Toucan: Many-to-Many Translation for 150 African Language Pairs

AbdelRahim Elmadany[ξ,⋆] **Ife Adebara**[ξ,⋆] **Muhammad Abdul-Mageed**[ξ,λ]

[ξ]The University of British Columbia   [λ] Invertible AI

{a.elmadany,ife.adebara,muhammad.mageed}@ubc.ca

Abstract

We address a notable gap in Natural Language Processing (NLP) by introducing a collection of resources designed to improve Machine Translation (MT) for low-resource languages, with a specific focus on African languages. First, We introduce two language models (LMs), `Cheetah-1.2B` and `Cheetah-3.7B`, with 1.2 billion and 3.7 billion parameters respectively. Next, we finetune the aforementioned models to create Toucan, an Afrocentric machine translation model designed to support 156 African language pairs. To evaluate Toucan, we carefully develop an extensive machine translation benchmark, dubbed AfroLingu-MT, tailored for evaluating machine translation. Toucan significantly outperforms other models, showcasing its remarkable performance on MT for African languages. Finally, we train a new model, spBLEU[1K], to enhance translation evaluation metrics, covering 1K languages, including 614 African languages. This work aims to advance the field of NLP, fostering cross-cultural understanding and knowledge exchange, particularly in regions with limited language resources such as Africa. The GitHub repository for the *Toucan* project is available at https://github.com/UBC-NLP/Toucan.

1 Introduction

Machine Translation (MT) is an important technology that bridges linguistic divides and enables communication across the globe. Although transfer learning methods (Zoph et al., 2016) employing multilingual language models (mLM) (Xue et al., 2021a; Liu et al., 2020) have benefited the field specially for languages with limited resources (Liu et al., 2023), a significant gap remains for many African languages. In particular, although a handful of mLMs finetuned for MT for African languages (Adelani et al., 2022; Oladipo et al., 2023;

Figure 1: Toucan is a powerful MT model, proficiently trained on 156 language pairs. It covers a wide spectrum of 43 African languages as well as Arabic, English, and French

Jude Ogundepo et al., 2022), these pioneering works only serve 31 out of the 2,000+ languages of the African continent. This lack of coverage means the issues of language barriers, the risk of language extinction, and the under-representation of diverse communities in global conversations (Koehn and Knowles, 2017). Language barriers in particular pose significant challenges, hindering the smooth exchange of ideas, information, and cultural nuances across diverse linguistic landscapes. In contexts characterized by limited resources, where access to proficient human translators is constrained, MT has the potential to be a transformative remedy that offer unparalleled advantages in dismantling linguistic obstacles and promote heightened cross-cultural comprehension.

In this paper, we address this gap by presenting a family of pretrained models that we also finetune for machine translation, an extensive evaluation benchmark, and an evaluation metric with wider

---

⋆ Authors contributed equally.

13189

coverage of African languages. By introducing these, we aim to contribute to the advancement of low-resources language technology especially African languages, unlocking new possibilities for cross-cultural understanding and knowledge exchange. Our main contributions can be summarized as follow:

1. **AfroLingu-MT.** We introduce AfroLingu-MT, a benchmark for African languages comprising 156 language pairs. To the best of our knowledge, AfroLingu-MTis the largest African MT benchmark to date. We design it to rigorously evaluate and advance the state of MT for a diverse host of African languages, addressing addressing a critical need in this area.

2. **Pretrained LLM.** We present a new sequence-to-sequence large language model (LLM) that covers 517 African languages and 10 foreign languages including - Arabic, English, French, German, Greek, Italian, Portuguese, Russian, Spanish, and Turkish. The model is available in two sizes, 1.2B and 3.7B parameters. We refer to these models as "Cheetah-1.2B" and "Cheetah-3.7B".

3. **Toucan models.** A versatile many-to-many family of MT models capable of translating between 46 different languages (43 African languages and the three non-Indigenous major languages in Africa: Arabic, English, and French). Our models cover 156 translation language pairs.

4. **Comprehensive evaluation.** We offer a comprehensive comparison between generative and sequence-to-sequence LMs by evaluating a wide range of models on our AfroLingu-MT benchmark under both few-shot and full finetuning scenarios. For our evaluation we use both existing models and also introduce new models as we outline next.

   spBLEU[1K]. We introduce spBLEU[1K], a sentencepiece model that covers $1,003$ languages, including 614 African languages, designed to improve translation evaluation quality. Our model aims to address the limitations found in traditional BLEU score evaluations (Peters et al., 2018) and the FLORES spBLEU model (Goyal et al., 2022) by expanding coverage to a vast array of languages

that have historically been underrepresented in translation models and benchmarks.

The rest of the paper is organized as follows: In Section 2, we discuss related work for MT benchmarks, models, and tools. In Section 3 and Section 4, we describe AfroLingu-MT evaluation benchmark and Toucan models respectively. We provide details of the empirical evaluation in Section 5, including experimental setup, baselines, and evaluation metrics. We present the results and discussion in Section 6. Finally, we conclude in Section 7, and outline a number of limitations, ethics and use cases for our work in Section 8 and Section 9 respectively.

2 Literature Review

MT has seen remarkable advancements, particularly in supporting underrepresented languages such as those spoken in Africa (Team et al., 2022; Jude Ogundepo et al., 2022; Adelani et al., 2022). The development of MT tools for African languages is vital for promoting linguistic diversity, fostering cross-cultural communication, and driving socio-economic progress. In this review, we briefly cover LLMs supporting African languages and related topics such as datasets and evaluation tools. Also, we provide additional details in Section A in the Appendix.

2.1 Datasets and Benchmarks

A major hindrance to developing MT models for African languages is the scarcity of data (Adda et al., 2016; Adebara and Abdul-Mageed, 2022). While web data collection typically provides ample fine quality datasets for high-resource languages, the corpora for many African languages obtained through similar methods are often constrained in both size and quality (Kreutzer et al., 2021; Alabi et al., 2020). To address these issues, a number of benchmarks for both training and evaluation have been developed. Notable among them are FLORES-101 (Goyal et al., 2021), FLORES-200 (Goyal et al., 2022), Menyo-20 (Adelani et al., 2021), Lafand-MT (Adelani et al., 2022), and Salt (Akera et al., 2022).

2.2 Models and Tools

Models such as "No Language Left Behind" (NLLB) (Team et al., 2022) cater to translation needs in 200 languages, including 23 African ones. Similarly, M2M-100 (Fan et al., 2020)

Type	Model	Lang/Total
CLM	ALMA (Xu et al., 2023)	English only
	ALMA-MT (Xu et al., 2023)	0/6
	Bloomz (Muennighoff et al., 2022)	14/101
	Bloomz-MT (Muennighoff et al., 2022)	Unknown/46
	Llama-2 (Touvron et al., 2023b)	English only
Seq2Seq	Afri-mT5 (Adelani et al., 2022)	17/17
	AfriTeVa (Oladipo et al., 2023)	10/10
	Aya (Üstün et al., 2024)	15/101
	Mistral (Jiang et al., 2023)	English only
	mT0 (Muennighoff et al., 2022)	14/101
	mT5 (Xue et al., 2020)	12/101
	Cheetah (Adebara et al., 2024)	517/527

Table 1: Models with African languages represented. **Lang/Total**: the number of African languages covered by the model/the number of total language covered.

and AfriTeVa (Jude Ogundepo et al., 2022) support 17 and 10 African languages, respectively, utilizing text-to-text architectures. AfriTeVa-V2 (Oladipo et al., 2023) has enhanced support for 16 African languages with improved quality pretraining data. AfroMT5 (Adelani et al., 2022) and AfriMBART (Adelani et al., 2022) each cover 17 African languages. Additionally, Cheetah (Adebara et al., 2024) extends its support to 517 African languages and 10 widely spoken global languages. Decoder-only models, exemplified by the Generative Pretrained Transformer (GPT) (OpenAI, 2023; Brown et al., 2020) and Llama (Touvron et al., 2023a,b), demonstrate outstanding performance across various tasks such as text comprehension, language translation, and content generation but usually fall short in terms of coverage of African languages as we will show in this work. Aya (Üstün et al., 2024) is a massively multilingual generative language model that supports instructions following in 101 languages including 24 African languages. Aya has been shown to outperform models like mT0 (Muennighoff et al., 2023) and BLOOMZ (Muennighoff et al., 2023) on a wide variety of automatic and human evaluations despite covering double the number of languages.

2.3 Evaluation Metrics

Prominent evaluation metrics such as BLEU (Papineni et al., 2002) and ChrF (Popović, 2015a) focus on assessing n-gram correspondence between model translations and human references, favoring precision. Meanwhile, METEOR (Banerjee and Lavie, 2005) emphasizes both precision and

recall by considering synonyms, stemming, and word order. TER (Agarwal and Lavie, 2008) measures edit distance between machine-generated and reference texts, while COMET (Rei et al., 2020; Stewart et al., 2020) leverages contextual embeddings for semantic similarity evaluation. These metrics aid in benchmarking machine translation (MT) systems, guiding improvements for higher quality translations. Goyal et al. (2022) introduced a new metric, SentencePiece BLEU (spBLEU), which extends coverage to 101 languages, including 23 African languages. Similarly, AfriCOMET Wang et al. (2023) is tailored for 17 African languages, addressing their unique challenges in MT evaluation. These developments signify ongoing efforts to create more robust and inclusive evaluation tools for MT across diverse languages worldwide.

3 AfroLingu-MT Benchmark

In this section, we describe our data collection and construction procedures, along with the features of our comprehensive benchmark for evaluating African MT systems, AfroLingu-MT. To create AfroLingu-MT , we execute several steps, encompassing data curation, quality evaluation, pair selection, determination of translation directions, and specification of the output format. We now explain each of these steps.

3.1 Data Collection

Our collection comprises data from a total of 43 datasets, encompassing 84 unique language pairs derived from 46 different languages. We also develop a new manually translated dataset useful for

evaluation in the government domain. In all, the data cover 43 African languages from five language families domiciled in 29 African countries. We also include Arabic, English, and French, since these are widely spoken in Africa. Table 2 in the Appendix provides detailed information about our collected data, including the number of pairs and data-points (i.e, examples) for each dataset. Table C.1 (Appendix C) on the other hand has details about each of the 46 languages in our dataset. These tables serve as a valuable resource for understanding the breadth and depth of our datasets, ensuring transparency and facilitating further research in the field of MT for African languages. We also translate into Yoruba a portion of the Arab-acquis data (Habash et al., 2017) and include it in the benchmark. We refer to this data henceforth as Legal-genre.

Datasets	#Pairs	#Examples
AraOPUS-20 (Nagoudi et al., 2022)	3	2.01M
Bamanakan Lexicon (Bamba, 2016)	2	5, 978
Corpora Ethiopia (Teferra Abate et al., 2018)	7	77, 739
ENGLISH-AKUAPEM TWI (azu, 2021)	1	25, 421
English-Luganda (muk, 2021)	1	15, 021
FFR-Dataset	1	136, 098
Flores-200 (Costa-jussà et al., 2022),	237	842
French-Ewe-fongbe (deg, 2020)	2	70, 000
Gamayun(Öktem et al., 2021)	7	75, 000
Global Voices (Tiedemann, 2012)	9	399, 178
Gnome (Tiedemann, 2012)	14	850, 068
Gourmet-MT	5	263, 882
Gov-ZA(Lastrucci et al., 2023; mar, 2023)	54	638, 737
Horn-MT	15	21, 848
Igbo-NLP	1	10, 008
Lafand-MT (Adelani et al., 2022)	20	5.2M
Legal-genre (ours)	1	3, 580
Masakhane Wazobia	2	66, 324
Menyo-20k (Adelani et al., 2021)	1	20, 100
Multi-paracrawl (Tiedemann, 2012)	3	46, 478
NCHLT(Tiedemann, 2012)	7	746, 646
Open Subtitles (Tiedemann, 2012)	1	44, 703
Opusinfopanki (Tiedemann, 2012)	1	47, 220
Opusmemat (Tiedemann, 2012)	1	139, 260
Paracrawl	2	147, 396
PidginUNMT_corpus	1	2, 101
Salt (Akera et al., 2022)	15	25, 000
TED (Reimers and Gurevych, 2020)	6	15, 401
Tico-19 (Tiedemann, 2012)	18	7, 740
Umsuka eng - zul Corpus (Mabuya et al., 2021)	1	10, 701
Vuk'uzenzele (Lastrucci et al., 2023; mar, 2023)	55	66, 318
Wikimedia (Tiedemann, 2012)	21	132, 213
Xhosanavy (Tiedemann, 2012)	1	49, 981
XNLI	1	1, 000
Yoruba Proverbs	1	5, 144

Table 2: Detailed description of datasets in AfroLingu-MT benchmark.

3.2 Data Quality

To ensure high quality, we follow a rigorous manual review process for each dataset. This involves a thorough examination of the original paper associated with each dataset to gain a clear understanding of its data collection methodology. Following this review, we classify the datasets into four quality

tiers: Synthetic, Human evaluated, Gold, and Unknown. (1) Synthetic datasets consist of translations generated solely by machine translation models, without any human quality evaluation. We exclude these datasets from further consideration and they are not part of our final collection. (2) Human evaluated quality translations are also generated by MT models, but typically undergo correction by human reviewers to improve their quality. (3) Gold quality translations are either directly translated or evaluated by human experts. Datasets in this category are also sourced from domains with a lower likelihood of containing noisy data. (4) Unknown quality datasets either lack associated publications or detailed information about their data quality and collection processes. Additionally, for certain languages, we go beyond paper analysis and conduct specific evaluations of translation quality. In particular, we manually assess translation pairs between English and languages such as Yoruba, Hausa, and Nigerian Pidgin.

3.3 Pair Selection and Translation Directions

We exclude "Unknown" and "Syntheic" datasets, retaining only "Human Evaluated" and "Gols" quality data and standardizing the language codes to ISO-693. Our objective is to facilitate development of robust MT models capable of translating between a wide range of African languages as well as Arabic, English, and French.[1] To achieve this, we adopt a multifaceted approach essentially enabling many-to-many translation. For instance, the user may need to specify only the target language thus allowing for versatile translation possibilities including translation from any language into English. This results in the selection of 156 distinct language pairs.

3.4 Data Selection

As depicted in Table 2 (Appendix), there is significant variation in data distribution among the language pairs, with some having a substantial number of data points and others having much fewer. To create a balanced training dataset, especially since we are targeting many-to-many translation, we aim to obtain data from each translation direction for each language pair. First, we maintain the original dataset splits where one exists from source; otherwise, we divide the language pairs into training, development, and testing datasets in an 80/10/10 ratio. Next we sample 5K/50/200 data points for

---

[1]Arabic, English and French are widely spoken in Africa.

```
{"langcode":"nyn-ach", "instruction":"Translate the following text to Acholi language. Return only the
translated sentence only. Do not repeat the instruction.", "input":"Bakakora omukoro gw'okuhendera
emishomo yaabo, orwakashatu oruhwaire.", "output":"Gubed ki yub me kwero tyeko kwan i ceng adek
ma okato"}
{"langcode":"ach-lug","instruction":"Translate the following text to Luganda language. Return only the
translated sentence only. Do not repeat the instruction.", "input":"Cal pa dako man onya i nyonyo pa
muno calo mac oro.", "output":"Ebifaananyi bye bibadde biyitangana ku mikuttu emigattabantu."}

{"langcode":"eng-teo","instruction":"Translate the following text to Ateso language. Return only the
translated sentence only. Do not repeat the instruction.","input":"The Joint Anti-Terrorist Task Force
car was never recovered.","output":"Mam aponi kodumunai emotoka loka Egurupu loka itunga luitijiete
itunga lukodwaratau."}
```

Figure 2: Examples from AfroLingu-MT benchmark train set.

training, development, and testing, respectively, for each language pair direction, such as English-to-Afrikaans (eng-afr) and Afrikaans-to-English (afr-eng). This approach enables us to facilitate translation for 46 unique languages. In addition, when dealing with abundant data, we ensure there is no overlap between source and target data points for each pair in either of the directions. However, for pairs with limited data, we swap the data points between the two directions to augment the dataset. AfroLingu-MT contains a total of $620,573$ parallel data points, with $586,261$ allocated for training, $7,437$ for development, and $26,875$ for the test data split. As mentioned, it comprises translations for 46 languages derived from 156 language pairs. We show the data distribution for each language pair in Table C.2 (Appendix C).

3.5 Data Format

Following the Alpaca style (Taori et al., 2023), we organize our dataset as follows:
Lang-code: This specifies the ISO-639-3 codes for both the source and target languages.
Instruction: This field provides a concise description of the task.
Input: This field contains the source text intended for translation.
Output: This includes the text translated into the target language.

We save each data point on a separate line in JSONL format. Figure 2 shows examples of translating from source language to target language in AfroLingu-MT.

3.6 AfroLingu-MT in Comparison

Table 3 presents a comparison between AfroLingu-MT and existing benchmarks. The table high-

Benckmark	Lang/Total
FLORES-101 (Goyal et al., 2021)	23 / 101
FLORES-200 (Goyal et al., 2022)	23 / 200
Menyo-20 (Adelani et al., 2021)	1/2
Lafand-MT (Adelani et al., 2022)	16/18
Salt (Akera et al., 2022)	5/5
AfroLingu-MT (ours)	43/46

Table 3: AfroLingu-MT benchmark (ours) in comparison with with other benchmarks with notable African language coverage. **Lang/Total** column describes the number of African languages comparing with the covered languages in the language models.

lights the total number of supported languages and language pairs in each benchmark. As Table 3 shows, compared to other benchmarks, AfroLingu-MT doubles the number of African languages covered and has an order of magnitude higher coverage in terms of language pairs/translation directions.

4 Toucan Models

In this work, we develop a number of many-to-many Afrocentric machine translation models dubbed Toucan. For this purpose, we first pretrain a number of Afrocentric sequence-to-sequence models that serve as the foundational backbone for our proposed machine translation Toucan models.

4.1 Cheetah Backbone LMs

To effectively train a MT language model for African languages, it is crucial to start with a powerful, Afrocentric pretrained language model. For this purpose, we select Cheetah (Adebara et al., 2024), a recently introduced SoTA model with extensive coverage encompassing 517 African languages. One limitation of Cheetah, however, is that

it is available only in a base architecture, featuring 580M parameters. Given our objective to develop a large-scale language model for machine translation capabale of serving 156 directions, this base model does not fully meet our requirements.

To address this limitation, we embark on training larger and more expansive Afrocentric sequence-to-sequence models. We focus on two sizes: one model with 1.2B parameters and another with 3.7B parameters. We refer to the new models "Cheetah-1.2B" and "Cheetah-3.7B", respectively, to reflect their enhanced capabilities and parameter scale. These models represent a significant advancement in our efforts to improve machine translation for African languages, offering greater capacities in handling the rich linguistic nuances of African languages. **Cheetah Pertaining**. To train the new Cheetah models, we utilize the same pre-training dataset employed in training the original Cheetah-base model (Adebara et al., 2024). This strategic choice ensures consistency in the foundational data across models, enabling the advanced Cheetah-1.2B and Cheetah-3.7B versions to build upon the rich linguistic diversity captured in the original dataset. We refer to (Adebara et al., 2024) for more information about the pretraining data of Cheetah models. We employ a learning rate of 0.01, a batch size of $1,024$ sequences, and a maximum sequence length of $1,024$. Each model undergoes pretraining for 1 million steps. The training process is conducted on Google Cloud TPU with 128 cores (v3 − 128) provided by the TensorFlow Research Cloud (TFRC). We provide additional details on pretraining in Section B in the Appendix.

4.2 Toucan Finetuning

We finetune the vanilla Cheetah-base model as well as the newly proposed architectures, Cheetah-1.2B and Cheetah-3.7B, on our AfroLingu-MT. As explained in Section 3, this dataset is the largest and most diverse African MT dataset. We refer to our new models fintuned for MT as Toucan-base, Toucan-1.2B and Toucan-3.7B. We provide more information about model finetuning in Section 5.2.

5 Empirical Evaluation

5.1 Evaluation Settings

We evaluate AfroLingu-MT in diverse scenarios, including both finetuning and zero-shot settings. **(1)** We conduct comparative analyses between **multilingual and Africa-centric pretrained language models by finetuning** these models on

our AfroLingu-MT dataset. Specifically, we utilize mT5 (Xue et al., 2020) and mT0 (Muennighoff et al., 2022) as our representative multilingual pretrained models. In contrast, we compare these with African-specific models such as Afri-mT5 (Adelani et al., 2022), AfriTeVa (Oladipo et al., 2023), and Cheetah (Adebara et al., 2024). **(2)** We also evaluate the performance of **instruction-following LLMs on AfroLingu-MT in a zero-shot** setting, employing a prompt-based technique. We utilize LLMs that have been trained on general instructions, such as LLaMA-2 (Touvron et al., 2023b), Mistral (Jiang et al., 2023), and ALMA (Xu et al., 2023). Additionally, we assess LLMs that have been specifically trained on machine translation data, including Bloomz-MT (Muennighoff et al., 2022) and mT0-XXL-MT (Muennighoff et al., 2022). Table 1 shows the comparison between these models based on the African languages represented.

5.2 Experimental Setup

We finetune all models on AfroLingu-MT for 5 epochs. For base and large architectures, we finetune the models using a learning rate of $5e^{-5}$, a batch size of 8, and a maximum sequence length of 512 tokens. For XL model (3.7B parameters), we use a learning rate $2e^{-5}$, a batch size of 2, and a maximum sequence length of 256 tokens. During the training process, we implement a linear learning rate scheduler, incorporating a warm-up phase that accounts for 10% of the total training steps.

In all finetuning experiments, we rigorously select the best-performing checkpoint for each model based on performance in the respective development set. We then report and analyze the performance of each model on the corresponding test set.

5.3 Evaluation Metrics

In this work, we present the performance outcomes of our proposed models as well as the baseline models each evaluated independently on the AfroLingu-MT benchmark. This evaluation employs three pertinent metrics specific to machine translation. These metrics are: SentencePiece BLEU (i.e., spBLEU) (Goyal et al., 2022), word-based Character n-gram F-score (i.e., ChrF++) (Popović, 2015b), and AfriCOMET (Wang et al., 2023). These metrics have been selected for their effectiveness in assessing the quality of machine translations from various perspectives, including lexical accuracy and fluency. We also introduce a wider coverage metric

Type	Models	#params	DEV				TEST			
			spBLEU	spBLEU[1K]	ChrF++	AfriCOMET	spBLEU	spBLEU[1K]	ChrF++	AfriCOMET
						Zero-Shot Setting				
CLM	ALMA 7B MT★	7B	1.78	2.15	12.3	25.10	1.93	2.24	12.31	24.69
	Bloomz 7B MT★	7B	2.02	2.63	11.71	23.89	2.29	2.85	12.04	23.67
	Llama-2 7B Chat	7B	1.87	2.20	13.42	23.25	1.79	2.17	13.29	22.81
	Mistral 7B Instruct v2	7B	2.00	2.49	14.79	22.33	1.89	2.40	14.95	22.14
	ALMA 13B MT★	13B	2.14	2.52	12.24	27.16	2.00	2.33	12.06	26.89
	Llama-2 13B Chat	13B	1.07	1.4	9.07	18.67	0.94	1.35	9.09	18.2
S2S	mT0 XXL MT★	13B	5.88	7.48	19.79	33.82	6.09	7.67	20.09	33.71
						Finetuned Setting				
CLM	Llama-2-7B-MT	7B	2.86	3.39	13.72	24.67	2.97	3.54	13.89	24.17
	Mistral-7B-MT	7B	2.04	2.61	12.57	20.21	1.95	2.57	12.66	21.97
S2S	Afri-mT5 Base	580M	3.23	3.43	16.33	29.69	3.28	3.47	16.46	30.01
	AfriTeVa Base	229M	0.00	0.01	8.12	12.18	0.00	0.01	8.04	12.13
	AfriTeVa V2 Base	428M	3.61	3.86	16.94	31.74	3.96	4.18	17.45	31.91
	mT0 Base	580M	10.96	12.12	28.55	48.51	11.66	12.88	29.48	48.40
	mT5 Base	580M	11.54	12.89	28.89	51.47	11.93	13.22	29.46	51.78
	🦜 Toucan Base (ours)	580M	<u>16.65</u>	<u>17.6</u>	<u>34.09</u>	<u>60.42</u>	<u>17.33</u>	<u>18.2</u>	<u>34.56</u>	<u>60.21</u>
	AfriTeVa Large	745M	3.35	3.51	17.05	29.41	3.31	3.42	17.14	29.25
	AfriTeVa V2 Large	1B	6.06	6.17	22.86	32.67	6.24	6.31	23.23	32.76
	mT0 Large	1.2B	12.03	14.94	30.93	50.22	12.10	12.01	30.2	50.24
	mT5 Large	1.2B	13.28	14.21	30.21	50.89	13.33	14.26	30.34	50.79
	🦜 Toucan 1.2B (ours)	1.2B	**18.30**	**19.39**	**35.89**	**62.58**	**18.78**	**19.73**	**36.41**	**62.36**
	mT0 XL	3.7B	14.56	16.30	35.54	53.81	14.21	16.32	34.34	53.76
	mT5 XL	3.7B	15.56	16.03	35.16	53.54	15.45	15.95	35.16	53.85
	🦜 Toucan 3.7B (ours)	3.7B	**22.11**	**22.53**	**38.91**	**66.63**	**22.67**	**23.15**	**39.53**	**66.73**

Table 4: Performance on our AfroLingu-MT benchmark across both the zero-shot and full finetuning scenarios. For all causal model, we use prompting to induce translations. For sequence-to-sequence models, we use target language-based prefixes. We offer results both for development (i.e., DEV) and test (i.e., TEST) datasets. ★Notably, these models are trained on multilingual MT datasets. Text highlighted in Bold Green indicates the highest scores across all settings. Text highlighted in Bold Orange indicates a new evaluation metric (ours).

Models	#params	DEV				TEST			
		spBLEU	spBLEU[1K]	ChrF++	AfriCOMET	spBLEU	spBLEU[1K]	ChrF++	AfriCOMET
NLLB-200-1.3B	1.3B	14.59	14.74	29.88	53.06	15.39	15.40	30.80	53.16
🦜 Toucan 1.2B (ours)	1.2B	**20.83**	**21.93**	**38.36**	**62.69**	**21.29**	**22.37**	**38.80**	**62.15**

Table 5: A comparison between NLLB and Toucan on 59 language pairs both models support.

modeled after spBLEU, dubbed spBLEU[1K], as we explain next.

5.3.1 spBLEU[1K]

Employing the BLEU metric for evaluating translations, particularly in the context of low-resource languages, is suboptimal due to its fundamental reliance on n-gram overlap. This reliance significantly impacts the metric's effectiveness, as it is heavily influenced by the specific tokenization method used. Notably, employing a more aggressive tokenization strategy can lead to artificially inflated BLEU scores Goyal et al. (2022). To address this, Goyal et al. (2022) proposed a novel metric, SentencePiece BLEU (i.e., spBLEU), designed to measure and analyze the performance of transla-

tions across 101 languages. This approach involves training a new SentencePiece-based tokenizer using monolingual data for 101 languages, replacing the default tokenizer typically used in SacreBLEU, known as 'mosestokenizer' (Post, 2018). This innovation aims to standardize the tokenization process, thus providing more accurate and comparable translation performance metrics across many languages.

Significantly, the spBLEU metric covers merely 23 out of the 43 languages present in our AfroLingu-MT benchmark. To address this limitation, we adopt a methodology similar to that of Goyal et al. (2022). Namely, we develop a new SentencePiece tokenizer that utilizes 1000+ monolingual data sources.

Category	spBLEU[1K]
Arabic ⟷ XX	17.27
XX ⟶ Arabic	16.77
Arabic ⟶ XX	17.78
Arabic ⟶ (not supported)	NA
Arabic ⟶ XX (supported)	17.78
English ⟷ XX	25.16
XX ⟶ English	21.48
English ⟶ XX	28.83
English ⟶ XX (not supported)	32.00
English ⟶ XX (supported)	25.22
French ⟷ XX	17.42
XX ⟶ French	17.77
French ⟶ XX	17.62
French ⟶ XX (not supported)	12.54
French ⟶ XX (supported)	23.07
African ⟷ African	22.43
French ⟶ African (not supported)	38.01
French ⟶ African (supported)	18.69
Total supported languages	**26.57**
Total unsupported languages	**22.36**

Table 6: The performance of our Toucan 3.7B model varies based on language categories on TEST dataset. We also compare the performance between the languages in our benchmark that spBLEU supports - 23 supported and not supported languages

Data. We collect monolingual data covering $1,003$ languages, including 614 African languages, 53 Indigenous American languages, and the remainder spanning the most resource-rich languages worldwide. We use a diverse data source, encompassing Wikipedia, Wikibooks, the Bible, newspapers, and common web sources. Additionally, we utilize the MADLAD dataset (Kudugunta et al., 2023), which covers 419 languages. A list with the $1,003$ we cover is available at Toucan.

Training a SentencePiece Model (SPM). One significant challenge is the uneven availability of monolingual data across various languages. This disparity is especially acute for low-resource languages, which often suffer from a lack of comprehensive coverage in subword units and may not possess a sufficiently large corpus of sentences to ensure a broad and diverse representation of content. To address this issue and enhance the training of our new SPM, we adopt a temperature upsampling technique similar to the methodology described in Conneau et al. (2019).

Integrating with SacreBleu. We integrate this newly created SPM into SacreBLEU, resulting in the formulation of our more inclusive metric spBLEU[1K]. Our metric is thus designed to provide a more comprehensive evaluation across a broader range of languages, including those that are under-

represented in existing metrics such as spBLEU.

6 Results and Discussion

spBLEU[1K] Metric. The results indicate that our new metric, spBLEU[1K], enhances the translation scores by an average of 0.74 and 0.62 BLEU points on the development and test datasets, respectively. We also evaluate the performance of our best model, Toucan-3.7B, based on language categories. Table 6 presents a comparison between spBLEU and our new evaluation metric across different categories of translation directions in our data. Results demonstrate that the two metrics are almost identical in the shared languages (the languages that both metrics supported), however, the new metric improves the translation results in languages not supported by spBLEU.

Exploring different pretraining settings allows us to derive unique insights. Examples of insights that can be gleaned from Table 4 include:

Sequence-to-sequence models with wider coverage perform better MT. Unsurprisingly, the findings show that sequence-to-sequence models pretrained with more languages enable better MT performance on our dataset. For example, Toucan, which supports 517 African languages and ten of the most spoken languages worldwide, outperforms all other models. On the other hand, AfriTeVa (which is pretrained on only ten languages) has the lowest performance. However, AfriTeVa-v2 outperforms Afri-MT5 even though it supports fewer languages. We assume that the size of the LM may plays a role here.

Models of the same size finetuned on more language pairs tend to perform better translations. Among the models we evaluate in zero-shot, four are already finetuned on external MT parallel data. These are ALMA-7B-MT, Bloomz-7B-MT, ALMA-13B-MT, and MT0-XXL-MT (13B). While we do not see a clear pattern for the 7B model size, our results for the 13B do carry a pattern: Results show that the MT0-XXL-MT model (13B), which is finetuned on pairs from 46 languages gives significantly better translations than ALMA-13B-MT, which is finetuned on pairs from six languages (i.e., 5.34 points in spBLEU[1k] on TEST).

Larger models perform better. Again, unsurprisingly, we observe that larger models outperform

Lang pair	Aya 13B		Toucan 3.7B (ours)		Lang pair	Aya 13B		Toucan 3.7B (ours)	
	spBLEU	ChrF++	spBLEU	ChrF++		spBLEU	ChrF++	spBLEU	ChrF++
afr→eng	31.1	55.2	**43.83**	**64.63**	eng→yor	4.8	19.6	**5.47**	**23.81**
amh→eng	**19.2**	**44.6**	16.50	40.10	eng→zul	**11.4**	**39.7**	6.54	35.42
eng→afr	27.8	51.8	**29.77**	**54.61**	hau→eng	19.3	42.7	**24.66**	**47.50**
eng→amh	**11.9**	**23.9**	3.14	19.74	ibo→eng	16.7	40.3	**19.76**	**42.28**
eng→hau	11.0	38.4	**18.24**	**43.13**	nso→eng	17.3	40.5	**26.88**	**48.88**
eng→ibo	**10.4**	**32.3**	9.93	31.27	sna→eng	16.6	39.4	**18.72**	**41.03**
eng→nso	**6.1**	**29.5**	2.53	13.58	som→eng	16.8	40.3	**20.65**	**43.98**
eng→sna	**5.6**	**33.2**	0.30	4.80	sot→eng	20.7	44.2	**26.72**	**48.80**
eng→som	7.3	35.0	**8.75**	**36.28**	swa→eng	23.0	47.4	**32.47**	**56.20**
eng→sot	**16.3**	**42.4**	14.63	36.93	xho→eng	20.5	43.7	**24.99**	**47.42**
eng→swa	19.5	46.7	**24.85**	**51.16**	yor→eng	11.1	34.6	**14.21**	**36.12**
eng→xho	**8.5**	**36.3**	6.08	33.80	zul→eng	20.5	44.4	**26.17**	**48.69**

Table 7: Comparing Aya 13.B with Toucan on their intersection of pairs. Toucan outperforms Aya on 16 of 28 pairs. Comparison is done with Flores devtest splits.

smaller ones. For example, models with 1.2B parameters outperform base models by approximately *two* points spBLEU[1k], whereas the larger models with 3.7B outperform both the base and 1.2B parameter models by approximately *five* and *three* points spBLEU[1k], respectively. Additionally, we note that our model, Toucan, outperforms the second best performing model (i.e., mT5) base, 1.2B, and 3.7B models by 4.98, 7.47, and 7.2 points spBLEU[1k] on TEST, respectively.

Additionally, we compare our model, Toucan-1.2B, to the Facebook's NLLB model (Team et al., 2022; Costa-jussà et al., 2022). Again, we find Toucan-1.2B outperforming NLLB-200-1.3B by 6.96 points in spBLEU[1K], as shown in Table 5.

Toucan outperforms Aya We compare the performance of Toucan with Aya (Üstün et al., 2024). We use results for Aya as they appear in the paper, hence, we do not compute spBLEU[1k] results in this analysis. Although Aya is a 13B parameter model, significantly larger than Toucan 3.7B, we report better performance in 16 of 28 pairs. In Table 7, we show the performance of Toucan and Aya on Flores200.

7   Conclusion

In this paper, we introduce a suite of resources aimed at enhancing MT for low-resource African languages. We present Cheetah-1.2B and Cheetah-3.7B, with 1.2B and 3.7B parameters, respectively. We further finetune these models into versatile many-to-many model capable of translating between 46 different languages, including 43 African languages and the three major non-Indigenous languages in Africa: Arabic, English,

and French. We also introduce AfroLingu-MT as the largest African MT benchmark to date. We provide a comprehensive comparison between various LLMs by evaluating a wide range of models on our AfroLingu-MT benchmark. Finally we extend spBLEU - spBLEU[1K] - a sentencepiece-based evaluation metric covering 1,003 languages, including 614 African languages. This aims to enhance translation evaluation quality, particularly for languages historically underrepresented in translation models and benchmarks.

8   Limitations

We can identify a number of limitations that are relevant to our work, as follows:

- Even though we cover the largest number of African languages in our MT models, compared to previous research, there is still a long list of African languages that lack MT support. It will take the community more work to develop new parallel datasets for these languages so that they can be supported. We believe that our new Cheetah models can be valuable in this regards, since they have a coverage of 517 languages and can be easily finetuned on new languages once parallel datasets are available.

- Another limitation is that metrics such as AfriCOMET can only cover 17 African languages, although we developed the spBLEU[1K] metric that allowed us to evaluate on our dataset in a dependable way since it is based on wide coverage vocabulary. Again, new parallel datasets can be helpful for extending COMET to more African languages.

9 Ethics Statement and Wider Impacts

Our model, Toucan, is rooted in Afrocentric NLP principles, prioritizing the technological needs of African communities. We anticipate that Toucan will not only benefit speakers of the supported languages but also aid researchers in African languages, including anthropologists and linguists. Below, we highlight some potential use cases for Toucan and discuss its broader impacts:

Addressing the technological disparity in approximately 90% of the world's languages, which disproportionately affects native speakers, Toucan focuses specifically on Africa. As the first massively multilingual pre-trained language model (PLM) developed for African languages and their variants, it encompasses knowledge of 517 African languages, making it the largest model for African NLP to date.

Toucan facilitates enhanced access to critical information for the African community through Indigenous African languages. Particularly beneficial for individuals with limited proficiency in other languages, this capability has the potential to foster greater global connectivity.

Toucan supports language preservation efforts for numerous African languages, many of which have not been utilized in NLP tasks before. We believe it can encourage the continued use of these languages across various domains and spur the development of language technologies tailored to their specific needs.

Despite their versatility, language models like Toucan can be susceptible to misuse. Developed using publicly available datasets that may contain biases, we strive to conduct analyses and diagnostic case studies to assess our model's performance. However, our investigations are not exhaustive and do not guarantee the absence of bias in the data, especially considering limited access to native speakers of most covered languages.

In summary, Toucan represents a significant step forward in Afrocentric NLP, addressing technological disparities, fostering language preservation, and promoting responsible deployment of language technologies in African contexts.

References

2020. *Parallel text dataset for Neural Machine Translation (French -> Fongbe, French -> Ewe)*. Zenodo.

2021. *ENGLISH-AKUAPEM TWI PARALLEL CORPUS*. Zenodo.

2021. *An English-Luganda parallel corpus*. Zenodo.

2023. *The South African Gov-ZA multilingual corpus*. Zenodo.

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélene Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, et al. 2016. Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science*, 81:8–14.

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, Dublin, Ireland. Association for Computational Linguistics.

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

David Adelani, Jesujoba Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen Muhammad, Guyo Jarso, Oreen Yousuf, Andre Niyongabo Rubungo, Gilles Hacheme, Eric Peter Wairagala, Muhammad Umair Nasir, Benjamin Ajibade, Tunde Ajayi, Yvonne Gitau, Jade Abbott, Mohamed Ahmed, Millicent Ochieng, Anuoluwapo Aremu, Perez Ogayo, Jonathan Mukiibi, Fatoumata Ouoba Kabore, Godson Kalipe, Derguene Mbaye, Allahsera Auguste Tapo, Victoire Memdjokam Koagne, Edwin Munkoh-Buabeng, Valencia Wagner, Idris Abdulmumin, Ayodele Awokoya, Happy Buzaaba, Blessing Sibanda, Andiswa Bukula, and Sam Manthalu. 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 Adelani, Dana Ruiter, Jesujoba Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of the 18th Biennial Machine Translation Summit (Volume 1: Research Track)*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

Abhaya Agarwal and Alon Lavie. 2008. Meteor, M-BLEU and M-TER: Evaluation metrics for high-correlation with human rankings of machine translation output. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 115–118, Columbus, Ohio. Association for Computational Linguistics.

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.

Benjamin Akera, Jonathan Mukiibi, Lydia Sanyu Naggayi, 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.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina Espana-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: The case of Yorùbá and Twi. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2754–2762.

Moussa Bamba. 2016. Bamanankan lexicon.

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.

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. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

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*.

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.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

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.

Nizar Habash, Nasser Zalmout, Dima Taji, Hieu Hoang, and Maverick Alzate. 2017. A parallel corpus for evaluating machine translation between Arabic and European languages. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 235–241, Valencia, Spain. Association for Computational Linguistics.

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*.

Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. 2022. AfriTeVA: Extending ?small data? pretraining approaches to sequence-to-sequence models. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 126–135, Hybrid. Association for Computational Linguistics.

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

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta

13199

Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*.

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.

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.

Richard Lastrucci, Isheanesu Dzingirai, Jenalea Rajab, Andani Madodonga, Matimba Shingange, Daniel Njini, and Vukosi Marivate. 2023. Preparing the vuk'uzenzele and ZA-gov-multilingual South African multilingual corpora. In *Proceedings of the Fourth workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 18–25, Dubrovnik, Croatia. Association for Computational Linguistics.

Shudong Liu, Xuebo Liu, Derek F. Wong, Zhaocong Li, Wenxiang Jiao, Lidia S. Chao, and Min Zhang. 2023. kNN-TL: k-nearest-neighbor transfer learning for low-resource neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1891, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation.

Rooweither Mabuya, Jade Abbott, and Vukosi Marivate. 2021. Umsuka english - isizulu parallel corpus.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

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.

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*.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2022. TURJUMAN: A public toolkit for neural Arabic machine translation. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 1–11, Marseille, France. European Language Resources Association.

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

OpenAI. 2023. Gpt-4 technical report.

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.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Maja Popović. 2015a. 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.

Maja Popović. 2015b. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395. 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, Belgium, Brussels. 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. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

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.

Craig Stewart, Ricardo Rei, Catarina Farinha, and Alon Lavie. 2020. COMET - deploying a new state-of-the-art MT evaluation metric in production. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 78–109, Virtual. Association for Machine Translation in the Americas.

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.

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

Solomon Teferra Abate, Michael Melese, Martha Yifiru Tachbelie, Million Meshesha, Solomon Atinafu, Wondwossen Mulugeta, Yaregal Assabie, Hafte Abera, Binyam Ephrem, Tewodros Abebe, Wondimagegnhue Tsegaye, Amanuel Lemma, Tsegaye Andargie, and Seifedin Shifaw. 2018. Parallel corpora for bi-directional statistical machine translation for seven Ethiopian language pairs. In *Proceedings of the First Workshop on Linguistic Resources for Natural Language Processing*, pages 83–90, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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).

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. 2023a. Llama: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Marek Masiak, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Hamam Mokayede, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, 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, Ayinde Hassan, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-Azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Samuel Njoroge, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Tosin Adewumi, 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 Brian, Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, and Pontus Stenetorp. 2023. Afrimte and africomet: Empowering comet to embrace under-resourced african languages.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models.

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*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021a. mt5: A massively multilingual pre-trained text-to-text transformer.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021b. mT5: A massively multilingual

pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Alp Öktem, Eric DeLuca, Rodrigue Bashizi, Eric Paquin, and Grace Tang. 2021. Congolese swahili machine translation for humanitarian response.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

Appendices

A   Literature Review

The recent years have witnessed a substantial evolution in LLMs, marked by numerous breakthroughs across various applications such as MT and text summarization. Particularly in MT, significant advancements have occurred, with a notable emphasis on supporting underrepresented languages, including those spoken in Africa (Team et al., 2022; Jude Ogundepo et al., 2022; Adelani et al., 2022). The development of MT tools tailored for African languages holds significant importance in promoting linguistic diversity, facilitating cross-cultural communication, and driving socio-economic progress within the region (Adebara et al., 2024).

In this section, we explore the current landscape of LLMs supporting African languages, covering topics such as datasets and benchmarks, instruction fine-tuning, evaluation metrics, challenges, and practical applications. It examines key contributions and advancements within this dynamic field, providing insights into its ongoing development and potential impact.

Datasets and Benchmarks One of the primary hindrances to developing MT models for African languages is the scarcity of data (Adda et al., 2016; Adebara and Abdul-Mageed, 2022). For many high resource languages, collecting data from the web often yields large and high quality datasets. However, resulting corpora for many African languages using similar methods are often limited in size and quality (Kreutzer et al., 2021; Alabi et al., 2020). Furthermore, texts from religious domains dominate the data landscape with most datasets coming from bibles (McCarthy et al., 2020) and other religious documents (Agić and Vulić, 2019). To address these issues, a handful of benchmarks for both training and evaluation have been developed. Notable among them are FLORES-101 (Goyal et al., 2021), FLORES-200 (Goyal et al., 2022), Menyo-20 (Adelani et al., 2021), Lafand-MT (Adelani et al., 2022), and Salt (Akera et al., 2022).

Models and Tools A few African languages have benefited from the recent advancement of LM. We now describe models and tools that support African languages. (1) No Language Left Behind (NLLB) is a suite of open-sources models capable of delivering evaluated, high-quality translations directly between 200 languages including 23 African lan-

guages (Team et al., 2022). (2) M2M-100 supports 100 languages including 17 African languages (Fan et al., 2020). (3) AfriTeVa (Jude Ogundepo et al., 2022) supports ten African languages using a T5-style model. (4) AfriTeVa V2 (Oladipo et al., 2023) supports 16 African languages using a better quality pre-training data. (5) AfroMT5 (Adelani et al., 2022) and AfriMBART (Adelani et al., 2022) each support 17 African languages. (6) T5 (Text-To-Text Transfer Transformer) based models such as mT5 (Xue et al., 2021b), mT0 and (Muennighoff et al., 2022). (7) Cheetah (Adebara et al., 2024), supports 517 African languages and 10 widely spoken languages in the world.

Decoder-only models have also shown remarkable improvements across multiple tasks. Now, we describe some of these modes. (1) Generative Pre-trained Transformer (GPT) (OpenAI, 2023; Brown et al., 2020), a transformer architecture model, pre-trained on expansive datasets, resulting in its proficiency in generating coherent, contextually rich text. (2) Llama (Touvron et al., 2023a,b) represents a state-of-the-art language model designed for nuanced natural language understanding and generation. LLama excels in tasks ranging from comprehensive text comprehension to intricate language translation and sophisticated content creation. Its versatility positions it as a pivotal tool within the realm of artificial intelligence, particularly in contexts requiring nuanced linguistic analysis. These models exhibit prowess in various facets of natural language processing, treating each task as a text-to-text challenge. Its capabilities span translation, summarization, and question answering, showcasing versatility and effectiveness.

Evaluation Metrics MT evaluation metrics are tools used to assess the quality of translations generated by MT models/systems. These metrics provide objective, quantifiable measures of how accurately and fluently a machine-translated text matches a reference translation, typically produced by humans. Among the most widely used BLEU (Papineni et al., 2002), ChrF (Popović, 2015a) evaluate the correspondence of n-grams between the machine-generated text and the reference, favoring precision. METEOR (Banerjee and Lavie, 2005), on the other hand, emphasizes both precision and recall, incorporating synonyms, stemming, and word order into its evaluation. TER (Agarwal and Lavie, 2008) measures the number of edits required to change a machine-translated text into a reference text. COMET (Rei et al., 2020; Stewart et al., 2020)

leverage contextual embedding to capture semantic similarities more effectively. These metrics facilitate the benchmarking of MT systems, guiding researchers and developers in improving translation technologies to achieve higher quality and more natural translations.

Recently, researchers have made strides in enhancing the capabilities of evaluation metrics like BLEU and COMET to better accommodate low-resource languages. Goyal et al. (2022) introduced a novel metric, SentencePiece BLEU (spBLEU), which extends coverage to 101 languages, including 23 African languages. This development marks a significant step forward in making language technology more inclusive, especially for languages that have traditionally been underrepresented in machine translation research. Similarly, Wang et al. (2023) developed AfriCOMET, a COMET-based evaluation metric, specifically designed to support 17 African languages. AfriCOMET represents another leap towards recognizing and addressing the unique challenges posed by African languages in machine translation. Both spBLEU and AfriCOMET exemplify the ongoing efforts within the research community to develop more robust and inclusive tools for evaluating machine translation quality across a broader spectrum of the world's languages.

B Cheetah Models Pretraining Details

Vocabulary. We employ SentencePiece (Kudo and Richardson, 2018) to encode text into Word-Piece tokens (Sennrich et al., 2016), utilizing 250K WordPieces. Additionally, our dataset encompasses the top ten globally spoken languages: Arabic, English, French, German, Greek, Italian, Portuguese, Russian, Spanish, and Turkish, sourced from Wikipedia dumps. Each language comprises 1 million sentences, solely included in the vocabulary.

Models Architecture. We pretrain AfroLingu-MTusing the encoder-decoder architecture (Xue et al., 2021b). Both the encoder and decoder components are structured similarly to T5, featuring 12 layers, each with 12 attention heads, and 768 hidden units for the base model. Consequently, the model comprises approximately ~ 580 million parameters.

Objective. We employ an unsupervised (denoising) objective, where the model is trained on masked (corrupted) versions of the original sentence to reconstruct the original sequence (Xue et al., 2021b). This objective involves randomly sampling and dropping out 15% of tokens in the input sequence. Subsequently, all consecutive spans of dropped-out tokens are replaced by a single sentinel token.

Pretraining Procedure During the pretraining of AfroLingu-MT, we employ a learning rate of 0.01, a batch size of 1,024 sequences, and a maximum sequence length of 1,024. Each model undergoes pretraining for 1 million steps. The training process is conducted on Google Cloud TPU with 128 cores $(v3 - 128)$ provided by the TensorFlow Research Cloud (TFRC).[2]

C AfroLingu-MT Benchmark

ISO	Name	Country(ies)	# of Speakers	Family	Script
aar	Afar	Ethiopia	2.36M	Afro-Asiatic	Latin
ach	Acholi	Uganda, South Sudan	1.58M	Nilo-Saharan	Latin
afr	Afrikaans	South Africa	17.67M	Indo-European	Latin
aka	Akan	Ghana	9.88M	Niger-Congo	Latin
amh	Amharic	Ethiopia	57.56M	Afro-Asiatic	Ethiopic
ara	Arabic	North Africa	372.56K	Afro-Asiatic	Arabic
bam	Bambara	Côte d'Ivoire, Mali	14.18M	Niger-Congo	Latin
bas	Bassa	Cameroon	300K	Niger-Congo	Latin
bem	Bemba	Zambia, Democratic Republic of Congo	4.11M	Niger-Congo	Latin
btg	Bhete	Côte d'Ivoire	329K	Niger-Congo	Latin
eng	English	Global	1.45B	Indo-European	Latin
ewe	Ewe	Ghana, Togo	5.5M	Niger-Congo	Latin
fan	Fang	Equatorial Guinea, Cameroon, Gabon	1.06M	Niger-Congo	Latin
fon	Fon	Benin, Togo	2.28M	Niger-Congo	Latin
fra	French	Global	310M	Indo-European	Latin
gez	Ge'ez	Ethiopia	Religious use	Afro-Asiatic	Ethiopic
hau	Hausa	Nigeria	78.52M	Afro-Asiatic	Latin
ibo	Ibo	Nigeria	30.89M	Niger-Congo	Latin
kau	Kanuri	Nigeria, Niger	9.47M	Nilo-Saharan	Latin
kbp	Kabiye	Togo, Benin	992K	Niger-Congo	Latin
kin	Kinyawanda	Rwanda	14.52M	Niger-Congo	Latin
kon	Kongo	Democratic Republic of Congo, Angola, Congo	7.01M	Niger-Congo	Latin
lgg	Lugbara	Uganda, Democratic Republic of Congo	1.94M	Nilo-Saharan	Latin
lin	Lingala	Democratic Republic of Congo, Congo	40.27M	Niger-Congo	Latin
lug	Luganda	Uganda	11.01M	Niger-Congo	Latin
mlg	Malagasy	Madagascar	25M	Austronesian	Latin
nnb	Nande	Democratic Republic of Congo	903K	Niger-Congo	Latin
nya	Chichewa	Malawi, Mozambique, Zambia	13.38M	Niger-Congo	Latin
nyn	Nyankore	Uganda, Rwanda	3.43M	Niger-Congo	Latin
orm	Oromo	Ethiopia	37.45M	Afro-Asiatic	Latin
pcm	Nigerian Pidgin	Nigeria	116M	Creole	Latin
som	Somali	Somalia	22.04M	Afro-Asiatic	Latin
sot	Sesotho	Lesotho	13.52M	Niger-Congo	Latin
ssw	Siswati	Eswatini	4.71M	Niger-Congo	Latin
swa	Swahili	Kenya, Tanzania	73M	Niger-Congo	Latin
swc	Swahili Congo	Democratic Republic of Congo	11.14M	Niger-Congo	Latin
teo	Ateso	Uganda, Kenya	2.77M	Nilo-Saharan	Latin
tir	Tigrinya	Eritrea, Ethiopia	8.82M	Afro-Asiatic	Ethiopic
tsn	Tswana	Botswana	13.75M	Niger-Congo	Latin
tso	Tsonga	South Africa	10M	Niger-Congo	Latin
twi	Twi	Ghana	9.88M	Niger-Congo	Latin
wal	Wolaytta	Ethiopia	2.49M	Afro-Asiatic	Latin
wol	Wolof	Senegal, Mauritania	12.39M	Niger-Congo	Latin
xho	Xhosa	South Africa	19.21M	Niger-Congo	Latin
yor	Yoruba	Nigeria	45.86M	Niger-Congo	Latin
zul	Zulu	South Africa	27.8M	Niger-Congo	Latin

Table C.1: Details of the Languages in our dataset.

Lang Pair	Train	Dev	Test	Lang Pair	Train	Dev	Test	Lang Pair	Train	Dev	Test
aar-amh	1166	50	145	eng-som	5,000	50	200	nyn-eng	5000	50	200
aar-eng	1146	50	143	eng-sot	4408	96	248	nyn-lgg	5000	50	200
aar-orm	1166	50	145	eng-ssw	580	36	72	nyn-lug	5000	50	200
aar-som	1166	50	146	eng-swa	5000	50	200	nyn-teo	5000	50	200
aar-tir	1165	50	145	eng-teo	5000	50	200	orm-aar	1166	50	145
ach-eng	5000	50	200	eng-tir	5000	50	200	orm-amh	5000	50	200
ach-lgg	5000	50	200	eng-tsn	5000	50	200	orm-eng	5000	50	200
ach-lug	5000	50	200	eng-tso	5000	50	200	orm-som	1170	50	146
ach-nyn	5000	50	200	eng-twi	5000	50	200	orm-tir	4980	50	200
ach-teo	5000	50	200	eng-wal	5000	50	200	orm-wal	2338	50	146
afr-eng	5000	50	200	eng-wol	5000	50	200	pcm-eng	1681	50	105
aka-eng	3145	50	196	eng-xho	5000	50	200	som-aar	1166	50	146
amh-aar	1166	50	145	eng-yor	5000	50	200	som-amh	1156	50	145
amh-eng	5000	50	200	eng-zul	5000	50	200	som-eng	5000	50	200
amh-gez	4618	50	200	ewe-eng	128	16	16	som-fra	4231	50	200
amh-mlg	693	43	87	ewe-fra	5000	50	200	som-orm	1170	50	146
amh-orm	5000	50	200	fan-btg	222	28	27	som-swa	1390	50	173
amh-som	1156	50	145	fon-fra	5000	50	200	som-tir	1153	50	144
amh-swa	112	14	14	fra-ara	5000	50	200	sot-eng	4408	96	248
amh-tir	5000	50	200	fra-ewe	5000	50	200	ssw-eng	580	37	72
amh-wal	3763	50	200	fra-fon	5000	50	200	swa-amh	112	14	14
ara-eng	5000	50	200	fra-lin	4000	50	200	swa-eng	5000	50	200
ara-fra	5000	50	200	fra-nnb	5000	50	200	swa-fra	5000	50	200
ara-yor	1397	50	100	fra-som	4231	50	200	swa-mlg	5000	50	200
bam-eng	4461	50	200	fra-swa	5000	50	200	swa-som	1390	50	173
bas-eng	5000	50	200	fra-swc	5000	50	200	swa-yor	24	3	3
bem-eng	5000	50	200	gez-amh	4619	50	200	swc-fra	5000	50	200
btg-fan	222	28	27	gez-eng	4724	50	200	teo-ach	5000	50	200
eng-aar	1146	50	143	hau-eng	5000	50	200	teo-eng	5000	50	200
eng-ach	5000	50	200	ibo-eng	5000	50	200	teo-lgg	5000	50	200
eng-afr	5000	50	200	kau-eng	4257	50	200	teo-lug	5000	50	200
eng-aka	3145	50	197	kbp-eng	5000	50	200	teo-nyn	5000	50	200
eng-amh	5000	50	200	kin-eng	5000	50	200	tir-aar	1165	50	145
eng-ara	5000	50	200	kon-eng	4357	50	200	tir-amh	5000	50	200
eng-bam	4462	50	200	lgg-ach	5000	50	200	tir-eng	5000	50	200
eng-bas	5000	50	200	lgg-lug	5000	50	200	tir-orm	4981	50	200
eng-bem	5000	50	200	lgg-nyn	5000	50	200	tir-som	1153	50	144
eng-ewe	128	16	16	lgg-teo	5000	50	200	tir-wal	2024	50	126
eng-gez	4724	50	200	lin-eng	246	31	31	tsn-eng	5000	50	200
eng-hau	5000	50	200	lin-fra	4000	50	200	tso-eng	5000	50	200
eng-ibo	5000	50	200	lug-ach	5000	50	200	twi-eng	5000	50	200
eng-kau	4257	50	200	lug-eng	5000	50	200	wal-amh	3763	50	200
eng-kbp	5000	50	200	lug-lgg	5000	50	200	wal-eng	5000	50	200
eng-kin	5000	50	200	lug-nyn	5000	50	200	wal-orm	2338	50	147
eng-kon	4357	50	200	lug-teo	5000	50	200	wal-tir	2024	50	127
eng-lin	246	31	31	mlg-amh	693	43	87	wol-eng	5000	50	200
eng-lug	5000	50	200	mlg-eng	5000	50	200	xho-eng	5000	50	200
eng-mlg	5000	50	200	mlg-swa	5000	50	200	yor-ara	1397	50	100
eng-nya	1052	50	132	mlg-yor	10	1	1	yor-eng	5000	50	200
eng-nyn	5000	50	200	nnb-fra	5000	50	200	yor-mlg	10	1	1
eng-orm	5000	50	200	nya-eng	1052	50	132	yor-swa	24	3	3
eng-pcm	1681	50	105	nyn-ach	5000	50	200	zul-eng	5000	50	200

Table C.2: Statistics of each language pair in our dataset.