Scaling Pre-training Data and Language Models for African Languages

by

Akintunde Oladipo

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2024

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Chapters 3, 4 and 5 are based on the co-authored work published in Ogundepo et al. (2022) [30] and Oladipo et al. (2023) [45]. I declare that I am responsible for the code contribution, conducting of experiments, and paper writing.

Abstract

Recent advancements in language models, particularly for high-resource languages, have not been paralleled in low-resource languages spoken across Africa. This thesis addresses this gap by scaling pre-training data and developing improved language models for African languages. We introduce Wura, a high-quality, document-level pre-training dataset encompassing 16 African languages along with four high-resource languages commonly spoken on the continent: Arabic, English, French, and Portuguese. Leveraging Wura, we pre-train new versions of the AfriBERTa (encoder-only) and AfriTeVa (encoder-decoder) model families. These new models demonstrate superior performance across a variety of natural language understanding and generation tasks compared to existing baselines. Notably, AfriTeVa V2 Large (1B) stands as the largest sequence-to-sequence model pre-trained for African languages to date.

Our methodology includes a meticulous three-stage curation process for Wura— auditing and filtering existing web crawls, initiating new web crawls, and integrating existing language resources. The experimental setup and evaluation encompass tasks like text classification, information retrieval, translation, summarization, and cross-lingual question answering. Our new models outperform their predecessors and other established models, even those with significantly more parameters, highlighting the efficacy of high-quality pre-training data. Furthermore, we study the generalization of our models to languages not deliberately included in their pre-training data.

Acknowledgements

I must thank Professor Jimmy Lin whose focus on natural language processing research for African languages has made this thesis possible. I would like to thank him for the interesting questions he posed, which kept me up many nights, guided my research efforts and spurred my collaboration with other exceptional researchers.

I want to express my gratitude to the readers of my thesis, Professor Charles Clarke and Professor Jian Zhao, for taking the time out to review my work and for their valuable insights.

I am privileged to have David Adelani, Orevaoghene Ahia, Kelechi Ogueji, Odunayo Ogundepo and Mofetoluwa Adeyemi: collaborators, senior colleagues and mentors. I also acknowledge members of the Data Systems Group (DSG) for many insightful discussions and useful resources.

Finally, my special thanks to my family for their unwavering support; to Esther Oshinaike and Sydney Okoroafor for their encouragement and love throughout these years. Àwarawa — you are family too.

Dedication

This is dedicated to Pelumi, Faisal, Bayo, Kenny, Blessing, Dayo. Our friendship changed our lives and inadvertently made this work possible.

Table of Contents

List of Tables

Chapter 1

Introduction

Language models have scaled up in size and improved in capability significantly in recent years. This progress is due to a combination of improved computing infrastructure, painstaking curation of quality pretraining data [19, 21, 52, 56, 58] and eventual human preference alignment of the trained models through instruction finetuning [57, 62, 69]. Commensurate work has also followed exploring mathematical laws that govern the scaling of these models [26, 31], particularly for the transformer-based language models.

These models are typically categorized into different architectures: encoder-only, decoder-only, and encoder-decoder models. Encoder-only models, like BERT [16], are primarily used for tasks requiring text understanding, such as classification and entity recognition. Decoder-only models, like GPT [50], excel at text generation and are popularly used for chatbots. Encoder-decoder models, like T5 [52] and BART [35], are versatile, performing well in both understanding and generation tasks, making them suitable for applications like translation and summarization. Today a plethora of open-source and proprietary language models exist, boasting billions of parameters, trained on trillions of tokens and excelling on benchmarks for natural language understanding, reasoning, code generation, etc.

Despite their impressive capabilities, these models primarily focus on high-resource languages which have abundant corpora to support the training data requirements of large language models. This limitation restricts their utility for low-resource languages [44]. Researchers have already demonstrated the viability of pre-training or adapting multilingual language models for African languages [9, 18, 42, 38]. These models are competitive with or outperform state-of-the-art language models on a large variety of natural language processing tasks despite having fewer parameters and being pre-trained on significantly smaller corpora. It is thus clear that scaling pre-training data as well as the size of these

specialized language models is the logical next step in the advancement of language models for African languages.

Already, this endeavour has been attempted in different forms. Adebara et al. (2023) [2] aggregate pre-training data up to 42GB by expanding to 517 languages, including 10 non-African languages. Jacaranda Health (2023) also introduce UlizaLlama [23], an 8B parameter language model adapted for Swahili from Llama 2 [60] by expanding its tokenizer vocabulary and pre-training on 321M Swahili tokens. In contrast to these efforts, we propose to scale pre-training data for 16 focus African languages and to pre-train language models from scratch to serve the populations that speak these languages.

In this thesis, we describe and expand our work improving language models for African languages. We scale the monolingual pre-training corpora available for 16 African languages and introduce improved second generations of the AfriBERTa [42] and AfriTeVa [30] families of language models. These new models are evaluated extensively on a diverse range of natural language understanding and generation tasks. We also study their generalization, in zero-shot and full fine-tuning scenarios, to 10 other African languages not deliberately included in their pre-training data.

Parts of this work were done in collaboration with other researchers and presented at the 2022 Workshop on Deep Learning for Low-Resource Natural Language Processing [43] and the 2023 Conference on Empirical Methods in Natural Language Processing [45].

1.1 Contributions

The main contributions of this thesis are summarized below:

1. We release Wura, a high-quality document-level pre-training dataset with 1.5× more data across 16 African languages compared to previously existing multilingual pre-training datasets. With it, we release the web crawling framework and data audit pipeline used for its curation.

2. We pre-train multilingual BERT-style language models from scratch as successors to the AfriBERTa model family.

3. We pre-train new multilingual sequence-to-sequence T5 models, AfriTeVa. With this we introduce, to our knowledge, the largest sequence-to-sequence models pre-trained for African languages.

4. We demonstrate the practical benefits of careful curation of quality pre-training data for African languages through improved effectiveness compared to existing baselines on a expanded set of evaluation tasks.

5. We release final checkpoints of our pre-trained AfriBERTa and AfriTeVa models. We also make intermediate checkpoints of our models available for researchers studying how learning, memorization, and other such properties emerge in language models pre-trained for African languages.

1.2 Thesis Organization

This thesis is organized as follows:

1. In Chapter 2, we cover background knowledge and related work that contextualizes our work.

2. Chapter 3 describes Wura: our vision for the pre-training corpora, its curation and audit, its properties and how it compares to other pretraining corpora available for African languages.

3. Chapter 4 describes our experimental setup, evaluation tasks and datasets, as well as implementation details of note for reproducibility.

4. Chapter 5 examines our experimental results in-depth and compares them to existing baselines. We also highlight important findings and discuss their implications.

5. Finally, Chapter 6 summarizes our work, highlights its main contributions and proposes future research directions.

Chapter 2

Background and Related Work

2.1 Language Modelling

The objective of language modelling is to learn a probability distribution over a defined vocabulary of tokens. This capability is crucial for a variety of natural language processing (NLP) tasks, including machine translation, text generation, and speech recognition. Early language models, such as n-gram models, relied on fixed-size context windows to predict the next word. These models faced limitations in capturing long-range dependencies due to their reliance on a limited context. To address these limitations, neural network-based approaches were introduced, such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, which can capture dependencies over longer sequences and were trained on large datasets [39].

A significant breakthrough in language modelling came with the introduction of the Transformer architecture [61]. Transformers rely on self-attention mechanisms to process entire sequences in parallel, enabling them to capture long-range dependencies more effectively than RNNs or LSTMs. The Transformer architecture can be categorized into three main types based on how its components are utilized: encoder-only models, decoder-only models, and encoder-decoder models.

- **Encoder-only models** focus solely on the encoding part of the Transformer architecture. These models, such as BERT [16], are designed to generate a contextualized representation of the input text. The encoder stack, consisting of multiple layers of self-attention and feed-forward networks, processes the input sequence to produce embeddings that capture the contextual information for each token. Encoder-only

models are particularly effective for tasks that require understanding the entire input context, such as text classification, named entity recognition, and sentence embedding.

- **Decoder-only models**, exemplified by GPT [50], utilize only the decoder component of the Transformer architecture. These models are auto-regressive, meaning they generate text one token at a time while conditioning on the previously generated tokens. The self-attention mechanism in the decoder allows the model to attend to all previous tokens in the sequence, making it adept at tasks like language modeling and text generation. Decoder-only models are highly effective for tasks that involve generating coherent and contextually relevant text based on a given prompt or context.

- **Encoder-decoder models** combine both the encoder and decoder components of the Transformer architecture, as seen in models like T5 [52]. In this configuration, the encoder processes the input sequence to produce a set of contextual embeddings, which are then fed into the decoder. The decoder generates the output sequence by attending to both the encoder's embeddings and the previously generated tokens. This architecture is versatile and well-suited for a wide range of sequence-to-sequence tasks, such as machine translation, text summarization, and question answering. The encoder-decoder setup allows the model to effectively transform an input sequence into a different output sequence, leveraging the strengths of both the encoder and decoder components.

In this work, we focus on encoder-only and encoder-decoder models, specifically AfriBERTa and AfriTeVa, which leverage the unique strengths of their architectures: encoder-only models excel at understanding and representing text, while encoder-decoder models offer flexibility and robustness for tasks requiring both comprehension and generation of text.

The versatility and effectiveness of the Transformer architecture has led to its widespread adoption and the development of numerous state-of-the-art models across various NLP tasks, as evidenced by the success of landmark language models, such as BERT, GPT, and T5. Equally important are the objectives with with these language models are trained. Objectives provide the model with learning mechanisms that it can generalize to downstream tasks. BERT was pre-trained with Masked Language Modelling (MLM) objective in which a percentage of the input tokens are randomly masked, and the model is trained to predict masked tokens based on surrounding context. This objective was combined with the next-sentence prediction (NSP) objective, enabling the model to learn bidirectional

representations for tokens as well as the relationship between sentences. For sequence-to-sequence models, Raffel et al. (2020) [52] introduce a *span corruption* version of MLM and perform an exhaustive comparison of multiple pre-training objectives. Lastly, generative models like GPT use Causal Language Modelling (CLM) in which the model is tasked to auto-regressively predict the next token in a sequence. We refer the interested reader to these works.

2.2 Pre-training Data for African Languages

Earlier multilingual models such as mBERT [16] and XLM-R [14] were trained on monolingual data from Wikipedia and/or other large-scale web crawls which included only a few African languages. OSCAR [1, 46], a deliberate effort towards multilingual web resources, includes 6 African languages, three of which have roughly 1000 documents. All 6 languages covered amount to less than 200MB.

The introduction of mC4 [64], a document-level dataset spanning 101 languages helped alleviate this coverage gap. However, previous work [33] has shown that mC4 and other existing large-scale pre-training corpora have numerous quality issues, particularly for the low-resource African languages they contain. Still, mC4 remains one of the largest sources for pre-training data for African languages.

While researchers have released smaller datasets for specific languages [9, 38, 42], scaling pre-training data for African languages remains an active research area. This research work aims to alleviate this issue. Table 2.1 provides information about the languages we cover in this work (through pre-training or evaluation)—their scripts, language families, number of native speakers and resource rating according to Joshi et al. (2020) [29].

2.3 Multilingual Language Models

The effectiveness and versatility of multilingual BERT (mBERT), pre-trained on Wikipedia corpus for 100 languages, laid the foundation for the development of more multilingual models able to learn cross-lingual representations [13, 14]. One of the primary advantages of these models is their ability to transfer knowledge across languages [27, 49]. This transfer learning ability is particularly valuable for language with few resources—a category into which African languages fall.

Prior to the release of multilingual T5 (mT5) [64], early multilingual models covered only a handful of African languages. Meanwhile, indigenous efforts to build language resources for Africa converged to two approaches:

1. **Small high-quality data** (e.g., 1GB) pretraining where most data are from the clean or verified sources like news domain [42].
2. **Large aggregation of all available data** (e.g., 15−42 GB) from noisy or unverified sources like CC-100 [15], and mC4, combined with high-quality sources like news corpora [2, 3, 9].

This trade-off between quantity and quality is forced by the unavailability of large, quality pre-training data for African languages.

Ogueji et al. (2021) [42] were first to demonstrate the viability of pre-training language models on small high-quality corpora for a subset of related African languages. The authors studied the effect of various architectural hyperparameters such as vocabulary size, number of layers, and number of attention heads on the effectiveness of multilingual models. Their work yielded AfriBERTa, a family of encoder models, competitive with both mBERT and XLM-R despite having more than 28% fewer parameters.

Dossou et al. (2022) [18] explored how data-efficiency through self-active learning can improve effectiveness of models pre-trained using the AfriBERTa approach. The authors applied iterative next-token prediction to generate new training examples for each active learning round. While they demonstrate improvement over AfriBERTa baselines, research is still needed to disentangle the effects of their increased model size (264M vs. 126M) from that of their proposed active learning approach.

In earlier work, we extended this small-data pre-training approach to sequence-to-sequence models, AfriTeVa [30], with modest success. Notably, our research demonstrated that including English in the pre-training data significantly improved the effectiveness of our models across multiple tasks.

Other researchers have taken the adaptation approach to creating multilingual language models for African languages. Adelani et al. (2022) [3] adapted mT5 [64], ByT5 [63] and mBART [37] through continual pre-training on 16 African languages. They demonstrated improved effectiveness on machine translation over the base models.

Alabi et al. (2022) [9] adopted a different approach when adapting AfriBERTa and XLM-R models to new languages. The authors specialized each model's tokenizer vocabulary before continual pre-training. While their adaptation approach yielded mixed results

for AfriBERTa, they achieve significant gains for XLM-R base and large on the evaluation tasks they consider. For many natural language processing tasks for African languages, Afro-XLMR-large is state-of-the-art [4, 5].

Finally, multilingual language models have been trained for specific regions or countries in Africa [10, 59] and foundation models such as GPT-4 and Llama which have billions of parameters, have been shown to exhibit varying levels of effectiveness on benchmark tasks for African languages [6, 44].

S/N	ISO Code	Language	Script	Family	Resources	# of Speakers
1	afr	Afrikaans	Latin	Indo-European	Mid	7M
2	amh	Amharic	Ge'ez	Afro-Asiatic	Low	35M
3	arz	Egyptian Arabic	Arabic	Afro-Asiatic	Low	78M
4	bbj	Ghomálá'	Latin	Atlantic-Congo	Low	350K
5	bem	Bemba	Latin	Atlantic-Congo	Low	4M
6	eng	English	Latin	Indo-European	High	380M
7	ewe	Ewe	Latin	Atlantic-Congo	Low	4M
8	fon	Fon	Latin	Atlantic-Congo	Low	2M
9	fra	French	Latin	Indo-European	High	74M
10	fuv	Nigerian Fufulde	Latin	Atlantic-Congo	Low	37M
11	hau	Hausa	Latin	Afro-Asiatic	Low	63M
12	ibo	Igbo	Latin	Atlantic-Congo	Low	27M
13	kin	Kinyarwanda	Latin	Niger-Congo	Low	15M
14	lin	Lingala	Latin	Niger-Congo	Low	20M
15	lug	Luganda	Latin	Niger-Congo	Low	6M
16	luo	Luo	Latin	Nilotic	Low	4M
17	mlg	Malagasy	Latin	Austronesian	Low	25M
18	mos	Mossi	Latin	Atlantic-Congo	Low	7M
19	nso	Northern Sotho	Latin	Atlantic-Congo	Low	5M
20	nya	Chichewa	Latin	Atlantic-Congo	Low	7M
21	orm	Afaan Oromoo	Latin	Afro-Asiatic	Low	46M
22	pcm	Nigerian Pidgin	Latin	English Creole	Low	5M
23	por	Portuguese	Latin	Indo-European	High	260M
24	sna	Shona	Latin	Indo-European	Low	7M
25	som	Somali	Latin	Afro-Asiatic	Low	24M
26	sot	Southern Sesotho	Latin	Atlantic-Congo	Low	6M
27	ssw	Swati	Latin	Atlantic-Congo	Low	2M
28	swa	Swahili	Latin	Niger-Congo	Low	98M
29	tir	Tigrinya	Ge'ez	Afro-Asiatic	Low	10M
30	tsn	Setswana	Latin	Atlantic-Congo	Low	4M
31	twi	Twi	Latin	Atlantic-Congo	Low	9M
32	wol	Wolof	Latin	Atlantic-Congo	Low	7M
33	xho	Xhosa	Latin	Atlantic-Congo	Low	8M
34	yor	Yorùbá	Latin	Atlantic-Congo	Low	42M
35	zul	Zulu	Latin	Atlantic-Congo	Low	27M

Table 2.1: Languages Information. Number of Speakers reported is from the Wikipedia page for each language.

Chapter 3

WURA

WURA[1] is an actively-maintained document-level multilingual dataset comprising 16 African languages and 4 high-resource languages popularly spoken on the African continent – Arabic, English, French, and Portuguese. The curation of WURA was carried out in a three-part process:

- Auditing and cleaning mC4.
- Crawling indigenous websites.
- Combination with existing language resources.

In this chapter, we provide an overview of the WURA pre-training dataset and a detailed description of the curation process. Its statistics and properties are also discussed.

3.1 Auditing and Cleaning mC4

Kreutzer et al. (2021) [33] reported high ratio of non-linguistic content and sentences in incorrect languages in mC4, with African languages being of particular concern. Further, they found that using automatic language classification [12] to filter out such contents yielded mixed results. The authors report significant loss (up to 50%) in recall of correct

[1]Wura means Gold in Yoruba – with more refining, the quality of our data and model improves.

in-language sentences as they increased precision of their automatic language classification. These findings guide our own audit and subsequent filtering of mC4.

We aim to tease out heuristics that are guaranteed to help us quickly and reliably extract high-quality monolingual text across the African languages in mC4. First, we reduce the source URL of each document to its hostname[2] and keep a list of unique hostnames that exist for each language. For each language, we first sample a hostname then sample 20 documents sourced from the sampled hostname. This sampling strategy not only allows to audit more documents and sources faster, it allows us trace existing quality issues to the source URLs that produced the documents. We follow non-expert auditing strategies proposed by Kreutzer et al. (2022) [33]. Additionally, we also visit the hostname URL to ascertain its purpose for speakers of the language and translate paragraphs in the document using Google Translate. Some hostnames may have moved to new addresses or shut down permanently. In such cases, we check the Internet Archive[3].

Our manual audit of mC4 corroborates the documented issues. In addition, we highlight three important findings:

- The distribution of mC4 document sources has a long tail. Many individual news publications yield thousands of documents in the mC4.
- Documents from news publications are more likely to be of higher quality i.e., both in-language and grammatical compared to documents from other web sources.
- Some documents are from websites which translate content using online translation tools. Such documents are often a mix of in-language and noisy or non-linguistic text, and may best be filtered at sentence-level.

These issues and findings inform our filtering pipeline which we discuss in the following sub-sections.

3.1.1 Corpus-level Filtering

We first rank unique websites in descending order of the number of documents they contribute to the mC4 corpus for each language. Then, we select the top 20% of websites for each language and collect documents sourced from websites in this list. This preserves high potential sources for further document-level filtering. Note that this trades off recall for precision.

[2]The hostname property of the URL interface is a string containing the domain name of the URL
[3]https://archive.org/

3.1.2 Document-level Filtering

In our work, we found that document sources in mC4 often contained documents in a mix of languages — usually the language mC4 reported and a high-resource language such as English or Portuguese which is spoken in the associated country. Note that this form of language contamination is relatively benign and has been reported to help cross-lingual generalization of models pre-trained on such data [11, 43, 66].

Nevertheless, our aim is to curate high-quality monolingual corpora across these languages so we filter out documents that do not contain at least 5 stopwords in them [12] using stopwords from Stopword Lists for African Languages dataset.[4] We found stopword filtering to be a low-cost yet effective method to filter in-language documents for our corpora.

3.1.3 Passage-level Filtering

After document-level filtering, we chunk the dataset into passages of roughly 512 tokens. Finally, we filter out passages that contain fewer than 4 unique words or contain repetition for more than 20% of its word length; have more than 40% of its characters are numeric or contain markers of possibly offensive content such as included in the Toxicity-200 dataset [41] for the relevant language. While Kreutzer et al. (2022) [33]'s audit of mC4 did not yield a significant amount of offensive content (0.06% of sentences they audited) and our web crawls mainly focused on verified news publications, these filters ensure that non-linguistic and offensive contents are removed at the passage level.

3.2 Web Crawling: mC4 is a Great Source!

The inclusion of each document's source URL makes the mC4 corpus even more useful as a data source. Commonly, multiple articles are collected from the same base website, e.g., news publications. For many news publications that provide a sitemap, we find that there are fewer articles in mC4 than is actually available on the websites. Further, mC4's cut-off data is August, 2020 so updating the crawls up to the current day yields more data.

We initiate focused crawls for such websites and this leads to significant increase ($>$ 100% for Hausa and Somali) in the amount of articles available per language. For all

[4] https://www.kaggle.com/datasets/rtatman/stopword-lists-for-african-languages

Language	# mC4 Articles	# After URL Filter	# After Stopword Filter	% Articles Retained
Afrikaans (afr)	2,152,243	1,898,152	978,740	45.5
Amharic (amh)	162,870	152,937	112,843	69.3
Chichewa (nya)	174,696	138,265	42,917	24.6
Hausa (hau)	247,507	223,290	147,028	59.4
Igbo (ibo)	92,909	72,779	34,802	37.5
Malagasy (mlg)	345,040	287,979	110,841	32.1
Sesotho (sot)	66,837	56,938	41,547	62.2
Shona (sna)	326,392	262,698	48,337	14.8
Somali (som)	893,012	843,936	513,028	57.5
Swahili (swa)	985,654	956,645	831,162	84.3
Xhosa (xho)	69.048	53,954	24,992	36.2
Yoruba (yor)	46,214	37,219	20,463	44.3
Zulu (zul)	555,458	447,494	61,387	11.1

Table 3.1: mC4 Filtering Statistics: We provide the count of crawled articles, Wikipedia articles, original mC4 articles, and final size before passage-level filtering for each language. In total, we have ∼4.7M articles, more than 1.5 times what mC4 contains across 16 African languages.

languages we consider except Chichewa, Sesotho, Xhosa and Zulu, we collect 1.39M articles (see Table 3.2) from credible sources found in mC4. For some Somali and Swahili news publications, we are able to collect up to 50 times more articles.

We open-source Otelemuye,[5] an extensible framework for large scale web-crawls. In our work, we crawl at a safe pace that does not degrade the website's performance and respect the rules websites publish in their robots.txt.[6]

3.3 Combination with Existing Language Resources and Non-African Languages

Our aim is to introduce broadly useful language models for Africa. Following previous works [2, 9], we include certain non-African languages in our pre-training data. Specifically,

[5] https://github.com/theyorubayesian/otelemuye

[6] https://developers.google.com/search/docs/crawling-indexing/robots/intro

we include over 240,000 articles newly crawled from 10 African news websites reporting in English, French and Portuguese. We also include a sample of 1.5M Wikipedia articles for English and French, as well as Wikipedia articles written in Egyptian Arabic. For the African languages, we include all Wikipedia articles. Finally, we de-duplicate using the document URLs. In doing this, we prioritize news articles in our focused crawls over their existing counterparts in mC4.

3.4 Final Dataset Statistics

Table 3.2 presents a statistical summary of our dataset. The combined dataset from crawling, combining with existing sources and deduplication amounts to ∼30GB of data across all languages and ∼19GB for African languages. WURA is publicly available through HuggingFace Hub[7]. We release both document-level and passage-level versions.

Where possible, we include the category under which each document was published. This information may be useful for identification of the domains in our dataset. We also release a list of the top document URLs for each language[8] and invite native speakers to audit these sources to help us improve the quality of WURA.

[7] https://huggingface.co/datasets/castorini/wura
[8] https://github.com/castorini/AfriTeVa-keji#dataset

	African Languages in mC4					
Language	**# Crawled Articles**	**# Wikipedia Articles**	**# mC4 Articles**	**# Combined Articles**	**# De-duped Articles**	**Size (GB) Articles**
Afrikaans (`afr`)	139,977	107,860	978,740	1,226,577	1,158,680	4.8
Amharic (`amh`)	22,831	15,713	112,843	151,387	150,958	1.2
Chichewa (`nya`)	—	1,135	42,917	44,052	44,052	0.4
Hausa (`hau`)	247,507	25,957	147,028	420,492	399,866	0.9
Igbo (`ibo`)	6,196	16,158	34,802	57,156	57,095	0.2
Malagasy (`mlg`)	35,839	95,612	110,841	242,292	240,233	0.5
Sesotho (`sot`)	—	1,076	41,547	42,623	42,623	0.2
Shona (`sna`)	10,637	10,847	48,337	69,821	67,762	0.5
Somali (`som`)	585,928	11,241	513,028	1,110,197	1,084,982	2.3
Swahili (`swa`)	265,733	77,017	831,162	1,173,912	1,151,393	3.5
Xhosa (`xho`)	—	1,554	24,992	26,546	26,546	0.1
Yoruba (`yor`)	28,463	32,915	20,463	81,841	81,632	0.1
Zulu (`zul`)	—	11,331	61,387	72,718	72,718	0.7
	African Languages not in mC4					
Afaan Oromoo (`orm`)	18,675	1,535	—	22,410	22,410	0.06
Kinyarwanda (`kin`)	17,218	7,423	—	32,437	32,437	0.10
Tigrinya (`tir`)	8,728	427	—	9,155	9,155	0.03
Total	1,393,097	422,536	2,968,087	4,793,623	4,652,549	18.9
	Other Languages					
Arabic (`arz`)	—	1,617,402	—	1,617,402	1,617,402	0.72
English (`eng`)	31,727	1,500,000	—	1,531,727	1,531,727	4.0
French (`fra`)	103,529	1,500,000	—	1,603,529	1,603,529	3.6
Portuguese (`por`)	107,670	1,102,551	—	1,210,221	1,210,221	2.3
Total	1,636,023	6,142,489	2,968,087	10,756,502	10,615,428	29.5

Table 3.2: WURA Dataset Statistics: We provide the count of crawled articles, Wikipedia articles, original mC4 articles, and final size before passage-level filtering for each language. In total, we have $\approx$ 4.7M articles, more than 1.5 times what mC4 contains across 16 African languages.

Chapter 4

Experimental Setup

4.1 Tokenization

In multilingual settings, the design of tokenizers has great impact on the downstream utility and cost of inference of language models across languages [8, 48]. We characterize the performance of our tokenizers using *fertility* [68], defined as the number of subwords created per word (or per dataset) by the tokenizer. We compute fertility on the subset of languages in Wura covered by MasakhanePOS [17].

We train multiple unigram language models on our dataset using Sentencepiece [34] with vocabulary sizes ranging from $100,000$ to $250,000$. As shown in Table 3.2 above, our dataset sizes varies over orders of magnitude between languages. To alleviate unfair treatment of the lowest-resourced of the languages we consider, we follow Conneau et al. (2019) [15] to learn the unigram language models on sentences sampled according to a multinomial distribution with probabilities q_i calculated as follows:

$$q_i = \frac{{p_i}^{\alpha}}{\sum_{j=1}^{N} p_j^{\alpha}} \quad \text{where} \quad p_i = \frac{n_i}{\sum_{k=1}^{N} n_k} \quad \text{and} \quad \alpha = 0.3 \tag{4.1}$$

N denotes the number of languages and n_i, the number of sentences in language i. We denote this as sampling configuration ①. We also investigate a sampling configuration ② in which we further upsample languages which still do not have adequate representation after sampling sentences with the calculated probabilities. Simply, after calculating probabilities using ①, we upsample by a factor of 10 for `ibo`, `kin`, `nya`, `sna`, `sot`, `tir`, `xho`, and a factor of 5 for `amh`, `arz`, `mlg`, `som`. We make this choice of upsampling factor taking into

Sampling	Vocab Size	hau	ibo	kin	nya	sna	swa	xho	yor	zul
Config ①	100,000	1.29	1.62	1.80	1.90	1.76	1.24	2.37	2.05	2.22
	150,000	1.25	1.53	1.67	1.74	1.64	1.21	2.20	1.97	2.06
	200,000	1.23	1.49	1.57	1.67	1.56	1.19	2.10	1.92	1.96
	250,000	1.22	1.47	1.54	1.63	1.53	1.19	2.03	1.90	1.91
Config ②	100,000	1.25	1.43	1.52	1.65	1.54	1.29	2.07	1.67	1.90
	150,000	1.21	1.39	1.43	1.51	1.45	1.25	1.94	1.59	1.77
	200,000	**1.20**	**1.37**	**1.38**	**1.45**	**1.38**	**1.23**	**1.86**	**1.55**	**1.69**

Table 4.1: Tokenizer Fertility: We measure the fertility of our tokenizers with varying vocabulary sizes using the MasakhanePOS dataset. The 150k tokenizer gives the best trade-off in size and fertility scores across all languages, especially in the second sampling configuration.

consideration the maximum amount of data we can train unigram language models with given our CPU resources.

The fertility of tokenizers trained on the sentences obtained by both sampling configurations are presented in Table 4.1. Sampling configuration ② already yields better ferility at $150,000$ vocabulary size than configuration ① does at $200,000$.

4.2 Models

We pre-train base (428M parameters) and large (1B parameters) AfriTeVa V2 models, successors to AfriTeVa, using t5x and seqio [54]. Unlike their predecessor, AfriTeVa V2 models utilize the T5 1.1 [52, 55] encoder-decoder transformer architecture. Pre-training is done with the span-corruption objective and the Adafactor optimizer.

For AfriBERTa V2 models, we pretrain base (173M parameters) and large (187M parameters) configurations using Flax [24] on the standard masked language modelling objective [15] with 15% of input tokens randomly masked. Architectural details are presented in Table 4.2.

All our models use the trained sub-word tokenizer of vocabulary size $150,000$ described in section 4.1. We pre-trained the models for $524,288$ steps, taking checkpoints every $50,000$ steps. Each training batch consists of 512 examples, each with an input of 512 tokens. For AfriTeVa V2 models, the output is 114 tokens.

Model	# of Layers	# Attention Heads	# of Parameters	Vocabulary Size	# of Languages
AfriBERTa Base	8	6	111M	70,000	10
AfriBERTa Large	10	6	126M	70,000	10
AfriBERTa V2 Base	8	6	173M	150,000	21
AfriBERTa V2 Large	10	6	187M	150,000	21
AfroXLMR Base	12	12	270M	250,002	100(17)[b]
AfroXLMR Large	16	24	550M	250,002	100(17)[b]
AfriTeVa Base	12	12	229M	70,000	10
AfriTeVa Large	24	16	745M	70,000	10
AfriTeVa V2 Base	12	12	428M	150,144	20
AfriTeVa V2 Large	24	16	1.2B	150,144	20
AfriMT5 Base	12	12	582M	250,112	101(17)[b]
AfriByT5 Base	18/6[a]	12	582M	384	101(17)[b]

Table 4.2: Model Configurations:
[a] ByT5 models are encoder-heavy: 18 encoder layers, 6 decoder layers; [b] These models were adapted to cover 17 African languages from base models which initially covered 100 or more languages

We employ data packing during training resulting in a training budget of 8.9M samples (roughly 4.6B tokens) per epoch. This training budget is 42× larger than that of the first generation AfriBERTa and AfriTeVa models, but still orders of magnitude smaller than the models we compare against during evaluation [9, 14, 52, 64]. Over $524,288$ steps, our models are trained on ≈136B tokens with significant repetition.

4.3 Evaluation

4.3.1 Downstream Tasks

Text Classification

We use the news topic classification dataset, MasakhaNews, recently introduced by Adelani et al. (2023) [5] for 16 African languages. The authors establish multiple baselines on the dataset using both classical machine learning models and finetuning or prompting language models.

For T5 models, we cast the classification task as a text generation task in which the model outputs two tokens: the class and the end-of-sentence token. We are fortunate that the classes included in MasakhaNews exist as individual tokens in our tokenizer.

Dense Retrieval

We evaluate the effectiveness of our AfriBERTa V2 models as backbones for dense passage retrieval models [32]. We fine-tune our models on the English MS MARCO [40] passage ranking dataset [1] and evaluate their effectiveness in monolingual and cross-lingual retrieval scenarios using MIRACL [67], Mr. TyDi [65] and CIRAL [7].

We follow established practices of training with 128 batch size and learning rate $1e^{-5}$ for 40 epochs [66]. The maximum length of queries and passages is set to 64 and 256, respectively, and we train using Tevatron [20].

Machine Translation

We evaluated using MAFAND-MT [3] – a machine translation benchmark in the news domain. MAFAND-MT contains few thousand parallel training sentences ($2,500$-$30,000$ sentences) for 16 African languages, ideal for evaluating the effective adaptation of pre-trained LMs to new languages and domains. The authors find that adapting existing pre-trained language models to low-resource languages (through continual pre-training) before fine-tuning on high-quality translation data is effective for transferring to new domains and languages.

Following their work, we truncate inputs to 512 tokens and outputs to 84 tokens. We fine-tune AfriTeVa V2 models for 3 epochs on each language using a constant learning rate $5e^{-5}$ and batch size of 10. For inference, we use beam search with beam size 10.

Summarization

For summarization, we use XL-Sum [22], an abstractive summarization dataset which covers 44 languages, including 9 African languages. The authors establish strong baselines on both low and high-resource languages in the dataset through multilingual fine-tuning of mT5.

Following their work, we fine-tune our AfriTeVa V2 models for $50,000$ steps on the subset of languages in XL-Sum covered by Wura. We truncate inputs to 512 tokens and outputs to 84 tokens. Each training run uses an inverse-square root learning rate schedule with $5,000$ warmup steps. For inference, we use beam search with beam size 4 and length penalty $\alpha = 0.6$.

[1] We make use of the MS MARCO dataset provided by Tevatron: https://huggingface.co/datasets/Tevatron/msmarco-passage

Cross-lingual Question Answering

We evaluate on the test set of AfriQA [43], a cross-lingual question answering dataset with questions in 10 African languages and gold passages in English or French. We fine-tune models on the train set of the SQuAD 2.0 dataset [53] and evaluate in zero-shot generative cross-lingual QA settings using in-language queries and the provided gold passages in English.

Chapter 5

Results and Analysis

In this chapter, we evaluate the effectiveness of our new pre-trained language models and their predecessors against existing baselines on our evaluation tasks. We highlight improvements, trends and questions raised by our experimental results.

Hereafter, we refer to the first generation models as AfriBERTa and AfriTeVa models. When discussing our new pre-trained models, we refer to them explicitly as AfriBERTa and AfriTeVa V2 models. In discussions that apply to an entire model family, we use the terms AfriBERTa-style and AfriTeVa-style models.

5.1 Downstream Task Results

5.1.1 Text Classification

AfriBERTa V2

We evaluate the effectiveness of AfriBERTa V2 models against the base and large configurations of XLMR and AfroXLMR models. Across all languages in MasakhaNews, AfroXLMR Large with 550M parameters attains the best F1 scores as shown in Table 5.1.

First, AfriBERTa-style models outperform XLMR models by 4.0 F1 points on average. In addition, AfriBERTa V2 models significantly narrow the performance gap between the original AfriBERTa models and AfroXLMR models. AfriBERTa V2 Base attains 90.2 F1 points on average compared to AfriBERTa Base's 88.5 and AfroXLMR Base's 91.7 F1

Model	Size	amh	eng	fra	hau	ibo	lin	lug	orm	pcm	run	sna	som	swa	tir	xho	yor	AVG	AVG^{SL}
XLMR Base	270M	90.9	90.6	90.4	88.4	82.5	87.9	65.3	82.2	97.8	85.9	88.9	73.8	85.6	54.6	78.6	84.5	83.0	85.4
AfroXLMR Base	270M	94.2	92.2	**92.5**	91.0	90.7	93.0	89.4	**92.1**	98.2	91.4	**95.4**	85.2	88.2	86.5	94.7	93.0	91.7	92.2
XLMR Large	550M	93.1	92.2	91.4	90.6	84.2	91.8	73.9	88.4	98.4	87.0	88.9	76.1	85.6	62.7	89.2	84.5	86.1	87.4
AfroXLMR Large	550M	**94.4**	**93.1**	91.1	**92.2**	**93.4**	**93.7**	**89.9**	**92.1**	**98.8**	**92.7**	**95.4**	**86.9**	**87.7**	**89.5**	**97.3**	**94.0**	**92.6**	**93.0**
AfriBERTa Base	111M	91.3	87.7	80.5	90.2	87.8	87.8	85.2	90.4	97.3	91.2	90.8	82.4	85.8	85.6	90.5	91.6	88.5	89.4
AfriBERTa Large	126M	90.6	88.9	76.4	89.2	87.3	87.0	85.1	89.4	98.1	91.3	89.3	83.9	83.3	87.0	86.9	90.3	87.8	89.2
AfriBERTa V2 Base	173M	92.0	90.0	89.5	90.4	89.8	92.8	88.2	87.1	97.5	90.4	93.0	85.0	86.5	83.6	95.0	92.0	90.2	89.6
AfriBERTa V2 Large	187M	91.9	89.0	89.1	89.7	90.0	92.5	86.7	88.1	97.4	91.0	93.1	84.5	85.8	85.0	94.1	92.5	90.0	89.7

Table 5.1: MasakhaNews classification results for encoder models: Evaluation is done using the weighted F1 score and the scores presented are averaged across 5 seeds. While AfroXLMR Large boasts the best F1 scores across most languages, AfriBERTa V2 models improve over AfriBERTa models. The average scores excluding languages not in each model's pre-training dataset are also provided in AVG^{SL}. The best scores for each language are in **bold**.

points. It should be noted that AfriBERTa-style models achieve this despite having more than 2× less parameters than XLMR and AfroXLMR models.

According to Table 5.1, both AfriBERTa V2 models outperform their corresponding AfriBERTa counterparts by an average of 2.0 F1 points. For languages such as English, French, Shona and Xhosa, this improvement can be attributed to their inclusion in the pre-training of AfriBERTa V2 models. However, AfriBERTa V2 models also demonstrate significantly better performance for Lingala and Luganda, which were not included in their pre-training data. Specifically, AfriBERTa V2 Large achieves an F1 score of 92.5 F1 for Lingala, compared to 87.0 for AfriBERTa Large.

The base and large configurations of AfriBERTa-style models differ only in the number of layers, with 8 layers for the base models and 10 for the large models — a difference of approximately 15M parameters. Interestingly, their effectiveness is not only comparable on MasakhaNews, but the base models actually achieve a slightly higher average F1 than the large models across the 16 languages: with 88.5 vs 87.8 for AfriBERTa and 91.5 vs 91.0 for AfriBERTa V2.

AfriTeVa V2

We evaluate the effectiveness of AfriTeVa V2 models on MasakhaNews against the base configurations of mT5, FlanT5 and AfriMT5 which all have 580M parameters. Given that AfriTeVa V2 Large (1B) has 42% more parameters than these baseline models, we primarily compare them against AfriTeVa V2 Base (428M), which has approximately 26% fewer parameters.

Model	Size	amh	eng	fra	hau	ibo	lin	lug	orm	pcm	run	sna	som	swa	tir	xho	yor	AVG	AVG^{SL}
AfriTeVa-base	229M	87.0	80.3	71.9	85.8	79.9	82.8	60.2	82.9	95.2	80.0	84.4	58.0	80.7	55.2	69.4	86.4	77.5	78.4
mT5-base	580M	78.2	89.8	59.0	82.7	76.8	80.8	75.0	79.2	96.1	85.7	90.4	75.0	76.1	65.1	71.8	86.2	79.2	78.6
FlanT5-base	580M	54.5	**92.4**	88.9	84.5	86.6	**90.6**	84.1	85.8	97.8	87.3	90.6	76.0	79.0	41.5	90.8	88.9	82.5	83.2
AfriMT5-base	580M	90.2	90.3	87.4	87.9	88.0	88.6	84.8	83.9	96.6	91.0	91.5	77.8	84.4	80.8	91.6	88.8	87.7	87.8
AfriTeVa V2 Base	428M	**93.6**	91.3	**90.2**	**92.3**	91.7	88.7	**90.5**	87.5	98.0	92.5	93.1	84.2	87.5	**86.5**	94.8	**94.2**	91.0	90.5
AfriTeVa V2 Large	1B	92.9	91.4	**90.2**	91.7	**92.9**	90.2	89.7	**89.0**	**98.4**	**92.7**	**94.1**	85.6	**88.1**	**88.3**	**95.4**	94.1	**91.5**	**91.1**

Table 5.2: MasakhaNews classification results for encoder-decoder models: Evaluation is done using the weighted F1 score and the scores presented are averaged across 5 seeds. AfriTeVa V2 Base surpasses mT5-base and FlanT5-base by up to 10 points despite having $\approx 26\%$ less parameters. The average scores excluding languages not in each model's pre-training dataset are also provided in AVG^{SL}. The best scores for each language are in **bold**.

The results in Table 5.2 highlight the strong performance of the AfriTeVa V2 models on the MasakhaNews dataset. AfriTeVa V2 Base outperforms all the models we compare to on 14 out of the 16 languages. Specifically, AfriTeVa V2 Base achieves an average F1 score of 91.0, surpassing the 79.2, 82.5, and 87.7 average F1 scores of mT5-base, FlanT5-base, and AfriMT5-base, respectively. AfriTeVa V2 Large, with 42% more parameters than the baseline models, further improves performance, achieving the highest average F1 score of 91.5.

Moreover, AfriTeVa V2 Base substantially improves over the original AfriTeVa-base model, which has an average F1 score of 78.4. On Lingala and Luganda, which were not included in pre-training data for AfriTeVa-style models, AfriTeVa V2 Base outperforms the original AfriTeVa Base model by 6 and 30 F1 points respectively.

5.1.2 Dense Retrieval

We compare the effectiveness of our AfriBERTa V2 models as backbones for dense retrieval models to mBERT (178M) and AfroXLMR Base (270M). It should be noted that we evaluate cross-lingual transfer here since all models are fine-tuned on English MS MARCO and evaluated on test collections covering African languages.

For monolingual retrieval on Mr. TyDi and MIRACL, AfriBERTa-style models consistently outperform mBERT and AfroXLMR Base. Comparing AfriBERTa V2 models (rows 5 & 6 in Table 5.3) to AfroXLMR Base (row 4), AfriBERTa V2 Base achieves 0.511 MRR@100 on Mr. TyDi compared to 0.333 and 0.479 for mBERT and AfroXLMR Base, respectively. Similarly, it secures 0.499 nDCG@10 on MIRACL for Swahili compared to 0.356 and 0.299 for mBERT and AfroXLMR Base. While AfroXLMR significantly out-

	Mr. TyDi		MIRACL				CIRAL							
	swa		swa	yor	swa	yor	hau	som	swa	yor	hau	som	swa	yor
	MRR@100	**Recall@100**	**nDCG@10**		**Recall@100**		**nDCG@20**				**Recall@100**			
	MS MARCO pFT													
(1) mBERT (178M)	0.333	0.637	0.299	0.441	0.616	0.832	0.052	0.070	0.148	0.156	0.082	0.114	0.210	0.327
(2) AfriBERTa Base (112M)	0.472	0.807	0.452	0.448	0.775	0.740	0.154	0.115	0.113	0.102	0.219	0.171	0.167	0.263
(3) AfriBERTa Large (126M)	0.477	0.800	0.435	0.440	0.759	0.733	0.185	0.128	0.121	0.16	0.264	0.185	0.178	0.341
(4) AfroXLMR Base (270M)	0.479	0.799	0.356	**0.502**	0.622	**0.840**	**0.243**	**0.220**	0.255	0.213	**0.347**	**0.300**	**0.335**	0.403
(5) AfriBERTa V2 Base (173M)	0.511	0.848	**0.499**	0.451	**0.825**	0.797	0.226	0.173	0.203	0.231	0.317	0.291	0.288	0.410
(6) AfriBERTa V2 Large (187M)	**0.518**	**0.859**	0.498	0.443	0.807	0.687	0.219	0.193	**0.259**	**0.245**	**0.347**	0.292	0.326	**0.437**

Table 5.3: Retrieval results: AfriBERTa V2 models outperform mBERT and AfroXLMR Base for monolingual retrieval on Mr. TyDi and MIRACL. For cross-lingual retrieval on CIRAL, AfriBERTa V2 models achieves competitive effectiveness with AfroXLMR Base as a backbone for dense retrieval. Across the test collections, the best scores for each language are in **bold**.

performs all models on Yoruba MIRACL, all AfriBERTa-style models continue to yield stronger performance compared to mBERT.

The improved effectiveness of AfriBERTa V2 models compared to their corresponding AfriBERTa models is most apparent in the context of cross-lingual retrieval on CIRAL. CIRAL entails retrieving relevant documents in African languages when issues English queries. For the original AfriBERTa models, which were not pre-trained on English, we see a performance drop-off in this cross-lingual retrieval scenario. In contrast, Table 5.3 shows that AfriBERTa V2 models outperform mBERT and remain competitive with AfroXLMR Base across all languages covered by CIRAL.

5.1.3 Machine Translation

We present the translation results for English/French to African languages (*en/fr-xx*) in Table 5.4, and for African languages to English/French (*xx-en/fr*) in Table 5.5. For each translation direction, we provide both the average and median BLEU and CHRF scores. Significant variation between these statistics indicates that the model in question performs substantially better for some languages than for others. This discrepancy is particularly noteable for English-pivot languages in Tables 5.4 and 5.5, as 4 of the 10 languages (Luganda, Luo, Setswana & Twi) were not included in the pre-training data of all the models we evaluate. Our primary comparison focuses on AfriTeVa V2 Base (428M) against mT5-base and ByT5-base, as well as their adapted counterparts, AfriMT5-base and AfriByT5-base.

Although all the models we evaluate were pre-trained on French, none of the six (6) languages with a French pivot in MAFAND-MT were included in their pre-training data.

Model	*en-xx*												*fr-xx*							
	hau	ibo	pcm	swa	yor	zul	lug	luo	tsn	twi	**AVG**	**MED**	bam	bbj	ewe	fon	mos	wol	**AVG**	**MED**
	BLEU																			
mT5-base	2.8	18.0	34.1	25.1	4.8	11.7	3.0	3.1	3.4	1.7	10.7	4.1	1.5	0.4	2.2	1.6	0.1	0.9	1.1	1.2
AfriMT5-base	5.1	19.6	**35.0**	26.7	6.2	13.2	5.2	4.6	7.0	2.7	12.5	6.6	2.1	0.8	3.7	2.5	0.1	1.8	1.8	2.0
ByT5-base	8.3	21.8	30.1	24.4	7.5	14.0	12.1	8.4	14.7	6.0	14.7	13.1	9.5	1.8	**5.5**	**3.8**	0.1	6.0	4.5	4.7
AfriByT5-base	9.3	**22.7**	30.0	24.7	7.6	15.3	**13.1**	**8.9**	17.0	**6.1**	**16.2**	**15.3**	**11.4**	**2.2**	5.2	3.7	0.2	**6.4**	**4.9**	**4.5**
AfriTeVa V2 Base	11.2	17.8	33.2	29.2	8.0	13.5	4.9	3.3	7.4	1.8	13.0	9.6	0.9	0.3	0.0	0.8	0.3	2.3	0.8	0.6
AfriTeVa V2 Large	**12.9**	20.7	34.5	**30.8**	**10.9**	**15.0**	7.0	5.6	**17.8**	2.9	15.8	14.0	4.4	0.6	0.9	1.6	**0.5**	2.0	1.7	1.3
	CHRF																			
mT5-base	23.6	41.1	64.1	53.7	20.8	36.0	24.9	21.6	22.8	17.8	32.6	24.3	10.0	7.4	9.7	11.5	7.9	9.1	9.3	9.4
AfriMT5-base	29.7	43.1	**64.7**	55.1	24.3	40.3	30.4	25.7	31.5	21.5	36.6	31.0	14.0	12.7	16.6	14.8	8.2	13.8	13.4	13.9
ByT5-base	31.3	46.5	58.1	52.5	25.5	40.3	40.0	32.2	38.6	27.9	39.3	39.3	27.8	17.7	23.8	16.1	8.8	22.9	19.5	20.3
AfriByT5-base	32.8	47.4	58.0	52.8	26.0	42.9	**42.2**	**33.6**	42.1	**29.0**	40.7	42.2	**31.4**	**19.9**	**24.1**	**16.5**	**9.8**	**23.8**	**20.9**	**21.9**
AfriTeVa V2 Base	37.6	46.8	63.6	57.2	26.8	46.3	24.5	19.0	31.1	15.6	36.9	34.4	9.2	8.4	3.7	11.0	8.9	11.8	8.8	9.1
AfriTeVa V2 Large	**41.7**	**50.2**	64.6	**58.9**	**32.8**	**50.6**	34.6	25.8	**45.7**	15.9	**42.1**	**43.7**	17.4	12.0	11.3	14. 6	**9.8**	12.7	12.6	12.0

Table 5.4: MAFAND-MT *en/fr-xx* results: Evaluation is done using the BLEU and CHRF scores. AfriBERTa V2 models obtain significantly higher scores for languages included in their pre-training data. On other languages except `mos`, AfriByT5 is consistently the most effective model.

Consequently, translations involving French as a pivot require the models to generalize to a previously unseen language. Unsurprisingly, all models perform better in *xx-fr* translations than in *fr-xx* due to greater exposure to French than to these low-resource languages. AfriTeVa V2 Base averages 0.8 BLEU for *fr-xx* translations and 1.8 BLEU for *xx-fr* translations. In comparison, mT5-base averages 1.1 BLEU for *fr-xx* translations and 1.5 BLEU for *xx-fr* translations. Thus, while AfriTeVa V2 Base underperforms mT5-base in *fr-xx* translations, it outperforms mT5-base in *xx-fr* translations. AfriMT5-base shows great improvement over both mT5-base and AfriTeVa V2 Base, especially in the *xx-fr* translation direction. However, the token-free models (ByT5 & AfriByT5) exhibit the strongest capability in translating languages that were not seen during pre-training. As shown in Table 5.5, AfriByT5 achieves an average BLEU score of 6.6 for *xx-fr* translations. This performance is more than double that of mT5-base (1.5), AfriMT5-base (3.1), AfriTeVa V2 Base (1.8), and even the larger AfriTeVa V2 Large (4.1). We observe a similar trend in *fr-xx* translation results as presented in Table 5.4.

When translating to and from English, AfriTeVa V2 models are more competitive. AfriTeVa V2 Base outperforms all other models in 4 out of 6 of the languages included in its pre-training data, especially in the *xx-en* translation direction. Specifically, AfriTeVa V2 Base achieves 33.7 BLEU when translating from Swahili to English and 29.2 BLEU when translating into Swahili. For Luganda, Luo, Setswana and Twi—languages not included in its pre-training data—AfriTeVa V2 Base performs better than it does with the French-pivot languages. Remarkably, it even surpasses mT5-base and AfriMT5-base in both translation

Model	*xx-en* hau	ibo	pcm	swa	yor	zul	lug	luo	tsn	twi	AVG	MED	*xx-fr* bam	bbj	ewe	fon	mos	wol	AVG	MED
	BLEU																			
mT5-base	5.8	18.9	42.2	29.5	12.3	22.4	12.6	6.4	9.5	4.6	16.4	12.5	2.5	0.9	1.1	2.4	0.7	1.3	1.5	1.2
AfriMT5-base	10.4	19.5	44.6	30.6	13.8	24.0	15.5	9.7	16.1	8.4	19.3	15.8	6.4	2.0	2.1	4.2	1.2	2.9	3.1	2.5
ByT5-base	12.9	21.0	39.4	27.1	11.5	22.8	19.8	12.1	9.8	11.5	18.8	16.4	10.0	2.7	4.1	4.9	1.5	7.2	5.1	4.5
AfriByT5-base	13.5	20.7	39.5	27.0	11.9	24.0	**21.1**	**12.5**	19.7	**10.5**	20.0	20.2	**13.8**	**4.4**	**4.5**	**5.8**	2.2	**9.0**	**6.6**	**5.2**
AfriTeVa V2 Base	11.9	20.1	43.8	33.7	16.1	25.9	13.8	5.2	15.5	4.6	19.1	15.8	2.7	1.3	1.7	2.5	1.2	1.2	1.8	1.5
AfriTeVa V2 Large	**18.4**	**23.7**	**47.0**	**36.1**	**20.9**	**33.6**	19.4	10.1	**25.1**	7.9	**24.2**	**22.3**	10.0	2.0	2.6	4.3	**2.5**	3.0	4.1	2.8
	CHRF																			
mT5-base	26.3	43.5	66.9	53.7	31.1	43.9	36.3	26.1	32.2	25.2	38.5	34.3	19.4	15.1	17.0	17.9	10.9	16.2	16.1	16.6
AfriMT5-base	32.5	44.9	68.4	54.5	33.9	45.9	40.2	32.2	39.6	31.2	42.3	39.9	27.7	19.6	21.1	21.4	13.2	21.6	20.8	21.3
ByT5-base	33.2	46.4	62.0	50.6	31.4	42.5	45.4	34.1	42.4	32.9	42.1	42.5	31.2	21.8	24.8	20.5	15.4	26.2	23.3	23.3
AfriByT5-base	33.9	46.4	62.1	50.5	32.0	43.7	**47.1**	**35.0**	43.4	**33.4**	42.8	43.6	**34.8**	**25.5**	**24.9**	**22.0**	16.2	**29.3**	**24.5**	**25.2**
AfriTeVa V2 Base	36.4	46.3	67.5	57.4	37.0	48.8	37.6	25.1	39.8	23.9	42.0	38.7	20.4	17.8	20.2	17.3	14.5	16.2	17.7	17.8
AfriTeVa V2 Large	**40.7**	**49.3**	**69.8**	**59.3**	**42.3**	**54.7**	45.5	33.9	**49.3**	29.8	**47.5**	**47.4**	32.3	19.3	23.7	21.6	**17.5**	21.9	22.7	21.8

Table 5.5: MAFAND-MT *xx-en/fr* results: Evaluation is done using the BLEU and CHRF scores. All models perform better in this *xx-en/fr* translation direction than in *en/fr-xx*. AfriTeVa V2 models consistently outperform other models for languages included in its pre-training data. For unseen languages, AfriTeVa V2 models perform better for languages with English pivot than for French-pivot languages. The best scores for each language are in **bold**.

directions for Setswana. As we can see in Table 5.4, AfriTeVa V2 Base achieves 7.4 BLEU for Setswana and AfriTeVa V2 Large, 17.8 — an improvement over AfriByT5's 17.0 BLEU. This strong performance is likely due to the close relation of Setswana to Sesotho, as both languages belong to the Sotho language group.

Overall, AfriTeVa V2 Large's strong improvements over AfriTeVa Base suggest the importance of scale. Table 5.5 shows that the 1B parameter model obtains a median BLEU of 22.3 for translations into English, an improvement over AfriTeVa V2 Base (15.8) and AfriByT5 (20.2). Its strongest improvements over the baseline models come for Hausa, Nigerian Pidgin, Setswana, Swahili and Zulu. When translating from English into African languages, it achieves a median BLEU of 14.0 compared to AfriByT5's 15.3. Its effectiveness for this translation direction is hampered by languages not included in its pre-training data, achieving on average 3 BLEU points less than AfriByT5 for these languages.

5.1.4 Summarization

Table 5.6 presents the performance metrics (Rouge-1/Rouge-2/Rouge-L) of AfriTeVa V2 Base and Large, and baseline models mT5-base and AfriMT5-base, across 8 of the African languages included in our work. AfriTeVa V2 models generally outperform mT5 and AfriMT5, with AfriTeVa V2 Large consistently achieving the highest scores across all

	mT5	AfriMT5	AfriTeVa V2 Base	AfriTeVa V2 Large
hau	39.4/17.7/31.7	39.4/17.3/31.6	41.6/19.9/33.9	**43.2/20.7/35.0**
ibo	31.6/10.2/24.5	33.9/12.9/25.7	35.4/13.9/26.9	**37.3/15.0/28.2**
orm	18.7/6.2/16.2	18.7/6.4/16.3	20.0/7.0/17.5	**22.1/7.9/19.0**
pcm	38.0/15.1/29.9	38.7/15.3/30.1	39.6/16.6/31.4	**41.1/17.2/32.0**
run	32.0/14.4/25.8	31.5/14.4/25.6	34.3/17.0/28.5	**36.2/18.1/29.7**
som	31.6/11.6/24.2	31.2/11.2/23.9	33.2/12.7/25.7	**34.1/13.3/26.4**
swa	37.7/17.9/30.9	37.5/17.7/30.6	39.7/19.4/32.8	**40.5/20.0/33.4**
yor	31.7/11.7/25.1	38.4/16.2/29.0	39.3/17.6/30.3	**40.7/18.5/31.4**
AVG	32.7/12.9/26.1	33.7/13.9/26.6	35.4/15.5/28.4	**36.9/16.3/29.4**

Table 5.6: XL-SUM results: Results reported are Rouge-1/Rouge-2/Rouge-L. While AfriMT5 improvements over mT5 are modest, AfriTeVa V2 Base achieves more than 2.0 points improvements on average over both models across all rouge metrics. The best scores for each language are in **bold**.

languages. This continued trend suggests that the enhancements and larger capacity of AfriTeVa V2 Large contribute significantly to better handling of these languages.

While AfriMT5-base only marginally improves over the mT5-base rouge scores for most languages, AfriTeVa V2 Base is consistently better than both models for every language by 1.5 Rouge-L points on average. As we can see in Table 5.6, the largest improvements over mT5-base come for Yorùbá with AfriMT5-base achieving 38.4 Rouge-1 compared to mT5-base's 31.7 and its 29.0 Rouge-L compared to mT5-base's 25.1. AfriTeVa V2 Base further improves over AfriMT5-base with 39.3 Rouge-1 and 30.3 Rouge-L scores for Yorùbá.

Across all models, rouge scores for Afaan Oromo are noticably lower than for other languages. This warrants investigation and could indicate that the language presents unique challenges, possibly due to its linguistic features.

5.1.5 Cross-lingual Question Answering

AfriTeVa V2 Base delivers impressive results in the cross-lingual question-answering task, especially for languages in our pre-training data. This task evaluates cross-lingual transfer by fine-tuning on English data and testing on the AfriQA test set. We report F1 scores and Exact Math (EM) accuracy of generative gold passage answer prediction.

As shown in Table 5.7, AfriTeVa V2 Base achieves significantly higher F1 scores and Exact Match accuracies (≈2×) across 6 out of 7 languages compared to both mT5-base

Metric	Model	bem	hau	ibo	kin	twi	yor	zul	AVG
F1	mT5-base	2.9	25.8	41.7	25.5	**5.3**	11.9	24.7	17.6
	AfriTeVa Base	3.5	4.6	5.5	4.8	5.4	6.1	4.4	4.9
	AfriMT5-base	**6.4**	39.7	40.7	30.3	5.3	21.8	31.9	25.2
	AfriTeVa V2 Base	5.7	**45.4**	**57.1**	**45.4**	2.1	**37.6**	**45.9**	**34.2**
EM	mT5-base	1.1	22.3	34.7	20.2	**3.5**	7.8	20.9	13.9
	AfriTeVa Base	2.0	2.7	4.2	3.2	3.1	3.9	3.1	3.2
	AfriMT5-base	4.2	33.0	33.0	23.1	2.9	15.7	25.5	19.6
	AfriTeVa V2 Base	**5.2**	**36.7**	**47.7**	**33.7**	1.4	**29.5**	**37.8**	**27.4**

Table 5.7: Cross-lingual Question Answering Results: F1 and Exact Match (EM) Accuracy scores on the test set of AfriQA. For both metrics, AfriTeVa V2 outperforms mT5 for all languages except for `twi`. The best scores for each language are in **bold**.

and AfriMT5-base. The exceptions are Bemba, for which AfriTeVa V2 Base remains competitive with AfriMT5-base, and Twi, where it underperforms compared to all other models, including AfriTeVa Base.

Finally, AfriTeVa V2 Base exhibits superior cross-lingual transfer ability with an average F1 score of 34.2 across all languages, compared to 4.9 for the original AfriTeVa Base model.

5.2 Discussion

In this section, we discuss implications of the results observed in the previous section and highlight the contributions of this work. Our findings in section 5.1 show that increasing data and model size significantly enhances the effectiveness of language models for African languages. This advancement benefits both the speakers of these languages (see Table 2.1) and the NLP community striving to advance low-resource languages.

5.2.1 Data Quality Versus Scale

In the development of language models today, significant work goes into deliberate curation of high-quality, diverse pre-training data [28, 47]. Previous works have shown the correlation between the quality of pre-training data and the performance of the trained

model [33, 51]. The improvement of AfriTeVa V2 Base over baselines in multiple downstream tasks suggests that this is true. Our experiments show that AfriTeVa V2 Base outperforms larger models like AfriMT5-base & AfriByT5-base [9] which were trained on unfiltered mC4 corpus. However, our pre-training dataset, WURA, contains ≈1.5× more data than mC4 contains across 16 African languages. Therefore, additional experiments are necessary to distinguish the effects of data scale from data quality, extending to models of a similar size to AfriTeVa V2 Large (1B) to properly contextualize its strong performance.

In their extensive study of performance degradation in language models due to repeated data, Hernandez et al. (2022) [25] found that 0.1% of pre-training data repeated 100 times can degrade the performance of a model to that of a model half its size. This degradation occurs due to a shift from generalization to memorization. However, in the data-constrained context of African languages, such data repetition is both common and necessary for developing performant language models. In our work, we observed continued improvement from training our language models on multiple epochs (≈15) of our pre-training dataset. Nevertheless, it is crucial to study these mechanisms as they manifest for African languages to advance research in this area.

In line with this, we make available intermediate checkpoints of our AfriBERTa V2 and AfriTeVa V2 models, taken every 50,000 steps during pre-training. [1]

5.2.2 The Importance of High-Resource Languages

With AfriBERTa, Ogueji et al. (2021) [42] challenged the previously accepted belief that lower-resource languages need higher-resource languages in multilingual language models trained to serve them. Their work demonstrated the viability of pre-training exclusively on related low-resource languages, ushering in a wave of similarly "small but competitive" language models. Revisiting this discussion is pertinent for two reasons:

- The range of downstream tasks for African language is expanding, and multilingual models pre-trained on high-resource languages can leverage cross-lingual transfer to perform well without task-specific data.
- Specialized models for African languages must reflect the continent's multilingual settings to be broadly useful.

[1] https://github.com/castorini/AfriTeVa-keji?tab=readme-ov-file#datasets

In our work pre-training the original AfriTeVa models [42], the inclusion of English during pre-training improved BLEU scores by up to 3.0 points during translation from African languages into English. While the benefit is immediately obvious for translation, our experiment results for information retrieval (subsection 5.1.2) and cross-lingual question answering (subsection 5.1.5) highlight the importance of pre-training on "carefully selected" high-resource languages on the utility and downstream effectiveness of these models.

The second reason is necessitated by the bilingual nature of many African populations. High-resource languages such as English and French often serve as the language of education and administration while the local language is used for broader communication. This bilingual context means that effective language models must handle both local and high-resource languages to be truly useful. By incorporating such high-resource languages during pre-training, we can better capture the linguistic realities of these populations, resulting in models that are more practical and effective in real-world applications.

5.2.3 Generalizing to Unseen Languages

When discussing generalization to languages not seen during pre-training, we hypothesized that the models evaluated in subsection 5.1.3 performed well on Setswana due to its close relation to Sesotho, a Bantu language included in their pre-training data. It is also possible that this particular generalization indicates language contamination in the mC4 dataset. Kreutzer et al. (2022) [33] attributed this benign form of language contamination in their quality audit of mC4 to the inability of language identification models to differentiate some low-resource languages from their closely-related, higher-resource counterparts, with Bantu languages being of particular concern. We believe that advances in language identification for African languages are necessary for the proper attribution of generalization mechanisms in models pre-trained for the continent.

The strong generalization of ByT5 and AfriByT5—the two token-free language models we evaluated in this work—to languages not included in their pre-training data suggests their suitability and practicality for the low-resource context of African languages. Our evaluation results on MAFAND-MT (subsection 5.1.3) show how well these models are able to learn from a few thousand examples. This robustness is likely due to their byte tokenizer which is better equipped to handle noisy text and *out-of-vocabulary* words [63].

It should be noted that both token-free models were pre-trained on mC4 so language contamination could theoretically contribute to the strong generalization to new languages

that we observe. However, mT5 and AfriMT5, which were also pre-trained on mC4, do not replicate this generalization ability outside of Setswana.

Typically, byte-level encoding leads to larger token sequences and, in turn, greater training and inference cost of byte-level models [63]. Recently, Limisiewicz et al. (2024) [36] proposed morphology-driven byte encoding to alleviate these issues. They achieve equitable segmentation of words across 99 languages, competitive performance with ByT5 across evaluation tasks and improved training efficiency. Continued research in this area will benefit low-resource languages.

5.2.4 Opportunities for Decoder-Only Models

Our research primarily concentrated on encoder and encoder-decoder models. However, AfriTeVa V2 Large's impressive performance in generative tasks such as summarization, translation, and question answering indicates potential for using decoder-only models for African languages. This opens up promising opportunities to enhance natural language processing applications, making them more accessible and effective for diverse linguistic communities across the African continent.

5.2.5 Towards Agency in African AI Development

The question of agency in advancing artificial intelligence research and applications in Africa has gained prominence as language models became more multilingual, capable and commercial. The transformative impact of these models is evident in the widespread adoption of systems like Open AI's ChatGPT and the integration of AI assistants into various software.

However, the current landscape raises concerns about the inclusion of African languages and contexts in these models. There remains a significant gap in the development and deployment of AI technologies that cater sufficiently or specifically to the African context. This gap underscores the need for African researchers, developers and policymakers to take an active role in shaping the future of AI on the continent.

Our work directly contributes to this agency by developing language models tailored to African languages and contexts. We aim to bridge the gap in representation and provide tools that are more accurate and broadly useful for the African population. This aim informed decisions such as the inclusion of Arabic, English, French and Portuguese in our pre-training data. These four high-resource languages are popularly spoken and used

on the continent, serving as lingua francas that facilitate communication across diverse linguistic groups and eventually show up in code-mixing and lexical borrowing scenarios. By incorporating these languages, we ensure that our models are not only capable of understanding and generating text in widely-used languages but also better positioned to support applications in practical African multilingual settings.

Chapter 6

Conclusion and Future Work

In this thesis, we scale the pre-training data available for African languages and pre-train improved language models to serve the populations that use these languages. We introduce WURA, a document-level pre-training dataset covering 16 African languages and four high-resource languages popularly spoken on the continent: Arabic, English, French and Portuguese. We pre-train new versions of the AfriBERTa (encoder-only) [42] and AfriTeVa (encoder-decoder) [30] model families on WURA and demonstrate their improved effectiveness on a diverse range of natural language understanding and generation tasks. To our knowledge, our AfriTeVa V2 Large (1B) is the largest sequence-to-sequence model pre-trained for African languages.

In Chapter 3, we provide details of the three-stage curation process of WURA: auditing and filtering existing web crawls, initiating web crawls of our own, and combining with existing language resources. Our process is informed by prior work in this area [33] and we mention trade-offs that we make to ensure quality of our dataset while keeping resource requirements manageable. Our aim is for WURA to be actively maintained and extended to support more African languages thus we highlight measures we put in place to ensure this and to garner community contribution. We detail our experimental setup in Chapter 4, including all details necessary for its reproduction, and discuss our results in Chapter 5. We compare the effectiveness of our models against existing baselines, many of which are larger models, and study the generalization of our models to languages not deliberately included in our pre-training data.

We evaluate our trained models on a broad range of NLP tasks — text classification, information retrieval, translation, summarization, and cross-lingual question answering. Our evaluation results in Chapter 5 indicate that our models significantly outperform their

predecessors across all tasks, highlighting the value of extensive, high-quality pre-training data. Our AfriBERTa V2 models surpass existing encoder-only models (mBERT [16], XLMR [14]) and are competitive with AfroXLMR, despite having over 50% fewer parameters. Additionally, our AfriTeVa V2 Base consistently outperforms existing encoder-decoder baselines (ByT5-base [63], mT5-base [64]) across all tasks, and AfriTeVa V2 Large shows potential for even greater performance gains at scale. Notably, we achieve these results despite pre-training on a significantly smaller token budget due to the low-resource nature of African languages.

In future work, we aim to use synthetic data to expand pre-training datasets for African languages and develop recall-focused pipelines for filtering large web crawls while maintaining precision like Wura. Given the strong performance of AfriTeVa V2 Large (1B) on generative tasks, we plan to explore instruction fine-tuning and the model's open-ended generation capabilities. Additionally, we intend to compare our models' performance to the 13B parameter mT5 model fine-tuned on an instruction mixture covering 101 languages, as done by Üstün et al. (2024) [69].

References

[1] Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *arXiv preprint arXiv:2201.06642*, 2022.

[2] Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. SERENGETI: Massively Multilingual Language Models for Africa. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1498–1537, Toronto, Canada, July 2023. Association for Computational Linguistics.

[3] 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. 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, July 2022. Association for Computational Linguistics.

[4] David Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Alabi, Shamsuddeen Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya,

Bonaventure F. P. Dossou, Blessing Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris Chinenye Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Mboning Tchiaze Elvis, Tajuddeen Gwadabe, Tosin Adewumi, Orevaoghene Ahia, and Joyce Nakatumba-Nabende. MasakhaNER 2.0: Africa-centric transfer learning for named entity recognition. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4488–4508, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[5] David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris Chinenye Emezue, Sana Al-azzawi, Blessing Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Ajayi, Tatiana Moteu, Brian Odhiambo, Abraham Owodunni, Nnaemeka Obiefuna, Muhidin Mohamed, Shamsuddeen Hassan Muhammad, Teshome Mulugeta Ababu, Saheed Abdullahi Salahudeen, Mesay Gemeda Yigezu, Tajuddeen Gwadabe, Idris Abdulmumin, Mahlet Taye, Oluwabusayo Awoyomi, Iyanuoluwa Shode, Tolulope Adelani, Habiba Abdulganiyu, Abdul-Hakeem Omotayo, Adetola Adeeko, Abeeb Afolabi, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Ogbu, Chinedu Mbonu, Chiamaka Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola Awosan, Tadesse Kebede, Toadoum Sari Sakayo, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Kanda Tshinu, Ussen Kimanuka, Thina Diko, Siyanda Nxakama, Sinodos Nigusse, Abdulmejid Johar, Shafie Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, Jules Jules, Ivan Ssenkungu, and Pontus Stenetorp. MasakhaNEWS: News topic classification for African languages. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi, editors, *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 144–159, Nusa Dua, Bali, November 2023. Association for Computational Linguistics.

[6] David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Zhuang Yun Jian, Jesujoba Oluwadara Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing K. Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Bridget Odu, Rooweither Mabuya, Shamsud-

deen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, and Pontus Stenetorp. IrokoBench: A New Benchmark for African Languages in the Age of Large Language Models. 2024.

[7] Mofetoluwa Adeyemi, Akintunde Oladipo, Xinyu Zhang, David Alfonso-Hermelo, Mehdi Rezagholizadeh, Boxing Chen, and Jimmy Lin. CIRAL at FIRE 2023: Cross-Lingual Information Retrieval for African Languages. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, page 4–6, New York, NY, USA, 2024. Association for Computing Machinery.

[8] Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David Mortensen, Noah Smith, and Yulia Tsvetkov. Do All Languages Cost the Same? Tokenization in the Era of Commercial Language Models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923, Singapore, December 2023. Association for Computational Linguistics.

[9] Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 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, October 2022. International Committee on Computational Linguistics.

[10] Israel Abebe Azime, Mitiku Yohannes Fuge, Atnafu Lambebo Tonja, Tadesse Destaw Belay, Aman Kassahun Wassie, Eyasu Shiferaw Jada, Yonas Chanie, Walelign Tewabe Sewunetie, and Seid Muhie Yimam. Enhancing amharic-llama: Integrating task specific and generative datasets. *arXiv preprint arXiv:2402.08015*, 2024.

[11] Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The Belebele Benchmark: A Parallel Reading Comprehension Dataset in 122 Language Variants. *arXiv preprint arXiv:2308.16884*, 2023.

[12] Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. Language ID in the Wild: Unexpected Challenges on the Path to a Thousand-Language Web Text Corpus. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics.

[13] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3576–3588, Online, June 2021. Association for Computational Linguistics.

[14] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics.

[15] Alexis Conneau and Guillaume Lample. *Cross-lingual Language Model Pretraining*. Curran Associates Inc., Red Hook, NY, USA, 2019.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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, June 2019. Association for Computational Linguistics.

[17] Cheikh M. Bamba Dione, David Ifeoluwa Adelani, Peter Nabende, Jesujoba Alabi, Thapelo Sindane, Happy Buzaaba, Shamsuddeen Hassan Muhammad, Chris Chinenye Emezue, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, Jonathan Mukiibi, Blessing Sibanda, Bonaventure F. P. Dossou, Andiswa Bukula, Rooweither Mabuya, Allahsera Auguste Tapo, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Fatoumata Ouoba Kabore, Amelia Taylor, Godson Kalipe, Tebogo Macucwa, Vukosi Marivate, Tajuddeen Gwadabe, Mboning Tchiaze Elvis, Ikechukwu Onyenwe, Gratien Atindogbe, Tolulope Adelani, Idris Akinade, Olanrewaju Samuel, Marien Nahimana, Théogène Musabeyezu, Emile Niyomutabazi, Ester Chimhenga, Kudzai Gotosa, Patrick Mizha, Apelete Agbolo, Seydou Traore, Chinedu Uchechukwu, Aliyu Yusuf, Muhammad Abdullahi, and Dietrich Klakow. Masakha-POS: Part-of-Speech Tagging for Typologically Diverse African languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

Papers), pages 10883–10900, Toronto, Canada, July 2023. Association for Computational Linguistics.

[18] Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. AfroLM: A Self-Active Learning-based Multilingual Pretrained Language Model for 23 African Languages. In Angela Fan, Iryna Gurevych, Yufang Hou, Zornitsa Kozareva, Sasha Luccioni, Nafise Sadat Moosavi, Sujith Ravi, Gyuwan Kim, Roy Schwartz, and Andreas Rücklé, editors, *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[19] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint, abs/2101.00027*, 2020.

[20] Luyu Gao, Xueguang Ma, Jimmy J. Lin, and Jamie Callan. Tevatron: An efficient and flexible toolkit for dense retrieval. *ArXiv*, abs/2203.05765, 2022.

[21] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio Cesar, Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks Are All You Need, June 2023.

[22] Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 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, August 2021. Association for Computational Linguistics.

[23] Jacaranda Health. Ulizallama. https://huggingface.co/Jacaranda/UlizaLlama, 2023. Accessed: 2023-10-01.

[24] Jonathan Heek, Anselm Levskaya, Avital Oliver, Marvin Ritter, Bertrand Rondepierre, Andreas Steiner, and Marc van Zee. Flax: A neural network library and ecosystem for JAX, 2023.

[25] Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al.

Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

[26] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2024. Curran Associates Inc.

[27] Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR, 13–18 Jul 2020.

[28] Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the Potential of Small Language Models With Scalable Training Strategies. *arXiv preprint arXiv:2404.06395*, 2024.

[29] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics.

[30] Odunayo Jude Ogundepo, Akintunde Oladipo, Mofetoluwa Adeyemi, Kelechi Ogueji, and Jimmy Lin. AfriTeVA: Extending ?small data? pretraining approaches to sequence-to-sequence models. In Colin Cherry, Angela Fan, George Foster, Gholamreza (Reza) Haffari, Shahram Khadivi, Nanyun (Violet) Peng, Xiang Ren, Ehsan Shareghi, and Swabha Swayamdipta, editors, *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 126–135, Hybrid, July 2022. Association for Computational Linguistics.

[31] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models. *arXiv preprint, abs/2001.08361*, 2020.

[32] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense Passage Retrieval for Open-Domain Question Answering. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, November 2020. Association for Computational Linguistics.

[33] Kreutzer, Julia and Caswell, Isaac and Wang, Lisa and Wahab, Ahsan and van Esch, Daan and Ulzii-Orshikh, Nasanbayar and Tapo, Allahsera and Subramani, Nishant and Sokolov, Artem and Sikasote, Claytone and Setyawan, Monang and Sarin, Supheakmungkol and Samb, Sokhar and Sagot, Benoît and Rivera, Clara and Rios, Annette and Papadimitriou, Isabel and Osei, Salomey and Suarez, Pedro Ortiz and Orife, Iroro and Ogueji, Kelechi and Rubungo, Andre Niyongabo and Nguyen, Toan Q. and Müller, Mathias and Müller, André and Muhammad, Shamsuddeen Hassan and Muhammad, Nanda and Mnyakeni, Ayanda and Mirzakhalov, Jamshidbek and Matangira, Tapiwanashe and Leong, Colin and Lawson, Nze and Kudugunta, Sneha and Jernite, Yacine and Jenny, Mathias and Firat, Orhan and Dossou, Bonaventure F. P. and Dlamini, Sakhile and de Silva, Nisansa and Çabuk Ballı, Sakine and Biderman, Stella and Battisti, Alessia and Baruwa, Ahmed and Bapna, Ankur and Baljekar, Pallavi and Azime, Israel Abebe and Awokoya, Ayodele and Ataman, Duygu and Ahia, Orevaoghene and Ahia, Oghenefego and Agrawal, Sweta and Adeyemi, Mofetoluwa. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72, 01 2022.

[34] Taku Kudo and John Richardson. SentencePiece: A Simple and Language Independent Subword Tokenizer and Detokenizer for Neural Text Processing. In Eduardo Blanco and Wei Lu, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[35] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[36] Tomasz Limisiewicz, Terra Blevins, Hila Gonen, Orevaoghene Ahia, and Luke Zettle-

moyer. Myte: Morphology-driven byte encoding for better and fairer multilingual language modeling, 2024.

[37] Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020.

[38] Gati Martin, Medard Edmund Mswahili, Young-Seob Jeong, and Jiyoung Woo. SwahBERT: Language model of Swahili. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 303–313, Seattle, United States, July 2022. Association for Computational Linguistics.

[39] Tomáš Mikolov. *Statistical Language Models Based on Neural Networks*. Ph.d. thesis, Brno University of Technology, Faculty of Information Technology, 2012.

[40] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. November 2016.

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

[42] Kelechi Ogueji, Yuxin Zhu, and Jimmy Lin. Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin, editors, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 116–126, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[43] Odunayo Ogundepo, Tajuddeen Gwadabe, Clara Rivera, Jonathan Clark, Sebastian Ruder, David Adelani, Bonaventure Dossou, Abdou Diop, Claytone Sikasote, Gilles Hacheme, Happy Buzaaba, Ignatius Ezeani, Rooweither Mabuya, Salomey Osei, Chris Emezue, Albert Kahira, Shamsuddeen Muhammad, Akintunde Oladipo, Abraham Owodunni, Atnafu Tonja, Iyanuoluwa Shode, Akari Asai, Anuoluwapo Aremu, Ayodele Awokoya, Bernard Opoku, Chiamaka Chukwuneke, Christine Mwase, Clemencia Siro, Stephen Arthur, Tunde Ajayi, Verrah Otiende, Andre Rubungo, Boyd Sinkala, Daniel Ajisafe, Emeka Onwuegbuzia, Falalu Lawan, Ibrahim Ahmad, Jesujoba Alabi, Chinedu Mbonu, Mofetoluwa Adeyemi, Mofya Phiri, Orevaoghene Ahia, Ruqayya Iro, and Sonia Adhiambo. Cross-lingual Open-Retrieval Question Answering for African Languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14957–14972, Singapore, December 2023. Association for Computational Linguistics.

[44] Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David Ifeoluwa Adelani. How Good Are Large Language Models on African Languages?, 2024.

[45] Akintunde Oladipo, Mofetoluwa Adeyemi, Orevaoghene Ahia, Abraham Owodunni, Odunayo Ogundepo, David Adelani, and Jimmy Lin. Better Quality Pre-training Data and T5 Models for African Languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 158–168, Singapore, December 2023. Association for Computational Linguistics.

[46] Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9 – 16, Mannheim, 2019. Leibniz-Institut für Deutsche Sprache.

[47] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. The fineweb datasets: Decanting the web for the finest text data at scale. *arXiv preprint arXiv:2406.17557*, 2024.

[48] Aleksandar Petrov, Emanuele La Malfa, Philip H. S. Torr, and Adel Bibi. Language Model Tokenizers Introduce Unfairness Between Languages. In *Advances in Neural Information Processing Systems*, 2023.

[49] Telmo Pires, Eva Schlinger, and Dan Garrette. How Multilingual is Multilingual BERT? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings*

of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[50] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding with unsupervised learning. 2018.

[51] Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, H. Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant M. Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake A. Hechtman, Laura Weidinger, Iason Gabriel, William S. Isaac, Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. Scaling language models: Methods, analysis insights from training gopher. *CoRR*, abs/2112.11446, 2021.

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

[53] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Jian Su, Kevin Duh, and Xavier Carreras, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics.

[54] Adam Roberts, Hyung Won Chung, Gaurav Mishra, Anselm Levskaya, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob

Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Kehang Han, Michelle Casbon, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. Scaling Up Models and Data with t5x and seqio. *Journal of Machine Learning Research*, 24(377):1–8, 2023.

[55] Noam Shazeer. GLU Variants Improve Transformer. *arXiv preprint, abs/2002.05202*, 2020.

[56] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: An Open Corpus of Three Trillion Tokens for Language Model Pretraining Research. *arXiv preprint, abs/2402.00159*, 2024.

[57] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford Alpaca: An Instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

[58] Together Computer. RedPajama: An Open Dataset for Training Large Language Models, October 2023.

[59] Atnafu Lambebo Tonja, Israel Abebe Azime, Tadesse Destaw Belay, Mesay Gemeda Yigezu, Moges Ahmed Ah Mehamed, Abinew Ali Ayele, Ebrahim Chekol Jibril, Michael Melese Woldeyohannis, Olga Kolesnikova, Philipp Slusallek, Dietrich Klakow, and Seid Muhie Yimam. EthioLLM: Multilingual large language models for Ethiopian languages with task evaluation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6341–6352, Torino, Italia, May 2024. ELRA and ICCL.

[60] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023.

[61] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[62] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[63] Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a Token-Free Future with Pre-trained Byte-to-Byte Models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022.

[64] Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics.

[65] Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. Mr. TyDi: A Multi-lingual Benchmark for Dense Retrieval. In Duygu Ataman, Alexandra Birch, Alexis Conneau, Orhan Firat, Sebastian Ruder, and Gozde Gul Sahin, editors, *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 127–137, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[66] Xinyu Zhang, Kelechi Ogueji, Xueguang Ma, and Jimmy Lin. Toward Best Practices for Training Multilingual Dense Retrieval Models. *ACM Trans. Inf. Syst.*, 42(2), sep 2023.

[67] Xinyu Zhang, Nandan Thakur, Odunayo Ogundepo, Ehsan Kamalloo, David Alfonso-Hermelo, Xiaoguang Li, Qun Liu, Mehdi Rezagholizadeh, and Jimmy Lin. MIRACL: A Multilingual Retrieval Dataset Covering 18 Diverse Languages. *Transactions of the Association for Computational Linguistics*, 11:1114–1131, 09 2023.

[68] Judit Ács. Exploring BERT's Vocabulary. 2019.

[69] Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model. *arXiv preprint arXiv:2402.07827*, 2024.