Ibom NLP: A Step Toward Inclusive Natural Language Processing for Nigeria's Minority Languages

Oluwadara Kalejaiye[1], Luel Hagos Beyene[2,3], David Ifeoluwa Adelani[4,5,6], Mmekut-Mfon Gabriel Edet[7], Aniefon Daniel Akpan[8], Eno-Abasi Urua[9], Anietie Andy[1]

[1]Electrical Engineering and Computer Science, Howard University,
[2]AIMS Research and Innovation Centre, [3]NM-AIST, [4]Mila - Quebec AI Institute,
[5]McGill University, [6]Canada CIFAR AI Chair, [7]Korapay,
[8]National Institute for Nigerian Languages, [9]University of Uyo

Correspondence: anietie.andy@howard.edu

Abstract

Nigeria is the most populous country in Africa with a population of more than 200 million people. More than 500 languages are spoken in Nigeria and it is one of the most linguistically diverse countries in the world. Despite this, natural language processing (NLP) research has mostly focused on the following four languages: Hausa, Igbo, Nigerian-Pidgin, and Yorùbá (i.e $<1\%$ of the languages spoken in Nigeria). This is in part due to the unavailability of textual data in these languages to train and apply NLP algorithms. In this work, we introduce IBOM—a dataset for machine translation and topic classification in four Coastal Nigerian languages from the Akwa Ibom State region: Anaang, Efik, Ibibio, and Oro. These languages are not represented in Google Translate or in major benchmarks such as Flores-200 or SIB-200. We focus on extending Flores-200 benchmark to these languages, and further align the translated texts with topic labels based on SIB-200 classification dataset. Our evaluation shows that current LLMs perform poorly on machine translation for these languages in both zero-and-few shot settings. However, we find the few-shot samples to steadily improve topic classification with more shots.

1 Introduction

Significant progress has been made towards developing and applying Natural Language Processing (NLP) and Machine Learning (ML) algorithms for translating textual data for low resource African languages (Kuwanto et al.; Nwafor and Andy, 2022; Adelani et al., 2022c,a). However, so far, these NLP and ML algorithms have been applied to only a few low-resource African languages. This is in part due to the unavailability of textual data in some of these languages (Adelani et al., 2021a). Some African languages are not written; instead, they are orally passed down from one generation to the next, and they are not part of the educational system in their respective countries. The languages that receive attention are typically the most widely spoken, official or national languages (Adelani, 2025), which largely coincide with the top 50 African languages included in massively multilingual datasets (NLLB-Team, 2022; Conneau et al., 2023; Adelani et al., 2024).

Colonialism is in part responsible for the exclusion of these languages from the educational system. [1] In the colonial times, only a few languages were encouraged in the educational system in the colonized countries; thereby ensuring that these languages were considered prestigious in comparison to other languages. Even when the colonized countries became independent, this practice was maintained in the respective countries. This has led to the endangerment and near extinction of some of these languages.

Nigeria is the most populous African country, with a population of more than 200 million people [2]. There are more than 500 languages spoken in Nigeria [3], making Nigeria one of the most linguistically diverse countries in the world [4]. Despite the large number of languages spoken in Nigeria, very few of these languages have received sufficient attention as it relates to documentation and description. Nigerian languages are classified as either (a) "major" / "majority" or (b) "minor" / "minority". The "majority" languages are Hausa, Yoruba, and

[1] hhttps://www.goethe.de/prj/zei/en/art/22902448.html
[2] https://datacommons.org/place/country/NGA
[3] www.ethnologue.com
[4] https://www.weforum.org/stories/2023/04/worlds-most-multilingual-countries/

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 372–382

Igbo. These majority languages have been taught in schools for decades and have significantly been documented; however, in comparison, the "minority" languages have received scant attention, and little has been documented in these languages.

This work introduces IBOM—a new dataset consisting of two major NLP tasks (machine translation and topic classification) for four "*minority*" languages (Anaang, Efik, Ibibio, and Oro) in Nigeria by extending the massively multilingual Flores-200 (NLLB-Team, 2022) and SIB-200 (Adelani et al., 2024) datasets. Our evaluation shows that current LLMs perform poorly on machine translation for these languages in both zero-and-few shot settings. We find that leveraging cross-lingual transfer from M2M-100 (Fan et al., 2021) on religious parallel texts and a related language (Efik) improves performance for the other languages. On topic classification, we find that few-shot prompting steadily improve topic classification performance, and even exceed performance of supervised fine-tuning baseline with AfroXLMR (Alabi et al., 2022). Specifically, this work makes the following contributions:

(A) We develop IBOM-MT, a collection of translated sentences from English Flores-200 dataset to four low resource Nigerian languages (i.e. Ibibio, Efik, Anaang, and Oro) that are not represented on Google translate. To our knowledge, this is the first parallel language resource created for Anaang and Oro languages, respectively.

(B) We develop IBOM-TC, a topic classification dataset created by aligning IBOM-MT and SIB-200 topic labels.

(C) We apply several fine-tuned baselines and LLM prompting for machine translation and topic classification on these introduced benchmarks.

(D) We release the datasets from this work and make it available to the research community. [5]

2 Related Works

African machine translation data Flores evaluation datasets (Goyal et al., 2022; NLLB-Team, 2022) have emerged as critical resources in several low-resource languages. Beyond the large scale evaluation datasets, there have been several community-driven data collection for Nigerian languages including: MENYO-20k for Yorùbá (Adelani et al., 2021b), Igbo-English MT (Ezeani et al., 2020), MAFAND-MT (Adelani et al., 2022a) for Hausa and Nigerian-Pidgin, NollySenti (Shode et al., 2023)—a translated benchmark for movie sentiment, and IrokoBench (Adelani et al., 2025)—a translated benchmark for knowledge QA, natural language inference and math reasoning. The last two benchmarks only cover Hausa, Igbo, Yorùbá and Nigerian-Pidgin. Initiatives like the WMT Shared Task have also played an important role in boosting the evaluation of MT on African languages (Adelani et al., 2022b).

Available Ibom languages data In the understudied languages, only Efik has some available bilingual and monolingual data. JW300 (Agić and Vulić, 2019)—a multilingual parallel corpus based on Jehovah Witness publications contains over 331K parallel sentences with English. Other parallel resources includes SMOL (Caswell et al., 2025), and Gatitos dictionary (Jones et al., 2023). Aside bitexts, some monolingual data exists in large filtered Common Crawl data such as GlotCC (Kargaran et al., 2024) and FineWeb-2 (Penedo et al., 2025) but they often smaller in size.

3 Ibom languages characteristics

3.1 Geographical information

The Ibibio, Efik, Anaang, and Oro languages are predominantly spoken in Akwa Ibom State in the South-South geopolitical region of Nigeria. Figure 1 shows the map of the regions in Akwa Ibom State, Nigeria, where these four languages are spoken. This map indicates the indigenous languages of each Local Government Area in Akwa Ibom State. Ibibio, Anaang, and Oro are shown to be the major languages in Akwa Ibom State as observed by the number of Local Government Areas where the languages are indicated as indigenous languages. In addition, Efik is mainly spoken in the Local Government Areas of Akwa Ibom State which border the Cross River State. The following is information about each of these languages.

Ibibio The Ibibio language is primarily spoken in Akwa Ibom State and in some Local Government Areas in Cross River State, in Nigeria. Ibibio serves as the lingua franca of Akwa Ibom State and is spoken as a first language in fifteen Local Government Areas in Akwa Ibom State namely;

[5] https://huggingface.co/collections/howard-nlp/ibom-nlp

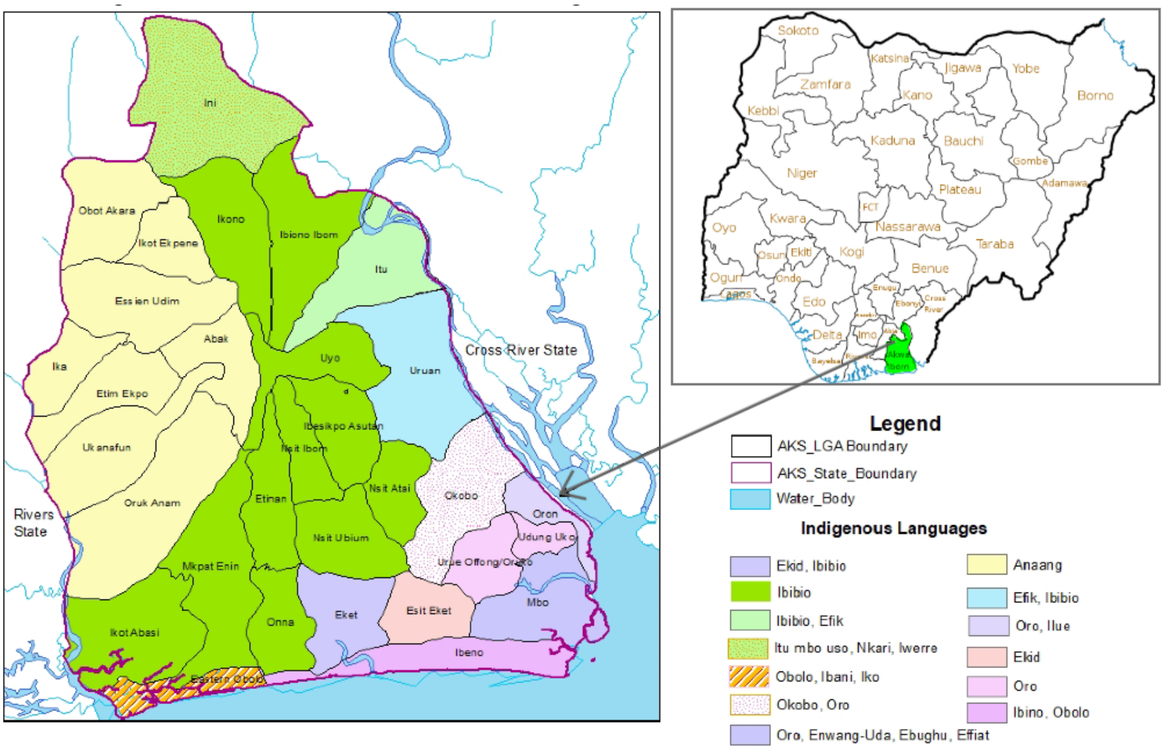

Figure 1: **Map showing the regions where the IBOM languages are spoken in Nigeria** including Anaang, Efik, Ibibio, and Oro. The languages are native to the South-South region of Nigeria, primarily spoken in the Akwa-Ibom state painted in Green.

Uyo, Itu, Uruan, Etinan, Nsit Ibom, Nsit Atai, Nsit Ubium, Ibesikpo Asutan, Ikono, Ini, Ikot Abasi, Mkpat Enin, Ibiono Ibom, Onna, and Eket (Urua, 2021). There are approximately 3.7 million speakers of Ibibio (Mensah et al., 2024).

Efik The Efik language is spoken in the following local government areas in Akwa Ibom State: Itu, Uruan, and Oron local government area, and Southwest Cameroun. Efik is also spoken in the following local government areas in Cross River State, Nigeria: Calabar Municipality, Calabar South, Akpabuyo, Bakassi, Odukpani, and parts of Akamkpa (Offiong and Ansa, 2013). Efik has approximately 3.5 million speakers. [6]

Anaang The Anaang language is primarily spoken in the North West part of Akwa Ibom State, Nigeria. Anaang is predominantly spoken in the following eight Local Government Areas in Akwa Ibom State: Abak, Ikot Ekpene, Essien Udim, Ukanafun, Etim Ekpo, Ika, Obot Akara, and Oruk Anam (Udoh, 2012). There are approximately 1,400,00 Anaang speakers. [7]

Oro The Oro language is predominantly spoken in the following Local Government Areas in Akwa Ibom State: Oron, Urue Offong/Oruko, Okobo, Udung Uko, and Mbo. There are more than 400,000 speakers of Oro. [8]

3.2 Linguistic features

All the languages belong to the Lower Cross branch of the Cross River Division of the (New) Benue-Congo family (Williamson and Bendor-Samuel, 1989; Urua, 2021). Ibibio, Efik, Anaang belong to the Efik-Ibibo sub-family while Oro is in a different sub-family (Nsíŋ Oro) Phonologically, all the languages are tonal and have a similar tonal system of high, low, downstepped and contour tones (Udoh, 2012; Ansa, 2017; Ukpe, 2018). The vowels, consonants, and syllable structure of these four languages are presented in Table 1.

Morphologically, Ibibio, Efik, and Anaang are considered agglutinating languages to a large extent and inflecting languages to some extent (Essien, 2008; MMensah, 2010; Offiong and Ansa, 2013). Syntactically, all languages have flexible Subject Verb Object (SVO) sentence structure as shown in the following examples:

[6] https://www.britannica.com/

[7] https://people.umass.edu/nconstan/representatives/

[8] https://www.101lasttribes.com/tribes/oron.html

English: I am going to school.
Ibibio: ami ñ-ka ufọkñwed.
Efik: ami n-ka ufọkñwed.
Anaang: ami n-ka ufọkñgwed.
Oro: ami n-ga uvọkñwid.

4 IBOM dataset

In this work, we used the data from the Flores-200 dataset (NLLB-Team, 2022), which is an evaluation benchmark that consists of sentences obtained from English Wikipedia covering various topics and domains. We used the data splits from the Flores-200 dataset i.e. DEV (997 sentences) and DEVTEST (1,012 sentences); we also collected 1,000 out of the 6,000 sentences in the NLLB-SEED training dataset. Finally, we extended the DEV and DEVTEST set for topic clasification based on SIB-200 recipe. We provide the data statistics for both tasks in Table 2.

The lead translator for each of the languages reviewed the translations in their respective languages for errors and made corrections where necessary.

4.1 IBOM-MT: Machine translation

For each language of focus in this work (i.e. Ibibio, Efik, Anaang, and Oro), we identified and worked with three (3) linguists who speak, studied, and reside in a State in Nigeria in which these languages are spoken. These linguists were the translators for the dataset collected in this work. Each of these linguists had at least a Bachelors degree in Linguistics. Some of the linguists had a PhD in Linguistics—one is a Full Professor of Linguistics and another is a Lecturer of Linguistics. For each language, one of the translators was the lead translator. Translators received appropriate remuneration.

Given that for each language, we have three translators and there are three sets of data i.e. DEV (997), DEVTEST (1,012), and training (1,000), for each language, one translator translated one set of data. When they were done with the translations, the lead translator for each of the languages reviewed the translations in their respective languages for errors and made corrections where necessary. The translation for this work was done over a period of four months. During this time period, the research team consisting of linguists, NLP experts, and students met weekly to discuss the progress of the project and resolve any issues that arose during the data collection process and afterwards.

4.2 IBOM-TC (Topic classification)

We extended SIB-200 topic classification dataset (Adelani et al., 2024) to the Ibom languages by aligning translated sentences with topic labels in English SIB-200. Since, SIB-200 only used DEV and DEVTEST portion of Flores-200 to develop a topic classification, the alignment was straightforward. We used the same script released by the SIB-200 authors to automatically align the translated sentences with English topic labels. Thus, we have IBOM-TC for the four Ibom languages.

5 Experimental setup

Here, we describe the experiments used to evaluate the four Ibom-NLP languages. We conduct an extensive evaluation of fine-tuning baselines and LLM prompting on two tasks: machine translation (MT) and topic classification (TC).

5.1 Fine-tuning baselines

5.1.1 Machine translation

We fine-tuned two massively multilingual MT models: M2M-100 (418M parameters) (Fan et al., 2021) and NLLB-200 (600M parameters) (NLLB-Team, 2022), covering many-to-many translation to/from 100 and over 200 languages, respectively. For ease of fine-tuning, we trained only the smaller parameter versions of the M2M-100 and NLLB-200 models. We fine-tuned the models for 3 epochs using an NVIDIA GH200 120GB GPU machine. We measure the performance of the MT models using three metrics: ChrF++ (Popović, 2017), BLEU (Papineni et al., 2002), and SSA-COMET (Li et al., 2025)—an extension of COMET (Rei et al., 2020) embedding-based metric to African languages.

2-stage MT fine-tuning We explore the 2-stage fine-tuning specifically for the MT task. Given the availability of medium-sized religious parallel data ($\sim 331K$ sentences) for English–Efik from MT560 (Gowda et al., 2021), we designed a 2-stage fine-tuning approach. In the first stage, we fine-tuned separately on English-to-Efik (`en-efi`) and Efik-to-English (`en-efi`) data. In the second stage, we fine-tuned on 1,000 parallel sentences for each language pair. Since the fine-tuning data is limited, we leverage the effectiveness of cross-lingual transfer to improve performance on related

Language	Vowels	Consonants	Syllable Structure	Tone Pattern
Ibibio	a, e, i, ị, o, ọ, u, ụ, Λ, ə	b, d, f, gh, h, k, kp, m, n, ñ, ñw, ny, p, s, t, w, y	N, V, CV, CGV (CrV), CVC, CVV, CVVC	High(H), Low(L), Downstepped(D), Rising(R), Falling(F)
Efik	a, e, i, ị, o, ọ, u, Λ	p, b, d, f, g, h, k, kp, kw, n, ny, ñ, ñy,m, n, p, r, s, t, w, y	N, V, CV, CVC, CGV (CrV), CVV, CVVC	High(H), Low(L), Downstepped (D), Rising (R), Falling (F)
Anaang	a, e, i, o, ọ, u, ụ	b, ch, d, f, gh, gw, j, k, kp, kw, l, m, n, ñ, ñw, ny, p, r, s, t, w, y	N, V, CV, CVV, CVC, CVVC	High (H), Low (L), Downstepped (D), Rising (R), Falling (F)
Oro	a, e, ẹ, i, ị, o, ọ, u	b, d, f, g, gb, gh, gw, j, k, kp, kw, l, m, n, ny, ñ, ñw, r, s, t, v, w, y, z	N, V, CV, CGV (CrV), CVC, CVV, CVVC	High (H), Low (L), Downstepped (D), Rising (R), Falling (F)

Table 1: **Phonological characteristics of the Ibom languages**

Split	IBOM-MT (# sents)	IBOM-TC (# sents)
Train	1000	701
Dev	997	99
Test	1012	204
Total	3,009	1,004

Table 2: **IBOM dataset**. We report the data statistics for both IBOM-MT and IBOM-TC tasks

Ibom languages that are not covered in existing multilingual models.

5.1.2 Topic classification

We fine-tuned six multilingual encoder models: XLM-R (Conneau et al., 2020), Glot500 (Imani et al., 2023), AfriBERTa (Ogueji et al., 2021), Serengeti (Adebara et al., 2023), AfroXLMR (Alabi et al., 2022), and AfroXLMR-61L (Adelani et al., 2024). XLM-R and Glot500 cover 100 and 511 languages, respectively, while the remaining four encoders are African-centric. The multilingual models were created using two approaches: (1) pre-training from scratch, and (2) adapting from pre-trained encoders such as XLM-R via continued pre-training. XLM-R, Serengeti, and AfriBERTa were pre-trained from scratch, although Serengeti covers significantly more languages—primarily from Africa—i.e., 500 vs. 100 (XLM-R) and 11 (AfriBERTa). AfroXLMR was created through multilingual adaptive fine-tuning (i.e., continued pre-training) for 20 widely spoken African languages, while AfroXLMR-61L extends this to 61 African languages. Glot500 first performed vocabulary extension before continued pre-training on 511 languages and has one of the widest coverages of low-resource languages.

IBOM languages coverage in LLMs Despite the broad language coverage of some multilingual encoders such as Glot500, AfroXLMR-61L, and Serengeti, only one or two of the Akwa Ibom languages were included during pre-training. Specifically, Glot500 and AfroXLMR-61L include only Efik, while Serengeti includes both Efik and Ibibio. Given the linguistic closeness of the Akwa Ibom languages, we believe they can benefit from cross-lingual transfer for both tasks.

5.2 LLM prompting

We performed zero-shot and few-shot evaluations on both machine translation and topic classification tasks. We primarily focused on proprietary models, as they have been shown to achieve better overall performance than open models and to provide broader coverage of low-resource languages—for example, Gemini claims to support 400 undisclosed languages (Comanici et al., 2025). We evaluated the GPT-4.1, o4-mini, Gemini 2.0 Flash, and Gemini 2.5 Flash (Thinking) models. For these four models, we conducted zero-shot and few-shot evaluations (5-shot, 10-shot, and 20-shot), where each 5-shot set consists of the first five examples from the training data split.

6 Results

6.1 MT results

Table 3 shows the MT results for fine-tuning baselines and LLM prompting.

Overall zero-shot poor results by LLMs While some LLMs, such as Gemini, claim to support over 400 languages, we observe extremely low ChrF++ scores for the Akwa Ibom languages—particularly in the en→xx direction. The performance is slightly better in the xx→en direction, especially for Anaang (anw), which achieved a ChrF++ score

Model	*eng → X* anw	efi	ibb	oro	Avg.	*X → eng* anw	efi	ibb	oro	Avg.	AVG
Encoder-Decoder											
M2M-100FT	14.7	12.7	10.5	10.7	12.2	35.1	25.0	23.9	23.5	26.9	19.5
NLLB-200 FT	27.4	16.7	15.9	17.4	19.4	32.2	24.0	22.8	23.6	25.7	22.5
M2M-100 FT: 2-stage	27.6	36.2	24.5	18.2	26.6	37.6	32.0	27.7	24.0	30.3	28.5
NLLB-200 FT: 2-stage	22.0	35.5	20.6	18.2	24.1	37.6	34.6	29.3	25.8	31.8	28.0
Decoder-only											
LLM eval: 0-shot											
GPT-4.1	25.9	23.0	21.1	16.1	21.5	37.1	27.3	26.6	28.3	29.8	25.7
o4-mini (`thinking`)	26.7	21.1	19.4	10.5	19.4	36.2	26.4	26.8	29.0	29.6	24.5
Gemini 2.0 Flash	25.8	31.1	24.2	15.1	24.1	**38.8**	**38.6**	**32.1**	**29.2**	**34.7**	29.3
Gemini 2.5 Flash (`thinking`)	17.7	31.4	24.3	19.6	23.3	31.8	36.2	30.7	25.2	31.0	27.1
LLM eval: 5-shots											
GPT-4.1	37.6	24.3	22.4	20.3	26.1	37.6	29.2	27.3	**29.2**	30.8	28.5
o4-mini (`thinking`)	28.1	20.3	19.5	18.7	21.7	34.3	26.0	26.8	28.4	28.9	25.3
Gemini 2.0 Flash	**28.0**	31.0	24.0	20.4	25.9	38.5	35.8	30.5	22.2	31.6	28.8
Gemini 2.5 Flash (`thinking`)	26.0	44.4	25.2	22.2	29.5	31.8	36.2	30.7	25.2	31.0	**30.2**
LLM eval: other-shots											
Gemini 2.5 Flash (`10-shots`)	25.6	**44.5**	25.5	**24.9**	30.1	26.7	35.3	28.7	25.2	29.0	29.6
Gemini 2.5 Flash (`20-shots`)	35.9	42.6	**27.1**	20.9	**31.6**	11.4	20.1	20.4	20.8	18.2	24.9

Table 3: **MT Performance of fine-tuned models and LLM prompting on Ibom-MT**. We report ChrF++ score. We highlighted the best result in each experimental setup in gray.

of 37.1 with GPT-4.1 and 38.8 with Gemini 2.0 Flash. Surprisingly, thinking models such as o4-mini and Genini 2.5 Flash [9] performed worse than their non-thinking counterparts.

2-stage fine-tuning provides a stronger baseline Fine-tuning with only 1,000 parallel sentences yields very low performance, with ChrF++ scores below 25 in both translation directions. Leveraging cross-lingual transfer by first fine-tuning on 331K en⇄efi examples, followed by 1K examples, results in a significant performance boost—especially for Efik and related languages (Anaang and Ibibio). Oro showed a more modest improvement compared to the other languages due to being linguistically farther away from the others. We find the 2-stage approach to outperform all zero-shot transfer results, except for the xx→en direction using Gemini 2.0 Flash.

Few-shot prompting is more effective for 5-shots and 10-shots We find that 5-shots improve performance across all LLMs. The Genini 2.5 Flash (thinking) model achieved the best overall score based on the ChrF++ metric, outperforming the other LLMs. Thus, we assess the performance of Genini 2.5 Flash with more shots. We observe that 10-shots and 20-shots further improve performance in the en→xx direction. However, in the xx→en direction, 20-shots lead to a significant drop in performance. In general, we find few-shot to be more useful for Efik than the other Ibom languages achieving up to 44.5 ChrF++ score, we attribute this to the resource level of Efik since it has more monolingual data than the other languages (Gowda et al., 2021).

Inconsistency in MT metrics for the Ibom languages We find that, at times, ChrF++ and BLEU scores do not fully align. We further evaluated using the African-centric COMET metric (Rei et al., 2020), SSA-COMET (Li et al., 2025) and observed that it appears to be more consistent in the en→xx direction than in the xx→en direction. We believe it is important to invest in metric development alongside MT development for low-resource languages. SSA-COMET results are reported in Table 6.

Since the metrics do not fully align, we provide human direct assessment evaluation (Human DA) for some of the results in (§6.2)

6.2 Human evaluation for MT results

To verify the automatic metrics, we performed human evaluation on 50 test examples and the MT outputs based on the best two systems identified by ChrF++ in Table 3 i.e. Gemini 2.5 Flash 10-shots and M2M-100 FT (2-stage). We only evaluated the Ibom languages to English direction for this evaluation. We collected human direct assessment (DA) using the same tool used by the AfriCOMET (Wang et al., 2024). For each of the languages, we re-

[9] We set the thinking budget to "-1"

Model	*eng → X* anw	efi	ibb	oro	Avg.	*X → eng* anw	efi	ibb	oro	Avg.	AVG
Encoder-Decoder											
M2M-100	5.4	3.1	0.8	3.0	3.1	12.7	5.6	4.1	3.8	6.6	4.8
NLLB-200	6.0	2.0	1.5	3.2	3.2	8.1	4.0	2.7	2.7	4.4	3.8
M2M-100 FT: 2-stage	6.5	8.8	2.7	4.0	5.5	14.4	9.2	5.2	3.8	8.2	6.8
NLLB-200 FT: 2-stage	2.1	8.5	1.3	1.4	3.3	13.2	12.2	6.3	3.4	8.8	6.1
Decoder-only											
LLM eval: 0-shot											
GPT-4.1	5.4	3.2	2.0	2.5	3.3	13.1	5.5	3.7	5.9	7.1	8.7
o4-mini (thinking)	4.4	2.8	1.9	0.5	2.4	10.4	4.5	3.3	5.2	5.9	4.1
Gemini 2.0 Flash	5.2	7.2	4.0	3.1	4.9	13.5	13.9	7.7	6.4	10.4	7.6
Gemini 2.5 Flash (thinking)	2.3	25.3	7.7	4.6	9.9	8.9	20.4	10.4	7.0	11.7	10.8
LLM eval: 5-shots											
GPT-4.1	11.6	3.5	2.3	3.2	5.1	12.0	5.2	3.5	5.2	6.5	5.8
o4-mini (thinking)	6.6	1.3	2.2	2.7	3.2	8.1	4.4	3.5	5.2	5.3	4.3
Gemini 2.0 Flash	5.2	7.2	4.0	3.1	4.9	13.5	13.9	7.7	6.4	10.4	7.6
Gemini 2.5 Flash (thinking)	2.7	14.7	5.2	4.9	6.9	10.2	15.3	16.6	9.3	12.8	9.9
LLM eval: other-shots											
Gemini 2.5 Flash (10-shots)	12.0	22.8	4.4	6.8	11.5	4.7	17.8	11.4	13.0	11.7	11.6
Gemini 2.5 Flash (20-shots)	11.0	11.4	5.6	5.2	8.3	1.4	3.9	3.3	3.3	3.0	5.6

Table 4: **MT Performance of fine-tuned models and LLM prompting on Ibom-MT**. We report BLEU score. We highlighted the best result in each experimental setup in gray.

Language	Gemini 2.5 Flash 10-shots)	M2M-100 2-stage FT	Ave.
anw	9.44	31.04	20.24
efi	51.11	71.27	61.19
ibb	16.8	9.81	13.11

Table 5: **Human Direct Assessment (DA) score** of the best two MT results

cruited three annotators who are bilingual native speakers of the languages.

After annotation, we only make use of annotators ratings if it has a high spearman correlation with another annotator, and the spearman correlation is more than 0.5. Out of the four languages, only Oro did not meet this criteria (<0.2), so, we excluded it from the final evaluation in Table 5. For the others, they are between 0.5 and 0.65

Table 5 shows the Human DA results where most annotators prefer the output of M2M-100 2-stage to that of Gemini 2.5 Flash (10-shots) for the Anaang and Efik languages with more than +20 points. However, for Ibibio, annotators slighly prefer Gemini 2.5 Flash. This evaluation highlights the weakness of the current evaluation metrics for many low-resource languages. Surprisingly, we find the human evaluation to slightly correlate with the SSA-COMET (Li et al., 2025) evaluation metric in Table 6 where SSA-COMET judged Gemini 2.5 Flash to be better than M2M-100 2-stage only for Ibibio, while for other languages, it gave very similar scores for the two models.

6.3 Topic classification results

Table 7 shows the result for the TC using six fine-tuned multilingual encoders and four LLMs.

African-centric encoders excels the other multilingual encoders We find that African-centric models such as AfroXLMR-61L and AfroXLMR achieve significantly better performance than the XLM-R model, which does not support many African languages. Although AfroXLMR-61L supports only Efik, it can leverage cross-lingual transfer to improve performance on other Akwa-Ibom languages. While Serengeti covers an additional language, Ibibio, it still performs worse than the AfroXLMR (-61L) variants—likely due to its smaller parameter size or the curse of multilinguality, as it covers 500 low-resource languages.

Fine-tune baselines is better than LLMs in zero-shot settings Overall, we find that the best fine-tuned baseline, AfroXLMR-61L, delivered better overall performance than prompting LLMs in zero-shot settings. However, we find Genini 2.5 Flash to be competitive with fine-tuned models that have seen more than 700 training examples. Nevertheless, there remains a large performance gap compared to the English language, which achieves up to 92.7 points with o4-mini.

Leveraging few-shot is highly effective for Gemini LLMs For 5-shot settings, Gemini 2.0 Flash and Genini 2.5 Flash improved performance by

Model	Size	*eng → X*					*X → eng*					AVG
		anw	efi	ibb	oro	Avg.	anw	efi	ibb	oro	Avg.	
Encoder-Decoder												
M2M-100	480M	6.5	8.0	-1.2	0.5	3.45	38.9	32.9	29.9	26.2	32.0	17.7
NLLB-200	600M	39.9	43.4	29.5	36.3	37.3	38.8	30.3	29.5	28.8	31.9	34.6
M2M-100 FT: 2-stage		34.5	49	35.3	24.6	35.9	42.1	41.9	37.3	28.0	37.3	36.6
NLLB-200 FT: 2-stage		34.1	46.1	38.8	36.4	38.9	44.9	47.9	38.8	31.8	40.9	39.9
Decoder-only												
LLM eval: 0-shot												
GPT-4.1	–	25.8	26.9	27.7	22.5	25.7	42.5	40.1	41.6	33.2	39.4	32.5
o4-mini (`thinking`)	–	27.1	18.8	18.1	19.2	20.8	28.1	28.5	29.5	28.5	28.7	24.7
Gemini 2.0 Flash	–	21.9	25.0	24.0	28.1	24.5	25.5	28.3	27.9	28.1	27.5	26.1
Gemini 2.5 Flash (`thinking`)	–	36.2	48.9	45.9	33.3	41.1	41.6	50.4	45.7	41.7	44.9	43.0
LLM eval: 5-shots												
GPT-4.1	–	22.1	26.4	24.0	26.6	24.8	41.8	30.7	32.9	37.2	35.7	30.2
o4-mini (`thinking`)	–	17.0	20.6	16.8	19.6	18.5	24.2	25.6	27.8	25.6	25.8	22.2
Gemini 2.0 Flash	–	22.1	24.7	23.5	24.5	23.7	24.0	27.4	26.8	26.7	26.2	25.0
Gemini 2.5 Flash (`thinking`)	–	34.5	38.6	45.3	37.9	39.1	23.3	29.5	31.1	26.1	27.5	33.3
LLM eval: other-shots												
Gemini 2.5 Flash (`10-shots`)	–	34.1	50.5	45.2	38.8	42.2	20.2	25.2	26.6	24.0	24.0	33.1
Gemini 2.5 Flash (`20-shots`)	–	34.0	51.0	44.1	32.5	40.4	20.1	22.7	21.6	20.4	21.2	30.8

Table 6: **MT Performance of fine-tuned models and LLM prompting on Ibom-MT**. We report SSA-COMET score.

Model	eng	anw	efi	ibb	orx	Avg.
Encoder models						
XLM-R	**91.8**	65.0	57.5	54.9	46.1	52.8
AfriBERTa	80.6	59.1	65.1	59.4	58.9	61.5
Serengeti	86.4	67.4	59.1	55.9	55.4	57.1
Glot-500	82.6	51.8	38.2	37.9	35.2	37.1
AfroXLMR	90.7	64.7	67.0	61.5	60.7	63.0
AfroXLMR-61L	90.4	69.6	71.3	66.6	**66.5**	68.1
Decoder-only						
LLM eval: 0-shot						
GPT-4.1	89.2	60.8	44.1	42.1	50.0	49.3
o4-mini	92.7	57.8	41.7	47.6	49.5	49.2
Gemini 2.0 Flash	87.7	60.3	67.6	61.8	61.8	62.9
Gemini 2.5 Flash	87.8	70.1	76.5	74.0	51.0	67.9
LLM eval: 5-shots						
GPT-4.1	87.3	56.9	45.1	41.7	47.6	47.8
o4-mini	85.3	55.9	42.2	41.2	46.1	46.4
Gemini 2.0 Flash	84.8	66.2	73.5	66.7	57.8	66.1
Gemini 2.5 Flash	89.2	73.0	77.0	77.0	58.8	71.4
LLM eval: other-shots						
Gemini 2.5 (10-sh)	89.2	72.6	78.4	79.4	57.4	71.9
Gemini 2.5 (20-sh)	88.7	**73.5**	**79.4**	**80.9**	65.2	**74.8**

Table 7: **Topic classification performance of fine-tuned models and LLM prompting on Ibom-TC**. We report accuracy metric. We highlighted the best result in each experimental setup in gray.

+3.2 and +3.5 points, respectively, over their zero-shot prompting. Similarly, increasing the number of shots to 10 and 20 for Genini 2.5 Flash led to further improvements of +4.0 and +6.9 points, respectively, compared to the zero-shot result. While the Gemini LLMs show performance gains, we find that GPT-4.1 and o4-mini did not benefit significantly from the few-shot examples. This suggests that Gemini is likely more multilingual than the OpenAI models, although further investigation is needed.

7 Conclusion

In this paper, we develop new datasets for Akwa-Ibom languages, which are truly low-resource Nigerian languages. While many AfricaNLP papers have focused on the big three national languages of Nigeria—Hausa, Igbo, and Yorùbá, our paper is one of the first to extend to other low-resource languages in Nigeria. We performed evaluation on both machine translation (a text generation task) and topic classification (a natural language understanding task) by extending Flores-200 and SIB-200 to these languages. Our evaluation shows that LLMs are difficult to adapt for these low-resource languages for the machine translation task, however, we find a more positive adaptation with few-shot prompting for topic classification using Genini 2.5 Flash LLM.

In the future, we plan to extend the training data size for the Ibom-MT languages to further boost performance, and to extend COMET evaluation support for these languages. We hope our paper will encourage more investment in NLP beyond the top-10 most spoken languages in Africa.

8 Limitations

Our work has some limitations. In this section we address these limitations.

(A) This study focused on four low resourced Nigerian languages i.e. Ibibio, Efik, Anaang, and Oro. While this work has contributed to the development of parallel language resources for these languages, the results from the experiments conducted in this work may not generalize to other languages.

(B) There are more than 500 languages spoken in Nigeria; however, our study covers only four of these languages. We hope that this work and the strategies used for collecting data in these four languages will inform linguists who speak and study the various languages in Nigeria and NLP experts to collect translation data in these languages, conduct NLP research, and share the data and findings from their research with the NLP research community.

(C) For the LLM prompting experiments, we evaluated GPT-4.1, o4-mini, Gemini 2.0 Flash, and Gemini 2.5 Glash (Thinking) models. In the future, we will investigate the performance on other LLMs.

9 Acknowledgments

David Adelani acknowledges the funding of IVADO and the Canada First Research Excellence Fund through IVADO R3AI Regroupement Grant. Luel acknowledges the support of Carnegie Corporation of New York provided through AIMS-RIC.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. SERENGETI: Massively multilingual language models for Africa. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537, Toronto, Canada. Association for Computational Linguistics.

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, and 1 others. 2022a. A few thousand translations go a long way! leveraging pre-trained models for african news translation. In *Conference of the North-American-Chapter-of-the-Association-for-Computational-Linguistics (NAAACL)-Human Language Technologies, JUL 10-15, 2022, Seattle, WA*, pages 3053–3070. Association for Computational Linguistics.

David Ifeoluwa Adelani. 2025. Natural language processing for african languages. *arXiv preprint arXiv:2507.00297*.

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 H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, and 42 others. 2021a. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Md Mahfuz Ibn Alam, Antonios Anastasopoulos, Akshita Bhagia, Marta R Costa-jussà, Jesse Dodge, Fahim Faisal, Christian Federmann, Natalia Fedorova, Francisco Guzmán, and 1 others. 2022b. Findings of the wmt'22 shared task on large-scale machine translation evaluation for african languages. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 773–800.

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

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen H Muhammad, Peter Nabende, and 1 others. 2022c. 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.

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

David Ifeoluwa Adelani, Dana Ruiter, Jesujoba O. Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021b. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research*

Track, pages 61–75, Virtual. Association for Machine Translation in the Americas.

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 Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

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

Asuquo I. Essien-Eyo A. Ansa, S. 2017. *(Fonetik: nwed ukpepnkpo abanade mme uyom iko Efik.* . Excel Publishers, Calabar.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Koulako Moussa Doumbouya, and 1 others. 2025. Smol: Professionally translated parallel data for 115 under-represented languages. *arXiv preprint arXiv:2502.12301*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

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, Online. Association for Computational Linguistics.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Okon Essien. 2008. *A Grammar of the Ibibio Language*. University Press PLC, Ibadan.

Ignatius Ezeani, Paul Rayson, Ikechukwu Onyenwe, Chinedu Uchechukwu, and Mark Hepple. 2020. Igbo-english machine translation: An evaluation benchmark. *arXiv preprint arXiv:2004.00648*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

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

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Alex Jones, Isaac Caswell, Ishank Saxena, and Orhan Firat. 2023. Bilex rx: Lexical data augmentation for massively multilingual machine translation. *arXiv preprint arXiv:2303.15265*.

Amir Hossein Kargaran, François Yvon, and Hinrich Schütze. 2024. Glotcc: An open broad-coverage commoncrawl corpus and pipeline for minority languages. *Advances in Neural Information Processing Systems*, 37:16983–17005.

Garry Kuwanto, Eno-Abasi Urua, Priscilla Amuok, Shamsuddeen Hassan Muhammad, Anuoluwapo Aremu, Verrah Otiende, Loice Nanyanga, Teresiah Nyoike, Aniefon Akpan, Nsima Udouboh, and 1 others. Mitigating translationese in low-resource languages: The storyboard approach.

Senyu Li, Jiayi Wang, Felermino DMA Ali, Colin Cherry, Daniel Deutsch, Eleftheria Briakou, Rui Sousa-Silva, Henrique Lopes Cardoso, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. Ssa-comet: Do llms outperform learned metrics in evaluating mt for under-resourced african languages? *arXiv preprint arXiv:2506.04557*.

Eyo Mensah, Kirsty Rowan, and Mfon Ekpe. 2024. The representation of people in the ibibio anthroponymic system: A socio-onomastic investigation. *Languages*, 9(6):188.

E. MMensah. 2010. On efik prefixing morphology. *The Humanities Journal: An International Journal of Research and Development*, 1(1):49–71.

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

Ebelechukwu Nwafor and Anietie Andy. 2022. A survey of machine translation tasks on nigerian languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6480–6486.

Offiong Ani Offiong and Stella Ansa. 2013. The efik language: A historical profile. *Research on humanities and Social Sciences*, 3(21):23–31.

Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. 2021. Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic. 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.

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

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

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.

Iyanuoluwa Shode, David Ifeoluwa Adelani, JIng Peng, and Anna Feldman. 2023. NollySenti: Leveraging transfer learning and machine translation for Nigerian movie sentiment classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 986–998, Toronto, Canada. Association for Computational Linguistics.

I Udoh. 2012. Anaang phonology: A descriptive sketch. *Germany: LAP LAMBERT Academic Publishing*.

Q Ukpe. 2018. Aspects of efik phonology.

Eno-Abasi Urua. 2021. Tbibio phonetics and phonology.

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

Kay Williamson and J Bendor-Samuel. 1989. The niger-congo languages. *J. Bendor-Samuel. Lanham, MD: University Press of America*.