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

Anonymous ACL submission

Abstract

The digital revolution has left behind hundreds of millions of speakers of low-resource languages (LRLs), particularly in Africa, creating a critical gap in global information access and technological representation. We introduce the African Languages Lab (All Lab), a systematic approach to advancing NLP capabilities for African LRLs through a coordinated research framework. Our work introduces (1) a quality-controlled data collection pipeline, yielding the largest validated multi-modal speech and text dataset for African LRLs spanning 40 languages, encompassing 500 GB of combined parallel text and 4,000 hours of aligned speech data and (2) experiments demonstrating how our custom dataset, combined with QLoRA, achieves improvements across multiple metrics (up to +49.8 chrF++, +60.2 BLEU, and +0.28 COMET points) compared to the base model. Our work establishes a sustainable framework for expanding NLP capabilities to historically underserved languages while fostering local research capacity through structured mentorship and collaboration. We will release our data for research.

1 Introduction

Most of the artificial intelligence (AI) and natural language processing (NLP) research today focuses on about 20 of the 7000 languages of the world, leaving the vast majority of languages understudied (Magueresse et al., 2020). 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 (3) insufficient representation in global research communities (Nigatu et al., 2024; Issaka et al., 2024; Magueresse 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 shown in Figure 1, about 814 African languages are in danger of extinction. Countries like Nigeria, Cameroon, and Ivory Coast have 171, 75, and 65 languages facing extinction, respectively. 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 community. Analyzing mentions of the top 10 global languages versus the top 10 African languages across major academic databases reveals that, on average, for every paper discussing African languages in multilingual LLM contexts, there are 20 papers on global languages in Google Scholar, 23 in CORE, 34 in arXiv, and 70 in IEEE (Table 1) and (Table 4 in the Appendix). This systematic underrepresentation reinforces a cycle of technological marginalization.

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. Started in 2020, the All Lab operates through a tightly coordinated team of dedicated researchers who combine three innovative elements:

1. a systematic, quality-controlled data collection framework powered by our "All Voice" platform, which has enabled the creation of 500GB of validated multi-modal speech and text data,

2. state-of-the-art multilingual modeling tech-

niques optimized for low-resource scenarios, achieving an average +26.7 BLEU, +0.22 COMET, and chrF++ +17.31 improvement over existing baselines, and

3. a sustainable research-first approach that cultivates and empowers young researchers through formal mentorship and collaboration structures.

Our framework represents a step toward digital inclusion for millions of African language speakers, demonstrating that systematic, research-driven, and community-involved approaches can effectively bridge the technological divide while preserving linguistic diversity.

2 Related Work

Our work builds on three pillars of African NLP research: community-driven initiatives, advances in multilingual LLMs, and benchmarks and evaluation frameworks. We examine how these connected areas have shaped the landscape of African NLP.

2.1 Community-Driven Research Initiatives

The development of African NLP has been shaped by several complementary community and institutional efforts. Masakhane, comprising 2,947 Slack members as of January 2025, represents the largest, sustained, community-driven NLP initiatives for African languages (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 development for 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 to advancing multilingual modeling capabilities (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[1] 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).

[1] https://deeplearningindaba.com/

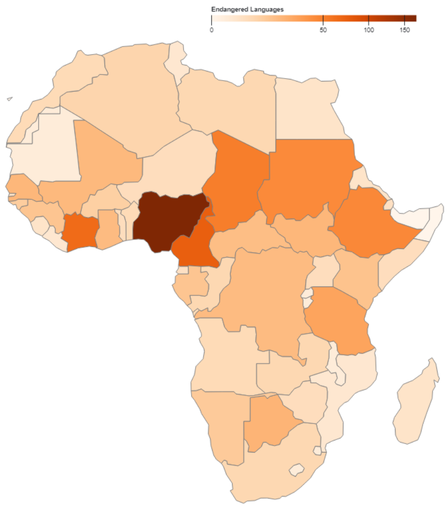

Figure 1: Illustrates the number of endangered languages in each African country using data from Ethnologue

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), 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 (Ope-

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: Research visibility analysis comparing publication volumes for top 10 global languages versus top 10 African languages across major academic databases (Google Scholar (GS), arXiv, IEEE, and CORE) from 2020 to 2024. The stark contrast in publication volumes highlights the digital divide in academic research visibility.

nAI et al., 2024), Gemini (Bao et al., 2023), 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.

2.3 Benchmarks and Evaluation Frameworks

The development of evaluation frameworks has enabled systematic measurement of progress in African NLP. They span multiple task domains: MasakhaNER provides Named Entity Recognition datasets for 10 languages (Adelani et al., 2021), AfriSenti offers sentiment analysis benchmarks across 14 languages (Muhammad et al., 2023), and AFROMT establishes standardized translation benchmarks for 8 languages (Reid et al., 2021).

More targeted evaluation resources include NaijaSenti for Nigerian languages (Muhammad et al., 2022) and Kencorpus for Kenyan languages (Wanjawa et al., 2023). IrokoBench provides human-translated datasets across 17 typologically-diverse African languages (Adelani et al., 2025). 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, 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.

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. Also, All Voices distinguishes itself through its multimodal capabilities, supporting text and audio data collection and validation. 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 [2] is open and free to everyone, aligning with our mission to democratize language technology development.

The platform's architecture, built using React-Native [3] and Firebase [4], 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 manage-

[2] inserted_after_accept
[3] https://reactnative.dev/
[4] https://firebase.google.com/

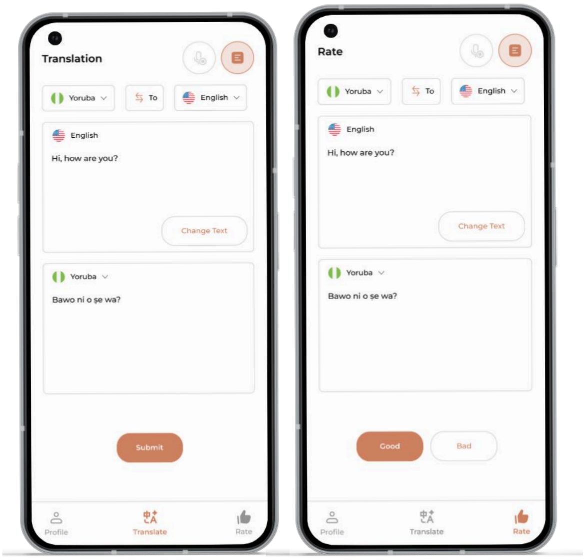

Figure 2: The All Voices platform interface demonstrating its dual functionality: direct text translation from English to Yoruba (left panel) and community-driven translation validation system (right panel).

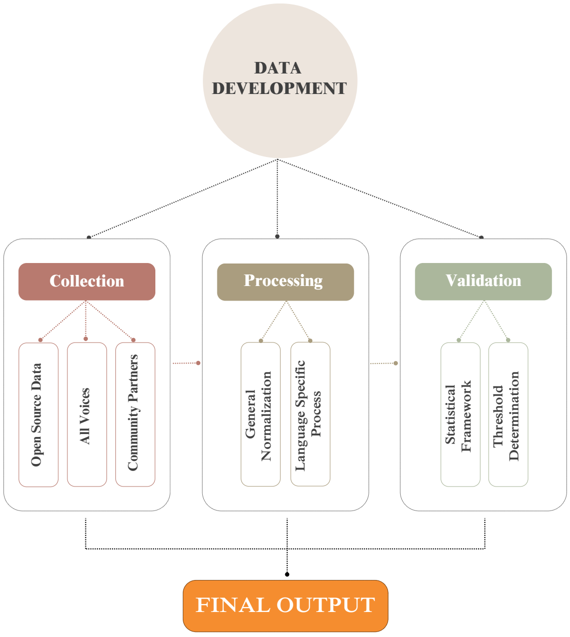

Figure 3: End-to-end data processing pipeline for African language corpus development, illustrating: (a) multi-source data collection, (b) language-specific preprocessing, and (c) validation frameworks for ensuring dataset integrity.

ment provides centralized storage for textual and audio translations with associated metadata, securing all data through AES-256 encryption at rest and TLS 1.3 in transit. 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 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 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, script variant normalization, and tone mark standardization, 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 exhibiting 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 datasets are structured following HuggingFace [5] Dataset specifications, enabling seamless API integration.

3.2 Model Development

To evaluate our dataset's utility and establish baselines for 15 randomly selected African language translations, we experimented with Llama-3.2-3B-Instruct (Grattafiori et al., 2024). We chose this model as our base architecture due to its demonstrated multilingual capabilities and efficient parameter scaling, making it particularly suitable for low-resource scenarios. Given the computational constraints and the need to train on multiple languages efficiently, we implemented parameter-efficient fine-tuning using Quantization-aware Low-Rank Adaptation (QLoRA) (Üstün et al., 2024). Specifically, we employed 4-bit quantization with a LoRA rank of 4, which our preliminary experiments showed to provide an optimal balance between memory efficiency and performance.

The training pipeline was implemented using the Transformers library, with the model fine-tuned using a supervised approach with a consistent instruction format: "Translate the following English text to X:", where X represents the target African language. We configured the model with a maximum input sequence length of 512 tokens and limited the maximum generation length to 600 tokens. To manage computational resources effectively while ensuring comprehensive coverage, we capped the training data at 1 million random examples per language. The training process utilized an NVIDIA H100 GPU with a batch size 16, and gradient accumulation steps of 16. We trained each model for 2 epochs using a learning rate 5.0×10^{-5} with cosine decay, warm-up ratio of 0.15, and BF16 mixed precision. For generation, we used a beam size of 5, temperature of 0.7, and top-p of 0.9.

[5] 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. Together, these metrics comprehensively assess translation quality across different linguistic aspects. Also, the test set was capped at 5,000 random samples per language for consistent evaluation across all languages while maintaining diversity.

4 Results

4.1 Datasets

Through extensive data collection efforts spanning multiple sources, we have compiled a comprehensive dataset encompassing 40 African languages. Our analysis reveals a nuanced stratification with four distinct tiers of resource availability:

1. **Primary resource languages (>40 GB):** This tier includes Arabic, Swahili, and Hausa, representing languages that have benefited from sustained digitization efforts and strong institutional support.

2. **Established digital languages (20-40 GB):** This category includes Afrikaans, Bemba, Amharic, and Xhosa. These languages demonstrate robust digital presence, likely due to consistent documentation and preservation initiatives.

3. **Emerging digital languages (5-20 GB):** A substantial group including Luganda, Lingala, Yoruba, and Malagasy. These languages show emerging digital footprints but still face resource constraints.

Language	Size	Language	Size	Language	Size	Language	Size
Arabic	85.6 GB	Amharic	25.87 GB	Zulu	17.96 GB	Rundi	372.43 MB
Swahili	44.73 GB	Bemba	26.69 GB	Malagasy	16.41 GB	Kikongo	418.38 MB
Hausa	40.73 GB	Xhosa	24.65 GB	Tswana	16.54 GB	Mossi	245.45 MB
Afrikaans	35.70 GB	Ewe	18.52 GB	Luganda	15.96 GB	Tshiluba	282.56 MB
Twi	31.00 GB	Sesotho	18.55 GB	Lingala	13.39 GB	Bambara	200.60 MB
Yoruba	12.05 GB	Kinyarwanda	12.54 GB	Igbo	6.28 GB	Umbundu	147.64 MB
Wolof	8.77 GB	Kikuyu	5.15 GB	Oromo	5.22 GB	Berber	118.63 MB
Chewa	4.85 GB	Shona	4.19 GB	Somali	4.19 GB	Krio	40.51 MB
Fon	1.59 GB	Tigrinya	1.24 GB	Mandinka	33.13 MB	Ngambay	47.61 MB
Fula	404.58 MB	Kanuri	3.92 MB	Fang	36.89 KB	Kiluba	75.20 KB

Table 2: Dataset composition across our 40 African languages, showing the distribution of combined speech and text corpus in GB.

Language	chrF++		COMET		BLEU	
	Llama3B	Finetuned	Llama3B	Finetuned	Llama3B	Finetuned
Amharic	9.2	23.2	0.44	0.72	4.2	21.0
Fula	15.7	42.3	0.38	0.64	2.2	21.3
Yoruba	17.0	19.8	0.35	0.62	1.44	3.2
Igbo	11.5	61.3	0.43	0.68	11.0	71.2
Oromo	18.7	27.7	0.38	0.60	2.70	12.7
Swahili	29.4	43.0	0.59	0.70	19.0	21.0
Hausa	19.7	43.2	0.40	0.67	2.21	48.6
Twi	18.1	52.8	0.41	0.69	3.53	61.1
Shona	17.3	29.2	0.39	0.57	2.20	38.0
Somali	18.8	37.0	0.38	0.64	1.32	40.5
Kinyarwanda	19.1	27.7	0.41	0.57	2.82	8.23
Ewe	16.4	21.4	0.35	0.61	1.84	16.2
Bambara	14.8	34.0	0.36	0.56	1.48	28.3
Wolof	14.2	28.2	0.37	0.53	1.39	32.5
Luganda	24.2	32.9	0.39	0.57	7.44	41.5

Table 3: Performance comparison between base Llama3B and Finetuned Llama3B models across different metrics (chrf++, COMET, and BLEU) for various African languages. Higher scores indicate better performance.

4. **Resource-constrained languages (<5 GB):** The majority of languages in our dataset fall into this category, including widely spoken languages like Bambara, Kikongo, and Fang. This tier reveals critical gaps in digital infrastructure, even for languages with significant speaker populations.

Our analysis highlights a stark digital divide, with the top 3 languages accounting for about 35% of our data. This disparity underscores the urgent need for targeted resource development efforts, particularly for languages with substantial speaker populations but limited digital presence.

4.2 Models

Our evaluation of Llama-3.2-3B-Instruct and its fine-tuned variant across 15 African languages revealed nuanced patterns in translation performance. The base model exhibited varying degrees of effectiveness across languages, with chrF++ scores ranging from 9.2 to 29.4, COMET scores from 0.35 to 0.59, and BLEU scores from 1.32 to 19.0. Notably, Swahili demonstrated good performance across all metrics (chrF++ 29.4, COMET 0.59, BLEU 19.0), likely attributable to its substantial presence in the pre-training corpus and rich digital resources. Fine-tuning yielded substantial but heterogeneous improvements across the language set. The most remarkable gains were observed in Igbo, where performance increased dramatically across all metrics (chrF++ from 11.5 to 61.3, COMET from 0.43 to 0.68, BLEU from 11.0 to 71.2), and Twi (chrF++ from 18.1 to 52.8, COMET from 0.41 to 0.69, BLEU from 3.53 to 61.1). These improvements suggest that targeted fine-tuning can effectively bridge the performance gap for African LRLs. However, the impact of fine-tuning varied significantly across languages. While languages like Hausa and Fula showed substantial improvements (chrF++ increases of 23.5 and 26.6 points respectively), others like Yoruba and Ewe exhibited more modest gains (chrF++ increases of 2.8 and 5.0 points). This variability in improvement patterns suggests complex interactions between model architecture, pre-training data distribution, and language-specific characteristics that warrant further investigation. COMET scores, which correlate better with human judgments, showed consistent improvements across all languages after fine-tuning, with gains ranging from 0.11 to 0.28 points. This uniform improvement in COMET scores, despite varying changes in chrF++ and BLEU, indicates that fine-tuning enhances translation quality in ways that may not be fully captured by surface-level metrics.

5 Conclusion

We presented the African Languages Lab (All Lab), introducing a comprehensive initiative for advancing NLP capabilities in African languages through three key innovations:

1. the All Voices platform, which enabled the creation of the largest validated multi-modal dataset spanning 40 African languages,

2. experimental validation demonstrating improvements through QLoRA fine-tuning (average gains of +26.7 BLEU, +0.22 COMET, and +17.31 chrF++), and

3. a sustainable research development program that has successfully mentored fifteen early-career researchers.

Our results demonstrate that targeted data collection and parameter-efficient fine-tuning can significantly improve performance for low-resource African languages, while our open-source approach establishes a replicable framework for expanding NLP capabilities to other underserved languages.

6 Limitations

6.1 Computational and Model Constraints

Our experimental evaluation was conducted using a single base model (Llama-3.2-3B-Instruct) with QLoRA fine-tuning, primarily due to computational constraints. While this approach allowed us to demonstrate the potential of our dataset, it may not fully represent the optimal architecture for African language processing. Additionally, we were only able to evaluate on a subset of our collected languages and data, which may not fully represent the diversity of African languages.

6.2 Data Quality and Validation

Our data cleaning and standardization procedures rely heavily on automated approaches due to limited access to native speakers across all 40 languages. While we implement grounded statistical validation methods, this automation may miss subtle linguistic nuances, dialectal variations, and cultural contexts that human validators would catch. Also, the quality assessment of our dataset, may not fully capture potential biases or quality issues.

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.

These limitations inform our ongoing work and highlight important 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

7.1 Research Capacity Building

The All Lab maintains a structured research development program which has successfully mentored fifteen early-career researchers across four institutions through a comprehensive twelve-week curriculum. The program implements a four-phase model: foundation building, guided research, independent project development, and project completion. Each researcher receives one-on-one mentorship from experienced researchers, with high-performing participants transitioning into extended research roles. This approach has shown measurable success, evidenced by achievements such as the Ozy Genius Award recognition, while establishing a sustainable pipeline for African NLP talent development.

7.2 Sustainable Societal Impact

The All Lab's work advances several United Nations Sustainable Development Goals, particularly in education and inequality reduction. Our platform democratizes access to digital resources for millions of speakers of African LRLs, enabling communities to preserve their linguistic heritage while participating in the digital economy. This work addresses historical technological disparities through three key mechanisms: increasing digital representation of marginalized languages, enabling community-led content creation, and facilitating open knowledge transfer.

7.3 Challenges and Mitigation Strategies

We acknowledge several critical challenges in our work. The commercial viability of LRL technologies remains limited, affecting sustainable development. Data collection and evaluation face significant hurdles due to limited digital presence and infrastructure constraints. To address these challenges, we implement: (1) partnerships with academic and industry stakeholders to ensure resource sustainability, (2) a quality-controlled data collection framework that balances automation with human validation, and (3) structured community engagement programs to ensure cultural and linguistic authenticity.

7.4 Future Directions and Long-term Impact

Our work establishes a framework for sustainable LRL technology development guided by the African philosophy of Ubuntu (a philosophy that affirms the positive values of inclusivity, community, difference, anti-racism, hospitality, and openness to others). We outline a clear roadmap for future development, including expanding language coverage, optimizing model architectures for African LRLs, and strengthening research partnerships. This approach not only advances technical capabilities but also contributes to cultural preservation, educational advancement, and economic inclusion for hundreds of millions of speakers of low-resource African languages. Through these initiatives, we aim to contribute to addressing the systematic underrepresentation of African languages in NLP while establishing replicable methodologies for LRL technology development globally. Our work demonstrates that technical innovation in NLP can directly contribute to broader societal goals while maintaining rigorous research.

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.

Guangsheng Bao, Zebin Ou, and Yue Zhang. 2023. GEMINI: Controlling the sentence-level summary style in abstractive text summarization. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 831–842, Singapore. 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 *Find-*

ings 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, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, 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.

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).

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 Appendices

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"