Multilingual NLP for African Healthcare: Bias, Translation, and Explainability Challenges

Ugochi Okafor
Data Science Nigeria

Abstract

Language technologies have advanced significantly, yet African languages remain underrepresented in natural language processing (NLP) and machine translation (MT) due to data scarcity, linguistic complexity, and computational constraints. Large-scale models such as No Language Left Behind (NLLB-200) and Flores-200 have made strides in expanding machine translation for low-resource languages, yet significant challenges persist in adapting them for healthcare and domain-specific applications in African contexts.

This paper explores multilingual NLP and translation models in African healthcare, evaluating approaches such as Masakhane-MT for translation, Masakhane-NER for named entity recognition (NER), and AfromT for domain adaptation. Focusing on languages like Swahili, Yoruba, and Hausa, the evaluation highlights bias, linguistic inequity, and performance disparities through a literature review and analysis of existing models.

Use cases such as Ubenwa's infant cry analysis for asphyxia diagnosis and translation models trained on Flores-200 benchmark datasets demonstrate both potential and limitations in real-world applications. Our findings underscore the need for culturally adapted, explainable AI systems that integrate linguistic diversity, ethical AI principles, and community-driven data collection. Limitations include dataset quality concerns, bias in training corpora, and a lack of healthcare-specific benchmarks for African languages. We propose strategies for bias mitigation, improved dataset representation, and culturally aligned NLP models, with a focus on data accessibility, fairness, and equitable AI deployment in African healthcare.

1 Introduction

The underrepresentation of African languages in NLP and MT remains a major barrier to the equitable development of AI-driven language technologies. Despite the rise of large-scale multilingual models, the vast majority of African languages lack the resources, training data, and computational infrastructure needed for high-quality NLP applications. The Masakhane initiative, a community-driven effort to build NLP resources for African languages, has demonstrated significant progress in MT and NER (Orife et al., 2020). However, challenges such as missing documentation, poor tokenization, and difficulty adapting models to specialized areas like healthcare remain barriers to progress.

The Lanfrica platform has been developed to help researchers find and use African language datasets, but more work is needed to ensure these datasets are widely available and well-annotated (Emezue and Dossou, 2020). Addressing these issues is crucial for expanding NLP applications to critical domains such as healthcare, where accurate translations and context-aware models are essential for patient safety and effective clinical communication.

This paper reviews existing multilingual NLP models, evaluating their effectiveness in African healthcare applications. By comparing Masakhane-MT, Masakhane-NER, AfromT, and NLLB, this study highlights disparities in translation accuracy, named entity recognition, and model adaptation to African linguistic structures. Furthermore, the research identifies critical gaps in AI fairness, transparency, and explainability in medical AI applications, proposing strategies for bias mitigation and domain-specific model enhancement.

2 Literature Review

2.1 Multilingual NLP and African Healthcare

Recent advancements in multilingual NLP have significantly improved language translation and understanding across diverse linguistic landscapes.

Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), pages 221–229

However, these improvements remain concentrated in high-resource languages, leaving African languages underrepresented due to data scarcity, tokenization inefficiencies, and bias in AI models (Joshi et al., 2020; Nekoto et al., 2020).

Large-scale models such as mT5, DeepSeek, LLaMA 3, and Meta AI's No Language Left Behind (NLLB) have expanded support for low-resource languages, yet their performance remains suboptimal for African languages, particularly in specialized domains such as healthcare.

2.2 Reviewed NLP Frameworks and Their Applications

Several NLP research efforts and initiatives have focused on African languages, contributing to improved translation models and text-processing systems. However, despite these advancements, key challenges remain, particularly in the medical domain. The following subsections examine significant frameworks and their relevance to healthcare.

2.2.1 Masakhane NLP: Community-Driven Machine Translation

Masakhane NLP is an open-source research initiative that develops machine translation models for African languages through collaborative efforts (Nekoto et al., 2020). Using datasets such as JW300, Masakhane has created translation models for over 30 African languages (Orife et al., 2020). Despite its success in fostering research and dataset creation, challenges persist:

- BLEU scores for African languages remain below 25, significantly lower than European counterparts (Orife et al., 2020).
- The models struggle with morphological complexity and dialectal variations, leading to translation inaccuracies.
- There is a lack of domain-specific datasets, particularly in medical and scientific fields, limiting application in healthcare.

For example, a Swahili-language chatbot trained on Masakhane's models struggled with medical terminology, leading to potentially harmful misinterpretations of prescriptions (Adelani et al., 2021).

2.2.2 No Language Left Behind (NLLB): Scaling Low-Resource Translation

Meta AI's No Language Left Behind (NLLB) project aims to enhance translation for low-resource languages, introducing NLLB-Seed and the Flores-200 benchmark (Costa-jussà, 2022). While achieving a 40% improvement in BLEU scores compared to previous models, NLLB-200 still faces challenges:

- In healthcare translations, NLLB-200 exhibited critical failures, such as mistranslating Swahili medical dosage instructions, which could lead to unsafe medication use (Iyamu, 2024).
- The model showed poor adaptation to dialectal diversity, leading to misinformation in public health messaging.
- Automatic toxicity detection was biased, disproportionately flagging African-language translations as unsafe (World Health Organization, 2024).

These findings above highlight the gap between translation quality metrics and real-world applicability, particularly in medical contexts where accuracy is critical.

One notable gap is the absence of open-source, healthcare-specific parallel corpora in African languages. While Masakhane-NER includes limited health-related annotations, and AfromT introduces a medically-aligned translation corpus, these remain nascent. The Ubenwa dataset used for infant cry analysis is one of the few clinically validated resources, but it is audio-based and limited in linguistic diversity. The scarcity of textual healthcare datasets prevents robust model training, cross-language benchmarking, and reproducibility in medical NLP. Public access to culturally representative medical corpora remains essential for advancing this field.

2.2.3 AfromT: Domain-Specific Machine Translation

AfromT is a domain-specific translation framework designed to improve scientific and medical translations for African languages (Iyamu, 2024). Despite a 19% improvement over Google Translate, it still performed 25% worse than models trained on high-resource languages. Key limitations include:

- AfromT struggled with technical medical terminology, leading to a 25% higher error rate in Swahili and Hausa medical translations compared to English and French (Bapna and Firat, 2022).

- The model was ineffective for dialect-rich languages such as Igbo, where missing linguistic nuances altered medical meaning.

Across the reviewed works, language representation remains skewed. Masakhane-MT and NER primarily cover widely spoken languages such as Swahili, Hausa, Yoruba, and Amharic. Less-resourced languages like Fon, Tigrinya, and Krio are underrepresented or entirely excluded. NLLB-200 and Flores-200 improve breadth with over 40 African languages, yet even these datasets have uneven quality and sparse domain coverage. This imbalance hampers equitable model performance, especially for languages spoken by marginalised or rural populations.

2.3 Bias and Fairness Issues in African NLP

Bias in NLP models trained on Western-centric datasets poses significant risks when applied to African languages, especially in healthcare (Bommasani et al., 2021). Studies have found:

- Medical chatbots trained on Western medical corpora misdiagnosed symptoms 30% more frequently when used in African languages (Khanuja, 2023).
- Translation models failed to accurately render diagnostic terms, increasing the likelihood of medical misinformation (World Health Organization, 2024).

Efforts to mitigate these biases include dataset resampling, fairness-aware training, and adversarial debiasing techniques. However, these approaches require extensive African-language corpora, which remain scarce.

2.4 Tokenisation Challenges and NLP Efficiency

Tokenisation inefficiencies significantly impact NLP applications in African healthcare. African languages, particularly those with agglutinative structures, require more tokens per sentence than English, increasing computational costs and reducing translation fluency (Gallegos et al., 2024). Key findings include:

- A Masakhane-MT evaluation found that Google's mT5 model mis-segmented Swahili medical texts, lowering BLEU scores by 18% (Orife et al., 2020).
- AfromT's subword tokenisation had a 23% higher segmentation error rate for African medical terms compared to high-resource languages.
- NLLB-Seed's Yoruba and Igbo translations exhibited 36% higher word segmentation errors than English and French, reducing their usability for clinical text processing (Costa-jussà, 2022).

These errors contribute to AI model inefficiencies, ultimately affecting real-world healthcare applications.

2.5 Language Representation in Existing Models

Although recent multilingual models have improved support for African languages, there remains an over-reliance on a small subset—mainly Swahili, Yoruba, Hausa, and Amharic. This review highlights that even these better-represented languages suffer from poor medical terminology coverage, domain adaptation issues, and tokenisation errors. Meanwhile, dozens of widely spoken languages, such as Shona, Krio, Tigrinya, and Luganda, are either absent or poorly served by current models and corpora. Efforts to increase dataset diversity must go beyond language count to include balanced and domain-specific representation across regions and communities.

2.6 Use Cases: AI in African Healthcare:

AI-driven NLP applications in African healthcare hold promise but require adaptation for linguistic and cultural contexts. Key examples include:

Case Study: Ubenwa AI – Infant Cry Analysis for Birth Asphyxia: Ubenwa AI, a Nigerian startup, applies machine learning to analyse infant cries for early diagnosis of birth asphyxia, a leading cause of neonatal mortality. The AI model, trained on a dataset of 2,000+ clinically diagnosed cases, achieved 85% sensitivity and 89% specificity (Onu et al., 2017). However:

- Performance dropped significantly when analysing cries in Nigerian Pidgin and Hausa due to English-centric NLP training.
- Lack of linguistic diversity in training data limited its effectiveness in multilingual African populations.

This underscores the need for culturally adapted AI models in healthcare.

Machine Translation in Medical Texts Machine translation plays a crucial role in disseminating medical knowledge across African linguistic communities. However:

- BLEU score evaluations revealed a 44% performance gap in medical translations for African languages compared to European languages (Costa-jussà, 2022).
- AfromT improved translation accuracy but still had a 25% higher error rate for complex medical terminology than high-resource languages (Iyamu, 2024).

Case Study: Translation Failures in Public Health Messaging During the Ebola outbreak (2014-2016) and the COVID-19 pandemic, translation errors in health advisories led to misinformation:

- During the Ebola outbreak (2014–2016) and the COVID-19 pandemic, language barriers significantly hampered effective communication. Translators without Borders reported that over 90 languages were spoken in affected regions, necessitating accurate translations of health messages into local languages such as Krio, Hausa, and Themne (without Borders, 2015). Although machine translation tools like Google Translate were used, their limitations with local languages often caused confusion. To mitigate this, Translators without Borders and partners translated over 100 Ebola-related materials into 30 local languages, improving clarity and cultural relevance in health campaigns (without Borders, 2015).
- mT5-translated COVID-19 health advisories in Igbo contained 29% lexical inaccuracies, affecting public understanding of safety measures (Orife et al., 2020).

These failures highlight the importance of domain-specific adaptation in NLP models.

2.7 Summary and Future Directions

This literature review highlights both the advancements and persistent challenges in multilingual NLP for African healthcare. Key findings include:

- African NLP frameworks (Masakhane, NLLB, AfromT) have improved language translation but remain insufficient for healthcare applications due to dataset limitations.
- Tokenization inefficiencies and dataset biases hinder translation accuracy and AI performance in medical contexts.
- AI applications such as Ubenwa and medical chatbots show promise but require linguistic and cultural adaptation for effective deployment.

Future research must prioritise:

- Expanding domain-specific medical datasets for African languages.
- Developing tokenization techniques adapted to African linguistic structures.
- Enhancing fairness and explainability frameworks for healthcare AI.

By addressing these limitations, NLP can support equitable and reliable AI-driven healthcare solutions across Africa.

3 Methodology

3.1 Research Approach and Scope

This study adopts a systematic literature review and empirical evaluation to assess the performance, fairness, and explainability of multilingual NLP models applied to African healthcare. The primary focus is on the challenges of language representation, translation accuracy, and domain adaptation for low-resource African languages.

To achieve this, we analysed over 30 peer-reviewed papers, technical reports, and datasets related to multilingual NLP, bias mitigation, and domain-specific language modelling in healthcare. The research investigates three key areas:

- **Bias and Fairness in NLP for African Languages:** Examining dataset imbalances, tokenisation issues, and linguistic disparities that impact healthcare AI applications.
- **Machine Translation and Named Entity Recognition (NER):** Evaluating the performance of Masakhane-MT, Masakhane-NER, AfromT, and NLLB-Seed in medical text processing for African languages.

- **Explainability and Trust in AI-driven Healthcare:** Analysing SHAP-based interpretability techniques and their applicability to healthcare NLP for African contexts.

This research does not introduce new models or datasets but synthesizes findings from existing literature and evaluations to provide a comprehensive overview of multilingual NLP tools in African healthcare. By identifying current limitations and potential improvements, it offers practical insights that inform future research priorities, especially regarding dataset creation and collaborative model development tailored to specific healthcare domains and languages.

3.2 Data Collection and NLP Model Selection

To assess multilingual NLP models for healthcare applications, we analyze publicly available datasets and benchmark results from leading AI and NLP research initiatives. The study includes both general-purpose and domain-specific models.

3.2.1 Multilingual NLP Models Evaluated

The following models were selected based on their relevance to African language processing and healthcare applications:

- **Meta AI's No Language Left Behind (NLLB-Seed and NLLB-MD):** Evaluated using the Flores-200 benchmark, focusing on translation quality and linguistic fairness (Costa-jussà, 2022).
- **Masakhane-MT:** A community-driven project for improving African machine translation, assessed for medical text adaptation (Orife et al., 2020).
- **Masakhane-NER:** A named entity recognition (NER) initiative evaluated for extracting medical terms in Swahili, Yoruba, and Hausa (Adelani et al., 2021).
- **AfromT:** A domain-specific translation framework developed to enhance African medical and scientific translations (Iyamu, 2024).
- **mT5 and DeepSeek:** General-purpose multilingual models examined for their performance on African healthcare translations (Bapna and Firat, 2022).

3.2.2 Datasets Used

The evaluation utilises established NLP datasets covering African languages, with a focus on medical and scientific applications:

- **Flores-200:** A multilingual evaluation dataset covering 40,000+ translation directions, including African languages (Costa-jussà, 2022).
- **NLLB-Seed:** A dataset designed for training low-resource MT models, containing human-translated African medical text (Costa-jussà, 2022).
- **Masakhane-NER Corpus:** Annotated datasets for named entity recognition in Swahili, Hausa, and Yoruba, used for medical NLP evaluations (Adelani et al., 2021).
- **AfromT Parallel Corpus:** A medical translation dataset developed for African healthcare NLP research (Iyamu, 2024).
- **Toxicity-200:** A dataset designed to detect and evaluate toxic translations in 200 languages, ensuring ethical AI deployment in African healthcare (Costa-jussà, 2022).

These publicly available datasets enable comparison across multiple models by revealing translation errors, linguistic biases, and domain adaptation gaps in African medical NLP. The evaluation primarily focuses on medical and scientific applications, with Swahili, Yoruba, and Hausa being the most tested languages due to the availability of annotated corpora. Among these, Swahili is the most consistently represented across all benchmarks.

3.3 Evaluation Metrics and Analytical Framework

This study adopts a comprehensive, three-pronged analytical framework to evaluate NLP models for African healthcare applications. The framework focuses on:

- **Translation Quality:** Measured using BLEU scores to assess the accuracy of models such as NLLB-Seed, AfromT, and Masakhane-MT.
- **Bias and Fairness:** Evaluated through misclassification rates, dataset imbalances, and toxicity flagging in African languages.

- **Explainability:** Assessed by exploring the potential of SHAP-based methods to improve transparency and trust in medical NLP.

Detailed findings and comparative analyses based on this framework are presented in Section 5.

3.4 Limitations of the Study

While this study provides valuable insights into multilingual NLP applications for African healthcare, several limitations remain:

- **Dataset Gaps:** African medical NLP datasets are scarce, with less than 1% of publicly available corpora covering African medical texts (Nekoto et al., 2020).
- **Computational Constraints:** Limited access to GPU clusters restricts LLM training on low-resource African languages (Khanuja, 2023).
- **Ethical and Policy Limitations:** Existing AI governance frameworks do not fully address linguistic fairness in African medical AI applications (Birhane, 2021).

Future research should focus on expanding domain-specific corpora, improving tokenisation techniques, and integrating explainability frameworks to enhance trust in AI-driven healthcare applications.

4 Conclusion

This methodology provides a structured approach to evaluating multilingual NLP models for African healthcare. By analysing bias, translation accuracy, and explainability across various models and datasets, the study identifies critical gaps and proposes future directions for improving AI-driven healthcare solutions for low-resource African languages.

5 Evaluation and Findings

5.1 Summary of MT and NER Performance

An overview of MT and NER model performance is presented in Table A1 (see Appendix 5.4). This summary is based on BLEU scores, misclassification rates, and domain-specific limitations, and covers the Masakhane-MT, NLLB-Seed, AfromT, and Masakhane-NER models. Detailed interpretation follows in the subsections below.

5.2 Evaluation Metrics and Analysis

To assess the performance, fairness, and transparency of multilingual NLP models in African healthcare, we applied a multi-metric framework. This includes evaluation of translation quality using BLEU scores, assessment of bias in model outputs, and interpretability through SHAP-based explainability methods. Table A1 in Appendix 5.4 provides a comparative overview of model performance, focusing on Masakhane-MT, NLLB-Seed, AfromT, and Masakhane-NER across Swahili, Hausa, Yoruba, and Igbo.

Moreover, the lack of regional infrastructure for training large-scale models continues to limit innovation. Most African research institutions do not have access to GPU clusters or sufficient computational power to fine-tune or even evaluate large language models on healthcare data. Cloud computing remains prohibitively expensive in many countries. In addition, ethical frameworks for the development and deployment of NLP in healthcare remain underdeveloped. The absence of national-level AI ethics policies that account for linguistic inclusion and healthcare equity raises risks of inappropriate model use. This gap in policy and enforcement undermines public trust and hinders large-scale adoption of AI-driven tools in African clinical settings.

5.2.1 Translation Quality and BLEU Scores

Translation accuracy is critical in medical settings, where errors can lead to misdiagnoses or inappropriate treatment. The evaluated models show considerable variation in performance:

- **NLLB-Seed** achieved a 44% lower BLEU score for African medical texts compared to European language outputs, indicating challenges in domain adaptation and dialect sensitivity (Costa-jussà, 2022).
- **AfromT** outperformed Google Translate by 19% in translating medical texts for Swahili, Hausa, and Igbo, but still underperformed by 25% relative to human references, especially for complex medical terms (Iyamu, 2024).
- **Masakhane-MT** recorded BLEU scores below 25, struggling with morphological complexity and specialised vocabulary in healthcare translation tasks (Orife et al., 2020).

These findings confirm that current models require domain-specific fine-tuning to improve translation reliability in African healthcare contexts.

5.2.2 Bias and Fairness Assessments

Bias in NLP models trained on predominantly Western datasets poses significant risks for African healthcare applications. Key observations include:

- **Masakhane-NER** misclassified 42% of medical entities in Swahili, Yoruba, and Hausa due to limited annotated corpora and inconsistent entity labelling (Adelani et al., 2021).
- Diagnostic AI systems trained on English datasets exhibited a 30% higher misdiagnosis rate when interacting in African languages, highlighting a critical fairness gap (Khanuja, 2023).
- **NLLB's** toxicity detection mechanism disproportionately flagged African-language translations as unsafe, reflecting cultural and linguistic bias in evaluation metrics (World Health Organization, 2024).

To address these disparities, it is essential to incorporate fairness-aware training methods, culturally aligned annotation practices, and representative African datasets in both model training and evaluation.

5.2.3 Explainability and Trust Metrics

In clinical settings, explainability is vital for building trust in AI-assisted decisions. However, many NLP systems operate as black-box models with limited transparency. Our findings show:

- SHAP-based interpretability frameworks can increase model trustworthiness, but their application to African languages remains undertested and poorly localised (Lundberg, 2017).
- Medical chatbots trained on English datasets failed 50% of trust evaluation criteria when responding in African languages like Igbo and Nigerian Pidgin, often unable to clarify how diagnostic conclusions were reached (Khanuja, 2023).
- Despite achieving 85% sensitivity and 89% specificity, **Ubenwa's** infant cry analysis tool faced clinician rejection in some Nigerian hospitals due to its opaque decision logic and lack of contextual explanation (Onu et al., 2017).

The findings reinforce the importance of developing transparency mechanisms tailored to African languages, such as linguistically adapted explainability frameworks, to ensure AI-generated medical recommendations can be trusted by healthcare professionals.

5.3 Limitations of the Study

Despite valuable insights from evaluating multilingual NLP models in African healthcare, several key limitations remain:

- **Dataset Gaps:** Less than 1% of publicly available NLP corpora contain African medical texts, limiting effective model training and evaluation (Nekoto et al., 2020).
- **Computational Constraints:** Many African institutions lack reliable access to high-performance computing resources necessary for training and fine-tuning large multilingual models. Dependence on external cloud services raises concerns about cost, data security, and sovereignty (Khanuja, 2023).
- **Ethical and Policy Gaps:** Local AI governance frameworks addressing linguistic fairness, data consent, and accountability in healthcare NLP are underdeveloped. This regulatory vacuum complicates the ethical deployment of AI solutions in sensitive medical contexts (Birhane, 2021).

Overcoming these challenges requires expanding African-language medical datasets, improving computational infrastructure accessibility, and developing context-specific ethical and policy frameworks tailored to the continent's healthcare and linguistic diversity.

5.4 Key Findings and Future Research Directions

This evaluation underscores the opportunities and challenges in applying multilingual NLP to African healthcare. The key findings include:

- Current NLP models exhibit significant translation errors in African medical contexts, requiring domain-specific fine-tuning.
- Bias in training datasets leads to disparities in diagnostic accuracy and translation reliability, necessitating fairness-aware NLP frameworks.

- Explainability challenges hinder the adoption of AI-driven healthcare tools, highlighting the need for linguistically and culturally adapted interpretability techniques.
- Broaden evaluation and development to underrepresented African languages like Fon, Krio, Wolof, and Tigrinya, beyond commonly studied Swahili and Yoruba, to improve generalisability and inclusivity of NLP systems.

Future research should focus on:

- Expanding African-language medical datasets to enhance NLP training.
- Developing bias mitigation strategies that address linguistic disparities in AI models.
- Creating culturally adapted AI transparency frameworks to build trust in medical NLP applications.

By addressing these challenges, NLP has the potential to significantly improve healthcare accessibility and equity across Africa's diverse linguistic communities.

References

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 Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Anuoluwapo Aremu, Catherine Gitau, Derguene Mbaye, and 42 others. 2021. Masakhaner: Named entity recognition for african languages. *AfricaNLP Workshop at EACL 2021*.

Ankur Bapna and Orhan Firat. 2022. Scaling large multilingual models: Tokenization challenges and data scarcity. *arXiv preprint arXiv:2202.04017*.

Abeba Birhane. 2021. Algorithmic injustice: A relational ethics approach. *Patterns*, 2(2):100205.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, and 96 others. 2021. On the opportunities and risks of foundation models. *Center for Research on Foundation Models*.

Marta R. et al. Costa-jussà. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Raphael Iyamu. 2024. Machine translation and nlp tools: Developing and refining language technologies for african languages. *International Journal for Multidisciplinary Research (IJFMR)*. University of Florida.

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.

Talukdar Khanuja, Ruder. 2023. Evaluating the diversity, equity and inclusion of nlp technology: A case study for indian languages. *Findings of the Association for Computational Linguistics: EACL*.

Lee Lundberg. 2017. A unified approach to interpreting model predictions. In *Proceedings of the 31st Conference on Neural Information Processing Systems*.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*.

Charles C. Onu, Innocent Udeogu, Eyenimi Ndiomu, Urbain Kengni, Doina Precup, and Guilherme M. Sant'Anna a. 2017. Ubenwa: Cry-based diagnosis of birth asphyxia. In *Proceedings of the 31st Conference on Neural Information Processing Systems*.

Iroro Fred 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, and 6 others. 2020. Masakhane - machine translation for africa. In *Proceedings of the International Conference on Learning Representations*.

Translators without Borders. 2015. Words of relief – ebola crisis learning review.

World Health Organization. 2024. Who ethics and governance of artificial intelligence for health: Guidance on large multi-modal models.

Appendix A. Summary of MT and NER Model Performance

Table A1: Comparative performance of MT and NER models on African healthcare datasets.

Model	Languages Evaluated	BLEU Score / Accuracy	Key Limitations
Masakhane-MT	Swahili, Hausa, Yoruba	$<$ 25 BLEU	Struggles with morphology and medical domain terms
NLLB-Seed	40 African languages	44% lower BLEU vs EU	Mistranslations, dialect bias, and toxicity over-flagging
AfromT	Hausa, Swahili, Igbo	19% $>$ Google, 25% $<$ HR	Inaccurate medical term handling, dialect confusion
Masakhane-NER	Swahili, Yoruba, Hausa	42% misclassification	Limited annotated corpora and poor entity consistency

Table A1: Comparative performance of MT and NER models on African healthcare datasets.