epic 2024 Ethnographic Praxis in Industry Conference Proceedings

# Decolonizing LLMs: An Ethnographic Framework for AI in African Contexts

LINDSEY DEWITT PRAT, *Bold Insight*
OLIVIA NERCY NDLOVU LUCAS, *Mantaray Africa*
CHRISTOPHER GOLIAS, *Google*
MIA LEWIS, *Independent Scholar*

*This paper examines LLM deployment through African lenses that engage content's complex social, technological, and linguistic landscape. We propose an adaptable framework for LLM research based on an extensive synthesis of research literature and our primary research in Ethiopia, Ghana, Kenya, Nigeria, and South Africa. Our work unearths instances where LLMs could perpetuate digital colonialism or exacerbate existing sociopolitical tensions, as well as how they are contested and adapted. We emphasize the imperative for an embodied ethnographic approach that engages the inherently fluid, flexible, and multicultural nature of language, and connects LLM technologies with the complex social contexts in which it is built and deployed.*

## Introduction

*In Addis Ababa, Ethiopia, a Gen-Z student is assigned to research local folktales for a heritage and culture project. After briefly searching online and asking around their community with limited success, the student turns to a digital tool powered by a large language model (LLM). They use it in English because it doesn't work well in Amharic, one of their first languages. The LLM generates a story about Abba Otho, a supposed folk hero from the Amhara region known for his bravery and selflessness. The student includes this narrative in their project and receives a failing grade: neither the story nor the character exists in Ethiopian folklore. The LLM had fabricated a convincing yet entirely fictional tale, one infused with a Westernized concept of heroism.*

This hypothetical example, derived from an experiment with a widely available LLM and local insights from Ethiopian researchers and community members, brings into relief the stakes of language technology today and some of the potential hazards for language and culture: the erosion of linguistic diversity, the spread of misinformation, and the loss of traditional knowledge transmission methods, to name a few. This paper aims to help researchers anticipate and diffuse some of the risks posed by LLMs and to support culturally informed deployments that enhance local agency. We adopt an African optic to synthesize observations and, grounding them in primary research, propose a two-phased framework to help guide ethnographic work in the area.

Dewitt Prat, Lindsey, Olivia Nercy Ndlovu Lucas, Christopher Golias, and Mia Lewis. "Decolonizing LLMs: An Ethnographic Framework for AI in African Contexts." EPIC Proceedings (2024). 45–84. https://www.epicpeople.org/decolonizing-llms-ethnographic-framework-for-ai-in-african-contexts. ISSN 1559-8918. License: CC-BY-NC-ND.

Note that when we refer to "LLM research in Africa," we primarily mean using an ethnographic lens to study widely available LLMs (i.e., post-release) by considering the contextual interactions between real people and these technologies.

Language is one of the principal ways humans communicate, make meaning, and embed individual identities within communities. Until recently, computers could not reliably create linguistic content at a level comparable to native, fluent, human speakers. That changed in late 2022 with improvements to OpenAI's ChatGPT, a type of "generative artificial intelligence" (AI) called an LLM. These models are advanced deep learning algorithms, a subset of machine learning (ML) that can learn from its own errors and process vast amounts of data. LLMs can perform a variety of natural language processing (NLP) tasks, such as machine translation (MT), as well as other language-related tasks (e.g., text generation, summarization, question answering).

Our languages mold our understanding of the world (Koerner, 1992), and our technologies embody the assumptions of their makers (Tamkin et. al., 2023). Language is inherently fluid, flexible, and multicultural. Today's largest and most widely used LLMs are not. They are overwhelmingly trained by teams located in the US (Talat et al., 2022), and trained to operate on unspecified variants of English text (Bender, 2019).[1] The global deployment of LLMs, a product of WEIRD—Western, Educated, Industrialized, Rich, and Democratic—Silicon Valley (Henrich, 2020), therefore risks perpetuating its creators' unconscious biases and blind spots and presenting their worldview as the worldview to a global user base expecting results from an unbiased computer program.

But what about the rest of the world's population, whose written and spoken language repertoires do not include English? And what about those aspects of culture, language, and experience that remain undigitized? That have not yet been disembodied? The human operating system has thousands of coding languages, so to speak, and early attempts at localizing LLMs to other languages have produced mixed results. Many researchers practicing in the technology industry actively engage the challenges of conducting ethnography and user experience (UX) research with the goal of improving the technologies and practices enabled by LLMs. As this push broadens to other languages, there is a growing urgency to examine how these technologies are, or might be, adapted to diverse linguistic and cultural contexts, specifically African ones.

Many African languages face numerous challenges for language technology development, and are therefore sometimes classified as low-resource languages

(LRLs). The challenges they face include limited text corpora for digital computation, annotated speech data, linguistic annotations (e.g., part-of-speech tagging, named entities), standardized writing systems, and parallel text data for translation (Adelani et al., 2021; Magueresse et al., 2020). Broader resource limitations, such as funding, infrastructure, human expertise, and educational support, further impede progress (Leong et al., 2023; Nekoto et al., 2020). The lack of adequate natural language data online exacerbates the issue, and the text that is available often does not reflect everyday language use (Zupon et al., 2021). Consequently, most African languages receive little to no support from widely available technological tools and the support that does exist is of varying quality.

The number of languages supported by available tools remains extremely limited compared to Africa's linguistic diversity. The Niger-Congo language family, one of the largest commonly cited groups in the world, includes some 1500 languages (Good, 2020). Machine translation tools with the most expansive language support (e.g., Google Translate, Niutrans, Alibaba Translate) cover less than fifty African languages as of June 2024 (Machine Translate, n.d.). This limited support, even by some of the programs and apps that work in the largest number of languages, stands to further emphasize the dominance of English in the digital world. For those who simply use the internet to find and consume information, just over half of all websites are written in English, with Russian being the second most prevalent language at a mere 5% (Kemp, 2024, p. 107; Web Technology Surveys, n.d.). The only African language without modern colonial roots that makes up greater than .1% of website content today is Arabic. Yet Arabic has a long and layered history on the continent, shaped by various waves of contact and imposition, and some communities in North Africa and South Sudan regard Arabic as a colonial language.

The example of Arabic in Africa, or African Arabic, raises important questions about what constitutes an African language and the role of self-determination in linguistic decolonization—topics we'll explore in this paper. African self-representations and identities play a decisive role in the discussion as well. For instance, many North Africans identify more as Arab or Arab-Muslim than as "African." This single snapshot opens our eyes to Africa's layered partitionings—linguistic, geographic, ethnic, national, racial—and their interconnections across time and space (Aidi et al., 2021). African experiences are not monolithic or one-dimensional when it comes to language or anything else.

The contrast between pluralistic, multidimensional, multilingual Africa and today's monocultural, English-centric LLMs is stark. As one of the most linguistically diverse continents, Africa looms as a challenge, a gut check, and an opportunity for both ethnography and LLM technology. Ethnography's embodied engagement,

through direct interaction with people and researcher reflexivity, uniquely connects LLM technology with the complex social contexts it is built from and deployed into.

## A Global and African Overview of Cultural Awareness in LLM Research

The challenges LLMs pose regarding cultural representation and bias are already a major topic of interest for researchers, industry professionals, and various other stakeholders. Recent efforts in this space employ diverse frameworks and heuristics, with the shared goal of creating language technologies that serve the needs of all users, regardless of their cultural or linguistic background. Some initiatives specifically aim to understand the convergence of scale and power asymmetry posed by the global deployment of LLMs. Our review of current academic and gray literature in this dynamic field can be broadly categorized into three areas: (1) evaluative studies that identify and analyze cultural bias in LLMs and related AI systems beyond Western contexts; (2) efforts to develop technical solutions to enhance cultural representation in LLMs and language technologies; and (3) initiatives promoting AI education and responsible governance. While many research and development efforts have a global scope, there is a growing body of work specifically addressing African contexts, recognizing (as we do) the continent's unique linguistic and cultural landscapes and the need for contextualized solutions. We situate our research within the first two areas.

Evaluative studies published in the past year lay bare the significant gaps in cultural awareness and representation in LLMs and related AI systems as a first step towards redressing them. Studies assess LLM capabilities across cultures, languages, and geographies (e.g., Ahuja et al., 2024; Liu et al., 2024; Masoud et al., 2024; Moayeri et al., 2024; Shafayat et al., 2024; Watts et al., 2024), as well as cultural competence in vision-language models (e.g., Karamolegkou et al., 2024) and text-to-image models (e.g., Kannen et al., 2024). In specific sectors like healthcare, researchers examine equity-related biases in AI applications globally (e.g., Pfohl et al., 2024) and in Africa specifically, as intersected by contextualized axes of disparities, including colonialism (Asiedu et al., 2024).

Technical solutions focus on creating culturally rich datasets and benchmarks for LLMs and related language models. Global initiatives range from stereotype benchmarks (Jha et al., 2023) to solutions for integrating cultural differences through LLM augmentation (Li et al., 2024) and datasets covering thousands of ethnolinguistic groups (Fung et al., 2024). Some studies, like ours, push towards more meaningful participation, addressing issues like the expropriation of intellectual

property through collaborative dataset creation (Nayebare et al., 2023; Suresh et al., 2024). A growing network of research labs, open-source projects, and AI initiatives is strengthening the foundation for LLM research and technology relevant to African languages and cultures. While not all exclusively focused on LLMs, these efforts develop crucial language resources and tools, many crowd-sourced, that support culturally aware and linguistically diverse AI development. Examples include the pan-African open-source project Masakhane, GhanaNLP, Digital Umuganda in Rwanda, and EthioLLM in Ethiopia (Tonja et al., 2024). Other important contributors include the Distributed AI Research Institute (DAIR), Algorine, and Lelapa AI. These initiatives, through various approaches, collectively advance ML and NLP capabilities that underpin LLM development for and about African contexts.

The third category in our review encompasses governance and education initiatives. Collaborative frameworks bringing together researchers in academia and industry, civil society, and governments are establishing guidelines for inclusive AI governance (African Union, 2024; AUDA-NEPAD, 2023; O'Neill et al., 2024; UNESCO & Moroccan International Centre for Artificial Intelligence, 2024). While these frameworks often address AI broadly, they have significant implications for LLM development and use. Examples of educational efforts in Africa include AI literacy research (e.g., Oyewusi, 2024) and gatherings like the Deep Learning Indaba and the African Computer Vision Summer School. Alongside these efforts are government-backed initiatives, such as those in Nigeria (Nigeria Federal Ministry of Communications, Innovation & Digital Economy, 2024), focused on creating nationally relevant AI capabilities, often partnering with local industry and research talent to realize that vision.

Our study builds upon and complements this rich range of ongoing efforts through our focus on ethnography at the intersection of LLMs, language use, and their cultural impact in African settings.

## Paper Organization

The paper has two parts: a research paper and an adaptable framework for conducting LLM research in Africa. The research section provides context for the framework through four key areas. We begin with an introduction to Africa's techno-linguistic terrain and an exploration of the relationship between technology and language in Africa. We then examine African decolonization theory as it relates to language. Finally, we present selected research examples from UX studies conducted in Africa between 2023 and 2024. Our insights draw from collaborative projects and a targeted survey of African UX researchers, which we conducted in May 2024 specifically to inform this paper.

The two-part framework follows. We start with research planning, emphasizing reflexivity, historical awareness, and co-creation with African perspectives, all underscored by ethical practices such as informed consent and data governance. Then we move to research execution, outlining practical steps for data collection and analysis within a sustainable, mutually beneficial framework that emphasizes and integrates local expertise and inclusivity. By advancing ethnographic approaches that engage with Africa's unique sociocultural and technological landscapes, we aim to promote research that is thoughtful, collaborative, supports local agency, and contributes to the ethical, culturally relevant, and sustainable implementation of LLM technologies—an approach we frame as decolonizing.

## Africa's Techno-linguistic Terrain, Agency, and Innovation with LLMs

Africa pulses with linguistic diversity, hosting over 2,000 living languages across more than fifteen language families (Ethnologue, 2022). Among these, seventy-five African languages are said to number at least one million speakers. Nigeria reportedly has over 500 spoken languages and Ethiopia more than eighty, though these figures and language classifications are dynamic and open to interpretation (Mwinlaaru & Xuan, 2016) and have been understudied in the case of Africa (Lüpke & Storch, 2013). The South African constitution recognizes eleven official languages, with dozens more living alongside them.

Africa's rich linguistic diversity parallels, and arises from, its geography. The savanna grasslands of Sudan and the South African fynbos. The tropical rainforests of the Congo Basin. The vast Sahara stretching across all of northern Africa. The patchwork of Afromontane forest enclaves in the West and that follow the East African rift systems from the Red Sea down to the Great Escarpment of the continent's southern tip. African lands, like the continent's languages, defy generic characterization and insist on multiplicities and interrelations.

Similarly, Africa's technological landscapes, especially its digital realms, are a vast system of interconnected yet distinct segments. These segments are characterized by pockets of high innovation alongside regions of technological absence, and they form a dynamic whole often subject to unpredictable disruptions. Thinking about a system of "connectivity rifts" captures the uneven and unequal distribution of digital infrastructure and access, where advanced technologies like LLMs coexist with areas lacking internet connectivity entirely. In sub-Saharan Africa, about 600 million

people (more than half of the region's population) lack access to electricity, and many more have limited or unreliable power (United Nations, 2023).

Viewed from one perspective, we can understand Africa in sociologist Immanuel Wallerstein's terms as a "periphery"—a place that provides resources for lucrative technologies and also a market for selling them (Eades, 2012). This dynamic creates continuous cycles of enrichment and benefit for the "core" countries that extract the resources and sell the technologies. For centuries, Africa has been both a destination and resource for global (non-African) technological aspirations. Take cobalt as an example. Cobalt is key to lithium-ion batteries and many other technologies, and 70% of current production occurs in the Democratic Republic of the Congo (Gulley, 2022). While Western companies have historically dominated the extraction and utilization of African resources, recent shifts have seen significant Chinese investment. Currently, 80% of the DRC's cobalt output is owned by Chinese companies. Numerous reports detail human rights abuses in its mining, including child labor and forced labor, causing harm rather than providing benefit to some Congolese people (United States, 2023). Meanwhile, those who rely on informal cobalt mining to get by often live in extreme poverty (World Economic Forum, 2020).

And yet, Africa is fundamentally a core of technological innovation and adaptation, and has been for many millennia. A historical snapshot of the continent features myriad instances of technological development: innovation, importation, assimilation, adaptation, and imposition in the fields of textiles, manufacturing, mining, marketing, transport, currency, commodity exchange, and so much more. These historical systems laid the groundwork for today's wealth of tech hubs, startups, and platforms that, as ever, "create synergies between imported and locally invented modes of innovation and entrepreneurship" (Mavhunga, 2017, p. 18). Writing about the incorporation of new technologies into hunting practices among the vaShona and maTshangana in Zimbabwe, science and technology historian Clapperton Chakanetsa Mavhunga compels us to see users as designers and innovators full of creative resilience, rather than merely recipients of technology and innovation (Mavhunga, 2014). Such a lens allows us to see the capacities people already have (and have always had) to engage and deploy technologies, and in doing so to assign those technologies new meanings and purposes.

It is a near-truism that people shape and are shaped by technologies. Technologies are driven by human "passions and politics and pitiful calculations" (Latour, 1996, p. viii), and heavily influenced by broader structural and institutional factors. Technologies evolve through cycles of experimentation, development, socialization, consumption, and consequence. Advancements in hardware and

software are resisted and embraced in an ongoing negotiation between resources, innovation, power, and the multiple, overlapping contexts—social, cultural, geographic—that both enable and constrain their adoption (Williams, 1974). Technology's integration into African life, life anywhere on the planet, is contingent and contextual. So too is it open-ended, with multiple uses and outcomes that depend on agency.

Historically, technologies have served as tools of production and pleasure, and as instruments of domination and displacement—sometimes all at once. Contemporary views often equate technology with tools designed for convenience and efficiency. Such a utilitarian perspective, Western in origin, contrasts with philosophical ideas of technology as a way to reveal truth (as touted by European philosophers from Aristotle and Plato to Heidegger). Today, technology tends to be synonymous with the latest and the finest; continuous development and progress is implied. What do these Western perspectives mean to, and for, the rest of the world? *What about Africa?*

Enter "AI"—two letters that encapsulate a vast array of computational techniques and emerging machine capabilities, often shrouded in mystery and perceived (or even promoted) as magical. Generative AI tools like LLMs blur the lines between reality and imagination, creation and regurgitation, human and machine, transmission and truth (Yiu et al., 2023). The allure of LLMs lies in their ability to produce human-like text, images, and speech almost instantly, confounding users with its capabilities. Anyone with access to the tool can witness the show, but behind the curtain is a model tethered to its predominantly English-language text corpora and encoded with the opinions, knowledge, biases, and assumptions—the Western "habitus" (Bourdieu, 2022)—of its creators.[2]

One multinational company's academic paper likens MT, a technology related to LLMs, to a wizard's "magical device" from 1970s science fiction, capable of relating "all possible vocables to every conceivable system of meaning" (Costa-jussà et al., 2022, p. 4). The public-facing presentation introducing their dataset and model for multilingual MT promises "high-quality results" in fifty-five African languages. However, researchers Asmelash Teka Hadgu, Paul Azunre, and Timnit Gebru (2023) identified critical quality concerns with the dataset (e.g., potential inclusion of machine-translated content) and the model's performance (e.g., significant performance gaps for colloquial language).[3] These researchers argue that promoting such flawed models and datasets as high-quality solutions for underrepresented

languages undermines the indigenous African language research ecosystem (Hadgu et al., 2023, pp. 1–2).

LLMs, MT models, and other advanced language technologies see global use, including across Africa for those who can access them. The notion that these tools should primarily serve their development communities has merit, yet it ignores the reality of worldwide adoption and the potential for both positive and negative impacts in varied cultural and linguistic settings. Consider these complexities and the ethnographic imperative comes into clear relief. Researchers are tasked with investigating the social contexts and power dynamics of technology use, probing how systems and narratives are constructed and enacted, and ultimately illuminating the agency of those who create, use, and adapt these tools of "imported magic" (Medina et al., 2014).

Researching LLMs in Africa demands collaboration with African people, from the ground up. It requires awareness of all stakeholders' assumptions, historical contexts, and prevailing narratives. And it means starting with African vernaculars and local engagements to establish an equal ground, or, in ML parlance, to create the "ground truth": the directly observable reality (which developers aim to model). For ethnographers studying LLMs in Africa, the ground truth emerges from reflexive inquiry and culturally sensitive interactions, whether on-site, online, or anywhere else. This foundation leads us to deeper explorations of linguistic agency and decolonization.

## (De)colonizing Language

In addition to being linguistically diverse and rich in technological innovation and adaptation, Africa is highly multilingual. Some scholars hold that multilingualism itself is the continent's lingua franca (Fardon & Furniss, 1994). Most Africans speak several languages, often including a regional language that might be used as a medium of instruction in higher education (Brock-Utne, 2017). Our May 2024 survey with African UX researchers in Ethiopia, Ghana, Kenya, Nigeria, South Africa, and Tanzania, introduced in detail below, found that nearly half of respondents (13 of 27) have a linguistic repertoire that includes more than one language, and some (4 of 27) have a repertoire of three or more languages.

Africa's multilingualism is shaped by the frictions of its physical landscapes and, paradoxically, sustained by poor infrastructure and investment. These conditions can act like isolation in biology, encouraging linguistic diversity by creating barriers that prevent linguistic homogenization. Conversely, more resources and interactions between communities—equating to less friction—typically lead to less linguistic

diversity. In many regions, multilingualism is particularly widespread among the urban lower classes, motivated by mass rural exodus and resulting language contact (Lüpke, 2010). Consider the linguistic landscape of Nima, Ghana's largest *zongo* (a Hausa word meaning "strangers' quarters"), or enclave of intra-African migrants (Moro, 2023). Ethnographic research by sociologist Kwesi Kwaa Prah (2009) found that almost 70% of Nima residents he interviewed spoke at least four languages, and about 40% spoke five or more. In Ghana, and elsewhere, multilingualism is especially pronounced in oral domains compared to written ones, although the linkages vary and the correlations are not one-to-one (Lüpke, 2013).

Africa's linguistic resilience is also exemplified by the adaptation of ex-colonial languages into forms that reclaim and Africanize them, as seen in distinctive local styles or mixed colloquial registers of Pidgins and Creoles (and Pidgincreoles like Juba Arabic discussed below). "English in Africa comes in many shapes and sizes," as South African sociolinguist Rajend Mesthrie puts it, "all of which look set to grow" (2019, p. 19). The versatile and adaptive aspects of Africa's multilingualism raise the important question of when and how foreign languages become indigenous—just what counts as an African language, and who gets to decide?

Sociolinguist Jan Blommaert reminds us that "the horizontal distribution of languages…rarely matches the vertical distribution of languages as codified in language policies" (2007, p. 11). This observation holds true both in the postcolonial present and the colonial past. Many African countries designate "national" or "official" languages that are or include ex-colonial languages (Afrikaans, Arabic, English, French, Portuguese, Spanish), although there are notable exceptions, such as Ethiopia (Amharic), Somalia (Somali), and Tanzania (Kiswahili). Over time, several countries have strengthened local language policies—Kiswahili in Kenya and Tanzania (Robinson, 2024), and Amazigh (or Tamazight) in Morocco and Algeria (Alalou, 2023).

National languages and linguistic unity are often state-driven projects aimed at building national identities and a coherent national populace. The "official language problem" (Pool, 1991), intersected by ideological, political, economic, and other factors, typically means that resources are rarely allocated to support indigenous languages beyond those designated as "official." The functional hierarchies established through these policies in postcolonial Africa have led to a variety of outcomes: some positive, such as increased communication and administrative efficiency; and some negative, such as the marginalization of autochthonous languages and the loss of cultural heritage.

The struggle to balance vernacular and vehicular languages is marked by the legacies of colonialism. Colonial experiences in Africa, plural and diverse, triggered profound changes in language use and most other dimensions of life. Indigenous languages were suppressed (or pressed into the service of the colonizers) as European languages like English, French, and Portuguese were forcibly imposed and institutionalized. Sociologist Bonaventura de Sousa Santos (2014) termed this widespread dispossession of local languages and knowledge "epistemicide."

The case of African Arabic showcases the complexities of language use and the power dynamics impinging upon it. Arabic, like Amharic, Tigrinya, and other Semitic languages, evolved through ancient interactions—many characterized by dominance—between Western Asia and Northeastern Africa. Across centuries, Arabic exerted a profound influence on major indigenous African languages throughout the continent, such as Hausa, Wolof, Kiswahili, and Somali. In North Africa, French was imposed to the discouragement or outright exclusion of local vernaculars—including Arabic. However, in the postcolonial era, the linguistic policies of the governments of Algeria, Tunisia, Morocco, and Sudan, guided by a strong ideological push for "Arabization," themselves imposed Arabic on public domains, sometimes at the expense of local vernaculars (Tilmatine, 2015). Today, Arabic is considered less colonial than French in parts of North Africa but more colonial than English in South Sudan. When South Sudan in northeastern Africa gained independence in 2011, the government recognized English as the national language and dismissed Arabic as a colonial imposition—including homegrown variants like Juba Arabic. As a result, that widely spoken Arabic-based pidgincreole remains neither recognized as an official language nor officially acknowledged as an indigenous language (Manfredi & Tosco, 2016). Decolonization "unleashed many ironic and unintended effects," in the apt words of scholar of African American literature and postcolonial theory Vaughn Rasberry (2021). Some effects, like Arabization, recapitulate the dynamics of past colonialism in new forms.

Power asymmetries, both broadly and specifically in the context of language, have evolved into new forms in today's digital age, often characterized as data colonialism (Couldry & Mejias, 2024; Benyera, 2021; Birhane, 2020; Kwet, 2019). Where traditional colonialism concerned the acquisition of physical and epistemic territories for economic extraction, data coloniality revolves around the extraction and control of data. Design and research executive Ovetta Sampson reminds us that all data is produced by people, "it is the manifestation of who we are as people" (2023). In Africa, people produce data in contexts of limited infrastructure (the "connectivity rifts" introduced above) and limited data protection laws. Various other structural and relational challenges, many stemming from diverse local norms,

complicate interactions between external entities and African communities (Abebe et al., 2021). All of this is set against a backdrop of enduring social, political, and economic inequalities and a history of resource exploitation (Coleman, 2019).

Power asymmetries extend to the process of cultural and linguistic translation. As anthropologist Talal Asad noted, and our research examples below give shape to, cultural translation involves navigating "asymmetrical power dynamics and understanding the pressures within dominated and dominant societies" (1986). The task of exploring these processes and determining what Asad called "the limits and possibilities of effective translation" falls on the researcher.

Asad's call to action takes on added significance in the context of LLMs, where struggles for control increasingly extend to language data—the vast amounts of digitized linguistic resources necessary for their development. For example, Te Hiku Media in New Zealand recorded over 300 hours of annotated audio in te reo Māori, only to fend off corporate entities trying to use these recordings for their own data sets. The data governance plan drafted in the wake of the experience stipulates that "the project must directly benefit the Māori people and any project created using Māori data belongs to the Māori people" (Graham, 2021).

Considering the wide-ranging experiences of colonization and the ongoing colonialesque exploitation through modern means, a decolonizing framework is relevant and necessary. Decolonization literally means the negation of colonization, a concept built upon recognizing colonialism's legacies and aspiring to counteract them. Decolonization is multifaceted, comprising a continuum of meanings, practices, and definitions. Language plays a central role.

Kenyan writer and academic Ngũgĩ wa Thiong'o places indigenous African languages and cultural expressions at the heart of decolonization. For Ngũgĩ, language is not only a means of communication but also an essential "carrier of culture" (1986, p. 4). Similarly, African linguist Irina Turner (2023) writes that "languages contain not only semantic concepts describing life worlds but also ontologies that relate to holistic spiritual and philosophical existence" (p. 73). Words are a legitimate and necessary means for liberating the human mind from oppression. Reclaiming and revitalizing indigenous languages, it follows, is essential to decolonization. Zimbabwean historian Ngwabi Bhebe writes of oral traditions, cultures, and languages as "powerful instruments for the restoration and assertion of African pride and identity after colonial devastation" (Bhebe, 2002, p. 37).

Others emphasize that decolonization is primarily a matter of self-determination, where people have the capacity to choose their languages and systems, regardless of

origin. Philosopher Olúfémi Táíwò (2022) argues for a move "beyond linguistic decolonization" (p. 101), advocating for a pragmatic, instrumentalist approach to language. Táíwò joins Amílcar Cabral, writer, intellectual, and founding leader of the PAIGC (African Party for the Independence of Guinea and Cape Verde), in affirming African agency in decisions of whether and what to borrow or adapt—and in doing so own—from colonial languages and cultural artifacts (Cabral, 2016).[4] Táíwò and Cabral recognize the knotty ties binding colonial and indigenous languages, such that they cannot be simply disentangled. Their emphasis on individual and community autonomy in language choice bears deeply upon LLM research in Africa for two reasons. First, because ethnography bridges theoretical frameworks and on-ground realities. Second, because industry ethnographers connect the complex social world of human behavior and action to the development of "complicated products and services that their work informs" (Hoy et al., 2023).

Ethnographers can throw light on diverse expressions of African agency in digital contexts by studying how Africans navigate technologies designed primarily for ex-colonial languages. To support African agency, we must first acknowledge the multidimensionality and heterogeneity of African languages, lands, and lives. African experiences, as with African languages, are many and varied; they encompass the influences of colonialism and also extend beyond them. Agency can be expressed and autonomy exercised in many ways, including through language. Many Africans use English or other colonial languages instrumentally, as tools to interact with technology and the world, regardless of fluency.

Researchers therefore need to investigate:

1. Language use: What languages do people use, when, and why?
2. Language-technology interaction: How do language choices shape technological use, and vice versa?
3. Digital communication preferences: When might African individuals or communities prefer using English for digital communication over local vernaculars? Do those preferences differ across demographics or geographic regions?

Language choices profoundly shape technology use, making their study central to LLM research in Africa. Attention to agency—the ability to choose, adapt, innovate, or abandon tools to meet local needs and contexts—transforms the research into a decolonizing endeavor. A direct way to bolster African agency is to integrate it throughout the research process. This means researching *with* African communities, researchers, and scholars, not just talking *to* them. The framework presented later in the paper provides guidance on how to implement this approach. First, we explore

practical examples illustrating how technology, language, and agency intersect in several African contexts.

## Beyond Big Knowledge: Local Insights from Africa

Exploring Africa's techno-linguistic terrain and the contours of linguistic decolonization helps us clarify that advanced language technologies like LLMs are not merely technical innovations; they are catalysts for social change that stand to reshape how people communicate, express their culture, and interact with the digital world. As we consider the path to decolonizing LLMs, we must understand the complex interplay between these technologies and local African contexts. The following examples of real-world research scenarios where language, technology, and cultural dynamics converge illustrate both the challenges and opportunities that arise when global tools meet local African realities.

### Research Example 1: Kiswahili Language Localization (2023)

**Table 1. Research Example 1 Methodology**

| | |
|---|---|
| Location | Nairobi, Kenya; Dar es Salaam, Tanzania |
| Sample | N=10 each market, 90-min in-lab IDIs |
| Participants | Native Kiswahili speakers who use digital devices set to Swahili, balanced age, gender, and SEC mix |
| Tasks | Feedback on overall app language based on usage recall, task navigation in live applications, detailed language feedback on steps of these tasks using screenshots |

Two Kiswahili studies, conducted with identical methodologies and stimuli but using localized imagery, highlighted significant issues with the language use in digital platforms. Participants in Kenya expressed that the Kiswahili they found online tended to be excessively formal and academic, and unreflective of their everyday language. Additionally, there was a perceived bias towards those with more proficiency speaking formal Kiswahili in Tanzania and certain Kenyan regions, primarily inland. Several Kenyan participants said they opted to use their devices in English to avoid these issues. While participants in Tanzania generally found the language appropriate and aligned with their expectations for digital Kiswahili, overly formal expressions sometimes led to feelings of exclusion even for them. These varied responses direct attention to the nuances of language use and the challenge of creating universally suitable solutions in the linguistically diverse, postcolonial

contexts of Africa, where imposed linguistic norms can marginalize local dialects and the ways of speaking associated with certain socioeconomic classes.

Because LLMs are trained on digital language, which is often gleaned from the news, Wikipedia, and other sources that may prioritize more formal writing norms, they carry the risk of creating outputs so formal as to inadvertently exclude those who speak exclusively in more casual registers. That is, they risk reproducing the linguistic exclusion observed in these studies. Just as online Kiswahili felt overly formal to Kenyan participants, LLM outputs might alienate speakers of casual registers. This digital linguistic divide echoes historical patterns: formal English in postcolonial contexts often signaled education and rank, while Kiswahili's inclusive evolution incorporated Arabic loanwords for practical communication, often in the context of trade. If LLMs fail to adapt to diverse linguistic needs, they may inadvertently reinforce socioeconomic and regional language disparities in the digital realm.

## Research Example 2: isiZulu Language Localization (2023/2024)

**Table 2. Research Example 2 Methodology**

| | |
|---|---|
| Location | Johannesburg, South Africa |
| Sample | N=15, 90-min in-lab IDIs |
| Participants | Native isiZulu speakers, balanced gender and SEC mix, 50% Gen Z |
| Tasks | Feedback on brand language preferences, design walkthrough with transcreated stimuli |

This study assessed communication styles on an online content platform available via app and web browser with young native isiZulu speakers. The first challenge was recruiting these speakers, who predominantly use devices set to English and prefer the browser version over the app. Translating, or more specifically transcreating (i.e., adapting content for cultural relevance while maintaining the original intent), test materials from English to isiZulu was another significant hurdle. The process involved collaboration between local moderators, experts, linguists, ethnographers, and UX researchers. A local Zulu expert who had previously served a central governmental role in culture preservation played a crucial role. He felt compelled to help the team "get it right" after experiencing sleepless nights worrying about potential missteps. We observed instances where isiZulu language, written and checked by professional translators, was considered outdated or strange by participants, who reported feeling baffled by the mock content they

were shown. This study demonstrates how even carefully wrought human translations can unintentionally alienate native speakers.

The challenges we encountered with isiZulu translations mirror and amplify in LLMs. Currently, major LLMs struggle with language dynamism—regional diversity, borrowing, code-switching, slang, and other elements of living languages common in multilingual African contexts (Orji & Umeobi, 2023). Researchers argue that most major LLMs use an English-based backend for key aspects of their workflow, creating invisible translation layers that, by design, distance the model from non-Anglophone users' language and knowledge base. For example, Zhao et al. (2023) and Wendler et al. (2024) separately hypothesize that multilingual inputs are initially processed in English. Kew et al. (2023) discuss another technique, cross-lingual transfer learning, which leverages data from resource-rich languages (primarily English) to improve performance for LRLs. Such approaches can result in LLM outputs as distant or unintelligible as the carefully crafted isiZulu translations that perplexed our study participants. To mitigate these issues, LLMs should be trained on extensive multilingual data, with rigorous validation by native speakers from various regions and backgrounds.

## Research Example 3: Survey of African UX Researchers' Perspectives on LLMs (2024)

**Table 3. Research Example 3 Methodology**

| Location | Remote |
|---|---|
| Sample | n=27, 20-min survey with multiple choice, scale, and open-ended questions (10 open ends) |
| Participants | African UX researchers in Ethiopia, Ghana, Kenya, Nigeria, South Africa, and Tanzania |

We surveyed UX researchers across six African countries to assess LLM perceptions and use, focusing on language representation, accuracy, cultural impact, and associated ethical concerns. Respondents reported their native languages as Afaan Oromo, Afrikaans, Amharic, English, Hausa, Ibibio, Igbo, isiZulu, Kikuyu, Kiswahili, Nigerian Pidgin English, and Yorùbá. Most researchers (19 of 27) were familiar with LLMs, and over 80% had used them. The results paint a nuanced picture of benefits and challenges in multilingual African settings. English dominated LLM use (22 of 27), with half rating LLM performance in English highly. Several praised LLMs, used in English, for helping them clarify technical terms, summarize

and paraphrase text, and edit content. A few called them "indispensable" in their personal and professional lives already.

Half of our survey participants reported using LLMs for information-gathering tasks. While some saw potential for cultural knowledge preservation, many expressed concerns about misrepresentations and misinformation. Cultural sensitivity—defined here as the ability to preserve the nuances of diverse cultural contexts, languages, and historical narratives in LLM outputs—emerged as the top ethical concern (20 of 27), followed by data privacy (16 of 27) and data rights (14 of 27). Respondents stressed the importance of LLMs being able to identify and accurately interpret local expressions and practices. Several highlighted instances where LLM output failed in this regard—for example, incorrect biographical details about local political figures in Ethiopia or output that felt culturally "off the mark," as one researcher in South Africa put it (recalling the Western habitus noted earlier).

LLM performance with African languages presented significant challenges. Surveyed researchers identified dialects and variations (15 of 27), idiomatic expressions (14 of 27), cultural nuances (13 of 27), and syntax/grammar (13 of 27) as major areas of concern and poor performance. Experiments with Yorùbá, isiZulu, Amharic, Afaan Oromo, Nigerian Pidgin, and Kiswahili revealed substantial inaccuracies in LLM performance with these languages. Examples include:

### *Untranslated words*

A researcher in Lagos, Nigeria, reported using English-centric LLMs effectively for simple writing tasks in English. However, including even a single Yorùbá word led to inaccurate outputs. For instance, "*mo fe je iresi*?" ("I want to eat rice") yielded responses ranging from partial translations ("you want to eat some iresi?") to unrelated texts about Nigerian towns or general meal statements. A Johannesburg-based researcher likewise noted LLMs often leave colloquial expressions untranslated, citing the isiZulu word "Mzansi" (a colloquial name for South Africa) as an example.

### *Incorrect translations*

An Addis Ababa researcher prompted an English-centric LLM to explain the Afaan Oromo word *bukke*. The LLM responded:

> In Afaan Oromo, the word "Bukke" can refer to a "shelter" or "shade." It is used to describe a place or structure that provides protection or cover, typically from the sun or weather.

The Ethiopian researcher called out the LLM response as "totally different" from locally known meanings of *bukke* as "beside" or, in some dialects, "sensual organ." Notably, the very first result on a classic Google Search for "Afaan Oromo 'bukke'" yields a 1975 Peace Corps document titled *Oromo for beginners* that gives "beside" or "by the side of" as correct translations (Peace Corps, 1975).

Nigerian Pidgin translations yielded the poorest results reported. Several researchers noted that LLMs tended to either default to English or produce erroneous interpretations. Supporting these anecdotal observations, Ojo and Ogueji (2023) found incorrect English meanings when translating from Pidgin and retained English usage when translating into Pidgin. Variation in Pidgin seems to matter, too. Adelani et al. (2024) suggest major LLMs favor Nigerian Pidgin variants used by more educated groups.

### *Kiswahili Performance*

In contrast, Kiswahili showed more promising results. Native speakers in Kenya and Tanzania reported satisfactory experiences in-language on ChatGPT 3.5, noting reasonably accurate knowledge of current affairs and public opinion, with only minor translation challenges. This LLM's better performance stems from Kiswahili's vast speaker base (approximately 200 million) and prominence in education, entertainment, and politics. Kiswahili is the only African language officially recognized by name by the African Union alongside Arabic, English, French, Portuguese, and Spanish (African Union, n.d.). Crucially, Kiswahili has a higher online presence and has been included in several benchmarks for assessing LLM performance. However, even with Kiswahili, a significant performance gap separates high-resource languages (e.g., English and French) and African languages on current open as well as proprietary models (Alabi et al., 2024). That performance gap highlights the importance of local and multilingual initiatives (see examples in our earlier section "A Global and African Overview of Cultural Awareness in LLM Research").

## Research Takeaways

In many ways, translation is both the heart of ethnography and its central challenge, as ethnographies translate words, ideas, experiences, and lives (Clifford & Marcus, 1986). Translation is also at the heart of all human interactions. As Grice (1989) observes, some degree of translation effort is required in all conversations, even between two people using the same language. Working across languages and

cultures scales the effort and raises the stakes. The first two research examples above on dynamism *within* African languages spotlight the need to plan and execute research aimed at understanding not only how and why people use digital technologies but also the nuances of their language use—which languages in what contexts and why, and the specific dialect(s) that are intelligible and appropriate to them in those contexts.

Working with languages like isiZulu goes beyond simply having a native speaker translate. It requires an awareness of just how much comprehension and perception can vary across speakers, registers, tones—nuances current LLMs often struggle to capture. To give another example, in Igbo, *akwa* can mean egg, cloth, or cry (as in tears) depending on pronunciation and context, even though it has the same spelling. LLMs, typically prioritizing frequency, might lack the contextual understanding to distinguish such homonyms accurately.

Proceeding from a position of respect for the dynamic, contextual use of language, we can work to uncover the needs of our subjects and surface insights that may provide practical guidance on how to create products that are useful and relevant. Failure to do so risks misguided technological development and undermines efforts to preserve multilingualism, especially among younger generations who liberally use English. Additionally, there is a need for awareness of the new generation that has grown up with two linguistic realities: family and community memories of indigenous knowledge and linguistic wealth, and the postcolonial, Westernized world they now live in. This duality colors their use of technology.

The third research example, findings from our survey, lays bare LLM translation challenges *across* languages, and in doing so point to the broader risks of linguistic dilution and the unintended imposition of colonial languages, not to mention poor product uptake. Simple translation tests, as seen in the above examples, serve as a common "litmus test" for end users to gauge how well an LLM accommodates their languages. Poor translation accuracy can lead users to exercise their agency in several ways, each with distinct implications:

1. Avoid using the LLM in their first language; switch to another African language or opt to use English.
2. Abandon LLM use entirely, believing the technology doesn't work for them.
3. Accept incorrect translations, potentially abandoning local vernaculars—and their embedded knowledge systems—for perceived more authoritative, often Anglocentric terms.

For those who choose to use LLMs, and as users increasingly rely on LLMs for learning and writing, the model's linguistic features may not only shape their own

language choices and style but also influence those who consume and interact with their content. This impact could manifest in shifts in word choice or register—for example, encouraging the use of specific terms or phrasing that might not otherwise be employed. Language use often evolves under the influence of dominant linguistic frameworks, potentially leading to the gradual displacement of local terminology and its embedded knowledge systems. As Ghanaian philosopher Kwasi Wiredu (1997) warned, thinking *about* things in English "almost inevitably becomes thinking *in* English" (p. 12). However, LLMs can also lower barriers to working in English and boost confidence, offering micro steps towards leveling a steep linguistic playing field.

LLMs, while primarily language models, are increasingly used as information-gathering tools and de facto knowledge repositories (Fletcher & Nielsen, 2024). Half of our survey respondents reported using them this way, as already noted, and mostly in English. This usage pattern raises concerns in the African context, where without close human collaboration, LLMs may produce outputs that not only contain discernible errors, as seen in our translation examples, but also potentially distort or omit marginalized histories and cultures. Wachter et al. (2024), Oxford Internet Institute researchers at the University of Oxford, describe such outputs as "careless speech," content that appears plausible and confident but includes factual inaccuracies, misleading references, and biased information. In Africa, where linguistic and cultural diversity is vast, this misinformation risk is particularly acute, carrying both immediate dangers and the long-term risk of homogenizing diverse histories and cultures.

In a similar vein, Shah & Bender (2024) stress that while LLMs may provide access to information, they often lack provenance and remove the valuable process of manually searching through materials, which offers a range of choices and promotes critical thinking. The healthy friction of searching for information manually, whether via a search engine on a smartphone or perusing a physical space such as a library or a bazaar (Roberts et al., 2023), enables transparency and user agency.

These challenges and risks necessitate close attention by researchers on the ground and echo the calls we heard for foundational improvements to LLM technology by African researchers. Specifically, they voiced the need for broader representation of African languages in training data and improved contextual understanding tailored to local needs. And they pinpointed community driven research as key to gathering representative, multilingual data beyond widely spoken

languages or social media vernaculars. The framework we present next provides a roadmap for research that will help to actualize these calls.

## Framework for Ethnographic LLM Research in Africa

We invite the ethnographic community to view Africa with "many eyes" (Mavhunga, 2017) and recognize the innovations its peoples and governments contribute alongside the challenges they face. To provide practical know-how, we lay out below a grounded, inclusive approach to research that centers African agency, vernaculars, and epistemologies. This framework is designed to guide LLM research in African settings but is adaptable and can benefit various other settings. It covers each stage of the research process from study design to dissemination.

### Research Planning

*1. Identify Assumptions and Research Team Positionality*

Start by conducting researcher reflexivity exercises to identify and challenge personal assumptions. Reflect on how your knowledge of the African subject has framed your perspective, and recognize the limits of your access and knowledge despite intentions. This foundational step is critical in crafting a research framework that benefits the studied communities and not just the entities developing or using LLMs.

Engage with relevant African scholars to understand existing narratives and incorporate them into the research framework. Conduct desk research, focusing on works by Africans to cultivate historical awareness and assess risks and opportunities. Preparatory impact analyses should grasp the knowns and the potential benefits and harms of the new research and should consider colonial histories, language landscapes, and the research landscape.

Key Questions:

- How do my assumptions shape the research questions and approach?
- What does the team hypothesize about the forthcoming research? What do these hypotheses demonstrate about the team's assumptions?
- What are the socio-political and economic statuses of the language communities of study, and how does LLM research and technology play into these power dynamics?
- What are the tensions between the interests of users, societies, and the owners of LLM technology?

- Does the research consider how the LLM technology could address specific challenges identified by African communities?
- How will success be measured, especially if the ultimate goal is intangible or long-term?

*2. Co-Design to Empower Local Voices*

Involve community members and local experts throughout the research process using participatory research methods. Start the process of building equal partnerships with local researchers and communities by first listening to their stories and experiences as Africans and as African researchers. This approach requires researcher modesty and an acknowledgment that there might be cultural protocols to follow or limits to access and knowledge, or both.

Brainstorm together how you might integrate local methodologies and ethical frameworks to enhance the research's relevance and remain locally grounded. Be imaginative and creative as you scope the project—what culturally relevant methods will capture the most holistic view of participants' truths? Consider factors like regional language variation, geographic and generational language use in sampling strategies, device constraints, and connectivity disruptions during recruitment. Prioritize inductive reasoning in interviews, using semi-structured scripts or embedded ethnography over deductive hypothesis orientation. Favor longer-term engagements like participant observation over rapid assessments to gain deeper insights.

Two examples of ethnography in Africa illustrate the power of culturally sensitive, participatory research methods and the insights (and learnings) they can yield. First, Hanover (2014) used storytelling during co-creation research in rural Ghana to "diffuse our authority as researchers" and stimulate judgment-free conversation. The approach worked well, except with school children, where they added an object-based dimension using action figures. Hanover's team learned that the Western "superhero" trope did not resonate locally; instead, the nearest cultural analog was black magic—a source of fear rather than empowerment. Second, Abebe et al. (2021) employed storytelling with fictional personas, developed through iterative interviews with African data experts, to challenge dominant narratives on data sharing in Africa. The team's decolonizing approach brought local perspectives to the forefront and surfaced important counternarratives.

To further enhance the participatory approach, tap into the principles of value-sensitive design (e.g., Friedman & Hendry, 2019), co-design (e.g., Zamenopoulos &

Alexiou, 2018), and participatory synergy (Bennett et al., 2021; Eglash et al., 2024), all of which emphasize the importance of integrating stakeholder values and collaborative creation. Map stakeholder networks and aim to include as many voices as possible in the research.

Make "AI" a topic of discussion from the start to create space for all stakeholders to voice their perspectives and concerns, especially because the most publicly visible AI in 2024 are LLMs. Recognize that each stakeholder and participant may have a different understanding of those two letters. These understandings can vary in terms of what AI is, how it has been developed, its technical processes, and its contextual history. They may also differ on how AI might be managed and the opportunities and risks it presents for them individually, organizationally, and societally as producers of data, as consumers, as Africans, and as global citizens. Share your knowledge and answer questions to the extent possible, while acknowledging the limits to your knowledge. Be transparent about the gaps in understanding, and actively seek input from experts to address the more thorny or technical questions.

Everyone involved in the research will benefit from the transparency, especially those in the "hidden part of the iceberg" of Africa's data-sharing ecosystem—Global South researchers, data subjects (individuals and communities), data workers, activists, and others (Abebe et al., 2021). This "hard needed first step" (Dignum, 2023) to find shared ground will encourage mutually respectful relationships and help to demystify the processes, possibilities, and risks at hand.

Key questions:

- How can data collection methods be adapted to respect cultural norms?
- How can African-centric methodologies, such as the ubuntu principles of respectful relationships (Khupe & Keane, 2017), be integrated?
- How might I or my research contribute to building local expertise?
- How am I talking about LLMs and AI with the local research team and participants? To the extent known, what are the materials used in the LLM gathered from and who is involved?
- How can I work within legal and organizational bounds most effectively and ethically?
- What cultural protocols do I need to observe, and what are the budgetary and time implications?

*3. Recruiting & Informing Participants*

Building on the co-design and collaborative mentality from the research framework design phase, approach recruitment and data collection with a similarly attentive, caring, and compassionate mindset.

Consult with local partners to understand demographic aspects for creating a balanced participant pool. At the same time, recognize that efforts to include participants will inevitably exclude some individuals (Spivak, 2003). The most disadvantaged are often the most difficult participants to recruit for studies, such as those with low literacy, limited time, or lack of access to appropriate networks. Consider that database recruitment, though commonly used in Western contexts, can bias samples to an English-proficient, literate, digital, and legally identified participant pool. In-field recruitment can increase the diversity of the participant pool and extend participation to vulnerable groups.

Use best practices for recruitment to avoid pressuring participation through high incentives (Teixeira da Silva, 2022). Make efforts to identify the excluded and those who choose not to participate, and think about the implications of their absence (Birhane et al., 2022). Beyond direct participants, consider those who will be impacted by the end product, including communities affected by LLM-generated content. Consider ways to reach out to these groups in future research.

Informed consent is more than a form! Prioritize truly informed consent where participants are given as much information as possible about what the study is looking to understand, how it will be conducted, and what the data will potentially be used for. Tailor consent forms and information sheets to be culturally appropriate and easily understandable. Plan your study in a way that gives participants power over the data collected and knowledge of the results to the extent possible within legal frameworks and organizational constraints. Data governance policies have historically been extractive, with acute consequences for marginalized and indigenous communities.

Researchers, as the bridge between institutions and participants, have a responsibility to do more than share a form. They need to deeply understand the consent process, ask questions, and make sure local partners and participants do too. Neglecting ethical data use at the planning stage can cause "irreparable harm" (Abebe et al., 2021) during as well as after a project. Conversely, when researchers act as responsible data stewards (Ada Lovelace Institute, 2021), ethical data practices build trust and increase the likelihood of further collaboration—they become empowering.

Key questions:

- How can I select African communities to work with, prioritizing representativeness?
- Who am I including and excluding in the research process, and why? How can I ensure that efforts to include participants do not unintentionally exclude others, especially the most disadvantaged?
- Do participants truly understand what this research—and their participation—will be used for? Do I?
- How can I confirm that participants fully understand what they are consenting to, particularly in contexts with limited literacy or digital literacy?
- What measures can be taken to provide genuinely informed consent, beyond just ticking a box?
- How can we account for the power asymmetries in consent dynamics, especially in regions with limited market competition and high service dependency?

## Research Execution

*4. Grounded Input (Data Collection), Grounded Output (Synthesis)*

Research execution should be anchored by local teams' active involvement, with a deep connection to the community and nuanced understanding of cultural contexts. Use local moderators and interpreters whenever the situation permits for effective data collection. Choose moderators carefully to match participant profiles in terms of gender, age, ethnicity, language, and other factors. Their rapport and sense of trustworthiness and relatability will elicit more discerning feedback and, above all, an inquirer that will be sensitive to cultural protocols for the given discussion.

Prioritize local languages for data collection and analysis to surface the richest insights. Employ translation at the final stages for non-native language researchers and readers. Translation is not neutral and can strip context or distort meaning. Stress the importance of high-quality simultaneous interpretation if used. Always use human transcription and translation unless a tool is vetted by a local team and known to perform well with the African languages in question. Even then, use the tool as a secondary or tertiary complement. Transcription tools often miss nuances between speech and action, akin to deciphering dialogue in literature without narrative cues. Emphasize the importance of translations and transcriptions that

resonate contextually and capture the essence of what participants say, which is often conveyed subtly or between the lines.

During synthesis, continually reflect on how you will craft a narrative that responsibly translates the data, recognizing it as knowledge obtained through face-to-face human contact in historically specific conditions (de Pina-Cabral, 2011). Review your analysis at multiple points, both internally and with local research partners and stakeholders. In other words, validate results-in-progress using cultural context debrief sessions with local teams, for example, and iterate based on their feedback. Relate stories and situations—the ethnographic data—with proper context and analysis to build reliable and robust knowledge. Take into consideration that participants typically express themselves more comfortably in their first language(s), so even if research outputs are aimed at improving a global, English-centric interface (e.g., social media platform, e-commerce site, search engine), the value of local insights that convey an understanding of local vernaculars remains essential.

Remember that ethnographies can and should be artful and evocative narratives (Stoller, 2023), but never novels or fictions. While you can never present the full picture, aim to provide as representative a slice as possible, knowing it will be received and subsequently applied in ways beyond your control. And remember that ethnographers practicing in industry are custodians of knowledge, tasked with communicating insights to stakeholders in a way that captures the dimensionality and texture of user experiences. Acting as a clear channel for relating local situations to non-local audiences is challenging, as the isiZulu example above demonstrates, but getting it right matters most. Keeping these fundamentals top of mind during data collection and synthesis is paramount, as they underpin the narratives and deliverables that will follow in the later stages of the research process.

Key questions:

- What additional training might need to be conducted to prepare local moderators to extract the data needed?
- What balance of rigidity and flexibility is ideal to achieve rich results?
- What methods, including unconventional methods, will likely produce the most insightful results?
- What stories do participants tell themselves to make sense out of their realities?
- How might observations and interviews unfold in a way that allows both the research goals and participants' agency/desire to be heard? Can both be achieved?

- How will you identify and handle unarticulated needs in languages you do not personally have competence in?

*5. Create Research Artifacts, Socialize Them, and Drive Impact*

By the time the research project progresses to write-up and dissemination, the research team must take special care to simultaneously continue engagement with the local participants and local team, while re-engaging with the stakeholders who are tasked with implementing the research's findings and recommendations.

Researchers can continue collaboration with the local community or participants by asking them if they would be willing to review the eventual findings and recommendations. Continuing the feedback loop can give the researchers reaffirmation that their interpretation of trends in the ethnographic data, and what to do about them, remain consistent with African participants' wishes and emic interpretations.

Ongoing contact with the local team is more straightforward. Local partners typically collaborate with researchers on final deliverables by contributing insights, analysis, and recommendations. However, to avoid abstracting their knowledge in research documents, we recommend including the local team in any working sessions, readouts, or design sprints where product team stakeholders consume and integrate research outcomes. Visibility and inclusion not only empower the team to clarify any misinterpretations but give visibility to their embodied knowledge.

LLMs necessitate a different path to impact than other digital products. If the research uncovers opportunities with the interface itself, then the approach is similar to other static products, i.e., working with product, design, and engineering to make fixes. However, if the research uncovers opportunities in model training, language use, veracity, or other aspects of LLM-powered generation, then other paths to impact must materialize, such as rewrites, novel training sets, or model modifications.

Rewrites constitute direct corrections to LLM output to improve future outputs, and they can be a powerful way for those impacted by a model to influence it directly and in terms it can "understand." If rewrites are useful and practical for stakeholders, researchers can assess the interest and willingness of participants or local team members to submit them. Additional training sets are another similar way to influence models, but because the learning may be less supervised, and the ratio of new material to old material small, the impact may be less pronounced. If the model itself needs to be modified, then the researchers will likely need to have a thorough discussion with the product and engineering teams. If local team members or participants can be present, all the better, but when proprietary discussions are

occurring, the researcher is often the best positioned person to represent participants' and local teams' concerns and insights in an informed manner.

Key questions:

- How might we keep participants involved, or at least informed, of how their involvement is affecting change?
- How might we create project deliverables using a co-creation/co-ownership model?
- How might we avoid alienating members of the local team from the knowledge they help to create?
- How might we facilitate contact between product team members, local team members, participants, and their communities?
- How can we change the LLM in accordance with the wishes, desires, and aspirations of African participants?
- How might we make sure things get done after the readout occurs and the project is closed?

## Conclusion

LLM research in Africa bridges two complex worlds. It means engaging with the continent's unique techno-linguistic terrain and confronting the layered impacts of colonial history. And it means grappling with LLMs—the most significant tech development since the smartphone.

LLMs pose substantial risks, both narrow and broad, real and potential. Models with English-centric designs may be particularly prone to embedding biases, blind spots, and omissions. Access to LLM-powered tools varies by language and socioeconomic status, potentially reinforcing power asymmetries and further fueling data colonialism. The environmental toll of LLMs adds to existing climate concerns. Across it all, the negative consequences of these risks and others disproportionately affect the Global South, including Africa. And they can manifest in subtle yet profound ways—recall the example of the fictional Abba Otho tale at the paper's opening.

Yet, turning from critique to possibilities, global and African initiatives across interconnected sectors—spanning grassroots movements, research labs, tech startups, academic institutions, government agencies, civil society, and educational programs—showcase the potential of developing useful applications of LLM technology that serve African people. These efforts address crucial preconditions like

infrastructure, governance, and digital literacy while expanding the possibility space for decolonized LLMs by exemplifying a vision of users as agents who are active shapers of technology. When we lean into all the complexities, ambiguities, and contradictions of LLMs and Africa, and LLMs *in* Africa, challenges become entry points for fresh perspectives—a necessary "different optic" (Mavhunga, 2017). And when we view LLMs less as fixed products than as sets of open questions (Moss & Schuur, 2018), we enable co-creation and collaborative innovation.

Researchers, developers, policymakers, and community leaders are working in concert to understand LLMs culturally and contextually, develop solutions for gaps and risks, and build towards a responsible AI future that reflects and is driven by African voices. Ethnographers play a vital role in this ecosystem: Our proposed research framework builds two-way knowledge bridges between tech teams and local communities. Sturdy bridges pave the way for generative partnerships capable of influencing model development and creating technologies that work for Africans on their terms, in their languages. By centering Africa, a continent whose languages and knowledge have long been marginalized, we advance a collaborative vision and contribute to the crucial work of decolonizing LLMs.

## About the Authors

**Lindsey DeWitt Prat** is a Senior UX Researcher at Bold Insight, based in Paris. A humanities-trained ethnographer, author, translator, and compassionate human, she has conducted user research focused on localization and cultural insights in over 20 countries and contributes thought leadership on AI and UX research. Her published work explores gender exclusion, cultural heritage, and religion in Japan through a combined ethnographic and historical approach. Lindsey holds a PhD in Asian Languages & Cultures from UCLA and an MA in International Studies and Comparative Religion from the University of Washington. lindsey.dewittprat@boldinsight.com

**Olivia Nercy Lucas** is a user experience researcher at Mantaray Africa specializing in the African cultural sphere. A driven and creative researcher with a keen eye for detail and complexity, her recent research focuses on how we can create responsible AI for Africa. She holds a degree from Monash South Africa focused on communication, media and global studies, and leverages her past immersion in various African countries as a Mozambican to amplify her research curiosities. olivia@mantaray.africa

**Christopher Golias, Ph.D.**, is a technology anthropologist, currently with Google, who has conducted applied anthropological research across various areas including retail, healthcare, indigenous rights, substance use, ecommerce, governance, machine learning, localization and information technology. He holds a Ph.D. in Anthropology from the University of Pennsylvania. cgolias@google.com

**Mia Lewis, PhD**, is a user experience researcher and independent scholar. She has worked on human-centered research across nearly twenty countries, delivering insights that are linguistically and culturally attuned to the needs of subjects. Her recent professional research has focused on the intersection of language, LLMs, and users' mental models. Her published research focuses on modes of communication and gender(ed) representation in Japanese media. Mia holds a PhD in Japanese from Stanford University. mia.e.lewis@gmail.com

# Notes

We thank the researchers and teams in Africa for their generous participation, the study participants, and the many friends and interlocutors who contributed to the emergence of this paper. Special thanks to Mantaray Africa and Helga Stegmann for their generous and enduring support, as well as Bold Insight, Google, and our paper curator, Maya Costa-Pinto. While the paper has benefited greatly from their guidance, the views it contains are solely those of the authors and may not necessarily reflect the views of our employers and reviewers. We would also like to express our profound appreciation to Lindsey DeWitt Prat, whose extraordinary storytelling and substantial contributions have been instrumental in the creation of this paper. We position this work as a first step in a much larger project of research and collaboration, and acknowledge any shortcomings as our own.

1. Additionally, there is evidence that some LLMs perform better and cost less with Latin-script languages that have shorter token lengths (Ahia et al., 2023), specifically American English (Yao, 2024), and that some LLMs use English as an "internal pivot" language in certain representation layers (Wendler et al., 2024).

2. Habitus refers to the ingrained preferences and dispositions shaping how individuals perceive and interact with the world, their social positioning and subjectivity, their "feel for the game" (Bourdieu, 2002, p. 27).

3. Hadgu is co-founder and CTO of Lesan, a MT system for LRLs developed in Ethiopia. Azunre is a Ghanaian-American AI researcher and entrepreneur, founder of GhanaNLP and Aglorine. Gebru is founder and director of the Distributed AI Research Institute (DAIR).

4. Today, the interconnectedness Cabral championed is being channeled, literally, through the "Amilcar Cabral Submarine Cable Project," a $90 million initiative supported by the Economic Community of West African States (ECOWAS) and the World Bank seeking to expand broadband connectivity across West Africa 3,130 km from Praia, Cabo Verde, to Monrovia, Liberia (Front Page Africa, 2024).

# References Cited

Abdelhay, A., Abu-Manga, A. A., & Miller, C. (2015). Language policy and planning in Sudan. In B. Casciarri, M. Assal, & F. Ireton (Eds.), Multidimensional change in the Republic of Sudan (1989–

2011): Reshaping livelihoods, conflicts and identities (pp. 263–280). Berghahn. https://shs.hal.science/halshs-01410901f

Abebe, R., Aruleba, K., Birhane, A., Kingsley, S., Obaido, G., Remy, S. L., & Sadagopan, S. (2021, March). Narratives and counternarratives on data sharing in Africa. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 329–341. https://arxiv.org/pdf/2103.01168

Ada Lovelace Institute. (2021). Participatory data stewardship. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/report/participatorydata-stewardship/

Adejoh, F. E., Areji, A. C., & Odey, J. J. (2024, January). An appraisal of Kwasi Wiredu's conceptual decolonization. *Aquino Journal of Philosophy*, 4(1), 76–87.

Adelani, D. I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., ... & Osei, S. (2021). MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, (9), 1116–1131.

Adelani, D. I., Doğruöz, A. S., Shode, I., & Aremu, A. (2024). Which Nigerian-Pidgin does generative AI speak?: Issues about representativeness and bias for multilingual and low resource languages [Preprint]. https://arxiv.org/abs/2404.19442

African Computer Vision Summer School (ACVSS). (n.d.). Retrieved August 5, 2024, from https://www.acvss.ai/

African Union. (n.d.). AU languages. Retrieved July 16, 2024, from https://au.int/en/about/languages

African Union. (2024, June 17). African ministers adopt landmark continental artificial intelligence strategy, African digital compact to drive Africa's development and inclusive growth [Press release]. https://au.int/en/pressreleases/20240617/african-ministers-adopt-landmark-continental-artificial-intelligence-strategy

African Union Development Agency–NEPAD (AUDA-NEPAD). (2023, June). Regulation and responsible adoption of AI in Africa towards achievement of AU agenda 2063. Retrieved August 5, 2024, from https://onedrive.live.com/?authkey=%21AKJcwcnXeRGANKQ&id=14DDAD979C3656DF%2145404&cid=14DDAD979C3656DF

Ahia, O., Kumar, S., Gonen, H., Kasai, J., Mortensen, D. R., Smith, N. A., & Tsvetkov, Y. (2023). Do all languages cost the same? Tokenization in the era of commercial language models [Preprint]. https://arxiv.org/abs/2305.13707

Ahmad, S. F., Han, H., Alam, M. M., et al. (2023). Impact of artificial intelligence on human loss in decision making, laziness and safety in education. *Humanities and Social Sciences Communications*, 10(311). https://doi.org/10.1057/s41599-023-01787-8

Ahuja, S., Aggarwal, D., Gumma, V., Watts, I., Sathe, A., Ochieng, M., ... & Sitaram, S. (2023). Megaverse: Benchmarking large language models across languages, modalities, models and tasks [Preprint]. https://arxiv.org/abs/2311.07463

Aidi, H., Lynch, M., & Mampilly, Z. (2021, September). Racial formations in Africa and the Middle East: A transregional approach. *POMEPS Studies*, 44. https://pomeps.org/wp-content/uploads/2021/09/POMEPS_Studies_44_Web-rev3.pdf

Alalou, A. (2023). The sociolinguistic situation in North Africa: Recognizing and institutionalizing Tamazight and new challenges. *Annual Review of Linguistics*, 9(1), 155–170.

Algorine. (n.d.). Retrieved on August 1, 2024, from https://www.algorine.com/

Asad, T. (1986). The concept of cultural translation. In British Social Anthropology. In J. Clifford & G. Marcus (Eds.), Writing culture (pp. 141–164). University of California Press. https://doi.org/10.1525/9780520946286-009

Asiedu, M., Dieng, A., Haykel, A., Rostamzadeh, N., Pfohl, S., Nagpal, C., ... & Heller, K. (2024). The case for globalizing fairness: A mixed methods study on colonialism, AI, and health in Africa [Preprint]. https://arxiv.org/abs/2403.03357

Bender, E. (2019, September 14). The #benderrule: On naming the languages we study and why it matters. The Gradient. https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/

Bennett A, Eglash, R, Graf, R, Butoyila, D, Johnson, K, Low, J, Andréia, R. (2021). Towards radical synergy for more just and equitable futures. In LEARNxDESIGN 2021: 6th International Conference for Design Education Researchers. https://doi.org/10.21606/drs_lxd2021.02.188

Benyera, E. (2021). The fourth industrial revolution and the recolonisation of Africa: The coloniality of data. Taylor & Francis.

Bhebe, N. (2002). Oral tradition in Southern Africa. Gamsberg Macmillan Publishers.

Birhane, A. (2020). Algorithmic colonization of Africa. *SCRIPTed*, 17(2), 389–409. https://doi.org/10.2966/scrip.170220.389

Birhane, A., Isaac, W., Prabhakaran, V., Diaz, M., Elish, M. C., Gabriel, I., & Mohamed, S. (2022, October). Power to the people? Opportunities and challenges for participatory AI. *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–8.

Blommaert, J. (2007, January). Sociolinguistic scales. *Intercultural Pragmatics*, 4(1), 1–19. https://doi.org/10.1515/IP.2007.001

Bourdieu, P. (2002). Habitus. In J. Hillier & E. Rooksby (Eds.), Habitus: A sense of place (pp. 27–34). Ashgate.

Brock-Utne, B. (2017). Multilingualism in Africa: Marginalisation and empowerment. In H. Coleman (Ed.), Multilingualisms and development (pp. 61–77).

Cabral, A. (2016). Resistance and decolonization. (D. Wood, Trans.). Rowman & Littlefield.

Clifford, J., & Marcus, G. E. (Eds.). (2023). Writing culture: The poetics and politics of ethnography. University of California Press.

Coleman, D. (2018). Digital colonialism: The 21st century scramble for Africa through the extraction and control of user data and the limitations of data protection laws. *Michigan Journal of Race & Law*, 24, 417–439. https://repository.law.umich.edu/mjrl/vol24/iss2/6

Costa-jussà, M. R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., ... & NLLB Team. (2022). No language left behind: Scaling human-centered machine translation [Preprint]. https://arxiv.org/abs/2207.04672

Couldry, N., & Mejias, U. A. (2019). Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4), 336–349.

Deep Learning Indaba. (n.d.) Retrieved on August 5, 2024, from https://deeplearningindaba.com/

de Pina-Cabral, J. (2011). Ethnography as tradition in Africa. *Etnográfica*, 15(2), 379–394. http://journals.openedition.org/etnografica/991

de Sousa Santos, B. (2014). Epistemologies of the South: Justice against epistemicide. Routledge.

Digital Umuganda. (n.d.). Retrieved on August 5, 2024, from https://digitalumuganda.com/

Dignum, V. (2023). Responsible artificial intelligence: Recommendations and lessons learned. In D. Eke, K. Wakanuma, & S. Akintoye (Eds.), Responsible AI in Africa: Challenges and opportunities (pp. 195–214). Springer International Publishing.

Distributed AI Research Institute (DAIR). (n.d.). Retrieved on August 1, 2024, from https://www.dair-institute.org/

Eades, J. S. (2012). Anthropology, political economy and world-system theory. In A Handbook of Economic Anthropology, Second Edition. Edward Elgar Publishing.

Eberhard, D., Simons, G. F., & Fennig, C. D. (Eds.). (2022). Ethnologue. Languages of Africa and Europe (25th ed.). SIL International Publications.

Eglash, R., Robinson, K. P., Bennett, A., Robert, L., & Garvin, M. (2024). Computational reparations as generative justice: Decolonial transitions to unalienated circular value flow. *Big Data & Society*, 11(1). https://doi.org/10.1177/20539517231221732

Elish, M. C., & Boyd, D. (2018). Situating methods in the magic of Big Data and AI. *Communication monographs* 85(1), 57–80. https://doi.org/10.1080/03637751.2017.1375130

Ethio NLP. (n.d.). Retrieved on July 17, 2024, from https://ethionlp.github.io/

Fardon, R., & Furniss, G. (Eds.). (1994). African languages, development and the state. Routledge.

Fletcher, R., & Nielsen, R. K. (2024, May 28). What does the public in six countries think of generative AI in news? Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-4zb8-cg87

Friedman, B., & Hendry, D. G. (2019). Value sensitive design: shaping technology with moral imagination. MIT Press.

Fung, Y., Zhao, R., Doo, J., Sun, C., & Ji, H. (2024). Massively multi-cultural knowledge acquisition & lm benchmarking [Preprint]. https://arxiv.org/abs/2402.09369

GhanaNLP. (n.d.). Retrieved on July 17, 2024, from https://ghananlp.org/

Good, J. (2020). Niger-Congo, with a special focus on Benue-Congo. In G. J. Dimmendaal & R. Vossen (Eds.), The Oxford handbook of African language (pp. 139–160). Oxford University Press.

Graham, T. (2021, April 28). Maori are trying to save their language from Big Tech. *Wired.* https://www.wired.co.uk/article/maori-language-tech

Grice, P. (1989). Studies in the way of words. Harvard University Press.

Gulley, A. L. (2022). One hundred years of cobalt production in the Democratic Republic of the Congo. *Resources Policy*, 79(103007), 1–10. https://doi.org/10.1016/j.resourpol.2022.103007

Hadgu, A. T., Azunre, P., & Gebru, T. (2023). Combating harmful hype in natural language processing. Paper presented at the International Conference on Learning Representations (ICLR) 2023, Kigali, Rwanda. https://pml4dc.github.io/iclr2023/pdf/PML4DC_ICLR2023_39.pdf

Hanover, E. (2014). Co-creating your insight: A case from rural Ghana. *Ethnographic Praxis in Industry Conference Proceedings*, 2014, 50–57. https://www.epicpeople.org/co-creating-your-insight-a-case-from-rural-ghana/

Henrich, J. P. (2020). The weirdest people in the world: How the West became psychologically peculiar and particularly prosperous (1st ed.). Farrar, Straus and Giroux.

Hoy, T., Bilal, I. M., & Liou, Z. (2023). Grounded models: The future of sensemaking in a world of generative AI. *Ethnographic Praxis in Industry Conference Proceedings*, 2023, 177–200. https://www.epicpeople.org/grounded-models-future-of-sensemaking-and-generative-ai/

Jha, A., Davani, A., Reddy, C. K., Dave, S., Prabhakaran, V., & Dev, S. (2023). SeeGULL: A stereotype benchmark with broad geo-cultural coverage leveraging generative models [Preprint]. https://arxiv.org/abs/2305.11840

Kannen, N., Ahmad, A., Andreetto, M., Prabhakaran, V., Prabhu, U., Dieng, A. B., ... & Dave, S. (2024). Beyond aesthetics: Cultural competence in text-to-image models [Preprint]. https://arxiv.org/abs/2407.06863

Karamolegkou, A., Rust, P., Cao, Y., Cui, R., Søgaard, A., & Hershcovich, D. (2024). Vision-language models under cultural and inclusive considerations [Preprint]. https://arxiv.org/abs/2407.06177

Kemp, S. (2024, January 31). Digital 2024: Global overview report. DataReportal. https://datareportal.com/reports/digital-2024-global-overview-report

Khupe, C., & Keane, M. (2017). Towards an African education research methodology: Decolonising new knowledge. *Educational Research for Social Change*, 6(1), 25–37. http://dx.doi.org/10.17159/2221-4070/2017/v6i1a3

Kwet, M. (2019). Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class*, 60(4), 3–26.

Latour, B. (1996). Aramis, or the love of technology. Harvard University Press.

Lelapa AI. (n.d.). Retrieved on August 1, 2024, from https://lelapa.ai/

Leong, C., Shandilya, H., Dossou, B. F., Tonja, A. L., Mathew, J., Omotayo, A. H., ... & Adewumi, T. (2023). Adapting to the low-resource double-bind: Investigating low-compute methods on low-resource African languages [Preprint]. https://arxiv.org/abs/2303.16985

Li, C., Chen, M., Wang, J., Sitaram, S., & Xie, X. (2024). CultureLLM: Incorporating cultural differences into large language models [Preprint]. https://arxiv.org/abs/2402.10946

Front Page Africa (2024, April 29). Liberia: Amilcar Cabral submarine cable project to boost connectivity and digital integration across West Africa [Press Release]. Retrieved July 16, 2024, from https://frontpageafricaonline.com/opinion/press-release/amilcar-cabral-submarine-cable-project-to-boost-connectivity-and-digital-integration-across-west-africa/

Liu, Z., Currier, K., & Janowicz, K. (2024). Making geographic space explicit in probing multimodal large language models for cultural subjects. Accepted as a Provocation at the Global AI Cultures workshop, International Conference on Learning Representations (ICLR) 2024, Vienna, Austria. https://globalaicultures.github.io/pdf/15_making_geographic_space_explic.pdf

Lüpke, F. (2010). Multilingualism and language contact in West Africa: Towards a holistic perspective. *Journal of Language Contact*, 3(1), 1–12.

Lüpke, F., & Storch, A. (2013). Repertoires and choices in African languages (Language Contact and Bilingualism, Vol. 5). De Gruyter Mouton.

Machine Translate. (n.d.). Retrieved July 16, 2024, from https://machinetranslate.org/

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges [Preprint]. https://arxiv.org/abs/2006.07264

Manfredi, S., & Tosco, M. (2016). A new state, an old language policy, and a Pidgincreole: Juba Arabic in South Sudan. Preprint accessed at HAL Open Science. https://shs.hal.science/halshs-01357537

Masakhane. (n.d.). Retrieved on July 17, 2024, from https://www.masakhane.io/home

Masoud, R. I., Liu, Z., Ferianc, M., Treleaven, P., & Rodrigues, M. (2023). Cultural alignment in large language models: an explanatory analysis based on Hofstede's cultural dimensions [Preprint]. https://arxiv.org/abs/2309.12342

Mastel, P. M, Namara, E., Munezero, A., Kagame, R., Wang, Z., Anzagira, A., Gupta, A., & Ndibwile, J. D. (2023). Natural language understanding for African languages. *Proceedings of AfricaNLP 2023*, 1–9. https://openreview.net/pdf?id=gWuvdFMqHM

Mavhunga, C. C. (2017). What do science, technology, and innovation mean from Africa? The MIT Press.

Mavhunga, C. C. (2014). Transient workspaces: Technologies of everyday innovation in Zimbabwe. The MIT Press.

Medina, E., da Costa Marques, I., Holmes, C., & Cueto, M. (2014). Beyond imported magic. The MIT Press.

Mejias, U. A., & Couldry, N. (2024). Data grab: The new colonialism of big tech and how to fight back. University of Chicago Press.

Mesthrie, R. (2019, February 25). African Englishes from a sociolinguistic perspective. Oxford Research Encyclopedia of Linguistics. https://doi.org/10.1093/acrefore/9780199384655.013.225

Moayeri, M., Tabassi, E., & Feizi, S. (2024, June). WorldBench: Quantifying geographic disparities in LLM factual recall. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1211–1228).

Moro, F. (2023, September 19). Yan Zongo: A research note on Accra's strangers. *The Metropole: The Official Blog of the Urban History Association*. https://themetropole.blog/2023/09/19/yan-zongo-a-research-note-on-accras-strangers/

Moss, E., & Schüür, F. (2018). How modes of myth-making affect the particulars of DS/ML adoption in industry. *Ethnographic Praxis in Industry Conference Proceedings*, 2018 264–280. https://epicpeople.org/myth-making/

Mwinlaaru, I. N., Xuan, W. W. (2016). A survey of studies in systemic functional language description and typology. *Functional Linguist* 3(8). https://doi.org/10.1186/s40554-016-0030-4

Nayebare, M., Eglash, R., Kimanuka, U., Baguma, R., Mounsey, J., & Maina, C. (2023, September 29). Interim report for Ubuntu-AI: A bottom-up approach to more democratic and equitable training and outcomes for machine learning. Democratic Inputs for AI, OpenAI Foundation, San Francisco.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T., Fagbohungbe, T., Akinola, S. O., Muhammad, S., Kabenamualu, S. K., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L. (2020). Participatory research for low-resourced machine translation: A case study in African languages. *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2144–2160. https://aclanthology.org/2020.findings-emnlp.195

Nigeria Federal Ministry of Communications, Innovation & Digital Economy (2024, April 19). Dr. Tijani wraps up artificial intelligence strategy workshop; announces groundbreaking partnerships with Cisco, 21st entury Technologies, others [Press release]. Retrieved August 5, 2024, from https://fmcide.gov.ng/hon-minister-dr-bosun-tijani-wraps-up-artificial-intelligence-strategy-workshop-announces-groundbreaking-partnerships-with-cisco-21st-century-technologies-others/

OER Africa. (2023). Artificial intelligence and the underrepresentation of African languages. https://www.oerafrica.org/content/artificial-intelligence-and-underrepresentation-african-languages

Ojo, J., & Ogueji, K. (2023). How good are commercial large language models on African languages? [Preprint]. https://arxiv.org/abs/2305.06530

O'Neill, J., Marivate, V., Glover, B., Karanu, W., Tadesse, G. A., Gyekye, A., Makena, A., Rosslyn-Smith, W., Grollnek, M., Wayua, C., Baguma, R., Maduke, A., Spencer, S., Kandie, D., Maari, D., Mutangana, N., Axmed, M., Kamau, N., Adamu, M., & Nyairo, S. (2024, June). AI and the future of work in Africa [White Paper]. https://doi.org/10.13140/RG.2.2.10251.91683

Orji, D. M., & Umeobi, C. A. (2023). Language dynamism and nation building: A focus on the English language. *Journal of Linguistics, Language and Culture*, 10(2), 120–141. https://www.nigerianjournalsonline.com/index.php/jollc/article/view/4431/4295

Oyewusi, W. (2024). AI literacy in low-resource languages: Insights from creating AI in Yoruba videos [Preprint]. https://arxiv.org/abs/2403.04799

Peace Corps. (1975). Oromo for beginners. https://fsi-languages.yojik.eu/languages/PeaceCorps/Oromo/ED226615.pdf

Pfohl, S. R., Cole-Lewis, H., Sayres, R., Neal, D., Asiedu, M., Dieng, A., ... & Singhal, K. (2024). A toolbox for surfacing health equity harms and biases in large language models [Preprint]. https://arxiv.org/abs/2403.12025

Pool, J. (1991). The official language problem. *American Political Science Review*, 85(2), 495–514. https://doi.org/10.2307/1963171

Prah, K. K. (2009). A tale of two cities: Trends in multilingualism in two African cities: The cases of Nima-Accra and Katutura-Windhoek. In K. K. Prah and B. Brock-Utne (Eds.), Multilingualism—An African Advantage: A Paradigm Shift in African Language of Instruction Policies. *CASAS*, pp. 250–275).

Rasberry, V. (2021, September 30). In search of African Arabic. *New Lines Magazine*. https://newlinesmag.com/essays/in-search-of-african-arabic/

Ritchie, S., Cheng, Y. C., Chen, M., Mathews, R., van Esch, D., Li, B., & Sim, K. C. (2022). Large vocabulary speech recognition for languages of Africa: Multilingual modeling and self-supervised learning [Preprint]. https://arxiv.org/abs/2208.03067

Roberts, S., Hackett, E., Karol, S., & Baron, D. (2023). Favoring friction: Examining hotel browsing and buying through the lens of the bazaar. *Ethnographic Praxis in Industry Conference Proceedings*, 2023, 136–162. https://epicpeople.org/favoring-friction-hotel-browsing-and-buying/

Robinson, M. J. (2024). A language for the world: the standardization of Swahili. Ohio University Press.

Sampson, O. (2023). Pulp friction: Creating space for cultural context, values, and the quirkiness of humanity in AI engagement [Conference presentation]. Ethnographic Praxis in Industry Conference, Chicago, IL, United States. https://www.epicpeople.org/pulp-friction-creating-space-for-cultural-context-values-humanity-in-ai/

Shah, C., & Bender, E. M. (2024). Envisioning information access systems: What makes for good tools and a healthy web? *ACM Transactions on the Web*, 18(3), 1–24. https://dl.acm.org/doi/10.1145/3649468

Shafayat, S., Kim, E., Oh, J., & Oh, A. (2024). Multi-FAct: Assessing multilingual LLMs' multi-regional knowledge using FActScore [Preprint]. https://arxiv.org/abs/2402.18045

Spivak, G. (2003). Can the subaltern speak? *Die Philosophin* 14(27), 42–58.

Stoller, P. (2023). Wisdom from the edge: Writing ethnography in turbulent times. Cornell University Press.

Suresh, H., Tseng, E., Young, M., Gray, M., Pierson, E., & Levy, K. (2024, June). Participation in the age of foundation models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1609–1621). https://dl.acm.org/doi/pdf/10.1145/3630106.3658992

Táíwò, O. (2022). Against decolonisation: Taking African agency seriously. Hurst & Company.

Talat, Z., Névéol, A., Biderman, S., Clinciu, M., Dey, M., Longpre, S., ... & Van Der Wal, O. (2022, May). You reap what you sow: On the challenges of bias evaluation under multilingual settings. Proceedings of BigScience Episode #5 -- Workshop on Challenges & Perspectives in Creating Large

Language Models. Association for Computational Linguistics, pp. 26–41. https://aclanthology.org/2022.bigscience-1.3/

Tamkin, A., Askell, A., Lovitt, L., Durmus, E., Joseph, N., Kravec, S., ... & Ganguli, D. (2023). Evaluating and mitigating discrimination in language model decisions [Preprint]. https://arxiv.org/abs/2312.03689

Teixeira da Silva, J. A. (2022). Handling ethics dumping and neo-colonial research: From the laboratory to the academic literature. *Bioethical Inquiry*, 19, 433–443. https://doi.org/10.1007/s11673-022-10191-x

Tilmatine, M. (2015). Arabization and linguistic domination: Berber and Arabic in the North of Africa. In C. Stolz (Ed.), *Language Empires in Comparative Perspective* (pp. 1–16). De Gruyter, https://doi.org/10.1515/9783110408362.1

Tonja, A. L., Azime, I. A., Belay, T. D., Yigezu, M. G., Mehamed, M. A., Ayele, A. A., ... & Yimam, S. M. (2024). EthioLLM: Multilingual large language models for Ethiopian languages with task evaluation [Preprint]. https://arxiv.org/abs/2403.13737

Turner, I. (2023). Decolonisation through digitalisation? African languages at South African universities. *Curriculum Perspectives* 43, Suppl 1, 73–82. https://doi.org/10.1007/s41297-023-00196-w

UNESCO & Moroccan International Centre for Artificial Intelligence (2024, June 4). Rabat consensus on artificial intelligence: A call to action [Program and meeting document]. https://unesdoc.unesco.org/ark:/48223/pf0000390399

United States Congressional-Executive Commission on China. (2023, November 14). From cobalt to cars: How China exploits child and forced labor in DR Congo. Hearing before the Congressional-Executive Commission on China, One Hundred Eighteenth Congress, first session. U.S. Government Publishing Office. https://www.congress.gov/118/chrg/CHRG-118jhrg54083/CHRG-118jhrg54083.pdf

Wachter, S., Mittelstadt, B., & Russell, C. (2024, January 31). Do large language models have a legal duty to tell the truth? Royal Society Open Science. http://dx.doi.org/10.2139/ssrn.4771884

Watts, I., Gumma, V., Yadavalli, A., Seshadri, V., Swaminathan, M., & Sitaram, S. (2024). PARIKSHA: A large-scale investigation of human-LLM evaluator agreement on multilingual and multi-cultural data [Preprint]. https://arxiv.org/abs/2406.15053

Web Technology Surveys. (n.d.). Usage statistics of content languages for websites. Retrieved July 16, 2024, from https://w3techs.com/technologies/overview/content_language

Wendler, C., Veselovsky, V., Monea, G., & West, R. (2024). Do llamas work in English?: On the latent language of multilingual transformers [Preprint]. https://arxiv.org/abs/2402.10588

Williams, R. (1974). Television: Technology and the cultural form. Fontana.

Wiredu, K. (1997). The need for conceptual decolonization in African philosophy. In Philosophy and Democracy in Intercultural Perspective, 11–22. Brill. https://doi.org/10.1163/9789004457997_002

World Economic Forum. (2020, September). Making mining safe and fair: Artisanal cobalt extraction in the Democratic Republic of the Congo [White paper]. https://www3.weforum.org/docs/WEF_Making_Mining_Safe_2020.pdf

Yao, X. (2024, May 6). The superpower of "en-US": "en" vs. the under-represented languages [Post]. LinkedIn. https://www.linkedin.com/pulse/superpower-en-us-en-vs-under-represented-languages-xuchen-yao-61vmf/

Yiu, E., Kosoy, E., & Gopnik, A. (2023). Transmission versus truth, imitation versus innovation: What children can do that large language and language-and-vision models cannot (yet). *Perspectives on Psychological Science*, 19(5), 874–883. https://doi.org/10.1177/17456916231201401

Zamenopoulos, T., & Alexiou, K. (2018). Co-design as collaborative research. Connected Communities Foundation Series. Bristol University/AHRC Connected Communities Programme.

Zupon, A., Crew, E., & Ritchie, S. (2021). Text normalization for low-resource languages of Africa [Preprint]. https://arxiv.org/abs/2103.15845