Why AI Is WEIRD and Shouldn't Be This Way: Towards AI For Everyone, With Everyone, By Everyone

Rada Mihalcea[1*], Oana Ignat[2*], Longju Bai[1], Angana Borah[1], Luis Chiruzzo[3], Zhijing Jin[4], Claude Kwizera[5], Joan Nwatu[1], Soujanya Poria[6],Thamar Solorio[7]

[1]University of Michigan USA, [2]University of Santa Clara USA, [3]Universidad de la Republica Uruguay, [4]Max Plank Institute Germany, [5]CMU Africa, [6]SUTD Singapore, [6]MBZUAI United Arab Emirates

Abstract

This paper presents a vision for creating AI systems that are inclusive at every stage of development, from data collection to model design and evaluation. We address key limitations in the current AI pipeline and its WEIRD[1] representation, such as lack of data diversity, biases in model performance, and narrow evaluation metrics. We also focus on the need for diverse representation among the developers of these systems, as well as incentives that are not skewed toward certain groups. We highlight opportunities to develop AI systems that are for everyone (with diverse stakeholders in mind), with everyone (inclusive of diverse data and annotators), and by everyone (designed and developed by a globally diverse workforce).

Introduction

AI, and especially Large Language and Multimodal Models (LLMs and LMMs), have taken the world by storm, and yet much of the world is not represented in the data, models, and evaluations used in their development (Hershcovich et al. 2022; Moayeri, Tabassi, and Feizi 2024; Nayak et al. 2024). This lack of representation has two major implications. First, it can lead to numerous mistakes, misconceptions, and even harms, which can propagate to the growing number of applications powered by these models, and can limit the ability of AI systems to effectively serve diverse groups and contexts.

Second, as anthropologists have pointed out, our success as a human species is not as much due to our intelligence, as it is to our "collective cultural brains" that allow us to learn from one another over generations and across cultures (Henrich 2015). Cultural evolution has led to many innovations and entire bodies of knowledge – a form of collective intelligence that explains our species' uniqueness and success.

With the rapid growth of AI, we are now facing an evolution dilemma. On one side, we have our "recipe for success" learned over tens of thousands of years, where our collective cultural brains lead to innovation and evolution. On the other side, we have these very large AI models which, despite encompassing enormous bodies of information, act as a single "super-human" that homogenizes and erases entire bodies of cultural knowledge (Schwöbel et al. 2023; McVeety 2024; Byrd 2023; Perez et al. 2024; Wachter, Mittelstadt, and Russell 2024; Naous et al. 2024).

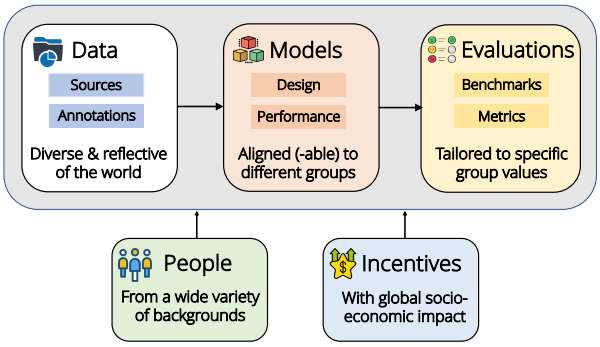

Figure 1: Desiderata and areas of research to expand the reach and impact of AI to everyone.

The goal of this paper is to present a vision for addressing limitations throughout the entire AI pipeline, including data, models, and evaluations, as well as the people driving this process and the incentives that shape its development. The aim is to ensure that these AI systems are: (1) **For Everyone**: AI systems that represent everyone (§1), with models with even performance across groups (§3.2), inclusive evaluation metrics (§4.1) and culturally diverse benchmarks (§4.2), and incentives that promote inclusive AI and balance profits and social impact (§5); (2) **With Everyone**: relying on diverse data sources (§2.1), diverse annotators and inclusive annotation standards (§2.2); (3) **By Everyone**: model designs that are unbiased (§3.1), built by a diverse group of developers and leading to applications grounded in real-life (§6).

We assembled a team of authors with broad expertise in the field of AI, who through their current or native countries, bring insights from twelve different cultures (China, Germany, India, Mexico, Nigeria, Romania, Rwanda, Singapore, Switzerland, United Arab Emirates, United States, Uruguay).

1 Representation in AI

Just as it is critical to see ourselves represented in our communities, we also strive for representation in the AI systems we use. Yet, current models often fail to grasp the characteristics

*Equal contributions. All other authors are listed alphabetically.

[1]WEIRD, an acronym coined by (Henrich, Heine, and Norenzayan 2010) to highlight the coverage limitations of many psychological studies, refers to populations that are Western, Educated, Industrialized, Rich, and Democratic. While we do not fully adopt this term for AI, as its current scope does not perfectly align with the WEIRD dimensions, we believe that today's AI has a similarly "weird" coverage, particularly in terms of who is involved in its development and who benefits from it.

of different cultures, either because they lack this cultural knowledge or because they fail to recognize when and where to apply it. To further complicate this, the definition of "culture" encompasses many aspects, along multiple semantic and demographic axes (Thompson, Roberts, and Lupyan 2020; Adilazuarda et al. 2024), including among others social norms (e.g., gift giving), beliefs and habits (e.g., daily routines), artistic taste (e.g., traditional music), or subjective perceptions (e.g., emotional connotations). Further, the narratives produced by these models are frequently seen through an outsider's lens, missing the essential culture-centric perspectives that give depth and authenticity, as illustrated in the vignette in Figure 2.

Petru is an 11-year-old boy living in his home country of Romania, at the age when he looks for role models to emulate. He asks an AI system for examples of male role models, and the reply comes back confidently: Nicolae Ceauşescu, with the justification that "he played a significant role in the Romanian history [...] and his regime had a lasting impact." This answer misleads Petru to believe that a dictator who was one of the darkest figures in Romanian history should be a model to follow.

Figure 2: Outsider perspectives on the history and culture of groups not represented in AI models often conflict with the insider perspectives and can be misleading.

Lack of Cultural Knowledge. The majority of resources used in the development of AI models often consider only general knowledge and disregard the cultural aspect (Shen et al. 2024). This knowledge permeates the entire AI pipeline, from data distribution and annotation (§2), to model development and alignment (§3), to evaluation metrics and benchmarks (§4). Lack of representation can inevitably introduce biases, which can stem from how the data collection is framed by the dataset developers (Parmar et al. 2022), from the annotators' belief (Sap et al. 2021) or the background of the contributors (Aguinis, Villamor, and Ramani 2020), from the strategies used to fine-tune or align AI models (Ouyang et al. 2022; Zhang et al. 2023), or from the benchmarks used to assess model performance (Shen et al. 2024).

Opportunities. Recent work included efforts to expand the cultural knowledge encompassed by AI systems along dimensions of time Shwartz (2022); food-related customs (Palta and Rudinger 2023); factual knowledge (Yin et al. 2022; Nguyen et al. 2022; Keleg and Magdy 2023; Romero et al. 2024); or cultural norms (Fung et al. 2023). Scaling these diversification efforts to include the numerous cultures worldwide and their many cultural dimensions is an ongoing challenge, particularly for those cultures with limited online representation. This may require close engagement with cultural experts or anthropologists who have an insider view into the culture of a group and understand its heritage (Olugbade et al. 2024). Additionally, it may necessitate developing tailored strategies to engage with cultural groups "where they are," acknowledging that not everyone has an online presence, and that much of this cultural knowledge floats openly in forms that may not be easily captured in digital form. Finally, people often affiliate with more than one culture, which requires solutions for cultural compositionality (Welch et al. 2020) or pluralistic alignment (Sorensen et al. 2024).

2 Data

Data is an essential component of the AI pipeline, used to train or fine-tune models but also to evaluate their performance. As illustrated in the vignette in Figure 3, a lack of diverse representation in AI datasets can have a negative impact. In the process of creating datasets, two aspects play a major role: the source of the data and its annotation.

Chinasa wants her small business to cater to a global audience, so she creates a website to sell African bath sponges and uploads several photos of her products to the site. Seeing that sales do not take off, she decides to do a simple image search for the word "sponge" on an AI-powered web search engine, just to find that out of a hundred image results, there is only one image that looks like the kind of sponge she sells.

Figure 3: Models trained on non-inclusive datasets hinder the representation of stakeholders in mainstream media.

2.1 Data Sources

Lack of Data Transparency. AI models, and especially LLMs and LMMs, rely on increasingly large collections of data. However, because of the current fast speed of research with minimal or non-existent regulations and standards, we have witnessed a crisis in data transparency (Longpre et al. 2023), often associated with unwanted societal biases and unexpected behaviors (Buolamwini and Gebru 2018; Gebru 2020). In response to these challenges, several strategies have emerged to enhance data documentation in AI. Gebru et al. (2021) introduced the concept of datasheets for datasets, which was followed by the development of data statements and data cards for NLP data (Bender and Friedman 2018; Holland et al. 2020; Pushkarna, Zaldivar, and Kjartansson 2022; Longpre et al. 2023).

Opportunities. Future efforts must prioritize the development of transparent, responsible, and data-centric AI models. Data licenses have the potential to promote more responsible, inclusive, and transparent machine-learning practices, but work still needs to be done to understand how to define and interpret license terms for data usage, and how to adapt them to AI data and models Longpre et al. (2023). AI researchers need to collaborate with policymakers and legal experts to develop tools that empower and educate users and creators about dataset documentation from various perspectives, including provenance, identifiers, and characteristics (Longpre et al. 2023). Inspiration can be drawn from the database community (Bhardwaj et al. 2014) and more regulated fields such as medicine (Moher et al. 2010). Additionally, innovative ideas from the NLP community, such as leveraging LLMs to document data and models, can further complement these efforts (Liu et al. 2024a).

Lack of Data Diversity. The predominant methods for data collection involve scraping vast amounts of text and image data from the Web (Villalobos et al. 2022). While effective for Western communities, this approach excludes minority groups with limited or no internet access, representing 37% of the global population (Nwatu, Ignat, and Mihalcea 2023). Moreover, there are many communities for which producing written data is a major challenge (Wiechetek et al. 2024).

Further, these uneven distribution and biases in data are perpetuated by the current practice of self-supervised learning where the data is labeled by AI models trained from a Western perspective (Oquab et al. 2023; Ramaswamy et al. 2023a).

Opportunities. We need to reevaluate our current data collection practices, and collect data that covers a wide range of perspectives across demographic and cultural dimensions. Recent work has shown that even small amounts of diverse data can improve performance (Ramaswamy et al. 2024). Efforts to improve representation in AI and reduce the need for large datasets include active learning (Hady and Schwenker 2013), domain adaptation (Kalluri, Xu, and Chandraker 2023; Wang and Russakovsky 2023), similarity-based (Ignat et al. 2024) and grammar-based (Lucas et al. 2024) strategies for data augmentation. Additionally, the development of diverse datasets should involve people from various demographics to guide data collection and annotation, foster collaboration, and empower stakeholders (Nwatu, Ignat, and Mihalcea 2023).

2.2 Data Annotations

Lack of Inclusive Annotation Standards. Most annotation standards fail to account for the diversity and subjectivity inherent in global data (Nwatu, Ignat, and Mihalcea 2023). Most benchmarks adhere to popular and conventional annotation systems such as ImageNet (Deng et al. 2009), which however are not a definitive standard (Beyer et al. 2020; Fang, Kornblith, and Schmidt 2023; Shankar et al. 2020) and have not been designed with cultural representation in mind. As a result, models trained on these benchmarks tend to overfit to an incomplete 'gold standard' that does not accurately represent the world as we see it (Shankar et al. 2020; Mayfield et al. 2019), and can lead to models that do not generalize well to real-world tasks and out-of-distribution data (Fang, Kornblith, and Schmidt 2023; Taori et al. 2020).

Opportunities. A significant challenge is defining and promoting standards that address common annotation issues encountered when collecting and annotating globally diverse data. Several studies (Beyer et al. 2020; Yun et al. 2021; Shankar et al. 2020; Faghri et al. 2023) have identified issues such as label noise, errors, ambiguity, and restrictiveness in current datasets, proposing various methodologies to improve these labels. Solutions include generating robust labels using strong pre-trained models, integrating human feedback, incorporating multiple labels per image, or explicitly accounting for the diversity of annotators (Deng et al. 2023). Flexible data annotation structures that use e.g., function-based labeling could also offer a strategy for improving representation in AI datasets (Nwatu, Ignat, and Mihalcea 2023).

Lack of Diverse Annotators and Curators. An demographic report of the most popular annotation platform (Amazon Mechanical Turk) (Difallah, Filatova, and Ipeirotis 2018) found that the majority (75%) of annotators are from the United States and 16% are from India. This is concerning, since the demographic background was found to significantly influence the people's ratings and performance Pei and Jurgens (2023). Similar statistics are observed in many existing AI datasets (Kirk et al. 2024).

Opportunities. New annotation frameworks may be required for inclusive data annotations that address the unique aspects of a problem. The MaRVL dataset (Liu et al. 2021) and CVQA dataset (Romero et al. 2024) are good examples of projects that prioritize inclusive data annotation by involving native speakers in selecting concepts, questions, and images. Additionally, it is important to maintain transparency on the annotation process (Gebru et al. 2021), and share aggregate statistics for the distribution of workers by region and demographics (Kirillov et al. 2023). This information, combined with methods that identify demographic blind spots across datasets (Dominguez-Catena, Paternain, and Galar 2023) can help users make informed decisions about which datasets to use for their specific applications.

3 Models

AI models are the engines that drive the capabilities of modern AI systems, enabling them to recognize patterns, make predictions, and generate content. Their effectiveness is heavily influenced by their architecture, and as illustrated in Figure 4, their performance can be uneven and limited to specific contexts and settings.

Maya is a high school administrator in a multicultural urban area in Canada. She decides to use a new AI-driven educational tool to help her personalize the learning experiences for her students. She soon notices that the tool performs poorly when students input text in the local French dialect, often misunderstanding the context or giving the wrong output, an issue not faced by her English-speaking students.

Figure 4: Uneven model performance for different languages leads to incorrect output and can favor those speaking the languages for which the model performs better.

3.1 Model Design

Models Exhibit Biases in Pre-training and Alignment. AI models acquire their *knowledge* primarily during the pre-training phase; later, the specification for how to interact with this knowledge is provided by the process of *alignment*. The encoded knowledge is mainly determined by the dataset used in the model pre-training and thus a crucial factor for potential bias in the model; we refer to this bias as **knowledge bias**. The alignment process allows us to interact with the encoded knowledge by acting as a knowledge retriever, and can shift the model toward certain preferences; we refer to this as **alignment bias**. However, several recent investigations have demonstrated that AI models consistently excel in US-specific queries while struggling with underrepresented cultures (Shen et al. 2024; Masoud et al. 2024).

Opportunities. To mitigate knowledge bias, the primary approach is to pre-train or fine-tune LLMs on culture-specific data (Li et al. 2024a), or to rely on prompt engineering (Wang et al. 2024; Kovač et al. 2023; Rao et al. 2023). These approaches raise the important question of how to effectively gather preference data that is reflective of diverse cultures. One strategy is to manually collect preference data based on Hofstede's Cultural Alignment Test (Masoud et al. 2024). Alternatively, LLMs could engage in conversations with individuals from various cultures to generate preference data, for instance by using multi-agent settings (Li et al. 2024b). An additional challenge is the integration of diverse preferences into a single model using strategies for pluralistic alignment (Sorensen et al. 2024). or meta-reward models.

3.2 Model Performance

Poor Model Generalization. AI models often struggle to maintain performance when encountering diverse and dynamic real-world settings (Malik et al. 2024; Gustafson et al. 2023). These performance issues are further complicated by differences in geographical locations and income levels, leading to disparities in model performance across different regions and social strata (Rojas et al. 2022; Ramaswamy et al. 2023b; Nwatu, Ignat, and Mihalcea 2023). Similarly, we see biased outcomes and reduced usability across diverse linguistic communities, affecting fairness and inclusivity in AI applications (Ziems et al. 2022; Blodgett, Wei, and O'Connor 2018; Xiao et al. 2023; Malmasi et al. 2016; Zhou et al. 2021), or potentially generating responses that are inappropriate or offensive in different cultural contexts (Peterson and Gärdenfors 2023; Tay et al. 2020; Santurkar et al. 2023; Anonymous 2024; Moore, Deshpande, and Yang 2024).

Opportunities. Improving the model generalization across diverse contexts can involve active learning or other data-driven strategies that can help supplement data for underrepresented groups (Ignat et al. 2024). Alternatively, specialized benchmarks (e.g., VALUE (Ziems et al. 2022) or DADA Liu, Held, and Yang (2023)) can handle the languages spoken by diverse groups more effectively (Xiao et al. 2023; Sun et al. 2023; Hofmann et al. 2024a; Faisal et al. 2024).Another direction can consider benchmarks and training processes that account for multilingual and multicultural nuances, following initiatives such as the ETHICS dataset (Hendrycks et al. 2021), or frameworks to assess and enhance AI alignment with diverse human values (Peterson and Gärdenfors 2023; Choenni, Lauscher, and Shutova 2024).

Security Vulnerabilities and Propagation of Harmful Stereotypes. AI models are increasingly vulnerable to security breaches and the propagation of harmful stereotypes, and *jailbreaking* attacks pose a significant threat by circumventing the safety mechanisms designed to prevent models from generating unethical, harmful, or dangerous content (Ouyang et al. 2022; Rafailov et al. 2023). This is paricularly true for low-resource languages, making underrepresented groups especially vulnerable (Yong, Menghini, and Bach 2023). Similarly, AI models are prone to perpetuating the stereotypes embedded in their training data, reinforcing societal prejudices related to race, gender, and ethnicity, among others (Ferrara 2023; Hofmann et al. 2024b).

Opportunities. Addressing these intertwined challenges requires more robust alignment and defense strategies, and specialized benchmarks (Shu et al. 2024; Mazeika et al. 2024; Luo et al. 2024; Liu et al. 2024b) that can lead to a deeper understanding of the underlying mechanisms of harmful outputs and security breaches. Mechanistic interpretability for alignment algorithms (Lee et al. 2024) also offers promising avenues for controlling model responses and mitigating the success of both jailbreak attacks and the propagation of harmful stereotypes (Arditi et al. 2024; Ball, Kreuter, and Rimsky 2024). Combating the propagation of harmful stereotypes will require innovations in both the training (Li et al. 2023; Kumar et al. 2023) and post-training (Ravfogel et al. 2022; Cheng et al. 2021) phases of model development. Additionally, direct adjustments to model architectures, such as integrating awareness of harmful outputs into operational frameworks or adjusting the attention to different social groups (Gaci et al. 2022; Kim, Kim, and Johnson 2024) can also help ensure the fairness across different social groups.

4 Evaluation

Evaluating AI systems is essential for ensuring their accuracy and reliability across various applications. This process typically focuses on metrics and benchmarks, which are used to measure performance and identify any biases or limitations.

Aarav, a developer passionate about empowering Indian children through education, wants to deploy an AI-powered education tool in several Indian villages. He soon realizes that the tool was evaluated using metrics designed for Western learning styles, focusing on individual achievement and competition. When deployed in India's collectivist society that values group collaboration and shared success, the tool fails to resonate with students and produces misleading performance results.

Figure 5: A misalignment between evaluation metrics and cultural values can lead to misleading estimation of a tool's effectiveness.

4.1 Evaluation Metrics

Lack of Inclusive Metrics. While several of the metrics used to evaluate AI models may be considered generic, such as accuracy or F1 scores, there are also many metrics that only reflect the reality of specific populations. For example, reading comprehension metrics may assume familiarity with Western literary references (Steffensen, Joag-Dev, and Anderson 1979; Kolisko and Anderson 2024). Similarly, bias detection metrics, such as the Word Embedding Association Test (WEAT) (Caliskan, Bryson, and Narayanan 2017) and Sentence Association Embedding Test (SEAT) (May et al. 2019) rely on Western-centric norms (Greenwald, McGhee, and Schwartz 1998). Additionally, the human evaluations sometime used to assess system performance –through crowdsourcing, expert evaluations, or user studies – while offering a more nuanced understanding of model performance, they also heavily depend on and can be biased by the background of the evaluators (Song, Cohn, and Specia 2013; Reiter 2018).

Opportunities. To ensure a comprehensive and fair assessment, generic metrics such as accuracy and F1 scores should be coupled with diverse and inclusive datasets and metrics that assess performance across all majority and minority classes. For metrics that are inspired by the reality of specific groups, it is critical to adjust them to reflect the needs and values of those directly impacted. Combining human evaluations with automatic metrics can enhance the reliability of assessments (van der Lee et al. 2021; Schuff et al. 2023), particularly when the human evaluators represent a diverse range of backgrounds. Additionally, fairness metrics such as demographic parity (Dwork et al. 2012), equalized odds, and statistical parity (Hardt, Price, and Srebro 2016) should be used alongside traditional performance metrics to identify and address biases in AI models.

4.2 Evaluation Benchmarks

Lack of Culturally Diverse Benchmarks. Most benchmarks are heavily biased towards English-speaking and Western cultures, often focusing on limited datasets that do not

consider the visual or language diversity across cultures. Furthermore, numerous languages are frequently overlooked in benchmark construction, and this is especially true for languages that do not have a written form. Particularly challenging are sign languages (Baker 2015), and benchmarks based solely on spoken languages (Pine and Turin 2017).

Opportunities. Culturally and linguistically diverse benchmarks are crucial for developing models that can be applied globally. This involves diversifying data sources (§2.1), and ensuring the annotators are familiar with the cultural or linguistic context (§2.2). Collaboration among researchers, linguists, and cultural experts is also key to developing benchmarks that are truly representative (ÓhÉigeartaigh et al. 2020). Multimodal benchmarks are particularly promising as they can provide valuable context for models. In languages that rely solely on speech, preserving oral literature often entails documentation and collaboration with native speakers to annotate images and videos (Bird 2010; Leedom Shaul and Shaul 2014). For sign languages, ongoing efforts are being made to develop writing systems like HamNoSys (Hanke 2004) or SignWriting (Sutton 2010), although there is currently no universally accepted standard.

5 Incentives

The incentives that drive the prioritization of certain cultures and languages in current AI technology development can vary widely (Joshi et al. 2020; Bommasani et al. 2021; Bird 2020; Rogers 2021), and are often connected to the source of funding behind the AI initiatives: economic drives, government support, or philanthropic initiatives. These, in turn, can influence individual decisions, as illustrated in Figure 6.

> *Maria is the CEO of a small tech startup in Romania, with limited staff and research resources. She wants to build an AI assistant that can help people with their daily chores. However, as she delves deeper, she quickly faces a hard problem: Should she target a high purchasing power demographic like the United States and Western Europe? Or should she focus on a market where cultural values align more closely with her vision, such as Romania and Eastern Europe, even if the financial returns are less certain?*

Figure 6: Selecting the right market for AI products requires balancing cultural values and financial returns in a high-stakes decision.

Economic Drives Prioritize Rich Countries and Major Languages. One of the largest root causes for AI development lies in its economic value, both immediate and potential (Furman and Seamans 2019; Chui et al. 2023). There is generally a lower perceived profit rate for developing AI applications for smaller groups or those from lower social and economic levels (Joshi et al. 2020; Blasi, Anastasopoulos, and Neubig 2022). However, this raises the question of whether the perceived profit matches the actual potential profit, as people may underestimate the unique opportunities within smaller communities (Bird 2020; Nekoto et al. 2020). Moreover, it introduces the efficiency versus fairness dilemma, as pursuing market-driven profits can exacerbate social inequalities and stability, often referred to as the "rich get richer" effect (Hovy and Spruit 2016; Bender et al. 2021).

Opportunities. To mitigate the concentration of economic investments in highly profitable areas, we need strategies to encourage companies to balance fairness with profits (Crawford 2021; Bender et al. 2021). Governments and philanthropic organizations can create incentives to address fairness by bridging the investment gap in less profitable areas, and ensuring that the perceived economic value of an investment is not the only drive behind AI development decisions.

Inconsistent Government Support for Inclusive AI. Governments often support technologies with immediate or potential socio-economic impact, as well as those essential for national defense (Wolff and Wessner 2012; Weiss 2014; Fleming et al. 2019). While some support, e.g., for fundamental science research, can promote the development of AI technologies adaptable to underrepresented languages and communities, this support is not always consistent (Shibayama 2011; National Science Foundation 2024). Some governments lack the resources for such investments, requiring justifications for their expenditure (Bird 2020; Nekoto et al. 2020). Furthermore, prioritizing technologies that maintain a competitive national advantage can create a dilemma, as nations that invest in "good causes" to help underrepresented communities, domestic or international, may fall behind those that do not (Okun 2010; Berg and Ostry 2011).

Opportunities. It is essential to improve decision-making systems to justify expenditures on fairness and support for underprivileged groups. Research demonstrating that such investments also enhance the quality of life for most citizens can bolster support. Moreover, these decisions should take into account sensitive international contexts and strive to balance maintaining national strength with ensuring equitable AI benefits for diverse populations.

Insufficient Philanthropic Initiatives for Promoting Inclusive AI. Philanthropic organizations play a vital role in counterbalancing the market-driven allocation of resources by supporting needed areas where funding does not naturally flow (Brest and Harvey 2018; Brass et al. 2018). Although philanthropies are well-positioned to address resource distribution issues, advocating for increased support for underrepresented AI can be challenging (Jammulamadaka and Varman 2010), as these philanthropic organizations must weigh investments to help various other needs (GiveWell 2022).

Opportunities. Encouraging more philanthropic support of AI development requires carefully thought-out justifications regarding the long-term value of research and development (Pennings and Lint 1997; Pisano 2012). Helping disadvantaged groups can be approached by either providing immediate aid, or by enhancing their skills and education through technology, which leads to increased productivity (Kuhn 2020; Saiz and Donald 2018). Therefore, advocating for AI technology investment should emphasize its potential to unlock significant long-term benefits for these groups.

6 People

Although AI has the potential for global impact, research and development are dominated by a few countries (Kleinberg and Raghavan 2021; Hershcovich et al. 2022; AlKhamissi et al. 2024). Our limited experiences can blind us to biases

against the very communities we aim to help, even with the best intentions, as shown in Figure 7.

> *A healthcare AI company launches an app to streamline patient diagnoses in a rural Nigerian community without accounting for local medical practices and resources. The app misinterprets common regional symptoms, like skin rashes from heat and sun exposure, as severe conditions, resulting in unnecessary treatments. By not involving local healthcare providers and patients in its development, the app ultimately causes more harm than good.*

Figure 7: People affected by AI systems are often not involved in their development, often leading to applications that fail to addresses real problems or even cause harm.

Lack of Global Diverse Representation and Agency in AI Research. If AI systems reinforce dominant cultures, whether implicitly or explicitly, they might lead to a cycle of cultural homogeneity (Schramowski et al. 2022; Vaccino-Salvadore 2023). To ensure that the systems and resulting applications reflect people's authentic culture, representatives from these groups should also participate in the design, data construction, and model development process in a participatory approach (Bondi et al. 2021). Having a diverse group of developers is essential to foster innovation that addresses the needs and values of different cultures (Page 2010). Reciprocity and mutual learning are central to successful research engagements (Brereton et al. 2014; Taylor et al. 2019; St John and Akama 2022).

Opportunities. We need to reevaluate the power dynamics between technologists and community members and establish equal research partnerships with the community (Bird and Yibarbuk 2024; Mignolo 2012). This approach follows a decolonizing practice that respects the sovereignty of local communities and prioritizes their input (Bird 2020). Additionally, several workshops and events have begun to explore how to empower stakeholders in the development and deployment of technology (Vaccaro et al. 2019; Givens and Morris 2020) and how to help researchers and practitioners consider when not to build systems at all (Barocas et al. 2020). Other examples of offering mentorship are open-source and free initiatives such as Masakhane,[2] Black in AI,[3], AmericasNLP,[4] and the ACL Mentorship,[5] just to mention a few.

Lack of Practical Applications and Real-Life Grounding in AI Research. Language is inherently situated. However, many AI systems and AI research fail to clearly articulate what problems they tackle. For instance, a survey of 146 papers on bias in NLP systems shows that their motivations are often vague, inconsistent, and lack normative reasoning (Blodgett et al. 2020). Furthermore, different social groups, especially those at the intersections of multiple axes of oppression, have different lived experiences due to their different social positions (Sassaman et al. 2020; Field et al. 2021).

Opportunities. To ensure our AI models effectively serve the communities they are intended to help, it is crucial to first understand their specific needs and how they plan to use the technology. This requires focusing on the lived experiences of those directly impacted by these systems (Blodgett et al. 2020). Cross-disciplinary collaboration is particularly important for real-world impact. For example, sociolinguists and anthropologists have studied how language varieties are perceived—whether as standard, correct, or uneducated (Reaser et al. 2018; Roche 2019; Craft et al. 2020). This research reveals that beliefs about language often mirror deeper beliefs about the speakers themselves (Rosa and Flores 2017). Understanding the role of language in maintaining social hierarchies is vital for improving bias analysis in NLP systems and addressing how racial ideologies both shape and are shaped by technology (Ruha 2024).

7 Conclusion

One of the key concerns about the rapid integration of AI technologies into everyday life is the risk of widening the socio-economic gap by disproportionately benefiting certain groups while marginalizing others, as well as the potentially negative impact it can have on individuals and society. On an individual level, it affects people's perception of themselves and others, influencing their interactions and the opportunities they have access to. On a societal level, the widespread use of non-inclusive AI systems can shape cultural norms and social structures and support discriminatory narratives that hinder efforts toward equality and inclusivity.

The vision we lay out in this paper highlights opportunities to develop AI for everyone, with everyone, by everyone. Our key recommendations focus on improving inclusivity along the entire AI pipeline, specifically targeting the five main areas we addressed in this paper: **Data**: Prioritize the development of transparent, responsible, and data-centric AI models; Reevaluate current data collection practices to ensure coverage of diverse perspectives across demographic and cultural dimensions; Define inclusive annotation standards to improve the representation in training data; Adopt a participatory approach to AI, engaging community members early in the research process. **Models**: Architect model designs that minimize knowledge and alignment bias; Improve model generalization to ensure performance across diverse groups and contexts; Build models that are robust and not vulnerable when used by underrepresented groups. **Evaluations**: Create inclusive metrics that assess performance across both majority and minority groups, and accurately reflect the reality of the target users; Develop culturally and linguistically diverse evaluation benchmarks. **Incentives:** Devise economic strategies to encourage companies to balance fairness with profits; Shape government agendas that promote both national strength and equitable AI benefits for all; Encourage philanthropic support of long-term AI research agendas focused on inclusivity. **People.** Foster the growth of a diverse AI workforce; Establish equal research partnerships with communities; Focus on the lived experiences of those directly impacted by these systems.

Each of these recommendations requires concrete steps to ensure the AI systems are equitable, robust, and representative of all populations. By addressing these core areas, we can advance towards AI systems that serve everyone, are built with input from a wide range of perspectives, and reflect the contributions of a diverse group of stakeholders.

[2]https://masakhane.io

[3]https://blackinai.github.io

[4]https://turing.iimas.unam.mx/americasnlp

[5]https://mentorship.aclweb.org

References

Adilazuarda, M. F.; Mukherjee, S.; Lavania, P.; Singh, S.; Dwivedi, A.; Aji, A. F.; O'Neill, J.; Modi, A.; and Choudhury, M. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Aguinis, H.; Villamor, I.; and Ramani, R. S. 2020. MTurk Research: Review and Recommendations. *Journal of Management*, 47: 823 – 837.

AlKhamissi, B.; ElNokrashy, M.; AlKhamissi, M.; and Diab, M. 2024. Investigating Cultural Alignment of Large Language Models. arXiv:2402.13231.

Anonymous. 2024. From Instructions to Basic Human Values: A Survey of Alignment Goals for Big Model. In *Submitted to ACL Rolling Review - June 2024*. Under review.

Arditi, A.; Obeso, O.; Syed, A.; Paleka, D.; Rimsky, N.; Gurnee, W.; and Nanda, N. 2024. Refusal in Language Models Is Mediated by a Single Direction. *arXiv preprint arXiv:2406.11717*.

Baker, A. 2015. Sign languages as natural languages. In *Anne Baker, Beppie van den Boegarde, Roland Pfau, and Trude Schermer, editors, Sign Languages of the World: A Comparative Handbook, chapter 31, De Gruyter, Berlin*, 729–770.

Ball, S.; Kreuter, F.; and Rimsky, N. 2024. Understanding Jailbreak Success: A Study of Latent Space Dynamics in Large Language Models. *arXiv preprint arXiv:2406.09289*.

Barocas, S.; Biega, A. J.; Fish, B.; Niklas, J.; and Stark, L. 2020. When not to design, build, or deploy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 695–695.

Bender, E. M.; and Friedman, B. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics*, 6: 587–604.

Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Berg, A.; and Ostry, J. D. 2011. Equality and efficiency. *Finance & Development*, 48(3): 12–15.

Beyer, L.; Hénaff, O. J.; Kolesnikov, A.; Zhai, X.; and Oord, A. v. d. 2020. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*.

Bhardwaj, A.; Bhattacherjee, S.; Chavan, A.; Deshpande, A.; Elmore, A. J.; Madden, S.; and Parameswaran, A. G. 2014. Datahub: Collaborative data science & dataset version management at scale. *arXiv preprint arXiv:1409.0798*.

Bird, S. 2010. A scalable method for preserving oral literature from small languages. In *International Conference on Asian Digital Libraries*, 5–14. Springer.

Bird, S. 2020. Decolonising Speech and Language Technology. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics*, 3504–3519. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Bird, S.; and Yibarbuk, D. 2024. Centering the Speech Community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 826–839.

Blasi, D.; Anastasopoulos, A.; and Neubig, G. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5486–5505. Dublin, Ireland: Association for Computational Linguistics.

Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics.

Blodgett, S. L.; Wei, J.; and O'Connor, B. 2018. Twitter Universal Dependency Parsing for African-American and Mainstream American English. In Gurevych, I.; and Miyao, Y., eds., *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1415–1425. Melbourne, Australia: Association for Computational Linguistics.

Bommasani, R.; Hudson, D. A.; Adeli, E.; Altman, R. B.; Arora, S.; von Arx, S.; Bernstein, M. S.; Bohg, J.; Bosselut, A.; Brunskill, E.; Brynjolfsson, E.; Buch, S.; Card, D.; Castellon, R.; Chatterji, N. S.; Chen, A. S.; Creel, K.; Davis, J. Q.; Demszky, D.; Donahue, C.; Doumbouya, M.; Durmus, E.; Ermon, S.; Etchemendy, J.; Ethayarajh, K.; Fei-Fei, L.; Finn, C.; Gale, T.; Gillespie, L.; Goel, K.; Goodman, N. D.; Grossman, S.; Guha, N.; Hashimoto, T.; Henderson, P.; Hewitt, J.; Ho, D. E.; Hong, J.; Hsu, K.; Huang, J.; Icard, T.; Jain, S.; Jurafsky, D.; Kalluri, P.; Karamcheti, S.; Keeling, G.; Khani, F.; Khattab, O.; Koh, P. W.; Krass, M. S.; Krishna, R.; Kuditipudi, R.; and et al. 2021. On the Opportunities and Risks of Foundation Models. *CoRR*, abs/2108.07258.

Bondi, E.; Xu, L.; Acosta-Navas, D.; and Killian, J. A. 2021. Envisioning communities: a participatory approach towards AI for social good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 425–436.

Brass, J. N.; Longhofer, W.; Robinson, R. S.; and Schnable, A. 2018. NGOs and international development: A review of thirty-five years of scholarship. *World Development*, 112: 136–149.

Brereton, M.; Roe, P.; Schroeter, R.; and Lee Hong, A. 2014. Beyond ethnography: engagement and reciprocity as foundations for design research out here. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1183–1186.

Brest, P.; and Harvey, H. 2018. *Money well spent: A strategic plan for smart philanthropy*. Stanford University Press.

Buolamwini, J.; and Gebru, T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, 77–91. PMLR.

Byrd, A. 2023. Truth-Telling: Critical Inquiries on LLMs and the Corpus Texts That Train Them. *Composition studies*, 51(1): 135–142.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.

Cheng, P.; Hao, W.; Yuan, S.; Si, S.; and Carin, L. 2021. FairFil: Contrastive Neural Debiasing Method for Pretrained

Text Encoders. In *International Conference on Learning Representations*.

Choenni, R.; Lauscher, A.; and Shutova, E. 2024. The Echoes of Multilinguality: Tracing Cultural Value Shifts during Language Model Fine-tuning. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15042–15058. Bangkok, Thailand: Association for Computational Linguistics.

Chui, M.; Hazan, E.; Roberts, R.; Singla, A.; and Smaje, K. 2023. The economic potential of generative AI.

Craft, J. T.; Wright, K. E.; Weissler, R. E.; and Queen, R. M. 2020. Language and discrimination: Generating meaning, perceiving identities, and discriminating outcomes. *Annual Review of Linguistics*, 6: 389–407.

Crawford, K. 2021. *Atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Deng, N.; Zhang, X.; Liu, S.; Wu, W.; Wang, L.; and Mihalcea, R. 2023. You Are What You Annotate: Towards Better Models through Annotator Representations. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12475–12498. Singapore: Association for Computational Linguistics.

Difallah, D.; Filatova, E.; and Ipeirotis, P. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the eleventh ACM international conference on web search and data mining*, 135–143.

Dominguez-Catena, I.; Paternain, D.; and Galar, M. 2023. DSAP: Analyzing Bias Through Demographic Comparison of Datasets. *arXiv preprint arXiv:2312.14626*.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, 214–226. New York, NY, USA: Association for Computing Machinery. ISBN 9781450311151.

Faghri, F.; Pouransari, H.; Mehta, S.; Farajtabar, M.; Farhadi, A.; Rastegari, M.; and Tuzel, O. 2023. Reinforce data, multiply impact: Improved model accuracy and robustness with dataset reinforcement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17032–17043.

Faisal, F.; Ahia, O.; Srivastava, A.; Ahuja, K.; Chiang, D.; Tsvetkov, Y.; and Anastasopoulos, A. 2024. DIALECT-BENCH: An NLP Benchmark for Dialects, Varieties, and Closely-Related Languages. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14412–14454. Bangkok, Thailand: Association for Computational Linguistics.

Fang, A.; Kornblith, S.; and Schmidt, L. 2023. Does progress on ImageNet transfer to real-world datasets? In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 25050–25080. Curran Associates, Inc.

Ferrara, E. 2023. Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies. *Sci*, 6(1): 3. ArXiv:2304.07683 [cs].

Field, A.; Blodgett, S. L.; Waseem, Z.; and Tsvetkov, Y. 2021. A Survey of Race, Racism, and Anti-Racism in NLP. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1905–1925. Online: Association for Computational Linguistics.

Fleming, L.; Greene, H.; Li, G.; Marx, M.; and Yao, D. 2019. Government-funded research increasingly fuels innovation. *Science*, 364(6446): 1139–1141.

Fung, Y.; Chakrabarty, T.; Guo, H.; Rambow, O.; Muresan, S.; and Ji, H. 2023. NORMSAGE: Multi-Lingual Multi-Cultural Norm Discovery from Conversations On-the-Fly. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 15217–15230. Singapore: Association for Computational Linguistics.

Furman, J.; and Seamans, R. 2019. AI and the Economy. *Innovation policy and the economy*, 19(1): 161–191.

Gaci, Y.; Benatallah, B.; Casati, F.; and Benabdeslem, K. 2022. Debiasing Pretrained Text Encoders by Paying Attention to Paying Attention. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9582–9602. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Gebru, T. 2020. Race and gender. *The Oxford handbook of ethics of AI*, 251–269.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.

Givens, A. R.; and Morris, M. R. 2020. Centering disability perspectives in algorithmic fairness, accountability, & transparency. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 684–684.

GiveWell. 2022. GiveWell Metrics Report 2022. Report, GiveWell. Accessed: 2024-09-02.

Greenwald, A. G.; McGhee, D. E.; and Schwartz, J. L. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6): 1464.

Gustafson, L.; Richards, M.; Hall, M.; Hazirbas, C.; Bouchacourt, D.; and Ibrahim, M. 2023. Pinpointing Why Object Recognition Performance Degrades Across Income Levels and Geographies. arXiv:2304.05391.

Hady, M. F. A.; and Schwenker, F. 2013. Semi-supervised learning. *Handbook on Neural Information Processing*, 215–239.

Hanke, T. 2004. HamNoSys–representing sign language data in language resources and language processing contexts. In *sign-lang@ LREC 2004*, 1–6. European Language Resources Association (ELRA).

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. arXiv:1610.02413.

Hendrycks, D.; Burns, C.; Basart, S.; Critch, A.; Li, J.; Song, D.; and Steinhardt, J. 2021. Aligning {AI} With Shared Human Values. In *International Conference on Learning Representations*.

Henrich, J. 2015. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton University Press.

Henrich, J.; Heine, S. J.; and Norenzayan, A. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3): 61–83.

Hershcovich, D.; Frank, S.; Lent, H.; de Lhoneux, M.; Abdou, M.; Brandl, S.; Bugliarello, E.; Cabello Piqueras, L.; Chalkidis, I.; Cui, R.; Fierro, C.; Margatina, K.; Rust, P.; and Søgaard, A. 2022. Challenges and Strategies in Cross-Cultural NLP. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6997–7013. Dublin, Ireland: Association for Computational Linguistics.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024a. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv:2403.00742.

Hofmann, V.; Kalluri, P. R.; Jurafsky, D.; and King, S. 2024b. Dialect prejudice predicts AI decisions about people's character, employability, and criminality. *Computing Research Repository*, arXiv:2403.00742.

Holland, S.; Hosny, A.; Newman, S.; Joseph, J.; and Chmielinski, K. 2020. The dataset nutrition label. *Data Protection and Privacy*, 12(12): 1.

Hovy, D.; and Spruit, S. L. 2016. The Social Impact of Natural Language Processing. In Erk, K.; and Smith, N. A., eds., *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 591–598. Berlin, Germany: Association for Computational Linguistics.

Ignat, O.; Bai, L.; Nwatu, J. C.; and Mihalcea, R. 2024. Annotations on a Budget: Leveraging Geo-Data Similarity to Balance Model Performance and Annotation Cost. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 1239–1259. Torino, Italia.

Jammulamadaka, N.; and Varman, R. 2010. Is NGO development assistance mistargeted? An epistemological approach. *Critical Review*, 22(2-3): 117–128.

Joshi, P.; Santy, S.; Budhiraja, A.; Bali, K.; and Choudhury, M. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6282–6293. Online: Association for Computational Linguistics.

Kalluri, T.; Xu, W.; and Chandraker, M. 2023. Geonet: Benchmarking unsupervised adaptation across geographies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15368–15379.

Keleg, A.; and Magdy, W. 2023. DLAMA: A Framework for Curating Culturally Diverse Facts for Probing the Knowledge of Pretrained Language Models. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Findings of the Association for Computational Linguistics: ACL 2023*, 6245–6266. Toronto, Canada: Association for Computational Linguistics.

Kim, M. Y.; Kim, J.; and Johnson, K. 2024. ABLE: Agency-BeLiefs Embedding to Address Stereotypical Bias through Awareness Instead of Obliviousness. In Calzolari, N.; Kan, M.-Y.; Hoste, V.; Lenci, A.; Sakti, S.; and Xue, N., eds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 43–56. Torino, Italia: ELRA and ICCL.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Kirk, H. R.; Whitefield, A.; Röttger, P.; Bean, A.; Margatina, K.; Ciro, J.; Mosquera, R.; Bartolo, M.; Williams, A.; He, H.; et al. 2024. The PRISM Alignment Project: What Participatory, Representative and Individualised Human Feedback Reveals About the Subjective and Multicultural Alignment of Large Language Models. *arXiv preprint arXiv:2404.16019*.

Kleinberg, J.; and Raghavan, M. 2021. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22): e2018340118.

Kolisko, S.; and Anderson, C. J. 2024. Exploring Social Biases of Large Language Models in a College Artificial Intelligence Course. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(13): 15825–15833.

Kovač, G.; Sawayama, M.; Portelas, R.; Colas, C.; Dominey, P. F.; and Oudeyer, P.-Y. 2023. Large Language Models as Superpositions of Cultural Perspectives. arXiv:2307.07870.

Kuhn, H. 2020. Reducing inequality within and among countries: Realizing SDG 10—A developmental perspective. *Sustainable development goals and human rights*, 5: 137–153.

Kumar, D.; Lesota, O.; Zerveas, G.; Cohen, D.; Eickhoff, C.; Schedl, M.; and Rekabsaz, N. 2023. Parameter-efficient Modularised Bias Mitigation via AdapterFusion. In Vlachos, A.; and Augenstein, I., eds., *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2738–2751. Dubrovnik, Croatia: Association for Computational Linguistics.

Lee, A.; Bai, X.; Pres, I.; Wattenberg, M.; Kummerfeld, J. K.; and Mihalcea, R. 2024. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*.

Leedom Shaul, D.; and Shaul, D. L. 2014. Language Preservation Begets Language Documentation. *Linguistic Ideologies of Native American Language Revitalization: Doing the Lost Language Ghost Dance*, 11–21.

Li, C.; Chen, M.; Wang, J.; Sitaram, S.; and Xie, X. 2024a. CultureLLM: Incorporating Cultural Differences into Large Language Models. arXiv:2402.10946.

Li, C.; Teney, D.; Yang, L.; Wen, Q.; Xie, X.; and Wang, J. 2024b. CulturePark: Boosting Cross-cultural Understanding in Large Language Models. arXiv:2405.15145.

Li, Y.; Du, M.; Wang, X.; and Wang, Y. 2023. Prompt Tuning Pushes Farther, Contrastive Learning Pulls Closer: A Two-Stage Approach to Mitigate Social Biases. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the*

61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 14254–14267. Toronto, Canada: Association for Computational Linguistics.

Liu, F.; Bugliarello, E.; Ponti, E. M.; Reddy, S.; Collier, N.; and Elliott, D. 2021. Visually Grounded Reasoning across Languages and Cultures. In Moens, M.-F.; Huang, X.; Specia, L.; and Yih, S. W.-t., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 10467–10485. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.

Liu, J.; Li, W.; Jin, Z.; and Diab, M. 2024a. Automatic Generation of Model and Data Cards: A Step Towards Responsible AI. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 1975–1997. Mexico City, Mexico: Association for Computational Linguistics.

Liu, X.; Zhu, Y.; Gu, J.; Lan, Y.; Yang, C.; and Qiao, Y. 2024b. MM-SafetyBench: A Benchmark for Safety Evaluation of Multimodal Large Language Models. arXiv:2311.17600.

Liu, Y.; Held, W.; and Yang, D. 2023. DADA: Dialect Adaptation via Dynamic Aggregation of Linguistic Rules. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13776–13793. Singapore: Association for Computational Linguistics.

Longpre, S.; Mahari, R.; Chen, A.; Obeng-Marnu, N.; Sileo, D.; Brannon, W.; Muennighoff, N.; Khazam, N.; Kabbara, J.; Perisetla, K.; et al. 2023. The data provenance initiative: A large scale audit of dataset licensing & attribution in ai. *arXiv preprint arXiv:2310.16787*.

Lucas, A.; Baladón, A.; Pardiñas, V.; Agüero-Torales, M.; Góngora, S.; and Chiruzzo, L. 2024. Grammar-based Data Augmentation for Low-Resource Languages: The Case of Guarani-Spanish Neural Machine Translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 6385–6397.

Luo, W.; Ma, S.; Liu, X.; Guo, X.; and Xiao, C. 2024. Jailbreakv-28k: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks. *arXiv preprint arXiv:2404.03027*.

Malik, H. S.; Huzaifa, M.; Naseer, M.; Khan, S.; and Khan, F. S. 2024. ObjectCompose: Evaluating Resilience of Vision-Based Models on Object-to-Background Compositional Changes. arXiv:2403.04701.

Malmasi, S.; Zampieri, M.; Ljubešić, N.; Nakov, P.; Ali, A.; and Tiedemann, J. 2016. Discriminating between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task. In Nakov, P.; Zampieri, M.; Tan, L.; Ljubešić, N.; Tiedemann, J.; and Malmasi, S., eds., *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, 1–14. Osaka, Japan: The COLING 2016 Organizing Committee.

Masoud, R. I.; Liu, Z.; Ferianc, M.; Treleaven, P.; and Rodrigues, M. 2024. Cultural Alignment in Large Language Models: An Explanatory Analysis Based on Hofstede's Cultural Dimensions. arXiv:2309.12342.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 622–628. Minneapolis, Minnesota: Association for Computational Linguistics.

Mayfield, E.; Madaio, M.; Prabhumoye, S.; Gerritsen, D.; McLaughlin, B.; Dixon-Román, E.; and Black, A. W. 2019. Equity Beyond Bias in Language Technologies for Education. In Yannakoudakis, H.; Kochmar, E.; Leacock, C.; Madnani, N.; Pilán, I.; and Zesch, T., eds., *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 444–460. Florence, Italy: Association for Computational Linguistics.

Mazeika, M.; Phan, L.; Yin, X.; Zou, A.; Wang, Z.; Mu, N.; Sakhaee, E.; Li, N.; Basart, S.; Li, B.; et al. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *arXiv preprint arXiv:2402.04249*.

McVeety, S. 2024. DIGITAL ALLOTMENT AND VANISHING INDIANS: IDSOV AND LLMS. *American Indian Law Journal*, 12(2): 4.

Mignolo, W. D. 2012. Local histories-global designs: Coloniality, subaltern knowledges, and border thinking.

Moayeri, M.; Tabassi, E.; and Feizi, S. 2024. WorldBench: Quantifying Geographic Disparities in LLM Factual Recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1211–1228. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.

Moher, D.; Hopewell, S.; Schulz, K. F.; Montori, V.; Gøtzsche, P. C.; Devereaux, P. J.; Elbourne, D.; Egger, M.; and Altman, D. G. 2010. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ*, 340.

Moore, J.; Deshpande, T.; and Yang, D. 2024. Are Large Language Models Consistent over Value-laden Questions? arXiv:2407.02996.

Naous, T.; Ryan, M.; Ritter, A.; and Xu, W. 2024. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16366–16393. Bangkok, Thailand: Association for Computational Linguistics.

National Science Foundation. 2024. NSF Award Search: Underrepresented. `https://www.nsf.gov/awardsearch/simpleSearchResult?queryText=underrepresented`. Accessed: 2024-08-15.

Nayak, S.; Jain, K.; Awal, R.; Reddy, S.; van Steenkiste, S.; Hendricks, L. A.; Stańczak, K.; and Agrawal, A. 2024. Benchmarking Vision Language Models for Cultural Understanding. *arXiv preprint arXiv:2407.10920*.

Nekoto, W.; Marivate, V.; Matsila, T.; Fasubaa, T.; Fagbohungbe, T.; Akinola, S. O.; Muhammad, S.; Kabongo Kabenamualu, S.; 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.; Tajudeen,

K.; Degila, K.; Ogueji, K.; Siminyu, K.; Kreutzer, J.; Webster, J.; Ali, J. T.; Abbott, J.; Orife, I.; Ezeani, I.; Dangana, I. A.; Kamper, H.; Elsahar, H.; Duru, G.; Kioko, G.; Espoir, M.; van Biljon, E.; Whitenack, D.; Onyefuluchi, C.; Emezue, C. C.; Dossou, B. F. P.; Sibanda, B.; Bassey, B.; Olabiyi, A.; Ramkilowan, A.; Öktem, A.; Akinfaderin, A.; and Bashir, A. 2020. Participatory Research for Low-resourced Machine Translation: A Case Study in African Languages. In Cohn, T.; He, Y.; and Liu, Y., eds., *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2144–2160. Online: Association for Computational Linguistics.

Nguyen, T.-P.; Razniewski, S.; Varde, A. S.; and Weikum, G. 2022. Extracting Cultural Commonsense Knowledge at Scale. *Proceedings of the ACM Web Conference 2023*.

Nwatu, J.; Ignat, O.; and Mihalcea, R. 2023. Bridging the Digital Divide: Performance Variation across Socio-Economic Factors in Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 10686–10702.

ÓhÉigeartaigh, S. S.; Whittlestone, J.; Liu, Y.; Zeng, Y.; and Liu, Z. 2020. Overcoming barriers to cross-cultural cooperation in AI ethics and governance. *Philosophy & technology*, 33: 571–593.

Okun, A. M. 2010. *Equality and efficiency: The big tradeoff*. Brookings Institution Press.

Olugbade, T.; Bieńkiewicz, M.; Barbareschi, G.; D'amato, V.; Oneto, L.; Camurri, A.; Holloway, C.; Björkman, M.; Keller, P.; Clayton, M.; Williams, A. D. C.; Gold, N.; Becchio, C.; Bardy, B.; and Bianchi-Berthouze, N. 2024. Human Movement Datasets: An Interdisciplinary Scoping Review. *ACM Computing Surveys*, 55.

Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H. V.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; HAZIZA, D.; Massa, F.; El-Nouby, A.; et al. 2023. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744.

Page, S. 2010. Diversity and complexity. In *Diversity and complexity*. Princeton University Press.

Palta, S.; and Rudinger, R. 2023. FORK: A Bite-Sized Test Set for Probing Culinary Cultural Biases in Commonsense Reasoning Models. In *Annual Meeting of the Association for Computational Linguistics*.

Parmar, M.; Mishra, S.; Geva, M.; and Baral, C. 2022. Don't Blame the Annotator: Bias Already Starts in the Annotation Instructions. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Pei, J.; and Jurgens, D. 2023. When do annotator demographics matter? measuring the influence of annotator demographics with the popquorn dataset. *arXiv preprint arXiv:2306.06826*.

Pennings, E.; and Lint, O. 1997. The option value of advanced R & D. *European Journal of Operational Research*, 103(1): 83–94.

Perez, J.; Léger, C.; Kovač, G.; Colas, C.; Molinaro, G.; Derex, M.; Oudeyer, P.-Y.; and Moulin-Frier, C. 2024. When LLMs Play the Telephone Game: Cumulative Changes and Attractors in Iterated Cultural Transmissions. *arXiv preprint arXiv:2407.04503*.

Peterson, M.; and Gärdenfors, P. 2023. How to measure value alignment in AI. *AI and Ethics*, 1–14.

Pine, A.; and Turin, M. 2017. Language revitalization.

Pisano, G. P. 2012. Creating an R&D strategy. *Boston, MA: Harvard Business School*.

Pushkarna, M.; Zaldivar, A.; and Kjartansson, O. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Ramaswamy, V. V.; Lin, S. Y.; Zhao, D.; Adcock, A.; van der Maaten, L.; Ghadiyaram, D.; and Russakovsky, O. 2024. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36.

Ramaswamy, V. V.; Lin, S. Y.; Zhao, D.; Adcock, A. B.; van der Maaten, L.; Ghadiyaram, D.; and Russakovsky, O. 2023a. Beyond web-scraping: Crowd-sourcing a geographically diverse image dataset. *arXiv preprint arXiv:2301.02560*.

Ramaswamy, V. V.; Lin, S. Y.; Zhao, D.; Adcock, A. B.; van der Maaten, L.; Ghadiyaram, D.; and Russakovsky, O. 2023b. GeoDE: a Geographically Diverse Evaluation Dataset for Object Recognition. arXiv:2301.02560.

Rao, A.; Khandelwal, A.; Tanmay, K.; Agarwal, U.; and Choudhury, M. 2023. Ethical Reasoning over Moral Alignment: A Case and Framework for In-Context Ethical Policies in LLMs. arXiv:2310.07251.

Ravfogel, S.; Twiton, M.; Goldberg, Y.; and Cotterell, R. D. 2022. Linear Adversarial Concept Erasure. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 18400–18421. PMLR.

Reaser, J.; Wilbanks, E.; Wojcik, K.; and Wolfram, W. 2018. *Language variety in the new South: Contemporary perspectives on change and variation*. UNC Press Books.

Reiter, E. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3): 393–401.

Roche, G. 2019. Articulating language oppression: Colonialism, coloniality and the erasure of Tibet's minority languages. *Patterns of prejudice*, 53(5): 487–514.

Rogers, A. 2021. Changing the World by Changing the Data. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2182–2194. Online: Association for Computational Linguistics.

Rojas, W. A. G.; Diamos, S.; Kini, K. R.; Kanter, D.; Reddi, V. J.; and Coleman, C. 2022. The Dollar Street Dataset: Images Representing the Geographic and Socioeconomic Diversity of the World. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Romero, D.; Lyu, C.; Wibowo, H. A.; Lynn, T.; Hamed, I.; Kishore, A. N.; Mandal, A.; Dragonetti, A.; Abzaliev, A.; Tonja, A. L.; et al. 2024. CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark. *arXiv preprint arXiv:2406.05967*.

Rosa, J.; and Flores, N. 2017. Unsettling race and language: Toward a raciolinguistic perspective. *Language in society*, 46(5): 621–647.

Ruha, B. 2024. *Race after technology: Abolitionist tools for the new Jim code*. Oxford: Polity.

Saiz, I.; and Donald, K. 2018. Tackling inequality through the Sustainable Development Goals: human rights in practice. In *The sustainable development goals and human rights*, 7–27. Routledge.

Santurkar, S.; Durmus, E.; Ladhak, F.; Lee, C.; Liang, P.; and Hashimoto, T. 2023. Whose Opinions Do Language Models Reflect? arXiv:2303.17548.

Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 2021. Annotators with Attitudes: How Annotator Beliefs And Identities Bias Toxic Language Detection. In *North American Chapter of the Association for Computational Linguistics*.

Sassaman, H.; Lee, J.; Irvine, J.; and Narayan, S. 2020. Creating community-based tech policy: case studies, lessons learned, and what technologists and communities can do together. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 685–685.

Schramowski, P.; Turan, C.; Andersen, N.; Rothkopf, C. A.; and Kersting, K. 2022. Large pre-trained language models contain human-like biases of what is right and wrong to do. *Nature Machine Intelligence*, 4(3): 258–268.

Schuff, H.; Vanderlyn, L.; Adel, H.; and Vu, N. T. 2023. How to do human evaluation: A brief introduction to user studies in NLP. *Natural Language Engineering*, 29(5): 1199–1222.

Schwöbel, P.; Golebiowski, J.; Donini, M.; Archambeau, C.; and Pruthi, D. 2023. Geographical Erasure in Language Generation. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12310–12324. Singapore: Association for Computational Linguistics.

Shankar, V.; Roelofs, R.; Mania, H.; Fang, A.; Recht, B.; and Schmidt, L. 2020. Evaluating Machine Accuracy on ImageNet. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 8634–8644. PMLR.

Shen, S.; Logeswaran, L.; Lee, M.; Lee, H.; Poria, S.; and Mihalcea, R. 2024. Understanding the Capabilities and Limitations of Large Language Models for Cultural Commonsense. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5668–5680.

Shibayama, S. 2011. Distribution of academic research funds: A case of Japanese national research grant. *Scientometrics*, 88(1): 43–60.

Shu, D.; Jin, M.; Zhu, S.; Wang, B.; Zhou, Z.; Zhang, C.; and Zhang, Y. 2024. AttackEval: How to Evaluate the Effectiveness of Jailbreak Attacking on Large Language Models. arXiv:2401.09002.

Shwartz, V. 2022. Good Night at 4 pm?! Time Expressions in Different Cultures. In *Findings*.

Song, X.; Cohn, T.; and Specia, L. 2013. BLEU Deconstructed: Designing a Better MT Evaluation Metric. *Int. J. Comput. Linguistics Appl.*, 4(2): 29–44.

Sorensen, T.; Moore, J.; Fisher, J.; Gordon, M.; Mireshghallah, N.; Rytting, C. M.; Ye, A.; Jiang, L.; Lu, X.; Dziri, N.; et al. 2024. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*.

St John, N.; and Akama, Y. 2022. Reimagining co-design on Country as a relational and transformational practice. *CoDesign*, 18(1): 16–31.

Steffensen, M. S.; Joag-Dev, C.; and Anderson, R. C. 1979. A cross-cultural perspective on reading comprehension. *Reading research quarterly*, 10–29.

Sun, J.; Sellam, T.; Clark, E.; Vu, T.; Dozat, T.; Garrette, D.; Siddhant, A.; Eisenstein, J.; and Gehrmann, S. 2023. Dialect-robust Evaluation of Generated Text. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 6010–6028. Toronto, Canada: Association for Computational Linguistics.

Sutton, V. 2010. The SignWriting Alphabet. *Read and Write any Sign Language in the World. ISWA Manual*.

Taori, R.; Dave, A.; Shankar, V.; Carlini, N.; Recht, B.; and Schmidt, L. 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33: 18583–18599.

Tay, Y.; Ong, D.; Fu, J.; Chan, A.; Chen, N.; Luu, A. T.; and Pal, C. 2020. Would you Rather? A New Benchmark for Learning Machine Alignment with Cultural Values and Social Preferences. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5369–5373. Online: Association for Computational Linguistics.

Taylor, J. L.; Council, W. W. A. S.; Soro, A.; Roe, P.; and Brereton, M. 2019. A relational approach to designing social technologies that foster use of the Kuku Yalanji language. In *Proceedings of the 31st Australian conference on human-computer-interaction*, 161–172.

Thompson, B.; Roberts, S. G.; and Lupyan, G. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4(10): 1029–1038.

Vaccaro, K.; Karahalios, K.; Mulligan, D. K.; Kluttz, D.; and Hirsch, T. 2019. Contestability in algorithmic systems. In *Companion Publication of the 2019 Conference on Computer Supported Cooperative Work and Social Computing*, 523–527.

Vaccino-Salvadore, S. 2023. Exploring the ethical dimensions of using ChatGPT in language learning and beyond. *Languages*, 8(3): 191.

van der Lee, C.; Gatt, A.; van Miltenburg, E.; and Krahmer, E. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Computer Speech & Language*, 67: 101151.

Villalobos, P.; Sevilla, J.; Heim, L.; Besiroglu, T.; Hobbhahn, M.; and Ho, A. 2022. Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv preprint arXiv:2211.04325*.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2024. Do large language models have a legal duty to tell the truth? *Royal Society Open Science*, 11(8): 240197.

Wang, A.; and Russakovsky, O. 2023. Overwriting pretrained bias with finetuning data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3957–3968.

Wang, W.; Jiao, W.; Huang, J.; Dai, R.; tse Huang, J.; Tu, Z.; and Lyu, M. R. 2024. Not All Countries Celebrate Thanksgiving: On the Cultural Dominance in Large Language Models. arXiv:2310.12481.

Weiss, L. 2014. *America Inc.?: innovation and enterprise in the national security state*. Cornell University Press.

Welch, C.; Kummerfeld, J. K.; Pérez-Rosas, V.; and Mihalcea, R. 2020. Compositional Demographic Word Embeddings. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4076–4089. Online: Association for Computational Linguistics.

Wiechetek, L.; Pirinen, F. A.; Gaup, B.; Trosterud, T.; Kappfjell, M. L.; and Moshagen, S. 2024. The Ethical Question–Use of Indigenous Corpora for Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 15922–15931.

Wolff, A. W.; and Wessner, C. W. 2012. Rising to the challenge: US innovation policy for the global economy.

Xiao, Z.; Held, W.; Liu, Y.; and Yang, D. 2023. Task-Agnostic Low-Rank Adapters for Unseen English Dialects. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7857–7870. Singapore: Association for Computational Linguistics.

Yin, D.; Bansal, H.; Monajatipoor, M.; Li, L. H.; and Chang, K.-W. 2022. GeoMLAMA: Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models. In *Conference on Empirical Methods in Natural Language Processing*.

Yong, Z. X.; Menghini, C.; and Bach, S. 2023. Low-Resource Languages Jailbreak GPT-4. In *Socially Responsible Language Modelling Research*.

Yun, S.; Oh, S. J.; Heo, B.; Han, D.; Choe, J.; and Chun, S. 2021. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2340–2350.

Zhang, S.; Dong, L.; Li, X.; Zhang, S.; Sun, X.; Wang, S.; Li, J.; Hu, R.; Zhang, T.; Wu, F.; et al. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Zhou, X.; Sap, M.; Swayamdipta, S.; Choi, Y.; and Smith, N. 2021. Challenges in Automated Debiasing for Toxic Language Detection. In Merlo, P.; Tiedemann, J.; and Tsarfaty, R., eds., *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 3143–3155. Online: Association for Computational Linguistics.

Ziems, C.; Chen, J.; Harris, C.; Anderson, J.; and Yang, D. 2022. VALUE: Understanding Dialect Disparity in NLU. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 3701–3720. Dublin, Ireland: Association for Computational Linguistics.