Mind the (Language) Gap:

Mapping the Challenges of LLM Development in Low-Resource Language Contexts

Juan N. Pava, Caroline Meinhardt, Haifa Badi Uz Zaman,
Toni Friedman, Sang T. Truong, Daniel Zhang, Elena Cryst,
Vukosi Marivate, Sanmi Koyejo

Stanford University
Human-Centered
Artificial Intelligence

The Asia Foundation

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Authors

Juan N. Pava is a research fellow in the Tech Ethics and Policy Rising Scholars Program at Stanford University's McCoy Family Center for Ethics in Society. At the Stanford Institute for Human-Centered Artificial Intelligence (HAI), he works at the intersection of AI, the social sector, and the Global South. His research interests include the political economy of emerging economies and its relationships with political philosophy and ethics. He holds a bachelor's degree in philosophy and economics from New York University.

Caroline Meinhardt is the policy research manager at Stanford HAI, where she develops and oversees policy research initiatives. Her research focuses on the implementation challenges of AI regulation, the governance of large-scale AI models, and global AI governance approaches. Prior to joining HAI, she worked as a China-focused consultant and analyst, managing and delivering in-depth research and strategic advice regarding China's development and regulation of emerging technologies, including AI. She holds a bachelor's degree in Chinese studies from the University of Cambridge and a master's degree in international policy from Stanford University.

Haifa Badi Uz Zaman is a manager of AI partnerships on the innovation team at The Rockefeller Foundation. Previously, she was a program manager at Stanford HAI, where she designed and led bespoke initiatives to empower social sector organizations and leaders with AI knowledge through publications, convenings, and education programs. She also managed research programs and grants at Stanford University's Center on Philanthropy and Civil Society and the Cyber Policy Center. She holds a master's degree in international education policy from the Harvard Graduate School of Education.

Toni Friedman is an assistant director of digital technology, policy, and innovation at The Asia Foundation, where she supports projects on AI policy, internet governance, and cybersecurity across the foundation's country offices. Her work explores understanding the impacts of AI on lower- and middle-income countries. She previously lived and worked in Shanghai, managing charity programs. She holds a bachelor's degree in international studies from the University of Chicago and a master's degree in international policy from Stanford University.

Sang T. Truong is a PhD candidate in computer science at Stanford University. He develops foundational and applied methods in probabilistic machine learning, emphasizing questions on measurement, preference, and decision. Specifically, Truong leverages generative models and adaptive algorithms to design efficient and reliable systems capable of operating in real-world environments. He has applied these methods downstream to challenging applications in machine learning systems and education.

Daniel Zhang is the senior manager for policy initiatives at Stanford HAI, where he leads the institute's policy research, outreach, and education initiatives. He is also a member of the High-Level Expert Group on AI Ethics at UNESCO, advising the agency on the implementation of its Recommendation on the Ethics of AI. Previously, he was the manager of the AI Index, where he lead-authored the 2021 and 2022 annual reports that measure and evaluate the rapid rate of AI advancement. Before Stanford, he worked at the Center for Security and Emerging Technology. He holds a master's degree in security studies from Georgetown University and a bachelor's degree in politics and international affairs from Furman University.

Authors (cont'd)

Elena Cryst is the director of policy and society at Stanford HAI, where she leads the institute's efforts to bring Stanford's cutting-edge AI research to policymakers worldwide. She also builds collaborations with civil society, philanthropy, and social impact leaders to understand how to better identify the concerns and passions of these communities with the development of these technologies. Cryst previously served in director roles at both the Stanford Institute for Economic Policy Research and the Freeman Spogli Institute for International Studies. She received her BA with honors in international relations and an MA in Latin American studies, both from Stanford University, and an MBA from the University of California, Berkeley, Haas School of Business.

Vukosi Marivate is a professor of computer science and holds the ABSA UP Chair of Data Science at the University of Pretoria, where he also leads the Data Science for Social Impact (DSFSI) lab. His research is dedicated to improving the methods, tools, and availability of data for local or low-resource languages. Marivate is a co-founder of Lelapa AI, an African startup focused on AI for Africans by Africans. He is a co-founder of the Masakhane Research Foundation, which aims to develop NLP technologies for African languages. Vukosi is also a co-founder of the Deep Learning Indaba, the leading grassroots machine learning and artificial intelligence conference on the African continent dedicated to empowering and supporting African researchers and practitioners in the field.

Sanmi Koyejo is an assistant professor in the department of computer science at Stanford University and a faculty affiliate at Stanford HAI. Koyejo leads the Stanford Trustworthy AI Research (STAIR) lab, which works to develop the principles and practice of trustworthy AI, focusing on applications to science and health care. Koyejo has received the Skip Ellis Early Career Award, Presidential Early Career Award for Scientists and Engineers (PECASE), and Sloan Fellowship, among other honors. Koyejo serves on the Neural Information Processing Systems Foundation Board and as president of the Black in AI organization.

Acknowledgments

The authors would like to thank Thomas Mullaney, Tracy Navichoque, Tolúlọpẹ́ Ògúnrẹ̀mí, Diyi Yang, and participants in the Rising Scholars Workshop organized by the McCoy Family Center for Ethics in Society for their valuable comments and feedback; and Jeanina Casusi, Chris Ellis, Nancy King, Carolyn Lehman, Shana Lynch, and Michi Turner for their help preparing the publication.

Table of Contents

Executive Summary

- Large language model (LLM) development suffers from a digital divide: Most major LLMs underperform for non-English—and especially low-resource—languages; are not attuned to relevant cultural contexts; and are not accessible in parts of the Global South.

- Low-resource languages (such as Swahili or Burmese) face two crucial limitations: a scarcity of labeled and unlabeled language data and poor quality data that is not sufficiently representative of the languages and their sociocultural contexts.

- To bridge these gaps, researchers and developers are exploring different technical approaches to developing LLMs that better perform for and represent low-resource languages but come with different trade-offs:

    - **Massively multilingual models**, developed primarily by large U.S.-based firms, aim to improve performance for more languages by including a wider range of (100-plus) languages in their training datasets.

    - **Regional multilingual models**, developed by academics, governments, and nonprofits in the Global South, use smaller training datasets made up of 10-20 low-resource languages to better cater to and represent a smaller group of languages and cultures.

    - **Monolingual or monocultural models**, developed by a variety of public and private actors, are trained on or fine-tuned for a single low-resource language and thus tailored to perform well for that language.

- Other efforts aim to address the underlying data scarcity problem by focusing on generating more language data and assembling more diverse labeled datasets:

    - Advanced **machine translation models** enable the low-cost production of raw, unlabeled data in low-resource languages, but the resulting data may lack linguistic precision and contextual cultural understanding.

    - **Automated or semi-automated approaches** can help streamline the process of labeling raw data, while **participatory approaches** that engage native speakers of low-resource languages throughout the entire LLM development cycle empower local communities while ensuring more accurate, diverse, and culturally representative LLMs.

- It is crucial to understand both the underlying reasons for and paths to addressing these disparities to ensure that low-resource language communities are not disproportionately disadvantaged by and can equally contribute to and benefit from these models.

- We present three overarching recommendations for AI researchers, funders, policymakers, and civil society organizations looking to support efforts to close the LLM divide:

    - **Invest strategically in AI development for low-resource languages**, including subsidizing cloud and computing resources, funding research that increases the availability and quality of low-resource language data, and supporting programs to promote research at the intersection of these issue areas.

    - **Promote participatory research** that is conducted in direct collaboration with low-resource language communities, who contribute to and even co-own the creation of AI resources.

    - **Incentivize and support the creation of equitable data ownership frameworks** that facilitate access to AI training data for developers while protecting the data rights of low-resource language data subjects and creators.

1. Introduction: The AI Language Divide

Scholars have long studied the digital divide and its impact on global inequality. Initially considered with regard to disparities in access to and usage of the internet and digital technologies such as laptops or smartphones, the term has since come to encompass the broader socioeconomic inequalities associated with many technological breakthroughs.

When it comes to AI, the digital divide permeates every layer of the AI stack.[1] It manifests, for example, in systemic barriers to accessing foundational infrastructure needed for AI development such as computing resources,[2] the limited availability of high-skilled technical talent,[3] as well as biased and uneven model performance across different linguistic and cultural contexts.[4]

This divide is especially pronounced in the development of large language models (LLMs). Most major LLMs are predominantly trained using English (or other high-resource language) data and not attuned to Global South[i]—or Majority World—contexts. As a result, they underperform for many non-English languages,[5] are often biased in favor of native English speakers,[6] and are not accessible in parts of the Global South.[7] A root cause is the vast gap in the availability and cost of data resources: While two-thirds of the world's languages are spoken in Africa and Asia,[8] most of these languages lack substantial digital support and other language resources.[9] Data is both more expensive to work with and less representative for many of these languages.[10] This disparity is commonly referred to as the "resourcedness gap," whereby low-resource languages, also referred to as under-resourced or digitally disadvantaged languages, are characterized by two key limitations:

Most major LLMs are predominantly trained using English (or other high-resource language) data and not attuned to Global South —or Majority World—contexts.

- **Quantity gap:** Low-resource languages suffer from an insufficient quantity of both labeled and unlabeled data. The data that does exist is often mislabeled,[11] not reflective of the number of speakers of said language,[12] or unsuitable for natural language processing (NLP) purposes (e.g., pornographic, nonsensical, or non-linguistic content).[13] Languages with non-Latin scripts face additional digitization challenges.[14] This data scarcity is a key reason English models consistently outperform non-English ones.[15]

- **Quality gap:** Data in low-resource languages is often of poor quality due to a lack of diverse sources. It is typically restricted to the Bible and other religious texts,[16] legal documents,

i. We choose to use the broad term Global South (over terms such as Global Majority) for simplicity and readability. While we acknowledge its limitations, we adopt this commonly used term as a shorthand to highlight disparities in LLM development for low-resource languages in technologically under-resourced geographies. These challenges transcend borders, of course, and also affect low-resource language speakers in more technologically well-resourced countries (such as Cherokee speakers in the United States), but these are not the focus of this paper.

and Wikipedia articles—many of which are hardly reflective of casual speech and/or themselves machine-translated. This challenge is exacerbated by the well-documented lack of non-English languages in high-quality scientific and educational literature.[17] The problem is even more pronounced with non-Latin scripts and languages spoken in the Global South.[18]

On a more fundamental level, the quantity and quality of labeled and unlabeled data for any given language is determined by a range of sociopolitical factors, including the state of digital infrastructure and technology adoption in relevant language communities. For example, the inclusion of a language in script encoding standards (such as the Unicode Standard), the availability of usable digital typefaces and keyboards for that language, as well as speakers' access to computers, mobile devices, and the internet more broadly all affect the availability of language data.[19] Less than 5 percent of the roughly 7,000 languages spoken around the world today have meaningful online representation, leading to what has been called a "digital language death"—when the lack of technological support for a particular language precipitates its decline.[20] Today, languages that are not fully digitally supported are less likely to be recorded—never mind included in AI model training datasets.

As a result, language resourcedness varies widely. At one end of the spectrum, there is English: Spoken by approximately 1.52 billion people, it functions as both the global and digital lingua franca. English is extremely well resourced: It dominates the internet, with nearly half of all websites written in English,[21] and, consequently, many of the most advanced NLP tools were built using and tailored for English.[22] At the other end, there are languages like numma guhooni—an Indigenous, endangered language spoken by fewer than 400 individuals in Kenya[23]—that have no significant digital presence or technological support.[24] Between these extremes, there is a large number of low-resource languages with varying data availability. For instance, languages like Vietnamese (with around 97 million speakers) and Hausa (94 million speakers) are considered low-resource languages despite having a large speaking population because they lack the depth and breadth of digital resources to support most advanced computational tasks.[25] Other low-resource languages like Nahuatl (around 1.5 million speakers) possess even fewer digitized language resources—mostly unlabeled data—in part because these languages are not broadly sustained by institutions beyond the home and community.[26]

Understanding the gap in LLM development is crucial to ensuring that low-resource language communities are not disproportionately disadvantaged by and can equally contribute to and benefit from these models.

Low-resource languages are not exclusive to the Global South; for instance, Western Europe and the United States are home to various low-resource languages and dialects (such as Welsh, Basque, or Cherokee). However, the resource gap is often compounded in Global South countries by a lack of sufficient AI literacy, talent, and computing resources.[27] Low digital literacy and limited internet access restrict data generation, while a shortage of advanced expertise stifles model development. Computing resource constraints make local NLP communities financially unsustainable. The resource gap is particularly severe when the causes are a combination of concerns such as data limitations and compute constraints,[28] which are seldom explored jointly. The result? With some exceptions,[29] most NLP research on Global South languages is conducted in Global North institutions,[30] where research biases often lead to low-resource language research needs being overlooked.[31]

In this paper, we focus on the challenges faced by low-resource languages in the Global South. Understanding the gap in LLM development is crucial to ensuring that low-resource language communities are not disproportionately disadvantaged by and can equally contribute to and benefit from these models. Yet addressing the LLM development gap requires a careful consideration of the various trade-offs. Given the resource intensity and vast environmental impact of LLM development, localized LLM development may not always be the most feasible or suitable path forward.

2. Closing the Model Gap

To bridge the wide-ranging LLM resource gaps, researchers and developers around the world are exploring different technical approaches to developing language models that better perform for and represent a greater variety of languages and cultural contexts. These models vary greatly in their language capabilities, training data size, technical performance, and level of openness (i.e., whether developers provide access to underlying code and/or data or restrict access to their models).

In this section, we explore three of the most prominent approaches currently being pursued. We summarize our findings in Table 1 before expanding on each approach.

Table 1: Summary of key approaches to closing the LLM divide

	Massively multilingual models	Regional multilingual models	Monolingual & monocultural models
Developers	Private industry (primarily large U.S.-based firms)	Academia, government, nonprofits	Academia, government, nonprofits, private industry
Languages in training data	100+	10-20	1-2
Examples	mBERT, mT5, XLM-R, mDeBERTa, Aya	AfriBERTa, BantuBERTa, SEA-LION, IndicBERT, Sailor	PuoBERTa, SwahBERT, UlizaLlama, IndoBERT, URA-LLaMa, MixSUra, GemSUra, Typhoon
Pros	• Enable cross-lingual transfer learning • More functional than maintaining multiple models for each language • Scalable across language families • Including higher-resource languages enables complex NLP tasks • Proven track record in speech LLMs	• Enable cross-lingual transfer learning • More cost-efficient than maintaining multiple models for each language • Scalable across language families • Competitive performance for basic NLP tasks • More culturally tailored and reduce linguistic biases • Lower compute costs due to smaller size	• Avoid the curse of multilinguality (where better multilingual performance may hurt performance in individual languages) • Perform better for languages with minimal data availability • Capture nuanced linguistic features specific to a language
Cons	• Suffer from the curse of multilinguality • Tend to be more opaque • Prone to cross-lingual vulnerabilities • Less culturally nuanced, open to bias	• Trade-off between adding similar languages and the curse of multilinguality remains unresolved • Underperform in domain-specific NLP tasks • Lack of comprehensive benchmarks • Early stage of development, effectiveness not yet fully proven	• Lack of cross-lingual transfer learning • Insufficient data for training, especially for higher-order NLP tasks • Training from scratch is computationally expensive

Massively multilingual models developed in the Global North

Researchers have in recent years prioritized the development of multilingual language models that perform better for a wider variety of languages. Thus far, the largest multilingual models, often referred to as massively multilingual models,[ii] have been developed primarily by large firms in the United States using multilingual data from over 100 languages.[32] They include Google's mBERT and its more recent mT5, Meta's XLM-R, and Microsoft's mDeBERTa,[33] as well as Aya, developed by Cohere's nonprofit research lab.[34]

Researchers have in recent years prioritized the development of multilingual language models that perform better for a wider variety of languages.

These models aim to improve performance for more languages by including a wider range of languages in their training datasets. They rely on cross-lingual transfer learning, whereby a model improves its low-resource language performance by inferring universal or shared linguistic patterns from high-resource languages.[35] In principle, the larger size of the training dataset gives the models more advanced AI capabilities for low-resource languages, such as coding or complex mathematical reasoning.

One early example of this approach is mBERT, the multilingual version of Google's BERT foundation model. Unlike BERT, which was trained solely on English data, the subsequently released mBERT was trained on Wikipedia content in the 104 languages with the largest Wikipedia repositories.[36] When training mBERT, Google purposefully over-sampled data from low-resource languages to account for the massive disparity in Wikipedia content available.[37]

Research has shown that multilingual *speech* models (i.e., models such as Meta's XLSR–53 that can process and/or generate speech)[38] perform better for individual languages than their monolingual counterparts with similarly sized training datasets.[39] However, the results for *text* models (i.e., models that only process or generate text) have been mixed.[40] While some multilingual text models do outperform their low-resource monolingual counterparts—especially when comparing performance across more complex NLP tasks—these successes have not yet been conclusive or are not generalizable to all models.[41] Nonetheless, a key advantage of the multilingual model approach is operational: Maintaining a single massively multilingual model can often be simpler than managing multiple monolingual ones.[42]

Still, multilingual models come with their own set of challenges. They are even more opaque than monolingual foundation models[43]—researchers still do not fully understand how models connect different languages, making model failures harder to diagnose and exposing models to additional cross-lingual vulnerabilities.[44] Also, these models face the so-called "curse of multilinguality," whereby, after a given point, improved performance in more languages comes at the expense of performance in others,[45] including

ii. Other names commonly used to refer to these models include "extremely large language models," "massively multilingual transformers (MMTs)," and "massively multilingual language models."

low-resource languages.[46] This partially explains why these massive models are typically capped at around 100 languages. Increasing model size can mitigate this concern and may explain the recent explosive growth in multilingual model scale. But this expansion makes an already computationally expensive and data-hungry approach even more costly and inaccessible to small, under-resourced research teams.[47] In light of these challenges, a series of alternative solutions have begun to emerge.

Regional multilingual models developed in the Global South

Efforts to build multilingual models have begun to emerge in the Global South in recent years, often motivated by local actors wanting to fill a language capability gap and maintain a certain level of autonomy over LLMs that will be used in their countries and contexts. These multilingual models, which are developed by a wide range of actors, frequently outside of the private sector, tend to be smaller: They are generally trained on multilingual data from around 10 to 20 languages, grouped by geographical or linguistic proximity.

Regional multilingual model development efforts generally follow two broad approaches. The first involves using the architecture of foundation models, often BERT models, to train a new multilingual model from scratch, specifically using language data from a small group of low-resource languages. Examples include AfriBERTa (for African languages),[48] the original version of SEA-LION (for Southeast Asian languages),[49] and IndicBERT (for Indian languages).[50] Alternatively, projects may take an off-the-shelf foundation model—already pre-trained on large-scale, unlabeled data—and further train it on a small group of low-resource languages in a process known as fine-tuning. For example, the second and third versions of SEA-LION fine-tuned Meta's Llama 3 and Google's Gemma 2 models, respectively, on data from several Southeast Asian languages.[51] Meanwhile, Sailor, another LLM developed for Southeast Asian languages, was built on Qwen1.5, a foundation model developed by Chinese company Alibaba, through a process that extended its pre-training with region-specific data.[52]

SEA-LION (short for Southeast Asian Languages In One Network) is a noteworthy example of a regional multilingual model project. Spearheaded by the Singaporean government's national AI R&D program, AI Singapore, with the support of private players including Amazon Web Services, Google Research, and IBM, aims to better cater to under-represented Southeast Asian population groups and their diverse languages and cultures.[53] It offers a regional alternative to models developed by large and well-resourced companies in the United States or China. The project encompasses a family of multilingual models for the Southeast Asian region. It now covers 13 high- and low-resource languages prevalent in the region, including English, Chinese, Indonesian, Malay, Thai, Vietnamese, Filipino, Tamil, Burmese, Khmer, and Lao, as well as Javanese and Sundanese. The project claims that the newest and largest version of SEA-LION performs better than or equal to Gemma 2 (the model it was fine-tuned on) in regional languages while still retaining Gemma 2's general capabilities.[54] Using a holistic evaluation benchmark,[55] researchers found that a prior version of the model performs better in the Indonesian language than Llama 2 and other models of the same size.[56] But when tested on English tasks, SEA-LION achieved middling results, showing that it is specialized for Southeast Asian use cases.

By comparison, AfriBERTa is an academic research project that came out of Canada's University of Waterloo.[57] The model was trained on 11 African languages—Afaan Oromoo, Amharic, Gahuza (consisting of Kinyarwanda and Kirundi), Hausa, Igbo, Nigerian Pidgin, Somali, Swahili, Tigrinya, and Yorùbá—spoken by more than 400 million people. Its developers found that overall AfriBERTa is competitive across all languages and that the model outperforms mBERT and XLM-R in several languages on several NLP tasks.[58]

The private sector has also been active within the African region. InkubaLM, for example, was trained on five African languages along with English and French by Lelapa, an Africa-centric AI research and product lab co-founded by one of our authors.[59] Other notable examples from the region are AfroXLMR-large,[60] which adapted Meta's XLM-R-Large for 17 African languages, and AfroLM,[61] a multilingual language model pre-trained from scratch on 23 African languages.

These examples show that it is possible to develop competitive multilingual language models using smaller training datasets made up solely of low-resource languages.[62] But regional multilingual language model projects have also encountered performance challenges. For example, while BantuBERTa[63]—a multilingual language model developed for the family of Bantu languages spoken across Central, Southern, Eastern, and Southeast Africa—was shown to perform well across several Bantu languages on simpler NLP tasks, it underperformed on more complex NLP tasks compared to other models.[64]

It is possible to develop competitive multilingual language models using smaller training datasets made up solely of low-resource languages.

As many of these models are still in the early phases of development, it remains to be seen whether they can remain technically competitive in the long term, how they will be integrated economically in their respective regions, and how access to and governance of these models will be managed.

Monolingual and monocultural models

Public and private institutions have also moved to develop models tailored to individual low-resource languages to improve model performance on these languages. However, monolingual efforts come with significant trade-offs as model performance remains constrained by the quality and quantity of available data in these languages.

Generally speaking, developers follow the same two approaches mentioned for regional multilingual models. Some use the architecture of BERT or other models to train a new model specialized in one target low-resource language. Examples include PuoBERTa (for Setswana),[65] SwahBERT or UlizaLlama (for Swahili),[66] Typhoon (for Thai),[67] and IndoBERT (for Indonesian).[68] Training models from scratch on local data offers flexibility for researchers to build their own custom pre-trained models in low-resource languages, which can later be fine-tuned for specific downstream tasks within the target language.

Others take an off-the-shelf pre-trained model and fine-tune it on additional data from the target lower-resource language. For example, a joint initiative by Stanford University (including two of our authors) and VNU-HCM University of Technology has released five Vietnamese-centric large language models fine-tuned on Meta's LLaMa 2, Mistral AI's Mix-tral 8×7B, and Google's Gemma.[69] This approach allowed the researchers to train high-performing models despite a scarcity of training datasets in Vietnamese and limited computational resources.

Monolingual models can, at least in principle, avoid some of the linguistic and cultural biases found in English-centric models.

Like the regional multilingual models discussed earlier, these monolingual models can, at least in principle, avoid some of the linguistic and cultural biases found in English-centric models,[70] such as the overrepresentation of Western concepts or a hardwired representational bias toward high-resource languages. Unlike the multilingual approach, the advantage of monolingual models is that they avoid having their capacity diluted across multiple languages.[71] That is, they are able to dedicate the entire model capacity to the target language, thereby steering clear of the "curse of multilinguality" discussed above.

This does not mean, however, that monolingual models are uniformly superior. Research has shown that even LLMs tailored to one non-English language can exhibit cultural biases toward entities associated with Western culture.[72] And evidence suggests that in contexts with low data availability, massively multilingual models outperform monolingual models fine-tuned from foundation models.[73] When comparing model performance across a wide array of NLP tasks, researchers found that low-resource language performance is not necessarily tied to model size but rather to the limited quantity and diversity of training in fine-tuning datasets,[74] underscoring data scarcity as a key bottleneck in advancing AI capabilities in the Global South. Ultimately, while having more monolingual data could potentially diminish the advantages of training a multilingual model, more research is needed to identify where this threshold lies.[75] Unfortunately, for the lowest-resource languages, there may not be enough data to effectively train a monolingual model, and even if sufficient data were available, it may not be adequately diverse or representative of casual speech. Many pre-training datasets are not well suited to build culturally aware language models.[76]

Training with more data may also be prohibitively expensive without guaranteeing superior performance.[77] Not only does it cost more to process every unit of data for low-resource languages, but oftentimes Global South researchers have to collect their own tailored local language data. Given these limitations, research is increasingly focused on solving the underlying data availability problem.

3. Closing the Underlying Data Gap

A variety of research and community initiatives have recently emerged that aim to directly address the data scarcity problem. These efforts focus on large-scale data production in low-resource languages through machine translation, creating bilingual datasets, and developing more data-efficient techniques.[78] A wide range of actors, including large U.S.-based firms, Global South NLP communities, and governments, are leading these projects to bridge the language divide. Below, we introduce a few such efforts.

By automatically translating text from English (or other high-resource languages) into low-resource languages, machine translation helps generate additional training material for AI systems.

Machine translation: An alternative to LLMs tailored to low-resource languages?

In recent years, major advances in the capabilities of machine translation technology have made them effective and low-cost tools for increasing the production of raw, unlabeled data in low-resource languages.[79] By automatically translating text from English (or other high-resource languages) into low-resource languages, machine translation helps generate additional training material for AI systems. Some researchers even argue that these tools could render monolingual low-resource language models obsolete.[80]

There are two main, relatively nascent, approaches to machine translation for model training.

The *translate-train* approach uses existing machine translation models with strong low-resource language translation capabilities (such as Meta's No Language Left Behind models)[81] to translate English text into a target low-resource language. The translated texts are then used to fine-tune a multilingual model for a given task, often in combination with real-world data from that target language. For instance, researchers aiming to analyze emotions on Swahili social media platforms could use this translate-train approach to train an emotion classification AI model using predominantly English language data.

The second approach, the *translate-test* method, involves translating text from a low-resource language into English and then employing an English-only language model for the desired tasks. Using the same example as above, researchers could translate Swahili social media posts into English and use them to fine-tune a pre-trained English model like BERT for emotion classification. This method is much more reliant on the quality of the translation model and tends to underperform for languages that are typologically distant from English,[82] which may explain its limited use in Global South contexts.

However, despite their potential to supplement existing low-resource language training data, these approaches come with significant limitations. Crucially, machine translation methods often miss important local contextual knowledge and linguistic nuances. They may, for example, produce "translationese," which refers to unnatural language patterns created when machine translation models oversimplify or overcomplicate sentences.[83] Other limitations range from gender bias to the flattening of connotations in different languages.[84] These problems also tend to be inconsistent across languages, making them difficult to address systematically.[85] Furthermore, models trained on machine-translated data risk creating a self-reinforcing cycle, as flawed outputs threaten to pollute future models' training data, further degrading model performance over time.[86]

Assembling more labeled data

While machine-translated text has been useful in generating data, it often lacks the precision and cultural understanding needed to train reliable AI models. These models learn from patterns in data, but, without well-labeled examples, they can struggle to accurately process language—especially in low-resource languages, where data is often scarce or inconsistent. Labeling is the process of categorizing and structuring raw text or speech data so that models can recognize meanings, contexts, and relationships between words.

To address this, other efforts have focused on producing diverse, high-quality annotated datasets using different machine learning approaches or by engaging native speakers of low-resource languages through crowdsourcing.

Machine learning approaches

Researchers have been employing a series of automated or semi-automated approaches to streamline the process of labeling raw, unlabeled data.[87] One approach, known as distant supervision, leverages alternative data sources such as dictionaries and bilingual datasets to label data. For example, after gathering a large corpus of unlabeled text, researchers may use a dictionary's list of adjectives to ensure that all relevant instances in the dataset are automatically labeled as adjectives.

Alternatively, domain experts or low-resource language speakers may develop a set of labeling rules for automatically annotating data. For example, such rules might identify date expressions based on known patterns of how speakers of the language use date keywords (day, month, year, etc.). This approach has been applied to temporal tagging across languages and for identifying dates in Hausa and Yorùbá.[88]

Finally, labeled data in one language (such as English) can be leveraged to annotate unlabeled data in another. This approach requires a means of projecting the labels into the low-resource language, which is typically done via machine translation engines or bilingual datasets.

All of these approaches still rely heavily on additional data sources such as bilingual datasets, domain experts, geographical directories, or task-specific labels which may not always be available. This reliance partly explains the frequent use—and associated challenges—of religious texts in low-resource NLP, since the Bible is one of the most widely translated works. The quantity and quality of these supplementary data sources are key to ensuring the transferability of approaches and results between low-resource languages.[89]

Participatory approaches

In contrast, a growing number of research groups are adopting a community-centric, participatory approach to strengthening low-resource language communities. Responding to the known challenges faced by data annotators in the Global South (including lax labor regulations and low wages),[90] research groups are increasingly involving native speakers throughout the entire AI development life cycle, enabling them to contribute more holistically to the creation of monolingual and bilingual models in their respective languages, rather than only the data labeling process.[91] This collaborative model empowers communities while ensuring more accurate, diverse, and culturally representative NLP technologies. It is also crucial to creating long-term mechanisms for collaboration between developers, researchers, and the communities that will adopt and be affected by the resulting AI models.

For example, Masakhane[92]—an African grassroots organization co-founded in 2019 by one of our authors—has built an inclusive distributed community of African NLP researchers who work together to build datasets and tools to facilitate NLP research on African languages.[93] The community, which now consists of more than 1,000 members from 30 African countries, finds creative ways to build more datasets, benchmarks, and models for African languages, but it also has a wide range of other ongoing research initiatives related to low-resource African languages.[94]

Similarly, large U.S.-based companies are taking a community-centric approach to building digital resources for low-resource languages. Microsoft's ELLORA (Enabling Low Resource Language) project,[95] for example, has researchers working closely with low-resource language communities across India

Research groups are increasingly involving native speakers throughout the entire AI development life cycle, enabling them to contribute more holistically to the creation of monolingual and bilingual models in their respective languages.

to collaboratively design and collect language data. Vaani, a Google-funded project, is compiling an open-source dataset of speech recordings from across India that includes local low-resource languages.[96]

While there are ongoing, successful efforts to crowdsource more diverse data for lower-resource languages, scaling remains a challenge. For example, evaluating data quality is a key challenge when trying to scale up crowdsourced data collection,[97] especially when contributors come from communities that have not previously been exposed to much digitized work. Given the widely known and deeply concerning prevalence of exploitative labor conditions for workers or volunteers who collect or label data,[98] crowdsourcing efforts will continue to pose ethical challenges that must be thoughtfully addressed by actors in this space.

4. Implications for Researchers, Funders, Policymakers, and Civil Society

This paper shows that there is a wide variety of technical and community-driven approaches to addressing the digital divide in LLM development and that each comes with its own trade-offs that differ significantly from one region to another depending on the linguistic, sociocultural, and economic context. It is crucial for AI researchers, policymakers, development agencies, funders, and civil society to understand both the underlying reasons for and paths to addressing these disparities in LLM development.

First, a large body of research has shown that major LLMs perform poorly for many non-English and especially low-resource languages.[99] In particular, users with lower English proficiency and education levels are disproportionately affected by undesirable model behavior such as hallucinations and bias.[100] More broadly, LLMs developed in the Global North are often not attuned to the sociocultural context of many Global South communities and instead represent the worldviews,[101] cultural norms, and unconscious biases of their predominantly U.S.-based developers. In the absence of LLMs tailored to low-resource languages, speakers of such languages must use unreliable and biased models, making them more vulnerable to misinformation, misconceptions, and representational harm. Crucially, if low-resource languages continue to be overlooked by NLP research communities, LLMs are likely to become even more linguistically biased against non-standard dialects and non-English speakers.[102]

Second, on a more fundamental level, many powerful LLMs remain inaccessible in parts of the Global South.

Beyond foundational access barriers, such as reliable internet connectivity and mobile device availability, non-English language speakers also face higher costs when using powerful models compared to English speakers.

Beyond foundational access barriers, such as reliable internet connectivity and mobile device availability, non-English language speakers also face higher costs when using powerful models compared to English speakers. Accessing OpenAI's GPT models via APIs, for example, can cost up to six times more for non-English speakers due to how these models process language inputs.[103] This access gap means that low-resource language communities disproportionately miss out on the economic and productivity benefits of LLM-driven innovation.[104] This dynamic not only restricts opportunities for wealth generation but also reinforces existing power imbalances, whereby Global North actors continue to dominate AI development, governance, and NLP research.[105]

More broadly, as low-resource language speakers are continuously overlooked in LLM model development, their perspectives are more likely to be excluded from other decision-making processes that shape AI deployment. The pervasiveness of exploitative labor practices and the uneven distribution of AI's environmental costs are prominent examples.[106] Developers may, for example, fail to consider how inefficiencies in applying LLMs to low-resource language models can heighten their environmental toll.[107]

Below, we present three overarching recommendations for researchers, funders, policymakers, and civil society organizations looking to support efforts to close the LLM divide and ensure that low-resource language communities can both benefit from and be represented in LLM development.

1. Make strategic investments in AI research and development for/with low-resource languages

Investing strategically in AI development for less-resourced languages is key to bridging the digital language divide.

Access to computing and cloud resources: One key approach is subsidizing access to cloud and computing resources. Not only are the costs for these resources prohibitive for many researchers in low-resource language contexts (one top-tier GPU can cost as much as $30,000),[108] but LLM inefficiencies and foreign exchange volatility also make cloud service costs even more burdensome and unpredictable for developers working with these languages.[109] With subsidized access to computing resources or cloud credits, local researchers and businesses can develop and experiment with more sophisticated AI models at more affordable and stable prices—effectively lowering barriers to entry and democratizing AI access.

Investing in the development and adoption of reliable, efficient, and culturally sensitive methods to evaluate model performance is equally important.

Culturally nuanced data sources and model evaluations: Funding research initiatives that increase the availability and quality of low-resource language data can drive significant progress. There is a pressing need to develop more efficient data collection methods that ensure datasets for such languages are diverse, representative, and free from biases that could perpetuate existing inequities. Additionally, investing in the development and adoption of reliable, efficient, and culturally sensitive methods to evaluate model performance is equally important, as prevailing benchmarks often fail to adequately take into account the sociocultural contexts of low-resource languages.[110] A robust evaluation ecosystem is essential to driving future LLM development and adoption.

Local research initiatives and cross-disciplinary partnerships: Governments and funders can support and establish grant programs, research challenges, and research workshops[111] to incentivize cross-disciplinary work on innovative solutions at the intersection of data and compute limitations.[112] And while Global North-South partnerships remain critical,[113] fostering South-South cooperation is equally important. Grassroots research communities like AmericasNLP and GhanaNLP are already making significant strides in developing AI for underrepresented languages.[114] Supporting and expanding these networks through joint meetings, research symposiums, and data-sharing initiatives can amplify resource sharing, enhance efficiency in model development, and foster innovation tailored to Global South contexts.

Sustainable approach: Investing in AI research and development must take into account the attendant environmental toll. Funding should prioritize AI research that employs techniques that minimize the environmental impact, such as geographical load balancing[115]—which enables the flexible deployment of AI computing across data centers in different regions. More fundamentally, investors and researchers alike must critically reflect on the broader trade-offs of resource-intensive localized LLM development and whether existing solutions or collaborative, regional initiatives can instead meet the needs specific to a given language community.[116] Civil society continues to demonstrate that AI development must be centered on lived realities.[117]

2. Promote participatory research for inclusive AI development

"Low resourcedness" is not solely a data problem[118] but a phenomenon rooted in societal problems such as non-diverse, exclusionary, and even exploitative AI research practices. The involvement of local communities throughout the LLM development and adoption process is therefore crucial. The above-mentioned participatory approaches to dataset construction, data annotation,[119] and LLM development are important case studies that demonstrate the immense value of ensuring that the experiences and perspectives of underserved and neglected language communities around the world are reflected not only in the training datasets but also in the approaches to model development and access.

"Low resourcedness" is not solely a data problem but a phenomenon rooted in societal problems such as non-diverse, exclusionary, and even exploitative AI research practices.

Direct collaboration with low-resource language communities: Researchers (and those funding research) in and outside of Global South contexts should ensure that research is conducted in direct collaboration with low-resource language communities who can help co-design datasets, labeling schemes, and evaluation methods. Such participatory approaches, in which communities actively contribute to and even co-own the creation of AI resources, help ensure that the models developed reflect the nuances of the language, culture, and specific needs of those communities. Collaborations between researchers and social sector organizations can further amplify this impact by connecting research efforts with social applications that directly benefit local populations. These collaborations must be guided by principles of fairness and ethical labor practices to ensure crowdsourcing partnerships are not exploitative.

3. Ensure equitable data ownership

The equitable collection and management of low-resource language datasets should be another priority. Data ownership and consent in AI development and deployment is a global problem. The use of vast amounts of public data crawled from the web for AI training has raised widespread concerns over consent, copyright, and fair compensation. Many researchers argue that data consent is in crisis for both developers and creators[120] and that copyright risks are real and largely unresolved.[121] Global South data subjects and creators are particularly vulnerable to data exploitation, with little to no pathways for being informed about, influencing, or seeking redress for the use of their data.

Rights-respecting licensing frameworks that facilitate AI development: Data rights organizations, policymakers, and funders should work together to incentivize and support the creation of equitable licensing frameworks that facilitate access to AI training data for developers while protecting the data rights of low-resource language data subjects and creators. Establishing fair compensation structures for data contributors is paramount, whether through monetary means or by providing shared benefits such as access to the resulting technologies.[122] While opinions differ on whether such licensing frameworks should be government-mandated,[123] several grassroots research collectives and trade groups have already been working to protect data rights by improving dataset transparency and documentation through large-scale data audits and creating opt-in systems for data consent.[124] Innovative, flexible approaches to compensation structures, such as token-based systems or shared revenue models, may offer promising pathways to achieving a balance between local data sharing and the protection and recognition of contributions from individuals and communities.

Establishing fair compensation structures for data contributors is paramount, whether through monetary means or by providing shared benefits such as access to the resulting technologies.

Endnotes

1. Jai Vipra and Sarah M. West, "Computational Power and AI," *AI Now Institute*, September 27, 2023, https://ainowinstitute.org/publication/policy/compute-and-ai.
2. Bridget Boaykye et al., "State of Compute Access: How to Bridge the New Digital Divide," *Tony Blair Institute for Global Change,* December 7, 2023, https://institute.global/insights/tech-and-digitalisation/state-of-compute-access-how-to-bridge-the-new-digital-divide#conclusion.
3. Ruihan Huang et al., "MacroPolo's The Global AI Talent Tracker 2.0," Paulson Institute, accessed April 1, 2025, https://macropolo.org/interactive/digital-projects/the-global-ai-talent-tracker/.
4. Gábor Bella et al., "Towards Bridging the Digital Language Divide," preprint arXiv, July 25, 2023, https://arxiv.org/abs/2307.13405; Damian Blasi et al., "Systematic Inequalities in Language Technology Performance across the World's Languages," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022, 5486–5505, https://aclanthology.org/2022.acl-long.376/?ref=ruder.io.
5. Viet Dac Lai et al., "ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning," preprint arXiv, April 12, 2023, https://arxiv.org/abs/2304.05613; Andrew Deck, "We Tested ChatGPT in Bengali, Kurdish, and Tamil. It failed," *Rest of World,* September 6, 2023, https://restofworld.org/2023/chatgpt-problems-global-language-testing/.
6. Paresh Dave, "ChatGPT Is Cutting Non-English Languages Out of the AI Revolution," *WIRED*, May 31, 2023, https://www.wired.com/story/chatgpt-non-english-languages-ai-revolution/.
7. Prabha Kannan, "Improving Equity and Access to Non-English Large Language Models," Stanford Institute for Human-Centered Artificial Intelligence, April 22, 2024, https://hai.stanford.edu/news/improving-equity-and-access-non-english-large-language-models.
8. David M. Eberhard et al. eds., "Ethnologue: Languages of the World" *SIL International*, 2024, http://www.ethnologue.com.
9. Derivation.co, "What is the Digital Language Divide," *Derivation,* 2024, https://derivation.co/digital-language-divide/.
10. Aivin V. Solatorio et al., "Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-Economic Disparities and Reduced Utility for Non-English Speakers," preprint arXiv, October 14, 2024, https://arxiv.org/abs/2410.10665; Aubra Anthony et al., "Advancing a More Global Agenda for Trustworthy Artificial Intelligence," Carnegie Endowment for International Peace, April 30, 2024, https://carnegieendowment.org/research/2024/04/advancing-a-more-global-agenda-for-trustworthy-artificial-intelligence?lang=en.
11. Constantine Lignos et al., "Toward More Meaningful Resources for Lower-Resourced Languages," *Findings of the Association for Computational Linguistics*, May 2022, 523–32, https://aclanthology.org/2022.findings-acl.44/.
12. Barry Haddow et al., "Survey of Low-Resource Machine Translation," *Computational Linguistics*, 48, no. 3 (2022): 673-732, https://direct.mit.edu/coli/article/48/3/673/111479/Survey-of-Low-Resource-Machine-Translation.
13. Julia Kreutzer et al., "Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets," *Transactions of the Association for Computational Linguistics,* no. 10: 50-72 (2022), https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00447/109285/Quality-at-a-Glance-An-Audit-of-Web-Crawled; Zeyi Yang, "GPT-4o's Chinese Token-Training Data Is Polluted by Spam and Porn Websites," *MIT Technology Review,* May 17, 2024, https://www.technologyreview.com/2024/05/17/1092649/gpt-4o-chinese-token-polluted/.
14. Anmol Irfan, "The Fight to Preserve the Urdu Script in the Digital World," *Time,* September 27, 2023, https://time.com/6317817/urdu-nastaliq-digital/.
15. Gabriel Nicholas and Aliya Bhatia, "Lost in Translation: Large Language Models in Non-English Content Analysis," *Center for Democracy & Technology,* May 23, 2023, https://cdt.org/insights/lost-in-translation-large-language-models-in-non-english-content-analysis/.
16. Ben Hutchinson, "Modeling the Sacred: Considerations when Using Religious Texts in Natural Language Processing," preprint arXiv, June 25, 2024, https://arxiv.org/html/2404.14740v2.
17. Anees Bahji et al., "Exclusion of the Non-English-Speaking World from the Scientific Literature: Recommendations for Change for Addiction Journals and Publishers," *Nordic Studies on Alcohol and Drugs*, 40, no. 1 (2023): 6–13, https://doi.org/10.1177/14550725221102227.
18. Julia Kreutzer et al., "Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets."
19. Hellina Hailu Nigatu et al., "The Zeno's Paradox of 'Low-Resource' Languages," *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, November 2024, 17753–74, https://aclanthology.org/2024.emnlp-main.983/.
20. András Kornai, "Digital Language Death," PLOS ONE 8, no. 10 (October 2013): e77056, https://doi.org/10.1371/journal.pone.0077056.
21. "What Are the Most Used Languages on the Internet?," Internet Society Foundation, May 15, 2023, https://www.isocfoundation.org/2023/05/what-are-the-most-used-languages-on-the-internet/.
22. Michael A. Hedderich et al., , "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios," preprint arXiv, last revised April 9, 2021, https://arxiv.org/abs/2010.12309.
23. The language is commonly known as "Dahalo," though native speakers use the term numma guhooni. "Dahalo," Endangered Languages Project, https://www.endangeredlanguages.com/lang/4064.
24. "Dahalo," *Ethnologue*, https://www.ethnologue.com/language/dal/.
25. Michael A. Hedderich et al., "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios."
26. Pratik Joshi et al., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World," *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, July 2020, 6282–93, https://aclanthology.org/2020.acl-main.560/.
27. Chinasa T. Okolo, "AI in the Global South: Opportunities and Challenges Towards More Inclusive Governance," *Brookings,* November 1, 2023, https://www.brookings.edu/articles/ai-in-the-global-south-opportunities-and-challenges-towards-more-inclusive-governance/.
28. Orevaoghene Ahia et al., "The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation," preprint arXiv, October 6, 2021, https://arxiv.org/abs/2110.03036.
29. Chijioke Okorie and Vukosi Marivate, "How African NLP Experts Are Navigating the Challenges of Copyright, Innovation, and Access," Carnegie Endowment for International Peace, April 30, 2024, https://carnegieendowment.org/research/2024/04/how-african-nlp-experts-are-navigating-the-challenges-of-copyright-innovation-and-access?lang=en.
30. Aubra Anthony et al., "Advancing a More Global Agenda for Trustworthy Artificial Intelligence."
31. Sebastian Ruder et al., "Square One Bias in NLP: Towards a Multi-Dimensional Exploration of the Research Manifold," preprint arXiv, June 20, 2022, https://arxiv.org/abs/2206.09755; Orevaoghene Ahia et al., "The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation."
32. See: Pranaydeep Singh et al., "Distilling Monolingual Models from Large Multilingual Transformers," *Electronics*, 12, no. 4 (2023): 1022, https://www.mdpi.com/2079-9292/12/4/1022; Alan Ansell et al., "Distilling Efficient Language-Specific Models for Cross-Lingual Transfer," preprint arXiv, June 2, 2023, https://arxiv.org/abs/2306.01709; Andrea Gregor de Varda and Marco Marelli, "Data-Driven Cross-lingual Syntax: An Agreement Study with Massively Multilingual Models," *Computational Linguistics*, 49, no. 2 (2023): 261–99, https://doi.org/10.1162/coli_a_00472.
33. Jacob Devlin et al., "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," preprint arXiv, last revised May 24, 2019, https://arxiv.org/abs/1810.04805; Linting Xue et al., "mT5: A Massively Multilingual Pre-Trained Text-to-Text Transformer," June 2021, 483–98, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics*, https://research.google/pubs/mt5-a-massively-multilingual-pre-trained-text-to-text-transformer/; Alexis Conneau et al., "Unsupervised Cross-Lingual Representation Learning at Scale," preprint arXiv, last revised April 8, 2020, https://arxiv.org/abs/1911.02116; Pengcheng He et al., "DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing," preprint arXiv, last revised March 24, 2023, https://arxiv.org/abs/2111.09543.
34. Cohere for AI Team, "C4AI Launches Aya, an LLM Covering More Than 100 Languages," Cohere (blog), February 12, 2024, https://cohere.com/blog/aya.
35. As of the time of writing, the mechanics of how cross-lingual transfer learning works are still poorly understood, with experts pointing to several different theories on how LLMs improve their language performance. See, for example, Gabriel Nicholas, "Foundation Models and Non-English Languages: Towards Better Benchmarking and Transparency," *Center for Democracy & Technology*, May 1, 2024, https://cdt.org/insights/foundation-models-and-non-english-languages-towards-better-benchmarking-and-transparency/.
36. Wikimedia, "List of Wikipedias," https://meta.wikimedia.org/wiki/List_of_Wikipedias.
37. Jacob Devlin et al.,"BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding."
38. Alexei Baevski et al., "XLS-R: Self-Supervised Speech Processing for 128 Languages," Meta (blog), November 18, 2024, https://ai.meta.com/blog/xls-r-self-supervised-speech-processing-for-128-languages/.
39. Hemant Yadav and Sunayana Sitaram, "A Survey of Multilingual Models for Automatic Speech Recognition," *Proceedings of the 13th Conference on Language Resources and Evaluation*, June 25, 2022, 5071–79, https://aclanthology.org/2022.lrec-1.542.pdf.
40. Sumanth Doddapaneni et al., "A Primer on Pretrained Multilingual Language Models," preprint arXiv, last revised December 23, 2021, https://arxiv.org/abs/2107.00676.
41. Shijie Wu and Mark Dredze, "Are All Languages Created Equal in Multilingual BERT?" *Proceedings of the 5th Workshop on Representation Learning for NLP*, July 2020, 120–30, https://aclanthology.org/2020.repl4nlp-1.16/; Nathaniel Robinson et al., "ChatGPT MT: Competitive for High- (but Not Low-) Resource Languages," *Proceedings of the Eighth Conference on Machine Translation*, December 2023, 392–418, https://aclanthology.org/2023.wmt-1.40/.

42. Gabriel Nicholas and Aliya Bhatia, "Lost in Translation: Large Language Models in Non-English Content Analysis."
43. Ibid.
44. Zheng-Xin Yong et al., "Low-Resource Languages Jailbreak GPT-4," preprint arXiv, last revised January 27, 2024, https://arxiv.org/abs/2310.02446.
45. Alexis Conneau et al., "Unsupervised Cross-Lingual Representation Learning at Scale."
46. Zirui Wang et al., "On Negative Interference in Multilingual Models: Findings and a Meta-Learning Treatment," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, November 2020, 4438–50, https://aclanthology.org/2020.emnlp-main.359/.
47. Naman Goyal et al., "Larger-Scale Transformers for Multilingual Masked Language Modeling," *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, August 2021, 29-33, https://aclanthology.org/2021.repl4nlp-1.4/.
48. Kelechi Ogueji et al., "Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-Resourced Languages," *Proceedings of the 1st Workshop on Multilingual Representation Learning*, November 2021, 116-26, https://aclanthology.org/2021.mrl-1.11/.
49. AI Singapore, "SEA-LION.AI," https://sea-lion.ai/our-models/#sealionv1.
50. Divyanshu Kakwani et al., "IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-Trained Multilingual Language Models for Indian Languages," *Findings of the Association for Computational Linguistics,* November 2020, 4948–61, https://aclanthology.org/2020.findings-emnlp.445/.
51. AI Singapore, "SEA-LION (Southeast Asian Languages In One Network)," *Github*, 2024, https://github.com/aisingapore/sealion.
52. Longxu Dou et al., "Sailor: Open Language Models for South-East Asia," preprint arXiv, April 4, 2024, https://arxiv.org/html/2404.03608v1.
53. AI Singapore, "SEA-Lion V3," SEA-LION.AI, accessed April 1, 2025, https://sea-lion.ai/our-models/.
54. AI Singapore, "SEA-LION (Southeast Asian Languages In One Network)."
55. Wei Qi Leong et al., "BHASA: A Holistic Southeast Asian Linguistic and Cultural Evaluation Suite for Large Language Models," preprint arXiv, last revised September 19, 2023, https://arxiv.org/abs/2309.06085v2/.
56. AI Singapore, "SEA-LION (Southeast Asian Languages In One Network)."
57. Kelechi Ogueji et al., "Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages."
58. Ibid.
59. Atnafu Lambebo Tonja et al., "InkubaLM: A Small Language Model for Low-Resource African Languages," preprint arXiv, last revised September 3, 2024, https://arxiv.org/abs/2408.17024.
60. Jesujoba O. Alabi et al., "Adapting Pre-Trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning," *Proceedings of the 29th International Conference on Computational Linguistics,* October 2022, 4336–49, https://aclanthology.org/2022.coling-1.382/.
61. Bonaventure F. P. Dossou et al., "AfroLM: A Self-Active Learning-Based Multilingual Pretrained Language Model for 23 African Languages," preprint, last revised November 23, 2022, https://arxiv.org/abs/2211.03263.
62. Kelechi Ogueji et al., "Small Data? No Problem! Exploring the Viability of Pretrained Multilingual Language Models for Low-resourced Languages."
63. Jesse Parvess et al., "BantuBERTa Model," *Hugging Face,* 2024, https://huggingface.co/dsfsi/BantuBERTa.
64. Jesse Parvess, "BantuBERTa: Using Language Family Grouping in Multilingual Language Modeling for Bantu Languages" (master's thesis, University of Pretoria, 2023), https://www.proquest.com/docview/2925081241?sourcetype=Dissertations%20&%20Theses.
65. Vukosi Marivate et al., "PuoBERTa: Training and Evaluation of a Curated Language Model for Setswana," *Communications in Computer and Information Science*, vol. 1976, 2023, https://doi.org/10.1007/978-3-031-49002-6_17.
66. Gati Martin et al., "SwahBERT: Language Model of Swahili," *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, July 2022, 303-13, https://aclanthology.org/2022.naacl-main.23/; Jacaranda Health (blog), "Jacaranda Launches First-in-Kind Swahili Large Language Model," October 31, 2023, https://jacarandahealth.org/jacaranda-launches-first-in-kind-swahili-large-language-model/.
67. Kunat Pipatanakul et al., "Typhoon: Thai Large Language Models," preprint arXiv, December 21, 2023, https://arxiv.org/abs/2312.13951.
68. Fajri Koto et al., "IndoLEM and IndoBERT: A Benchmark Dataset and Pre-Trained Language Model for Indonesian NLP," preprint arXiv, November 2, 2020, https://arxiv.org/abs/2011.00677.
69. Sang T. Truong et al., "Crossing Linguistic Horizons: Finetuning and Comprehensive Evaluation of Vietnamese Large Language Models," preprint arXiv, last revised May 26, 2024, https://arxiv.org/abs/2403.02715.
70. Gábor Bella et al., "Towards Bridging the Digital Language Divide," preprint arXiv, July 25, 2023, https://arxiv.org/abs/2307.13405.
71. Sumanth Doddapaneni et al., "A Primer on Pretrained Multilingual Language Models."
72. Tarek Naous et al., "Having Beer after Prayer? Measuring Cultural Bias in Large Language Models," *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* 1, August 2024, 16366-16393, https://aclanthology.org/2024.acl-long.862/.
73. Shijie Wu et al., "Are All Languages Created Equal in Multilingual BERT?"
74. Sang T. Truong et al., "Crossing Linguistic Horizons: Finetuning and Comprehensive Evaluation of Vietnamese Large Language Models."
75. Sumanth Doddapaneni et al., "A Primer on Pretrained Multilingual Language Models."
76. Tarek Naous et al., "Having Beer after Prayer? Measuring Cultural Bias in Large Language Models."
77. Russell Brandom, "What the AI Boom Is Getting Wrong (and Right), According to Hugging Face's Head of Global Policy," Rest of World, June 3, 2024, https://restofworld.org/2024/hugging-face-ai-boom/#:~:text=Something%20that%20I,a%20big%20deal.
78. Orevaoghene Ahia et al., "The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation."
79. Kai Zhu and Dylan Walker, "Machine-Assisted Content Creation on Peer Production Platforms," *SSRN*, last revised February 21, 2025, https://ssrn.com/abstract=4708614.
80. Tim Isbister et al., "Should We Stop Training More Monolingual Models, and Simply Use Machine Translation Instead?" preprint arXiv, April 21, 2021, https://arxiv.org/abs/2104.10441.
81. Marta R. Costa-jussà et al., "No Language Left Behind: Scaling Human-Centered Machine Translation," preprint arXiv, last revised August 25, 2022, https://arxiv.org/abs/2207.04672.
82. Mikel Artetxe et al., "Revisiting Machine Translation for Cross-Lingual Classification," *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, December 2023, 6489–99, https://aclanthology.org/2023.emnlp-main.399/; Tim Isbister et al., "Should We Stop Training More Monolingual Models, and Simply Use Machine Translation Instead?"
83. Vered Volansky et al., "On the Features of Translationese," *Digital Scholarship in the Humanities,* 30, no. 1 (April 2015): 98–118, https://doi.org/10.1093/llc/fqt031.
84. Gabriel Stanovsky et al., "Evaluating Gender Bias in Machine Translation," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* July 2019, 1679–84, https://aclanthology.org/P19-1164/; Gabriel Nicholas and Aliya Bhatia, "Lost in Translation: Large Language Models in Non-English Content Analysis"; Vered Volansky et al.,"On the Features of Translationese."
85. Sicheng Yu et al., "Translate-Train Embracing Translationese Artifacts," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* 2, 362–70, https://aclanthology.org/2022.acl-short.40/.
86. Ilia Shumailov et al., "AI Models Collapse When Trained on Recursively Generated Data," *Nature* 631 (2024): 755-59, https://www.nature.com/articles/s41586-024-07566-y.
87. Michael A. Hedderich et al., "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios."
88. Jannik Strötgen and Michael Gertz, "Multilingual and Cross-Domain Temporal Tagging," *Language Resources & Evaluation* 47 (2013): 269–98, https://doi.org/10.1007/s10579-012-9179-y; Michael A. Hedderich et al., "Transfer Learning and Distant Supervision for Multilingual Transformer Models: A Study on African Languages," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, November 2020, 2580–91, https://aclanthology.org/2020.emnlp-main.204/.
89. Michael A. Hedderich et al., "A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios."
90. Rina Chandran et al., "AI Boom Is Dream and Nightmare for Workers in Global South," *Context,* March 14, 2023, https://www.context.news/ai/ai-boom-is-dream-and-nightmare-for-workers-in-global-south.
91. Caleb Ziems et al., "VALUE: Understanding Dialect Disparity in NLU," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* 1, May 2022, 3701-3720, https://aclanthology.org/2022.acl-long.258/.
92. Masakhane, https://www.masakhane.io/.
93. Iroro Orife et al., "Masakhane—Machine Translation for Africa," preprint arXiv, March 13, 2020, https://arxiv.org/abs/2003.11529.
94. Masakhane, "Masakhane—A Living Collection of NLP Projects for Africans, by Africans," GitHub repository, https://github.com/masakhane-io/masakhane-community/blob/master/README.md.
95. Microsoft, "ELLORA: Enabling Low Resource Languages," accessed April 1, 2025, https://www.microsoft.com/en-us/research/project/ellora/.
96. Vaani, https://vaani.iisc.ac.in.
97. Basil Abraham et al., "Crowdsourcing Speech Data for Low-Resource Languages from Low-Income Workers," *Proceedings of the Twelfth Language Resources and Evaluation Conference*, May 2020, 2819–26, https://aclanthology.org/2020.lrec-1.343/.
98. Billy Perrigo, "Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic," *Time,* January 18, 2023, https://time.com/6247678/openai-chatgpt-kenya-workers/.
99. Aivin V. Solatorio et al., "Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-Economic Disparities and Reduced Utility for Non-English Speakers"; Andrew Deck, "We Tested ChatGPT in Bengali, Kurdish, and Tamil. It Failed"; Lai et al., "ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning."

100. Elinor Poole-Dayan et al., "LLM Targeted Underperformance Disproportionately Impacts Vulnerable Users," preprint arXiv, June 25, 2024. https://arxiv.org/abs/2406.17737.
101. Yan Tao et al., "Cultural Bias and Cultural Alignment of Large Language Models," *PNAS Nexus,* 3, no. 9 (September 2024), https://academic.oup.com/pnasnexus/article/3/9/pgae346/7756548.
102. Sebastian Ruder, "Why You Should Do NLP Beyond English," *Ruder.io* (blog), August 1, 2020, https://www.ruder.io/nlp-beyond-english/?ref=ruder.io#the-societal-perspective; Gábor Bellaet et al., "Towards Bridging the Digital Language Divide"; Andrew Myers, "AI-Detectors Biased Against Non-Native English Writers," Stanford Institute for Human-Centered Artificial Intelligence, May 15, 2023, https://hai.stanford.edu/news/ai-detectors-biased-against-non-native-english-writers.
103. Aivin V. Solatorio et al., "Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-Economic Disparities and Reduced Utility for Non-English Speakers."
104. Anran Xiao et al., "Bridging the Digital Divide: The Impact of Technological Innovation on Income Inequality and Human Interactions," *Humanities and Social Sciences Communications* 11, no. 809 (2024), https://doi.org/10.1057/s41599-024-03307-8; Anand S. Rao and Gerard Verweij, "Sizing the Prize: What's the Real Value of AI for Your Business and How Can You Capitalise?" PwC, 2017, https://www.pwc.com/gx/en/issues/analytics/assets/pwc-ai-analysis-sizing-the-prize-report.pdf; Prabha Kannan, "Improving Equity and Access to Non-English Large Language Models," Stanford Institute for Human-Centered Artificial Intelligence, April 22, 2024, https://hai.stanford.edu/news/improving-equity-and-access-non-english-large-language-models.
105. Shaleen Khanal et al., "Why and How Is the Power of Big Tech Increasing in the Policy Process? The Case of Generative AI," *Policy and Society* 44, no. 1 (January 2025): 52–69, https://doi.org/10.1093/polsoc/puae012; Sumaya Nur Adan, "The Case for Including the Global South in AI Governance Discussions," Centre for the Governance of AI, October 20, 2023, https://www.governance.ai/post/the-case-for-including-the-global-south-in-ai-governance-conversations; Pratik Joshi et al., "The State and Fate of Linguistic Diversity and Inclusion in the NLP World."
106. Adrienne Williams et al., "The Exploited Labor Behind Artificial Intelligence," *Noema,* October 13, 2022, https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence/; Shaolei Ren and Adam Wierman, "The Uneven Distribution of AI's Environmental Impacts," *Harvard Business Review*, July 15, 2024, https://hbr.org/2024/07/the-uneven-distribution-of-ais-environmental-impacts.
107. Aivin V. Solatorio et al., "Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-Economic Disparities and Reduced Utility for Non-English Speakers."
108. Stefan French, "The Cost of Cutting-Edge: Scaling Compute and Limiting Access," Mozilla.AI (blog), March 20, 2024, https://blog.mozilla.ai/the-cost-of-cutting-edge-scaling-compute-and-limiting-access/?utm_source=chatgpt.com.
109. Aivin V. Solatorio et al., "Double Jeopardy and Climate Impact in the Use of Large Language Models: Socio-Economic Disparities and Reduced Utility for Non-English Speakers."
110. Gabriel Nicholas, "Foundation Models and Non-English Languages: Towards Better Benchmarking and Transparency."
111. See, for example, the upcoming NAACL 2025 Workshop on May 4, 2025, on "Language Models for Underserved Communities," https://lm4uc.github.io/#.
112. U.S. Department of State, "Partnership for Global Inclusivity on AI," https://www.state.gov/advancing-sustainable-development-through-safe-secure-and-trustworthy-ai/; Stefan French, "The Cost of Cutting-Edge: Scaling Compute and Limiting Access"; Orevaoghene Ahia et al., "The Low-Resource Double Bind: An Empirical Study of Pruning for Low-Resource Machine Translation."
113. U.S. Department of Commerce, "Joint Statement on Harnessing Artificial Intelligence, Facilitating Data Flows and Empowering Digital Upskilling Between the United States Department of Commerce and the Nigerian Ministry of Communications, Innovation and Digital Economy," July 25, 2024, https://www.commerce.gov/news/press-releases/2024/07/joint-statement-harnessing-artificial-intelligence-facilitating-data.
114. AmericasNLP, https://turing.iimas.unam.mx/americasnlp/; GhanaNLP, https://ghananlp.org/.
115. Pengfei Li et al., "Towards Environmentally Equitable AI via Geographical Load Balancing," UC Riverside, June 2023, https://escholarship.org/uc/item/79c880vf.
116. Elina Noor and Binya Kanitroj, "Speaking in Code: Contextualizing Large Language Models in Southeast Asia," Carnegie Endowment for International Peace, January 6, 2025, https://carnegieendowment.org/research/2025/01/speaking-in-code-contextualizing-large-language-models-in-southeast-asia?lang=en.
117. Angela Oduor Lungati, "Here's Why We Need Inclusive AI: A View from the African Continent," *Context*, December 10, 2024, https://www.context.news/ai/opinion/heres-why-we-need-inclusive-ai-a-view-from-the-arican-continent.
118. "'Low-Resourcedness' Beyond Data," posted November 20, 2020, by Masakhane, YouTube, https://www.youtube.com/watch?v=Xbc_g_OknqA.
119. Constantine Lignos et al., "Toward More Meaningful Resources for Lower-Resourced Languages."
120. Shayne Longpre et al., "Consent in Crisis: The Rapid Decline of the AI Data Commons," preprint arXiv, last revised July 24, 2024, https://arxiv.org/abs/2407.14933.
121. Peter Henderson et al., "Foundation Models and Copyright Questions," Stanford Institute for Human-Centered Artificial Intelligence, November 2, 2023, https://hai.stanford.edu/policy/policy-brief-foundation-models-and-copyright-questions.
122. Billy Perrigo, "The Workers Behind AI Rarely See Its Rewards. This Indian Startup Wants to Fix That," *Time*, July 27, 2023, https://time.com/6297403/the-workers-behind-ai-rarely-see-its-rewards-this-indian-startup-wants-to-fix-that/.
123. Kate Knibbs, "A New Group Is Trying to Make AI Data Licensing Ethical," *Wired*, September 4, 2024, https://www.wired.com/story/dataset-providers-alliance-ethical-generative-ai-licensing/.
124. See Data Provenance Initiative, https://www.dataprovenance.org/; Dataset Providers Alliance, "Shaping the Future of AI," September 4, 2024, https://5a5ee099-3141-4217-af47-c61b445c2269.filesusr.com/ugd/6112c3_4c700dd417044c4aa268a4a4a9080c88.pdf.

Stanford University
Human-Centered
Artificial Intelligence

The Asia Foundation

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA