Cheetah: Natural Language Generation for 517 African Languages

Ife Adebara[ξ,⋆] **AbdelRahim Elmadany**[ξ,⋆] **Muhammad Abdul-Mageed**[ξ,Ω,λ]
[ξ]The University of British Columbia [Ω]MBZUAI [λ] Invertible AI
{ife.adebara,a.elmadany,muhammad.mageed}@ubc.ca

Abstract

Low-resource African languages pose unique challenges for natural language processing (NLP) tasks, including natural language generation (NLG). In this paper, we develop Cheetah, a massively multilingual NLG language model for African languages. Cheetah supports 517 African languages and language varieties, allowing us to address the scarcity of NLG resources and provide a solution to foster linguistic diversity. We demonstrate the effectiveness of Cheetah through comprehensive evaluations across six generation downstream tasks. In five of the six tasks, Cheetah significantly outperforms other models, showcasing its remarkable performance for generating coherent and contextually appropriate text in a wide range of African languages. We additionally conduct a detailed human evaluation to delve deeper into the linguistic capabilities of Cheetah. The findings of this study contribute to advancing NLP research in low-resource settings, enabling greater accessibility and inclusion for African languages in a rapidly expanding digital landscape. The GitHub repository for the *Cheetah* project is available at https://github.com/UBC-NLP/Cheetah.

Figure 1: Cheetah is trained on 517 African languages and language varieties across 14 language families. The languages are domiciled in 50 of 54 African countries and are written in six different scripts.

1 Introduction

The linguistic diversity present in African languages poses unique challenges for NLG systems. With over 2,000 languages spoken across the African continent (Eberhard et al., 2021), the need for effective NLG solutions that can accommodate this rich linguistic ecosystem cannot be over-emphasized. This is especially important because traditional NLG approaches have primarily focused on high-resource languages, such as English and French due to the availability of large-scale datasets and resources. Consequently, low-resource languages, including numerous African languages, have been marginalized in NLG research and development. Developing robust NLG systems for the diverse needs of African communities is challenging due to the scarcity of extensive language datasets, limited linguistic research, and variations across these languages. To address these challenges, recent advancements in language modeling and transfer learning techniques have shown promise in supporting NLG in low-resource languages. Pretrained language models, such as GPT-3 (Radford et al., 2018, 2019; Brown et al., 2020), mT5 (Xue et al., 2021), and mT0 (Muennighoff et al., 2022), have demonstrated remarkable capabilities in understanding and generating human-like text. These models capture the statistical regularities and syntactic structures of the languages they are trained on, making them valuable starting points for supporting NLG in low-resource settings.

In this paper, we present a pioneering work on NLG in African languages by introducing Cheetah: a novel language model (LM) specifically designed to support 517 African languages and language varieties. To the best of our knowledge, Cheetah supports the largest number of African languages and language varieties. Leveraging a vast

⋆ Authors contributed equally.

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12798–12823

corpus of text data collected from diverse sources, Cheetah learns intricate linguistic information that characterizes each African language. The contributions of this work are three fold. **First**, we address the scarcity of NLG resources for African languages by providing a comprehensive language model that covers a wide range of languages spoken on the continent. **Second**, we demonstrate the efficacy of our approach through extensive evaluations across six downstream task clusters. Each cluster includes multiple languages, showcasing the model's ability to generate coherent and contextually appropriate text in different African languages. **Third**, we perform fine grained human analysis of Cheetah using a controlled machine translation (MT) test set. This uncovers model behaviour that is not visible with automatic metrics. By supporting NLG in African languages, we foster linguistic diversity, empower African communities to express themselves in their native languages, and bridge the digital divide. This paper serves as a foundational step towards promoting Afrocentric NLP (Adebara and Abdul-Mageed, 2022) that prioritizes inclusivity and cultural preservation in language technology, emphasizing the importance of catering to the unique linguistic needs of diverse populations.

The rest of the paper is organized as follows: In Section 2, we discuss related work. In Section, 4 we describe AfroNLG, the benchmark we create for evaluation. We provide details of Cheetah in Section 3. We present performance of Cheetah in Section 5 and compare it to other multilingual models. We present controlled test sets in Section 5.1. We conclude in Section 6, and outline a number of limitations and use cases for our work in Section 7 and Section 8.

2 Literature Review

One of the challenges in NLG is to generate coherent and semantically meaningful text. Various approaches have been proposed, including template-based (Becker, 2002; Van Deemter et al., 2005), rule-based (Dušek and Jurčíček, 2015; van Miltenburg et al., 2020), and statistical approaches (Li et al., 2016). More recently, deep learning approaches (Sutskever et al., 2014) including the transformer model (Vaswani et al., 2017) have achieved SoTA results in various NLG tasks such as text summarization (Shi et al., 2021) and machine translation (Vaswani et al., 2017).

While these models have shown impressive results, they often require a large amount of training data and computing resources. However, only a few African languages have benefited from these advancements due to inadequate data. To address this issue, researchers have proposed transfer learning-based approaches, where a pretrained model is finetuned for a specific NLG task. Transfer learning (Raffel et al., 2020; He et al., 2022; Ruder et al., 2019) has enabled the use of low-resource languages on various NLP tasks. Due to lack of adequate (or good quality) pretraining data (Kreutzer et al., 2021), transfer learning is often the most accessible method for only a few low-resource languages leaving behind a vast majority of extremely low-resource languages. This is because about 90% of the world's languages is claimed to be either *left-behinds*, in that it is probably impossible to build NLP resources for them, or *scraping-bys* with no labelled datasets (Joshi et al., 2020). For the left-behinds, labelled and unlabelled data are unavailable and even transfer learning approaches are beyond reach while the scraping-by languages have no labelled data with which to evaluate model performance.

2.1 Language Models

Only a few African languages have benefited from the recent advancement of language models (LM) due to inadequate data sizes. We now describe encoder-decoder LMs that support NLP tasks in African languages. We describe these under two broad headings: massively multilingual models and African models. We summarize the models and African languages they cover in Table 1.

Multilingual Models: The massively multilingual models such as mBART (Liu et al., 2020), MT0 (Muennighoff et al., 2022), and mT5 (Xue et al., 2021) are trained on several languages. However, in most cases, only a few African languages are represented. Among the mentioned models, mT0 is pretrained on the highest number of African languages (n=13).

African Models. Adelani et al. (2022) use pretrained T5, mT5, and mBART models and develop AfriByT5, AfriMT5, AfriMBART respectively to investigate machine translation in zero-shot and out-of-domain settings. The authors experiment on 17 African languages and demonstrate that further pretraining is effective for adding new languages to pretrained models. Jude Ogundepo et al.

(2022) train AfriTeVa, an encoder-decoder language model from scratch on ten African languages and English using similar training objectives like T5 model. AfriTeVa-V2 (Oladipo et al., 2023) has enhanced support for 16 African languages with improved quality pretraining data. [1]

African Natural Language Understanding. Several works attempt to improve the performance on African NLU tasks by proposing multilingual and African-dedicated models such as mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), AfriBERTa (Ogueji et al., 2021), AfroLM (Dossou et al., 2022), Afro-XLM-R (Alabi et al., 2022), KINYaBERT (Nzeyimana and Niyongabo Rubungo, 2022), and SERENGETI (Adebara et al., 2023).

2.2 Benchmarks

Multiple benchmarks have been developed for NLG. However, only a few of Africa's $2,000$ languages have been supported to date. In most cases, the benchmarks support only the machine translation task. We provide a brief overview under two headings: African and multilingual. We summarize key information about each benchmark in Table C.1.

African Benchmarks. `AfroMT` (Reid et al., 2021a) is a multilingual machine translation benchmark. It consists of translation tasks between English and eight African languages — Afrikaans, Xhosa, Zulu, Rundi, Sesotho, Swahili, Bemba, and Lingala. `Menyo-20k` (Adelani et al., 2021) is an MT evaluation benchmark for English-Yorùbá.

Multilingual with African Languages. `FLoRES-200` (Costa-jussà et al., 2022; Guzmán et al., 2019) is an evaluation benchmark that provides MT evaluation support in 200 languages including 52 African languages. `GEM` (Gehrmann et al., 2021, 2022) referenced as "living" benchmark, comprises of 40 tasks and supports 52 languages including 10 African languages. `NLLB Seed Data` (Costa-jussà et al., 2022) is a set of professionally-translated sentences sampled from Wikipedia. It consists of around six thousand sentences in 39 languages which include 8 African language. Similarly, `NLLB Multi Domain` (Costa-jussà et al., 2022) is an MT evaluation benchmark made from a set of professionally-translated sentences in the news and health domains. It consists of approximately $3,000$ sentences in each domain and supports 8 languages including 2 African languages. `Toxicity-200` (Costa-jussà et al., 2022) is an evaluation benchmark to evaluate the presence of toxic items in the MT text. It provides support for 50 African languages. `XGLUE` (Liang et al., 2020) is a cross-lingual, multi-task benchmark created with multilingual and bilingual corpora. It supports 19 languages and one African language, i.e., Swahili.

3 Cheetah

3.1 Pretraining Data

We are guided by three main principles in developing this data: quality, linguistic diversity, and coverage.

Quality. Developing NLP technologies for low resource languages poses a significant challenge due to the limited availability of high-quality training data. To address this issue, we undertook the task of manually curating a diverse corpus spanning multiple domains, including news articles, health documents, religious texts, legal documents, and social media feeds. This manual curation approach was necessary because there were no existing datasets available for the majority of the languages we aimed to support, and we wanted to ensure the utilization of reliable and high-quality data.

Coverage. In all, we train Cheetah using a 42G multi-domain corpus across 517 African languages and language varieties. The languages are spoken in 50 of 54 African countries and they are written with five scripts. This provides support to at least 500M Africans.

Linguistic Diversity. The inclusion of languages from various domains, geographical regions, and linguistic typologies, along with the utilization of reliable data sources, contributes to enhancing the robustness and quality of Cheetah. Our data consists of languages from 14 language families in Africa written in five different orthographies. Furthermore, our data spans languages with a vast array of exotic linguistic features including tone, vowel and consonant harmony, reduplication, word orders, and word classes.

More details about pretraining is in Appendix A.

3.2 Implementation Details

Vocabulary. We use SentencePiece (Kudo and Richardson, 2018) to encode text as WordPiece

[1] After we finished our work, we became aware of a new version of AfriTeva, AfriTevaV2 (Oladipo et al., 2023) that was released only recently (December, 2023). We plan to evaluate AfriTevaV2 in our camera-ready version of this work.

Category	LM	Lang/Total	African Languages	Families
Multilingual	MBART	3/50	afr, swh, yor.	2
	MT0	14/101	afr, amh, hau, ibo, lin, mlg, nyj, orm, sot, sna, som, swh, xho, yor, and zul	4
	MT5	12/101	afr, amh, nya, hau, ibo, mlg, sna, som, swh, xho, yor, and zul	3
African	AfriVeTa	10/10	gaz, amh, Gahuza, hau, ibo, pcm, som, swa, tir, and yor.	3
	AfriMT5	17/17	bam, bbj, ewe, fon, hau, ibo, lug, luo, pcm, mos, swa, tsn, twi, wol, yor, zul.	3
	AfriByT5	17/17	bam, bbj, ewe, fon, hau, ibo, lug, luo, pcm, mos, swa, tsn, twi, wol, yor, zul.	3
	AfriMBART	17/17	afr, amh, nya, hau, orm, som, swh, xho.	3
	Cheetah	517/517	Includes 517 African languages.	14

Table 1: Comparing with available encoder-decoder models with African languages represented. **Lang/Total**. describe the number of African languages comparing with the covered languages in the pretrained language models. **Families**. describes the number of covered language families.

tokens (Sennrich et al., 2016) with 250K WordPieces. We also include data covering the ten top spoken languages globally: Arabic, English, French, German, Greek, Italian, Portuguese, Russian, Spanish, and Turkish. We use Wikipedia dumps for these ten languages. We use 1M sentences for each language. However, we only include it in the vocabulary.

Models Architecture. We pretrain Cheetah using the encoder-decoder architecture (Xue et al., 2021). Each of the encoder and decoder components is similar in size and configuration to T5, with 12 layers each with 12 attention heads, and 768 hidden units for the base model. In total, this results in a model with $\sim$ 580 million parameters. We provide further details in Table B.1.

Objective. We use an unsupervised (denoising) objective. The main idea is to feed the model with masked (corrupted) versions of the original sentence, and train it to reconstruct the original sequence. The denoising objective (Xue et al., 2021) works by randomly sampling and dropping out 15% of tokens in the input sequence. All consecutive spans of dropped-out tokens are then replaced by a single sentinel token.

Pretraining Procedure For pretraining Cheetah, we use a learning rate of 0.01, a batch size of $1,024$ sequences, and a maximum sequence length of $1,024$. We pretrain each model for 1M steps. We train our models on Google Cloud TPU with 128 cores ($v3-128$) from TensorFlow Research Cloud (TFRC).[2]

4 AfroNLG Benchmark

We create AfroNLG, a multi-lingual, multi-task benchmark comprising 67 test sets across six task clusters. AfroNLG includes cloze tasks, machine translation, paraphrase, question answering, summarization, and title generation. It supports 527 languages, including 517 African languages and language varieties and the top 10 world languages. To the best of our knowledge, this is the most extensive benchmark till date for African languages. Table C.1 shows, how our benchmark compares to others. We provide the details of each task cluster and datasets in what follows. For detailed statistics about the task clusters, refer to Appendix C.

Cloze Test. In order to comprehensively evaluate Cheetah across all the languages it was pretrained on, we employ cloze-tasks as our evaluation approach and perform two cloze tasks experiments. These tasks assess the model's ability to fill in missing information. In the first cloze task, which we henceforth call **`mask-one`**, we randomly mask only one token in each sentence. In the second cloze-task, which we call **`mask-at-least-one`**, we randomly mask at least one token and not more than 10% of the tokens in each sentence. For each of the 517 languages, we construct a cloze-task dataset comprising 200 data points for each language in the Train set, 100 examples for each language in the Test set, and 50 data points for each language in the Dev set. We ensure that there is no overlap between the data used for the cloze tasks and the pretraining data. We show an example of our cloze task in Figure C.1.

Machine Translation. We include only datasets pertaining African languages in our benchmark. In selecting the languages for our MT benchmark, we strive to keep datasets that have been used in any published machine translation task. This allows us to cover a diverse set of languages and compare our models to existing SoTA across a large number of

[2]https://sites.research.google/trc/about/

language pairs. Our benchmark thus contains data from Afro-MT[3] (Reid et al., 2021b), Lafand-MT[4] (Adelani et al., 2022), PidginUNMT[5] (Ogueji and Ahia, 2019), and SALT[6] (Akera et al., 2022). The datasets we consider make up 35 language pairs.

Paraphrase. A paraphrase task aims to create semantically similar and fluent paraphrases given an input text (Chen et al., 2023; Palivela, 2021). We use the TaPaCo dataset (Scherrer, 2020) for our paraphrase generation benchmark. TaPaCo is a freely available paraphrase corpus for 73 languages extracted from the Tatoeba database. The dataset has four African languages: Afrikaans, Berber (a macro-language), Amazigh, and Kirundi.

Question Answering. The QA task aims to provide answers to questions based on a knowledge base also referred to as contexts. We use TYDIA[7] QA dataset (Clark et al., 2020). The dataset has a primary task and a gold passage task. In our benchmark, we only include the gold passage task, where a correct answer is predicted from a passage containing one answer, similar to the existing reading comprehension task.

Summarization. Summarization is the task of generating an abridged version of a text, while capturing the salient ideas and the intended information from the original text (Nallapati et al., 2016; King et al., 2022). We use the subset of XL-Sum (Hasan et al., 2021), an abstractive summarization dataset, that consists of African languages including Amharic, Hausa, Igbo, Kirundi, Oromo, Pidgin, Somali, Swahili, Tigrinya, and Yorùbá. We also develop new test sets using data we crawled from the web, which are non-overlapping with XL-Sum. Specifically, we crawl data from BBC and Voice of Africa (webpages) for Hausa, Ndebele, and Swahili.

Title Generation. The title generation task returns a single sentence title for a given article. Similar to the summarization task, we use XL-SUM to create a news title generation dataset. We also collect a new test set for title generation across 15 languages. The dataset comprises $\sim 6,000$ BBC and Voice of Africa articles, non-overlapping with XL-Sum, and is particularly useful for zero-shot title generation.

[3]https://github.com/machelreid/afromt

[4]https://github.com/masakhane-io/lafand-mt

[5]https://github.com/keleog/PidginUNMT

[6]https://github.com/SunbirdAI/salt

[7]https://github.com/google-research-datasets/tydiqa

5 Evaluation and Results

We evaluate Cheetah on six task clusters of AfroNLG benchmark and compare to performance on mT0, mT5, Afri-MT5, and AfriTeVa. We report results in Table 2. For all models, we finetune on the training data split (Train) for 20 epochs with an early stopping of 5 epochs, learning-rate of $5e-5$, batch size of 16, and sequence length of 512. All experiments were performed on 4 GPUs (Nvidia V100). We report the results of each experiment as an *average of three runs*, each with a different seed.[8] We show evaluation results per language and provide information of model performance next.

Cloze Test. Cheetah outperforms all other models on both cloze tasks as in Table 2. We show the results for each language that is supported by the models compared in Table D.1 and Table D.2. The performance of all models on mask-one is better than the performance on mask-at-least-one, reflecting how increasing the number of masked tokens makes the task more challenging. It is also important to mention that since evaluation is based on BLEU it does not reflect correct synonyms that each model may have generated to replace the masked tokens.

Machine Translation. Cheetah sets a new SOTA on 23 tasks surpassing previous models. The mT0 and AfriTEVA models also demonstrate strong performance on six languages. Notably, pairs with French as the source language tend to yield the lowest BLEU scores, indicating relatively lower translation quality. On the other hand, the language pair involving English to Nigerian Pidgin, specifically on LafandMT and PidginUNMT, showcases the highest BLEU scores. We assume that the similarity between the Nigerian Pidgin and English contributes favourably to translation quality in these scenarios. We also report CHRF and CHRF++ results in Table C.4 and Table C.5 in the Appendix.

Paraphrase. In the three paraphrase tasks, Cheetah demonstrates remarkable superiority over all other models. Specifically, we achieve an impressive ROUGE score of 46.0 on the Berber paraphrase task, surpassing the second-best model by a margin of approximately two points.

Question Answering. In the task of question answering, mT0 exhibits superior performance compared to other models. While mT5 achieves the second-highest performance, Cheetah attains the third-highest performance in this task.

[8]Specifically, we use seed values 41, 1512, and 20235.

Cluster	Task	Metric	mT0	mT5	Afri-MT5	AfriTeVa	Cheetah
Machine Translation (MT)	English → Afrikaans	Bleu	$\mathbf{20.38}^{\pm0.3}$	$12.35^{\pm1.1}$	$7.12^{\pm2.67}$	$7.75^{\pm1.67}$	$19.72^{\pm0.75}$
	English → Bemba	Bleu	$19.19^{\pm0.3}$	$12.28^{\pm0.48}$	$11.73^{\pm12.3}$	$\mathbf{20.5}^{\pm0.87}$	$18.9^{\pm1.22}$
	English → Lingala	Bleu	$\mathbf{15.98}^{\pm1.16}$	$14.12^{\pm0.56}$	$14.32^{\pm12.74}$	$13.88^{\pm1.04}$	$9.64^{\pm1.11}$
	English → Rundi	Bleu	$\mathbf{12.26}^{\pm0.47}$	$8.82^{\pm0.43}$	$9.57^{\pm0.42}$	$7.83^{\pm1.04}$	$10.54^{\pm0.54}$
	English → Sesotho	Bleu	$11.04^{\pm1.2}$	$12.74^{\pm0.75}$	$10.0^{\pm1.79}$	$10.76^{\pm1.4}$	$\mathbf{13.3}^{\pm1.38}$
	English → Swahili	Bleu	$10.59^{\pm1.84}$	$9.33^{\pm0.58}$	$3.08^{\pm0.57}$	$7.24^{\pm0.46}$	$\mathbf{11.08}^{\pm0.61}$
	English → Xhosa	Bleu	$10.04^{\pm0.98}$	$8.25^{\pm0.7}$	$3.86^{\pm1.35}$	$7.5^{\pm0.32}$	$\mathbf{12.34}^{\pm0.51}$
	English → Zulu	Bleu	$17.65^{\pm1.86}$	$17.97^{\pm1.69}$	$1.9^{\pm1.11}$	$13.45^{\pm1.81}$	$\mathbf{19.49}^{\pm1.16}$
	English → Hausa	Bleu	$5.06^{\pm0.21}$	$4.96^{\pm0.16}$	$0.85^{\pm0.04}$	$7.32^{\pm0.00}$	$\mathbf{9.22}^{\pm0.08}$
	English → Igbo	Bleu	$13.05^{\pm0.17}$	$11.57^{\pm0.23}$	$1.12^{\pm0.09}$	$12.34^{\pm0.23}$	$\mathbf{16.75}^{\pm0.26}$
	English → Luganda	Bleu	$2.17^{\pm2.77}$	$3.33^{\pm0.35}$	$0.09^{\pm0.01}$	$4.21^{\pm0.77}$	$\mathbf{9.75}^{\pm0.01}$
	English → N. Pidgin	Bleu	$\mathbf{33.17}^{\pm0.28}$	$32.65^{\pm0.19}$	$2.39^{\pm0.23}$	$9.39^{\pm0.18}$	$32.64^{\pm0.14}$
	English → Swahili	Bleu	$22.04^{\pm2.89}$	$23.2^{\pm0.23}$	$2.79^{\pm0.08}$	$22.39^{\pm0.28}$	$\mathbf{28.11}^{\pm0.14}$
	English → Zulu	Bleu	$6.83^{\pm0.29}$	$0.58^{\pm1.37}$	$0.4^{\pm0.03}$	$4.45^{\pm0.37}$	$\mathbf{11.75}^{\pm0.38}$
	English → Twi	Bleu	$3.4^{\pm0.12}$	$1.23^{\pm0.03}$	$0.03^{\pm0.0}$	$1.68^{\pm0.94}$	$\mathbf{4.64}^{\pm0.13}$
	English → Yoruba	Bleu	$5.42^{\pm0.85}$	$2.58^{\pm3.1}$	$0.04^{\pm0.0}$	$3.63^{\pm4.01}$	$\mathbf{7.83}^{\pm0.14}$
	English → Zulu	Bleu	$10.28^{\pm0.49}$	$1.31^{\pm2.26}$	$0.14^{\pm0.03}$	$3.8^{\pm4.2}$	$\mathbf{12.13}^{\pm0.1}$
	French → Bambara	Bleu	$2.0^{\pm2.6}$	$0.37^{\pm0.19}$	$0.15^{\pm0.01}$	$\mathbf{3.18}^{\pm0.18}$	$3.06^{\pm0.27}$
	French → Ghomálá’	Bleu	$0.4^{\pm0.09}$	$0.33^{\pm0.01}$	$0.07^{\pm0.0}$	$\mathbf{0.96}^{\pm0.01}$	$0.28^{\pm0.25}$
	French → Ewe	Bleu	$0.7^{\pm0.35}$	$0.31^{\pm0.36}$	$0.09^{\pm0.07}$	$0.84^{\pm0.16}$	$\mathbf{3.47}^{\pm0.03}$
	French → Fon	Bleu	$0.69^{\pm0.31}$	$0.8^{\pm0.13}$	$1.52^{\pm0.06}$	$\mathbf{1.73}^{\pm0.53}$	$1.29^{\pm0.16}$
	French → Moore	Bleu	$0.27^{\pm0.06}$	$0.12^{\pm0.05}$	$0.19^{\pm0.02}$	$0.47^{\pm0.04}$	$\mathbf{1.66}^{\pm0.86}$
	French → Wolof	Bleu	$4.02^{\pm0.12}$	$0.3^{\pm0.05}$	$0.11^{\pm0.01}$	$\mathbf{3.08}^{\pm0.25}$	$3.01^{\pm0.07}$
	English → N. Pidgin (UNMT)	Bleu	$\mathbf{27.44}^{\pm0.26}$	$23.42^{\pm1.61}$	$7.05^{\pm1.37}$	$22.54^{\pm0.84}$	$26.56^{\pm0.04}$
	Acholi → English	Bleu	$16.41^{\pm0.08}$	$11.16^{\pm4.77}$	$4.9^{\pm0.11}$	$8.37^{\pm8.12}$	$\mathbf{19.33}^{\pm0.1}$
	Acholi → Lugbara	Bleu	$2.57^{\pm0.21}$	$1.48^{\pm1.31}$	$2.44^{\pm0.37}$	$\mathbf{8.29}^{\pm0.14}$	$7.21^{\pm0.69}$
	Acholi → Luganda	Bleu	$3.64^{\pm0.07}$	$1.74^{\pm0.12}$	$0.92^{\pm0.01}$	$5.53^{\pm0.34}$	$\mathbf{8.03}^{\pm0.38}$
	Acholi → Nyankore	Bleu	$2.17^{\pm0.14}$	$0.79^{\pm0.51}$	$0.46^{\pm0.03}$	$4.26^{\pm0.54}$	$\mathbf{5.1}^{\pm0.14}$
	Acholi → Ateso	Bleu	$1.64^{\pm2.34}$	$1.94^{\pm0.25}$	$4.9^{\pm0.11}$	$\mathbf{7.74}^{\pm0.33}$	$6.33^{\pm0.6}$
	English → Lugbara	Bleu	$6.19^{\pm6.33}$	$8.38^{\pm0.49}$	$5.93^{\pm0.22}$	$10.95^{\pm0.32}$	$\mathbf{11.61}^{\pm0.28}$
	English → Luganda	Bleu	$12.08^{\pm0.03}$	$10.58^{\pm0.25}$	$2.59^{\pm0.73}$	$12.41^{\pm0.35}$	$\mathbf{17.12}^{\pm0.16}$
	English → Nyankore	Bleu	$6.46^{\pm0.08}$	$5.69^{\pm0.02}$	$1.4^{\pm0.39}$	$7.88^{\pm0.18}$	$\mathbf{9.04}^{\pm0.24}$
	English → Ateso (salt)	Bleu	$10.24^{\pm0.06}$	$8.28^{\pm0.19}$	$4.91^{\pm0.59}$	$\mathbf{11.64}^{\pm0.49}$	$11.12^{\pm0.38}$
	Lugbara → Ateso	Bleu	$2.21^{\pm0.35}$	$1.5^{\pm0.2}$	$2.22^{\pm0.15}$	$\mathbf{6.67}^{\pm0.32}$	$3.68^{\pm0.31}$
	Luganda → Lugbara	Bleu	$3.96^{\pm0.57}$	$2.61^{\pm0.12}$	$3.44^{\pm0.32}$	$\mathbf{8.05}^{\pm0.23}$	$7.99^{\pm0.47}$
	Luganda → Ateso	Bleu	$4.47^{\pm0.08}$	$3.01^{\pm0.16}$	$2.5^{\pm0.22}$	$\mathbf{8.17}^{\pm0.18}$	$8.13^{\pm0.33}$
	Nyankore → Lugbara	Bleu	$3.45^{\pm0.29}$	$2.1^{\pm0.32}$	$2.6^{\pm0.29}$	$\mathbf{7.5}^{\pm0.09}$	$7.29^{\pm0.09}$
	Nyankore → Luganda	Bleu	$8.54^{\pm0.17}$	$6.91^{\pm0.23}$	$2.01^{\pm0.25}$	$\mathbf{6.77}^{\pm6.73}$	$6.25^{\pm10.26}$
	Nyankore → Ateso	Bleu	$3.33^{\pm0.11}$	$2.25^{\pm0.23}$	$2.12^{\pm0.4}$	$6.27^{\pm0.12}$	$\mathbf{6.36}^{\pm0.4}$
Paraphrase	Multilingual	Bleu	$41.79^{\pm0.28}$	$41.75^{\pm0.21}$	$34.72^{\pm0.51}$	$43.02^{\pm1.25}$	$\mathbf{43.23}^{\pm0.09}$
	Berber	Bleu	$44.84^{\pm0.31}$	$44.03^{\pm0.24}$	$36.08^{\pm0.83}$	**$\mathbf{46.41}^{\pm0.71}$	$46.0^{\pm0.27}$
	Kabyle	Bleu	$25.91^{\pm0.13}$	$25.32^{\pm0.46}$	$11.56^{\pm0.73}$	$16.06^{\pm14.79}$	$\mathbf{26.27}^{\pm0.56}$
Question Answering	QA Swahili	F1	$\mathbf{79.84}^{\pm0.19}$	$72.04^{\pm0.54}$	0	$62.64^{\pm0.78}$	$71.98^{\pm1.18}$
Summarization	Multilingual	RougeL	$22.31^{\pm0.12}$	$22.23^{\pm0.04}$	$5.34^{\pm0.48}$	$18.97^{\pm0.06}$	$\mathbf{24.86}^{\pm0.02}$
	Amharic	RougeL	$13.81^{\pm0.04}$	$13.09^{\pm0.03}$	$4.4^{\pm1.07}$	$8.29^{\pm0.51}$	$\mathbf{15.09}^{\pm0.1}$
	Igbo	RougeL	$\mathbf{18.9}^{\pm0.73}$	$13.22^{\pm0.46}$	$14.24^{\pm0.39}$	$16.05^{\pm0.49}$	$17.36^{\pm0.43}$
	Oromo	RougeL	$11.28^{\pm0.03}$	$10.51^{\pm0.07}$	$3.52^{\pm0.49}$	$7^{\pm1.73}$	$\mathbf{14.53}^{\pm0.1}$
	Rundi	RougeL	$19.63^{\pm0.01}$	$18.02^{\pm0.13}$	$11.82^{\pm0.39}$	$16.13^{\pm0.03}$	$\mathbf{22.57}^{\pm0.04}$
	Swahili	RougeL	$26.38^{\pm0.02}$	$24.81^{\pm0.11}$	$15.07^{\pm0.17}$	$21.59^{\pm0.13}$	$\mathbf{29.05}^{\pm0.13}$
	Yoruba	RougeL	$21.57^{\pm0.05}$	$20.06^{\pm0.12}$	$13.52^{\pm0.18}$	$17.3^{\pm0.11}$	$\mathbf{22.49}^{\pm0.0}$
	Hausa	RougeL	$26.46^{\pm0.06}$	$25.76^{\pm0.02}$	$19.96^{\pm0.26}$	$25.19^{\pm0.11}$	$\mathbf{30.07}^{\pm0.31}$
	Nigerian Pidgin	RougeL	$26.54^{\pm0.05}$	$25.79^{\pm0.1}$	$14.28^{\pm1.23}$	$20.29^{\pm0.12}$	$\mathbf{27.08}^{\pm0.02}$
	Somali	RougeL	$20.69^{\pm0.08}$	$19.21^{\pm0.06}$	$13.62^{\pm0.81}$	$19.27^{\pm0.18}$	$\mathbf{23.92}^{\pm0.04}$
	Tigrinya	RougeL	$15.84^{\pm0.13}$	$13.93^{\pm0.11}$	$6.53^{\pm0.42}$	$10.07^{\pm0.09}$	$\mathbf{16.88}^{\pm0.12}$
Title Generation	Multilingual	Bleu	$6.53^{\pm0.02}$	$6.65^{\pm0.08}$	$0.1^{\pm0.02}$	$5.2^{\pm0.02}$	$\mathbf{7.52}^{\pm0.07}$
	Amharic	Bleu	$3.13^{\pm0.23}$	$2.65^{\pm0.68}$	$0.34^{\pm0.14}$	$2.31^{\pm0.14}$	$\mathbf{4.34}^{\pm0.34}$
	Igbo	Bleu	$6.95^{\pm0.13}$	$6.9^{\pm0.22}$	$0.77^{\pm0.12}$	$4.61^{\pm0.14}$	$\mathbf{8.47}^{\pm0.07}$
	Oromo	Bleu	$1.1^{\pm1.84}$	$2.66^{\pm0.19}$	$0.21^{\pm0.06}$	$1.54^{\pm0.17}$	$\mathbf{3.26}^{\pm0.21}$
	Rundi	Bleu	$4.4^{\pm0.28}$	$4.13^{\pm0.22}$	$0.84^{\pm0.07}$	$3.33^{\pm0.23}$	$\mathbf{6.05}^{\pm0.5}$
	Swahili	Bleu	$9.1^{\pm0.23}$	$9.31^{\pm0.11}$	$1.22^{\pm0.09}$	$7.01^{\pm0.09}$	$\mathbf{10.59}^{\pm0.6}$
	Yoruba	Bleu	$6.8^{\pm0.16}$	$7.23^{\pm0.59}$	$0.34^{\pm0.05}$	$5.04^{\pm2.0}$	$\mathbf{7.97}^{\pm0.32}$
	Hausa	Bleu	$8.11^{\pm0.24}$	$7.3^{\pm0.34}$	$2.59^{\pm0.01}$	$6.69^{\pm0.18}$	$\mathbf{8.48}^{\pm0.23}$
	Nigerian Pidgin	Bleu	$\mathbf{6.75}^{\pm0.6}$	$3.96^{\pm4.3}$	$0.89^{\pm0.02}$	$4.72^{\pm0.84}$	$6.22^{\pm0.28}$
	Somali	Bleu	$3.37^{\pm0.21}$	$3.31^{\pm0.16}$	$0.38^{\pm0.11}$	$2.82^{\pm0.47}$	$\mathbf{5.25}^{\pm0.14}$
	Tigrinya	Bleu	$2.99^{\pm0.1}$	$2.94^{\pm1.09}$	$0.7^{\pm0.18}$	$1.92^{\pm0.26}$	$\mathbf{5.1}^{\pm0.05}$
Cloze-task	Mask-one	Bleu	$13.61^{\pm0.91}$	$8.18^{\pm3.94}$	$0.00^{\pm0.00}$	$8.36^{\pm3.42}$	$\mathbf{13.98}^{\pm0.32}$
	Mask-at-least-one	Bleu	$2.36^{\pm0.11}$	$2.66^{\pm0.09}$	$0.93^{\pm0.12}$	$0.68^{\pm0.09}$	$\mathbf{7.07}^{\pm0.09}$
		AfroNLG Score	12.56	11.05	5.15	10.84	**14.25**

Table 2: Average performance of finetuned African and multilingual models across three runs on AfroLNG benchmark test sets.

Summarization. Cheetah sets a new SOTA on 11 languages, outperforming other models by an average margin of at least three points. Detailed results can be found in Table 2.

Title Generation. On the Title generation task, Cheetah sets a new SOTA on 11 languages. We report results in Table 2.

5.1 Investigating linguistic capabilities

In order to further test the utility of our models, we use grammar templates to construct test data in English. We use nine linguistic rules and 19 lexical items to generate 152 sentences. Next, we use our model to translate from source to target and manually evaluate the quality of the generated data. We design new evaluation metrics, *faithfulness* and *fluency*, for the manual evaluation (see Section 5.2). A detailed description follows.

Grammar templates. We use grammar templates (McCoy et al., 2019) developed with context-free grammars (CFG) on the source side to construct controlled test sets in English. We use CFG on the source side alone because constituents and constituent order differs across languages. We adopt this method for two reasons. First, utilizing grammar templates provides a standardized framework that ensures that the same grammatical phenomena are tested consistently. By employing a uniform approach, we can effectively isolate and evaluate specific linguistic features, facilitating a more rigorous and meaningful comparison of language model performance. Second, grammar templates exhibit a high degree of flexibility, allowing for easy modification and extension to encompass a wide range of linguistic phenomena. This adaptability not only facilitates the incorporation of new linguistic features but also enables the evolution of our test sets to match the dynamic landscape of natural language processing research.

Other alternatives to templates include using parsed corpora (Bender et al., 2011) or naturally occurring sentences. For the languages we explore, there are no good quality parsers, making automatic parsing inaccessible for this analysis. Furthermore, when a corpus is parsed automatically, the likelihood of encountering parsing errors escalates with the intricacy of the sentence structure (Bender et al., 2011; Marvin and Linzen, 2018). Conversely, if the test set exclusively comprises sentences with accurate gold parses, sourcing an ample quantity of instances showcasing syntactic complexities becomes an arduous task (Marvin and Linzen, 2018). Furthermore, the utilization of naturally occurring sentences introduces potential complications that might confound the interpretation of experiments (Ettinger et al., 2018). The templates include transitive and intransitive structures, negative and affirmative structures, and structures with gender and number. Table E.1 provides examples of generated sentences using the templates.The entire generated grammar is available at our GitHub:[9].

Inference. We test three of our finetuned machine translation models with the generated dataset. This allows us to evaluate how much linguistic information the models have acquired during pretraining and finetuning. Specifically, we use the English→Hausa, English→Swahili, and English→Yorùbá based on MT0, MT5, AfriTEVA, and Cheetah models that were finetuned on the LafandMT dataset. We do not include Afri-MT5 in this analysis because it has very low scores across several tasks as shown in Table 2. Notably, Hausa, Swahili, and Yorùbá have distinct typologies and the performance of each model on each language gives further insight of performance across varying typological features. Table E.2 shows a number of linguistic differences between the three languages. This method can be generalized to any African language.

5.1.1 Linguistic Details

Morphology Morphologically, both Hausa and Swahili are classified as agglutinative languages (Jaggar, 2017; Dryer and Haspelmath, 2013), characterized by the systematic addition of prefixes, suffixes, and affixes to root words or stems. This process imparts precise grammatical meanings, encompassing tense, case, mood, person, number, and more. Conversely, Yorùbá exhibits an analytic structure, relying on word order and discrete function words to denote grammatical relationships, with minimal use of inflections or affixes. The following are examples from the generated (1) Hausa, (2) Swahili, and (3) Yorùbá, respectively.

(1)	a.	*Bai*	*barshi*	*ba*
		neg.masculine	leave	at-all

'he did not leave him'

[9] generatedsentences

b. Bata barshi ba
Neg.feminine leave at-all

'she did not leave him'

(2) a. *Ha-ku-mu-a-cha*
3pl.sg.sub-neg-3pl.sg.obj-leave

'He did not leave him'

b. Ha-ku-mu-a-cha
3pl.sg.sub-neg-3pl.sg.obj-leave

'She did not leave him'

(3) a. *Òhun ò kúrò lỌ́dỌ̀ ẹ̀*
3pl.sg.sub neg leave from 3pl.sg.obj

'He did not leave him'

b. Òhun ò kúrò lỌ́dỌ̀ ẹ̀
3pl.sg.sub neg leave from 3pl.sg.obj

'She did not leave him'

Phonology In terms of phonology, Yorùbá and Hausa are tonal languages, where pitch distinctions contribute to word differentiation. However, Hausa features a relatively simpler tone system compared to Yorùbá and in most cases tone is not marked in Hausa orthography. Only dictionaries and pedagogical materials indicate tone in text. Yorùbá on the other hand has three tones and indicating tones in orthography significantly reduces ambiguity (Adebara and Abdul-Mageed, 2022). Swahili, in contrast, is devoid of tones altogether.

5.2 Human evaluation

To evaluate the effectiveness of each model across different languages, we assess the generated output's faithfulness and fluency using a five-point Likert scale. *Faithfulness* measures how accurately a model's output corresponds to the input sentence, while *fluency* assesses the grammatical coherence and plausibility of the generated output. We use both metrics because a model can produce coherent output that may not be faithful to the input sentence. This way, if faithfulness penalizes a model for outputs that are not true to the input or that include additional unnecessary information, fluency complements our evaluation of the quality of the same model if the output is fluent. For each grammar category, we return the average Likert point for each language and across the different models model.

5.3 Annotation

We annotated each model's output for faithfulness and fluency. For Hausa and Yorùbá, two expert annotators evaluated the model's output for faithfulness and fluency. We ensured that each annotator has native speaker competency in reading and writing (while some had a linguistic background). We gave specific annotation instructions (see Section E in the Appendix) to ensure the values are not assigned arbitrarily. We also ensured that the annotators do not know who created which models to prevent any biases. We report the Cohen's Kappa agreement scores in Table E.3 (Appendix). For Swahili, only one annotator made it to the final annotation task since we could not acquire high quality annotations from other annotators. The Swahili annotator who did the final annotation is a university lecturer with a Ph.D. in linguistics.

5.4 Fluency and Faithfulness Performance

We report the distribution of faithfulness and fluency scores across all models and languages in Figure 2. Overall, Cheetah produces more faithful and more fluent outputs than other models on all languages. We now go on to provide a brief analysis of model performance. More details are in in Appendix F.

Intransitives In the case of Hausa examples, all three models can generate intransitive sentences with varying levels of fluency and faithfulness with Cheetah outperforming other models. In the context of Swahili, errors primarily relate to tense, possibly because Swahili has an agglutinative structure, and the models may lack exposure to a comprehensive range of grammatical features during training. In the case of Yorùbá, all models consistently incorporate at least one object in each intransitive case. This could be because intransitive sentences in Yorùbá lack a clear direct object, making it challenging for machine translation models to select the accurate translation. Additionally, some intransitive phrases in Yorùbá can be polysemous, further complicating the translation process. We report the distribution of scores in Figure G.1.

Transitives Cheetah demonstrates the capability to provide three distinct semantic senses for the polysemous transitive verb treated whereas the other models typically produce only a single semantic in-

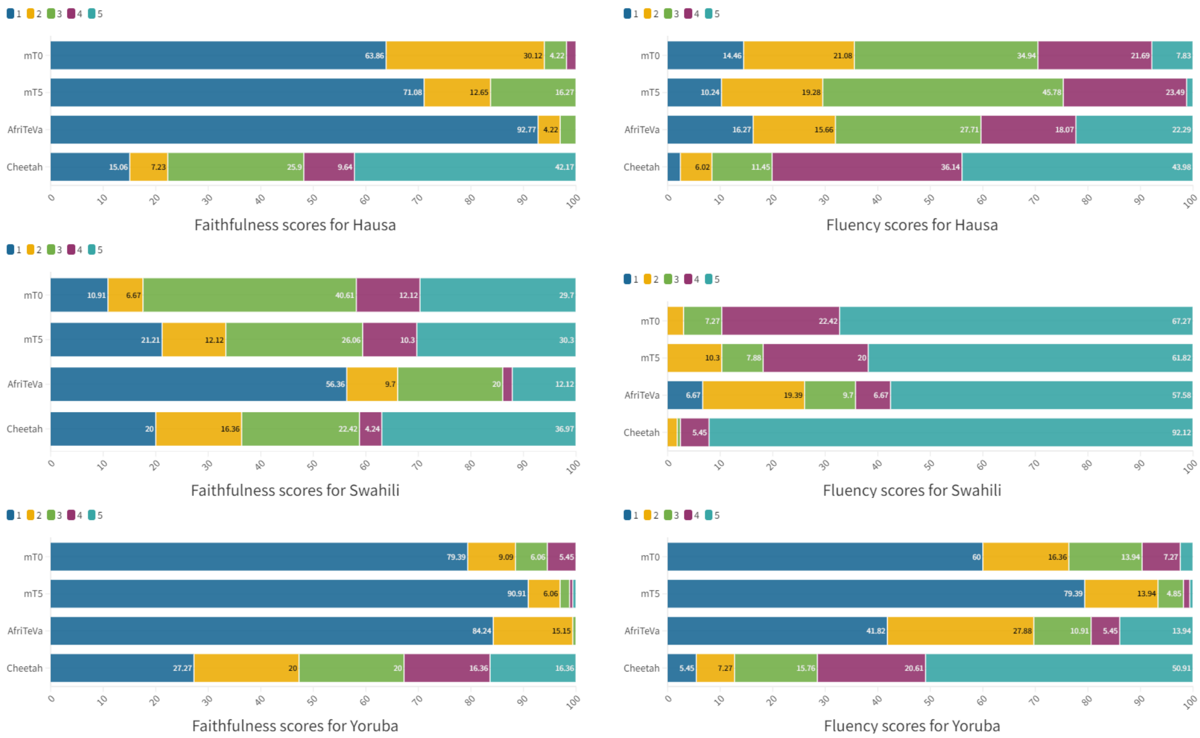

Figure 2: Faithfulness and fluency for Hausa, Swahili, and Yorùbá

terpretation. For Swahili, certain instances exhibit the deletion or simplification of object markers in an ungrammatical manner. Figure G.2 shows the distribution of model performance on transitives.

Negative mT0, MT5, and AfriTeVa have a tendency to output the negation of the antonym of the verb in each sentence rather than the negation of the verb. Cheetah also makes similar mistakes about 5% of the time. **Affirmative.** The models generally perform better in the context of the affirmative examples than on the negated examples. However, in the context of Hausa, mT5, mT0, and AfriTeVa consistently output the antonym of negated verb. For instance, the models return “Sara left" rather than “Sara did not leave". In the Swahili examples, we also find cases of double negation (which is not grammatically correct in Swahili). We show the distribution of results in Figure G.5 and Figure G.4.

Gender/Agreement Yorùbá does not distinguish gender, yet Cheetah uses *Arábìrin* (female) before every occurrence of the name “Sara" to indicate that the it has a high probability of being feminine (see Figure G.3). However, “Fred" is not annotated this way. For Hausa, which requires agreement between the gender of the noun and the verb, we find Cheetah outperforming both mt0 and mt5 significantly. AfriTeVa, however, has very low accuracy in the context of gender. Furthermore, mt0, mt5, and Cheetah return connotations for love and relationships for each examples where a male and female pronoun co-occur cross-lingually.

6 Conclusion

We introduced Cheetah, a massively multilingual language model designed for African natural language generation. We also propose a new African language generation benchmark, dubbed AfroNLG, that is both sizeable and diverse. We evaluate Cheetah on AfroNLG comparing it to three other models, two multilingual and one dedicated to African languages. The performance of Cheetah surpasses that of all other models we evaluate. This is demonstrated by its superior AfroNLG score, which is approximately three times better than the combined performance of other models. Furthermore, Cheetah outperforms all other models across 48 out of 65 test sets spanning six task clusters. We further analyze our model’s robustness to lexical complexity and carry out human evaluation to inspect the model’s perform on a controlled test set. Again, our results underscore superiority of our model.

7 Limitations

We identify the following limitations for our work:

1. The limitations of our language model include

the limited scope of our evaluation. Future work should focus on increasing the subset of languages evaluated manually in order to ensure quality. We believe automatic analyses are not sufficient for development of models that get deployed in particular applications.

2. Another limitation is related to our inability to perform extensive analysis of biases and hateful speech present in our pretraining data. Again, this is due to relatively restricted access to native speakers (and even automated tools) to perform this analysis. As a result, we cannot fully ensure that our models are free from biases and socially undesirable effects. Therefore, it is important that these models be used with care and caution, and be analyzed for biases and socially undesirable effects before use.

3. Additionally, due to unavailability of sufficient computing resources, we were unable to evaluate larger multilingual language models.

8 Ethics Statement and Wider Impacts

Cheetah aligns with Afrocentric NLP where the needs of African people is put into consideration when developing technology. We believe Cheetah will not only be useful to speakers of the languages supported, but also researchers of African languages such as anthropologists and linguists. We discuss below some use cases for Cheetah and offer a number of broad impacts.

1. Cheetah aims to address the lack of access to technology in about 90% of the world's languages, which automatically discriminates against native speakers of those languages. More precisely, it does so by focusing on Africa. To the best of our knowledge, Cheetah is the first massively multilingual PLM developed for African languages and language varieties. A model with knowledge of 517 African languages, is by far the largest to date for African NLP.

2. Cheetah enables improved access of important information to the African community in Indigenous African languages. This is especially beneficial for people who may not be fluent in other languages. This will potentially connect more people globally.

3. Cheetah affords opportunities for language preservation for many African languages. To the best of our knowledge, Cheetah consists of languages that have not been used for any NLP task until now. We believe that it can help encourage continued use of these languages in several domains, as well as trigger future development of language technologies for many of these languages.

4. Although LMs are useful for a wide range of applications, they can also be misused. Cheetah is developed using publicly available datasets that may carry biases. Although we strive to perform analyses and diagnostic case studies to probe performance of our models, our investigations are by no means comprehensive nor guarantee absence of bias in the data. In particular, we do not have access to native speakers of most of the languages covered. This hinders our ability to investigate samples from each (or at least the majority) of the languages.

5. We emphasize the ethical use of these models.

Acknowledgments

We acknowledge support from Canada Research Chairs (CRC), the Natural Sciences and Engineering Research Council of Canada (NSERC; RGPIN-2018-04267), the Social Sciences and Humanities Research Council of Canada (SSHRC; 435-2018-0576; 895-2020-1004; 895-2021-1008), Canadian Foundation for Innovation (CFI; 37771), Digital Research Alliance of Canada,[10] and UBC Advanced Research Computing-Sockeye.[11]

References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), pages 1 – 9, Mannheim. Leibniz-Institut für Deutsche Sprache.

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

[10] https://alliancecan.ca
[11] https://arc.ubc.ca/ubc-arc-sockeye

3814–3841, Dublin, Ireland. Association for Computational Linguistics.

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2023. Serengeti: Massively multilingual language models for africa.

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 Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

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 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.

Tilman Becker. 2002. Practical, template–based natural language generation with TAG. In *Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+6)*, pages 80–83, Universitá di Venezia. Association for Computational Linguistics.

Emily M. Bender, Dan Flickinger, Stephan Oepen, and Yi Zhang. 2011. Parser evaluation over local and non-local deep dependencies in a large corpus. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 397–408, Edinburgh, Scotland, UK. 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.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiaao Chen, Derek Tam, Colin Raffel, Mohit Bansal, and Diyi Yang. 2023. An Empirical Survey of Data Augmentation for Limited Data Learning in NLP. *Transactions of the Association for Computational Linguistics*, 11:191–211.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Chinenye Emezue. 2022. Afrolm: A self-active learning-based multilingual pretrained language model for 23 african languages.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Ondřej Dušek and Filip Jurčíček. 2015. Training a natural language generator from unaligned data. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 451–461, Beijing, China. Association for Computational Linguistics.

David M Eberhard, F Simons Gary, and Charles D Fennig (eds). 2021. Ethnologue: Languages of the world. *Twenty-fourth edition*, Dallas, Texas: SIL International.

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120, Online. Association for Computational Linguistics.

Sebastian Gehrmann, Abhik Bhattacharjee, Abinaya Mahendiran, Alex Wang, Alexandros Papangelis, Aman Madaan, Angelina Mcmillan-major, Anna Shvets, Ashish Upadhyay, Bernd Bohnet, Bingsheng Yao, Bryan Wilie, Chandra Bhagavatula, Chaobin You, Craig Thomson, Cristina Garbacea, Dakuo Wang, Daniel Deutsch, Deyi Xiong, Di Jin, Dimitra Gkatzia, Dragomir Radev, Elizabeth Clark, Esin Durmus, Faisal Ladhak, Filip Ginter, Genta Indra Winata, Hendrik Strobelt, Hiroaki Hayashi, Jekaterina Novikova, Jenna Kanerva, Jenny Chim, Jiawei Zhou, Jordan Clive, Joshua Maynez, João Sedoc, Juraj Juraska, Kaustubh Dhole, Khyathi Raghavi Chandu, Laura Perez Beltrachini, Leonardo F . R. Ribeiro, Lewis Tunstall, Li Zhang, Mahim Pushkarna, Mathias Creutz, Michael White, Mihir Sanjay Kale, Moussa Kamal Eddine, Nico Daheim, Nishant Subramani, Ondrej Dusek, Paul Pu Liang, Pawan Sasanka Ammanamanchi, Qi Zhu, Ratish Puduppully, Reno Kriz, Rifat Shahriyar, Ronald Cardenas, Saad Mahamood, Salomey Osei, Samuel Cahyawijaya, Sanja Štajner, Sebastien Montella, Shailza Jolly, Simon Mille, Tahmid Hasan, Tianhao Shen, Tosin Adewumi, Vikas Raunak, Vipul Raheja, Vitaly Nikolaev, Vivian Tsai, Yacine Jernite, Ying Xu, Yisi Sang, Yixin Liu, and Yufang Hou. 2022. GEMv2: Multilingual NLG benchmarking in a single line of code. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 266–281, Abu Dhabi, UAE. Association for Computational Linguistics.

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Junxian He, Chunting Zhou, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Towards a unified view of parameter-efficient transfer learning. In *International Conference on Learning Representations*.

Philip J. Jaggar. 2017. *The Hausa "Grade 5" verb: Morphosyntactic preliminaries*, 1 edition, pages 18–27. Harrassowitz Verlag.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

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.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S. Weld, Iz Beltagy, and Doug Downey. 2022. Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 555–571, Abu Dhabi, United Arab Emirates (Hybrid). 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 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.

Aman Kumar, Himani Shrotriya, Prachi Sahu, Amogh Mishra, Raj Dabre, Ratish Puduppully, Anoop Kunchukuttan, Mitesh M. Khapra, and Pratyush Kumar. 2022. IndicNLG benchmark: Multilingual datasets for diverse NLG tasks in Indic languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5363–5394, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Xiao Li, Kees van Deemter, and Chenghua Lin. 2016. Statistics-based lexical choice for NLG from quantitative information. In *Proceedings of the 9th International Natural Language Generation conference*, pages 104–108, Edinburgh, UK. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, Daniel Campos, Rangan Majumder, and Ming Zhou. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. 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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. 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, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany. Association for Computational Linguistics.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5347–5363, Dublin, Ireland. Association for Computational Linguistics.

Kelechi Ogueji and Orevaoghene Ahia. 2019. Pidginunmt: Unsupervised neural machine translation from west african pidgin to english. *arXiv preprint arXiv:1912.03444*.

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.

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.

Hemant Palivela. 2021. Optimization of paraphrase generation and identification using language models in natural language processing. *International Journal of Information Management Data Insights*, 1(2):100025.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021a. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 african languages. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Punta Cana, Dominican Republic.

Machel Reid, Junjie Hu, Graham Neubig, and Yutaka Matsuo. 2021b. AfroMT: Pretraining strategies and reproducible benchmarks for translation of 8 African languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1306–1320, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sebastian Ruder, Matthew E. Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials*, pages 15–18, Minneapolis, Minnesota. Association for Computational Linguistics.

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6868–6873, Marseille, France. European Language Resources Association.

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.

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan K. Reddy. 2021. Neural abstractive text summarization with sequence-to-sequence models. *ACM/IMS Trans. Data Sci.*, 2(1).

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112. MIT Press.

Kees Van Deemter, Emiel Krahmer, and Mariët Theune. 2005. Real versus template-based natural language generation: A false opposition? *Comput. Linguist.*, 31(1):15–24.

Emiel van Miltenburg, Chris van der Lee, Thiago Castro-Ferreira, and Emiel Krahmer. 2020. Evaluation rules! on the use of grammars and rule-based systems for NLG evaluation. In *Proceedings of the 1st Workshop on Evaluating NLG Evaluation*, pages 17–27, Online (Dublin, Ireland). Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages

483–498, Online. Association for Computational Linguistics.

Appendices

A Pretraining Data

We provide details of our pretraining data below: **Religious Domain.** Our religious data is taken from online Bibles, Qurans, and data crawled from the Jehovah's witness website. We also include religious texts from the book of Mormon.

News Domain. We collect data from online newspapers (Adebara and Abdul-Mageed, 2022) and news sites such as Voice of America, Voice of Nigeria, BBC, Global voices, and DW news sites. We collect local newspapers from 27 languages from across Africa.

Government Documents. We collect government documents South African Centre for Digital Language Resources (SADiLaR), and the Universal Declaration of human rights (UDHR) in multiple languages.

Health Documents. We collect multiple health documents from the Department of Health, State Government of Victoria, Australia. We collect documents in Amharic, Dinka, Harari, Oromo, Somali, Swahili, and Tigrinya.

Existing Corpora. We collect corpora available on the web for different African languages, including from Project Gutenberg for Afrikaans, South African News data. for Sepedi and Setswana, OSCAR (Abadji et al., 2021) for Afrikaans, Amharic, Somali, Swahili, Oromo, Malagasy, and Yoruba. We also used Tatoeba for Afrikaans, Amharic, Bemba, Igbo, Kanuri, Kongo, Luganda, Malagasy, Sepedi, Ndebele, Kinyarwanda, Somali, Swahili, Tsonga, Xhosa, Yoruba, and Zulu; Swahili Language Modelling Data for Swahili; Ijdutse corpus for Hausa; Data4Good corpora for Luganda, CC-100 for Amharic, Fulah, Igbo, Yoruba, Hausa, Tswana, Lingala, Luganada, Afrikaans, Somali, Swahili, Swati, North Sotho, Oromo, Wolof, Xhosa, and Zulu; Afriberta-Corpus for Afaan / Oromo, Amharic, Gahuza, Hausa, Igbo, Pidgin, Somali, Swahili, Tigrinya and Yoruba; mC4 for Afrikaans, Amharic, Hausa, Igbo, Malagasy, Chichewa, Shona, Somali, Sepedi, Swahili, Xhosa, Yoruba and Zulu.

B Models

C AfroNLG Benchmark

We report statistics of AfroNLG benchmark in Table C.2 and C.1 respectively.

Category	Source	Target
Mask-one	Àwọn Ìbéèrè Táwọn Èèyàn Máa Ń <extra_id_0>	Béèrè
	Rimwe na rimwe baranyugaranira hanze canke bakarya ntibansigire <extra_id_0> .	Namba
Mask-at-least-one	Nilaa ala ma ya <extra_id_0> atënë pepe si ala zia ngbene lege ti sarango <extra_id_1> ti ala pepe , tongaso si mbi kai kobela ti ala pepe .	ti ye
	አኸ <extra_id_0> <extra_id_1> እያ ቢፍታኼ ቄም ም የኸኖ የጨቄስከኸማ ኸስብ ፤ ኸኖ ጭን እያ ኸወጥቅወ ጐጀ ኸነቦም ። 21 ቶ !	በኸ አትማችኸማ

Figure C.1: Examples from the mask-one and mask-at-least-one cloze task data.

C.1 CHRF and CHRF++ Results

D Cloze Task Results

We provide results on the performance of each model on individual languages. We use a dash '-' to indicate that a specific model does not support a language.

E Annotation

We gave the following annotation rules to our annotators: Faithfulness refers to how close to the English sentence the model output is. It should be annotated with values between 1 and 5. Faithfulness should be evaluated independently of the fluency of the model output. Below are some detailed explanations for the scale for faithfulness:

- Give a value **1** if model output is not related to the source sentence.
- Give a value **2** if the model output is the opposite of the source sentence.
- Give a value **3** if the model output is somewhat related to the source sentence. It should have some words or phrases that make it related to the source.
- Give a value **4** if the model output is closely related but changes the meaning slightly (e.g difference in gender, number etc)
- Give a value **5** if the model output is an exact translation

Fluency is how grammatically correct the model is. Faithfulness and fluency should be judged independently. That is, even if the output is not faithful, don't use it to determine the fluency score and vice versa. Here are some detailed explanations on how to assign the values:

- Give a value **1** if model output is completely ungrammatical and nonsensical.

Model	Size	Params	No._heads	No._layers	D_model	Vocab	S._Len	B. Size	#Train_Steps	#Langs	#A.Langs
mT0	base	580M	12	12	768	∼250k	1024	1024	UNK	101	13
mT5	base	580M	12	12	768	250K	1024	1024	1M	101	13
AfriMT5	base	580M	UNK	UNK	UNK	UNK	UNK	2048	UNK	17	17
AfriTeVa	base	229M	12	12	768	40K	512	256	500K	10	10
Cheetah	base	580M	12	12	768	250K	1024	1024	1M	527	517

Table B.1: Parameters of Cheetah compared with other models.

Category	Benchmark	Reference	Task	Lang/Total	Datasets	Tasks
Multilingual	FLoRES200	(Costa-jussà et al., 2022)	52/200	MT	Wiki	1
	GEM_{v1}	(Gehrmann et al., 2021)	DRG, DT, RES, TS, SMP	10/52	18	13
	GEM_{v2}	(Gehrmann et al., 2021)	DRG, DT, PPH, QA, RES, TS, SLG, SMP, TS	10/52	50	9
	IndicNLG	(Kumar et al., 2022)	BG, HG, SUM, PARA, QA	0/11	5	5
	IndoNLG	(Cahyawijaya et al., 2021)	SUM, QA, Chit-Chat	0/3	5	3
	NLLB M.D.	(Costa-jussà et al., 2022)	MT	2/8	Wiki	1
	NLLB S.D.	(Costa-jussà et al., 2022)	MT	2/8	Wiki	1
	Toxicity200	(Costa-jussà et al., 2022)	MT	50/200	Wiki	1
	XGLUE	(Liang et al., 2020)	NER, POS, MLQA, PAWS-X, XLNI, NC, QADSM, WPR, QAM, QG, NTG	1/19	19	11
African	AfroMT	(Reid et al., 2021a)	MT	8/8	5	1
	Menyo-20k	(Adelani et al., 2021)	MT	1/2	6	1
	AfroNLG	Our Work	Cloze, CS, MT, QA, TG, SUM, PARA	517/527	67	7

Table C.1: A Comparison of AfroNLG with other multilingual Benchmarks. **MT**: Machine translation, **QA:** Question Answering, **CS:** Code-Switching, **TG**: Title Generation, **SUM**: Summarization, **PARA**: Paraphrase, **NER**: Named Entity Recognition, **POS**: Part-Of-Speech Tagging, **MLQA:** Multilingual Question Answering, **PAWS-X**: Parallel Aggregated Word Scrambling for Cross-Lingual Understanding, **XNLI**: Cross-Lingual Natural Language Interference, **NC**: News Classification, **QADSM**: Query-AD Matching, **WPR:** Web Page Ranking, **QAM**: QA Matching, **NTG**: News Title Generation, **BG**: WikiBio Biography Generation, and **HG:** Headline Generation. **SD**: Seed Data, **MD**: Multi Domain. **DRG:** Dialogue Response Generator, **DT:** Data-to-Text, **RES:** Reasoning, **TS:** Text Summarization, **SMP:** Text Simplification, **PPH:** Paraphrase, **SLG:** Slide Generation

- Give a value **2** if the model output is reasonable but includes some foreign words or gibberish.
- Give a value **3** if the model output contains some grammatical phrases but also contains some ungrammatical phrases.
- Give a value **4** if the model output is almost grammatical (but may have a few errors like spelling mistakes)
- Give a value **5** if the model output is very fluent and sounds looks like what a native speaker will say.

To prepare for the official annotation process, each evaluator annotated 10 random samples for the language pair they were assigned to. Following the individual evaluations, we reviewed any annotation errors and inconsistencies in assessments and assigned another random pair of 10 samples. In the second phase, we removed evaluators for which the quality of their annotations were deemed of a poor quality.

F Fluency and Faithfulness Performance

We report the distribution of faithfulness and fluency scores across all models and languages in Figure 2. Overall, Cheetah produces more faithful and more fluent outputs than other models on all languages. We now go on to provide detailed analysis of model performance.

Intransitives In the case of Hausa examples, all three models manage to produce intransitive examples. However, Cheetah consistently appends objects to these intransitive examples. This inclination to add objects might stem from biases within the data used for pretraining or finetuning Cheetah. Nevertheless, it is worth noting that Cheetah outperforms other models by generating more fluent and more faithful Hausa outputs. In the Swahili context, all models successfully generate intransitive translations, with model errors primarily related to tense. This performance discrepancy in Swahili can be attributed to its agglutinative structure, with models potentially lacking exposure to a comprehensive range of grammatical features during pretraining

Dataset	Pairs	Train	Dev	Test
Lafand	eng-hau	5,866	1,301	1,501
	eng-ibo	6,945	1,457	1,412
	eng-lug	4,076	1,501	1,501
	eng-pcm	4,791	1,485	1,565
	eng-swa	30,783	1,792	1,836
	eng-tsn	2,101	1,343	1,501
	eng-twi	3,338	1,285	1,501
	eng-yor	6,645	1,545	1,559
	eng-zul	3,540	1,462	1,001
	fra-bam	3,014	1,501	1,501
	fra-bbj	2,233	1,134	1,431
	fra-ewe	2,027	1,415	1,564
	fra-fon	2,638	1,228	1,580
	fra-mos	2,494	1,493	1,575
	fra-wol	3,361	1,507	1,501
AfroMT	eng-afr	25,799	3,226	3,226
	eng-bem	12,043	1,506	1,506
	eng-lin	17,679	2,211	2,210
	eng-run	12,475	1,560	1,560
	eng-sot	28,844	3,607	3,606
	eng-swa	28,084	3,511	3,512
	eng-xho	26,091	3,263	3,262
	eng-zul	29,127	3,641	3,642
PidginUNMT	eng-pcm	1,682	211	211
SALT	All-pairs	20,006	2,501	2,502

Table C.2: Statistics of the MT data in our benchmark. All-pairs each have the same size of data. They include ach-eng, ach-lgg, ach-lug, ach-nyn, ach-teo, ach-teo, eng-lgg, eng-lug, eng-nyn, eng-teo, lgg-teo, lug-lgg, lug-teo, nyn-lgg, nyn-lug, and nyn-teo

or finetuning. In the context of Yorùbá, all models consistently incorporate at least one object in each intransitive case. Notably, mT0 generates an output without an object approximately 5.88% of the time. This may be because intransitive sentences inherently lack a clear direct object, making it more challenging for machine translation models to grasp context and select the accurate translation. In certain instances, some intransitive phrases can be polysemous, further complicating the translation process. Intransitive English verbs do not always retain their intransitive nature in Yorùbá. Furthermore, transitives with optional/truncated objects tend to have a compulsory object in Yorùbá. This phenomenon potentially contributes to the models' tendency to append objects to intransitive Yorùbá phrases. For instance, whereas the intransitive "slept" in "John slept" maps to the intransitive form "John sùn" in Yorùbá, the intransitive verb "prayed", in "John prayed" becomes "John gbàdúrà", a transitive verb in Yorùbá. On the other hand, the transitive verb "ate" in "John ate", has an optional/truncated object in English but becomes "John jẹun", a transitive with an obligatory object. In Yorùbá, both "ate" and "prayed" are transitive verbs that require an object. They are derived from "jẹ" (eat) and "oúnjẹ" (food), which give rise to "jẹun" and "gbà" (collect) and "àdúrà" (prayer), resulting in "gbàdúrà" respectively.

Transitives In the context of transitives, Cheetah stands out as the top-performing model across all three languages, as illustrated in Figure 2. Cheetah demonstrates the capability to provide three distinct semantic senses for the polysemous transitive verb treated whereas the other models typically produce only a single semantic interpretation. In Swahili examples, certain instances exhibit the deletion or simplification of object markers in an ungrammatical manner. For a visual representation of the annotation of intransitive sentences in Yorùbá, please refer to Figure G.3. Figure G.2 shows the distribution of model performance on transitives.

Negative In the context of Yorùbá, all models are able to produce the correct negation marker including the correct tone marks. The tone patterns on negation markers may vary based on the context of words before and after the negation marker and it was interesting to see these variations in the models outputs. Despite this, mT0, MT5, and AfriTeVa have a tendency to output the negation of the antonym of the verb in each sentence rather than the negation of the verb. Cheetah also makes similar mistakes about 5% of the time.

Gender/Agreement We find interesting cases of gender in the model's output. For example, whereas Yorùbá grammar does not distinguish gender, Cheetah uses *Arábìrin* (female) before every occurrence of the name "Sara" to indicate that the it has a high probability of being feminine (see Figure G.3). It is important to mention that "Fred" is not annotated this way. For Hausa, which requires agreement between the gender of the noun and the verb, we find Cheetah outperforming both mt0 and mt5 significantly. AfriTeVa, however, has very low accuracy in the context of gender. Furthermore, mt0, mt5, and Cheetah return connotations for love and relationships for each examples where a male and female pronoun co-occur cross-lingually.

Number Cheetah significantly outperforms all three models in accurately assigning appropriate number markers. We also find that when translating the word "you" into Hausa, Swahili, or Yorùbá, all four models use either singular or plural forms. We assume that this is due to the fact that the second person in English (i.e., "you") can be either singular or plural while each of these languages have a different word for the singular and plural forms.

Task Cluster	Test Set	Source	Train	Dev	Test
Cloze test	517 languages	Ours	103,400	25,850	51,700
Paraphrase	Multilingual††	(Scherrer, 2020)	22,390	2,797	2,794
	Berber		17,607	2,200	2,200
	Kabyle		4,441	555	555
Question Answering	Swahili	(Clark et al., 2020)	49,881	499	n/a
Summarization	Multilingual†	(Hasan et al., 2021)	63,040	7,875	7875
	Amharic		5,761	719	719
	Igbo		4,183	522	522
	Oromo		6,063	757	757
	Rundi		5,746	718	718
	Swahili		7,898	987	987
	Yorùbá		6,350	793	793
	Hausa		6,418	802	802
	Nigerian Pidgin		9,208	1,151	1,151
	Somali		5,962	745	745
	Tigrinya		5,451	681	681
	Multilingual⋆†	Ours			428
Title Generation	Multilingual†	(Hasan et al., 2021)	63,040	7,875	7875
	Amharic		5,761	719	719
	Igbo		4,183	522	522
	Oromo		6,063	757	757
	Rundi		5,746	718	718
	Swahili		7,898	987	987
	Yorùbá		6,350	793	793
	Hausa		6,418	802	802
	Nigerian Pidgin		9,208	1,151	1,151
	Somali		5,962	745	745
	Tigrinya		5,451	681	681
	Multilingual⋆	Ours			5899

Table C.3: Statistics of the data in our benchmark. †† includes amh, ber, kab, run. † has amh, ibo, orm, run, swa, yor, hau, pcm, som, and tir. ⋆† is a newly created summarization test set including 'hau', 'nde' (zero-shot), and 'swa'. ⋆ is a newly created test set across 15 languages: 'amh', 'gag' (zero-shot), 'hau', 'ibo', 'pcm', 'som', 'swa', 'tir', 'yor', 'kin' (zero-shot), 'afr', 'mlg' (zero-shot), 'orm', 'nde' (zero-shot), 'sna'(zero-shot)

Task	Metric	mT0	mT5	afri-mt5	AfriTeVa	Cheetah
Translate English to Afrikaans	Chrf	$26.97^{\pm 4.75}$	$26.11^{\pm 4.12}$	$14.66^{\pm 8.79}$	$20.75^{\pm 4.02}$	$\mathbf{39.88^{\pm 0.81}}$
Translate English to Bemba	Chrf	$10.27^{\pm 0.89}$	$6.39^{\pm 1.96}$	$20.23^{\pm 13.97}$	$9.94^{\pm 10.05}$	$\mathbf{15.76^{\pm 0.19}}$
Translate English to Rundi	Chrf	$21.51^{\pm 1.39}$	$17.56^{\pm 3.13}$	$24.91^{\pm 3.59}$	$\mathbf{31.58^{\pm 2.33}}$	$28.65^{\pm 3.55}$
Translate English to Sesotho	Chrf	$21.08^{\pm 3.54}$	$12.08^{\pm 10.91}$	$23.75^{\pm 4.77}$	$\mathbf{29.57^{\pm 1.61}}$	$29.05^{\pm 2.41}$
Translate English to Swahili	Chrf	$23.26^{\pm 0.16}$	$20.35^{\pm 4.87}$	$24.60^{\pm 0.2}$	$20.5^{\pm 4.88}$	$\mathbf{37.24^{\pm 0.04}}$
Translate English to Xhosa	Chrf	$27.44^{\pm 3.1}$	$25.88^{\pm 4.94}$	$\mathbf{34.97^{\pm 2.49}}$	$20.25^{\pm 15.35}$	$33.45^{\pm 0.21}$
Translate English to Zulu	Chrf	$27.12^{\pm 3.49}$	$21.54^{\pm 2.16}$	$37.8^{\pm 1.41}$	$25.39^{\pm 16.55}$	$\mathbf{43.75^{\pm 0.11}}$
Translate English to Hausa	Chrf	$28.53^{\pm 0.26}$	$27.65^{\pm 0.53}$	$19.99^{\pm 0.42}$	$31.68^{\pm 0.29}$	$\mathbf{34.9^{\pm 0.32}}$
Translate English to Igbo	Chrf	$40.31^{\pm 0.17}$	$37.18^{\pm 0.34}$	$22.01^{\pm 0.7}$	$33.24^{\pm 0.23}$	$\mathbf{44.37^{\pm 0.31}}$
Translate English to Luganda	Chrf	$25.94^{\pm 2.41}$	$23.33^{\pm 0.31}$	$15.57^{\pm 1.45}$	$24.16^{\pm 2.55}$	$\mathbf{36.22^{\pm 0.09}}$
Translate English to N. Pidgin	Chrf	$63.49^{\pm 0.05}$	$\mathbf{63.9^{\pm 0.1}}$	$24.79^{\pm 0.68}$	$53.76^{\pm 0.01}$	$62.95^{\pm 0.17}$
Translate English to Swahili	Chrf	$50.52^{\pm 3.33}$	$51.76^{\pm 0.12}$	$21.00^{\pm 0.7}$	$44.84^{\pm 0.33}$	$\mathbf{56.36^{\pm 0.15}}$
Translate English to Setswana	Chrf	$30.89^{\pm 0.36}$	$16.62^{\pm 0.22}$	$13.17^{\pm 1.73}$	$23.75^{\pm 0.45}$	$\mathbf{35.87^{\pm 0.64}}$
Translate English to Twi	Chrf	$23.56^{\pm 0.24}$	$15.8^{\pm 1.29}$	$12.74^{\pm 1.33}$	$17.47^{\pm 3.26}$	$\mathbf{25.89^{\pm 0.2}}$
Translate English to Yoruba	Chrf	$19.41^{\pm 1.97}$	$16.51^{\pm 0.38}$	$11.49^{\pm 0.29}$	$20.62^{\pm 0.36}$	$\mathbf{25.09^{\pm 0.07}}$
Translate English to Zulu	Chrf	$35.4^{\pm 1.27}$	$16.13^{\pm 7.84}$	$15.04^{\pm 1.1}$	$12.75^{\pm 0.56}$	$\mathbf{38.81^{\pm 0.21}}$
Translate French to Bambara	Chrf	$16.49^{\pm 0.39}$	$7.44^{\pm 1.12}$	$10.16^{\pm 1.58}$	$19.41^{\pm 0.53}$	$\mathbf{19.91^{\pm 0.05}}$
Translate French to Ghomálá'	Chrf	$8.3^{\pm 0.76}$	$6.53^{\pm 0.57}$	$6.72^{\pm 3.75}$	$\mathbf{13.16^{\pm 0.4}}$	$8.57^{\pm 3.15}$
Translate French to Ewe	Chrf	$10.19^{\pm 2.32}$	$5.46^{\pm 3.02}$	$6.96^{\pm 3.02}$	$13.44^{\pm 1.64}$	$\mathbf{21.6^{\pm 0.22}}$
Translate French to Fon	Chrf	$5.67^{\pm 2.65}$	$6.09^{\pm 0.72}$	$5.82^{\pm 1.58}$	$11.88^{\pm 1.83}$	$\mathbf{12.71^{\pm 0.41}}$
Translate French to Moore	Chrf	$7.86^{\pm 1.43}$	$5.16^{\pm 2.20}$	$7.79^{\pm 0.97}$	$11.42^{\pm 0.7}$	$\mathbf{12.34^{\pm 0.56}}$
Translate French to Wolof	Chrf	$17.55^{\pm 0.2}$	$3.15^{\pm 0.12}$	$11.26^{\pm 1.91}$	$\mathbf{17.58^{\pm 0.44}}$	$16.67^{\pm 0.21}$
Translate English to N. Pidgin (pidginUNMT)	Chrf	$41.83^{\pm 0.17}$	$37.12^{\pm 0.77}$	$21.65^{\pm 1.33}$	$39.04^{\pm 0.50}$	$\mathbf{40.2^{\pm 0.17}}$
Translate Acholi to English	Chrf	$39.12^{\pm 0.1}$	$33.07^{\pm 5.49}$	$21.65^{\pm 1.33}$	$34.19^{\pm 0.06}$	$\mathbf{42.17^{\pm 0.05}}$
Translate Acholi to Lugbara	Chrf	$25.05^{\pm 0.85}$	$20.61^{\pm 5.92}$	$28.71^{\pm 0.34}$	$\mathbf{34.01^{\pm 0.29}}$	$32.31^{\pm 1.11}$
Translate Acholi to Luganda	Chrf	$22.13^{\pm 0.63}$	$25.75^{\pm 0.02}$	$24.31^{\pm 0.1}$	$32.77^{\pm 0.68}$	$\mathbf{37.34^{\pm 0.47}}$
Translate Acholi to Nyankore	Chrf	$27.52^{\pm 0.45}$	$20.03^{\pm 3.88}$	$24.50^{\pm 0.02}$	$32.39^{\pm 0.92}$	$\mathbf{35.0^{\pm 0.33}}$
Translate Acholi to Ateso	Chrf	$26.0^{\pm 1.99}$	$22.16^{\pm 1.63}$	$28.33^{\pm 0.01}$	$\mathbf{35.37^{\pm 0.61}}$	$34.62^{\pm 1.05}$
Translate English to Lugbara	Chrf	$38.84^{\pm 0.01}$	$37.12^{\pm 0.77}$	$39.11^{\pm 0.01}$	$38.94^{\pm 0.3}$	$\mathbf{40.2^{\pm 0.17}}$
Translate English to Luganda	Chrf	$43.71^{\pm 0.08}$	$41.05^{\pm 0.19}$	$35.34^{\pm 1.11}$	$43.14^{\pm 0.22}$	$\mathbf{49.38^{\pm 0.02}}$
Translate English to Nyankore	Chrf	$40.43^{\pm 0.21}$	$38.38^{\pm 0.13}$	$36.8^{\pm 0.07}$	$40.36^{\pm 0.17}$	$\mathbf{43.67^{\pm 0.32}}$
Translate English to Ateso (salt)	Chrf	$41.98^{\pm 0.13}$	$38.91^{\pm 0.05}$	$39.76^{\pm 1.35}$	$42.1^{\pm 0.42}$	$\mathbf{42.96^{\pm 0.48}}$
Translate Lugbara to Ateso	Chrf	$22.67^{\pm 1.51}$	$20.47^{\pm 0.7}$	$28.13^{\pm 0.58}$	$\mathbf{34.3^{\pm 0.64}}$	$29.04^{\pm 0.3}$
Translate Luganda to Lugbara	Chrf	$28.65^{\pm 1.5}$	$25.74^{\pm 0.5}$	$30.87^{\pm 0.12}$	$34.26^{\pm 0.24}$	$\mathbf{34.94^{\pm 0.6}}$
Translate Luganda to Ateso	Chrf	$31.74^{\pm 0.22}$	$27.66^{\pm 0.64}$	$34.04^{\pm 0.01}$	$37.19^{\pm 0.07}$	$\mathbf{39.05^{\pm 0.49}}$
Translate Nyankore to Lugbara	Chrf	$27.47^{\pm 0.45}$	$24.63^{\pm 0.76}$	$15.01^{\pm 0.01}$	$\mathbf{33.17^{\pm 0.21}}$	$33.2^{\pm 0.19}$
Translate Nyankore to Luganda	Chrf	$39.34^{\pm 0.14}$	$37.34^{\pm 0.16}$	$35.26^{\pm 0.13}$	$40.48^{\pm 0.63}$	$\mathbf{45.29^{\pm 0.01}}$
Translate Nyankore to Ateso	Chrf	$28.6^{\pm 0.11}$	$24.64^{\pm 1.05}$	$30.69^{\pm 0.16}$	$34.37^{\pm 0.14}$	$\mathbf{35.52^{\pm 0.64}}$
	Average	28.07	23.88	22.62	28.77	**34.08**

Table C.4: Performance of various models on MT data using CHRF

Task	Metric	mT0	mT5	afri-mt5	AfriTeVa	Cheetah
Translate English to Afrikaans	Chrf++	$22.86^{\pm 3.74}$	$22.32^{\pm 2.80}$	$11.62^{\pm 6.72}$	$17.27^{\pm 2.91}$	$\mathbf{34.02}^{\pm \mathbf{0.7}}$
Translate English to Bemba	Chrf++	$9.04^{\pm 0.79}$	$5.46^{\pm 1.78}$	$\mathbf{23.65}^{\pm \mathbf{1.87}}$	$7.85^{\pm 7.45}$	$13.9^{\pm 0.13}$
Translate English to Rundi	Chrf++	$18.06^{\pm 1.16}$	$14.41^{\pm 2.53}$	$20.36^{\pm 2.88}$	$\mathbf{25.39}^{\pm \mathbf{1.57}}$	$23.94^{\pm 3.03}$
Translate English to Sesotho	Chrf++	$17.34^{\pm 3.09}$	$10.2^{\pm 8.75}$	$19.31^{\pm 3.94}$	$23.85^{\pm 1.43}$	$\mathbf{23.9}^{\pm \mathbf{2.03}}$
Translate English to Swahili	Chrf++	$18.5^{\pm 0.31}$	$16.28^{\pm 4.48}$	$19.42^{\pm 2.2}$	$16.16^{\pm 3.93}$	$\mathbf{30.6}^{\pm \mathbf{0.11}}$
Translate English to Xhosa	Chrf++	$21.34^{\pm 2.66}$	$19.96^{\pm 4.05}$	$26.94^{\pm 1.92}$	$15.76^{\pm 11.49}$	$\mathbf{27.0}^{\pm \mathbf{1.01}}$
Translate English to Zulu	Chrf++	$21.14^{\pm 2.6}$	$17.32^{\pm 3.17}$	$28.97^{\pm 1.14}$	$19.29^{\pm 12.69}$	$\mathbf{40.97}^{\pm \mathbf{1.10}}$
Translate English to Hausa	Chrf++	$25.98^{\pm 0.27}$	$25.22^{\pm 0.5}$	$18.28^{\pm 0.41}$	$28.56^{\pm 0.22}$	$\mathbf{32.23}^{\pm \mathbf{0.29}}$
Translate English to Igbo	Chrf++	$37.82^{\pm 0.15}$	$34.8^{\pm 0.32}$	$20.25^{\pm 0.68}$	$29.89^{\pm 0.22}$	$\mathbf{41.87}^{\pm \mathbf{0.31}}$
Translate English to Luganda	Chrf++	$23.15^{\pm 2.19}$	$20.74^{\pm 0.36}$	$13.43^{\pm 1.28}$	$20.27^{\pm 2.21}$	$\mathbf{33.12}^{\pm \mathbf{0.08}}$
Translate English to N. Pidgin	Chrf++	$60.57^{\pm 0.15}$	$\mathbf{60.12}^{\pm \mathbf{0.07}}$	$23.85^{\pm 0.64}$	$49.72^{\pm 0.36}$	$59.74^{\pm 0.18}$
Translate English to Swahili	Chrf++	$47.67^{\pm 3.33}$	$48.95^{\pm 0.13}$	$19.01^{\pm 1.69}$	$40.84^{\pm 0.31}$	$\mathbf{53.67}^{\pm \mathbf{0.15}}$
Translate English to Setswana	Chrf++	$29.02^{\pm 0.35}$	$14.87^{\pm 0.16}$	$11.77^{\pm 1.61}$	$21.25^{\pm 0.36}$	$\mathbf{34.05}^{\pm \mathbf{0.64}}$
Translate English to Twi	Chrf++	$21.25^{\pm 0.22}$	$13.63^{\pm 1.18}$	$11.7^{\pm 1.13}$	$15.39^{\pm 3.02}$	$\mathbf{23.96}^{\pm \mathbf{0.2}}$
Translate English to Yoruba	Chrf++	$18.41^{\pm 1.89}$	$15.47^{\pm 0.4}$	$10.19^{\pm 0.25}$	$18.99^{\pm 0.27}$	$\mathbf{24.1}^{\pm \mathbf{0.06}}$
Translate English to Zulu	Chrf++	$30.99^{\pm 1.13}$	$13.86^{\pm 6.85}$	$11.34^{\pm 2.1}$	$10.58^{\pm 0.77}$	$\mathbf{34.31}^{\pm \mathbf{0.2}}$
Translate French to Bambara	Chrf++	$15.75^{\pm 0.36}$	$6.8^{\pm 0.97}$	$10.2^{\pm 1.41}$	$18.28^{\pm 0.49}$	$\mathbf{19.65}^{\pm \mathbf{0.14}}$
Translate French to Ghomálá'	Chrf++	$7.0^{\pm 0.77}$	$5.64^{\pm 0.44}$	$5.84^{\pm 3.04}$	$\mathbf{11.13}^{\pm \mathbf{0.34}}$	$7.28^{\pm 2.83}$
Translate French to Ewe	Chrf++	$9.09^{\pm 2.21}$	$4.75^{\pm 2.76}$	$6.56^{\pm 3.19}$	$11.72^{\pm 1.4}$	$\mathbf{20.53}^{\pm \mathbf{0.23}}$
Translate French to Fon	Chrf++	$5.24^{\pm 2.33}$	$5.57^{\pm 0.63}$	$5.28^{\pm 1.38}$	$10.94^{\pm 1.93}$	$\mathbf{11.76}^{\pm \mathbf{0.45}}$
Translate French to Moore	Chrf++	$7.08^{\pm 1.33}$	$4.63^{\pm 2.02}$	$7.18^{\pm 0.79}$	$10.31^{\pm 0.64}$	$\mathbf{11.2}^{\pm \mathbf{0.54}}$
Translate French to Wolof	Chrf++	$16.27^{\pm 0.24}$	$2.65^{\pm 0.11}$	$10.23^{\pm 1.73}$	$\mathbf{15.73}^{\pm \mathbf{0.33}}$	$15.58^{\pm 0.19}$
Translate English to N. Pidgin (pidginUNMT)	Chrf++	$42.12^{\pm 0.18}$	$37.67^{\pm 1.64}$	$22.53^{\pm 1.31}$	$28.38^{\pm 0.98}$	$\mathbf{39.58}^{\pm \mathbf{0.49}}$
Translate Acholi to English	Chrf++	$37.96^{\pm 0.1}$	$27.18^{\pm 0.36}$	$28.24^{\pm 0.38}$	$31.83^{\pm 0.07}$	$\mathbf{41.06}^{\pm \mathbf{0.06}}$
Translate Acholi to Lugbara	Chrf++	$23.41^{\pm 0.84}$	$19.57^{\pm 5.04}$	$27.18^{\pm 0.36}$	$\mathbf{31.45}^{\pm \mathbf{0.29}}$	$30.68^{\pm 1.02}$
Translate Acholi to Luganda	Chrf++	$25.67^{\pm 0.34}$	$19.59^{\pm 0.56}$	$21.52^{\pm 0.02}$	$28.52^{\pm 0.63}$	$\mathbf{33.93}^{\pm \mathbf{0.48}}$
Translate Acholi to Nyankore	Chrf++	$24.02^{\pm 0.41}$	$17.35^{\pm 3.35}$	$21.38^{\pm 0.23}$	$27.73^{\pm 0.84}$	$\mathbf{31.04}^{\pm \mathbf{0.29}}$
Translate Acholi to Ateso	Chrf++	$23.65^{\pm 1.87}$	$20.07^{\pm 1.53}$	$25.81^{\pm 0.04}$	$31.56^{\pm 0.57}$	$\mathbf{31.83}^{\pm \mathbf{0.99}}$
Translate English to Lugbara	Chrf++	$36.83^{\pm 0.03}$	$\mathbf{38.3}^{\pm \mathbf{0.13}}$	$37.29^{\pm 0.12}$	$34.3^{\pm 0.77}$	$35.85^{\pm 0.01}$
Translate English to Luganda	Chrf++	$40.1^{\pm 0.06}$	$37.56^{\pm 0.19}$	$32.18^{\pm 1.05}$	$38.28^{\pm 0.2}$	$\mathbf{45.82}^{\pm \mathbf{0.04}}$
Translate English to Nyankore	Chrf++	$35.93^{\pm 0.18}$	$34.07^{\pm 0.12}$	$32.59^{\pm 0.05}$	$34.88^{\pm 0.15}$	$\mathbf{39.17}^{\pm \mathbf{0.33}}$
Translate English to Ateso (salt)	Chrf++	$37.98^{\pm 0.11}$	$38.93^{\pm 0.01}$	$36.83^{\pm 1.23}$	$37.85^{\pm 0.4}$	$\mathbf{39.87}^{\pm \mathbf{0.47}}$
Translate Lugbara to Ateso	Chrf++	$20.55^{\pm 1.38}$	$18.54^{\pm 0.65}$	$25.6^{\pm 0.64}$	$\mathbf{30.48}^{\pm \mathbf{0.59}}$	$26.43^{\pm 0.32}$
Translate Luganda to Lugbara	Chrf++	$26.79^{\pm 1.49}$	$23.94^{\pm 0.48}$	$29.13^{\pm 0.11}$	$31.56^{\pm 0.24}$	$\mathbf{33.04}^{\pm \mathbf{0.58}}$
Translate Luganda to Ateso	Chrf++	$28.94^{\pm 0.22}$	$25.11^{\pm 0.59}$	$31.26^{\pm 0.01}$	$33.18^{\pm 0.05}$	$\mathbf{35.99}^{\pm \mathbf{0.45}}$
Translate Nyankore to Lugbara	Chrf++	$22.89^{\pm 0.73}$	$25.75^{\pm 0.44}$	$12.07^{\pm 0.11}$	$30.54^{\pm 0.2}$	$\mathbf{31.35}^{\pm \mathbf{0.2}}$
Translate Nyankore to Luganda	Chrf++	$35.7^{\pm 0.12}$	$33.73^{\pm 0.15}$	$31.99^{\pm 0.07}$	$35.74^{\pm 0.54}$	$\mathbf{41.63}^{\pm \mathbf{0.0}}$
Translate Nyankore to Ateso	Chrf++	$26.03^{\pm 0.08}$	$22.35^{\pm 0.98}$	$28.05^{\pm 0.09}$	$30.53^{\pm 0.13}$	$\mathbf{32.65}^{\pm \mathbf{0.62}}$
	Average	25.58	21.67	20.50	25.16	**31.24**

Table C.5: Performance of various models on MT data using CHRF++

ISO	MT0	MT5	AfriMT5	AfriTeVa	Cheetah
afr	0	0	-	-	**20.45**
amh	0	0	-	0	0
bam	-	-	0	-	0
bbj	-	5.21	0	-	**8.45**
ewe	-	-	0	-	0
fon	-	-	0	-	0
hau	0	0	0	0	**13.41**
ibo	0	0	0	0	0
lin	0	-	-	-	**25.35**
lug	-	-	0	-	0
luo	-	-	0	-	**9.35**
mos	-	-	0	-	**14.53**
mlg	0	0	-	-	**15.65**
nya	-	-	-	-	**7.64**
nyj		-	-	-	
orm	0	-	-	-	0
pcm	-	-	0	0	**10.10**
sna	0	0	-	-	0
som	0	0	-	0	**10.39**
sot	4.69	-	-	-	**15.23**
swa	-	-	0	0	**7.02**
swh			-	-	
tir	-	-	-	-	**6.33**
tsn	-	-	0	-	0
twi	-	-	0	-	0
wol	-	-	0	-	0
xho	0	0	-	-	**6.92**
yor	0	3.61	0	0	**6.42**
zul	0	0	0	-	**8.05**

Table D.1: Bleu scores for mask-one cloze task on the union of languages represented in the four models we compare Cheetah with. Red describes zero-shot performance greater than 0.

ISO	MT0	MT5	AfriMT5	AfriTeVa	Cheetah
afr	0	0	-	-	0
amh	0	0	-	0	0
bam	-	-	0	-	0
bbj	-	-	0	-	0
ewe	-	-	0	-	0
fon	-	-	0	-	0
hau	0	-	0	0	**6**
ibo	0	-	0	0	**8**
lin	0	-	-	-	0
lug	-	-	0	-	0
luo	-	-	0	-	0
mos	-	-	0	-	0
mlg	0	-	-	-	0
nya	0	0	-	-	**12**
nyj		-	-	-	
orm	0	-	0	-	0
pcm	-	-	0	0	0
sna	0	0	-	-	0
som	0	0	-	0	**4**
sot	-	-	-	-	**10**
swa	-	-	0	0	**12**
swh			-	-	
tir	-	-	0	0	0
tsn	-	-	0	-	0
twi	-	-	0	-	0
wol	-	-	0	-	0
xho	0	0	0	-	**6**
yor	0	0	0	0	0
zul	0	0	0	-	0

Table D.2: Bleu scores for mask-at-least-one cloze task on the union of languages represented in the four models we compare Cheetah with.

Category	Example
Intransitive	He left
Intransitive + Negation	We did not leave
Transitive	You left Lagos
Transitive + Negation	She did not leave them

Table E.1: Some examples of sentences generated with the templates

Lang.	Family	# Tone	Gender	Morphology
Hausa	Afro-Asiatic	Two	Two	Isolating
Swahili	N.C. Bantu	None	Five	Agglutinative
Yourba	N.C. Non-Bantu	Three	None	Isolating

Table E.2: Some linguistic differences between Hausa, Swahili, and Yoruba. **N.C.** refers to Niger-Congo

	hau		yor	
Model	Faith.	Flu.	Faith.	Flu.
mT0	90.54	97.62	96.57	93.92
mT5	93.51	96.48	82.23	81.10
AfriTeVa	87.27	96.94	88.56	84.73
Cheetah	96.61	97.26	87.11	92.64

Table E.3: Kappa scores for Faithfulness (i.e., Faith.) and Fluency (i.e., Flu.) across the four models and three languages we evaluate.

G Results on Quality Evaluation

Figure G.1: Faithfulness and fluency for Intransitives in Hausa, Swahili, and Yorùbá

Figure G.2: Faithfulness and fluency for Transitives in Hausa, Swahili, and Yorùbá correct | plausible | wrong / irrelevant

English	Yoruba Gold	AfriTeVa	mt0	mt5	Cheetah
Sara did not leave	Sara ò kúrò	sara kò ní kúrò nínú oyè	Sara ò sì lọ.	Sara ò sì lọ sí ìlú.	Arábìnrin Sara kò fi ara rẹ̀ sílẹ̀.
Sara would not leave	Sara ò ní kúrò	sara yóò kojú àdánwò	Sara ò ní ìgbà yìí.	Sara ò sì lọ sílẹ̀.	Arábìnrin Sara ò ní lọ sílé.
Fred did not leave	Fred ò kúrò	omíyalé ò jáde	Fred ò sì lọ.	Ààrẹ Fred kò sí ìfẹ́ ẹ̀wọ̀n	Fred kò kúrò nílé.
Fred would not leave	Fred ò ní kúrò	ẹ máa wàásù láti ọwọ́ọ ṣìgún òfin	Fred ò ní ìgbà yìí.	Ààrẹ ò ní ìfẹ́ ẹ̀yìn àti ìfẹ́ ẹ̀yìn rẹ̀.	Ó ní Fred ò ní lọ kúrò nílé.
you did not leave	o ò kúrò	au kò ní kúrò	Ẹ ò sì lọ.	Àmọ́ kò sì lọ́wọ́ lọ́wọ́ àwọn ará wa.	Ó tiẹ̀ jẹ́ pé kò sí ìkankan lára àwọn tí wọ́n ń gbé.
you would not leave	o ò ní kúrò	ẹ máa wàásù níbi tí kò séwu	Ẹ ò ní ìgbà yìí.	Àmọ́ kò ní ìgbà yìí.	A ò ní lọ kúrò nílé
he did not leave	kò kúrò	ò rò òdodo kò ní kúrò	Ò sì lọ́wọ́.	Àmọ́ kò sí ìgbà tí ó sì lọ.	Ó kọ̀ láti kúrò nílé.
he would not leave	kò ní kúrò	ò tîì ṣì gbọdọ̀ jáde láyé	Ò ní ìgbà yìí.	Ìgbà tí ó sì ń lọ nínú ìjọba yìí.	Ó ò ní lọ kúrò nílé.
she did not leave	kò kúrò	a ò ní kúrò nínú ológun	Òun ò sì lọ.	Àmọ́ kò sí i lọ́wọ́ ìjọba.	Ó kọ̀ láti kúrò nílé.
she would not leave	kò ní kúrò	kò sí alẹ́ láti kúrò níbi tí wọ́n ti ń gbé.	Ò ní ìgbà yìí.	Ẹ̀rí ìwà rẹ̀ yóò sì lọ́wọ́ nínú ìjọba.	Ó ò ní lọ kúrò nílé.
they did not leave	àwọn ò kúrò	a ò ní kúrò lẹ́wọ̀n	Ẹ̀ sì sì lọ́wọ́.	Àwọn òun kò sì lọ́wọ́ wọn.	Wọ́n ò fi wọ́n sílẹ̀, wọ́n sì kọ̀ láti kúrò nílé
they would not leave	àwọn ò ní kúrò	ṣé o ò ní kúrò nílé	Ẹ ò ní ìgbà yìí.	Àwọn ará ò ní ìgbà wọn nínú ìjọba.	Wọ́n ò ní lọ sílé, wọ́n ò sì ní lọ.
I did not leave	Èmi ò kúrò	ò ní kúrò nínú oyè	Nígbà tí mo kò lọ.	À sì sì lọ́wọ́ lọ́wọ́ lọ́wọ́ lọ́wọ́ lọ́wọ́.	Mo ò fi ara mi sílẹ̀, mo sì kọ̀ láti lọ.
I would not leave	Èmi ò ní kúrò	ẹ máa bẹ̀rù.	Mo ò ní ìgbà yìí.	À ò ní ìlú yìí.	Mo ò ní lọ kúrò nílé
we did not leave	Àwa ò kúrò	a ò so	Ẹ ò sì lọ́wọ́.	Àmọ́ kò sì lọ́wọ́ àwọn ará wa.	A ò fi ara wa sílẹ̀, a ò sì fi ara wa sílẹ̀.
we would not leave	Àwa ò ní kúrò	a ò lè sọ̀rọ̀ yìí	Ẹ ò ní ìgbà yìí.	Ìjọba kò ní ìlú yìí.	A ò ní lọ kúrò nílé, a ò sì ní lọ kúrò nílé

Figure G.3: Performance on some intransitive examples in the Yorùbá test set. The correct words have no highlights, plausible words or phrases are highlighted with yellow ink while wrong words and phrases are highlighted with grey highlights. We use plausible to refer to words or phrases that can be used in place of the gold or which give additional information.

Figure G.4: Faithfulness and fluency for Intransitives + Negation in Hausa, Swahili, and Yorùbá

Figure G.5: Faithfulness and fluency for Transitives + Negation in Hausa, Swahili, and Yorùbá