the tables and figures throughout the paper

- Alwaly/whisper: Alwaly/whisper-medium-wolof
- CAYTU/whisper: CAYTU/whisper-large-v2
- Facebook/mms: Facebook/mms-1b-fl102
- bilalfaye/wav2vec2: bilalfaye/wav2vec2-large-mms-1b-wolof

3.5 Fine-tuning Approach

We fine-tuned three different speech recognition models on Wolof medical conversations using Low-Rank Adaptation (LoRA), a lightweight fine-tuning method that updates only a small part of each model (Hu et al., 2021). This approach enables the models to capture local speech patterns, accents, and domain-specific pronunciations even with a small amount of annotated data, while keeping computational requirements low and training times short. By updating only a subset of parameters, LoRA provides an efficient way to specialize large pre-trained models to underrepresented languages and healthcare-specific speech without the need for extensive hardware or massive datasets.

The maternal and reproductive health data set was divided into 80% for training (10 hours of

audio) and 20% for testing (2 hours). Fine-tuning was performed for 20 epochs with evaluation at each epoch, using mixed-precision training (fp16) and the AdamW optimizer (Loshchilov and Hutter, 2019). Training was conducted on an NVIDIA A100-SXM4 GPU (40 GB) with CUDA 12.4.

The results show that the Alwaly model achieved the lowest Word Error Rate (WER)

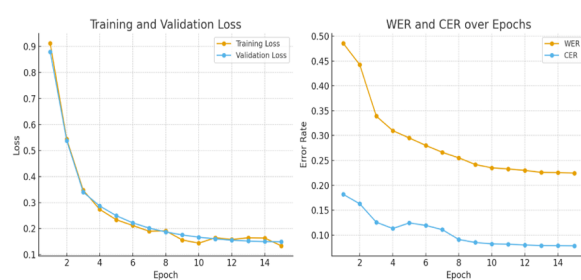


Figure 2: Training and validation loss curves of the Alwaly model after fine-tuning

The first graph shows the training and validation loss, both decreasing over epochs, indicating successful model convergence. The second graph illustrates the reduction in WER (orange) and CER (blue), reflecting improvements in the model’s speech transcription accuracy, particularly for maternal and reproductive health in Wolof.

Models	WER	CER
Alwaly/whisper	0.23	0.172
CAYTU/whisper	0.374	0.181
Facebook/mms	0.406	0.308

Table 3: Model performance after fine-tuning on maternal and reproductive health data

In addition to the main Whisper-based model, we also fine-tuned CAYTU/Whisper-large-v2 and Facebook/MMS. Although both benefited from LoRA adaptation, the performance gains were smaller and their higher inference latency makes them less suitable for lightweight deployment in low-resource healthcare settings. These results confirm that selecting a pretrained model with a reasonable baseline WER is crucial for effective domain adaptation and demonstrate that our conclusions are not limited to a single model architecture.

To assess the impact of data augmentation, we conducted an additional experiment fine-tuning the selected Alwaly/Whisper model using only the original 2-hour dataset without augmentation. This setting resulted in a WER of 29%, compared

to 23% WER when training with augmented data. This confirms that data augmentation plays a significant role in improving robustness under low-resource conditions. While all these augmentation techniques were applied collectively, we did not separately quantify the individual contribution of each method.

3.6 Evaluation of medical terms

Recent research has highlighted weaknesses in relying solely on the word error rate (WER) for morphologically rich languages (K et al., 2025). To capture domain-specific accuracy, we used the Keyword Error Rate (KER), focusing on maternal and reproductive health terms. KER evaluates transcription quality at the character level by aligning predicted and reference keywords. It is calculated using dynamic programming to estimate the edit distance of Levenshtein (Levenshtein, 1966), which calculates the difference between two strings by counting the insertions, deletions, and substitutions needed to transform one string into another.

Standard metrics such as WER and CER often fail to capture errors in domain-critical terms. To address this, we evaluated the keyword error rate (KER), a metric that quantifies the precision of recognition of a predefined set of maternal and reproductive health keywords, thereby assessing the model’s ability to capture domain-critical terminology. Keywords were detected using fuzzy string matching. Fuzzy matching compares strings by measuring edit distance, that is, the minimum number of character-level insertions, deletions, or substitutions required to transform one string into another, allowing minor differences between recognized and reference text. A sliding-window alignment was applied to match short segmented text with the target keywords. Fine-tuning reduced KER from 0.169 to 0.11, demonstrating improved recognition of medically important terms and highlighting the value of domain-sensitive evaluation in healthcare ASR.

$$\text{KER} = \frac{S + D + I}{N}$$

Where S, D, I and N correspond to the number substitutions, deletions and insertions in the total number of keyword found. To identify keywords in ASR outputs, we applied fuzzy string matching with text normalization and a sliding-window search, selecting the closest match based on Levenshtein similarity.

Sliding-Window Search is used to find matching substrings between the expected keyword and the ASR output by comparing small portions of both strings and selecting the best match as the window moves across the text. **Levenshtein Similarity** then calculates the inverse of the Levenshtein distance, quantifying the similarity between the two strings. This method played a key role in the model’s fine-tuning, which resulted in a significant reduction in the Keyword Error Rate (KER), dropping from 0.17 to 0.11. This improvement highlights the model’s increased ability to capture domain-specific, medically relevant terminology.

Model	WER	CER	KER
Alwaly	46.46%	17%	17%
Alwaly(fine-tuned)	23.16%	7.83%	11%

Table 4: Model performance evaluation based on the KER metric (Medical Keyword Error Rate)

3.7 Evaluation on F1-score

The F1-score was used to evaluate performance due to its effectiveness in health-related assessments, especially for measuring accurate identification of medical terminology. An F1-score of 0.809 reflects good model performance, indicating reliable recognition of medical keywords. (Sokolova and Lapalme, 2009)

3.8 Human-Centered Evaluation

Standard ASR benchmarks often fail to reflect real-world conditions, with performance degrading on noisy, out-of-domain speech (Likhomanenko et al., 2021). Following recommendations for community-grounded evaluation (Khan et al., 2024), we assessed our Wolof ASR using community-collected recordings capturing natural variation in accent, pronunciation, and recording environments.

Participants from multiple regions of Senegal produced their own maternal and reproductive health phrases and recorded them in everyday settings. The resulting dataset contains 1,340 unaugmented recordings (3 hours of speech), balanced by gender (50% female) and dialectally diverse. Compared to controlled data, real-world evaluation revealed a performance drop (WER: 0.23 → 0.31; CER: 0.08 → 0.12), driven by ambient noise, regional accents, and spontaneous speech. These results demonstrate that human-centered, ecolog-

ically valid evaluation is essential for deploying ASR systems in low-resource contexts.

Evaluation setting	WER	CER
Internal test set (clean)	23%	8%
Real-world evaluation	31%	12%

Table 5: Human Centered Evaluation

3.9 Error Analysis and Linguistic Observations

Qualitative analysis of Wolof ASR outputs reveals that many residual “errors” correspond to orthographic or phonetic variants rather than true recognition failures. These variants often arise from natural speech patterns, regional accents, and code-switching, while still preserving the intended meaning. As a result, conventional WER and CER metrics may overestimate error rates in such contexts. Table 6 provides representative examples of these variants.

These observations motivate the use of Keyword Error Rate (KER) alongside WER and CER for domain-specific ASR evaluation, and normalization prior to computing evaluation metrics or fine-tuning the model on Wolof data.

4 Experiments and Results

4.1 Benchmarking of Pretrained ASR Models

We evaluated several open-source ASR models on 120 Wolof maternal and reproductive health utterances using Word Error Rate (WER), Character Error Rate (CER), and inference latency. Domain-adapted and Wolof-specific models consistently outperformed general multilingual systems. Alwaly/whisper-medium-wolof achieved the best baseline performance (WER: 46.4%, CER: 17.2%) and was selected for further adaptation, while models with WER above 50% were excluded.

4.2 Fine-Tuning with LoRA

The selected Whisper model was fine-tuned using Low-Rank Adaptation (LoRA) on approximately 10 hours of augmented maternal health speech. Fine-tuning substantially reduced error rates (Table 4), confirming the effectiveness of lightweight domain adaptation for medical ASR. To reflect real-world usage, evaluation was performed with consistently reduced CER across models, indicating improved robustness to orthographic variation in Wolof.

To better capture domain-critical errors, we evaluated Keyword Error Rate (KER) on maternal and reproductive health terminology using normalized text and fuzzy string matching. Fine-tuning reduced KER from 0.169 to 0.11, demonstrating improved recognition of medically important terms and highlighting the limitations of WER and CER for domain-specific evaluation.

4.3 Human-Centered Real-World Evaluation

We further evaluated the model on 1,340 community-recorded utterances collected in everyday environments across Senegal. Compared to controlled test data, performance degraded (WER: 0.23 \rightarrow 0.31; CER: 0.08 \rightarrow 0.12), reflecting the impact of noise, regional accents, and spontaneous speech. This confirms that benchmark-only evaluation overestimates real-world ASR performance

5 Key contributions

Our work further demonstrates that, even with a small, carefully curated dataset, effective ASR performance can be achieved through data augmentation and LoRA fine-tuning, while evaluation on a larger, real-world dataset ensures robustness. This shows that in low-resource, domain-specific contexts, strategically combining high-quality limited data with larger evaluation corpora allows for reliable ASR deployment, making it possible to support critical applications such as maternal and reproductive healthcare in Wolof-speaking communities.

6 Limitations and Perspectives

This study relies on a small curated dataset and a predefined set of maternal health keywords, designed to prioritize linguistic validity, domain relevance, and speaker diversity rather than scale. As a result, findings may not fully generalize to unrestricted conversational speech. Future work will explore larger speaker pools, broader vocabularies, and systematic ablation studies. We plan to improve Wolof ASR using multilingual and cross-lingual strategies, leveraging related languages to enhance performance in extremely low-resource healthcare settings. Additional training data will include more speakers, regions, everyday environments, and advanced augmentation techniques to better reflect real-world variability. Future efforts will also focus on developing population-specific lexicons tailored to education level, region, and

Reference	Hyphothesis	Error category
Klinik	Kilinig	accent / code-switching
Pediatre	Pejiatre	accent / accent / code-switching
Dagnuy	Danu	standard variant
infection	infekcion	code-switching variant

Table 6: Qualitative analysis of model errors

health context, enabling finer-grained evaluation and adaptation. Finally, we aim to expand linguistic and dialectal coverage to include additional Wolof dialects and other low-resource African languages.

7 Conclusion

This study shows that lightweight LoRA fine-tuning of pretrained ASR models, combined with small but carefully curated domain-specific datasets and real-world evaluation, can substantially improve Wolof ASR performance for maternal and reproductive health. Although benchmark results show strong gains, performance degradation in naturalistic conditions underscores the need for human-centered and domain-aware evaluation when developing ASR systems for low-resource healthcare settings. By focusing on Wolof maternal health, this work serves as a case study demonstrating how targeted data collection and lightweight adaptation can yield effective domain-specific ASR. These findings provide a practical foundation for extending similar approaches to other low-resource languages and healthcare domains in future research.

References

- Tejumade Afonja, Tobi Olatunji, Sewade Ogun, Naome A. Etori, Abraham Owodunni, and Moshood Yekini. 2024. [Performant ASR Models for Medical Entities in Accented Speech](#). In *Interspeech 2024*, pages 2315–2319.
- Ashish Alex, Lin Wang, Paolo Gastaldo, and Andrea Cavallaro. 2023. [Data augmentation for speech separation](#). *Speech Communication*, 152:102949.
- Diuf Aliou. 2010. [Some morphological phenomena in suffixations of wolof language](#). *Russian Journal of Linguistics*, (3):50–53.
- Jaco Badenhurst and Febe de Wet. 2019. [The usefulness of imperfect speech data for asr development in low-resource languages](#). *Information*, 10(9):268.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. [wav2vec 2.0: A framework for self-supervised learning of speech representations](#). In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.
- Corentin Bourdeau. 2024. [The wolof basic clause and its information-structural derivatives](#). *Linguistique et langues africaines*, 10(2).
- Antoine Caubrière and Elodie Gauthier. 2024. [Représentation de la parole multilingue par apprentissage auto-supervisé dans un contexte subsaharien](#). In *Actes des 35èmes Journées d’Études sur la Parole*, pages 163–172, Toulouse, France. ATALA and AFPC.
- Thierno Ibrahima Cissé and Fatiha Sadat. 2023. [Automatic spell checker and correction for under-represented spoken languages: Case study on Wolof](#). In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 1–10, Dubrovnik, Croatia. Association for Computational Linguistics.
- T. I. Cissé and F. Sadat. 2023. [Automatic spell checker and correction for under-represented spoken languages: Case study on wolof](#). *arXiv*.
- Faje Fatu Diop. 2015. [Active processes in the senegalese linguoculture and the wolof language in the course of globalization](#). *RUDN Journal of Language Studies, Semiotics and Semantics*, 3:89–94. English translation of Russian original.
- Robert Flynn and Anton Ragni. 2024. [How much context does my attention-based asr system need?](#) In *Proceedings of Interspeech 2024*, Kos, Greece. ISCA.
- Elodie Gauthier, Aminata Ndiaye, and Abdoulaye Guissé. 2024. [Kallaama: A transcribed speech dataset about agriculture in the three most widely spoken languages in Senegal](#). In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages @ LREC-COLING 2024*, pages 10–19, Torino, Italia. ELRA and ICCL.
- Etienne Gauthier, Laurent Besacier, Stephanie Voisin, Mekuriaw Melese, and Ulrich P. Elingui. 2016. [Collecting resources in sub-saharan african languages for automatic speech recognition: A case study of wolof](#). In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.

- Michael A. Hedderich, Lukas Lange, Heike Adel, Jan-nik Strötgen, and Dietrich Klakow. 2021. [A survey on recent approaches for natural language processing in low-resource scenarios](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2545–2568, Online. Association for Computational Linguistics.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahmed, Ibrahim Said Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025. [Automatic speech recognition for African low-resource languages: Challenges and future directions](#). In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 89–94, Vienna, Austria. Association for Computational Linguistics.
- Dan Jurafsky and James H. Martin. 2009. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall, Upper Saddle River, N.J.
- Thennal D K, Jesin James, Deepa Padmini Gopinath, and Muhammed Ashraf K. 2025. [Advocating character error rate for multilingual ASR evaluation](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4926–4935, Albuquerque, New Mexico. Association for Computational Linguistics.
- Hania Khan, Aleena Fatima Khalid, and Zaryab Hassan. 2024. [Transcending controlled environments assessing the transferability of asrrobust nlu models to real-world applications](#). *Preprint*, arXiv:2401.09354.
- Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. [Audio augmentation for speech recognition](#). In *Interspeech 2015*, pages 3586–3589.
- Vladimir I. Levenshtein. 1966. [Binary codes capable of correcting deletions, insertions, and reversals](#). *Soviet Physics Doklady*, 10(8):707–710.
- Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Paden Tomasello, Jacob Kahn, Gilad Avidov, Ronan Collobert, and Gabriel Synnaeve. 2021. [Rethinking evaluation in asr: Are our models robust enough?](#) *Preprint*, arXiv:2010.11745.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Franco Mak, Avashna Govender, and Jaco Badenhorst. 2024. [Exploring asr fine-tuning on limited domain-specific data for low-resource languages](#). *Journal of the Digital Humanities Association of Southern Africa*, 5(1):—.
- Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. 2023. [Afrispeech-200: Pan-african accented speech dataset for clinical and general domain asr](#). *Preprint*, arXiv:2310.00274.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. [Scaling speech technology to 1,000+ languages](#). *Preprint*, arXiv:2305.13516.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *Preprint*, arXiv:2212.04356.
- Zahra Rahimi and Mohammad Mehdi Homayounpour. 2022. [The impact of preprocessing on word embedding quality: A comparative study](#). *Language Resources and Evaluation*.
- Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. 2021. [Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors](#). In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2021*.
- Muhammad A. Shah, David Solans Noguero, Mikko A. Heikkilä, Bhiksha Raj, and Nicolas Kourtellis. 2024. [Speech robust bench: A robustness benchmark for speech recognition](#). *Preprint*, arXiv:2403.07937.
- Marina Sokolova and Guy Lapalme. 2009. [A systematic analysis of performance measures for classification tasks](#). *Information Processing Management*, 45(4):427–437.

Eyaa-Tom 26, Yodi - Mantissa and Lom Bench: A Community Benchmark for TTS in Local Languages

Justin E. Bakoubolo¹ Catherine N. N. Essuman¹ Messan Agbobl^{1,2}

Ahoefa Kansiw^{1,2} Eli Jean DOUMEYAN¹ Julie PATO¹

Notou Your Timibe^{1,3} Agossou^{1,4} Guedela Bakouya^{1,2} and al.

¹Umbaji, Lomé, Togo

²University of Lomé, Togo ³University of Kara, Togo ⁴UCAO

{justin.bakoubolo, catherine.essuman}@umbaji.org

Abstract

We present an extension of our previous work on multilingual NLP for Togolese languages by introducing new datasets, improved models, and a community-driven evaluation benchmark for Text-To-Speech (TTS). We expand the Eyaa-Tom multilingual corpus with additional speech data of about 26.9k recordings (30.9 hours) across 10 local languages, and incorporated 64.6k clips (46.6 hours) of Mozilla Common Voice contributions for Adja, Nawdm, Mina, and Tem to strengthen Automatic Speech Recognition (ASR) and speech synthesis. We detail how community contributors – including collaboration with a national TV journalist – helped collect and validate the Kabyè and French text, with an ethical compensation model in place. We fine-tune state-of-the-art models: OpenAI Whisper and faster-whisper, and Meta’s NLLB-200 model for machine translation across 11 languages (achieving 19.4 BLEU score for French→Ewe and 26.1 BLEU score for Kabyè→French). We also introduce the Lom Bench, a community-based benchmark where native speakers rate TTS output, indicating promising preliminary results in Mina and Togolese lingua franca french although further data is needed. We provide a comparative analysis of our results with recent multilingual systems, including Simba, Meta’s Omnilingual ASR, and UBC Toucan. Our work emphasizes practical pathways and how FAIR data sourcing and community participation can drive sustainable NLP development for under-served languages.

1 Introduction

Togo is home to over forty languages, including prominent Niger-Congo tongues such as Ewè, Kabyè, Tem (Kotokoli) and Adja, as well as Moba, Losso, Mina and others. Despite their cultural importance, these languages remain under-represented in NLP resources: there are few parallel corpora, speech datasets or high-quality transla-

tion and speech models. This dearth of data hinders the development of inclusive technologies for news dissemination, education, and accessibility. Our previous work *YodiV3*, took a first step by releasing the Eyaa-Tom dataset and baseline ASR/MT systems for ten Togolese languages and introduced the *Lom* metric to assess language readiness (Bakoubolo et al., 2024).

Recent advances in multilingual NLP provide an opportunity to bridge the gap. Meta’s *No Language Left Behind* (NLLB) trained translation models covering 200 languages using sparse Mixture-of-Experts (Team, 2022), and *Massively Multilingual Speech* (MMS) scaled speech models to over 1,000 languages using self-supervised pre-training (Pratap et al., 2023a). *Omnilingual ASR* further expands ASR coverage to 1,600 languages by combining public resources with community sourced data and scaling to 7B parameters (Keren et al., 2025). The University of British Columbia’s *Toucan* models fine-tune large pretrained LMs to support 156 African language pairs and introduce *AfroLingua-MT* (Elmadany et al., 2024), demonstrating high quality translation for languages such as Ewè and Kabyè. These developments show that high-resource models can be adapted to low-resource languages when complemented by carefully curated data, yet they often overlook local dialects or fail to capture regional accents and proper names.

This paper presents *Yodi-Mantissa*, a comprehensive update that combines community-driven compensated data collection, model fine-tuning and human-centered evaluation. We expand Eyaa-Tom with new parallel text and recorded speech, integrate Common Voice data and additional monolingual corpora, and develop a community platform to engage contributors ethically. We fine-tune NLLB, Whisper and TTS models on these resources, achieving notable improvements. To evaluate synthetic speech, we use the Lom bench-

mark where native speakers rate TTS output. Our work emphasizes reproducibility, data ethics and comparative analysis with recent models (e.g., Omnilingual ASR, SimbaBench and Toucan). We hope to demonstrate a practical path for under-served languages to benefit from global advances while centering local voices.

2 Related Work

Research on African languages has gained momentum in recent years. Community-driven initiatives like Masakhane have mobilized researchers to create datasets and models for numerous African languages via participatory approaches. For example, the Masakhane MT project produced translation benchmarks for several African language pairs, and MasakhaNER by (Adelani et al., 2021) provided Named Entity Recognition (NER) data for 10 languages. These efforts demonstrate the power of grassroots collaboration, a philosophy we embraced in (Bakoubolo et al., 2024) by engaging local linguists and volunteers in data collection and validation. Surveys such as NLP in Kenya like (Amol et al., 2024), and for Ethiopian languages like (Tonja et al., 2023) highlight that many African languages still face severe data scarcity and technical challenges (e.g. complex orthographies, tonality). Our work is similar in spirit, focusing on Togolese languages which belong mostly to the Niger-Congo family (with tonal systems) and have even fewer existing resources. To our knowledge, prior to (Bakoubolo et al., 2024) there was no comprehensive NLP benchmark or dataset dedicated to Togo’s languages, making our expanded Eyaa-Tom corpus a first of its kind resource.

Participatory data collection. The Masakhane project by Nekoto et al. in 2020 showed that community-led efforts can produce translation datasets and benchmarks for over thirty African languages, proving the feasibility of participatory research. We build on this philosophy by recruiting native speakers through our Data Hub and compensating them for their contributions.

Multilingual MT models. Meta’s NLLB project introduced a 200-language MT model using sparsely gated Mixture-of-Experts and human-centered evaluation (Team, 2022). UBC’s *Toucan* extends this work by fine-tuning large LMs to create a many-to-many system covering 156 African language pairs (Elmadany et al., 2024).

Toucan trains on the AfroLingu-MT benchmark and achieves strong performance on language pairs including French–Ewè and French–Kabyè. Our paper fine-tunes NLLB on eleven Togolese languages and French, providing new evaluation results and comparison to existing systems.

African Speech Benchmarks and Models.

Whisper by Radford et al. (2022) and MMS Pratap et al. (2023b) are large-scale ASR systems trained on weakly supervised data. Whisper supports roughly 100 languages and can be fine-tuned for accented speech and proper names, while MMS extends coverage to over 1,000 languages. Meta’s Omnilingual ASR pushes coverage to 1,600+ languages by scaling self-supervised pretraining to 7B parameters and incorporating community-sourced data (Keren et al., 2025). SimbaBench by Elmadany et al. (2025) provides state-of-the-art speech models and benchmarks across African languages. Such benchmarks are crucial for mapping progress; in our context, we similarly compile Lom Bench to continually assess TTS quality. For TTS, Meta’s meta-learning TTS system trains voices for over 7,000 languages (Lux et al., 2024). We leverage Whisper and Faster–Whisper for ASR and we compare to MMS and Omnilingual ASR where possible.

Another notable effort is RobotsMali’s Bambara ASR work (Diarra et al., 2025), where they collected 612 hours of spontaneous speech in Bambara and trained ultra-compact models for that oral language. Their “Hard Facts” study offers practical guidelines for field data collection and showed that fine-tuning plus human-in-loop transcription can yield substantial WER reductions. Additionally, (Tapo et al., 2025) created Bayelemabaga, a 47k-sentence French→Bambara parallel corpus, and demonstrated that augmenting existing data with curated new data can boost MT performance by +4.5 BLEU on Bambara. We take inspiration from these works. Like Bambara, many Togolese languages are oral-first and benefit from careful curation; and like Simba, we aim for broad multi-task evaluation (covering translation, ASR, TTS, etc. for each language). Our models are not yet as advanced as Simba’s best systems or Omnilingual’s 7B-parameter encoder, but our fine-tuning approach and community evaluation help narrow the gap for our specific use cases.

3 Datasets

3.1 Eyaa-Tom Expansion

The original existing corpus contains around 8 hours of speech data and 30k parallel sentences across ten Togolese languages and French (Bakoubo et al., 2024). The previous work dataset was partially released under an open-source license. (Bakoubo et al., 2025) For this work, we expand both text and speech. Working with a national television journalist, we expand on the prior work, by collecting and verifying *Kabyè*-French text in domains such as news and public communication, adding over 3000 new sentence pairs. We also curated additional domain-specific texts for Ewè, Nawdm, Mina and other languages from climate, healthcare and financial sources. Overall each language now has at least 5000 new parallel sentences. The Eyaa-Tom 26 now contains:

1. +2000 NMT translation sentence for Kabyè, Ewè, Nawdem, Lama
2. +5000 NMT translation sentences for Tem, Mina

For speech, we organized recording campaigns on our *Yodi Data Hub*. Contributors recorded prompts in their native languages and French. Across the ten languages we gathered 26,909 validated recordings totaling 30.9 hours. Figure 1 shows the distribution of recordings by language; Nawdm has the most hours (8.6h), while Ewè, Adja, Bassar, Mina, Kabyè, Lamba and Moba each contributed 2–5h.

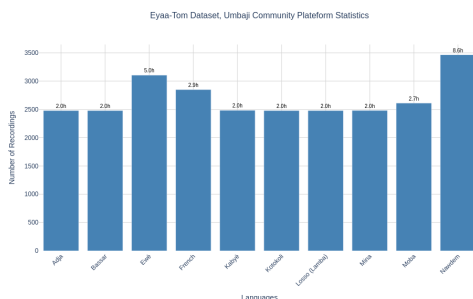


Figure 1: Number of recordings and total hours per language in the expanded Eyaa-Tom speech corpus. Bars show the number of recordings (y-axis), with hour labels above each bar (e.g., 8.6 h for Nawdm). Total recordings across all languages are 26,909 (30.9 h).

3.2 Mozilla Common Voice and Monolingual Text

To augment our acoustic training data, we imported validated recordings from Mozilla Common Voice for four Togolese languages: Adja (Foundation, 2025a), Nawdm (Foundation, 2025c), Mina (Foundation, 2025d) and Kotokoli (Foundation, 2025b). Through community campaigns organized by our team and Mozilla, these languages collectively contributed 64,580 clips (46.6 h) (Ardila et al., 2020), with an average clip length of 2.6 s. Figure 2 summarizes the distribution of hours and clips per language. This speech data provides clean acoustic coverage and complements the more varied existing corpus recordings.

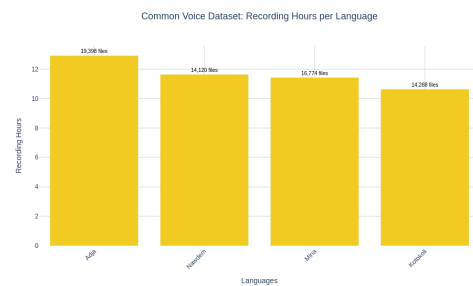


Figure 2: Common Voice contributions for Togolese languages. Each bar shows the total hours (top label) and number of validated clips collected. Adja, Nawdm, Mina and Kotokoli contribute 46.6 h (64,580 clips) in total.

3.3 Yodi Data Hub Platform

The *Yodi Data Hub* is our web platform for crowd-sourcing speech and text with fair compensation. Figure 3 illustrates the interface. Contributors log in, select their language and read or translate prompts. Each recording session shows progress and provides the option to add French translations when missing. Submitted clips go into a validation queue where other volunteers verify pronunciation and transcription. A reward system awards points for recording and validation, and contributors can request payments when a threshold is reached. A dashboard displays overall corpus statistics (e.g., 60,650 clips collected, 415 contributors from six countries) and a map of contributor distribution. Ethical participation is emphasized through a visible data policy and user agreement.

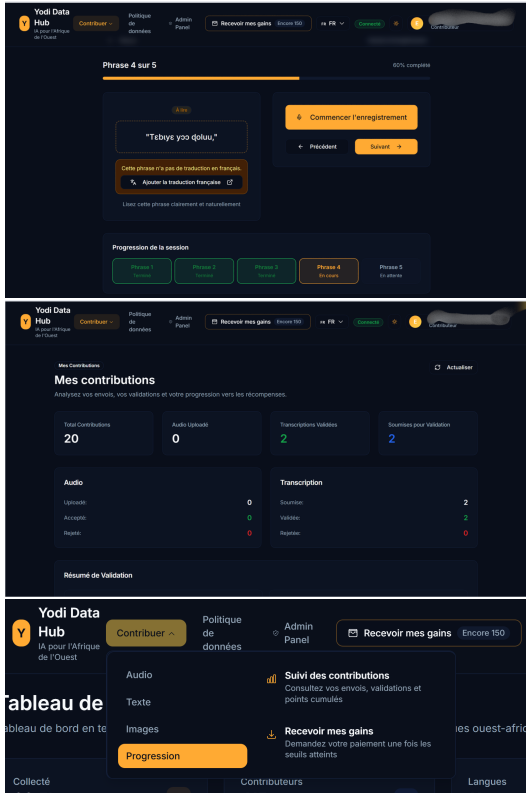


Figure 3: Screenshots of our Data Hub. Top: recording interface showing Kabyè prompt, record button and progress bar. Middle: contributor menu to track contributions and request payments. Bottom: dashboard summarizing collected clips, number of contributors, active languages and geographic distribution.

4 Methods

4.1 Automatic Speech Recognition

We selected *Whisper* by (Radford et al., 2022) as our baseline ASR because of its strong multilingual performance. We evaluated the off-the-shelf *small* model on Togolese-accented French, then fine-tuned it on a combination of 8 hours of accented French, including transcripts received from journalists and we used the same approach with the Mozilla dataset but with MMS-1b for comparison. Fine-tuning used Adam with a learning rate of $1e-5$ for 3 epochs. We also deployed the model using *Faster-Whisper*, an optimized inference engine running in near real time on CPU. Word error rate (WER), Character error rate (CER) and a proper-noun recognition test were computed.

4.2 Machine Translation

We performed a bidirectional fine-tuning with the 600M-parameter NLLB-medium model on our eleven-language parallel corpus (ten local languages plus French). Each dataset had an ini-

tial sentence pair of about 5000, which were augmented by transliteration and diacritic removal to create about 14,000 sentence pairs for the fine-tuning with an 80/20 train-test split, a learning rate of $3e-5$ run over 8 epochs. Evaluation used BLEU and METEOR on held-out test sets for each language pair direction. We compare our results with the base NLLB and with the UBC’s *Toucan* translation model (Team, 2022; Elmadany et al., 2024) where possible.

4.3 Text-to-Speech and Lom Bench

For TTS, we experimented with two toolkits: Meta’s MMS pre-trained voices and the Simba (Elmadany et al., 2024).

Lom Bench We established Lom Bench, a community evaluation platform integrated into our Data Hub 3.3. We record Mean Opinion Scores (MOS) adapted for low-resource settings and collect transcriptions for intelligibility from users based on some criteria; Quality, Naturalness, and Pronunciation. The platform currently supports Mina and Togolese lingua franca French. Evaluation is still ongoing with the aim to validate with a maximum of native speakers, with about 10 experts and community contributors who have evaluated the quality of the TTS so far.

5 Evaluation

5.1 ASR Results

Table 1 reports the WER and CER on Togolese French, and other Togolese languages test sets and the accuracy on a proper-noun list. Fine-tuning *Whisper* reduces WER from 50% to 10% on French. Proper name recognition improves from 18% to 68% after fine-tuning. *Faster-Whisper* gives identical accuracy but faster inference.

5.2 MT Results

Table 2 lists BLEU/METEOR scores for translation between French and each local language. Our fine-tuned model achieves 19.4 BLEU and 0.45 METEOR for French→Ewè and 24.8 BLEU / 0.53 METEOR for Ewè→French. Kabyè→French attains 26.1 BLEU. Scores for Adja, Tem (Kotokoli), Moba and other languages are lower (9–14 BLEU), reflecting their smaller training sets. Overall, fine-tuning yields a relative improvement of 50–60% over the base NLLB. We compare to *Toucan*’s reported scores where available (Elmadany et al.,

Model	Togolese French WER/CER (%)	Mina (GEJ) WER (%)	Adja (AJG) WER (%)	Kotokoli (KDH) WER (%)	Losso (NMZ) WER (%)
Whisper-small	$\approx 50.0 / \approx 62.56$	–	–	–	–
Whisper-small + FT	10.00 / 8.76	–	–	–	–
MMS-1b-all	–	100.0	100.0	100.0	100.0
MMS-1b-all + FT	–	70.51	83.53	94.30	95.92

Table 1: ASR performance from the provided reports. For Umbaji/Eyaa-Tom (Whisper), WER and CER are reported. For Mina/Adja/Kotokoli/Losso (MMS), only WER is reported in the report (CER not provided). Lower is better.

2024); our Ewè and Kabyè results are competitive though Toucan benefits from larger training data and model size.

6 Ethics and Data Governance

We adhered to ethical principles throughout this project. All contributors agreed to a data policy explaining the research purpose, usage rights (CC-BY) and prohibition of personal sensitive content. Recordings were validated by at least one other speaker for quality control. Contributors earn points for recording, validation and translation, redeemable for modest payments, ensuring fair compensation. The data platform provides transparent statistics and progress indicators, encouraging accountability. We strive to balance open access with community ownership, ensured by the Nwulite Obodo License (NOODL) system.

7 Conclusion

Yodi-Mantissa advances NLP resources and models for Togolese languages by combining community-driven data collection, model fine-tuning and participatory evaluation. We expanded the existing multilingual corpus, integrated Common Voice and monolingual text, and implemented a FAIR, ethical contribution platform. Our fine-tuned NLLB and Whisper models deliver improved translation and ASR performance, and our pilot TTS voices demonstrate the feasibility of Mina synthesis. The introduction of *Lom Bench* offers a new way to evaluate synthetic speech with native speakers. We compare Yodi-Mantissa to state-of-the-art multilingual systems and highlight complementarities.

8 Limitations and Future Work

The TTS component of this work is still in its preliminary stages, and hence we do not provide any details on the evaluation aspect as done with the other methods. Detailed work on the TTS approach

evaluation including the MOS will be made available in later publications.

Future work will extend coverage to the TTS approach, more languages, strengthen existing metrics with the new dataset version, provide metrics for further existing datasets, models benchmark for African languages and continue refining Lom Bench. We hope our open datasets, models and benchmark will facilitate inclusive AI development for low-resource languages and inspire similar community-centered projects worldwide.

References

- David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D’souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, and 42 others. 2021. Masakhaner: Named entity recognition for african languages. In *Transactions of the Association for Computational Linguistics*, volume 9, pages 167–184. MIT Press.
- Cynthia Jayne Amol, Everlyn Asiko Chimoto, Rose Delilah Gesicho, Antony M. Gitau, Naome A. Etori, Carington Kinyanjui, Steven Ndung’u, Lawrence Moruye, Samson Otieno Ooko, Kavengi Kitonga, Brian Muhia, Catherine Gitau, Antony Ndolo, Lilian D. A. Wanzare, Albert Njoroge Kahira, and Ronald Tombe. 2024. *State of nlp in kenya: A survey*. Preprint, arXiv:2410.09948.
- Rosana Ardila, Megan Branson, Kelly Huang, Will Wang, and 1 others. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*. Describes the Mozilla Common Voice corpus, a crowdsourced dataset of read speech across many languages.
- Justin E. Bakoubolo, Catherine Nana Nyaah Essuman, Messan Agbobli, Ahoefa Kansiwir, Kpona Sekpane Kpatika, Notou Your Timibe, Agossou, Guedela Bakouya, Bruno Koukoudjoe, and Samuel Kossi Mawouena Afolo. 2025. Eyaa-tom: A

Lang	Adja	Bassar	Ewè	Kabyè	Mina	Moba	Lamba	Konkomba	Kotokoli	Nawdm
Fr→Lang (BLEU/MET)	11.3/0.32	10.8/0.29	19.4/0.45	18.0/0.43	9.5/0.27	12.8/0.34	13.0/0.35	9.7/0.28	10.5/0.30	13.6/0.36
Lang→Fr (BLEU/MET)	17.5/0.48	16.0/0.42	24.8/0.53	26.1/0.55	13.8/0.39	18.2/0.49	19.0/0.50	14.2/0.40	15.9/0.44	20.4/0.51

Table 2: BLEU & METEOR scores for French–local language pairs. Fine-tuning NLLB on Yodi V3.1 data yields significant improvements over the base model. Ewè and Kabyè achieve the highest scores, while languages with less training data (e.g., Mina and Konkomba) remain challenging.

System	Language	Task	WER	CER	BLEU	Notes
Simba	Ewe	MT (text)	–	–	15–20	Community-driven MT corpora; BLEU varies by dataset size
MMS	Ewe	ASR/TTS	30–35%	15–20%	–	Speech-first coverage; robust phonetic modeling, no MT
Mantissa (ours)	Ewe	MT + ASR/TTS	–	–	20 - 30	Production-ready API; supports text + speech, but benchmarks unpublished
NLLB	Ewe	MT (text)	–	–	~25	Large-scale MT; mid-range BLEU for Ewe-English
Toucan (UBC)	Ewe	MT (text)	–	–	22–24	AfroLingu-MT benchmark; explicit African focus
Toucan (UBC)	Kabyè	MT (text)	–	–	~18	One of few models with Kabyè coverage; BLEU lower due to limited corpus
Mantissaa (ours)	Kabyè	MT + ASR/TTS	15 %	10 %	-	Practical deployment in Togo; no public evaluation metrics

Table 3: Detailed comparison of multilingual systems covering Togolese languages (Ewe and Kabyè). Metrics include WER (Word Error Rate), CER (Character Error Rate), and BLEU (MT quality). Simba, NLLB, and Toucan provide published BLEU scores for Ewe/Kabyè; MMS reports WER/CER for speech; YodiV3.1 supports both text and speech but remains proprietary.

multidomain parallel corpus for togolese languages. <https://huggingface.co/datasets/Umbaji/Eyaa-Tom>. GitHub/Hugging Face repository.

Justin E. Bakoubolo, Catherine Nana Nyaah Essuman, Messan Agbobli, Ahoefa Kansier, Kpona Sekpane Kpatika, Notou Your Timibe, and 1 others. 2024. YodiV3: NLP for togolese languages with Eyaa–Tom dataset and the Lom metric. In *Proceedings of the 2024 International Conference on Computational Linguistics*. Introduces the Eyaa–Tom dataset for multiple Togolese languages, baseline ASR/MT models, and the Lom metric for assessing language readiness.

Sebastien Diarra, Allahsera Auguste Tapo, Nouhoum Coulibaly, Seydou Diallo, Christopher M Homan, Mamadou K Keita, and Michael Leventhal. 2025. Gaife: Using genai to improve literacy in low-resourced settings. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 442–456. Association for Computational Linguistics.

AbdelRahim Elmadany, Ife Adebara, and Muhammad Abdul-Mageed. 2024. Toucan: Many-to-many translation for 150 african language pairs. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13189–13206. Introduces Cheeta pretraining and fine-tunes it to create Toucan, a multilingual translation model covering 156 African language pairs including Ewè and Kabyè; describes the AfroLingu-MT benchmark.

AbdelRahim Elmadany, Ife Adebara, Stéphane Clemens, Geoffrey Bate, Michael Auli, and Muhammad Abdul-Mageed. 2025. Voice of a continent: Simba – mapping africa’s speech technology frontier. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Presents SimbaBench and the Simba family of speech models for African languages, setting state-of-the-art results across ASR and TTS tasks.

Mozilla Foundation. 2025a. Common voice scripted speech 24.0 – adja (ajg). <https://datacollective.mozillafoundation.org/datasets/cmj8u3orn000hnxxbz6440x52>. Mozilla Data Collective, CC0 license.

Mozilla Foundation. 2025b. Common voice scripted speech 24.0 – kotokoli (kdh). <https://datacollective.mozillafoundation.org/datasets/cmj8u3paz00d9nxxb5gc0lfgm>. Mozilla Data Collective, CC0 license.

Mozilla Foundation. 2025c. Common voice scripted speech 24.0 – losso (nmz). <https://datacollective.mozillafoundation.org/datasets/cmj8u3pk800jlnxxbz1hvhe7u>. Mozilla Data Collective, CC0 license.

Mozilla Foundation. 2025d. Common voice scripted speech 24.0 – mina (gej). <https://datacollective.mozillafoundation.org/datasets/cmj8u3p4p0091nxxb1nbd3uzd>. Mozilla Data Collective, CC0 license.

Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, and 1 others. 2025. Omnilingual ASR: Open-source multilingual speech recognition for 1600+ languages. *arXiv preprint arXiv:2511.09690*. Introduces Omnilingual ASR, a 7B-parameter encoder-decoder ASR system that covers over 1,600 languages by combining public and community-sourced data, enabling communities to add new languages with limited data.

Hannes Lux, Daniel Eonghin, John Peters, Marcella Cornia, Antje Kruger, Sebastian Müller, and 1 others. 2024. Meta learning text-to-speech synthesis in over

- 7000 languages. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*. Presents a meta-learning approach to train TTS models across 7,000+ languages, enabling zero-shot synthesis and releasing code and data.
- Wilhelmina Nekoto, Vukosi Marivate, Carel Van Niekerk, Oluwafemi Fagbohunbe, Kabelo Mohale, René Niyongabo, Malte Onyango, and 1 others. 2020. Participatory research for low-resource machine translation: A case study of african languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Highlights the Masakhane initiative and demonstrates the feasibility of participatory MT research for African languages.
- Vineel Pratap, Jacob Kahn, Qiantong Xu, Pasquale Tomasello, Tatiana Likhomanenko, Awni Hannun, Wei-Ning Hsu, Michael Auli, Vitaly Liptchinsky, and Gabriel Synnaeve. 2023a. [Massively multilingual speech](#). *Preprint*, arXiv:2305.13516.
- Vineel Pratap, Kaushik Ram Sadagopan, Jean Mailard, Peter Liu, Michael Auli, and 1 others. 2023b. Scaling speech technology to 1,000+ languages. In *Proceedings of Interspeech 2023*. Presents the Massively Multilingual Speech project, building ASR and TTS models for over 1,000 languages using self-supervised pretraining and Bible readings.
- Alec Radford, Jong Wook Kim, Christopher Hallacy, Rachel McDowell, Aditya Ramesh, Gabriel Lawson, Jeff Clark, Chris Krueger, Alejandro Jay, Girish Prabhu, and 1 others. 2022. Robust speech recognition via large-scale weak supervision. *OpenAI Technical Report*. Introduces Whisper, a multilingual ASR model trained on 680,000 hours of weakly supervised data.
- Allahsera Auguste Tapo, Nouhoum Coulibaly, Seydou Diallo, Sebastien Diarra, Christopher M Homan, Mamadou K Keita, and Michael Leventhal. 2025. Gaife: Using genai to improve literacy in low-resourced settings. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 442–456. Association for Computational Linguistics.
- NLLB Team. 2022. No Language Left Behind: Scaling human-centered machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*. Introduces a 200-language translation model using sparsely gated mixture-of-experts and evaluates it on the FLORES benchmark.
- Atnafu Lambebo Tonja, Tadesse Destaw Belay, Israel Abebe Azime, Abinew Ali Ayele, Moges Ahmed Mehamed, Olga Kolesnikova, and Seid Muhie Yimam. 2023. [Natural language processing in ethiopian languages: Current state, challenges, and opportunities](#). *Preprint*, arXiv:2303.14406.

Using Subword-Embeddings for Bilingual Lexicon Induction in Bantu Languages

Adrian Breiding

adrian.johannes.breiding.1@hu-berlin.de

Alan Akbik

alan.akbik@hu-berlin.de

Abstract

Bilingual Lexicon Induction (BLI) is a valuable tool in machine translation and cross-lingual transfer learning, but it remains challenging for agglutinative and low-resource languages. In this work, we investigate the use of weighted sub-word embeddings in BLI for agglutinative languages. We further evaluate a graph-matching and Procrustes-based BLI approach on two Bantu languages, assessing its effectiveness in a previously underexplored language family. Our results for Swahili with an average P@1 score of 51.84% for a 3000 word dictionary demonstrate the success of the approach for Bantu languages. Weighted sub-word embeddings perform competitively on Swahili and outperform word embeddings in our experiments with Zulu.

1 Introduction

Bilingual Lexicon Induction (BLI) is the task of automatically generating translation pairs from monolingual corpora, typically using a small seed dictionary to align two separate semantic spaces. For an example of a generated output, see Figure 1. These lexicons serve as a critical bridge for many NLP tasks, such as providing translations for out-of-vocabulary words in Machine Translation (MT) (Irvine and Callison-Burch, 2017) and facilitating cross-lingual transfer learning for under-represented languages (Wang et al., 2022).

Despite its potential, current BLI research is constrained by a significant "resource and relatedness" bias. Most work focuses on high-resource, closely related language pairs where the underlying corpora are drawn from similar domains. Furthermore, evaluations claiming to address low-resource settings often do so by down-sampling massive datasets (Marchisio et al., 2022), rather than confronting the noise and sparsity of authentic low-resource environments. While recent efforts have begun to bridge this gap (Nakashole, 2019;

kununua	v.	buy
cheti	N.	certificate
vyeti	N.PL.	certificates
mjadala	N.	discussion
mijadala	N.PL.	dialogue (<i>discussions</i>)
takriban	ADV.	nearly
kusini	N.	south
		⋮

p. 42

Figure 1: Excerpt from a generated Swahili–English bilingual dictionary.

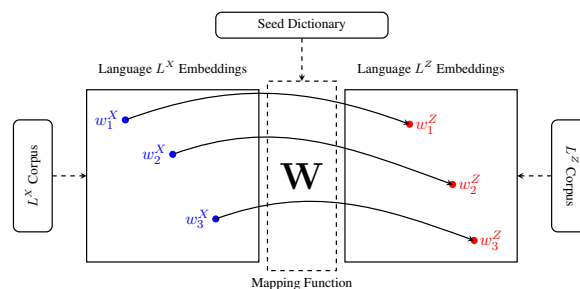


Figure 2: Schematic visualization of bilingual lexicon induction using embeddings

Bhowmik and Ralescu, 2023), African languages, specifically the Bantu language family, remain almost absent from the literature.

A more fundamental limitation, however, lies in the architectural assumptions of existing BLI frameworks. Nearly all current approaches are built on word-level embeddings, which treat words as indivisible, atomic units. This paradigm assumes a rough one-to-one correspondence between words in different languages. Additionally, approaches often assume the existence of a linear mapping between the embedding spaces of the two languages, a premise known as the isomorphism hypothesis (Søgaard et al., 2018). See Figure 2 for an example. While this may hold for morphologically simple or related languages (e.g., English and Spanish), it breaks down when applied to the agglutinative

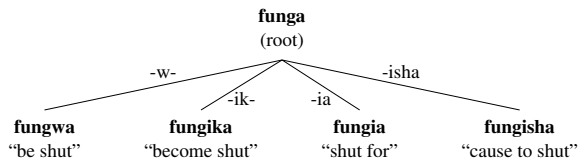


Figure 3: Nuances of In- and Suffixes in Swahili (Encyclopaedia Britannica Editors, 2025)

morphology of Bantu languages.

Challenges of Agglutinative Morphology. In Bantu languages, a single verb root can generate several unique surface forms through the addition of noun-class prefixes, tense markers, and derivational suffixes (see Figure 3). When BLI is performed at the word level, this leads to two critical failures:

- **Vocabulary Fragmentation:** The semantic meaning of a single root is scattered across dozens of unique "atoms" (e.g., *ninauza*, *utauza*, *walituuza*), each appearing with low frequency. This prevents the model from learning a cohesive representation for the core concept ("to sell"). See Figure 1 for an example.
- **Structural Misalignment:** A single complex word in Swahili often encodes the semantic equivalent of an entire English phrase. This creates a geometric divergence between embedding spaces, making linear mapping techniques like the orthogonal Procrustes approach mathematically insufficient.

In this paper, we address these challenges by investigating BLI for two representative Bantu languages: Swahili and Zulu. We leverage the Fundus library (Dallabetta et al., 2024) to build specialized newspaper-based corpora and generate seed dictionaries via Google Translate. To move beyond the limitations of word-level models, we propose and evaluate the use of sub-word level embeddings.

We further test a hybrid approach that combines graph-matching and Procrustes-based mapping (Marchisio et al., 2022), providing, to our knowledge, the first rigorous evaluation of this system on Bantu-English pairs. Finally, we analyze the degree of isomorphy between these spaces to quantify the difficulty of the Bantu-English BLI task.

Contributions. Our main contributions are:

- Identifying and demonstrating the failure modes of word-level BLI when applied to the agglutinative structures of the Bantu language family.
- Validating that sub-word embeddings mitigate these failures by capturing morphological nuances and improving alignment with English.
- Providing a benchmark for BLI on two low-resource African languages using authentic, non-simulated datasets.

2 Related Work

Harris (1954) theorizes that words in similar contexts often have similar meanings. Following this intuition, Mikolov et al. (2013b) presented a method, where a single-layer feed-forward neural network is used to predict a word given its context, and the embeddings can be extracted from the weights of the hidden layer. In their experiments, they found that these embeddings capture some semantic information by computing "Madrid" - "Spain" + "France" and extracting the nearest neighbor, which yields "Paris".

Experiments with Korean and Swahili, both (highly) agglutinative languages, show a higher performance when using syllable-aware embeddings (Shikali et al., 2019; Choi et al., 2017). In both settings, the researchers generated trained syllable vectors in combination with a convolutional neural network. Since these approaches require more resources, we focused on an alternative: Byte-Pair embeddings, which have the advantage of requiring a lower amount of resources compared to other sub-word unit embedding approaches (Sennrich et al., 2016; Heinzerling and Strube, 2018). It is based on Byte-Pair encoding (Gage, 1994), where a new symbol iteratively replaces the most common symbol pair. After encoding, the symbols should represent the most common character strings. Training embeddings on these symbols will result in them reflecting the most frequent subwords in the hope of corresponding to semantically significant morphemes. By varying the number of merge operations, the resulting embeddings can be manipulated to represent shorter character sequences or many frequently occurring words primarily.

2.1 The Procrustes Approach

Mikolov et al. (2013a) presented the original idea that later led to the Procrustes method after the

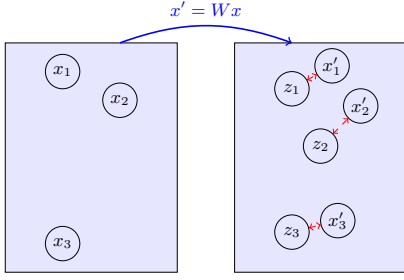


Figure 4: Visualization of the Procrustes approach. Adapted from Marchisio et al. (2021)

discovery that embeddings can capture linguistic regularities. Hence, postulating a mapping from the source embedding space to the target embedding space seemed reasonable. Concretely, given a set of word pairs and their corresponding embeddings $\{x_i, z_i\}_{i=0}^{n-1}$ with $x_i, z_i \in \mathbb{R}^d$, the goal is to find a matrix $W \in \mathbb{R}^{d \times d}$, such that Wx_i is a good approximation of z_i . This can be represented as the following optimization problem:

$$\arg \min_{W \in \mathbb{R}^{d \times d}} \sum_{i=0}^{n-1} \|Wx_i - z_i\|_2^2 \quad (1)$$

Restricting $W \in O(d)$ to be orthogonal and replacing the Euclidean norm with the Frobenius norm yields the orthogonal Procrustes problem, which is exactly solvable (Conneau et al., 2018). Having found a mapping between the embedding spaces, we can now use it to predict translations. In the original paper, this was done by extracting the nearest neighbors using the cosine similarity metric. In higher-dimensional spaces, it can be observed that so-called *hubs* and *anti-hubs* are formed. These hubs are the nearest neighbors for many other points with a high probability, whereas anti-hubs are not the nearest neighbors for any other point (Radovanović et al., 2010). Naturally, this causes issues when extracting potential translation candidates using a standard nearest-neighbor measure. To produce a more reliable matching, Conneau et al. (2018) introduce the Cross-Domain Similarity Local Scaling Measure, which intuitively combats hubness by penalizing points with dense neighborhoods while boosting the similarity values for anti-hubs.

The underlying assumption is that the embedding spaces are “isomorphic”, which is understood to be “geometrically similar” in this context. While it is reasonable for related languages, experiments have shown that it is less sensible in cases where

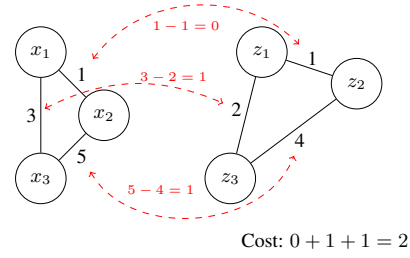


Figure 5: Visualization of the graph-based approach. Adapted from Marchisio et al. (2021)

the language similarity decreases or in cases of varying domains of the monolingual training data sets (Søgaard et al., 2018; Patra et al., 2019).

2.2 The Graph-Based Approach

To attempt BLI using graph matching (Marchisio et al., 2022), one can consider words as nodes in monolingual, weighted, undirected graphs $G_X = (V_X, E_X, w_X)$, $G_Z = (V_Z, E_Z, w_Z)$. The edge weights are computed using the cosine similarity metric. The fundamental idea is that relationships between words are reflected in the similarity of the corresponding embeddings, and these similarities remain relatively consistent across different languages. With this assumption, one can try to find the optimal permutation π aligning $\pi(V_Z)$ with V_X , minimizing the edge disagreement.

To quantify the edge disagreement of the two graphs, it is intuitive to choose a metric similar to:

$$\|A_X - PA_ZP^T\|_F = \sqrt{\sum_{i=1}^n \sum_{j=1}^n [(A_X)_{ij} - (PA_ZP^T)_{ij}]^2} \quad (2)$$

also known as the Frobenius norm. The problem is NP-hard, but functional approximations are available. For one such approximation, the optimization of equation 2 can be attempted by using an instance of the Optimal Transport (OT) Problem

$$\arg \min_{P \in U(r, c)} \text{tr}(P^T M) \quad (3)$$

with $U(r, c)$ the transportation polytope and a cost matrix $M \in \mathbb{R}^{n \times n}$. A well-performing algorithm for solving the OT problem in graph matching, also in large, non-isometric cases, is the *Graph Matching via Optimal Transport* (GOAT) algorithm (Saad-Eldin et al., 2021).

The performance of this approach and the translation matrix method increases with the number of translation pairs, so-called seeds, that are passed

into the algorithm. Experiments with various languages and varying numbers of seeds show that graph matching outperforms the nearest-neighbor method in most cases. The performance is further robust on dissimilar languages and low supervision (Marchisio et al., 2022).

2.3 Prior Work on Bantu Languages

Nakashole (2019) attempts BLI for Bantu languages by exploiting grammatical similarities. For one Bantu language L_1 with a small bilingual dictionary, they modify the corpus of a second Bantu language L_2 and merge the corpus of L_1 with the modified corpus of L_2 . They then try to find a projection matrix that maps the English embeddings to the combined Bantu word embeddings as introduced by Mikolov et al. (2013a), while only requiring the seed dictionary for L_1 . This approach scores 0.30, 0.56 and 0.58 in precision at top- k ($P@k$) for $k \in \{1, 5, 10\}$ respectively for L_1 and 0.10, 0.18 and 0.20 for L_2 .

2.4 Reliability of Google Translate

There is little work available on the current reliability of Google Translate for English-Swahili and English-Zulu translations. For Swahili, Okafor (2025) has reported significant improvement potential for Google Translate for text translation in the medical domain. Concretely, AfromT (Iyamu, 2024), a domain-specific translation framework for African languages, improves performance on scientific and medical texts by $\sim 19\%$ relative to Google Translate. However, AfromT remains $\sim 25\%$ below the performance of comparable models trained on high-resource languages. A qualitative study by Sangili (2024), based on a randomly selected sample, reports “excellent” performance on individual lexemes but notes a marked decline in translation quality for multiword phrases and figurative usages; the author does not report a formal sample size for that claim.

For Zulu, Khoboko et al. (2025) have assessed Google Translate using the BLEU (Papineni et al., 2002), G-Eval (Liu et al., 2023) and ChFr++ (Yu et al., 2021) metrics. They report scores of 5.93, 55.12 and 95.6% respectively. The low score for BLEU, an n-gram overlap metric, indicates a low surface-level agreement with reference translations. G-Eval and ChFr++ attempt to capture meaning preservation on different scales, suggesting that Google Translate often preserves semantics.

Taken together, these findings suggest that

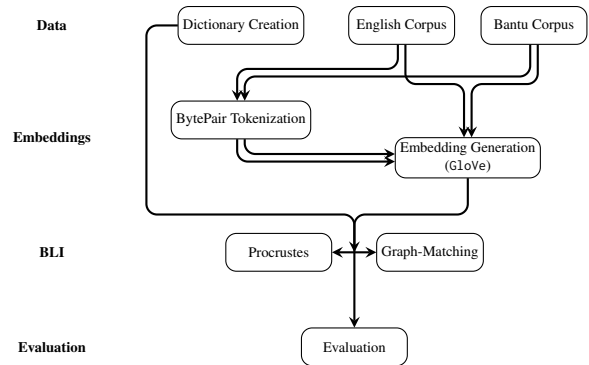


Figure 6: Overview of the four elemental steps in creating a dictionary using BLI.

Google Translate is sufficient for generating the limited set of common-word seeds required in our experiments.

3 Methodology

The process is split into four steps: data attainment, embedding training, BLI, and evaluation. Figure 6 provides an overview.

3.1 Data Acquisition

One of the main criteria for a possible data source in the selection process was its bilingual availability, since using two corpora from varying domains could drastically reduce the performance (Søgaard et al., 2018). Newspapers from a bilingual publisher may satisfy this constraint. Using the Fundus library (Dallabetta et al., 2024) we crawled [Daily News](https://dailynews.co.tz)¹ and [Habari Leo](https://habarileo.co.tz)² for Swahili - English and [Eyethu News](https://eyethunews.co.za)³ (only Zulu articles), [Ilanga News](https://ilanganews.co.za)⁴ and [The Citizen](https://citizen.co.za)⁵ for Zulu - English. The plain-texts of the articles are combined into a single file, converted to lower-case, stripped of non-letter characters, and all numbers replaced with zeroes.

This process generates a Swahili - English corpus with 194 605 and 589 460 sentences respectively and a Zulu - English corpus with 146 967 each. We also translate the Zulu corpus with Google Translate, resulting in the “translated” corpus.

Both the Procrustes and the graph-matching algorithm require seed translations as input. The 10 000 most frequent words in each corpus are compiled as a list, translated using Google Translate and

¹<https://dailynews.co.tz>

²<https://habarileo.co.tz>

³<https://eyethunews.co.za>

⁴<https://ilanganews.co.za>

⁵<https://citizen.co.za>

combined into a unidirectional dictionary for each language pair. This dictionary can be utilized as input seeds and for automatic evaluation.

In principle, reliance on Google Translate does not limit the applicability of this approach to additional low-resource languages, as it only requires a small number of seeds for competitive results. These amounts can be easily obtained from a single native speaker or a small existing dictionary.

3.2 Embeddings

A natural choice for a type of embeddings are subword embeddings, for intuitively, they have the potential to better reflect relationships between words like *andika* (write) and *andikia* (write **to**) (Mpiranya, 2023). For this work, we use Byte Pair embeddings (Sennrich et al., 2016), which are trained using the training scripts provided in the GitHub repository⁶ corresponding to the publication (Heinzerling and Strube, 2018). In the first step, the corpus is tokenized and encoded using SentencePiece (Kudo and Richardson, 2018). The embeddings are then trained on the encoded corpora using GloVe (Pennington et al., 2014). In the final step, the embeddings must be re-arranged in the same order as in the BPE vocabulary file generated in the tokenization phase.

Both of the approaches require word embeddings. A given word is encoded using the learned BPE model. The corresponding embedding is computed by calculating the (weighted) arithmetic mean of the subword embeddings. Given a word w^X with a tokenization (t_1, \dots, t_n) . Let τ_i be the embedding corresponding to the token t_i , then the mean word embeddings are computed by:

$$x = \frac{1}{n} \sum_{i=1}^n \tau_i \quad (4)$$

The weighted mean embeddings are calculated using the token length $|t_i|$ as weights.

$$x = \sum_{i=1}^n \frac{|t_i|}{\sum_{j=1}^n |t_j|} \tau_i \quad (5)$$

3.3 Bilingual Lexicon Induction

Generally, Marchisio et al. (2022) have shown that the best results are achieved using a combination of Procrustes and graph matching. Choosing the correct algorithm to start with (heavily) depends

⁶<https://github.com/bheinzerling/bpemb>

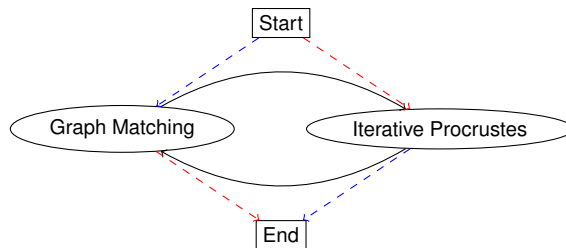


Figure 7: Overview of the system combination experimental setup. Adapted from Marchisio et al. (2022)

on the number of seeds used. Given the moderate amount of seeds available, starting with graph matching has proven to be the better choice, which is why the following introduction will follow this order.

First, graph matching is run in the forward and reverse directions. For each direction a hypothesized mapping $h_i : \{1, n\} \rightarrow \{1, n\}, i \in \{f, b\}$ is returned. The indices f, b represent forward and backward, respectively. The hypothesis sets for each direction can then be defined as $H_f = \{(a, b) | h_f(a) = b\}$ and $H_b = \{(a, b) | h_b(b) = a\}$. The intersection $H = H_f \cap H_b$ is used as an input for Procrustes.

Similarly, the Procrustes problem is solved in both directions using the gold seeds and the hypotheses from the graph-matching approach. The hypotheses are extracted using 1-nearest neighbors. Instead of the commonly used cosine metric, we will use Cross-Domain Similarity Local Scaling (Conneau et al., 2018) because it is more resistant against hubness. Multiple iterations of Procrustes are run in total, with the intersections of forward and backward hypotheses from the previous round used as additional seeds. After five iterations, the hypothesis intersection is passed into the graph-matching algorithm.

After 20 iterations, H_f is returned as the final hypothesis and can be used to generate the dictionary.

4 Results

The experiment’s goal was to create a dictionary for the 3000 most frequent words in the English and Bantu corpora, assuming that occurrences of a word and its translation are of a similar magnitude. Most parameters relevant to the graph matching or Procrustes processes were used as in the original paper. To optimize the final dictionary, we performed a parameter search over several parameters: (1) Merge operations in the training of

Experiment	Num. Seeds
Swahili	866
Zulu	929
Zulu (t)	1030

Table 1: Number of available seeds per experiment. Zulu (t) indicates Zulu with translated corpus.

the Byte Pair Embeddings (10 000, 20 000, 50 000), (2) Computation of word embeddings using the mean or weighted average, (3) Number of seeds (50, 75, 100), and (4) Ending with Procrustes or graph matching.

The experiment was run 10 times for each combination of parameters on a randomized seed input. The pseudorandom number generator was seeded with the same value for each iteration across all experiments. The evaluation was performed using the unused seeds. An overview of the total available seeds for testing and supervision is available in Table 1. The implementations of Marchisio et al. (2022) served as a foundation for our experiments. The code can be found on [GitHub](#)⁷.

4.1 BLI for Bantu Languages

Tables 2 and 3 contains the parameter combinations yielding the highest averaged score for each type of embedding and experiment with a small corpus. One observes that in our number of seeds regime, the best results are achieved with the highest number of seeds and when ending with Procrustes. A finding that coincides with the results of Marchisio et al. (2022).

The optimal vocabulary size appears to be dependent on the language, with 50 000 for Swahili and 20 000 for Zulu. Similarly, the optimal embedding types are language dependent. While the weighted outperform the mean Byte Pair embeddings, the basic word embeddings still provide better results for Swahili. Finally, the performance for Zulu is significantly improved by switching the crawled corpus with the translated corpus.

4.2 Isomorphism of Embedding Spaces

Another factor that helps understand the differences in performance is the degree of isomorphism of the two embedding spaces (Marchisio et al., 2022). Using the implementations by Vulić et al. (2020), we calculate the Gromov-Hausdorff (GH) Distance

and Laplacian Eigenvector Similarities (EVS) for selected embedding spaces. Similar to the process of BLI, the similarity measures are computed using the embeddings of the 3000 most frequent words from each language, using the best performing parameter combination for each experiment (see Tables 2 and 3).

Experiment	GH Distance	Eigenvector Sim.
Swahili	0.06	12.94
Zulu	0.13	9.04
Zulu (t)	0.18	16.49

Table 4: Eigenvector similarity and Gromov-Hausdorff (GH) distances for selected language pairs

5 Discussion

Averaging 51.84% for the best parameter combination, this approach is highly successful for Swahili, yielding better results than some results for distant language pairs in the original paper by Marchisio et al. (2022). Yet, our scores are not directly comparable with theirs because of the larger training corpus and larger created dictionary size in the original experiments. Nevertheless, they do provide an adequate point of reference and demonstrate a successful application of this approach for an African language.

The performance is significantly worse for Zulu, averaging 2.37% for the best parameter combination with the crawled corpus. The precision can be improved by replacing the crawled English corpus with the translated English corpus, averaging 23.35%, showing that BLI can also be successfully applied to Zulu. Though it remains unclear, why the performances vary so strongly. An indicator is the low GH and EVS distances for Swahili, which correlate with a stronger performance in BLI (Marchisio et al., 2022). Consequently, one would expect the crawled corpus for Zulu to perform better than the translated corpus, which contradicts our observations. This discrepancy suggests that GH and EVS distances alone are insufficient to explain the observed performance differences, and that additional factors, such as corpus quality or domain mismatch, may also be relevant.

For the two Zulu experiments an additional factor may be a sensitivity to varying domains. The Swahili corpora were crawled from a bilingual publisher, mostly publishing similar articles in Swahili on Habari Leo and in English on Daily News. In

⁷<https://github.com/addie9800/bantu-bli>

Emb. Type	Vocab.	Score (Avg.)
Mean	50 000	44.03 \pm 1.21
Weighted	50 000	49.46 \pm 0.76
Word	-	51.84 \pm 1.01

Table 2: BLI for Swahili, showing the best averaged score per experiment and embedding type. All experiments achieved best results with 100 seeds and ending with Procrustes.

absence of such a publisher for Zulu, we resorted to crawling from related publishers, which likely don’t show such an overlap. The subsequent translation then effectively generated a parallel corpus, boosting the performance. It remains unclear, why also the translated corpus for Zulu performs significantly worse than Swahili.

It further appears to be beneficial to use subword embeddings for Bantu languages. In the case of Zulu, it outperforms regular word embeddings while achieving competitive results in Swahili. Additionally, the experiments appear to confirm the intuition that longer tokens should be given more weight, as the weighted embeddings consistently score higher than the mean embeddings.

Limitations

The experiments were only performed on two representatives of the Bantu language family, which are comparatively high-resourced, as they have online newspapers available and are additionally also supported by Google Translate. Repeating the experiments with more languages may provide insights into causes of the varying performance. Additionally, the experiments were performed using 3000 embeddings, whereas Marchisio et al. (2022) work with 200 000, and it remains open how well this approach scales for larger systems in the Bantu setting.

Previous work suggests (see section 2.4) that translation quality is sufficient, especially for individual words, yet no quantitative analysis of the quality of the generated seeds was performed, which would provide greater confidence in the results.

Finally, the strong results for Zulu were obtained using an English corpus generated by translating the Zulu corpus, effectively creating a parallel cor-

Corpus	Emb. Type	Vocab.	Score (Avg.)
Crawled	Mean	20 000	2.28 \pm 0.46
Crawled	Weighted	20 000	2.37 \pm 0.46
Crawled	Word	-	1.75 \pm 0.47
Translated	Mean	20 000	18.26 \pm 1.40
Translated	Weighted	20 000	23.35 \pm 0.63
Translated	Word	-	13.89 \pm 0.99

Table 3: BLI for Zulu, showing the best averaged score per experiment and embedding type. All experiments achieved best results with 100 seeds and ending with Procrustes.

pus, which is not realistic for an arbitrary low-resource language.

References

- Kowshik Bhowmik and Anca Ralescu. 2023. Bridging the resource gap in cross-lingual embedding space. In *Modelling and Development of Intelligent Systems*, pages 122–135, Cham. Springer Nature Switzerland.
- Sanghyuk Choi, Taek Kim, Jinseok Seol, and Sang-goo Lee. 2017. [A syllable-based technique for word embeddings of Korean words](#). In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 36–40, Copenhagen, Denmark. Association for Computational Linguistics.
- Alexis Conneau, Guillaume Lample, Marc’ Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. [Word translation without parallel data](#). *Preprint*, arXiv:1710.04087.
- Max Dallabetta, Conrad Dobberstein, Adrian Breiding, and Alan Akbik. 2024. [Fundus: A simple-to-use news scraper optimized for high quality extractions](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 305–314, Bangkok, Thailand. Association for Computational Linguistics.
- Encyclopaedia Britannica Editors. 2025. [Swahili language](#). Accessed 19 December 2025.
- Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.
- Zellig S. Harris. 1954. [Distributional structure](#). *WORD*, 10(2-3):146–162.
- Benjamin Heinzerling and Michael Strube. 2018. [BPEmb: Tokenization-free pre-trained subword embeddings in 275 languages](#). In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

- Ann Irvine and Chris Callison-Burch. 2017. [A comprehensive analysis of bilingual lexicon induction](#). *Computational Linguistics*, 43(2):273–310.
- Raphael Iyamu. 2024. [Machine translation and nlp tools: Developing and refining language technologies for african languages](#). *International Journal For Multidisciplinary Research*.
- Pitso Walter Khoboko, Vukosi Marivate, and Joseph Sefara. 2025. [Optimizing translation for low-resource languages: Efficient fine-tuning with custom prompt engineering in large language models](#). *Machine Learning with Applications*, 20:100649.
- Taku Kudo and John Richardson. 2018. [SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Kelly Marchisio, Youngser Park, Ali Saad-Eldin, Anton Alyakin, Kevin Duh, Carey Priebe, and Philipp Koehn. 2021. [An analysis of Euclidean vs. graph-based framing for bilingual lexicon induction from word embedding spaces](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 738–749, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Kelly Marchisio, Ali Saad-Eldin, Kevin Duh, Carey Priebe, and Philipp Koehn. 2022. [Bilingual lexicon induction for low-resource languages using graph matching via optimal transport](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2545–2561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tomas Mikolov, Quoc V. Le, and Ilya Sutskever. 2013a. [Exploiting similarities among languages for machine translation](#). *Preprint*, arXiv:1309.4168.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. [Distributed representations of words and phrases and their compositionality](#). *Preprint*, arXiv:1310.4546.
- Ferdinand Mpiranya. 2023. *English-Swahili Swahili-English Immersive Dictionary*, 1 edition. Routledge.
- Ndapa Nakashole. 2019. [Bilingual dictionary induction for bantu languages](#). *Preprint*, arXiv:1811.07080.
- Ugochi Okafor. 2025. [Multilingual NLP for African healthcare: Bias, translation, and explainability challenges](#). In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 221–229, Vienna, Austria. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R Gormley, and Graham Neubig. 2019. [BLISS in non-isometric embedding spaces](#).
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [Glove: Global vectors for word representation](#). In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović. 2010. [Hubs in space: Popular nearest neighbors in high-dimensional data](#). *Journal of Machine Learning Research*, 11(86):2487–2531.
- Ali Saad-Eldin, Benjamin D. Pedigo, Carey E. Priebe, and Joshua T. Vogelstein. 2021. [Graph matching via optimal transport](#). *Preprint*, arXiv:2111.05366.
- Nabeta K. N. Sangili. 2024. [Digitising kiswahili for translation economy](#). *Journal of Kiswahili and Other African Languages*, 2(2):44–51.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Neural machine translation of rare words with subword units](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.
- Casper S. Shikali, Zhou Sijie, Liu Qihe, and Refuoe Mokhosi. 2019. [Better word representation vectors using syllabic alphabet: A case study of swahili](#). *Applied Sciences*, 9(18).
- Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. [On the limitations of unsupervised bilingual dictionary induction](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.
- Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. [Are all good word vector spaces isomorphic?](#) In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. [Expanding pretrained models to thousands more languages via lexicon-based adaptation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. 2021. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1077.

A Parameter Search Results

Tables 5, 6 and 7 show the results of the hyperparameter search across all parameters: embedding type, vocabulary size, number of seeds, whether or not to end with Procrustes (P-End) for Swahili, Zulu and Zulu with the translated corpus.

Emb. Type	#Seeds	Vocab.	P-End?	Score (Avg.)
Mean	50	10 000	True	41.04 ± 1.30
Mean	50	10 000	False	29.13 ± 2.54
Mean	75	10 000	True	41.73 ± 1.24
Mean	75	10 000	False	35.23 ± 1.84
Mean	100	10 000	True	43.07 ± 0.85
Mean	100	10 000	False	37.23 ± 0.92
Mean	50	20 000	True	42.33 ± 0.90
Mean	50	20 000	False	30.04 ± 4.14
Mean	75	20 000	True	43.12 ± 0.86
Mean	75	20 000	False	36.11 ± 1.61
Mean	100	20 000	True	43.02 ± 0.44
Mean	100	20 000	False	38.17 ± 1.45
Mean	50	50 000	True	43.53 ± 0.77
Mean	50	50 000	False	23.89 ± 4.36
Mean	75	50 000	True	43.35 ± 1.08
Mean	75	50 000	False	34.86 ± 1.92
Mean	100	50 000	True	44.03 ± 1.21
Mean	100	50 000	False	37.48 ± 1.18
Weighted	50	10 000	True	46.81 ± 1.36
Weighted	50	10 000	False	37.85 ± 2.35
Weighted	75	10 000	True	47.82 ± 0.79
Weighted	75	10 000	False	41.69 ± 0.73
Weighted	100	10 000	True	48.40 ± 0.81
Weighted	100	10 000	False	43.20 ± 1.20
Weighted	50	20 000	True	47.62 ± 0.95
Weighted	50	20 000	False	38.47 ± 1.13
Weighted	75	20 000	True	48.01 ± 0.41
Weighted	75	20 000	False	41.61 ± 1.29
Weighted	100	20 000	True	49.24 ± 1.15
Weighted	100	20 000	False	43.78 ± 0.78
Weighted	50	50 000	True	47.64 ± 0.72
Weighted	50	50 000	False	27.57 ± 8.96
Weighted	75	50 000	True	49.34 ± 0.93
Weighted	75	50 000	False	40.84 ± 1.01
Weighted	100	50 000	True	49.46 ± 0.76
Weighted	100	50 000	False	42.64 ± 1.22
Word	50	-	True	50.90 ± 0.98
Word	50	-	False	37.30 ± 2.36
Word	75	-	True	51.39 ± 1.05
Word	75	-	False	43.61 ± 1.20
Word	100	-	True	51.84 ± 1.01
Word	100	-	False	46.63 ± 1.46

Table 5: Parameter search results using the Swahili Corpus

Emb. Type	#Seeds	Vocab.	P-End?	Score (Avg.)
Mean	50	10 000	True	0.99 ± 0.49
Mean	50	10 000	False	0.54 ± 0.44
Mean	75	10 000	True	1.28 ± 0.45
Mean	75	10 000	False	0.80 ± 0.47
Mean	100	10 000	True	2.02 ± 0.31
Mean	100	10 000	False	0.94 ± 0.38
Mean	50	20 000	True	1.09 ± 0.72
Mean	50	20 000	False	0.48 ± 0.21
Mean	75	20 000	True	1.51 ± 0.48
Mean	75	20 000	False	0.73 ± 0.29
Mean	100	20 000	True	2.28 ± 0.46
Mean	100	20 000	False	0.69 ± 0.36
Mean	50	50 000	True	1.08 ± 0.33
Mean	50	50 000	False	0.31 ± 0.26
Mean	75	50 000	True	1.60 ± 0.58
Mean	75	50 000	False	0.44 ± 0.17
Mean	100	50 000	True	1.42 ± 0.33
Mean	100	50 000	False	0.88 ± 0.40
Weighted	50	10 000	True	1.40 ± 0.41
Weighted	50	10 000	False	0.45 ± 0.16
Weighted	75	10 000	True	1.98 ± 0.49
Weighted	75	10 000	False	0.79 ± 0.25
Weighted	100	10 000	True	2.31 ± 0.46
Weighted	100	10 000	False	1.05 ± 0.43
Weighted	50	20 000	True	1.16 ± 0.54
Weighted	50	20 000	False	0.39 ± 0.29
Weighted	75	20 000	True	1.60 ± 0.50
Weighted	75	20 000	False	0.65 ± 0.32
Weighted	100	20 000	True	2.37 ± 0.46
Weighted	100	20 000	False	0.84 ± 0.23
Weighted	50	50 000	True	1.11 ± 0.33
Weighted	50	50 000	False	0.27 ± 0.19
Weighted	75	50 000	True	1.47 ± 0.36
Weighted	75	50 000	False	0.55 ± 0.20
Weighted	100	50 000	True	1.82 ± 0.55
Weighted	100	50 000	False	0.68 ± 0.34
Word	50	-	True	0.89 ± 0.47
Word	50	-	False	0.30 ± 0.15
Word	75	-	True	1.36 ± 0.52
Word	75	-	False	0.41 ± 0.26
Word	100	-	True	1.75 ± 0.47
Word	100	-	False	0.74 ± 0.38

Table 6: Parameter search results using the crawled Zulu Corpus

Emb. Type	#Seeds	Vocab.	P-End?	Score (Avg.)
Mean	50	10 000	True	13.13 ± 1.49
Mean	50	10 000	False	8.76 ± 2.59
Mean	75	10 000	True	14.72 ± 0.72
Mean	75	10 000	False	11.49 ± 1.81
Mean	100	10 000	True	15.89 ± 1.12
Mean	100	10 000	False	11.61 ± 1.71
Mean	50	20 000	True	14.89 ± 0.87
Mean	50	20 000	False	10.30 ± 2.51
Mean	75	20 000	True	15.55 ± 1.45
Mean	75	20 000	False	12.39 ± 1.30
Mean	100	20 000	True	18.26 ± 1.40
Mean	100	20 000	False	14.54 ± 1.39
Mean	50	50 000	True	9.62 ± 0.91
Mean	50	50 000	False	5.82 ± 1.90
Mean	75	50 000	True	13.84 ± 0.75
Mean	75	50 000	False	7.99 ± 1.90
Mean	100	50 000	True	15.58 ± 1.10
Mean	100	50 000	False	9.24 ± 1.25
Weighted	50	10 000	True	21.08 ± 0.77
Weighted	50	10 000	False	19.94 ± 1.34
Weighted	75	10 000	True	22.77 ± 0.87
Weighted	75	10 000	False	20.75 ± 1.70
Weighted	100	10 000	True	23.01 ± 1.06
Weighted	100	10 000	False	20.29 ± 1.40
Weighted	50	20 000	True	21.64 ± 1.33
Weighted	50	20 000	False	19.13 ± 2.89
Weighted	75	20 000	True	21.27 ± 2.29
Weighted	75	20 000	False	20.44 ± 1.35
Weighted	100	20 000	True	23.35 ± 0.63
Weighted	100	20 000	False	20.52 ± 1.88
Weighted	50	50 000	True	16.70 ± 2.42
Weighted	50	50 000	False	15.92 ± 2.35
Weighted	75	50 000	True	19.67 ± 1.92
Weighted	75	50 000	False	17.26 ± 3.66
Weighted	100	50 000	True	20.41 ± 2.21
Weighted	100	50 000	False	16.34 ± 3.22
Word	50	-	True	7.55 ± 1.70
Word	50	-	False	1.35 ± 0.48
Word	75	-	True	11.52 ± 1.69
Word	75	-	False	2.86 ± 0.88
Word	100	-	True	13.89 ± 0.99
Word	100	-	False	3.98 ± 0.77

Table 7: Parameter search results using the translated Zulu Corpus

AfriNLLB: Efficient Translation Models for African Languages

Yasmin Moslem*

ADAPT Centre
Trinity College Dublin
Dublin, Ireland
yasmin.moslem@adaptcentre.ie

Aman Kassahun Wassie*

African Institute for
Mathematical Sciences (AIMS)
Addis Ababa, Ethiopia
awassie@aimsammi.org

Amanuel Gizachew Abebe*

Shaggar Institute of
Technology (SIT)
Shaggar city, Ethiopia
amanuel.g.abebe1@gmail.com

Abstract

In this work, we present AfriNLLB, a series of lightweight models for efficient translation from and into African languages. AfriNLLB supports 15 language pairs (30 translation directions), including Swahili, Hausa, Yoruba, Amharic, Somali, Zulu, Lingala, Afrikaans, Wolof, and Egyptian Arabic, as well as other African Union official languages such as Arabic (MSA), French, Portuguese, and Spanish. Our training data covers bidirectional translation between English and 13 languages, and between French and two languages (Lingala and Wolof).

AfriNLLB models are based on NLLB-200 600M, which we compress using iterative layer pruning and quantization. We fine-tune the pruned models on parallel corpora we curated for African languages, employing knowledge distillation from a larger teacher model. Our work aims at enabling efficient deployment of translation models for African languages in resource-constrained settings.

Our evaluation results demonstrate that AfriNLLB models achieve performance comparable to the baseline while being significantly faster. We release two versions of the AfriNLLB models, a Transformers version that allows further fine-tuning and a CTranslate2 version for efficient inference. Moreover, we release all the training data that we used for fine-tuning the baseline and pruned models to facilitate further research.

1 AfriNLLB: Background & Motivation

Africa is a linguistically rich continent, with over 2,000 native languages (Grimes, 1996; Heine and Nurse, 2000). Although African languages have millions of native speakers, most of them are low-resource languages (Azime et al., 2024; Wassie, 2024; Adelani et al., 2025b; Farouq et al., 2025; Ojo et al., 2025). This results in a scarcity of

African datasets and models for diverse natural language processing tasks, including machine translation (MT). Since MT resources for African languages are scattered across multiple sources, gathering these resources for fine-tuning open-source models is costly and time-consuming. Moreover, providing translation support for speakers of these low-resource languages in governmental and health sectors remains a significant challenge (Anastasopoulos et al., 2020; Wassie et al., 2025).

AfriNLLB seeks to bridge this gap by delivering efficient translation models and curated training data.^{1,2} Language selection for AfriNLLB considered several factors, including the number of native speakers in Africa and dataset availability. The AfriNLLB models are based on NLLB-200 (Costa-jussà et al., 2022), and support 15 language pairs (30 translation directions), including 10 native African languages: Swahili, Hausa, Yoruba, Amharic, Somali, Zulu, Lingala, Afrikaans, Wolof, and Egyptian Arabic (cf. Table 1). Additionally, we include 5 of the official languages of the African Union, namely Arabic (MSA), English, French, Portuguese, and Spanish. Since several African languages share some lexicon with these languages due to historical contact, multilingual models can leverage this linguistic overlap through transfer learning from high-resource languages to enhance the performance of low-resource languages (Liu et al., 2020; Fan et al., 2021).

AfriNLLB is a series of efficient multilingual open-source models for African languages, motivated by multiple goals:

- Gathering and curating bilingual training datasets for African languages
- Building lightweight MT models specialized in translating African languages, utilizing compression approaches such as pruning and quantization

¹<https://github.com/AfriNLP/AfriNLLB>

²<https://hf.co/collections/AfriNLP/afriNllb>

*Equal contribution

Family	Subfamily	Name	Code	Regions
Afro-Asiatic	Chadic	Hausa	hau_Latn	West Africa (Nigeria, Niger)
	Cushitic	Somali	som_Latn	Horn of Africa (Somalia, Ethiopia, Djibouti, Kenya)
	Semitic	Amharic	amh_Ethi	Horn of Africa (Ethiopia)
	Semitic	Egyptian Arabic	arz_Arab	North Africa (Egypt)
Indo-European	Germanic	Afrikaans	afr_Latn	Southern Africa (South Africa, Namibia)
Niger-Congo	Atlantic	Wolof	wol_Latn	West Africa (Senegal, Gambia, Mauritania)
	Bantu	Lingala	lin_Latn	Central Africa (Congo)
	Bantu	Swahili	swh_Latn	East Africa (Tanzania, Kenya)
	Bantu	Zulu	zul_Latn	Southern Africa (South Africa)
	Volta-Niger	Yoruba	yor_Latn	West Africa (Nigeria, Benin)

Table 1: African Languages in AfriNLLB

Family	Subfamily	Name	Code	Regions
Afro-Asiatic	Semitic	Arabic, Modern Standard	arb_Arab	North Africa (formal use)
Indo-European	Germanic	English	eng_Latn	Southern Africa (South Africa)
	Romance	French	fra_Latn	Africa-wide (mostly L2)
	Romance	Portuguese	por_Latn	Southern Africa (Angola, Mozambique)
	Romance	Spanish	spa_Latn	Central Africa (Equatorial Guinea)

Table 2: Non-Native Languages in AfriNLLB

- Open-sourcing the code, training data, and models we have created
- Sharing our approaches and lessons learned to facilitate future work in this area

2 Data

We employ multi-stage fine-tuning before and after model pruning. First, we fine-tune the baseline NLLB-200 600M to improve the performance for African languages. Afterwards, we fine-tune the pruned models again to restore the translation performance. For this purpose, we collect datasets primarily in African languages (Swahili, Hausa, Yoruba, Amharic, Somali, Zulu, Lingala, Afrikaans, Wolof, and Egyptian Arabic) and a few relevant high-resource languages (Arabic (MSA), French, Spanish, Portuguese).

2.1 Data Sources

We mainly collect the datasets from OPUS (Tiedemann, 2012) and Hugging Face (Lhoest et al., 2021), with additional data from GitHub and other publicly available online sources. This results in a total of 1.2M samples for 11 African language pairs (9 from/into English, and 2 from/into French). For high-resource languages (Arabic, French, Spanish, Portuguese), we focus on collecting only 1.5M for processing, filter the data, and then sample 200k from each language pair for training. Table 3 summarizes data before and after filtering, while Table 6 elaborates on data sources.

2.2 Data Processing

To ensure the quality of data, we process the datasets in a four-stage pipeline: (i) rule-based filtering, (ii) language detection, (iii) semantic filtering, and (iv) quality estimation. While rule-based filtering uses predefined rules, the other pipeline stages employ a model to generate scores and filter the data based on a threshold. We experimented with different threshold values and found 0.6 to be a reasonable choice.

Rule-based filtering involves deduplication, dropping empty segments, and removing HTML tags. We also filter out sentence pairs with lengths less than 3 or greater than 200 characters. Moreover, to avoid misaligned segments, we remove translation pairs exceeding the 2x source-target length ratio.

Language detection discards segments that are unlikely to be in the expected language. We use two language detector models, AfrolID (Adebara et al., 2022) for the African languages and fastText (Joulin et al., 2017) for the rest of the languages.

Semantic filtering evaluates the translation pairs with cosine similarity scores derived from sentence embedding models, using the Sentence-Transformers library (Reimers and Gurevych, 2019). To handle all the languages, we employ different embedding models based on language support. We use *DistilUSE* (Reimers and Gurevych,

2020; Yang et al., 2020) for all high-resource language pairs and *LaBSE* (Feng et al., 2022) for African languages. We apply semantic filtering for all languages except Lingala as we could not find an embedding model that supports it.

Quality estimation is the final stage of the filtering pipeline, in which we apply reference-free evaluation of the translation and exclude segments that are lower than the threshold. We use COMET (Rei et al., 2020) for high-resource language pairs, and Masakhane’s model AfriCOMET-QE-STL (Wang et al., 2024) for African languages.

After thoroughly processing the dataset, we merge the datasets and deduplicate the combined dataset to avoid repetition from different sources. We ended up with a total of 6.4M. However, to mitigate data imbalance, we downsampled the high-resource languages to only 200k per language pair. This results in a total of 1.6M samples (3.2M bidirectional samples, after reversing the dataset), which we use for training. The dataset size for each language direction is presented in Table 3, and elaborated in Table 6.

2.3 Validation and Test Data

We use Flores200³ (Costa-jussà et al., 2022) for validation and test, as it covers all the languages in our experiments. We use the *dev* split (997 segments) of Flores200 for validation during training, and for layer importance evaluation as part of iterative layer pruning (cf. Section 3), and use the *devtest* split (1,012 segments) for testing and evaluation of our models.

Language Pair	Initial	Processed	Sampled	
afr_Latn	192,541	161,644	161,644	
amh_Ethi	156,739	85,010	85,010	
arz_Arab	85,942	84,170	84,170	
hau_Latn	222,387	155,881	155,881	
som_Latn	87,521	43,657	43,657	
eng_Latn	swh_Latn	286,687	181,045	181,045
	wol_Latn	34,956	31,170	31,170
	yor_Latn	34,720	22,626	22,626
	zul_Latn	38,532	33,189	33,189
	arb_Arab	1,526,102	1,424,237	200,000
	fra_Latn	1,500,000	1,483,951	200,000
	por_Latn	1,500,000	1,401,671	200,000
spa_Latn	1,500,000	1,324,681	200,000	
fra_Latn	wol_Latn	10,745	9,071	9,071
	lin_Latn	8126	1,948	1,948
Total	7,184,998	6,443,951	1,609,411	

Table 3: Parallel corpus sizes before and after processing from and into English and French. Since all data is reversed to create the opposite translation direction, the final dataset size is effectively doubled.

³<https://hf.co/datasets/facebook/flores>

3 Methodology

In our experiments, we apply iterative layer pruning to the *NLLB-200 600M* model after fine-tuning it on the training dataset. This approach incrementally identifies and removes layers with minimal contribution to translation quality, one layer at a time. The pruned models resulting from this process are then fine-tuned again to restore most of the translation quality of the baseline model. The resulting models are smaller and faster while retaining or outperforming the quality of the baseline. The following points elaborate on the process.

Layer importance evaluation: We conduct layer importance evaluation by measuring translation performance without each layer. In this greedy layer pruning approach (Peer et al., 2022; Rostami and Dousti, 2024; Moslem et al., 2025; Moslem, 2025), to prune $n + 1$ layers, only a single optimal layer to prune must be added to the already known solution for pruning n layers. After identifying and removing the least critical layer, we repeat the layer importance evaluation on the remaining layers until reaching our n pruning target. We observe that while removing certain layers of the model (e.g. the first or last layer) substantially degrades translation performance, others result in minimal performance drops. Following Moslem (2025), we use the chrF++ metric for layer importance evaluation for both better efficiency and quality. We use the dev split of the Flores200 dataset, mainly where African languages are the target, to improve their translation quality. In the future, we plan to experiment with using both directions.

Layer pruning: We iteratively prune one decoder layer at a time, selecting the layer whose removal has the least negative impact on translation quality, measured by chrF++ scores. At each iteration, we evaluate the translation performance of the pruned model on the dev split of the Flores200 dataset, after removing each candidate layer. The layer whose removal yields the best performance is eventually pruned. This process continues until a predefined number of layers (4, 6, or 8 layers) have been removed. By iteratively removing the least important layers, this performance-guided method produces a more compact model that can be fine-tuned further to recover the translation quality of the original model. We also experimented with middle layer pruning and found that iterative layer pruning yields better results (cf. Section 4.1).

Direction	Model	BLEU \uparrow	chrF++ \uparrow	COMET \uparrow	Throughput (toks/s) \uparrow	Time (s) \downarrow
xx-en	NLLB 600M (Baseline)	33.81	56.22	71.11	1469.96	21.02
	NLLB 600M + FT	35.15	57.61	71.87	1530.94	20.39
	Pruned + FT	34.01	56.98	71.20	1807.61	17.38
	Pruned + FT (FP16)	34.05	56.99	71.19	3513.32	8.96
en-xx	NLLB 600M (Baseline)	22.70	47.89	69.36	1530.10	28.09
	NLLB 600M + FT	24.28	49.97	70.91	1610.23	26.98
	Pruned + FT	24.17	50.05	70.37	1946.61	22.51
	Pruned + FT (FP16)	24.15	50.06	70.41	3732.72	11.98
xx-fr	NLLB 600M (Baseline)	16.41	38.83	17.34	1475.48	26.46
	NLLB 600M + FT	17.91	40.45	18.47	1524.32	26.12
	Pruned + FT	17.43	40.21	14.52	1845.09	21.61
	Pruned + FT (FP16)	17.38	40.18	14.53	3569.23	11.17
fr-xx	NLLB 600M (Baseline)	9.44	33.42	19.25	1047.18	49.92
	NLLB 600M + FT	10.98	35.68	21.33	1081.84	51.56
	Pruned + FT	10.20	35.21	20.04	1261.66	49.91
	Pruned + FT (FP16)	10.11	35.13	20.03	2313.85	31.15

Table 4: Average Performance by Translation Direction. The category en \leftrightarrow xx includes 13 language pairs (26 translation directions), while the category fr \leftrightarrow xx includes 2 language pairs for Lingala and Wolof (4 translation directions). The pruned models are up to 20% faster than the baseline without quantization, and 57% faster with float16 quantization. While more efficient, the translation quality of the compressed models is comparable with the fine-tuned NLLB-200 model. Table 5 elaborates on the experimental results.

Fine-tuning: We employ multi-stage fine-tuning. First, we fine-tune the baseline NLLB-200 model on the training dataset to improve its quality for African languages. Since pruning the fine-tuned models results in performance degradation, the pruning step is followed by fine-tuning the pruned model for 1 epoch using the training dataset (cf. Section 2). During training, we use a learning rate of $5e-5$, a batch size of 8, gradient accumulation steps of 4, and early stopping with a patience value of 10 evaluation runs. The evaluation takes place every 1000 training steps. The final saved model is the best model based on the evaluation loss score. The training is conducted on one A40 48GB GPU. We use the *Transformers* framework⁴ (Wolf et al., 2020) for training. As illustrated by Table 4, this fine-tuning step successfully recovers the translation quality of the baseline model.

Knowledge distillation: To improve the quality of our models, we employ sequence-level knowledge distillation (Kim and Rush, 2016; Crego and Senellart, 2016; Gandhi et al., 2023), where the student model is fine-tuned on a combination of authentic data and synthetic data generated by the teacher model for the same training dataset. In this case, the teacher model is the NLLB-200 3.3B baseline while the students are the NLLB-200 600M baseline and then the pruned models based on our

fine-tuned version. After generating the data, we filter it by removing duplicates (exact matches in the target side of the authentic data), and we follow the filtering pipeline we use for processing the original training data (cf. Section 2). The knowledge distillation data after filtering is 568k segments for African languages.

4 Evaluation and Results

For inference, we use CTranslate2⁵ (Klein et al., 2020), with beam size of 3 and batch size of 1024 tokens, on a A40 48GB GPU.

To evaluate our systems, we calculated BLEU (Papineni et al., 2002), chrF++ (Popović, 2017), as implemented in the sacreBLEU library⁶ (Post, 2018). For semantic evaluation, we use AfriCOMET (Wang et al., 2024) for African languages, and COMET (Rei et al., 2020) for Arabic and European languages.⁷

The process of iterative layer pruning of 4 decoder layers created a 548M model that is 23% faster in average than the baseline. Moreover, the quality degradation caused by pruning has been mitigated through fine-tuning and knowledge distillation. As demonstrated by Table 4 and elaborated by Table 7, by the end of the process, the

⁵<https://github.com/OpenNMT/CTranslate2>

⁶<https://github.com/mjpost/sacrebleu>

⁷In particular, we used the “*africomet-ml*” model for AfriCOMET and the “*wmt22-comet-da*” model for COMET.

⁴<https://github.com/huggingface/transformers>

pruned model could recover most of the translation quality of the baseline model. Moreover, quantization (float16) of the pruned model further enhanced the inference performance, making the model 57% faster than the baseline.

4.1 Ablation Study

In this ablation study, we compare three scenarios: (i) removing middle layers⁸ instead of iteratively determining the layers to remove based on layer importance evaluation (cf. Section 3), (ii) pruning both encoder and decoder layers instead of pruning decoder layers only, and (iii) pruning various values of the decoder layers, namely 4, 6, and 8 layers.

We observe that iterative layer pruning clearly outperforms middle layer pruning in both cases of removing decoder layers only or both encoder and decoder layers. Fine-tuning after pruning is crucial in all cases, as it mitigates the effect of pruning on performance. Figure 1 illustrates four pruned models, both before and after fine-tuning:

- Middle pruning, 4 decoder layers (Mid 548M)
- Middle pruning, 4 encoder layers and 4 decoder layers (Mid 498M)
- Iterative pruning, 4 decoder layers (Iter 548M)
- Iterative pruning, 4 encoder layers and 4 decoder layers (Iter 498M)

When it comes to removing encoder layers in addition to decoder layers, it is not clear to what extent this affects the quality. Obviously, removing encoder layers reduces the size of the model further, which can cause performance degradation. Keeping encoder layers intact was recommended by previous work on speech (Gandhi et al., 2023; Moslem, 2025), which poses the question whether the same concept applies to text-based encoder-decoder models such as NLLB-200. We intend to investigate this further in future work.

Furthermore, we thoroughly studied the effect of keeping all 12 encoder layers intact while iteratively removing different numbers of decoder layers. We experimented with three pruning configurations, removing 4, 6, or 8 decoder layers, resulting in models with 12 encoder layers and 8, 6, or 4 decoder layers, respectively. As illustrated in Figure 2 and Figure 3, the effect of the number of decoder layers removed varies across language pairs, although removing up to 6 layers (50%) yields similar or better performance compared to the NLLB-200 600M baseline, thanks to

⁸For middle layer pruning, we remove layers 4 to 7 inclusively.

fine-tuning before and after pruning. Table 5 elaborates further on the performance results in terms of both translation quality and inference speed.

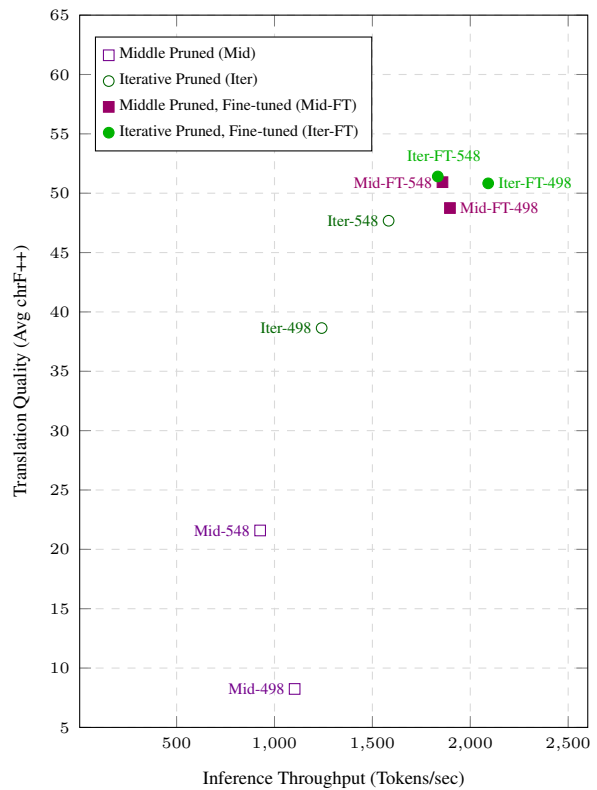


Figure 1: Quality-Efficiency Comparison. The iterative-pruned models demonstrate a superior balance of speed and quality compared to the middle-pruned variants. The 548M models include 12 encoder layers and 8 decoder layers (i.e. 4 decoder layers are pruned), while the 498M models include 8 encoder layers and 8 decoder layers (i.e. 8 layers are pruned, 4 from the encoder and 4 from the decoder). The chart reports the average chrF++ scores across all language pairs before and after fine-tuning the pruned models.

5 Conclusions and Future Work

In this work, we presented AfriNLLB, lightweight models for African languages, that achieve over 20–50% inference performance gains compared to their baseline NLLB-200 600M. We release models with various sizes to match different needs.

We have demonstrated that iterative layer pruning is an effective approach for model compression while retaining translation quality. The method relies on layer importance evaluation, followed by fine-tuning on a medium-sized dataset. This iterative layer pruning process reduces the model size and accelerates inference. We are open-sourcing AfriNLLB models and data. In addition, to ensure

reproducibility, we are making all the processing and training code publicly available.

In future versions of AfriNLLB, we plan to add more languages. Research directions include investigating data augmentation approaches besides knowledge distillation, such as back-translation. Moreover, we plan to expand our approach to other architectures, such as autoregressive large language models and encoder-only models.

We hope that by releasing AfriNLLB models, training data, and code, we facilitate further research on African languages, and support the African community worldwide.

References

- Idris Abdulmumin, Satya Ranjan Dash, Musa Abdullahi Dawud, Shantipriya Parida, Shamsuddeen Muhammad, Ibrahim Sa'id Ahmad, Subhadarshi Panda, Ondřej Bojar, Bashir Shehu Galadanci, and Bello Shehu Bello. 2022. [Hausa visual genome: A dataset for multi-modal English to Hausa machine translation](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6471–6479, Marseille, France. European Language Resources Association.
- Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Inciarte. 2022. [AfroLID: A neural language identification tool for African languages](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1958–1981, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruitter, Dietrich Klakow, Peter Nabende, and 14 others. 2022. [A few thousand translations go a long way! leveraging pre-trained models for African news translation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070. Association for Computational Linguistics.
- David Adelani, Dana Ruitter, Jesujoba Alabi, Damilola Adebajo, Adesina Ayeni, Mofetoluwa Adeyemi, Ayodele Awokoya, and Cristina Espina-Bonet. 2021. [MENYO-20k: A multi-domain English-Yor'ub'a corpus for machine translation and domain adaptation](#). In *Proceedings of the Second Workshop on African Natural Language Processing*, pages 27–34. Association for Computational Linguistics.
- David Ifeoluwa Adelani, Alison Chi, Simbiat Aderibigbe, Butoyi Beatrice, Tumaini Balikwisha, Barkwende Hugues Diallo, Tunde Oluwaseyi Ajayi, Joseph K. O. Oaminu, Ruqayya Nasir Iro, and 12 others. 2025a. [AFRIDOC-MT: Document-level MT Corpus for African Languages](#). *arXiv preprint arXiv:2501.06374*.
- David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, and 18 others. 2025b. [IrokoBench: A new benchmark for African languages in the age of large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2732–2757, Albuquerque, New Mexico. Association for Computational Linguistics.
- Rania Al-Sabbagh. 2024. [Arzen-multigenre: An aligned parallel dataset of egyptian arabic song lyrics, novels, and subtitles, with english translations](#). *Data in Brief*, 54:110271.
- Duarte Miguel Alves, Jose Pombal, Nuno M. Guerreiro, Pedro Henrique Martins, Joao Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, and 4 others. 2025. [WMT24++: Expanding the Language Coverage of WMT24 to 55 Languages & Dialects](#). *arXiv preprint arXiv:2502.12404*.
- Antonios Anastasopoulos, Alessandro Cattelan, Zi-Yi Dou, Marcello Federico, Christian Federmann, Dmitriy Genzel, Francisco Guzmán, Junjie Hu, Macduff Hughes, and 9 others. 2020. [TICO-19: the Translation Initiative for COvid-19](#). In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*, Online. Association for Computational Linguistics.
- Israel Abebe Azime, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Mitiku Yohannes Fuge, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Waleign Tewabe Sewunetie, and Seid Muhie Yimam. 2024. [Walia-LLM: Enhancing Amharic-LLaMA by integrating task-specific and generative datasets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 432–444, Miami, Florida, USA. Association for Computational Linguistics.
- Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, and 13 others. 2025. [SMOL: Professionally translated parallel data for 115 under-represented languages](#). *arXiv preprint arXiv:2502.12301*.
- Mauro Cettolo, Christian Girardi, and Marcello Federico. 2012. [WIT3: Web inventory of transcribed and translated talks](#). In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation*, pages 261–268. European Association for Machine Translation.
- Christos Christodouloupoulos and Mark Steedman. 2015. [A massively parallel corpus: the bible in 100 languages](#). *Language Resources and Evaluation*, 49(2):375–395.
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe

- Kalbassi, Janice Lam, Daniel Licht, and 29 others. 2022. [No Language Left Behind: Scaling human-centered machine translation](#). *arXiv [cs.CL]*.
- Josep Crego and Jean Senellart. 2016. [Neural Machine Translation from Simplified Translations](#). *arXiv [cs.CL]*.
- Andreas Eisele and Yu Chen. 2010. [MultiUN: A multilingual corpus from united nation documents](#). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA).
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Sibli, Dominik Krzemiński, and 77 others. 2025. [Mmteb: Massive multilingual text embedding benchmark](#). *arXiv preprint arXiv:2502.13595*.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, and 7 others. 2021. [Beyond English-Centric Multilingual Machine Translation](#). *Journal of Machine Learning Research*, 22(107):1–48.
- Muhammad Hazim Al Farouq, Aman Kassahun Wassie, and Yasmin Moslem. 2025. [Bemba Speech Translation: Exploring a Low-Resource African Language](#). In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 354–359, Vienna, Austria (in-person and online). Association for Computational Linguistics.
- Christian Federmann, Tom Kocmi, and Ying Xin. 2022. [NTREX-128 – a benchmark for evaluating machine translation performance](#). In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24. Association for Computational Linguistics.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. 2023. [Distil-Whisper: Robust knowledge distillation via large-scale pseudo labelling](#). *arXiv [cs.CL]*.
- Barbara F. Grimes. 1996. *Ethnologue: Languages of the World*, 13th edition. SIL International, Dallas, TX. Summer Institute of Linguistics.
- Bernd Heine and Derek Nurse, editors. 2000. *African Languages: An Introduction*. Cambridge University Press, Cambridge.
- Andreea Iana, Goran Glavočić, and Heiko Paulheim. 2023. [News without borders: Domain adaptation of multilingual sentence embeddings for cross-lingual news recommendation](#). In *Proceedings of the 17th ACM Conference on Recommender Systems*. ACM.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. [Bag of Tricks for Efficient Text Classification](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- Yoon Kim and Alexander M Rush. 2016. [Sequence-Level Knowledge Distillation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas. Association for Computational Linguistics.
- Guillaume Klein, Dakun Zhang, Clément Chouteau, Josep Crego, and Jean Senellart. 2020. [Efficient and high-quality neural machine translation with OpenNMT](#). In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 211–217, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Laban Kumbuga, Joyce Nakatumba-Nabende, Jonathan Mukiibi, and Andrew Katumba. 2024. [SALT: Sunbird African language translation corpus](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5462–5472. ELRA and ICCL.
- Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, and 23 others. 2021. [Datasets: A community library for natural language processing](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. [Multilingual Denoising Pre-training for Neural Machine Translation](#). *Transactions of the Association for Computational Linguistics*, 8:726–742.
- Yasmin Moslem. 2025. [Efficient speech translation through model compression and knowledge distillation](#). In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 379–388, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Yasmin Moslem, Muhammad Hazim Al Farouq, and John Kelleher. 2025. [Iterative layer pruning for efficient translation inference](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 1022–1027, Suzhou, China. Association for Computational Linguistics.

- Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. **AfroBench: How good are large language models on African languages?** In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. **Bleu: a Method for Automatic Evaluation of Machine Translation.** In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- David Peer, Sebastian Stabinger, Stefan Engl, and Antonio Rodríguez-Sánchez. 2022. **Greedy-layer pruning: Speeding up transformer models for natural language processing.** *Pattern Recognit. Lett.*, 157:76–82.
- Maja Popović. 2017. **chrF++: words helping character n-grams.** In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Matt Post. 2018. **A Call for Clarity in Reporting BLEU Scores.** In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. **COMET: A Neural Framework for MT Evaluation.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2019. **Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.** In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2020. **Making monolingual sentence embeddings multilingual using knowledge distillation.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Pedram Rostami and Mohammad Javad Dousti. 2024. **CULL-MT: Compression using language and layer pruning for machine translation.** *arXiv [cs.CL]*.
- Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufiş, and Dániel Varga. 2006. **The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages.** In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC’06)*, Genoa, Italy. European Language Resources Association (ELRA).
- Jörg Tiedemann. 2012. **Parallel Data, Tools and Interfaces in OPUS.** In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).
- Jörg Tiedemann. 2020. **The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT.** In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.
- Jiayi Wang, David Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, and 49 others. 2024. **AfriMTE and AfriCOMET: Enhancing COMET to embrace under-resourced African languages.** In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5997–6023, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Aman Kassahun Wassie. 2024. **Machine translation for ge’ez language.** *arXiv preprint arXiv:2311.14530*, arXiv:2311.14530.
- Aman Kassahun Wassie, Mahdi Molaei, and Yamin Moslem. 2025. **Domain-specific translation with open-source large language models: Resource-oriented analysis.** *arXiv preprint arXiv:2412.05862*, arXiv:2412.05862.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, and 13 others. 2020. **Transformers: State-of-the-art natural language processing.** In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, and 3 others. 2020. **Multilingual Universal Sentence Encoder for Semantic Retrieval.** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 87–94, Online. Association for Computational Linguistics.
- Biao Zhang, Philip Williams, Ivan Titov, and Rico Senrich. 2020. **Improving massively multilingual neural machine translation and zero-shot translation.** In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1628–1639. Association for Computational Linguistics.

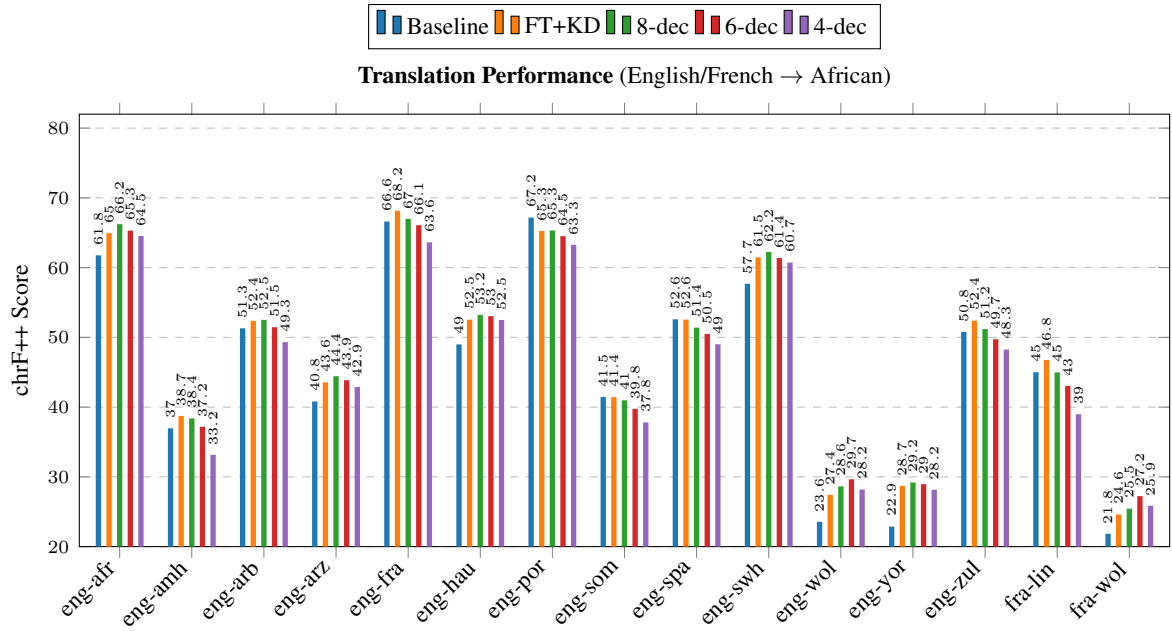


Figure 2: Translation performance (chrF++) from English/French to African languages.

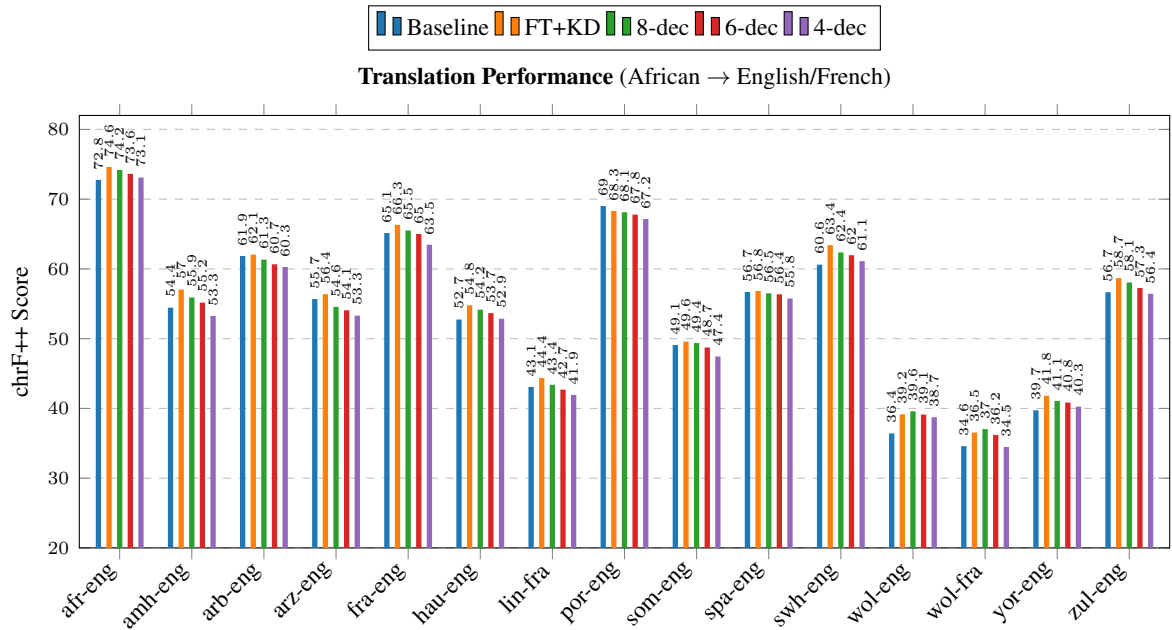


Figure 3: Translation performance (chrF++) from African languages to English/French.

Performance Comparison: Layer Pruning Configurations

Translation quality (BLEU, chrF++, COMET) and efficiency (throughput, inference time) across baseline, fine-tuned, and pruned configurations with optional float16 (FP16) quantization

Lang	Model	Enc	Dec	Quant	BLEU ↑	chrF++ ↑	COMET ↑	Throughput ↑	Time ↓
xx-en	NLLB	12	12	–	33.81	56.22	71.11	1469.96	21.02
				FP16	33.80	56.22	71.13	2834.69	10.92
	NLLB + FT	12	12	–	35.15	57.61	71.87	1530.94	20.39
				FP16	35.10	57.61	71.87	2808.90	11.15
	AfriNLLB	12	8	–	34.01	56.98	71.20	1807.61	17.38
				FP16	34.05	56.99	71.19	3513.32	8.96
12		6	–	33.35	56.48	70.79	2028.18	15.41	
			FP16	33.32	56.45	70.79	4000.25	7.82	
12	4	–	32.03	55.62	69.71	2257.03	13.77		
		FP16	32.01	55.60	69.71	4589.42	6.79		
8	8	–	30.89	54.32	68.08	1852.13	17.05		
		FP16	30.86	54.30	68.08	3550.50	8.91		
en-xx	NLLB	12	12	–	22.70	47.89	69.36	1530.10	28.09
				FP16	22.68	47.88	69.38	2898.38	15.33
	NLLB + FT	12	12	–	24.28	49.97	70.91	1610.23	26.98
				FP16	24.14	49.84	70.90	2811.34	18.82
	AfriNLLB	12	8	–	24.17	50.05	70.37	1946.61	22.51
				FP16	24.15	50.06	70.41	3732.72	11.98
12		6	–	23.48	49.34	68.98	2265.87	18.50	
			FP16	23.49	49.35	69.00	4428.68	9.65	
12	4	–	21.77	47.80	65.68	2489.35	17.31		
		FP16	21.77	47.81	65.68	4954.62	9.09		
8	8	–	23.59	49.64	69.90	2015.53	21.34		
		FP16	23.58	49.63	69.88	3851.13	11.34		
xx-fr	NLLB	12	12	–	16.41	38.83	17.34	1475.48	26.46
				FP16	16.33	38.83	17.23	2850.66	13.71
	NLLB + FT	12	12	–	17.91	40.45	18.47	1524.32	26.12
				FP16	17.83	40.42	18.37	2749.45	14.68
	AfriNLLB	12	8	–	17.43	40.21	14.52	1845.09	21.61
				FP16	17.38	40.18	14.53	3569.23	11.17
12		6	–	16.52	39.44	11.78	2044.27	19.21	
			FP16	16.54	39.42	11.68	3953.51	9.92	
12	4	–	14.96	38.21	5.67	2340.99	16.77		
		FP16	14.90	38.17	5.71	4766.12	8.24		
8	8	–	14.42	37.05	3.14	1866.26	21.84		
		FP16	14.34	36.97	3.14	3448.51	11.83		
fr-xx	NLLB	12	12	–	9.44	33.42	19.25	1047.18	49.92
				FP16	9.52	33.40	19.38	1920.41	29.05
	NLLB + FT	12	12	–	10.98	35.68	21.33	1081.84	51.56
				FP16	10.48	35.05	21.49	1700.25	51.31
	AfriNLLB	12	8	–	10.20	35.21	20.04	1261.66	49.91
				FP16	10.11	35.13	20.03	2313.85	31.15
12		6	–	10.07	35.14	19.83	1416.33	30.89	
			FP16	9.99	35.08	19.78	2465.60	18.68	
12	4	–	7.57	32.42	14.16	1207.06	38.75		
		FP16	7.57	32.38	14.29	2069.52	23.25		
8	8	–	9.75	35.23	20.05	1222.83	45.33		
		FP16	9.84	35.31	20.11	2186.73	25.97		

Table 5: Comprehensive performance evaluation across translation directions. AfriNLLB models use various encoder-decoder layer configurations (12-8, 12-6, 12-4, 8-8) with and without float16 quantization.

Datasets Sources and Sizes

Names, sources, and sizes of our training datasets before and after filtering for each language pair

Dataset	fra-eng	spa-eng	por-eng	arb-eng	swb-eng	amh-eng	son-eng	hau-eng	yor-eng	zul-eng	afz-eng	arz-eng	wol-fra	wol-eng	lin-fra
<i>OPUS Datasets</i>															
Tatoeba (Tielemann, 2020)	-	-	-	-	-	213/188	9/5	259/183	423/421	72/170	2.4K/2.1K	6.5K/1.3K	-	-	555/120
translatowiki	-	-	-	-	-	-	-	-	-	-	6.2K/244	111K/23K	1.7K/243	-	-
wikimedia	1.4M/1.1M	-	822K/610K	621K/374K	16.3K/11.3K	4043	942/425	1.1K/718	190K/121K	12.5K/4.8K	9.3K/5.5K	78.5K/66.5K	-	690/169	21/5
GNOME	-	-	21.2K/15.3K	150/41	4043	57.1K/26.9K	753/1.1K	5.8K/110	-	1K/590	4.5K/7.7K	12.7K/27.8K	-	-	-
Ubuntu	-	-	-	6K/2.5K	-	-	-	-	-	1410	-	-	-	220/38	222/26
GlobalVoices	-	-	-	-	32.3K/26.9K	-	-	-	-	136/61	-	-	-	-	-
bible-sodia (Christodoulosopoulos and Steedman, 2015)	-	-	-	62.2K/16.3K	-	1.8K/1.2K	6.2K/49.5K	-	-	-	62.1K/50.6K	-	-	7.9K/648	15.8K/2.6K
Neulab-TedTalks	212K/185K	215K/190K	81.2K/55K	-	-	-	-	-	-	-	-	-	-	-	-
EMEA	-	1.1M/235K	-	-	-	-	-	-	-	-	-	-	-	-	-
ELBookshop	-	-	-	-	1.7K/1.10	-	-	-	-	-	-	-	-	-	-
ELRC-wiki_health	4.4K/3.7K	-	-	15.1K/14.4K	608/501	-	-	-	-	-	-	404/312	-	-	-
News-Commentary	156K/125K	-	-	-	-	-	-	-	-	-	-	-	-	-	-
IRC-Aquis (Steinberger et al., 2006)	814K/65.3K	806K/398K	-	-	-	-	-	-	-	-	-	-	-	-	-
TED2020	-	-	-	408K/341K	9.7K/80.8K	1K/1.7K	2K/1.3K	27/21	-	-	-	2.3K/1.8K	-	-	-
KDE4	-	-	-	116K/25.6K	-	-	-	149/66	-	-	-	64.3K/29.8K	-	-	-
ELRC-EMEA	-	777K/614K	-	-	-	-	-	-	-	-	-	-	-	-	-
Books	-	93.5K/63.4K	-	-	-	-	-	-	-	-	-	-	-	-	-
Tanzil	-	-	-	-	138K/96.7K	93.5K/90.5K	93.8K/10.5K	128K/63.4K	-	-	-	-	-	-	-
OpenSubtitles	-	-	-	-	94.6K/95.8K	3K/1.8K	531/446	-	-	-	-	969K/11.8K	-	-	-
TICO-19 (Anastasopoulos et al., 2020)	-	-	-	-	3.1K/2.8K	3.1K/3.1K	3.1K/1.2K	3.1K/2.1K	-	3.1K/2.3K	-	-	-	-	2.9K/544
ELRC_2022	-	-	-	-	607/498	-	-	-	-	-	-	403/310	-	-	-
ELRC-3073-wiki_health	-	-	-	-	608/501	-	-	-	-	-	-	-	-	-	-
infopankki	-	-	-	-	-	-	47.2K/89.8K	-	-	-	-	-	-	-	-
QED	-	-	-	-	-	-	-	-	-	-	-	28.8K/17.5K	-	-	-
SPC	-	-	-	-	-	-	-	-	-	-	-	57.4K/47.3K	-	-	-
ELRC-monuments	-	-	-	-	-	-	-	-	-	-	-	54/41	-	-	-
ELRC-Museus	-	-	-	-	-	-	-	-	-	-	-	320	-	-	-
<i>HuggingFace Datasets</i>															
smol (Caswell et al., 2025)	-	-	-	-	863/719	862/551	863/548	863/153	863/552	863/610	-	-	-	-	7.4K/570
mafand (Adelani et al., 2022)	-	-	-	-	34.4K/29.9K	1.9K/1.4K	-	5.9K/4.4K	6.6K/4K	3.5K/2K	-	-	-	-	-
mafand-dev	-	-	-	-	-	-	-	1.3K/971	6.6K/4K	1.2K/636	-	-	-	-	-
mafand-test	-	-	-	-	-	-	-	1.5K/1.2K	6.6K/4K	998/596	-	-	-	-	-
Ponoso-Translations	-	-	-	-	6.1K/2.8K	-	-	3.2K/1.2K	4.4K/553	3.3K/735	13.1K/2.7K	-	-	-	6.8K/802
Wolbat-Translations	-	-	-	-	17.2K/7.2K	8K/2.1K	1.6K/310	143/54	164/333	66/53	23.2K/1.8K	-	-	-	-
nrxex (Federmann et al., 2022)	-	-	-	-	2K/1.7K	2K/1.5K	2K/1.9K	2K/1.2K	2K/602	2K/1.8K	-	-	-	-	-
AfriDocMT-health (Adelani et al., 2025a)	-	-	-	-	7K/6.6K	7K/6.6K	-	7K/5.9K	7K/3.7K	7K/5.4K	-	-	-	-	-
AfriDocMT-doc_health	-	-	-	-	2404	2406	-	-	-	-	-	-	-	-	-
AfriDocMT-doc_health_2	-	-	-	-	540/96	540/104	-	-	-	-	-	-	-	-	-
AfriDocMT-doc_health_5	-	-	-	-	1.5K/1.5K	1.5K/1.5K	-	-	-	-	-	-	-	-	-
AfriDocMT-doc_health_10	-	-	-	-	812/266	812/400	-	-	-	-	-	-	-	-	-
quran_multilingual	-	-	-	-	6.2K/5K	6.2K/3.7K	6.2K/740	6.2K/3.8K	6.2K/1K	-	-	-	-	-	-
Nazirali-Quran	-	-	-	-	6.2K/5K	6.2K/3.7K	6.2K/5K	-	-	-	-	-	-	-	-
OPUS-100 (Zhang et al., 2020)	-	-	-	-	-	-	-	-	-	10.4K/2.3K	-	-	-	-	-
OPUS-100-dev	-	-	-	-	-	-	-	-	-	10.4K/2.3K	-	-	-	-	-
OPUS-100-test	-	-	-	-	-	-	-	-	-	10.4K/2.3K	-	-	-	-	-
menyo20k_mt-train (Adelani et al., 2021)	-	-	-	-	-	-	-	-	-	10.1K/4.6K	-	-	-	-	-
menyo20k_mt-dev	-	-	-	-	-	-	-	-	-	3.4K/1.4K	-	-	-	-	-
menyo20k_mt-test	-	-	-	-	-	-	-	-	-	6.6K/3.7K	-	-	-	-	-
yoruba_audio_trans	-	-	-	-	-	-	-	-	-	9.2K/1.9K	-	-	-	-	-
arz-en-parallel	-	-	-	-	-	-	-	-	-	-	-	25K/22.6K	-	-	-
news-comm-eng-arz (Moslem et al., 2025)	-	-	-	-	-	-	-	-	-	-	-	832K/83.3K	-	-	-
mehnatatoha-bitext (Enevoldsen et al., 2025)	-	-	-	-	-	-	-	-	-	-	-	8.9K/2.9K	-	-	-
fr-wolof-trans-gs	-	-	-	-	-	-	-	-	-	-	-	-	10.4K/1.6K	-	-
wolof_en_fr	-	-	-	-	-	-	-	-	-	-	-	-	26.6K/6.5K	26.6K/7.6K	84.7K/17.2K
english_wolof_trans	-	-	-	-	-	-	-	-	-	-	-	-	-	7.5K/4K	13.4K/2.2K
comet_score_en_wo	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
wolof_en_bible	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
MuHiLiN (Eisele and Chen, 2010)	-	-	-	9.8M/9.8M	-	-	-	-	-	-	-	-	-	-	-
ted_talks_jwsh-14 (Cettolo et al., 2012)	-	-	-	-	52/42	-	-	-	-	-	-	-	-	-	-
ted_talks_jwsh-15	-	-	-	-	68/53	-	-	-	-	-	-	-	-	-	-
ted_talks_jwsh	-	-	-	-	-	-	188/730	-	-	-	-	-	-	-	-
WMT24pp (Alves et al., 2025)	-	-	-	-	998/691	-	-	-	-	-	-	-	-	-	-
sambird-salt (Kumbuga et al., 2024)	-	-	-	-	24.9K/23.1K	-	-	-	-	-	-	-	-	-	-
Hausa-VI (Abdelmounim et al., 2022)	-	-	-	-	-	-	-	28.9K/7.9K	-	-	-	-	-	-	-
polynnews-parallel (fana et al., 2023)	-	-	-	-	-	-	-	5.7K/4.4K	-	3.4K/2K	-	-	-	-	-
Quran	-	-	-	-	-	-	-	6.2K/3.7K	-	-	-	-	-	-	-
lit-spe	-	-	-	-	-	-	-	-	-	-	-	57.4K/47.4K	-	-	-
linguax_test	-	-	-	-	-	-	-	-	-	-	-	1.1K/649	-	-	-
subscene	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
opus_infopankki	-	-	-	-	-	-	9000	-	-	-	-	-	-	-	-
opus_infopankki	-	-	-	-	-	-	47.2K/89.8K	-	-	-	-	-	-	-	-
<i>other sources</i>															
ArzEn-MultiGenre (Al-Sabbagh, 2024)	-	-	-	-	-	-	-	-	-	-	-	25K/6.6K	-	-	-
ethiopian-legal	-	-	-	-	-	5.4K/3.7K	-	-	-	-	-	-	-	-	-
ethiopian-history	-	-	-	-	-	1.3K/737	-	-	-	-	-	-	-	-	-
ethiopian-news	-	-	-	-	-	5.4K/1.1K	-	-	-	-	-	-	-	-	-
ethiopian-bible	-	-	-	-	-	6.5K/3.8K	-	-	-	-	-	-	-	-	-
ethiopian-ethio_bible	-	-	-	-	-	11.7K/5.7K	-	-	-	-	-	-	-	-	-
ethiopian-jw_bible	-	-	-	-	-	31.1K/25.2K	-	-	-	-	-	-	-	-	-
ethiopian-jw_daily	-	-	-	-	-	4.7K/4.3K	-	-	-	-	-	-	-	-	-
horv-ent	-	-	-	-	-	2K/2K	-	-	-	-	-	-	-	-	-
mt-eval-am-amen	-	-	-	-	-	997/712	-	-	-	-	-	-	-	-	-
mt-eval-am-enum	-	-	-	-	-	1.9K/1.4K	-	-	-	-	-	-	-	-	-
skachumana	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
zenodo-training	-	-	-	-	-	-	-	-	-	-	26.7K/13.8K	-	-	-	-
zenodo-eval	-	-	-	-	-	-	-	-	-	-	4.7K/2.6K	-	-	-	-
Gampyan-fr-it	-	-	-	-	-	-	-	-	-	-	998/596	-	-	-	-
Total (Origin)	2.6M	3M	6.2M	11M	399K	319K	275K	398K	124K	113K	619K	244K	47.3K	162K	8.1K
After Filter	1.5M	1.5M	1.5M	10.5M	287K	157K	87.3K	222K	34.7K	38.5K	174K	85.9K	9.2K	35K	2K
After Dedup	1.48M	1.32M	1.4M	1.42M	181K	85K	87.5K	156K	22.6K	33.2K	174K	84.2K	9.1K	31.2K	1.9K

Table 6: Dataset statistics for all language pairs. Values shown as Original/Final (K=thousand, M=million), and “-” indicates dataset not used.

Author Index

- Abdulmumin, Idris, 197
Abebe, Amanuel Gizachew, 281
Abolade, Daud, 82
Adebanjo, Fatimo, 44
Adelani, David Ifeoluwa, 103, 243
Adeleke, Lateef, 149
Agbobli, Messan, 264
Agossou, Emile KOGBEDJI, 264
Ajao, Simbiat, 235
Akbik, Alan, 271
Akeru, Benjamin, 204, 211
Aliyu, Mahi Aminu, 44
Ankeli, Oche David, 256
Assowe, Houssein A, 179
Awak, Mbuotidem, 56
- Babatunde, Oreoluwa Boluwatife, 228
Bainomugisha, Engineer, 211
Bakouya, Guedela, 264
Basilwango, Ertony, 256
Bassey, Ita Etim, 56
Belay, Tadesse Destaw, 197
Bello, Muhammad Yahuza, 197
Berdys, Pierre Herve, 256
Beux, Yann Le, 256
Bolarinwa, Emmanuel, 228
Borisov, Vadim, 116
Borodo, Salisu Musa, 37
Breiding, Adrian, 271
- Carpuat, Marine, 163
Carroll, Cody, 103
Chen, Fei-Yueh, 149
Coulibaly, Nouhoum, 1, 190
- Dembélé, Aymane, 1
Diallo, Seydou, 17, 248
Diarra, Sebastien, 17
Diarra, Yacouba, 1, 190, 248
Doumeyan, Eli Jean, 264
Downey, C. M., 149
- Edet, Offiong Bassey, 56
Egbewale, Jesujuwon, 217
El Oirghi, Hiba, 163
Elhoseiny, Mohamed, 11
Essuman, Catherine Nana Nyaah, 264
Ezeani, Ignatius, 64
- Fokoa, Kevin Jeff Fogang, 171
- Gedi, Abdifatah Ahmed, 179
Gilbert, Yiga, 204
Gonzalez, Javier, 64
Gwadabe, Tajuddeen, 163
Gyamfi, Samuel, 116
- Han, Seung Hun Eddie, 11
Hassan, Fuad Mire, 179
Homan, Christopher M, 17
Hu, Tim Wenjie, 204
- Imam, Amina Abubakar, 37
Imam, Sukairaj Hafiz, 197
- Janssen, Maarten, 142
Jimmy, Edifon Emmanuel, 113
Justin, Bakoubolo Essowe, 264
- Kamaté, Panga Azazia, 1, 190, 248
Kampo, Adam Bouno, 248
Kansiwer, Ahoefa, 264
Karim, Nihal, 103
Kawalya, Deo, 211
Keita, Mamadou K., 17, 248
Khalid, Abubakar Auwal, 37
Kondoro, Alfred Malengo, 116
Kone, Emmanuel Elise, 1
- Leventhal, Michael, 1, 190
Louzan, Hamza, 103
Lundin, Jessica M., 103
- Maina, Samuel Chege, 64
Mireku, Prince, 44
Mohamed, Muhidin A., 179
Mohamed, Shafie Abdi, 179
Mohamed, Youssef, 11
Moshood, Kausar Yetunde, 228
Moslem, Yasmin, 281
Muhammad, Shamsuddeen Hassan, 171, 197
Mwebaze, Ernest Tonny, 204, 211
- Ngoufo, Grace Kevine, 171
Novikova, Jekaterina, 44
Nsumba, Solomon, 211

Nwokocha, Hannah Sopuruchi, 217
Nyong, Benjamin Okon, 56

O'Neill, Jacki, 64
Ochieng, Millicent, 64
Odeajo, Israel Olanrewaju, 113
Oduwole, Mardiyyah, 44
Ogunremu, Oluwapelumi, 217
Oko-odion, Terry, 217
Olajide, Oluwatosin, 44
Olufemi, Victor Tolulope, 228
Oluwademilade, Williams, 228
Ouattara, Aboubacar, 248
Ouma, Evelyn Nafula, 204, 211
Oyo-Ita, Emmanuel Ubene, 56
Öztürk, Yankı, 116

Pato, Julie, 264

Quinn, John, 204, 211

Ronen, Keshet, 64

Samil, Hadia Mohmmedosman Ahmed, 243

Schreiber, Richard Hans, 116
Ssentanda, Medadi E., 211

Tall, Madani Amadou, 1
Thieme, Anja, 64
Timibe, Notou Your, 264
Tiwari, Anuj, 217

Udeze, Nneoma C, 88
Ueno, Risa, 64
Umar, Hadiza Ali, 197

Voigt, Rob, 88

Walukagga, Patrick, 204
Wassie, Aman Kassahun, 281
Wei, Guohao, 103

Yimam, Seid Muhie, 197
Yusuf, Yusuf A., 179

Zhang, Ada, 103