Pula: Training Large Language Models for Setswana

Anonymous ACL submission

Abstract

In this work we present **Pula**, a suite of bilingual language models proficient in both Setswana and English. Leveraging recent advancements in data availability and efficient fine-tuning, Pula 8B and Pula 14B outperform GPT 4o and Gemini 1.5 Pro on English-Setswana translation tasks and achieve state-of-the-art performance on Setswana reasoning tasks for its size. We release the weights for Pula **1B**, **3B**, **8B**, and **14B** as well as training logs and training and evaluation code. Alongside Pula, we release the largest-ever Setswana text corpora, **Marothodi**, and the first Setswana instruction-tuning dataset, **Medupe**, consisting of reformatted datasets, translated data, and synthetic LLM-generated data. To accompany this data, we release the code used for dataset construction, formatting, filtering, and scraping. Last, we release two Setswana LLM-translated benchmarks, **MMLU-tsn** and **GSM8K-tsn**, to measure Setswana knowledge and reasoning capabilities.[1]

1 Introduction

Setswana, also known as Tswana, is a Bantu language spoken by an estimated five to ten million people worldwide (Bennett et al., 2016). Closely related to Northern Sotho and Southern Sotho, Setswana holds official status in Botswana (Government of Botswana, 2024), South Africa (The Republic of South Africa, 1996), and Zimbabwe (The Parliament of Zimbabwe, 2013), and is also used in countries like Namibia, often interchangeably with English (Norris, 2017). Despite its significance in the lives of millions, Setswana has been largely overlooked in natural language processing (NLP) research, despite being classified by some works as a high-resource language(NLLB Team et al., 2022). This work aims to bridge the gap between Setswana and other high-resource languages by making open generative large language models capable of high-quality Setswana available to the community for the first time, significantly increasing data availability, and laying the groundwork for future Setswana-centric research.

Large Language Models (LLMs) have demonstrated powerful capabilities across various domains after training on web and synthetic data (OpenAI et al., 2023; Dubey et al., 2024; Anthropic, 2024; Gunasekar et al., 2023), excelling in areas such as mathematics (Mistral AI, 2024), programming (Guo et al., 2024), creative writing (Wang et al., 2024), and translation tasks (Vaswani et al., 2017). However, developers continue to primarily target English and certain high-resource languages in training and evaluation. While existing approaches yield impressive capabilities, they may produce models which lack knowledge of certain cultures, limit production use-cases outside majority demographics, or prevent a significant portion of the global population from utilizing language models effectively. African languages like Setswana, with little textual data available compared to languages such as English or French, subsequently suffer in performance and are underutilized in research.

Recent progress has been made to address the lack of language diversity in language models. Releases such as mBART (Liu et al., 2020), XLM-RoBERTa (Conneau et al., 2019), and BLOOM (BigScience Workshop et al., 2022) were among the earliest and most influential advancements in multilingual language models. Building upon these technologies, newer models like GPT-4 (OpenAI et al., 2023), Claude (Anthropic, 2024), Gemini (Gemini Team et al., 2023), Llama (Dubey et al., 2024), and Gemma (Gemma Team et al., 2024) have also found success in multilingual do-

[1]Weights, data, benchmarks, code, and training logs will all be released upon acceptance.

mains, often demonstrating reasoning and translation capabilities in languages not officially supported. Releases such as Aya 101 (Üstün et al., 2024) and Aya 23 (Aryabumi et al., 2024) have continued to improve language coverage for translation and generative tasks, and open corpora such as ROOTS (Laurençon et al., 2023), OSCAR (Suárez et al., 2019), and mc4 (Caswell et al., 2021) have made multilingual pre-training data readily available. However, Setswana comprises only a small fraction of these datasets, with just 0.0002% of the ROOTS corpus being written in Setswana. Moreover, much of the available Setswana text in open multilingual corpora is of lower quality or predominantly religious in nature, resulting in significantly worse conversational, translation, and reasoning capabilities in current open models.

To help address this issue, we introduce the **Pula** series of language models. This series consists of LoRA (Hu et al., 2021) and QLoRA (Dettmers et al., 2023) fine-tuned versions of Llama 3.2 1B, Llama 3.2 3B,, Llama 3.1 8B, and Qwen 2.5 14B, and AfroXLMR-base (279M) (Yang et al., 2024; Dubey et al., 2024; Alabi et al., 2022). By training a range of models across different parameter counts, we aim to provide the research community and millions of Setswana speakers with models that are both highly performant and capable of running on various hardware configurations ranging from data centers to consumer laptops and mobile phones. To fuel the training behind these models and to provide the research community with resources to further improve future models and research, we hand-curate the largest-ever corpora of Setswana text. In doing so, we merge several existing corpora which have not yet been consolidated, restore document-level text in certain data subsets, locate new sources of Setswana text which to our knowledge have not yet been utilized in NLP research, develop several dedicated scrapers and parsers to obtain this new data, reformat existing datasets to function as instruction-tuning corpora, translate existing English instruction-tuning datasets to Setswana using state-of-the-art large language models and translation models, and utilize GPT-4o for the generation of entirely new synthetic Setswana text. We name our pre-training dataset **Marothodi** and our instruction-tuning dataset **Medupe**.

Following the open approach of projects like OMLo (Groeneveld et al., 2024) and Dolma (Soldaini et al., 2024), we perform a fully open release. This includes model weights, all training data, metadata on data sources, training logs, and code for training, evaluation, dataset curation, dataset translation, synthetic data generation, and web scraping. We release all model weights and data on Hugging Face Hugging Face and the Pula training logs can be accessed here.

2 Related Work

Although much of current NLP research is English-centric, there have been significant recent advancements in Setswana-centric language models and data access. TswanaBERT (Motsoehli, 2020) represents one of the earliest examples, trained on over ten thousand Setswana sentences from the Leipzig Corpora Collection (Goldhahn et al., 2012), SABC news headlines (Marivate et al., 2020), and various blogs and websites. More recently, the NCHLT Setswana RoBERTa model (Eiselen et al., 2023) was released, having been trained on over fourteen million tokens of Setswana text from the NCHLT (Eiselen and Puttkammer, 2014), Autshumato (McKellar et al., 2016), Leipzig (Goldhahn et al., 2012), and Common Crawl corpora, and an internal CTexT corpus. PuoBERTa (Marivate et al., 2023) marked a significant step forward in masked language modeling, achieving state-of-the-art performance while being the first langauge model trained from-scratch. PuoBERTa was released alongside PuoData, the largest collection of curated Setswana text at the time, totaling 4.5 million PuoBERTa tokens excluding the JW300 (Agić and Vulić, 2019).

Much of the literature on African languages has targeted massively multilingual developments with a focus on machine translation and transfer learning. Corpora such as OPUS (Tiedemann, 2012a) and MADLAD-400 (Kudugunta et al., 2023) provide access to large volumes of parallel web text data, while work such as MAFAND-MT (Adelani et al., 2022a) have enabled improved translation performance across many African languages through additional human-generated data. Meta's No Language Left Behind (NLLB) (NLLB Team et al., 2022) has facilitated high-quality machine translation between over 200 languages, including Setswana, although it is noted as one of the 21 languages with the lowest accuracy on FLORES-200 (Goyal et al., 2022). MADLAD-400 (Kudugunta et al., 2023) has allowed for in-

creased multilinguality but we found suffers from worse performance on Setswana-English translations, as seen in Table 2. Furthermore, models such as AfriBERTa (Ogueji et al., 2021) and AfroXLMR (Alabi et al., 2022) have seen success in training masked language models across multiple African and low-resource languages, and LlamaX (Lu et al., 2024) has demonstrated high degrees of translation performance across over 100 languages as a generative LLM while retaining generalization capabilities.

Thanks to the cumulative improvements made by the research community in Setswana-centric NLP and increased levels of multilinguality, we believe there are significant opportunities to enhance existing Setswana NLP systems. Through our investigations we find several smaller corpora of Setswana text are often underutilized in the literature, potentially due to differing distribution methods and vastly differing formats, making data difficult to locate and utilize. For example, the South African Center for Digital Language Resources (SADiLaR) [2] (SADiLaR, 2024) has made publicly available several datasets of Setswana text and audio. However, as the format for this distributed text is non-standardized and much of this data is not available on commonly used external platforms such as Hugging Face or Kaggle, much of this data is left out of massive public datasets. We also find many websites and associated PDFs written in Setswana are excluded from existing datasets, meaning a significant portion of the available high-quality, and especially educational, Setswana data is not being utilized. Through the development and release of Marothodi and Medupe, we aim to reduce this burden and make data significantly easier to access.

3 Data

One key consideration in selecting training data for Pula is identifying the target languages. In the case of translation, while several parallel corpora are available for Setswana, with some such as OPUS containing Setswana text alongside multiple other languages, there is significantly more English-Setswana parallel text available. For example, OPUS contains over four times more parallel Setswana-English sentences than parallel Setswana-French sentences [3]. In addition, the majority of publicly available, high-quality, educational, and instruction-tuning datasets are written in English. Consequently, many state-of-the-art large language models already excel in English tasks, including the models we select for continued pre-training. Additionally, South Africa and Botswana, the two countries with the largest Setswana-speaking populations, often utilize English in legal, official, and government documents, with English sometimes being spoken interchangeably with Setswana (Parliament of Botswana, 2022). We observe this trend continued in web text documents, where many data sources are written in both English and Setswana, either incorporating code-switching or providing direct translations (TRK, 2021). Due to the strong link between these two langauges and in an effort to reduce computational demands, we focus our primary efforts on curating a high-quality dataset consisting of Setswana and English texts. We refer to our corpus of raw Setswana text as **Marothodi**, and our instruction-tuning dataset as **Medupe**.

Our datasets build upon several prior works in the African NLP research community. For consistency, when measuring token counts, we utilize the Llama 3.1 tokenizer throughout this paper. We release both Marothodi and Medupe in their entirety on Hugging Face [4]. To assist researchers in future potential analysis and deduplication work, we provide alongside each sequence metadata on its source, including a corpus identifier and an exact source URL where available.

3.1 Marothodi

Our core focuses in the development of Marothodi are to dramatically increase the amount of publicly available Setswana text and to significantly increase ease of access. We believe a significant limitation is the sparsity of published data: resources are scattered across Hugging Face, SADiLaR, massive web crawls, public websites, and cloud storage providers. As a result, researchers who are not deeply familiar with the current Setswana data landscape or do not have ample time to allocate toward data collection are severely limited in data diversity, quality, and quantity. To assist in resolving this problem we develop Marothodi, a Setswana pre-training dataset pulling together data

[2] https://sadilar.org

[3] https://opus.nlpl.eu/results/tn&en/corpus-result-table

[4] https://example.com

from 32 different sources totalling 74 million tokens, nearly 8 times as many tokens compared to PuoData, making Marothodi the largest available single corpus of Setswana text.

We increase available Setswana text by selectively including various underutilized data sources. We prioritize document integrity and rich context by utilizing complete documents from various sources. To aid in this effort, we restore the original documents of certain subsets of PuoData, including sources such as Wikipedia (Foundation), Nalibali [5], and the Setswana Bible. In addition, we further increase the available Nalibali data by performing PDF text extraction on previously unutilized children's stories and educational material.

Prior works predominantly focus on certain commonly utilized Setswana corpora, limiting downstream model generalization. Marothodi addresses this by directly targeting a rich selection of diverse, manually curated sources. We manually identify web sources containing Setswana text and develop individual programs to scrape the data as needed. We found many of these sources contain previously underutilized Setswana text in the form of PDFs, making text extraction an important part of Marothodi. We utilize *pypdfium2* to extract text from less readily accessible collections such as various children's stories (Nalibali, 2024), educational material (Lingua, 2024), governmental documents [6], and rare books (Rahlao et al., 2021). We also include text from the Peace Corps' Intro to Spoken Setswana which, due to the lower quality of both the scan and the embedded text, required image processing, object character recognition (OCR) using Florence-2 (Xiao et al., 2023), and textual reformatting using Llama 3 70B.

A significant source of educational Setswana data in Marothodi is the National Education Collaboration Trust (National Education Collaboration Trust, 2024) and the Thutong South African Educational Portal (Thutong South African Education Portal, 2024). Both sources contain a plethora of Setswana exams, study material, and books covering a wide variety of target age demographics. This data required significant manual intervention to fully acquire, and we believe its inclusion makes Marothodi the largest single corpora of human-written Setswana educational material.

Furthermore, we scrape the contents of various websites containing either Setswana or code-switched Setswana-English text, such as iAfrika (iAfrika, 2024), Setswana Mo Botswana (TRK, 2021), Tlhalefang Communications (Tlhalefang, 2009), and the University of South Africa (Unisa, 2023). We also include a small corpus of parallel text consisting of monolingual English mathematical text translated into code-mixed English and Setswana (Mokoka, 2024). Furthermore, we include miscellaneous individual documents such as the United States Peace Corps' Intro to Spoken Setswana (Mistry and Gare, 1987), and the Setswana Universal Declaration of Human Rights (Nations, 1998).

Moreover, we include five separate corpora from SADiLaR in Marothodi. This includes a corpus of multilingual code-switched soap opera speech, transcripts from a high quality corpus of Setswana text-to-speech dataa, transcripts from the Lwazi Setswana ASR and TTS corpora, and transcripts from the NCHLT Setswana auxillary speech corpus. We also incorporate smaller existing corpora into Marothodi such as Bloom-lm, GlotCC, and MADLAD-400.

To further increase the size of our corpus, we finally employ machine translation using Meta's *NLLB-200-3.3B* (NLLB Team et al., 2022) to translate subsets of TinyStories to Setswana. We found through evaluating round-trip translations (Moon et al., 2020) that TinyStories' simplistic vocabulary made for generally higher-quality translations when compared to more educational text such as FineWeb-edu (Penedo et al., 2024).

3.2 Medupe

Curating an instruction-tuning dataset for Setswana required both data collection and a strong focus on data augmentation. Given the lack of available Setswana instruction-following corpora, we found success in augmenting existing sources to fit the user-assistant format expected by most LLMs. Medupe consists of data sourced in three ways: 1) Augmenting existing corpora such as parallel texts, named entity recognition (NER) and part-of-speech (POS) tagging, 2) Translating existing corpora using NLLB 200 3.3B, GPT-4o, Gemini 1.5 Pro, and quantized Llama 3.1 405B, and 3) Purely synthetic Setswana text generated with GPT-4o.

Medupe's largest source of Setswana text is the OPUS corpus (Tiedemann, 2012a), which in total includes over six million parallel English-

[5] https://nalibali.org/

[6] https://www.parliament.gov.bw/index.php?

Setswana sentences across the NLLB (Schwenk et al., 2021; Fan et al., 2021), CCAligned (El-Kishky et al., 2020), XLEnt (El-Kishky et al., 2021), Wikimedia (Tiedemann, 2012b), and Tatoeba (Tiedemann, 2012b; Tatoeba, 2024) corpora. However, we find training on this data results in catastrophic forgetting, which we attribute to the large quantity of training sequences which yield a greatly imabalanced dataset. In addition, we find many English-Setswana parallel sequences to be of low quality which may reduce both model generalization and translation capabilities. To mitigate these issues we translate every English sequence to Setswana using NLLB (NLLB Team et al., 2022) and calculate the CHRF score using the translated Setswana text as a prediction and the Setswana sequence from OPUS as our reference. We then filter to include only the top 33% of OPUS examples with the greatest CHRF scores. In doing so, we yield a subset of OPUS whose parallel sequences have greater levels of agreement with existing translation systems and we find are typically of greater quality.

Much of Medupe's data sources focus on translation tasks. In addition to OPUS, this includes the Autshumato (Groenewald and du Plooy, 2010), MAFAND-MT training split (Adelani et al., 2022a), PolyNews-Parallel (Iana et al., 2024), CaLMQA (Arora et al., 2024), SIB-200 (Adelani et al., 2023), Microsoft Terms (Microsoft, 2022), xP3x (Muennighoff et al., 2022), the SADiLaR corpus of COVID-19 multilingual terminology (City of Tshwane et al., 2021), the SADiLaR Corpus of Generic Multilingual Academic Word-lists (van Dyk, 2021), and the SADiLaR Autshumato Corpus of Multilingual Word and Phrase Transitions (Puttkammer and Hocking, 2021). Each training example is formatted to mimic a user-assistant interaction (e.g. "Can you translate the following from English to Setswana..."). To ensure prompt diversity and encourage generalization, we randomize the language of the system prompt, the verbiage used in the system prompt, whether the source text is provided before or after the user query, the translation direction, and the language used in the user query for all datasets where applicable.

We perform data augmentation on a variety of other data sources. This includes Daily News Dikgang for news classification (Marivate et al., 2023), MasakhaNER 2.0 for NER (Adelani et al., 2022b), and the SADiLaR NCHLT Setswana Annotated Text Corpora for lemmatization and POS tagging (Puttkammer et al., 2021). We similarly randomize aspects of the system and user prompt where applicable.

While these forms of data augmentation are important for data diversity, the number of non-translation sources available is rather low, especially in educational content. Given the plethora of translation tools available and the success seen in instruction-tuning on machine translation text in prior works, we sought to improve data diversity via the translation of existing instruction-following datasets. To accomplish this, we utilize NLLB 200 3.3B to translate 15,201 examples from OpenHermes-2.5 (Teknium, 2023), a collection of several commonly used chat datasets, and 5,201 examples from Wildchat-1M (Zhao et al., 2024), a collection of one million ChatGPT conversations. To encourage multi-turn conversations, we filter Wildchat-1M to only include examples with at least three user-assistant turns, and we exclude toxic and non-English conversations. Both datasets are further filtered after translation to remove translations containing signs indicative of low-quality translations such as repetitive text or non-Latin characters.

While machine translation models such as NLLB 200 (NLLB Team et al., 2022) and MADLAD 400 (Kudugunta et al., 2023) are useful for translating simpler texts, as we previously saw for TinyStories, we note that while these models often achieve state-of-the-art translations, they typically do not preserve the formatting of the original text. This can be a problem when user queries or assistant responses contain code, bulleted or numbered lists, mathematical equations, or tabular data. In addition, we find these models may output the input text verbatim when the input text contains code or certain technical jargon. To obtain additional high-quality Setswana user-assistant conversations while hopefully avoiding these problems, we translate subsets of several corpora using GPT-4o, Gemini 1.5 Pro, and AWQ INT4 quantized Llama 3.1 405B. We utilize GPT-4o to translate 69,585 sequences from Dolly (Conover et al., 2023), 74,185 sequences from Magpie (Xu et al., 2024), and 23,790 sequences from UltraChat 200k (Ding et al., 2023; Tunstall et al., 2023), Gemini 1.5 Pro to translate 24,525 sequences from OpenHermes 2.5 (Teknium, 2023) and 9,832 sequences from The Tome (Arcee AI, 2024), and quantized Llama 3.1 405B to translate 130,800 sequences

from The Tome. We utilize the same filters as previously discussed with NLLB to remove low-quality translations. In addition, the LLMs will occasionally corrupt the format of the translated conversation, such as by hallucinating multiple system prompts or multiple user or assistant responses in a row, making the conversation unusable. We remove these instances from our corpus. This was significantly more common with quantized Llama 3.1 405B.

To further improve the writing quality of the Pula models and to experiment with the creation of purely synthetic Setswana text, we utilize *gpt-4o-2024-05-13* (OpenAI, 2024) to generate 7,860 pieces of Setswana text covering several topics in various writing styles. To achieve data diversity between samples, we use existing web data as a seed. First, a random subset of FineWeb and FineWeb-edu (Penedo et al., 2024) is selected. Then, given a sequence from one of these datasets, Llama 3 70B is prompted to determine five new pieces of writing that cover similar topics. For example, given a web document discussing the mathematics behind gravity, Llama 3 may suggest "A children's book explaining the story of an apple landing on Isaac Newton's head." Additionally, Llama 3 is prompted to select the five most unique words present in the provided sequence. We then cross-reference the Setswana-English split of Google Research's GATITOS (Google Research, 2024), a human-translated lexicon, and filter to only include instances where Llama selected at least one English word which is present in GATITOS. This ensures we can obtain a high-quality translation of each selected word present in this lexicon. Finally, prompts are constructed to be provided to GPT-4o consisting of: 1) A system prompt to instruct the model to write in Setswana, 2) One of the five pieces of writing, randomly selected, and 3) Each English word selected by Llama 3 which is present in GATITOS as well as its Setswana translation. We find that utilizing this methodology allows GPT-4o to generate high-quality Setswana writings while maintaining a high degree of diversity between texts. We retain the instructions sent to GPT-4o as well as the corresponding response in Medupe.

4 Training

We train all Pula models for three epochs, totaling approximately 2.3 billion tokens. Training is performed utilizing the Hugging Face *transformers*, *trl*, and *peft* libraries. Our approach involves performing continued pre-training for each model to ensure robust learning without suffering catastrophic forgetting. We utilize Low Rank Adaptation (LoRA) (Hu et al., 2021) and Quantized Low Rank Adaptation (QLoRa) (Dettmers et al., 2023) during training across the Query, Key Value, Output, Gate, Up and Down projection matrices. We perform full-precision training on the language modeling head and the embedding layers. This allows for effective learning of the Setswana language without significant computational and memory requirements, but also mitigates the risk of over-fitting and catastrophic forgetting. Pula 1B, 3B, and 8B are trained using LoRA on sequences up to 4096 tokens in length whereas Pula 14B is trained using 4-bit QLoRa on sequences up to 2048 tokens in length.

Whereas training may typically undergo two phases, with continued pre-training being performed on a base model followed by supervised fine-tuning (SFT), we opt to train on a mixture of raw text and instruction data. We train existing instruction-following models on a corpus consisting of Marothodi, our raw pre-training data, and Medupe, our instruction-tuning data, at the same time. By utilizing a mixture of raw text and instruction data we yield models which are capable of instruction-following without the additional compute required for separate pre-training and instruction-tuning stages, and without the risk of losing the knowledge that may have been learned in the original model's SFT or RLHF phases.

To further increase the proportion of Setswana reasoning text present in our training dataset, we duplicate each of our synthetic LLM datasets (Dolly, Magpie, Ultrachat, OpenHermes, The Tome) three times. To retain English capabilities and encourage generalization, we include a 15% mixture of OpenHermes 2.5 (Teknium, 2023), Magpie Pro MT 300k(Xu et al., 2024), Aya Dataset(Singh et al., 2024), and Inkuba Instruct(Tonja et al., 2024). We do so to not only maintain English reasoning and conversational capabilities, but to retain other multilingual capabilities and encourage cross-lingual transfer learning.

All models are trained using DeepSpeed (Rasley et al., 2020) and ZeRO Stage 3 (Rajbhandari et al., 2020) on eight NVIDIA H100 GPUs. We provide the full set of hyperparameters utilized during training in Table 3.

Model	Avg.	MAFAND-MT				Lego-MT				FLORES-200			
		eng-tsn		tsn-eng		eng-tsn		tsn-eng		eng-tsn		tsn-eng	
		CHRF	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF	BLEU	CHRF	BLEU
Llama 3 Instruct (8B)	11.76	22.90	2.60	29.52	5.73	11.02	1.69	11.61	1.70	20.78	2.37	27.04	4.15
Llama 3 Instruct (70B)	15.48	34.97	7.26	37.33	6.42	11.37	0.58	11.93	0.75	31.47	5.50	33.10	5.07
Aya 23 (8B)	7.31	14.85	0.74	17.17	1.82	8.49	0.94	9.52	1.25	14.19	0.86	16.56	1.33
Aya 23 (35B)	9.77	14.95	0.88	5.96	28.88	8.37	0.89	10.16	1.19	13.05	0.63	27.09	5.13
LLaMAX3 (8B)	11.59	23.39	2.00	27.55	5.62	11.57	1.29	13.21	1.98	21.51	1.64	25.12	4.14
MADLAD-400 MT (10B)	17.16	22.06	6.07	34.86	14.06	16.85	9.66	22.33	14.60	19.39	3.51	31.40	11.14
NLLB-200 (3.3B)	28.46	57.64	28.15	46.99	20.66	23.76	13.53	17.10	4.36	50.15	21.73	41.30	16.16
GPT-4o	30.64	51.07	23.08	60.91	35.28	20.62	6.33	19.57	8.22	45.11	17.41	52.71	27.40
GPT-4o Mini	25.33	40.12	14.20	53.11	26.18	21.04	9.00	21.16	8.22	35.29	10.05	45.98	19.59
Gemini 1.5 Pro	30.67	55.71	26.29	58.87	34.10	18.10	3.99	17.20	4.72	49.81	21.59	51.16	26.49
Gemini 1.5 Flash	26.50	46.98	16.20	55.01	29.66	18.23	4.42	17.42	3.65	42.41	11.87	48.51	23.60
Pula-1B	16.35	28.06	5.65	38.21	13.69	15.25	3.04	16.21	4.40	28.18	4.65	31.03	7.84
Pula-3B	23.05	38.80	11.41	50.41	25.15	17.71	3.80	19.35	7.17	36.00	8.90	41.57	16.31
Pula-8B	32.24	53.79	26.81	57.06	32.43	23.92	12.02	23.33	14.91	47.38	20.14	49.55	25.55
Pula-14B	**33.07**	55.57	28.48	57.31	31.94	24.90	12.07	24.34	17.40	47.78	21.00	49.98	26.11

Table 1: Translation performance across open and closed models on the MAFAND-MT, Lego-MT, and FLORES-200 benchmarks. We report CHRF and BLEU scores for English-Setswana and Setswana-English translation as well as average overall score.

Model	Avg.	Belebele		MMLU		GSM8K	
		tsn	eng	tsn	eng	tsn	eng
Llama 3.1 Instruct 8B	50.32	31.04	89.67	27.73	63.75	6.61	83.09
Llama 3.1 Instruct 70B	67.69	50.69	**96.00**	39.95	**84.28**	40.24	**95.00**
Aya 23 8B	32.23	30.33	62.00	26.09	44.58	2.05	28.35
Aya 23 35B	39.99	29.16	80.85	29.29	58.42	2.81	39.42
LLaMAX3 8B	27.67	30.53	68.53	25.53	41.13	0.23	0.07
Pula 1B	21.06	27.03	36.58	24.38	28.58	3.49	6.29
Pula 3B	35.35	30.93	50.73	30.76	44.03	11.39	44.28
Pula 8B	59.81	51.24	85.43	38.66	61.86	43.28	78.39
Pula 14B	**69.40**	**66.02**	91.50	**46.15**	74.27	**57.36**	81.12

Table 2: Performance comparison of different large language models on the Belebele, MMLU, and GSM8K benchmarks. For Setswana, we evaluate using the Setswana Belebele subset, MMLU-tsn, and GSM8K-tsn. For English, we evaluate using the English Belebele split, MMLU, and GSM8K.

5 Evaluation

We evaluate our models on a variety of tasks in both Setswana and English including translation, natural language understanding, multiple choice question-answering, and mathematical reasoning. All evaluations are performed using bf16 precision across 8 NVIDIA H100 GPUs using vLLM (Kwon et al., 2023).

To measure translation performance we evaluate on the MAFAND-MT (Adelani et al., 2022a), Lego-MT (Yuan et al., 2023), and Flores-200

(NLLB Team et al., 2022) benchmarks. These benchmarks cover a variety of translation domains, including news headlines, web articles, and miscellaneous crawled web documents. To measure performance we utilize the BLEU (Papineni et al., 2002) and CHRF (Popović, 2015) metrics. Our translation performance results are presented in Table 1.

To evaluate Pula's knowledge and reasoning capabilities in Setswana we develop MMLU-tsn and GSM8K-tsn. These are Setswana translations of the entirety of the test splits of the original Massive Multitask Language Understanding (MMLU) (Hendrycks et al., 2021b,a) and Grade School Math 8K (GSM8K) (Cobbe et al., 2021) benchmarks, translated using GPT-4o and Gemini 1.5 Pro, respectively. We acknowledge this method's reliance on translation systems to translate this task, especially given its technical nature, is likely to suffer from "translationese" and other errors (Plaza et al., 2024). Many sequences may be incorrectly translated, biased, or otherwise impossible to solve without lucky guessing. However, we find these translated benchmarks to still be a useful proxy for a model's performance in reasoning, knowledge, and instruction-following capabilities. These results are provided in Table 2.

We find the Pula series excels at translating between English and Setswana, with both Pula 14B and Pula 8B on average outperforming significantly larger frontier models such as GPT-4o and Gemini 1.5 Pro. These models exceed the performance of all tested open LLMs as well as the NLLB 200 and MADLAD 400 machine translation models. We note the Pula series tends to yield higher quality translations when translating from Setswana to English, a trend we see in other tested LLMs but not machine translation models.

We evaluate reading comprehension using Meta's Belebele benchmark, a corpus of multiple-choice questions regarding passages sourced from FLORES-200. We evaluate world knowledge and question-answering via the Measuring Massive Multitask Language Understanding (MMLU) and MMLU-tsn benchmarks, containing multiple-choice questions on various topics such as mathematics, computer science, law, and more. Last, we evaluate mathematical reasoning via GSM8K and GSM8K-tsn, which consist of open-ended grade-school math word problems.

We find Pula 14B to, on average, outperform all tested open models across Setswana and English reasoning tasks. Pula 8B outperforms or is competitive with Llama 3.1 70B on all three benchmarks, and Pula 14B significantly outperforms Llama 3.1 70B on all Setswana benchmarks. In addition, we find Pula's performance steadily grows with scale in both Setswana and English tasks; a trend displayed in the performance of existing Llama and Aya models. However, we do note the roughly equivalent performance of Aya 23 8B and 35B on Setswana Belebele. Given the lack of intentional Setswana data present in Aya 23's training corpus, this indicates reading comprehension may be a less transferable skill for LLMs when working with largely unseen languages compared to question-answering and mathematical reasoning. We also note large discrepancies between Setswana and English versions of benchmarks, such as MMLU and MMLU-tsn. We attribute these differences to accumulative errors during translation, such as ambiguous or incorrect wording, impossible questions, or damaging modifications to the correct answer.

6 Conclusion

In this work we introduce Pula, the first series of large language models tailored for Setswana. Our models demonstrate significantly improved translation and reasoning performance, rivaling models much larger than themselves on Setswana reading comprehension, question-answering, and mathematical reasoning tasks while retaining existing performance on English tasks. Pula exceeds in translating between Setswana and English, with Pula 8B and 14B on average outperforming GPT-4o and Gemini 1.5 Pro. In addition, we introduce Marothodi, the largest-ever single corpus of raw Setswana text, and Medupe, the tirst-ever Setswana instruction-tuning dataset. Our results indicate not only the quality and diversity of these datasets, but also there may be significant performance gains not yet reached in other languages which may be available using existing underutilized data and synthetic data generation. To further support Setswana evaluations, we develop and release MMLU-tsn and GSM8K-tsn, Setswana translations of the MMLU and GSM8K benchmarks translated using GPT-4o and Gemini 1.5 Pro. To support future multilingual and Setswana-centric NLP research and production use-cases, we release all model weights, data, benchmarks, training and evaluation code, and training logs.

7 Limitations

A foundational source for Medupe is translated Setswana instructions and synthetic data using NLLB 200 3.3B, GPT-4o, Gemini 1.5 Pro, and Llama 3.1 405B. The quality of these translations directly influences the quality of Medupe and Pula's downstream performance on Setswana tasks. Any inaccuracies, biases, or nuances lost during translation may propagate into the training data and even become more pronounced (Gallegos et al., 2023).

On a similar note, utilizing translations for benchmarking may introduce "translationese", such as direct translations rather than natural language (Doshi et al., 2024). These errors may distort the benchmark's authenticity and reduce the number of answerable questions with corresponding correct answers. These errors may be especially the case for MMLU-tsn and GSM8K-ts, where certain domain-specific vocabulary may not have direct Setswana equivalents.

Last, despite extensive efforts to curate comprehensive corpora of Setswana text, certain cultural and contextual elements may be underrepresented. Dialectical variations, cultural narratives, or region-specific terms, phrases, or other terminology may be comparatively sparse in this corpora. This may lead to models that are less effective in certain situations which may require additional cultural insight or sensitivity (Mousi et al., 2024). Addressing these cultural nuances remains an ongoing challenge and an area for future research to ensure language models are properly culturally knowledgeable and configurable for Setswana speakers.

8 References

References

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. 2022a. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O. Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *Preprint*, arXiv:2309.07445.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Oluwadara Alabi, Shamsuddeen Hassan Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing K. Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris C. Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine W. Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen R. Gwadabe, Tosin P. Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius M Ezeani, Chiamaka Ijeoma Chukwuneke, Mofetoluwa Adeyemi, Gilles Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. 2022b. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. *ArXiv*, abs/2210.12391.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

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

Anthropic. 2024. Claude 3.5 sonnet.

Arcee AI. 2024. The tome. https://huggingface.co/datasets/arcee-ai/The-Tome.

Shane Arora, Marzena Karpinska, Hung-Ting Chen, Ipsita Bhattacharjee, Mohit Iyyer, and Eunsol Choi.

2024. Calmqa: Exploring culturally specific long-form question answering across 23 languages.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024. Aya 23: Open weight releases to further multilingual progress.

Wm. G. Bennett, Maxine Diemer, Justine Kerford, Tracy Probert, and Tsholofelo Wesi. 2016. Setswana (south african). *Journal of the International Phonetic Association*, 46(2):235246.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario ako, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León

Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Rubungo Andre Niyongabo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2021. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *arXiv e-prints*, arXiv:2103.12028.

City of Tshwane, South African Centre for Digital Language Resources (SADiLaR), Department of Science and Innovation (DSI), and Pan South African Language Board (PanSALB). 2021. Covid-19 multilingual terminology. Sponsored by the South African Centre for Digital Language Resources and the Department of Science and Innovation.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Zhi Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. *Preprint*, arXiv:2305.14233.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Do not worry if you do not have data: Building pretrained language models using translationese.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke

de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang,

Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

Roald Eiselen, Rico Koen, Albertus Kruger, and Jacques van Heerden. 2023. Nchlt setswana roberta language model.

Roald Eiselen and Martin Puttkammer. 2014. Developing text resources for ten South African languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3698–3703, Reykjavik, Iceland. European Language Resources Association (ELRA).

Ahmed El-Kishky, Vishrav Chaudhary, Francisco GuzmÃąn, and Philipp Koehn. 2020. CCAligned: A massive collection of cross-lingual web-document pairs. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 5960–5969, Online. Association for Computational Linguistics.

Ahmed El-Kishky, Adithya Renduchintala, James Cross, Francisco Guzmán, and Philipp Koehn. 2021. Xlent: Mining a large cross-lingual entity dataset with lexical-semantic-phonetic word alignment. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10424–10430.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Wikimedia Foundation. Wikimedia downloads.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2023. Bias and fairness in large language models: A survey.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogoziska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel,

Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lui, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çalar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakievi, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kpa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen,

Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Brainskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Pawe Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikoaj Rybiski, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan

Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Pluciska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2023. Gemini: A family of highly capable multimodal models.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikua, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani,

Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Google Research. 2024. Gatitos.

Government of Botswana. 2024. About our country. https://www.gov.bw/about-our-country. Accessed: 2024-07-30.

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

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models.

Hendrik J Groenewald and Liza du Plooy. 2010. Processing parallel text corpora for three south african language pairs in the autshumato project. *AfLaT 2010*, page 27.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need.

Daya Guo, Qihao Zhu, Dejian Yang, Zhenda Xie, Kai Dong, Wentao Zhang, Guanting Chen, Xiao Bi, Y. Wu, Y. K. Li, Fuli Luo, Yingfei Xiong, and Wenfeng Liang. 2024. Deepseek-coder: When the large language model meets programming – the rise of code intelligence.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021a. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021b. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

iAfrika. 2024. iafrika.

Andreea Iana, Fabian David Schmidt, Goran Glava, and Heiko Paulheim. 2024. News without borders: Domain adaptation of multilingual sentence embeddings for cross-lingual news recommendation. *Preprint*, arXiv:2406.12634.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. Madlad-400: A multilingual and document-level large audited dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 67284–67296. Curran Associates, Inc.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario ako, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2023. The bigscience roots corpus: A 1.6tb composite multilingual dataset.

Live Lingua. 2024. Learn setswana.

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

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. Llamax: Scaling linguistic horizons of llm by enhancing translation capabilities beyond 100 languages. *arXiv preprint arXiv:2407.05975*.

Vukosi Marivate, Moseli Mots'Oehli, Valencia Wagner, Richard Lastrucci, and Isheanesu Dzingirai. 2023. Puoberta: Training and evaluation of a curated language model for setswana. In *SACAIR 2023 (To Appear)*.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. In *Proceedings of the first workshop on Resources for African Indigenous Languages*, pages 15–20, Marseille, France. European Language Resources Association (ELRA).

Cindy McKellar, Roald Eiselen, and Wikus Pienaar. 2016. Autshumato english-setswana parallel corpora.

Microsoft. 2022. Ms terms. https://huggingface.co/datasets/microsoft/ms_terms/commits/main.

Mistral AI. 2024. Mathstral.

Karen S. Mistry and Grace Gare. 1987. An introduction to spoken setswana.

Keneilwe Mokoka. 2024. Exploring machine translation for code-switching between english and setswana in south african classrooms.

Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020. Revisiting round-trip translation for quality estimation. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 91–104, Lisboa, Portugal. European Association for Machine Translation.

Moseli Motsoehli. 2020. Tswanabert.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2024. AraDiCE: Benchmarks for dialectal and cultural capabilities in llms.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Nalibali. 2024. Nalibali. https://nalibali.org/.

National Education Collaboration Trust. 2024. National education collaboration trust. https://nect.org.za.

United Nations. 1998. Universal declaration of human rights - western sotho/tswana/setswana.

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

Kelsey Norris. 2017. Sound and silence. *Oxford American*. A conversation with Gothataone Moeng.

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

OpenAI. 2024. Hello gpt-4o.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey,

Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, ukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, ukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Parliament of Botswana. 2022. Hansard.

Guilherme Penedo, Hynek Kydlíek, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Preprint*, arXiv:2406.17557.

Irene Plaza, Nina Melero, Cristina del Pozo, Javier Conde, Pedro Reviriego, Marina Mayor-Rocher, and María Grandury. 2024. Spanish and llm benchmarks: is mmlu lost in translation?

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Martin Puttkammer and Justin Hocking. 2021. Autshumato multilingual word and phrase translations. Word and phrase lists aligned from English to the other official South African languages.

Martin Puttkammer, Martin Schlemmer, and Ruan Bekker. 2021. Nchlt setswana annotated text corpora. Lemmatized, part of speech tagged, and morphologically analyzed corpora developed during the NCHLT Text project.

Malebogo Rahlao, Nina Lewin, and Taariq Surtee. 2021. New uses for old books: Description of digitised corpora-based on the setswana language collection in the wits cullen africana collection. In *Proceedings of the International Conference of the Digital Humanities Association of Southern Africa (DHASA)*, Johannesburg, South Africa. The University of the Witwatersrand, Johannesburg.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20. IEEE Press.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 35053506, New York, NY, USA. Association for Computing Machinery.

SADiLaR. 2024. South african centre for digital language resources (sadilar).

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, Armand Joulin, and Angela Fan. 2021. CCMatrix: Mining billions of high-quality parallel sentences on the web. In *Proceedings of the*

59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 6490–6500, Online. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiski, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

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. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research.

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

Tatoeba. 2024. Tatoeba.

Teknium. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

The Parliament of Zimbabwe. 2013. Constitution of zimbabwe. https://parlzim.gov.zw/download/constitution-of-zimbabwe-amendment-no-20-24-05-2013/. Accessed: 2024-07-30.

The Republic of South Africa. 1996. The south african constitution. https://www.justice.gov.za/constitution/SAConstitution-web-eng.pdf. Accessed: 2024-07-30.

Thutong South African Education Portal. 2024. Thutong south african education portal. https://www.thutong.doe.gov.za/.

Jörg Tiedemann. 2012a. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012b. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Tlhalefang. 2009. Tlhalefang communications.

Atnafu Lambebo Tonja, Bonaventure FP Dossou, Jessica Ojo, Jenalea Rajab, Fadel Thior, Eric Peter Wairagala, Aremu Anuoluwapo, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, et al. 2024. Inkubalm: A small language model for low-resource african languages. *arXiv preprint arXiv:2408.17024*.

TRK. 2021. Setswana mo botswana.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment. *Preprint*, arXiv:2310.16944.

Unisa. 2023. Learn to speak an african language.

Tobie van Dyk. 2021. Generic multilingual academic wordlists with definitions. The resource contains 2,427 terms with part of speech and usage examples. Developed for students to assist with vocabulary building and decoding academic texts.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

Tiannan Wang, Jiamin Chen, Qingrui Jia, Shuai Wang, Ruoyu Fang, Huilin Wang, Zhaowei Gao, Chunzhao Xie, Chuou Xu, Jihong Dai, Yibin Liu, Jialong Wu, Shengwei Ding, Long Li, Zhiwei Huang, Xinle Deng, Teng Yu, Gangan Ma, Han Xiao, Zixin Chen, Danjun Xiang, Yunxia Wang, Yuanyuan Zhu, Yi Xiao, Jing Wang, Yiru Wang, Siran Ding, Jiayang Huang, Jiayi Xu, Yilihamu Tayier, Zhenyu Hu, Yuan Gao, Chengfeng Zheng, Yueshu Ye, Yihang Li, Lei Wan, Xinyue Jiang, Yujie Wang, Siyu Cheng, Zhule Song, Xiangru Tang, Xiaohua Xu, Ningyu Zhang, Huajun Chen, Yuchen Eleanor Jiang, and Wangchunshu Zhou. 2024. Weaver: Foundation models for creative writing.

Bin Xiao, Haiping Wu, Weijian Xu, Xiyang Dai, Houdong Hu, Yumao Lu, Michael Zeng, Ce Liu, and Lu Yuan. 2023. Florence-2: Advancing a unified representation for a variety of vision tasks.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2023. Lego-MT: Learning detachable models for massively multilingual machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11518–11533, Toronto, Canada. Association for Computational Linguistics.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild.

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

A Hyperparameters

Training hyperparameters were selected using a preliminary version of the Marothodi and Medupe datasets. The learning rate and embedding learning rate, which is used for the language modeling head and embedding layers, were both ablated over. The pair of hyperparameters that yielded the greatest MMLU-tsn accuracy when training a preliminary Pula 8B was selected. These hyperparameters were used for training Pula 1B, 3B, and 8B. When training Pula 14B we reduce the maximum trained sequence length, LoRA rank, LoRA alpha, and gradient accumulation steps due to memory restraints.

All training hyperparameters used, including the effective batch size, are listed in Table 3.

Parameter	Pula 1B	Pula 3B	Pula 8B	Pula 14B
GPUs	8	8	8	8
Max Seq Length	4096 tokens	4096 tokens	4096 tokens	2048 tokens
LoRA Alpha	32	32	32	16
LoRA Dropout	0.2	0.2	0.2	0.2
LoRA Rank	64	64	64	32
Bias	None	None	None	None
Precision	bf16	bf16	bf16	bf16
Optimizer	AdamW 8bit	AdamW 8bit	AdamW 8bit	AdamW 8bit
Weight Decay	0.0	0.0	0.0	0.0
Warmup Ratio	0.05	0.05	0.05	0.05
Learning Rate	2e-05	2e-05	2e-05	2e-05
Embed Learning Rate	8e-06	8e-06	8e-06	8e-06
LR Scheduler	Cosine	Cosine	Cosine	Cosine
Epochs	3.0	3.0	3.0	3.0
Packing	✓	✓	✓	✓
Per-Device Batch Size	1	1	1	1
Gradient Accumulation Steps	8	8	8	1
Effective Batch Size	64 (260k tokens)	64 (260k tokens)	64 (260k tokens)	8 (16k tokens)

Table 3: Training Hyperparameters