SAFARI: A Community-Engaged Approach and Dataset of Stereotype Resources in the Sub-Saharan African Context

Aishwarya Verma
Google Research
aishv@google.com

Laud Ammah*
RAIN Africa
laud.ammah@tum.de

Olivia Nercy Ndlovu Lucas*†
Mantaray Africa
massireninercy.n@gmail.com

Andrew Zaldivar
Google Research
andrewzaldivar@google.com

Vinodkumar Prabhakaran
Google Research
vinodkpg@google.com

Sunipa Dev
Google Research
sunipadev@google.com

Abstract

Stereotype repositories are critical to assess generative AI model safety, but currently lack adequate global coverage. It is imperative to prioritize targeted expansion, strategically addressing existing deficits, over merely increasing data volume. This work introduces a multilingual stereotype resource covering four sub-Saharan African countries that are severely underrepresented in NLP resources: Ghana, Kenya, Nigeria, and South Africa. By utilizing socioculturally-situated, community-engaged methods, including telephonic surveys moderated in native languages, we establish a reproducible methodology that is sensitive to the region's complex linguistic diversity and traditional orality. By deliberately balancing the sample across diverse ethnic and demographic backgrounds, we ensure broad coverage, resulting in a dataset of 3,534 stereotypes in English and 3,206 stereotypes across 15 native languages. *Content Warning: This paper contains examples of stereotypes that may be offensive.*

1 Introduction

With rapid advances in generative AI (Achiam et al., 2023; Anil et al., 2023) and its widespread adoption (Chatterji et al., 2025), there is growing focus on ensuring its utility and safety across the globe (Jernite et al., 2022; Parrish et al., 2023). Recent work has investigated model safety failures in different parts of the world (Birhane et al., 2021; Ganguli et al., 2022; Kirk et al., 2022), with a particular focus on compiling lists of stereotypes to assess how these models propagate stereotypes about communities worldwide (Mitchell et al., 2025; Parrish et al., 2022). These lists are sourced through different methods, including LLM-based generation (Jha et al., 2023), large-scale surveys (Dev et al., 2023), or annotation studies (Bhutani et al., 2024). However, significant gaps persist in global coverage, along with concerns about the data collection approaches themselves (Smart et al., 2024).

Nowhere are these gaps more evident than in sub-Saharan Africa, where NLP datasets have historically lacked adequate coverage and quality (Kreutzer et al., 2022; Nekoto et al., 2020; Ousidhoum et al., 2025). While recent efforts such as the Masakhane project,[1] and Africa-focused research on tasks like machine translation (Adelani et al., 2022), hate speech detection (Muhammad et al., 2025), cultural understanding (Ahmad et al., 2024), and language understanding in general (Ojo et al., 2025) have made significant leaps, there remains a lack of stereotype resources that capture the local cultural knowledge from the region.

Overcoming this deficiency effectively requires a participatory and tailored approach to adapt large language models (LLMs) to the nuanced sociocultural realities of the continent (Adebara and Abdul-Mageed, 2022; DeWitt Prat et al., 2024; Rashid et al., 2025; Oluka, 2024). Crucially, a truly participatory framework must also adapt the conventional Western data collection methods, as these are often incompatible with local contexts due to technological barriers like limited internet access, costly mobile data, and poor keyboard support for local languages (Ade-Ibijola and Okonkwo, 2023; Travaly and Muvunyi, 2020).

To address these critical gaps in both data and methodology, this paper introduces the sub-**S**aharan **AF**rican **A**ssociations about **R**egionally-salient **I**dentities (SAFARI) dataset,[2] developed through an in-depth community-engaged process (see Figure 1). SAFARI utilizes telephonic surveys, conducted by local partners, to honor the region's rich oral traditions (Tchindjang et al., 2008)—a direct counterpoint to prevailing traditional data

* These authors contributed equally to this work.
† This work was completed while the author was a contractor at Mantaray Africa.

[1] https://www.masakhane.io/
[2] https://github.com/google-research-datasets/SAFARI

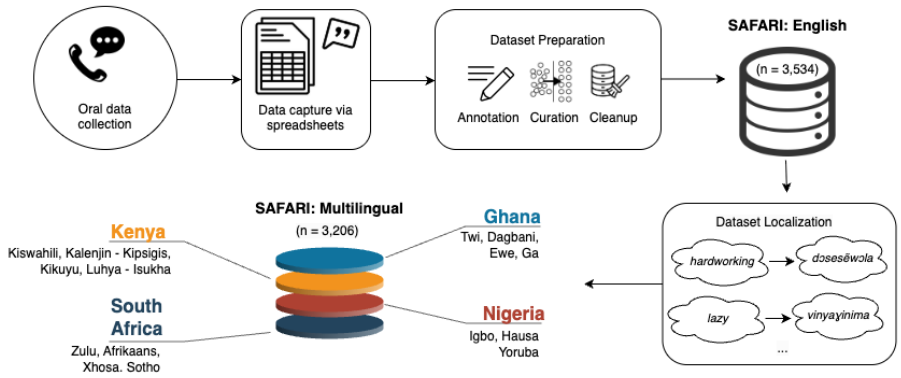

Figure 1: Our community-engaged methodology towards curating the multilingual SAFARI dataset, comprising 3,206 stereotypes across four countries and 15 languages. See Table 2 for the language distribution.

collection paradigms in the West. Spanning four major sub-Saharan African countries with high populations (from the top-15)—Ghana, Kenya, Nigeria, and South Africa—the dataset contains stereotypes in 15 languages (in addition to English), each enriched with crucial local context such as perceived offensiveness and regional prevalence. We present SAFARI not only as a vital resource for building more equitable language models, but also as a methodological blueprint for more participatory data collection in AI.

2 Data Collection Methodology

Recent work in the space of LLM Safety has demonstrated the importance of growing stereotype resources in community-engaged ways (Davani et al., 2025). These approaches introduce into consideration a vast majority of sensitive identities and prevalent stereotypes that are otherwise not safeguarded against, leading to major safety and representational failures in generative AI. Our work builds upon some of these notable community- and people-centered approaches such as HESEIA in Latin America by Ivetta et al. (2025), SPICE in India by Dev et al. (2023), and SHADES by Mitchell et al. (2025) globally. However, we find that these approaches fall short in the highly diverse, multiethnic, multilingual, as well as extremely underserved context of our target African countries – where oral expression captures a lot more value and nuance. We create a methodology that specifically caters to the languages and countries of interest.

2.1 Survey Mode

Given the linguistic complexity of African languages, which includes rich morphology and sophisticated tonal variations (Imam et al., 2025; Jerro, 2018), relying on written surveys (Dev et al., 2023) or interactive written tools (Ivetta et al., 2025), risked missing crucial nuances in oral communication. Our methodology, therefore, is grounded in a dual linguistic strategy (Oyekunle, 2025), wherein community knowledge is first gathered orally in the indigenous language and then recorded bilingually in the dataset, using both English and the native language. We leverage telephonic surveys with localized moderation in the respondent's preferred language, which allows a balance between broad reach and the ability to capture nuanced, authentic, and orally-expressed data.

2.2 Local Partnerships

This research study was conducted in collaboration with two organizations located in and operating within the countries of focus: Mantaray and RAIN Africa.[3,4] The recruitment of participants across communities was a collaborative effort between both of these organizations, with Mantaray also serving as the moderator for telephonic surveys in an array of languages commonly used in the countries of interest, and also preparing and translating the data. These partnerships were vital to ensure that local perspectives drive the salient questions of participant selection, demographic distribution, data quality, localization of survey questions, and more. Our study recruited 410 participants, at least 100 from each country. We employed a controlled quota sampling strategy based on the distribution of major ethnic groups within each country. We initiated the study with a screener to record demographic data and ensure adherence to the desired participant distribution.

[3] https://mantaray.africa/
[4] https://rainafrica.org/

2.3 Interview Protocol

We obtained informed consent from all participants and disclosed the monetary compensation amount (See details in Appendix A) beforehand. During the telephonic interview, participants were asked to share 10 to 15 prevalent stereotypes from their country (with a limit of three stereotypes per identity term to encourage diversity). They were also asked about the degree of offensiveness of the stereotype, its prevalence, and the axis of identity that the stereotype is about. These additional features will help identify the impact model failures can have when they perpetuate certain stereotypes. The survey in its entirety is shared in Appendix B.

2.4 Data Capture & Localization

Moderators were selected based on their fluency in the native languages as reported by the participants in the screener, in order to conduct interviews in each participant's preferred language. Despite this, we had to adopt a two-step approach that first captured the stereotypes in English which were then translated back to local languages by professional linguists. This was necessary because direct real-time transcription was often impractical (except for Swahili) due to linguistic and technical constraints. Moderators, while fluent speakers, often lacked the necessary written proficiency in native languages, a situation rooted in well-documented systemic challenges (Appiah and Ardila, 2021; Bamgboşe, 2000; Matavire, 2024; Mncwango and Sithole, 2017). Furthermore, standard electronic keyboards lack the specialized symbols and characters required for many languages from sub-Saharan Africa (Chimkono et al., 2021), and the unique phonetic signs inherent in these languages' orthographies cannot be represented by Roman characters.

Except for Swahili in Kenya, the moderators initially recorded all responses in written English, where applicable. To recover the original linguistic nuance, professional linguists were subsequently hired to translate these standardized English responses back into the participant's native language, or to translate Swahili entries to English for completeness of our English dataset.

Notes on Retention of Nuance While this multi-step localization process introduces some complexity and potential mediation (as noted under Limitations), it was a deliberate trade-off made to maximize language coverage across a diverse set of languages (Table 2), given systemic challenges and resource constraints. To mitigate the inevitable loss of nuance during this process, we prioritized interview moderation in participants' native languages and the involvement of professional linguists to ensure that the final dataset remained as faithful as possible to the original intent of the participants.

We offer this work as a foundational step and encourage future research to develop methods that can further minimize mediation and expand authentic, high-fidelity multilingual stereotype resources representing under-resourced African languages.

2.5 Data Preparation & Annotation

Following the initial data capture, the recorded stereotype statements were broken down and annotated into three key components as outlined in Table 1. A single identity term can relate to multiple axes. For instance, the term 'Hausa women' involves both the gender and ethnicity axes. Furthermore, identity axes can overlap. For example, the term 'Northerners' may indicate both a regional identity and a cluster of specific ethnicities associated with the Northern region of Ghana. Participants also provide additional social context about the stereotypes they share. They do so by rating each stereotype they shared on two perceived parameters: *Offensiveness*, rated on a 5-point scale (1 = Not Offensive; 5 = Extremely Offensive), and *Prevalence*, rated on a 4-point scale (1 = Rarely Used; 4 = Extremely Prevalent) (Appendix C). Our choice to use self-reported metrics for offensiveness and prevalence was an intentional effort to center the participant's voice, ensuring that the resulting dataset reflects a more granular and authentic range of perceptions regarding the stereotypes shared. We note that there will be inherent subjectivity in these ratings, which stem from individual calibrations of the assessment scales.

Name	Description	Example
Identity Term	A word or phrase used to describe a significant aspect of a personal or social identity.	Nigerians, Policemen, Zulu Women
Stereotype Attribute	The word associated with or used to describe the mentioned identity term.	Violent, Lazy, Ritualists, Liar
Identity Axis	Broader categories that group related identity terms.	Ethnicity, Race, Gender

Table 1: Key terms used in the SAFARI dataset

3 The SAFARI Dataset

We introduce SAFARI (sub-**S**aharan **AF**rican **A**ssociations about **R**egionally-salient **I**dentities), a dataset comprising 3,534 stereotypes in English and 3,206 stereotypes in 15 native languages (listed in Table 2), collected from 410 participants across four countries. Specifically, it contains 1,919 stereotypes tied to ethnic groups in Ghana, Kenya, Nigeria and South Africa, and 382 stereotypes related to regions or states within these countries, thereby highlighting its granular, local focus. We describe the distribution of the stereotypes multilingually in Table 2.

Country	Language	ISO 639-3	Count
Kenya	Kiswahili	swa	954
	Kalenjin (Kipsigis)	sgc	157
	Kikuyu	kik	132
	Luhya (Isukha)	ida	115
	Kenya Total		**1358**
Ghana	Twi	twi	305
	Dagbani	dag	211
	Ewe	ewe	202
	Ga	gaa	196
	Ghana Total		**914**
Nigeria	Igbo	ibo	184
	Hausa	hau	170
	Yoruba	yor	167
	Nigeria Total		**521**
South Africa	Zulu	zul	180
	Afrikaans	afr	86
	Xhosa	xho	75
	Sotho	sot	72
	South Africa Total		**413**
		Grand Total	**3206**

Table 2: Distribution of Multilingual Stereotypes by Country and Language

3.1 Dataset Characteristics

Diverse Participant Base The study cohort features a largely balanced gender distribution, with 52.4% Male (n=215), 46.8% Female (n=192), and 0.7% Non-Binary (n=3) participants. Age representation in the cohort spanned a wide range, from individuals aged 18-24 up to the 65+ age group. We particularly aimed for diversity across ethnicity, tribes, and linguistic identity, as these elements form the core social fabric of our target countries. Our study cohort collectively represents 56 ethnic groups and speaks over 120 unique languages (with the majority of participants being at least bilingual).

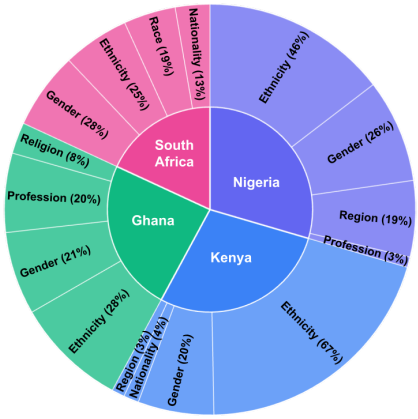

Figure 2: Country-wise distribution of dominant identity axes in the SAFARI stereotype dataset

Overlap with Other Resources Our dataset has minimal overlap with common stereotype resources, all of which exhibit geographical limitations: StereoSet (Nadeem et al., 2021), SPICE (Dev et al., 2023) and HESEIA (Ivetta et al., 2025) are limited to the perspectives from the USA, India, and Latin America, respectively. While SeeGULL Multilingual (Bhutani et al., 2024) covers Swahili (Kenya), the Swahili subset of SAFARI only shares four common identity terms with it; in fact, we introduce 168 new, unique identity terms, not previously covered. StereoSet does include seven African nationality-based identities, but the stereotypes were derived solely from US-based participants. In contrast, our resource features stereotypes shared by participants *from* sub-Saharan Africa.

Prominent Identity Axes: While ethnicity and gender remain the most dominant identity axes of stereotypes across the entire dataset, country-specific distributions reveal unique patterns, as depicted in Figure 2. In Nigeria, region is the third-most dominant identity axis after ethnicity and gender. Ghana's focus shifts to profession as its third major axis. The pattern in South Africa is unique, with race and nationality placing highly due to the country's distinct racial diversity.

3.2 Analysis and Evaluations

Common Stereotype Attributes: 128 stereotypes mentioning 'black magic,' 'witchcraft,' 'ritualistic', or 'juju' are found in the dataset, with 82.8% from Ghana and Kenya combined. These stereotypes primarily target ethnic groups, particularly the Ewe (Ghana), and the Kisii and Kamba (Kenya). 152 stereotypes related to financial status or display of wealth (with mentions of 'rich', 'love

Country	Gemini 2.5	GPT-4o	Claude Sonnet 4.5	Gemini 3 Pro	GPT-5.1	Claude Opus 4.5
Kenya (Swahili)	23.6 (19.8)	19.1 (16.8)	21.8 (20.6)	36.2 (38.8)	22.0 (17.0)	28.5 (26.9)
Ghana (Akan)	9.2 (9.6)	4.5 (7.8)	12.9 (27.2)	21.4 (34.4)	9.3 (19.5)	14.1 (29.5)
Nigeria (Hausa)	8.8 (17.0)	8.1 (9.3)	11.0 (16.0)	18.7 (30.7)	7.8 (16.8)	15.8 (30.2)
South Africa (Zulu)	5.7 (5.2)	2.5 (3.2)	7.2 (8.3)	16.0 (18.6)	4.5 (4.2)	12.1 (17.8)

Table 3: Rate of stereotype endorsement by industry leading models.The rate of stereotyping in English language is stated outside the braces and in the language mentioned within the braces respectively.

money', 'show off wealth', 'extravagant') emerge. 61.8% of these relate to ethnic groups, primarily associated with the Kikuyu (Kenya) and Igbo (Nigeria). Of the 327 profession-related stereotypes in the dataset, 74.6% emerge from Ghana, and most of these carry a negative connotation, targeting professions like police ('corrupt', 'take bribes'), politicians ('liars', 'corrupt'), and nurses ('unprofessional', 'disresepctful'). Out of the 1,034 gender-related stereotypes, 57.9% target women, and the remaining 42.1% target men.

Prevalence of Stereotypes in Generative AI: Stereotype datasets are used widely to both assess model skews and safety and to steer them towards fair answers. We thus test whether AI models endorse stereotypes from SAFARI in English or in languages from each of the four countries. We adopt the NLI based technique from Dev et al. (2020) for this assessment. In this test, each data point consists of two sentences: a premise and a hypothesis each of which invoke the attribute term and identity term respectively. The NLI based test checks if the model infers one from the other without sufficient context, which confirms model's tendency to stereotype. We create and employ human translators to carefully translate the NLI-based test templates from Dev et al. (2020) and evaluate the tendency to stereotype in generation (Jha et al., 2023) by Gemini 2.5 and Gemini 3 Pro (Comanici et al., 2025), GPT-4o and GPT 5.1 (OpenAI et al., 2024), and Claude Sonnet 4.5 and Claude Opus 4.5 (Anthropic). Table 3 describes our findings across all six frontier models. We see how it underscores that not only do models stereotype in the respective commonly used languages of each country, but also in English. In fact the extent of stereotyping is higher in English in many cases. We hypothesize that this discrepancy between the regionally popular language and English is stemmed in how the models do not particularly support the regional language and often fail to generate meaningfully. However, the overall prevalence of stereotyping across the board highlights the need to account for these stereotypes from underrepresented regions of the world to prevent potentially harmful generations.

4 Discussion and Future Work

We present the SAFARI dataset, that fills a critical data gap for AI safety in sub-Saharan Africa, while also contributing a novel methodology for culturally-situated data collection in NLP. Through a locally-situated and community-engaged approach, we collected over 3K stereotypes in English as well as 15 native languages from Ghana, Kenya, Nigeria, and South Africa. This dataset enables us to reveal previously undetected instances of stereotypes propagated by leading generative AI models, validating the urgent need for this work.

Beyond research evaluation, the SAFARI dataset can be plugged seamlessly into the many ways in which other West-focused stereotype resources are currently used, including for bias mitigation of models via counterfactual data generation (Zmigrod et al., 2019) and few shot debiasing, content moderation (Davani et al., 2023) and more. SAFARI can also be leveraged in approaches that semi automate stereotype curation based of few shot prompting (Bhutani et al., 2024).

Furthermore, our use of moderated, in-language telephonic interviews represents a novel data collection approach for NLP. This method provides a practical blueprint for navigating the complex techno-linguistic challenges prevalent in many global regions, effectively bypassing obstacles like low digital access and languages with orthographies unsupported by standard keyboards. Our work underscores the necessity of expanding data collection to more languages and global regions, while also accounting for such local realities, and we strongly encourage future research to build upon this foundation.

Limitations

Our work was aimed at improving the coverage of stereotype resources across languages prevalent in four countries in sub-Saharan Africa. While we attempt to carefully increase this coverage alongside maintaining demographic diversity of participants, we were limited by both our methodology and resources available to us. In particular, telephonic interviews allowed us to orally gather and record stereotypes, and thus reduce barriers related to writing, but we were still limited by connectivity issues. In particular, we may have failed to reach participants in remote areas who lack phones or reliable phone connectivity. For collecting perspectives from these specific, harder-to-reach groups, in-person surveys may be a better-suited methodology.

Our localization approach, described in Section 2, required a multi-step translation process, that may risk the loss of some subtle linguistic nuance and context as presented in the original spoken form. This focus on language preservation is critical, particularly because moderators observed that a stereotype's offensiveness was often diluted when translated into English, suggesting a valuable future research avenue concerning the loss of semantic and emotional nuance during cross-lingual transfer.

Further, our resource, while foundational for studying stereotypes in the sub-Saharan African context, is not exhaustive. It covers the major ethnic groups and native languages within our target countries, but omits many other ethnicities and languages due to limited sample size and resource constraints. Future research must expand upon this foundation to achieve comprehensive coverage of the numerous additional countries, languages, and ethnicities in sub-Saharan Africa. We recommend cross-referencing the findings from this expanded work with established traditional stereotype literature from the region (Fakunmoju and Bammeke, 2017; Harneit-Seivers, 2007; Lawson et al., 2015; Naituli and King'oro, 2018; Tewolde, 2024).

Ethical Considerations

Our research further highlights the high degree of linguistic complexity in sub-Saharan Africa. While our resource provides a significant subset of stereotypes towards ascertaining safety of generative models, they are in no way to be used to declare a model free of biases. For that, a lot more work and thought would be needed, including a much more exhaustive data collection. We also note that the recorded stereotypes should be only used for purposes of model evaluation and improvement as detailed in our Data Card (Appendix A) in order to prevent misuse.

Acknowledgements

First and foremost, we are extremely grateful to our data collection partners Mantaray and RAIN Africa, and their members and networks, whose expertise has been invaluable for this project. In particular, we thank Helga Stegmann, Liezel Stegmann, Michele Edwards, Partrick K. Wamuyu, Laeticia N. Onyejegbu, and Lavina Ramkissoon for sharing expert opinions. We thank all the moderators, data collection experts, project coordinators, and expert linguists who helped translate the stereotypes with socio-cultural grounding: Abiola Olaonipekun, Adisa Branice, Aisha Young, Alida Hillary, Ashleigh Burrell, Celeste Griesel, Darmack Kerubo, Deborah Agbeja, Dominic Abotchie, Emmanuel Mwadena, Feroza Khan, Geofrey Odhiambo, Gladys Allottey, Ilse Louw, John Ashihundu, Joyce Martins, Kennedy Mbithi, Kipkirui Ronald, Lindokuhle Modisane, Manoko Jegede, Margaret Muiruri, Martin Adjornor, Mary Obot, Mary Wangare, Mavis Dadzi, Naseega Adams, Nina Laubscher, Ogechi Anyanwu, Osumman Idrissu, Oyoe Quartey, Prince Danquah, Rahinatu Hamza, Richard Kueli, Samantha Lewis, Stephen Maina, Steve Amale, Sylvester Mukele, Victor Siele, and Yussif Abubakar. We are also very grateful to Debojit Ghosh, Charu Kalia, Ding Wang, Roxane Ghinet, Kwaku Agbesi, Taylor Montgomery, Ronnelle Castro, Jamelle Louise Javillonar, Jamila Smith-Loud, and Saška Mojsilović for their partnership, generous feedback, and advice. Finally, we also thank the reviewers and ACs at ARR and EACL for their very constructive feedback and discussion.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Abejide Ade-Ibijola and Chinedu Okonkwo. 2023. Artificial intelligence in africa: Emerging challenges. In *Responsible AI in Africa: Challenges and opportunities*, pages 101–117. Springer International Publishing Cham.

Ife Adebara and Muhammad Abdul-Mageed. 2022. Towards afrocentric NLP for African languages: Where we are and where we can go. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3814–3841, Dublin, Ireland. Association for Computational Linguistics.

David Ifeoluwa Adelani, Jesujoba Oluwadara Alabi, Angela Fan, Julia Kreutzer, Xiaoyu Shen, Machel Reid, Dana Ruiter, Dietrich Klakow, Peter Nabende, Ernie Chang, Tajuddeen Gwadabe, Freshia Sackey, Bonaventure F. P. Dossou, Chris Emezue, Colin Leong, Michael Beukman, Shamsuddeen H. Muhammad, Guyo D. Jarso, Oreen Yousuf, and 26 others. 2022. A few thousand translations go a long way! leveraging pre-trained models for African news translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3053–3070, Seattle, United States. Association for Computational Linguistics.

Ibrahim Said Ahmad, Shiran Dudy, Resmi Ramachandranpillai, and Kenneth Church. 2024. Are generative language models multicultural? a study on hausa culture and emotions using chatgpt. *Preprint*, arXiv:2406.19504.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, and 1 others. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Anthropic. The claude 3 model family: Opus, sonnet, haiku.

Samuel Opoku Appiah and Alfredo Ardila. 2021. The dilemma of instructional language in education: The case of ghana. *Hungarian Educational Research Journal*, 11(4):440–448.

Ayọ Bamgboşe. 2000. *Language and Exclusion: The Consequences of Language Policies in Africa*. Contributions to Studies on African Languages and Literatures. Lit Verlag, Münster ; London.

Mukul Bhutani, Kevin Robinson, Vinodkumar Prabhakaran, Shachi Dave, and Sunipa Dev. 2024. SeeGULL multilingual: a dataset of geo-culturally situated stereotypes. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 842–854, Bangkok, Thailand. Association for Computational Linguistics.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *Preprint*, arXiv:2110.01963.

Aaron Chatterji, Tom Cunningham, David Deming, Zoë Hitzig, Zoe Hitzig, Christopher Ong, Carl Shan, and Kevin Wadman. 2025. How people use chatgpt. *NBER Working Paper*.

Thokozani Chimkono, Eunice Mphako-Banda, Amelia Taylor, and Pascal Kishindo. 2021. A usability study of the central-bantu multilingual keyboards. *International Journal of Human–Computer Interaction*, 37(16):1489–1503.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Aida Mostafazadeh Davani, Sunipa Dev, Héctor Pérez-Urbina, and Vinodkumar Prabhakaran. 2025. A comprehensive framework to operationalize social stereotypes for responsible AI evaluations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30030–30043, Suzhou, China. Association for Computational Linguistics.

Sunipa Dev, Jaya Goyal, Dinesh Tewari, Shachi Dave, and Vinodkumar Prabhakaran. 2023. Building socioculturally inclusive stereotype resources with community engagement. *Preprint*, arXiv:2307.10514.

Sunipa Dev, Tao Li, Jeff M. Phillips, and Vivek Srikumar. 2020. On measuring and mitigating biased inferences of word embeddings. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7659–7666.

Lindsey DeWitt Prat, Olivia Nercy Ndlovu Lucas, Christopher Golias, and Mia Lewis. 2024. Decolonizing llms: An ethnographic framework for ai in african contexts. *EPIC Proceedings*, 2024(1):45–84.

Sunday B Fakunmoju and Funmi O Bammeke. 2017. Gender-based violence beliefs and stereotypes: Cross-cultural comparison across three countries. *International Journal of Asian Social Science*, 7:738–753.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, and 17 others. 2022. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *Preprint*, arXiv:2209.07858.

Axel Harneit-Seivers. 2007. Igbo history and society: The essays of adiele afigbo, edited by toyin falolamyth, history and society: The collected works of adiele afigbo, edited by toyin falola. *African Affairs*, 106:529–531.

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahamed, Ibrahim Said Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025. Automatic speech recognition for african low-resource languages: Challenges and future directions. *Preprint*, arXiv:2505.11690.

Guido Ivetta, Marcos J. Gomez, Sofía Martinelli, Pietro Palombini, M. Emilia Echeveste, Nair Carolina Mazzeo, Beatriz Busaniche, and Luciana Benotti. 2025. Heseia: A community-based dataset for evaluating social biases in large language models, codesigned in real school settings in latin america. *Preprint*, arXiv:2505.24712.

Yacine Jernite, Huu Nguyen, Stella Biderman, Anna Rogers, Maraim Masoud, Valentin Danchev, Samson Tan, Alexandra Sasha Luccioni, Nishant Subramani, Isaac Johnson, Gerard Dupont, Jesse Dodge, Kyle Lo, Zeerak Talat, Dragomir Radev, Aaron Gokaslan, Somaieh Nikpoor, Peter Henderson, Rishi Bommasani, and Margaret Mitchell. 2022. Data governance in the age of large-scale data-driven language technology. In *2022 ACM Conference on Fairness Accountability and Transparency*, FAccT '22, page 2206–2222. ACM.

Kyle Jerro. 2018. Linguistic complexity: A case study from swahili.

Akshita Jha, Aida Davani, Chandan K. Reddy, Shachi Dave, Vinodkumar Prabhakaran, and Sunipa Dev. 2023. Seegull: A stereotype benchmark with broad geo-cultural coverage leveraging generative models. *Preprint*, arXiv:2305.11840.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022. Handling and presenting harmful text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Victoria Wendy Lawson, Charity S. Akotia, and Maxwell Asumeng. 2015. Exploring Ethnic Stereotypes and Prejudice Among Some Major Ethnic Groups in Ghana. *Journal of Social Science Studies*, 2(1):17–35.

Juniel Shoko Matavire. 2024. Development of indigenous language orthographies: Setting up english as the torch bearer. *International Journal of Critical Diversity Studies*, 6(2).

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, and 35 others. 2025. SHADES: Towards a multilingual assessment of stereotypes in large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11995–12041, Albuquerque, New Mexico. Association for Computational Linguistics.

Elliot Mthembeni Mncwango and Nkosinathi Emmanuel Sithole. 2017. A diminished role of indigenous african languages in south african institutions. *International Journal of Humanities and Social Science*, 7(1):85–92.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, David Ifeoluwa Adelani, Ibrahim Said Ahmad, Saminu Mohammad Aliyu, Paul Röttger, Abigail Oppong, Andiswa Bukula, Chiamaka Ijeoma Chukwuneke, Ebrahim Chekol Jibril, Elyas Abdi Ismail, Esubalew Alemneh, Hagos Tesfahun Gebremichael, Lukman Jibril Aliyu, Meriem Beloucif, Oumaima Hourrane, Rooweither Mabuya, Salomey Osei, and 8 others. 2025. AfriHate: A multilingual collection of hate speech and abusive language datasets for African languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1854–1871, Albuquerque, New Mexico. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Joseph Gitile Naituli and Sellah Nasimiyu King'oro. 2018. An examination of ethnic stereotypes and coded language used in Kenya and its implication for national cohesion. *International Journal of Development and Sustainability*, 7(11):2670–2693.

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, and 28 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*, pages 2144–2160, Online. Association for Computational Linguistics.

Jessica Ojo, Odunayo Ogundepo, Akintunde Oladipo, Kelechi Ogueji, Jimmy Lin, Pontus Stenetorp, and David Ifeoluwa Adelani. 2025. AfroBench: How good are large language models on African languages? In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19048–19095, Vienna, Austria. Association for Computational Linguistics.

Alexander Oluka. 2024. Mitigating biases in training data: Technical and legal challenges for sub-saharan africa. *International Journal of Applied Research in Business and Management*, 5(1):209–224.

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, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Nedjma Ousidhoum, Meriem Beloucif, and Saif M. Mohammad. 2025. Building better: Avoiding pitfalls in developing language resources when data is scarce. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8881–8894, Vienna, Austria. Association for Computational Linguistics.

Akinpelu Ayokunnu Oyekunle. 2025. Exploring "linguistic complementarity" for intercultural communication in postcolonial african states. *Frontiers in Communication*, 10.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Alicia Parrish, Hannah Rose Kirk, Jessica Quaye, Charvi Rastogi, Max Bartolo, Oana Inel, Juan Ciro, Rafael Mosquera, Addison Howard, Will Cukierski, D. Sculley, Vijay Janapa Reddi, and Lora Aroyo. 2023. Adversarial nibbler: A data-centric challenge for improving the safety of text-to-image models. *Preprint*, arXiv:2305.14384.

Qazi Mamunur Rashid, Erin van Liemt, Tiffany Shih, Amber Ebinama, Karla Barrios Ramos, Madhurima Maji, Aishwarya Verma, Charu Kalia, Jamila Smith-Loud, Joyce Nakatumba-Nabende, Rehema Baguma, Andrew Katumba, Chodrine Mutebi, Jagen Marvin, Eric Peter Wairagala, Mugizi Bruce, Peter Oketta, Lawrence Nderu, Obichi Obiajunwa, and 3 others. 2025. Amplify initiative: Building a localized data platform for globalized ai. *Preprint*, arXiv:2504.14105.

Andrew Smart, Ben Hutchinson, Lameck Mbangula Amugongo, Suzanne Dikker, Alex Zito, Amber Ebinama, Zara Wudiri, Ding Wang, Erin van Liemt, João Sedoc, Seyi Olojo, Stanley Uwakwe, Edem Wornyo, Sonja Schmer-Galunder, and Jamila Smith-Loud. 2024. Socially responsible data for large multilingual language models. *Preprint*, arXiv:2409.05247.

Mesmin Tchindjang, Athanase Bopda, and Louise Angéline Ngamgne. 2008. Languages and cultural identities in africa. *Museum International*, 60(3):37–50.

Amanuel Isak Tewolde. 2024. Experiencing negative racial stereotyping: The case of coloured people in johannesburg, south africa. *Social Sciences*, 13:277.

Youssef Travaly and Kevin Muvunyi. 2020. The future is intelligent: Harnessing the potential of artificial intelligence in africa. The Brookings Institution.

Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

Appendix

A Dataset Details

A.1 Data

The dataset is released at https://github.com/google-research-datasets/SAFARI, and is accompanied by the data card, which specifies intended usage, data collection details, and more.

The dataset is structured into two tabs:

- The first tab, 'safari_english', contains 3,534 rows (excluding the header), each containing a stereotype and its corresponding annotations
- The second tab, 'safari_multilingual', includes 3,206 rows (excluding the header), each containing a stereotype from the multilingual subset of the dataset, covering 15 languages (following the distribution outlined in Table 2), along with corresponding annotations and the English counterpart for cross-reference.

The difference in the number of English (n=3,534) and multilingual (n=3,206) stereotypes is attributable to two factors: first, English-language

interviews were not translated into a native language when English was the participant's preferred language; and second, in cases of less commonly spoken languages with limited linguist availability.

A.2 Dataset Sample

Below is a sample data entry to illustrate key parameters recorded in the dataset:

- **Country:** Kenya
- **Stereotype Sentence (English):** *Mijikenda people are lazy*
- **Identity Term (English):** Mijikendas
- **Stereotype Attribute (English):** Lazy
- **Identity Axis (English):** Ethnicity
- **Offensiveness (1-5):** 3 - Slightly offensive
- **Prevalence (1-4):** 3 - Commonly used
- **Target Language:** Kiswahili
- **Identity Term (Kiwsahili):** Wamijikenda
- **Stereotype Attribute (Kiswahili):** Wavivu

A more detailed version is presented in the data card.

A.3 Data Cleaning Steps

Our data cleaning process was conservative and geared towards liberal inclusion of contributions from different participants. We focused solely on performing basic sanity checks to ensure the consistency needed for quantitative analyses, while deliberately avoiding extensive data tampering that could lead to further loss of contextual nuance collected from the field. No special software was used for this purpose. We cleaned the data manually and for aggregation purposes alone.

To clean the survey responses for precise counting and reporting, the following manual edits were applied:

- Instances of "woman" and "women" were standardized to "women". Similarly, "Igbos" and "Igbo" (when referring to the people) were standardized to "Igbos", and other identity terms were treated similarly.
- Language names were standardized (e.g. 'Kiswahili' and 'Swahili' updated to 'Kiswahili'; 'isiZulu' and 'Zulu' updated to 'isiZulu', and so on).
- Other minor cleaning steps included correcting spelling errors and trimming extra white space within cells to prepare the data for usage and analysis.

A.4 Review, Consent & Compensation

Dataset and data collection were reviewed by an internal, organizational, proprietary review board.

Prior to the study, informed consent was obtained from all participants, who confirmed they met the minimum age requirement (18+). The consent form explicitly stated their right to withdraw from the study at any time without penalty, and the option to decline any specific question asked to them. They were briefed on the study's purpose and the procedures for data retention, storage, and protection. In addition to the written consent form, the moderators ensured participants received these details verbally.

The dataset does not contain any Personally Identifiable Information (PII). All PII was scrubbed during collection. Unique, non-traceable identifiers (e.g., GH001, KE002, etc) were used solely for analysis and data counting purposes, but these cannot be used to trace or link back to the participating individuals.

For completing the telephonic survey, each participant was compensated in a fair and legally compliant manner with an amount that exceeded minimum local wage requirements. The incentive amount was mentioned in the screener shared beforehand. Duration of the telephonic interview was capped at 15 minutes per participant.

The two local organizations we partnered with were compensated either according to a pre-established Statement of Work (SOW) for services provided (in one case), or through a mutual agreement on the amount and deliverables (in the other). We will name both organizations and co-authors in the camera-ready version.

B Interview Design and Methodology

Section 1: Screener Demographic Questions

The following information was collected during the initial screening phase:

1. Country of Data Collection
2. Gender Identity
3. Age
4. Hometown / Region

5. Religion
6. Ethnicity / Tribe
7. Sub-Tribe (where applicable)
8. Language(s) Spoken
   (a) Primary (Native)
   (b) Secondary
   (c) Other
9. Highest Level of Completed Education
10. Occupation

Section 2: Stereotype Elicitation Questions

Respondents were asked to share at most 10-15 stereotypes known in their society and community (as much as they could comfortably list within the limit of a 15-minute interview).

1. **Question:** What stereotypes do you know in your society and community?
   (a) **Recording:** The stereotype was recorded, first as a sentence, and then broken down to an *identity term* and an *attribute term* (e.g., Nigerians, rich).
   (b) **Axis of Identity:** The identity term's axis was classified (e.g., gender, tribe, religion, language, race, age, profession).
   (c) **Offensiveness Rating:** Respondents were asked to rate the stereotype's offensiveness using a 1-5 scale (presented in Appendix C, Table 4)
   (d) **Prevalence Rating (Optional):** Respondents rated the stereotype's prevalence using a 1-4 scale (presented in Appendix C, Table 5)
2. **English Translation:** The closest English translation was provided by the moderator post-interview, not during the collection process.

Collection Guidelines

To ensure diversity of responses, the following limits were applied per participant:

- Not more than 3 stereotypes about any single identity term (e.g., Hausas).
- Not more than 5 stereotypes relating to any single identity axis (e.g., profession).

Rating (Score)	Definition
1	Not Offensive at all
2	Unsure
3	Slightly Offensive
4	Somewhat Offensive
5	Extremely Offensive

Table 4: Offensiveness Rating Scale

C Rating Scales

Participants were asked to rate the offensiveness of each shared stereotype on a 5-point scale (Table 4), and its prevalence on a 4-point scale (Table 5).

Rating (Score)	Definition
1	Rarely Used
2	Occasionally Used
3	Commonly Used
4	Extremely Prevalent/Popular

Table 5: Prevalence Rating Scale

D Data Details

D.1 Participant Diversity

The total number of stereotypes and participants per country is described in Table 6. The participants of our data collection effort were diverse along different demographic axes. Table 7 describes the gender distribution of our participants per country.

Country	Number of Stereotypes	Number of Respondents
Ghana	1037	101
Kenya	892	108
Nigeria	954	100
South Africa	651	101
Total	**3534**	**410**

Table 6: Data Collection Summary by Country

D.2 Language Distribution

The language distribution of the multilingual stereotype dataset is shown in Figure 3 and Table 2.

D.3 Demographics by Country

While our sampling strategy aimed to reflect the national ethnic distributions of each country, the

Country	Gender	Percentage (%)
Ghana	Male	54%
	Female	46%
Kenya	Female	50%
	Male	49%
	Non-binary	1%
Nigeria	Male	58%
	Female	42%
South Africa	Female	50%
	Male	49%
	Non-binary	2%

Table 7: Gender Distribution of Study Cohort by Country

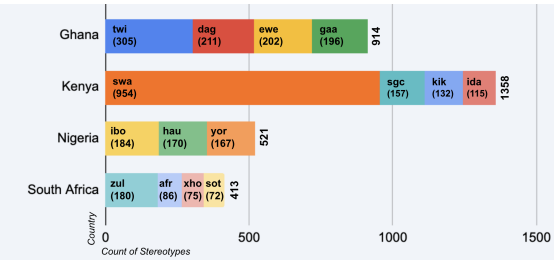

Figure 3: The SAFARI dataset represents 15 native languages from 4 countries. Table 2 provides the expanded ISO codes.

limited sample size necessitated a focus on ensuring broad diversity rather than exact proportional representation. We rely on self-identified ethnicity data as provided by participants during the screening process. The distribution of ethnic groups and tribes in the study cohort is detailed in the following country-wise tables: Table 8 (Ghana), Table 9 (Kenya), Table 10 (Nigeria) and Table 11 (South Africa).

Table 8: Distribution of Self-Identified Ethnicity of Participants in Ghana

Ethnicity / Tribe	Count	%
Akan	31	30.7
Ewe	20	19.8
Ga-Adangbe	19	18.8
Mole-Dagbani	17	16.8
Northern tribe	3	3.0
Wala	2	2.0
Kusasi	2	2.0
Frafrah	2	2.0
Kotokoli	1	1.0
Kassim	1	1.0
Hausa	1	1.0
Guan	1	1.0
Gonja	1	1.0
Total	**101**	**100.0**

Table 9: Distribution of Self-Identified Ethnicity of Participants in Kenya

Ethnicity / Tribe	Count	%
Kikuyu	19	17.6
Kalenjin	16	14.8
Luhya	15	13.9
Kamba	13	12.0
Mijikenda	11	10.2
Luo	11	10.2
Kisii	11	10.2
Meru	4	3.7
Maasai	3	2.8
Embu	3	2.8
Taita	1	0.9
Arab	1	0.9
Total	**108**	**100.0**

Table 10: Distribution of Self-Identified Ethnicity of Participants in Nigeria

Ethnicity / Tribe	Count	%
Hausa	22	22.0
Yoruba	20	20.0
Igbo	20	20.0
Ijaw	10	10.0
Tiv	5	5.0
Kanuri	5	5.0
Ibibio	5	5.0
Fulani	5	5.0
Urhobo	1	1.0
Ogoja	1	1.0
Igala	2	2.0
Benin	1	1.0
Nupe	1	1.0
Ekoi	1	1.0
Ebira	1	1.0
Total	**100**	**100.0**

Table 11: Distribution of Self-Identified Ethnicity of Participants in South Africa

Ethnicity / Tribe	Count	%
Zulu	25	24.8
Xhosa	12	11.9
Tswana	10	9.9
Afrikaans	10	9.9
Pedi	9	8.9
Basotho	7	6.9
Tsonga	5	5.0
Sesotho	4	4.0
Venda	3	3.0
Coloured	3	3.0
Ndebele	2	2.0
Koi-San	2	2.0
Indian-Asian	2	2.0
English	2	2.0
Bantu	2	2.0
Xitsonga	1	1.0
Ndele	1	1.0
African	1	1.0
Total	**101**	**100.0**

E AI Usage in Writing

We used a conversational AI tool to help lightly rephrase or shorten a few sentences, find alternate terms, and improve table formatting.