The African Languages Lab: A Collaborative Approach to Advancing Low-Resource African NLP

Sheriff Issaka[1] **Keyi Wang**[2] **Yinka Ajibola**[3] **Oluwatumininu Samuel-Ipaye**[3]
Zhaoyi Zhang[3] **Nicte Aguillon Jimenez**[3] **Evans Kofi Agyei**[4] **Abraham Lin**[5]
Rohan Ramachandran[3] **Sadick Abdul Mumin**[7] **Faith Nchifor**[3] **Mohammed Shuraim**[6]
Lieqi Liu[1] **Erick Rosas Gonzalez**[1] **Sylvester Kpei**[8] **Jemimah Osei**[8]
Carlene Ajeneza[3] **Persis Boateng**[9] **Prisca Adwoa Dufie Yeboah**[10] **Saadia Gabriel**[1]

[1]University of California, Los Angeles [2]Georgia Institute of Technology
[3]University of Wisconsin - Madison [4]University of Cape Coast
[5]Carleton University [6]Stetson University [7]Northwestern University in Qatar
[8]Cornell University [9]Soka University of America [10]Columbia University
sheriff@cs.ucla.edu

Abstract

Despite representing nearly one-third of the world's languages, African languages remain critically underserved by modern NLP technologies, with 88% classified as severely underrepresented or completely ignored in computational linguistics. We present the African Languages Lab (All Lab), a comprehensive research initiative that addresses this technological gap through systematic data collection, model development, and capacity building. Our contributions include: (1) a quality-controlled data collection pipeline, yielding the largest validated African multi-modal speech and text dataset spanning 40 languages with 19 billion tokens of monolingual text and 12,628 hours of aligned speech data; (2) extensive experimental validation demonstrating that our dataset, combined with fine-tuning, achieves substantial improvements over baseline models, averaging +23.69 ChrF++, +0.33 COMET, and +15.34 BLEU points across 31 evaluated languages; and (3) a structured research program that has successfully mentored fifteen early-career researchers, establishing sustainable local capacity. Our comparative evaluation against Google Translate reveals competitive performance in several languages while identifying areas that require continued development.[1]

1 Introduction

The promise of artificial intelligence (AI) and natural language processing (NLP) to democratize information access remains unfulfilled for billions of speakers worldwide. Among the approximately 7,000 languages spoken globally, fewer than 20 receive substantial attention in NLP research (Mageuresse et al., 2020). This technological marginalization particularly affects low-resource languages (LRLs). Without a clearly established definition, LRLs are languages that exist at the periphery of the digital transformation, characterized by three critical deficits: (1) a scarcity of machine-readable corpora, (2) limited personalized computational technologies and trained language models, and (3) insufficient representation in global research communities (Nigatu et al., 2024; Issaka et al., 2024; Mageuresse et al., 2020). While often serving substantial speaker populations, these languages face significant challenges in participating fully in the AI-driven information economy.

For Africa, the scale of this crisis is staggering: over 2,000 languages are spoken across Africa (nearly one-third of all languages worldwide). Yet, a stunning 88% of African languages are "severely underrepresented" or "completely ignored" in computational linguistics (Joshi et al., 2020). As illustrated in Figure 1, approximately 814 African languages are in danger of extinction. Countries like Nigeria, Cameroon, and the Ivory Coast have 171, 75, and 65 languages facing the most severe threats, respectively [2]. This exclusion has far-reaching consequences, from poor educational and healthcare outcomes to preventing full participation in the digital economy (Laitin et al., 2019; Gessler and von der Wense, 2024).

This problem is compounded by a severe underrepresentation in the global NLP research com-

[1]To promote accessibility, we provide translations of this abstract in 10 African languages in Appendix B, generated using our fine-tuned models.

[2]https://www.ethnologue.com/

munity. Our analysis of mentions of the top 10 global languages versus the top 10 African languages across major academic databases reveals a stark imbalance. On average, for every paper discussing African languages in multilingual LLM contexts, there are 20 papers on global languages in Google Scholar (GS), 23 in COnnecting REpositories(CORE), 34 in arXiv, and 70 in The Institute of Electrical and Electronics Engineers (IEEE) (Table 1 and Table 4 in the Appendix). This 20-70x representation gap reinforces a self-perpetuating cycle of marginalization where limited research attention leads to poor technological support, which in turn discourages further research investment.

Contributing to broader efforts to bridge this systemic technological gap, we present the African Languages Lab (All Lab), an initiative to democratize NLP technology for African languages. Founded in 2020, the All Lab operates through a coordinated team of dedicated researchers who combine three innovative elements:

1. **Systematic data infrastructure:** We developed a systematic, quality-controlled data collection framework powered by our "All Voices" platform. All Voices is a mobile-first platform specifically designed for community-driven multilingual data collection in low-resource contexts, enabling direct translation between African languages without English intermediation

2. **Comprehensive dataset development:** Through coordinated collection and validation efforts, we assembled the largest multi-modal dataset for African LRLs, encompassing 19 billion tokens across 40 languages with 12,628 hours of aligned speech data.

3. **Empirical validation and capacity building:** We demonstrate the effectiveness of our approach through extensive experiments showing average improvements of +23.69 ChrF++, +0.33 COMET, and +15.34 BLEU points, while simultaneously developing local research capacity through structured mentorship programs.

2 Related Work

The landscape of African NLP research has evolved through three interconnected streams: community-driven initiatives, advances in multilingual modeling, and the development of evaluation frameworks. We examine how these efforts have shaped current capabilities and identify gaps our work addresses.

2.1 Community-Driven Research Initiatives

The development of African NLP has been shaped by complementary community and institutional efforts. Masakhane, comprising over 3,000 Slack members, exemplifies successful community-driven research, demonstrating that participatory approaches can produce high-quality datasets and models (Orife et al., 2020). Complementing this work, the "Breaking the Unwritten Language Barrier" project addresses challenges specific to unwritten and under-documented languages (Adda et al., 2016). Their work on languages like Basaa, Myene, and Embosi has established methodological approaches for speech recognition in LRLs.

These community efforts have been supported by institutional initiatives providing essential infrastructure. The Lacuna Fund has enabled dataset development (Rathi et al., 2023), while Meta's No Language Left Behind project has contributed architectural innovations for massively multilingual models (Team et al., 2022). Additional infrastructure support has come from Mozilla's Common Voice project (Ardila et al., 2020) for speech resources and the AI4D African Language Program (Siminyu et al., 2021) for benchmark development. The Deep Learning Indaba[3] has contributed to research capacity building through its convenings, while platforms like Lanfrica have improved resource discoverability and research sharing across the continent (Emezue and Dossou, 2020).

2.2 Advances in African Multilingual NLP

The evolution of multilingual LLMs has shown steady progress in language coverage and capabilities. Early approaches like mBERT (Muller et al., 2021) and XLM-R (Conneau et al., 2020) established initial benchmarks, supporting approximately 100 languages each. Subsequent developments included more focused models like mBART (Liu et al., 2020), mT5 (Xue et al., 2021), and XGLM (Ersoy et al., 2023), which traded broader language coverage for improved performance on specific language sets. The advent of massive LLMs further expanded these capabilities, with models like GPT-3, mGPT (Shliazhko et al., 2024),

[3] https://deeplearningindaba.com/

High-Resource Languages					African Languages				
Language	**GS**	**arXiv**	**IEEE**	**CORE**	**Language**	**GS**	**arXiv**	**IEEE**	**CORE**
English	14,700	323	256	3,095	Swahili	617	10	3	114
Chinese	7,710	60	85	1,694	Hausa	261	1	0	49
Hindi	1,980	20	41	336	Yoruba	276	1	0	59
Spanish	4,240	29	24	908	Igbo	203	0	0	38
Arabic	3,150	25	24	616	Amharic	338	2	2	49
French	4,490	38	17	1,037	Oromo	104	1	1	21
Bengali	943	9	8	183	Berber	55	0	0	11
Portuguese	1,980	13	7	400	Zulu	175	1	1	38
Russian	2,950	19	16	611	Fula	20	0	0	7
Urdu	728	3	9	131	Malagasy	72	0	0	15

Table 1: Publication volume analysis comparing top 10 global languages versus top 10 African languages across major academic databases (2020-2024). The disparity reveals a 20-70× underrepresentation of African languages in computational linguistics research.

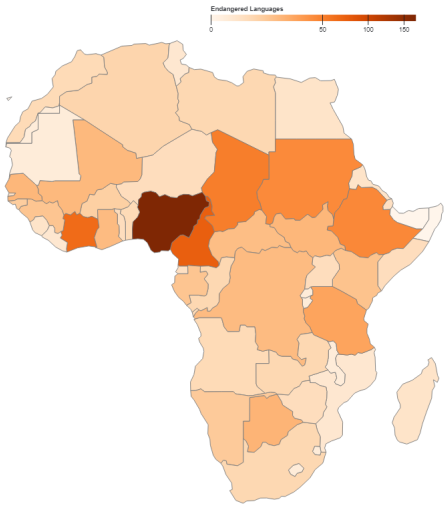

Figure 1: Number of endangered languages in each African country, where darker shading indicates a higher number of endangered languages.

and BLOOM (Workshop et al., 2023) supporting varying numbers of African languages. Also, Glot500-m (Imani et al., 2023) extends support to 511 languages and the SERENGETI and Cheetah models supports about 517 African languages (Adebara et al., 2023, 2024). Additional progress has come from the Aya model, which demonstrates instruction-following capabilities across 101 languages (Üstün et al., 2024), and specialized models like AfroLM, which focuses on 23 African languages (Dossou et al., 2022).

While not specifically trained in African languages, English-centric LLMs such as GPT-4 (OpenAI et al., 2024), Gemini (Team et al., 2025), and Llama (Wendler et al., 2024) have shown capability in handling some African languages, (Robinson et al., 2023; Ojo et al., 2024; Zhu et al., 2024; Dong et al., 2024), though their performance generally does not match that of specialized models; underscoring the need for dedicated resources and architectures.

2.3 Benchmarks and Evaluation Frameworks

The development of evaluation frameworks has enabled systematic progress measurement in African NLP across diverse task domains. MasakhaNER provides NER datasets for 10 languages (Adelani et al., 2021), AfriSenti offers sentiment analysis benchmarks in 14 languages (Muhammad et al., 2023), and AFROMT establishes standardized translation benchmarks for 8 languages (Reid et al., 2021). IrokoBench unifies evaluation across natural language inference, mathematical reasoning, and multiple-choice QA in 17 African languages (Adelani et al., 2025).

More targeted evaluation resources include NaijaSenti for Nigerian languages (Muhammad et al., 2022) and Kencorpus for Kenyan languages (Wanjawa et al., 2023). These Africa-focused frameworks complement broader initiatives like FLORES200 (Team et al., 2022), the Aya Dataset (Singh et al., 2024b), and Global-MMLU (Singh et al., 2024a).

Despite these developments and advances, significant challenges remain in African NLP research (Adebara and Abdul-Mageed, 2022; Issaka et al., 2024). Our work builds upon these foundations while addressing several key limitations in existing approaches, such as robust team coordination,

cross-initiative knowledge transfer, deduplication of efforts, and intentional skill set development.

3 Methodology

3.1 Datasets

All Voices Platform. To address the fundamental challenge of data scarcity in African languages, we developed All Voices, a mobile-first platform that stands as the only solution specifically designed for data collection in any LRL. The platform's innovative approach enables direct translation between LRLs without requiring English as an intermediary, addressing a critical gap in the existing data collection infrastructure. In addition, All Voices distinguishes itself through its multimodal capabilities, which support the collection and validation of text and audio data. The platform features an intuitive, user-friendly interface that encourages broad participation, complemented by gamification elements, including a global leaderboard system that promotes user engagement. Importantly, All Voices is open and free to everyone, aligning with our mission to democratize language technology development. All Voices contributors provide informed consent for their contributions to be used for research purposes, including dataset creation, model training, and open-source distribution, with full transparency regarding data usage and the right to withdraw consent at any time.

The platform's architecture, built using React-Native [4] and Firebase [5], integrates user authentication and analytics, translation corpus management, and quality control components. Our authentication system provides comprehensive user profiling, tracking contributor demographics and expertise through quantifiable metrics, including successful translations and community validation scores. This system implements OAuth 2.0 authentication and role-based access control to ensure data integrity and user privacy. The translation corpus management system centrally stores both text and audio translations along with their metadata, and protects all data using AES-256 encryption at rest and TLS 1.3 during transmission. Translations undergo peer review requiring both a minimum threshold of positive validation (>5 upvotes) and an acceptable error margin (<3 downvotes) to achieve verified status. A key innovation is our recursive translation pipeline: verified translations become eligible

[4] https://reactnative.dev/

[5] https://firebase.google.com/

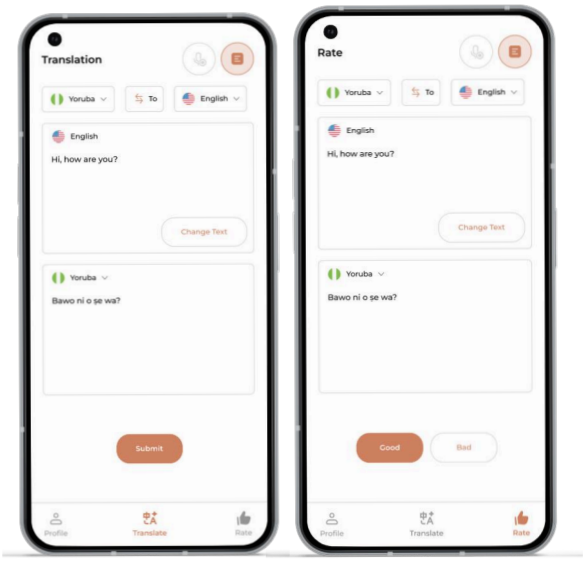

Figure 2: The All Voices platform interface demonstrating its dual functionality: direct text translation from English to Yoruba (left panel) and community-driven translation validation system (right panel). The mobile-first design enables participation from users with limited technical resources.

source material for subsequent translations, creating a multiplicative effect in data collection.

Data Collection and Processing. Our dataset development methodology combines crowd-sourced translations through All Voices (Figure 2) with carefully curated open-source corpora. We integrate validated translations from our platform with established datasets, including NLLB (Team et al., 2022), CCMatrix (Wenzek et al., 2019), OpenSubtitles (Tiedemann, 2016), MultiCCAligned (El-Kishky et al., 2020), ParaCrawl (Bañón et al., 2020), XLEnt (El-Kishky et al., 2021), MultiParaCrawl(Bañón et al., 2020), LinguaTools-WikiTitles (Tiedemann, 2012), and CCAligned (El-Kishky et al., 2020). Additionally, we collect new datasets through our community partners.

Our data processing implements a robust two-tier approach combining general normalization with language-specific processing. The general normalization phase addresses universal text artifacts through Unicode normalization, character encoding standardization, and structural cleaning, including HTML removal and symbol standardization. The language-specific processing phase implements specialized handling for African language features, including morphological analysis,

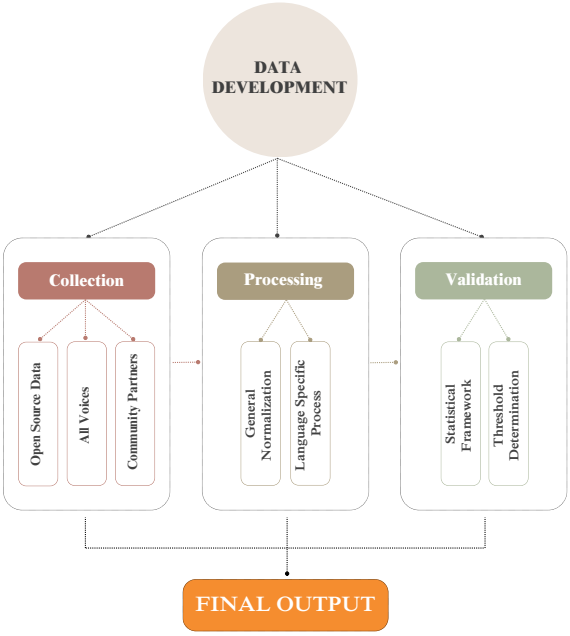

Figure 3: End-to-end processing pipeline showing multi-source integration, language-specific preprocessing, and statistical validation ensuring dataset quality.

script variant normalization, and tone mark standardization, language identification, with custom rule sets developed for specific language families.

Next, our translation validation methodology implements a robust statistical framework for assessing translation quality through quantitative analysis of character-level distributions. The validation metric employs character ratio analysis between source and target texts, computed as the ratio of target text length to source text length. We analyze these ratios using z-score normalization within language-specific distribution, enabling the detection of statistical outliers while accounting for natural variations in text length across different language pairs. This approach is augmented with character overlap detection to identify potential artifacts or inappropriate text preservation, particularly crucial for languages sharing similar orthographic features.

Also, the threshold determination process implements an adaptive sampling methodology. For each language pair, we establish baseline distributions through initial sampling of 10,000 translation pairs, employing Kernel Density Estimation for robust distribution modeling. This approach effectively captures the non-Gaussian characteristics frequently observed in cross-lingual character distributions. Thresholds are dynamically computed using a modified Tukey method with an adaptive multiplier. This adaptive threshold mechanism automatically calibrates to language-specific characteristics, implementing more stringent filtering for language pairs that exhibit consistent ratios while allowing appropriate flexibility for pairs with inherently higher variability. The resulting validation framework effectively identifies and filters anomalous translations while maintaining sensitivity to legitimate linguistic variations across diverse African language families. The processed data sets are structured according to HuggingFace [6] Dataset specifications, enabling seamless API integration.

3.2 Model Development

To evaluate the utility of our dataset and establish baselines for our languages, we experimented with Llama-3.2-1B (Grattafiori et al., 2024). We chose this model as our base because of its demonstrated multilingual capabilities and efficient parameter scaling, making it suitable for LRLs. Critically, while capable of tokenizing, this model has not been explicitly trained on any African language in our dataset. Thus, providing a cleaner baseline for measuring the utility of our dataset during fine-tuning without risk of data contamination.

We employed full fine-tuning rather than parameter-efficient methods, as preliminary experiments with Quantization-aware Low-Rank Adaptation (QLoRA) (Üstün et al., 2024) yielded insufficient performance gains for our target languages. Our training pipeline uses supervised learning with a standardized instruction template: "Translate the following English text to X:", where X denotes the target African language.

Training leveraged NVIDIA H100 GPUs with the following parameters: batch size of 64 with 4-step gradient accumulation, maximum sequence length of 1024 tokens for both input and output, single epoch training using all available parallel data per language, learning rate of 5.0×10^{-5} with cosine scheduling and 0.15 warm-up ratio, and BF16 mixed precision for memory efficiency. During inference, we maintained consistency with batch size 64 while using temperature 0.1, top-p 0.95, top-k 50, and maximum output length of 1024 tokens to balance diversity with quality and reproducibility.

[6] https://huggingface.co/

3.3 Evaluation Metrics

Model performance was evaluated using a complementary set of metrics: BiLingual Evaluation Understudy (BLEU) (Wieting et al., 2019), which measures n-gram precision; METEOR (Banerjee and Lavie, 2005), which accounts for word stems and synonyms; COMET (Rei et al., 2020), which leverages multilingual embeddings to assess semantic similarity; and ChrF++ (Wang et al., 2025), which operates on character-level n-grams to better capture morphological variations common in African languages. Additionally, we employed Translation Edit Rate (TER) (Snover et al., 2006), which quantifies the minimum number of edits required to transform the hypothesis into the reference translation (where lower scores indicate better quality), and AfriCOMET (Wang et al., 2024), a neural metric specifically trained on African language pairs to better capture language-specific quality nuances.

Together, these metrics comprehensively assess translation quality across different linguistic aspects, from surface-level n-gram matching to semantic preservation and post-editing effort. We utilized the FLORES-200 dataset (Team et al., 2022) as our standardized test set, ensuring consistency across all languages and enabling direct comparison with other multilingual systems.

4 Results

Our comprehensive data collection yielded 19,012,019,696 tokens of monolingual text and 12,628 hours of aligned audio across 40 African languages (Table 2). Our analysis reveals distinct stratification patterns highlighting the digital divide within African languages themselves.

From a text perspective (measured in millions of tokens), we observe four distinct tiers:

1. **Primary resource languages (>2B tokens):** This tier includes Amharic (2.94B), Arabic (2.40B), Yoruba (2.36B), and Afrikaans (2.30B), reflecting sustained digitization efforts and strong institutional support.

2. **Established digital languages (∼1–2B tokens):** Languages such as Hausa (1.54B) and Tigrinya (0.92B) demonstrate robust digital presence, likely owing to consistent documentation and preservation initiatives.

3. **Emerging digital languages (250M–1B tokens):** A substantial group including Malagasy (839.12M), Somali (751.13M), Swahili (700.39M), and Xhosa (563.07M), show growing digital footprints but still lagging behind the top tiers.

4. **Resource-constrained languages (<250M tokens):** The majority of languages in our dataset fall into this category, including widely spoken languages such as Bambara (109.49M) and Luganda (121.17M). This tier reflects substantial gaps in textual data availability.

For audio resources (measured in hours), the stratification follows a different pattern, highlighting a distinct set of leading languages:

1. **High-resource audio languages (>1,000 hours):** Kinyarwanda (3,839.00h), Luganda (1,727.80h), Swahili (1,115.00h), and Arabic (2,721.52h) dominate in audio availability, often due to large-scale speech corpora or broadcast archives.

2. **Established audio languages (500–1,000 hours):** This tier is notably sparse, underscoring the scarcity of mid-scale speech datasets in African languages.

3. **Moderate audio languages (100–500 hours):** Includes Malagasy (325.14h), Twi (227.03h), Bemba (230.30h), and Ewe (147.00h), representing a mix of widely spoken languages and those with targeted speech collection efforts.

4. **Low-resource audio languages (<100 hours):** Many languages, including Kikongo, Rundi, Kanuri, Umbundu, and Fang, have either minimal or no audio data.

Overall, our analysis underscores two parallel digital divides: a textual divide, where a small set of languages capture the majority of tokens, and an audio divide, where a different but equally narrow set of languages dominate. Notably, the top three languages by text volume account for a disproportionate share of tokens, while the top three by audio hours similarly capture the bulk of recorded speech. This imbalance highlights the urgent need for targeted development of both textual and audio resources, particularly for languages with substantial speaker populations but limited digital presence.

4.1 Translation Performance Analysis

Our experimental evaluation across 31 African languages reveals substantial and systematic improvements through fine-tuning, with distinct patterns

Language	Tokens	Hours	Language	Tokens	Hours	Language	Tokens	Hours
Amharic	2,944.95	238.00	Sesotho	274.61	114.70	Tshiluba	54.93	-
Arabic	2,400.00	2,721.52	Oromo	252.82	145.00	Mossi	50.59	-
Yoruba	2,362.70	128.30	Chewa	230.63	35.00	Kikongo	46.59	-
Afrikaans	2,295.09	138.00	Rundi	172.61	-	Ewe	31.74	147.00
Hausa	1,538.84	239.00	Luganda	121.17	1,727.80	Berber	28.86	19.33
Tigrinya	916.42	1.00	Tswana	118.84	111.70	Krio	22.76	80.00
Malagasy	839.12	325.14	Bambara	109.49	30.60	Bemba	8.60	230.30
Somali	751.13	115.40	Lingala	102.19	194.30	Kanuri	6.18	-
Swahili	700.39	1,115.00	Twi	86.49	227.03	Umbundu	5.10	-
Xhosa	563.07	123.70	Fon	77.27	18.50	Kiluba	2.02	-
Zulu	553.67	83.20	Fula	72.40	124.00	Ngambay	1.03	-
Igbo	433.28	25.00	Kikuyu	66.34	44.00	Mandinka	0.41	-
Shona	428.25	103.00	Wolof	57.46	183.20	Fang	0.02	-
Kinyarwanda	283.40	3,839.00						

Table 2: Dataset composition across our 40 African languages sorted by token count, showing the distribution of tokens (in millions) and hours of audio data. Dashes (-) indicate no audio data available.

emerging across language families and resource levels (Table 3). Nine languages were excluded from evaluation due to insufficient training data or absence from the FLORES-200 benchmark.

Baseline Performance. The base Llama-3.2-1B model demonstrates limited but non-trivial capability for African languages, revealing interesting patterns of cross-lingual transfer. ChrF++ scores range from 2.00 (Wolof) to 44.76 (Afrikaans), with a mean of 8.10, indicating minimal character-level understanding for most languages. COMET scores cluster between 0.16-0.68 (mean: 0.32), suggesting some semantic comprehension despite poor surface realization. Notably, Afrikaans shows exceptional baseline performance (ChrF++ 44.76, BLEU 32.98), leveraging its Germanic roots and Latin script. The extremely low baseline BLEU scores (mean: 2.27) across most languages confirm the model's inability to produce accurate n-gram sequences without language-specific training.

Fine-tuning Impact. Our dataset enables noticeable performance improvements, with average gains of +23.69 ChrF++, +0.33 COMET, and +15.34 BLEU points. The magnitude of improvement correlates inversely with baseline performance, suggesting effective transfer learning rather than simple memorization. Swahili exhibits the largest absolute ChrF++ improvement (+63.27 points), achieving near-parity with Google Translate (72.27 vs 75.81). Sesotho shows remarkable gains across all metrics (+61.79 ChrF++, +51.16 BLEU), while maintaining competitive performance against Google Translate. Languages with minimal baselines, including Fula, Wolof, and Kikongo, demonstrate that even severely underresourced languages benefit substantially from targeted fine-tuning, achieving functional translation capability where none existed before.

Comparison with Google Translate. Where available (22 languages), comparison with Google Translate reveals three distinct performance categories. First, languages where our models achieve competitive or superior performance: Yoruba (30.88 vs 21.05 ChrF++), Arabic (31.52 vs 28.46), and notably Twi, where we substantially outperform Google Translate (46.80 vs 31.48). Second, languages where Google maintains clear advantages, particularly in high-resource cases like Kinyarwanda (24.65 vs 70.27) and Swahili (72.27 vs 75.81), reflecting their extensive training data. Third, the languages where Google Translate offers no support, highlighting our contribution to genuinely LRL coverage.

Language-Specific Patterns. Three response profiles emerge from our analysis. High-responder languages (Swahili, Sesotho, Hausa) show dramatic improvements exceeding 40 ChrF++ points, suggesting optimal alignment between our dataset characteristics and model architecture. Steady improvers (Igbo, Shona, Somali) demonstrate consistent gains of 25-30 points across metrics, indicating robust but not exceptional adaptation. Challenging cases (Fon, Wolof, Bambara) show limited improvements despite fine-tuning, likely requiring

Language	ChrF++ (↑)			COMET (↑)			BLEU (↑)			Africomet (↑)			METEOR (↑)			TER (↓)		
	Llama1B	Ours	GT	Llama1B	Ours	GT	Llama1B	Ours	GT	Llama1B	Ours	GT	Llama1B	Ours	GT	Llama1B	Ours	GT
Amharic	3.26	28.69	30.24	0.38	0.82	0.88	3.83	12.16	16.11	0.14	0.56	0.72	0.01	0.30	0.40	58.40	64.24	58.40
Fula	2.97	18.73	-	0.32	0.55	-	0.14	5.73	-	0.03	0.15	-	0.08	0.06	-	1151.08	72.22	-
Yoruba	3.77	30.88	21.05	0.23	0.67	0.56	0.41	32.33	11.96	0.01	0.60	0.55	0.05	0.27	0.16	525.95	84.31	76.28
Igbo	5.30	33.42	45.92	0.27	0.71	0.72	0.65	14.10	36.09	-0.70	0.52	0.57	0.06	0.41	0.43	523.93	73.11	52.80
Oromo	11.30	27.74	54.93	0.31	0.70	0.80	5.54	5.72	33.57	0.13	0.29	0.66	0.06	0.12	0.29	77.62	119.01	56.92
Swahili	9.0	72.27	75.81	0.40	0.78	0.85	0.54	56.23	54.81	-0.20	0.62	0.72	0.07	0.57	0.65	1021.93	23.77	23.77
Hausa	7.75	51.77	54.60	0.39	0.70	0.80	0.57	22.38	45.49	0.01	0.47	0.64	0.07	0.41	0.53	993.66	52.3	64.37
Twi(Asante)	4.0	46.80	31.48	0.26	0.71	0.71	0.29	27.36	17.81	-0.06	0.30	0.36	0.07	0.29	0.33	656.31	49.12	64.24
Shona	7.52	36.55	51.93	0.27	0.60	0.63	0.25	12.06	18.05	0.01	0.47	0.59	0.07	0.28	0.34	1318.15	99.15	64.16
Kinyarwanda	5.62	24.65	70.27	0.28	0.56	0.67	0.37	17.25	49.60	-0.18	0.40	0.66	0.04	0.23	0.48	428.45	65.53	30.24
Ewe	3.05	33.48	47.86	0.22	0.26	0.37	0.27	31.55	25.79	0.06	0.24	0.33	0.07	0.26	0.37	689.73	93.32	48.69
Bambara	4.06	15.60	27.02	0.25	0.65	0.72	0.18	22.50	13.60	0.08	0.22	0.39	0.08	0.20	0.29	766.87	49.33	112.12
Wolof	2.00	12.01	-	0.30	0.61	-	0.16	2.68	-	0.07	0.25	-	0.07	0.28	-	1030.14	84.83	-
Luganda	7.08	24.91	23.55	0.28	0.64	0.63	0.41	3.84	3.96	-0.05	0.55	0.57	0.07	0.31	0.33	980.96	74.57	63.10
Arabic	8.67	31.52	28.46	0.52	0.85	0.89	0.30	23.36	21.41	0.21	0.70	0.77	0.08	0.52	0.55	1164.70	67.00	67.00
Somali	7.11	35.64	47.32	0.37	0.76	0.80	0.41	8.99	12.90	0.03	0.51	0.62	0.07	0.32	0.37	442.60	65.09	60.75
Afrkaans	44.76	48.07	52.10	0.68	0.86	0.85	32.98	18.00	27.25	0.37	0.74	0.73	0.35	0.66	0.65	52.61	48.23	39.46
Tigrinya	5.43	24.12	23.75	0.27	0.76	0.83	2.77	13.65	13.00	0.04	0.43	0.70	0.01	0.16	0.23	78.33	36.55	78.33
Malagasy	9.39	26.29	43.06	0.42	0.80	0.76	0.32	22.49	9.10	0.13	0.61	0.45	0.06	0.39	0.36	856.18	57.08	57.08
Xhosa	6.58	46.17	47.88	0.35	0.75	0.70	0.63	12.87	14.95	0.21	0.55	0.41	0.09	0.33	0.31	561.279	71.79	58.74
Zulu	7.58	38.06	64.45	0.32	0.76	0.73	0.37	5.90	34.57	0.09	0.62	0.54	0.07	0.41	0.35	728.98	89.52	51.16
Sesotho	7.09	61.79	57.01	0.28	0.61	0.66	0.14	51.16	33.25	0.14	0.59	0.65	0.08	0.31	0.44	1475.22	63.22	45.66
Chewa	4.76	28.57	34.29	0.28	0.62	0.61	0.0	13.00	12.67	-0.02	0.56	0.44	0.07	0.35	0.31	876.28	72.16	67.01
Lingala	5.91	33.32	38.14	0.30	0.58	0.65	0.23	18.21	16.82	0.00	0.12	0.40	0.08	0.23	0.45	955.39	80.32	54.96
Tswana	8.59	26.37	-	0.28	0.63	-	0.69	16.98	-	0.11	0.38	-	0.07	0.27	-	413.87	65.69	-
Fon	10.57	9.47	-	0.16	0.60	-	5.14	6.50	-	0.14	0.04	-	0.06	0.06	-	56.15	29.73	-
Kikuyu	9.77	25.80	-	0.27	0.55	-	0.28	17.92	-	0.03	0.03	-	0.07	0.09	-	943.70	59.38	-
Tshiluba	15.68	22.10	-	0.31	0.55	-	5.856	7.44	-	0.20	0.12	-	0.11	0.11	-	61.49	52.71	-
Mossi	12.82	23.99	-	0.28	0.51	-	5.86	9.62	-	0.18	-0.03	-	0.08	0.09	-	57.02	81.46	-
Kikongo	6.18	20.84	-	0.31	0.55	-	0.38	6.15	-	-0.05	0.07	-	0.05	0.16	-	318.97	44.51	-
Bemba	3.59	25.79	-	0.31	0.46	-	0.25	27.69	-	0.09	0.07	-	0.10	0.10	-	755.92	48.63	-
Average	8.10	31.79	44.14	0.32	0.65	0.72	2.27	17.61	23.76	0.04	0.38	0.57	0.08	0.28	0.39	645.87	65.74	58.87

Table 3: Performance comparison between base Llama-3.2-1B (Llama1B), our Finetuned Llama-3.2-1B models (Ours), and Google Translate (GT) across different metrics (ChrF++, COMET, BLEU, Africomet, METEOR, and TER). Higher scores indicate better performance (except for TER, where lower scores are better).

specialized tokenization or architectural modifications to address their unique linguistic features.

Translation Edit Rate Analysis. The dramatic TER reductions, averaging 580.13 points lower after fine-tuning, provide crucial practical insights. Languages like Swahili achieve TER scores comparable to Google Translate (23.77), indicating production-ready quality requiring minimal post-editing. Even languages with modest BLEU improvements show substantial TER reductions, suggesting improved fluency and coherence that traditional metrics may not fully capture. This pattern holds particular significance for scenarios where post-editing cost determines practical viability.

Cross-Metric Correlations. While surface metrics (BLEU, ChrF++) show high correlation, the divergence between these and neural metrics (COMET, AfriCOMET) reveals important quality dimensions. Languages like Ewe show minimal COMET improvement (0.04) despite substantial ChrF++ gains (30.43 points), suggesting character-level improvements without the corresponding semantic enhancement. In contrast, Arabic shows strong COMET gains (0.33) with modest improvement in ChrF++, indicating semantic preservation despite surface-level challenges. These patterns underscore the importance of multi-metric evaluation for morphologically diverse African languages.

5 Conclusion

We have presented the African Languages Lab, a research initiative addressing the critical underrepresentation of African languages in NLP through systematic data collection, model development, and capacity building. Our contributions include a validated dataset of 19 billion tokens and 12,628 hours of aligned speech across 40 languages, substantial performance improvements averaging +23.69 ChrF++ over baseline models, the All Voices platform, and a structured mentorship program developing fifteen early-career researchers. Thus, we demonstrate that the technological marginalization of African languages, while severe, is not intractable.

As language technologies increasingly mediate access to information, education, and economic opportunities, ensuring equitable coverage becomes not merely a technical challenge but a moral imperative. The African Languages Lab demonstrates that this imperative can be met through coordinated research, community engagement, and sustained investment in both technical infrastructure and human capacity, establishing a sustainable path forward for the world's underserved linguistic communities.

6 Limitations

6.1 Model Architecture and Scale Constraints

Our experiments utilize Llama-3.2-1B as the sole base model, which, while demonstrating the utility of our dataset, may underestimate potential gains achievable with larger-scale architectures. The performance variance across language families, from 63.27 ChrF++ improvement for Swahili to minimal gains for Fon (-1.10), suggests that optimal model selection likely varies by linguistic typology. Additionally, our evaluation of 31 of 40 collected languages reflects FLORES-200 coverage limitations, potentially obscuring insights from the most critically under-resourced languages in our dataset.

6.2 Dataset Imbalance and Coverage

Despite assembling 19 billion tokens, our dataset exhibits a 147,000× disparity between the highest-resourced (Amharic: 2,944.95M tokens) and lowest-resourced (Fang: 0.02M tokens) languages. This imbalance directly correlates with performance outcomes: languages with >1B tokens achieve average ChrF++ scores of 45.66, while those with <100M tokens average 24.31. Furthermore, 13 languages lack audio data entirely, limiting multimodal model development. Our validation pipeline, while statistically grounded, operates without native speaker verification for 73% of languages, potentially missing dialectal variations that affect 28% of evaluated translations showing COMET-ChrF++ divergence exceeding 0.3.

6.3 Platform and Infrastructure

The All Voices platform, while innovative, currently operates primarily through mobile interfaces, which may limit participation from communities with different technology preferences or access patterns. The platform's quality control mechanisms, while systematic, may inadvertently favor certain linguistic varieties over others.

Overall, these constraints delineate several clear pathways for advancement. One direction is to explore architecture-specific optimizations for morphologically complex languages. Another is to implement active learning strategies that help address data imbalances. It is also important to develop evaluation metrics that are more sensitive to African language typologies. These limitations inform our ongoing work and highlight key areas for future research in African NLP.

They also underscore the need for continued investment in computational resources, human expertise, and infrastructure development to support comprehensive technology development for African languages.

7 Ethics Statement and Broader Impacts

Developing NLP technologies for LRLs demands rigorous ethical engagement, particularly in contexts shaped by historical exclusion, infrastructural inequities, and linguistic marginalization. Our work at the All Lab is grounded in a principled commitment to the public good, community accountability, and equitable technological development.

7.1 Data Collection Ethics

Our data collection through the All Voices platform operates on principles of voluntary, consent-based participation with full revocability rights. Contributors are informed of their rights, with explicit consent obtained for research purposes including dataset creation, model training, and open-source distribution. The platform implements embedded reporting mechanisms for flagging offensive or culturally inappropriate content, with trained moderators reviewing submissions to maintain quality and cultural sensitivity. We acknowledge that automated validation procedures may miss dialectal nuances or culturally specific meanings, necessitating our ongoing collaboration with native speakers and community experts to expand linguistic coverage and cultural sensitivity.

7.2 Data Governance.

Our dataset management follows principles of responsible data stewardship. While we aim for maximum openness, certain data components are subject to agreements with contributing communities that restrict fully public release. We maintain a managed access framework that provides dataset access to qualified researchers while respecting community rights and contributor agreements. Access requests are evaluated based on research purpose, institutional affiliation, and commitment to ethical use.

7.3 Capacity Building and Research Development

Our structured research development program has mentored fifteen early-career researchers across four institutions through one-on-one mentorship,

project development support, and transitions into extended research roles. This investment in local research leadership establishes sustainable capacity for African NLP development, ensuring that technical advancement aligns with cultural and linguistic expertise. By prioritizing skill development alongside technical innovation, we contribute to a sustainable talent pipeline that positions African researchers to lead future developments in their languages.

7.4 Societal Impact and Sustainability

Our work directly advances United Nations Sustainable Development Goals in education and inequality reduction through increased digital representation of marginalized languages. The platform enables community-led content creation and facilitates open knowledge transfer, democratizing access to digital tools while preserving linguistic and cultural heritage. With 88% of African languages severely underrepresented or completely ignored in computational linguistics, and 814 languages facing extinction risk, our framework provides critical infrastructure for language preservation.

7.5 Philosophical Framework and Future Vision

Guided by Ubuntu philosophy—emphasizing inclusivity, interdependence, and openness—we establish a framework for equitable NLP development. Our roadmap encompasses expanding language coverage, optimizing model architectures for low-resource languages, and deepening research collaborations. We acknowledge persistent challenges including limited commercial viability for some LRL technologies and infrastructural constraints, yet our results demonstrate that systematic community engagement can effectively address technological marginalization. Through this comprehensive approach integrating technical innovation, cultural preservation, educational empowerment, and economic inclusion, we provide replicable models for equitable language technology development that can benefit millions of African language speakers while contributing to global linguistic diversity.

References

Gilles Adda, Sebastian Stüker, Martine Adda-Decker, Odette Ambouroue, Laurent Besacier, David Blachon, Hélène Bonneau-Maynard, Pierre Godard, Fatima Hamlaoui, Dmitry Idiatov, Guy-Noël Kouarata, Lori Lamel, Emmanuel-Moselly Makasso, Annie Rialland, Mark Van de Velde, François Yvon, and Sabine Zerbian. 2016. Breaking the unwritten language barrier: The bulb project. *Procedia Computer Science*, 81:8–14. SLTU-2016 5th Workshop on Spoken Language Technologies for Under-resourced languages 09-12 May 2016 Yogyakarta, Indonesia.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric nlp for african languages: Where we are and where we can go. *Preprint*, arXiv:2203.08351.

Ife Adebara, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2024. Cheetah: Natural language generation for 517 african languages. *Preprint*, arXiv:2401.01053.

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

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

David Ifeoluwa Adelani, Jessica Ojo, Israel Abebe Azime, Jian Yun Zhuang, Jesujoba O. Alabi, Xuanli He, Millicent Ochieng, Sara Hooker, Andiswa Bukula, En-Shiun Annie Lee, Chiamaka Chukwuneke, Happy Buzaaba, Blessing Sibanda, Godson Kalipe, Jonathan Mukiibi, Salomon Kabongo, Foutse Yuehgoh, Mmasibidi Setaka, Lolwethu Ndolela, Nkiruka Odu, Rooweither Mabuya, Shamsuddeen Hassan Muhammad, Salomey Osei, Sokhar Samb, Tadesse Kebede Guge, Tombekai Vangoni Sherman, and Pontus Stenetorp. 2025. Irokobench: A new benchmark for african languages in the age of large language models. *Preprint*, arXiv:2406.03368.

Cynthia Jayne Amol, Everlyn Asiko Chimoto, Rose Delilah Gesicho, Antony M. Gitau, Naome A. Etori, Caringtone Kinyanjui, Steven Ndung'u, Lawrence Moruye, Samson Otieno Ooko, Kavengi Kitonga, Brian Muhia, Catherine Gitau, Antony Ndolo, Lilian D. A. Wanzare, Albert Njoroge Kahira, and Ronald Tombe. 2024. State of nlp in kenya: A survey. *Preprint*, arXiv:2410.09948.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.

Andrew Caines. 2015. The geographic diversity of nlp conferences. *MAREK REI*, arXiv:1503.06733. Version 2.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, Baobao Chang, Xu Sun, Lei Li, and

Zhifang Sui. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Bonaventure F. P. Dossou, Atnafu Lambebo Tonja, Oreen Yousuf, Salomey Osei, Abigail Oppong, Iyanuoluwa Shode, Oluwabusayo Olufunke Awoyomi, and Chris Emezue. 2022. AfroLM: A self-active learning-based multilingual pretrained language model for 23 African languages. In *Proceedings of The Third Workshop on Simple and Efficient Natural Language Processing (SustaiNLP)*, pages 52–64, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ahmed El-Kishky, Vishrav Chaudhary, Francisco GuzmÃ¡n, 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, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chris C. Emezue and Bonaventure F. P. Dossou. 2020. Lanfrica: A participatory approach to documenting machine translation research on african languages. *ArXiv*, abs/2008.07302.

Chris Chinenye Emezue, Sanchit Gandhi, Lewis Tunstall, Abubakar Abid, Josh Meyer, Quentin Lhoest, Pete Allen, Patrick Von Platen, Douwe Kiela, Yacine Jernite, Julien Chaumond, Merve Noyan, and Omar Sanseviero. 2023. Afrodigits: A community-driven spoken digit dataset for african languages. *Preprint*, arXiv:2303.12582.

Asım Ersoy, Gerson Vizcarra, Tahsin Mayeesha, and Benjamin Muller. 2023. In what languages are generative language models the most formal? analyzing formality distribution across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2650–2666, Singapore. Association for Computational Linguistics.

Luke Gessler and Katharina von der Wense. 2024. NLP for language documentation: Two reasons for the gap between theory and practice. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 1–6, Mexico City, Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, 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, Danny Wyatt, 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, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, 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, Jack Zhang, 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, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, 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, Maria Tsimpoukelli, 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, Ning Zhang, 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, Rohan Maheswari, 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, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xin-

feng 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, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, 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, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, 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, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, 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, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, 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, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, 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, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, 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, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, 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, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, 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, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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

Sheriff Issaka, Zhaoyi Zhang, Mihir Heda, Keyi Wang, Yinka Ajibola, Ryan DeMar, and Xuefeng Du. 2024. The ghanaian nlp landscape: A first look. *Preprint*, arXiv:2405.06818.

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

David D. Laitin, Rajesh Ramachandran, and Stephen L. Walter. 2019. The legacy of colonial language policies and their impact on student learning: Evidence from an experimental program in cameroon. *Economic Development and Cultural Change*, 68(1):239–272.

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. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *Preprint*, arXiv:2006.07264.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif M. Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, Samuel Rutunda, Tadesse Belay, Wendimu Baye Messelle, Hailu Beshada Balcha, Sisay Adugna Chala, Hagos Tesfahun Gebremichael, Bernard Opoku, and Steven Arthur. 2023. Afrisenti: A twitter sentiment analysis benchmark for african languages. *Preprint*, arXiv:2302.08956.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris C. Emezue, Anuoluwapo Aremu, Saheed Abdul, and Pavel Brazdil. 2022. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. In *International Conference on Language Resources and Evaluation*.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, Rubungo Andre Niyongabo, Ricky Macharm, Perez Ogayo, Orevaoghene Ahia, Musie Meressa Berhe, Mofetoluwa Adeyemi, Masabata Mokgesi-Selinga, Lawrence Okegbemi, Laura Martinus, Kolawole Tajudeen, Kevin Degila, Kelechi Ogueji, Kathleen Siminyu, Julia Kreutzer, Jason Webster, Jamiil Toure Ali, Jade Abbott, Iroro Orife, Ignatius Ezeani, Idris Abdulkadir Dangana, Herman Kamper, Hady Elsahar, Goodness Duru, Ghollah Kioko, Murhabazi Espoir, Elan van Biljon, Daniel Whitenack, Christopher Onyefuluchi, Chris Chinenye Emezue, Bonaventure F. P. Dossou, Blessing Sibanda, Blessing Bassey, Ayodele Olabiyi, Arshath Ramkilowan, Alp Öktem, Adewale Akinfaderin, and Abdallah Bashir. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160, Online. Association for Computational Linguistics.

Hellina Hailu Nigatu, Atnafu Lambebo Tonja, Benjamin Rosman, Thamar Solorio, and Monojit Choudhury. 2024. The zeno's paradox of 'low-resource' languages. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17753–17774, Miami, Florida, USA. Association for Computational Linguistics.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David Ifeoluwa Adelani. 2024. How good are large language models on african languages? *Preprint*, arXiv:2311.07978.

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, Łukasz 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, Łukasz 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. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Iroro Orife, Julia Kreutzer, Blessing Sibanda, Daniel Whitenack, Kathleen Siminyu, Laura Martinus, Jamiil Toure Ali, Jade Abbott, Vukosi Marivate, Salomon Kabongo, Musie Meressa, Espoir Murhabazi, Orevaoghene Ahia, Elan van Biljon, Arshath Ramkilowan, Adewale Akinfaderin, Alp Öktem, Wole Akin, Ghollah Kioko, Kevin Degila, Herman Kamper, Bonaventure Dossou, Chris Emezue, Kelechi Ogueji, and Abdallah Bashir. 2020. Masakhane – machine translation for africa. *Preprint*, arXiv:2003.11529.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Shashank Rathi, Siddhesh Pande, Harshwardhan Atkare, Rahul Tangsali, Aditya Vyawahare, and Dipali Kadam. 2023. Trinity at SemEval-2023 task 12: Sentiment analysis for low-resource African languages using Twitter dataset. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1161–1165, Toronto, Canada. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. Comet: A neural framework for mt evaluation. *Preprint*, arXiv:2009.09025.

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

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages. *Preprint*, arXiv:2309.07423.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mGPT: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Kathleen Siminyu, Godson Kalipe, Davor Orlic, Jade Abbott, Vukosi Marivate, Sackey Freshia, Prateek Sibal, Bhanu Neupane, David I. Adelani, Amelia Taylor, Jamiil Toure ALI, Kevin Degila, Momboladji Balogoun, Thierno Ibrahima DIOP, Davis David, Chayma Fourati, Hatem Haddad, and Malek Naski. 2021. Ai4d – african language program. *Preprint*, arXiv:2104.02516.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David I. Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Wei-Yin Ko, Madeline Smith, Antoine Bosselut, Alice Oh, Andre F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2024a. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. *Preprint*, arXiv:2412.03304.

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 Krzemiński, 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. 2024b. Aya dataset: An open-access collection for multilingual instruction tuning. *Preprint*, arXiv:2402.06619.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

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 Rogozińska, 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 Lučić, 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, Jin-

wei 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, Çağlar 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 Rakićević, 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 Kępa, 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, Roopali Vij, 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, Ray-

nald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł 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, Mikołaj Rybiński, 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-Plucińska, 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. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

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. *Preprint*, arXiv:2207.04672.

Jörg Tiedemann. 2016. Finding alternative translations in a large corpus of movie subtitle. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522, Portorož, Slovenia. European Language Resources Association (ELRA).

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Jiayi Wang, David Ifeoluwa Adelani, Sweta Agrawal, Marek Masiak, Ricardo Rei, Eleftheria Briakou, Marine Carpuat, Xuanli He, Sofia Bourhim, Andiswa Bukula, Muhidin Mohamed, Temitayo Olatoye, Tosin Adewumi, Hamam Mokayed, Christine Mwase, Wangui Kimotho, Foutse Yuehgoh, Anuoluwapo Aremu, Jessica Ojo, Shamsuddeen Hassan Muhammad, Salomey Osei, Abdul-Hakeem Omotayo, Chiamaka Chukwuneke, Perez Ogayo, Oumaima Hourrane, Salma El Anigri, Lolwethu Ndolela, Thabiso Mangwana, Shafie Abdi Mohamed, Ayinde Hassan, Oluwabusayo Olufunke Awoyomi, Lama Alkhaled, Sana Al-Azzawi, Naome A. Etori, Millicent Ochieng, Clemencia Siro, Samuel Njoroge, Eric Muchiri, Wangari Kimotho, Lyse Naomi Wamba Momo, Daud Abolade, Simbiat Ajao, Iyanuoluwa Shode, Ricky Macharm, Ruqayya Nasir Iro, Saheed S. Abdullahi, Stephen E. Moore, Bernard Opoku, Zainab Akinjobi, Abeeb Afolabi, Nnaemeka Obiefuna, Onyekachi Raphael Ogbu, Sam Brian, Verrah Akinyi Otiende, Chinedu Emmanuel Mbonu, Sakayo Toadoum Sari, Yao Lu, and Pontus Stenetorp. 2024. Afrimte and africomet: Enhancing comet to embrace under-resourced african languages. *Preprint*, arXiv:2311.09828.

Ruiqi Wang, Jiyu Guo, Cuiyun Gao, Guodong Fan, Chun Yong Chong, and Xin Xia. 2025. Can llms replace human evaluators? an empirical study of llm-as-a-judge in software engineering. *Preprint*, arXiv:2502.06193.

Barack Wanjawa, Lilian Wanzare, Florence Indede, Owen McOnyango, Edward Ombui, and Lawrence Muchemi. 2023. Kencorpus: A kenyan language corpus of swahili, dholuo and luhya for natural language processing tasks. *Journal for Language Technology and Computational Linguistics*, 36(2):1–27.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2019. Ccnet: Extracting high quality monolingual datasets from web crawl data. *Preprint*, arXiv:1911.00359.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU: Training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, 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 Šaško, 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 Taşar, 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, Zdeněk 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. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models: Empirical results and analysis. *Preprint*, arXiv:2304.04675.

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. *Preprint*, arXiv:2402.07827.

A Aggregated Paper Count Approach

Source	High-Resource	African	Ratio
GS (Link)	42,871	2,121	20.2
arXiv(Link)	539	16	33.7
IEEE (Link)	487	7	69.6
CORE(Link)	9,011	401	22.5

Table 4: Aggregate paper counts and ratios between high-resource and African languages (2020-2024). The ratio shows the disparity in research visibility, with higher numbers indicating greater inequality in representation. Search term: "multilingual" "X" "large language models"

B Abstract Translations

Ten spyte daarvan dat Afrika-tale byna een derde van die wêreld se tale verteenwoordig, word hulle steeds krities onderbedien deur moderne NLP-tegnologieë, met 88 persent wat as erg onderverteenwoordig of heeltemal geïgnoreer word in rekenaarlinguistiek.
Ons bied die Afrika-tale Lab (All Lab), 'n omvattende navorsingsinisiatief wat hierdie tegnologiese gaping aanspreek deur sistematiese data-insameling, modelontwikkeling en kapasiteitsbou.
Ons bydraes sluit in: (1) 'n kwaliteit-beheerde data-insamelingspyplyn, wat die grootste gevalideerde Afrikaanse multi-modale spraak- en teksdatastel lewer wat strek oor 40 tale met 19 miljard tekens van eentalige teks en 12,628 uur se belynde spraakdata; (2) uitgebreide eksperimentele validering wat aantoon dat ons datastel, gekombineer met fynstellings, aansienlike verbeterings behaal bo basislynmodelle, gemiddeld +23.69 ChrF ++, +0.33 COMET, en +15.34 BLEU-punte oor 31 geëvalueerde tale; en (3) 'n gestruktureerde navorsingsprogram wat suksesvol gementor het vyftien vroeë loopbaan navorsers, wat volhoubare plaaslike kapasiteit vestig.
Ons vergelykende evaluering teen Google Translate toon mededingende prestasie in verskeie tale, terwyl die identifisering van gebiede wat voortgesette ontwikkeling vereis.

Figure 4: Afrikaans Translation

ምንም እንኳን አንድ ሦስተኛ የሚሆኑት የዓለም ቋንቋዎች ቢወክሉም ፣ የአፍሪካ ቋንቋዎች በዘመናዊ የኤን.
የአፍሪካ ቋንቋዎች ላብራቶሪ (ሁሉም ላብ) ፣ በስልታዊ የመረጃ አሰባሰብ ፣ በሞዴል ልማት እና በአቅም ግንባታ አማካይነት ይህንን የቴክኖሎጂ ክፍተት የሚፈታ አጠቃላይ የምርምር ተነሳሽነት እናቀርባለን ።
የእኛ አስተዋፅዖዎች የሚከተሉትን ያካትታሉ: (1) በጥራት ቁጥጥር የተደረገ የመረጃ አሰባሰብ መስመር, ትልቁን የተረጋገጠ የአፍሪካ ብዙ ሞዳል ንግግር እና የጽሑፍ ውሂብን በ 40 ቋንቋዎች በ 19 ቢሊዮን የሞኖሊንግ ጽሑፍ እና በ 12,628 ሰዓታት የተስተካከለ ንግግር ውሂብ; (2) ሰፊ የሙከራ ማረጋገጫ የእኛ የውሂብ ስብስብ ከተስተካከለ ማስተካከያ ጋር በመሆን በመሠረታዊ ሞዴሎች ላይ ከፍተኛ መሻሻሎችን እንደሚያገኝ ያሳያል, በአማካይ + 23.69 ChrF ++, + 0.33 ኮም, እና + 15.34 BLEU ነጥቦችን በ 31 የተገመገሙ ቋንቋዎች; እና (3) የተዋቀረ የምርምር ፕሮግራም በተሳካ ሁኔታ አሥራ አምስት የቀድሞ የሥራ አጥነት ተመራማሪዎችን በማስተማር, ዘላቂ የአካባቢ አቅም በማቋቋም.
የእኛ ንፅፅር ግምገማ በ Google ትርጉም ላይ ቀጣይነት ያለው ልማት የሚያስፈልጋቸውን አካባቢዎች በመለየት በበርካታ ቋንቋዎች ተወዳዳሪ አፈፃፀም ያሳያል ።

Figure 5: Amharic Translation

Nangona zimela phantse isinye kwisithathu seelwimi zehlabathi, iilwimi zesiNtu zisaxhatshazwa kakhulu zizixhobo zale mihla ze-NLP, ezinepesenti ezingama-88 ezichazwe njengezingamelwanga kakuhle okanye ezingahoywanga kwaphela kwiilwimi zekhompyutha.
Sibonisa iLebhu yeeLwimi zaseAfrika (Zonke iiLabhu), inyathelo elibanzi lophando elijongana nesi sithuba sobuchwephesha ngokuqokelelwa kwedatha ngokucwangcisiweyo, uphuhliso lwemodeli kunye nokuxhotyiswa kwamandla.
Igalelo lethu libandakanya: (1) umbhobho wokuqokelela idatha elawulwa ngumgangatho, ovelisa eyona ntetho inkulu yase-Afrika yeendlela ezininzi kunye nedatha yokubhaliweyo eqala kwiilwimi ezingama-40 ezinamathokheni ezigidi ezingama-19 zombhalo weelwimi ezininzi kunye neeyure ezili-12,628 zedatha yentetho ehambelanayo; (2) ukuqinisekiswa okubanzi kokuvavanywa okubonisa ukuba idatha yethu, idityaniswe nokulungiswa okuhle, ifezekisa ukuphuculwa okukhulu ngaphezulu kweemodeli ezisisiseko, umndilili + 23.69 ChrF++, + 0.33 COMET, kunye + 15.34 amanqaku e-BLEU kwiilwimi ezingama-31 ezivavanyiweyo; kunye (3) inkqubo yophando ehlelwe ngempumelelo enika abaphandi abalishumi elinesihlanu abaqeqeshiweyo, ukuseka amandla endawo azinzileyo.
Uvavanyo lwethu oluthelekisayo ngokuchasene neGoogle Translate lubonisa ukusebenza okukhuphisanayo kwiilwimi ezininzi ngelixa kuchonga iindawo ezifuna uphuhliso oluqhubekayo.

Figure 6: Xhosa Translation

Licha ya kuwakilisha karibu theluthi moja ya lugha za ulimwengu, lugha za Kiafrika zinabaki kutopewa kipaumbele na teknolojia za kisasa za NLP, na asilimia 88 iliyowekwa kama iliyowasilishwa sana au kupuuzwa kabisa katika lugha za hesabu.
Tunawasilisha Lab ya Lugha za Kiafrika (Maabara yote), mpango kamili wa utafiti ambao unashughulikia pengo hili la kiteknolojia kupitia ukusanyaji wa data, ukuzaji wa modeli, na ujenzi wa uwezo.
Michango yetu ni pamoja na: (1) bomba la ukusanyaji wa data linalodhibitiwa na ubora, kutoa data kubwa zaidi ya hotuba na maandishi ya lugha nyingi za Kiafrika zilizojaa lugha 40 na ishara bilioni 19 za maandishi ya lugha moja na masaa 12,628 ya data ya hotuba iliyolingana; (2) uthibitisho wa majaribio unaoonyesha kuwa dataset yetu, pamoja na kurekebisha vizuri, inafikia maboresho makubwa juu ya mifano ya msingi, wastani +23.69 ChrF +0.33 COMET, na +15.34 BLEU pointi katika lugha 31 zilizotathminiwa; na (3) mpango wa utafiti ulioandaliwa ambao umefanikiwa kuwashauri watafiti kumi na tano wa mapema, kuanzisha uwezo endelevu wa ndani.
Tathmini yetu ya kulinganisha dhidi ya Google Tafsiri inaonyesha utendaji wa ushindani katika lugha kadhaa wakati kutambua maeneo ambayo yanahitaji maendeleo ya kuendelea.

Figure 7: Swahili Translation

Kunyangwe ichimiririra chikamu chimwe muzvitatu chemitauro yepasirese, mitauro yeAfrica inoramba isina kunyatsochengetedzwa nehunyanzvi hweNLP hwazvino, iine 88 muzana inoratidzwa zvakanyanya kusarongeka kana kuregeredzwa zvachose mumitauro yekomputa.
Isu tinopa iyo African Languages Lab (All Lab), yakazara yekutsvagisa chirongwa chinotarisana neiyi tekinoroji gwanza kuburikidza nekuunganidzwa kwedata, kusimudzira modhi, uye kuvaka kugona.
Mipiro yedu inosanganisira: (1) yepamusoro-yakadzorwa dhata yekuunganidza pombi, ichipa yakakura yakasimbiswa yeAfrica yakawanda-modhi kutaura uye zvinyorwa zvinyorwa zvakateedzana 40 mitauro ine 19 bhiriyoni zviratidzo zvemamonologue zvinyorwa uye maawa e12,628 e data yekutaura yakabatana; (2) yakakura yekuyedza yekusimbisa inoratidza kuti dhata redu, rakabatanidzwa nekugadzirisa, rinowana kuvandudzwa kwakanyanya pamusoro pemamodeli ekutanga, avhareji + 23.69 ChrF ++, + 0.33 COMET, uye + 15.34 BLEU mapoinzi mumitauro makumi matatu nematanhatu akaongororwa; uye (3) chirongwa chakarongeka chekutsvagisa chakabudirira kudzidzisa gumi nevashanu vekutanga-basa, ichisimbisa kugona kwenzvimbo.
Kuenzanisa kwedu pamusoro Google Translate kunoratidza yemakwikwi kuita mumitauro akawanda apo kuziva nzvimbo kuti zvinoda mberi kukura.

Figure 8: Shona Translation

Na dia manana ny ampahatelon'ny fiteny manerantany aza ny fiteny afrikanina, dia mbola tsy voatsinjara amin'ny teknolojia NLP maoderina izy ireo, izay 88 isan-jato no sokajiana ho tena tsy voatsinjara na tsy noraharahiana tanteraka tamin'ny fiteny kajy.
Manolotra ny African Languages Lab (Lab rehetra) izahay, hetsika fikarohana feno izay miresaka ity hantsana teknolojia ity amin'ny alàlan'ny fanangonana angon-drakitra, ny fampandrosoana modely ary ny fananganana fahaiza-manao.
Ny fandraisana anjara ataonay dia ahitana: (1) fantsona fanangonana tahiry mifehy kalitao, manome ny angon-drakitra afrikanina maro karazana sy lahatsoratra maro izay maharitra amin'ny fiteny 40 miaraka amin'ny famantarana 19 miliara amin'ny lahatsoratra tokana sy ny ora 12,628 momba ny angon-drakitra mifanaraka amin'ny teny; (2) fanamarinana fanandramana goavana mampiseho fa ny angon-drakitra, miaraka amin'ny fanamafisana tsara, dia manatsara ny modely fototra, averina + 23.69 ChrF ++, + 0.33 COMET, ary + 15.34 BLEU hevitra manerana ny fiteny 31; ary (3) programa fikarohana miorina izay nahomby tamin'ny mpikaroka dimy ambin'ny folo tany am-piandohana, nametraka fahaiza-manao maharitra eo an-toerana.
Ny fampitahana ny vokatra azonay tamin'ny Google Translate dia mampiseho ny fahombiazan'ny fifaninanana amin'ny fiteny maro samihafa no sady mampahafantatra ireo faritra izay mila fampandrosoana mitohy.

Figure 9: Malagasy Translation

N'agbanyeghị na ọ na-anọchite ihe fọrọ nke nta ka ọ bụrụ otu ụzọ n'ụzọ atọ nke asụsụ ụwa, asụsụ ndị Africa ka na-enweghị nkwado site na teknụzụ NLP nke oge a, na 88 percent nkewa dị ka ndị na-anọchite anya ma ọ bụ na-eleghara anya kpamkpam na asụsụ kọmputa. Anyị na-eweta African Languages Lab (All Lab), atụmatụ nyocha zuru oke nke na-edozi ọdịiche nkà na ụzụ a site na nchịkọta data usoro, mmepe ihe nlereanya, na iwulite ikike. Onyinye anyị gụnyere: (1) pipeline nchịkọta data nke a na-achịkwa nke ọma, na-enye okwu na ederede kachasị ukwuu nke Africa nke nwere ihe ruru ijeri 40 na ijeri 19 nke ederede monolingual na awa 12,628 nke data okwu jikọtara; (2) nkwado nyocha sara mbara na-egosi na dataset anyị, jikọtara ya na nhazi nke ọma, na-enweta ọganihu dị ukwuu karịa ụdị ndị mbụ, na-eme ka +23.69 ChrF++, +0.33 COMET, na +15.34 BLEU gafee asụsụ 31 a nyochara; na (3) usoro nyocha ahaziri nke ọma nke duziri ndị nyocha iri na ise na-amalite ọrụ, na-eguzobe ikike mpaghara na-adịgide adịgide. Nnyocha anyị na-atụnyere megide Google Translate na-ekpughe arụmọrụ asọmpi n'ọtụtụ asụsụ ka ọ na-achọpụta ebe ndị chọrọ mmepe na-aga n'ihu.

Figure 10: Igbo Translation

Duk da wakiltar kusan kashi aya bisa uku na harsunan duniya, harsunan Afirka sun kasance masu mahimmanci ta hanyar fasahar NLP ta zamani, tare da kashi 88 cikin dari da aka rarraba a matsayin mai tsanani ko kuma ba a kula da su ba a cikin ilimin harshe. Mun gabatar da Labaran Harsunan Afirka (All Lab), cikakken shirin bincike wanda ke magance wannan rata na fasaha ta hanyar tattara bayanai na yau da kullum, samar da samfurin, da kuma ginin iyawa. Abubuwan da muka bayar sun hada da: (1) mai sarrafa bututun tattara bayanai mai inganci, samar da mafi girman ingantaccen magana da rubutu na Afirka da yawa da kuma bayanan rubutu wanda ya shafi harsuna 40 tare da alamomin biliyan 19 na rubutu guda aya da kuma sa'o'i 12,628 na bayanan magana masu daidaitawa; (2) gwajin gwaji mai yawa wanda ya nuna cewa dataset inmu, hae da daidaitawa, ya sami ci gaba mai mahimmanci akan samfuran tushe, matsakaicin +23.69 ChrF ++, +0.33 COMET, da +15.34 BLEU maki a cikin harsuna 31 da aka kimanta; da (3) tsarin bincike wanda ya samu nasarar jagorantar masu bincike na farko na farko, kafa karfin gida mai dorewa. Bincikenmu na kwatanta kan Google Translate ya nuna aikin gasa a cikin harsuna da yawa yayin da yake gano wuraren da ke bukatar ci gaba.

Figure 11: Hausa Translation

على الرغم من أن اللغات الأفريقية تمثل ما يقرب من ثلث لغات العالم ، إلا أنها لا تزال تعاني من نقص كبير في الخدمات التي تقدمها تقنيات البرمجة اللغوية الحديثة ، حيث يتم تصنيف 88٪ على أنها ممثلة تمثيلا ناقصا للغاية أو تم تجاهلها تمامًا في اللغويات الحسابية. نقدم مختبر (All Lab) اللغات الأفريقية ، وهي مبادرة بحثية شاملة تعالج هذه الفجوة التكنولوجية من خلال جمع البيانات بشكل منهجي وتطوير النماذج وبناء القدرات. تشمل مساهماتنا (1) خط أنابيب لجمع البيانات الخاضعة للرقابة ، مما يؤدي إلى أكبر مجموعة بيانات نصية ونصية متعددة الوسائط تم التحقق منها في 40 لغة مع 19 مليار رمز من النص أحادي اللغة و 12,628 ساعة من بيانات الكلام المتوافقة ؛ (2) التحقق التجريبي المكثف الذي يثبت أن مجموعة البيانات الخاصة بنا ، إلى جانب الضبط الدقيق ، تحقق تحسينات كبيرة على نماذج خط الأساس ، بمعدل + 23.69 ChrF ++ ، + 0.33 COMET ، 15.34 + و BLEU عبر 31 لغة تم تقييمها ؛ و (3) برنامج بحث منظم نجح في توجيه خمسة عشر باحثًا في مرحلة مبكرة من الحياة المهنية ، مما أدى إلى إنشاء قدرة محلية مستدامة. يكشف تقييمنا المقارن مقابل Google Translate عن الأداء التنافسي في عدة لغات مع تحديد المناطق التي تتطلب التطوير المستمر.

Figure 12: Arabic Translation

In kasta oo ay matalayaan ku dhowaad saddex-meelood meel ka mid ah luqadaha adduunka, haddana luqadaha Afrika waxay weli yihiin kuwo aan si dhab ah loo adeegsan teknoolojiyadda casriga ah ee NLP, iyada oo 88 boqolkiiba lagu tilmaamay inay yihiin kuwo si aad ah loo matalo ama gebi ahaanba loo iska indho-tiray luqadaha xisaabta.

Waxaan soo bandhigeynaa Luqadaha Afrika Lab (Dhammaan Lab), oo ah hindise cilmi baaris oo dhameystiran oo wax ka qabanaya farqiga farsamo ee ku yimaada xogta nidaamsan, horumarinta moodalka, iyo dhisidda awoodda. Waxyaabaha aanu ku soo koobnay waxaa ka mid ah: (1) dhuumaha xogta ee tayada leh ee la xakameeyo, oo dhalinaya xogta ugu badan ee la xaqiijiyay ee afrikaanka ah ee afrikaanka ah iyo xogta qoraalka ah ee ku baahsan 40 luqadood oo leh 19 bilyan oo calaamadood oo ah qoraalka monolingual iyo 12,628 saacadood oo xogta hadalka ah; (2) ansaxin tijaabo ah oo ballaadhan oo muujinaya in xogteena, oo ay weheliso isku-dhafka, ay ku guuleysato horumar la taaban karo oo ku saabsan moodooyinka aasaasiga ah, celcelis ahaan + 23.69 ChrF ++, + 0.33 COMET, iyo + 15.34 BLEU oo dhan 31 luqadood oo la qiimeeyay; iyo (3) barnaamij cilmi-baaris oo qaabeysan oo si guul leh u bartay shan iyo toban cilmi-baare oo hore, aasaasida awoodda maxalliga ah ee waara.

Our qiimaynta la barbardhigo ka dhanka ah Google Translate muujinaysaa qaab tartan dhowr luqadood halka aqoonsashada meelaha u baahan horumarinta sii.

Figure 13: Somali Translation