Social Bias in Multilingual Language Models: A Survey

Lance Calvin Lim Gamboa[1,2] **Yue Feng**[1] * **Mark Lee**[1]

[1]School of Computer Science, University of Birmingham

[2]Department of Information Systems and Computer Science, Ateneo de Manila University

llg302@student.bham.ac.uk, lancecalvingamboa@gmail.com
y.feng.6@bham.ac.uk
m.g.lee@bham.ac.uk

Abstract

Pretrained multilingual models exhibit the same social bias as models processing English texts. This systematic review analyzes emerging research that extends bias evaluation and mitigation approaches into multilingual and non-English contexts. We examine these studies with respect to linguistic diversity, cultural awareness, and their choice of evaluation metrics and mitigation techniques. Our survey illuminates gaps in the field's dominant methodological design choices (e.g., preference for certain languages, scarcity of multilingual mitigation experiments) while cataloging common issues encountered and solutions implemented in adapting bias benchmarks across languages and cultures. Drawing from the implications of our findings, we chart directions for future research that can reinforce the multilingual bias literature's inclusivity, cross-cultural appropriateness, and alignment with state-of-the-art NLP advancements.

1 Introduction

Multilingualism has grown to be a core property of recently released pretrained language models (PLMs), such as GPT-4 (OpenAI et al., 2023), Llama 3 (Meta, 2024), and Qwen 2 (Yang et al., 2024). The model release reports of these models all include evaluations on multilingual language understanding benchmarks and demonstrate the models' remarkable performances on these tests. These models' multilingual capabilities have been confirmed by independent assessments done by NLP researchers, such as Zhao et al. (2024) and Huang et al. (2023). Concurrently, there are also emerging endeavors to create models that specialize on handling tasks in multiple languages—e.g., Aya (Üstün et al., 2024) and BLOOMZ (Muennighoff et al., 2023)—or a specific non-English language—e.g., HyperCLOVA X for Korean (Yoo et al., 2024), ChatGLM for Chinese (Team GLM et al., 2024), and Vietcuna for Vietnamese (VILM, 2023).

Multilingual models are not exempt from the safety and bias issues that have been identified in models handling English. Pioneering studies calling attention to the biased behaviors of English models (e.g., Caliskan et al., 2017; Nangia et al., 2020) have been followed by replications demonstrating the presence of similar problems in models processing non-English languages (e.g., Lauscher et al., 2020; Névéol et al., 2022). As such, NLP scholars around the world have progressively expanded efforts to evaluate and ensure the fairness of multilingual and non-English models. This increasing efforts are demonstrated by Figure 1, which depicts the rising number of papers in this niche.

While these multilingual bias studies employ an eclectic selection of approaches, many utilize methods that have been criticized as being error-prone and culturally unaware—e.g., simply relying on automated translations in adapting English bias tests to non-English languages (Talat et al., 2022). These practices are concerning as they may lead not only to the underestimation of culturally specific biases within PLMs but also to a focus on Anglocentric concepts of fairness in the field of bias evaluation

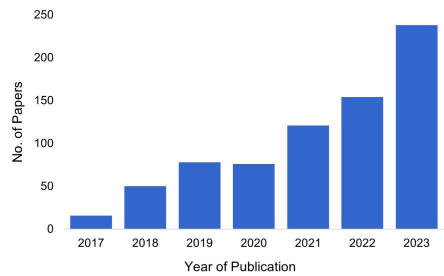

Figure 1: Number of ACL Anthology papers that contain the terms *bias*, *harm*, *stereotype*, *toxic*, or *fair* and *multilingual*, *cross-lingual*, *interlingual*, or a non-English language in the title or abstract.

*Corresponding author

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 27857–27880
November 4-9, 2025 ©2025 Association for Computational Linguistics

and mitigation. There thus stands a need to take stock of the multilingual PLM bias literature, consider the approaches it has been using, and check how successfully such approaches have been making multilingual models more inclusive. While multiple surveys of the general PLM bias scholarship have already been conducted (e.g., Gallegos et al., 2024; Gupta et al., 2024; Goldfarb-Tarrant et al., 2023; Navigli et al., 2023), most just list the investigation of non-English biases as a direction for future research and consequently fail to engage the growing number of studies in this area.

In this paper, we address this gap by systematically and critically reviewing studies on multilingual and non-English PLM bias. To identify these papers, we applied a systematic keyword-based search on ACL Anthology, IEEE Xplore, and the proceedings of the NeurIPS, FAccT, and AIES conferences. From these databases, we shortlisted NLP and language modeling articles that included the following strings in their titles or abstracts: *bias*, *fair*, *toxic*, *harm*, or *stereotyp* plus *multilingual*, *cross-lingual*, *interlingual*, *multiple languages*, or any language name enumerated in ISO 639 or the Codes for the Representation of Names of Languages (Library of Congress, 2017). We then constrained our selection to papers published on or before December 31, 2024. We also filtered out articles that fulfilled the above criteria but did not sustantially engage the concept of sociodemographic bias (e.g., papers about statistical, inductive, and positional bias). This process resulted in a final article set consisting of 106 articles relevant to multilingual PLM bias—97 from ACL Anthology, 7 from IEEE Xplore, 1 from FAccT, and 1 from AIES.

We examine these works using an annotation taxonomy we developed. Our taxonomy builds on extant PLM bias typologies (i.e., Gallegos et al., 2024; Gupta et al., 2024; Goldfarb-Tarrant et al., 2023) and extends these with categories relating to choice of languages and language families, dataset adaptation methods, and cultural awareness in methodological design. By clarifying these aspects among studies we inspected, our taxonomy allows us to explore issues in devising bias evaluation and mitigation protocols for multicultural contexts. Concurrently, we also highlight solutions that have been taken to navigate these concerns.

Our survey exposes the multilingual PLM bias literature's preference for Chinese, Indo-European, and other highly resourced languages. This partiality leads to a shortage of bias research on languages spoken by major sections of the global population and by cultures most active in the adoption of AI technologies. We also discover that more than half of non-English bias tests are accompanied by methodological protocols which do not explicitly document cultural considerations in the benchmark development process. This lack of transparency makes it unclear how (or if) these benchmarks engaged with the adaptation issues confronted by their more culturally aware counterparts—for example, the generalizability of some bias dimensions (e.g., race), the localization of universal biases, and the resolution of differences in linguistic gender across languages. Finally, our review also reveals that multilingual bias research largely stops at the evaluation stage and rarely crosses into the mitigation of biases. This finding highlights the urgency of developing debiasing approaches for multilingual models or, at least, of inspecting the applicability of English debiasing methods on non-English contexts.

Our contributions are threefold:

- We synthesize works on multilingual bias and pinpoint gaps and best practices in the field, thereby revealing and encouraging reflection about trends in the literature.
- Our review sheds light on common challenges encountered by multilingual bias researchers and the steps they have taken to solve these. This catalog of challenges and solutions can guide the design of future work in the field.
- We compile a concise agenda for future multilingual bias research based on issues and limitations we identified in our review.

The rest of this paper is structured as follows: we briefly describe our method for systematic review, particularly how we selected papers and examined them using our taxonomy (2). Next, we outline our findings and their implications, starting with our observations on the linguistic diversity of the PLM bias literature (3) and the cultural awareness of methods used to broaden this diversity (4). We continue with a review of evaluation and mitigation techniques applied on multilingual and non-English models (6). We conclude with a list of opportunities to improve future research in the field (7).

2 Annotation

We follow the conceptualization of social bias utilized by the survey papers of Gupta et al. (2024)

and Gallegos et al. (2024), who define PLM bias as inequalities in how PLMs generate outcomes or perform when handling data and inputs associated with diverse social demographics. We also recognize the distinction between biases leading to representational harms and those leading to allocational harms (Crawford, 2017; Barocas et al., 2017). The former occur when PLMs propagate stereotypes, toxic language, and disparate judgments that depict one social group more unfavorably than another (Blodgett et al., 2020; Crawford, 2017), while the latter emerge from PLMs distributing resources or opportunities unfairly across groups (Blodgett et al., 2020; Barocas et al., 2017). However, because most non-English languages lack labeled pretraining data and therefore have few to no predictive NLP systems that allocate resources (Joshi et al., 2020), the studies we scrutinize only analyze representationally harmful biases.

Given this conceptualization of bias we adopted, we took the bias dataset annotation scheme developed by Goldfarb-Tarrant et al. (2023) as a starting point in developing our own taxonomy. Rooted in an understanding of the potential representation harms of PLMs, their taxonomy notes (1) basic scope attributes about a paper and its accompanying dataset/s (e.g., language/s used, model/s tested, code availability) and (2) aspects about how a paper operationalizes bias evaluation (e.g., bias metric/s, benchmark format, proxies for demographic groups). We then refined the taxonomy with the typologies proposed by Gallegos et al. (2024) and Gupta et al. (2024), whose definitions of bias we also utilize. This led to a revision of the categories used to classify bias metrics and benchmark entries and the addition of a mitigation-related annotation attribute. We also leveraged our familiarity with the field of multilingual PLM bias to augment the taxonomy with elements relating to the originality of the non-English benchmarks, the benchmark development method, and the cultural nuances involved therein. Applying this initial taxonomy on the articles and revising it based on new categories and labels that emanated from the literature resulted in the final taxonomy in Appendix A.

The authors of this paper used this taxonomy to conduct annotations. Disagreements were infrequent and labeling was straightforward. We release[1] a consolidated list of the papers we examined and our annotations for each paper. Figure 2 illustrates a quantitative summary of our annotation and analysis, which we expound upon in the next three sections.

[1]https://github.com/gamboalance/multilingual_bias_survey

3 Language Choice and Diversity

Several of the studies we examined used more than one benchmark; therefore, we looked into a total of 124 bias benchmarks in this survey. Among these benchmarks, we further disaggregated multilingual ones into monolingual sub-benchmarks, resulting in a total of 376 single-language sub-benchmarks analyzed for this paper. All in all, the bias tests we annotated featured 67 different languages (listed in Appendix C), with Chinese taking the top spot in terms of frequency ($n = 30; 7.98\%$), followed by Spanish ($n = 28, 7.45\%$), French ($n = 24, 6.38\%$), German ($n = 24, 6.38\%$), and Arabic ($n = 20, 5.32\%$). As illustrated in Figures 2(a) and 2(b), grouping the languages into their language families and into NLP resource classes reveals that **the literature has an asymmetric focus on Indo-European languages** ($n = 242, 64.36\%$) **and on languages that Joshi et al. (2020) would classify as highly resourced in NLP** ($n_{Class5} = 137, 36.44\%; n_{Class4} = 126, 33.51\%$).

These disproportionate imbalances show that there is a *linguistic bias* in multilingual PLM bias research. Next to English, most PLM bias studies address problematic model behavior mainly in languages spoken by economically developed countries (as determined by GDP per capita data from International Monetary Fund, 2024). While bias studies in these languages are undoubtedly important, they are unable to address the negative repercussions of AI being used enthusiastically in less developed countries, like India, Indonesia, and the Philippines (Sarkar, 2023). Consequently, the statistics above lend empirical support to observations that the communities governing and regulating language model development and moderation are removed from the majority of the communities using these technologies (Talat et al., 2022).

Furthermore, 5 of the 35 most widely spoken languages in the world (Eberhard et al., 2023) have < 10 bias tests written in their languages (e.g., Bengali, Indonesian), while 9 have < 5 tests (e.g., Swahili, Persian) and 5 more are completely absent in the bias literature (e.g., Hausa, Javanese). This pattern in the literature risks values in only a limited number of cultures (i.e., white, Western, or

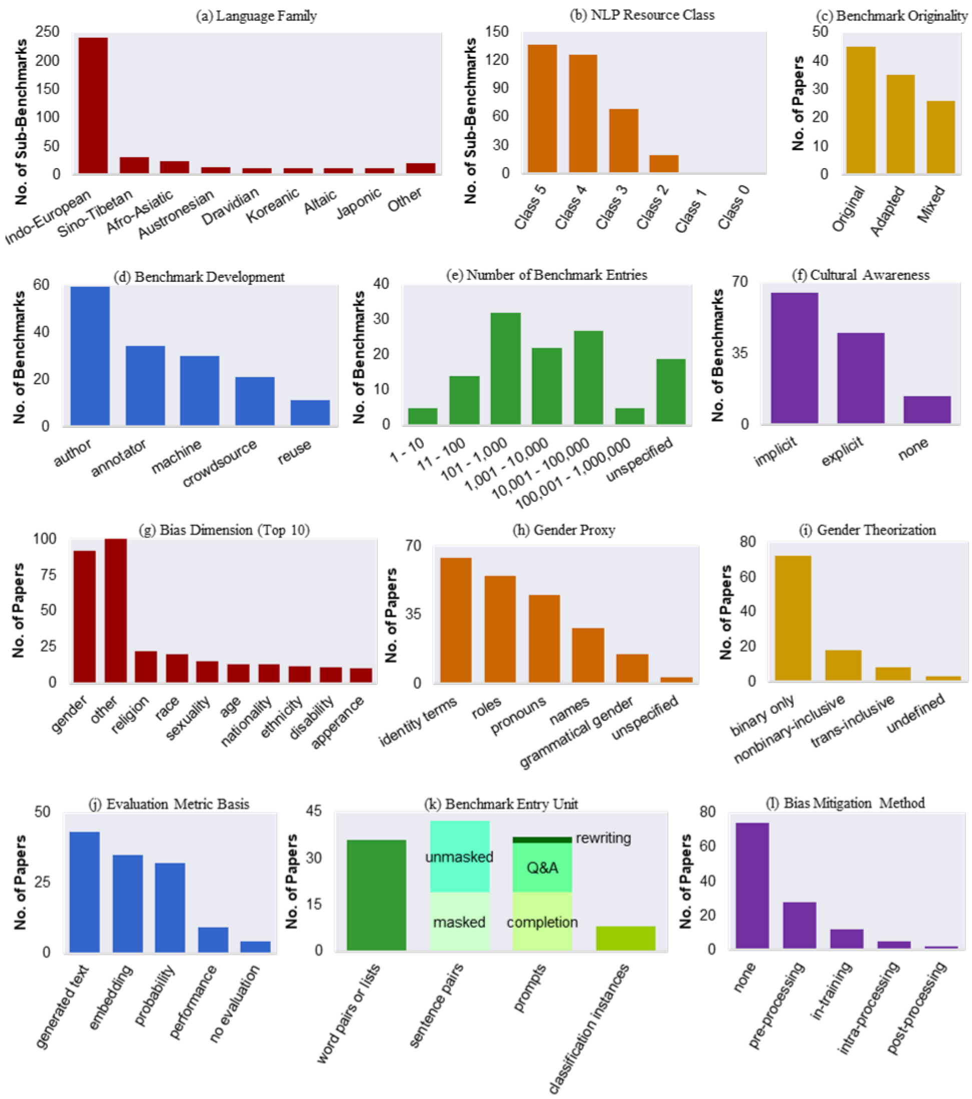

Figure 2: Results for annotating 106 multilingual bias articles using our taxonomy.

Chinese) being represented in endeavors to train and safeguard NLP systems (as previously found by Kreutzer et al., 2022 and Thylstrup and Talat, 2020). Despite conveying an impression of linguistic inclusivity in the PLM bias literature becoming, this *linguistic bias*—if left unchecked—may end up obscuring culturally specific issues in models and exacerbating inequities in what communities and perspectives are considered in the field. We therefore agree with appeals to empower technologically marginalized agents in contributing towards efforts to develop responsible AI (Talat et al., 2022).

4 Cultural Awareness in Benchmark Development

Figure 2(c) shows that the number of works that develop their own benchmark ($n = 45, 42.45\%$) is almost comparable to the number of papers that adapted a pre-existing bias test to their chosen language/s ($n = 35, 33.02\%$). Examples of adapted benchmarks are Multilingual HolisticBias (Costa-jussà et al., 2023) and AraWEAT (Lauscher et al., 2020), which were created through the translation of American-sourced stereotypes found in English HolisticBias (Smith et al., 2022) and English WEAT (Caliskan et al., 2017) respectively. In contrast, original benchmarks include CHBias (Zhao et al., 2023) and KoSBi (Lee et al., 2023b), the authors of which collected novel prejudices relevant to Chinese and Korean societies. A minority of studies employ a mix of original and adapted benchmarks ($n = 26, 24.53\%$)—e.g., the French CrowS-Pairs testbed (Névéol et al., 2022), which is composed of entries translated from the original CrowS-Pairs (Nangia et al., 2020) and new sentences sourced from French contributors.

Authors often take the lead in creating or adapting entries for non-English tests, with about half of the benchmarks ($n = 59, 47.58\%$) having some significant authorial contribution in their development—as demonstrated in Figure 2(d). Such contribution often comes in the form of the authors manually translating English entries (Fort et al., 2024; Gamboa and Estuar, 2023b) or personally constructing culturally appropriate prompts (Wang et al., 2024; Ibaraki et al., 2024). In writing the latter, they relied on stereotypes mined from pre-existing corpora (e.g., Wikidata and CommonCrawl in Naous et al., 2024), mass and social media (Zhu et al., 2024a; Huang and Xiong, 2024), academic articles (Gamboa and Estuar, 2023a), or government documents and statistics (e.g., job market data in Friðriksdóttir and Einarsson, 2024; state-owned name databases in Das et al., 2023).

Author-driven benchmark development, however, has been criticized for lacking diversity of perspective because of authors' limited familiarity with some biases (Goldfarb-Tarrant et al., 2023). To remedy this issue, a number of studies have employed cultural insiders—namely annotators or experts ($n = 34, 27.42\%$) and crowdsource workers ($n = 21, 16.94\%$)—to gather a multiplicity of viewpoints in creating their datasets. These cultural insiders helped in either writing the benchmark entries (e.g., Huang and Xiong, 2024; Touileb and Nozza, 2022) or validating translations and stereotypes provided by the authors, other annotators, and pre-existing benchmarks (e.g., Grigoreva et al., 2024; Mukherjee et al., 2023). For quite a few of the studies though, relying solely on humans was insufficient as they aimed to create benchmarks with entries numbering in the ten- and hundred-thousands—as shown in Figure 2(e). As such, they used NLP models to aid in their evaluation of PLMs ($n = 30, 24.19\%$). Some used translation technologies to adapt benchmarks into another language (e.g., NLLB and Google Translate in Sahoo et al., 2024), others leveraged generative PLMs to write prompts or stereotypes (e.g., Huang and Xiong, 2024), and several used algorithms to automatically populate templates and create a large number of test items (e.g., Jin et al., 2024).

Unfortunately, **no matter the benchmark development method, most works do not explicitly document the cultural nuances considered in creating bias tests**. As seen in Figure 2(f), about a tenth of the benchmarks examined ($n = 14, 11.29\%$) were created with no attention paid to pertinent cultural considerations; meanwhile, more than half ($n = 65, 52.42\%$) only implied a semblance of cultural awareness through the involvement of cultural insiders (e.g., annotators, experts, crowdsource workers) but did not record what adaptation issues and processes these participants engaged with.

This inclination to overlook aspects of benchmarks development linked to cultural interpretation and calibration is alarming, especially given the number of non-English datasets containing adapted elements ($n = 61, 57.55\%$). Nebulous or nonexistent descriptions of cultural concerns encountered in multilingual bias test construction make it impossible to assess how appropriate test items are in capturing biases in a particular culture. After all, biases and stereotypes are culture-dependent, and what might be a discriminatory statement in one societal context may be less significant in another (Gallegos et al., 2024; Talat et al., 2022). The lack of clarity among many multilingual bias studies on how they handle these cultural idiosyncrasies casts doubt on the validity of results and conclusions drawn from their culturally naïve benchmarks. There is thus a need to rectify this inadequacy in transparent cultural awareness in the PLM bias literature.

Fortunately, there is a non-negligible amount of research in the field ($n = 45, 36.29\%$) that is forthright in its cultural awareness and that can serve as basis for improving cultural transparency. We reviewed these works and identified common adaptation issues encountered by multilingual bias scholars and the solutions they implemented. We discuss these in the succeeding section and provide a quick summary at Table 1.

5 Issues in Adapting Bias Benchmarks

5.1 Bias Dimensions and Their Cultural Relevance

Across the studies we annotated, we identified 24 social dimensions (Table 5 in Appendix D) along which multilingual benchmarks measured bias. Among these attributes, a few consistently appeared in English benchmarks but were deemed by non-American scholars to be irrelevant to their contexts. Racism, in particular, was deemed to be an issue more central to primarily English-speaking countries and was therefore merged with ethnicity- and nationality-based bias in studies conducted in Sweden (Devinney et al., 2024) and South Korea

Paper/s	Adaptation Issue	Adaptation Practice
	General Adaptation Practices	
Fort et al. (2024); Gamboa and Estuar (2023b)		Authors manually translate entries from English benchmarks into other languages.
Wang et al. (2024); Ibaraki et al. (2024)		Authors construct entirely new prompts appropriate to their target culture, relying on the following to determine contextually relevant stereotypes:
Naous et al. (2024)		• Wikidata and CommonCrawl
Zhu et al. (2024a); Huang and Xiong (2024)		• mass and social media
Gamboa and Estuar (2023a)		• academic articles
Friðriksdóttir and Einarsson (2024); Das et al. (2023)		• government documents and statistics
Huang and Xiong (2024); Touileb and Nozza (2022)		Recruit cultural insiders (e.g., annotators, experts, or crowdsource workers) to write benchmark entries.
Grigoreva et al. (2024); Mukherjee et al. (2023)		Recruit cultural insiders to validate translations and stereotypes from the authors, other annotators, or existing benchmarks.
Sahoo et al. (2024)		Use machine translators to adapt benchmarks into another language.
Huang and Xiong (2024)		Use generative models to write new prompts or find culturally appropriate stereotypes.
	Adaptation Practices Pertaining to Bias Dimensions	
Devinney et al. (2024); Jin et al. (2024)	Some bias dimensions in English benchmarks (e.g., racism) are irrelevant to non-American or non-Western contexts.	Streamline race-, ethnicity-, and nationality-related biases into one bias dimension.
Sahoo et al. (2024); Bhatt et al. (2022); Malik et al. (2022); Huang and Xiong (2024); Lee et al. (2023b)	Bias dimensions relevant to specific cultures are absent in English benchmarks.	Add culturally specific bias dimensions (e.g., those related to caste, disease, and family structure) into the adapted benchmark.
	Adaptation Practices Pertaining to Contextualizing Universal Biases	
Jin et al. (2024); Névéol et al. (2022)	Some terms in English benchmarks are specfic to Western or American culture (e.g., *rugby* and *star quarterback* referencing sports prominent to the USA).	Replace these terms with contextually compatible equivalents (e.g., using *basketball* instead of *rugby* in cultures where the latter is not popular). If there are no equivalents, remove the entry containing the culturally irrelevant term.
Marinova et al. (2023); Hada et al. (2024)		Design completely novel evaluation frameworks and benchmarks based on how biases manifest locally.
	Adaptation Practices Pertaining to Gender and Sexuality	
Steinborn et al. (2022); Sahoo et al. (2024)	Benchmarks relying on counterfactual inputs are difficult to adapt into languages with gender-neutral pronouns and vocabularies (e.g., Finnish).	Translate pronouns into gendered names (e.g., John, Mary) or identity terms (e.g., man, woman).
Névéol et al. (2022)	In heavily gendered languages (e.g., French), gender inflections transform counterfactual pairs into practically different sentences.	Paraphrase the prompts to reduce the need for gender inflections while preserving meaning.
Chávez Mulsa and Spanakis (2020); Grigoreva et al. (2024); Matthews et al. (2021)	Some gendered words carry multiple meanings, making it impossible to disentangle whether bias effects arise from their inherent gender or from their other meanings.	Eliminate or find substitutes for these words. Alternatively, retain these words but warn readers and benchmark users of their potential impact on evaluation results.

Table 1: Benchmark adaptation practices utilized by multilingual bias researchers.

(Jin et al., 2024).

Other benchmark developers note that dimensions often included in English benchmarks do not capture the totality of biases present in their home societies. Caste-related stereotypes, for example, are highly salient in Indian society but are never featured in English bias datasets (Sahoo et al., 2024; Bhatt et al., 2022; Malik et al., 2022). Other culturally specific dimensions include disease, household registration (relevant to Chinese culture according to Huang and Xiong, 2024), pregnancy, family structure, and marital status (relevant to Korean Culture according to Lee et al., 2023b). In future, authors working on multilingual bias evaluation should therefore reflect on how suitable their benchmarks' dimensions are to the culture of the language they are working with. Furthermore, more efforts should be invested on non-gender-related biases. Sexism is the subject of the vast majority of multilingual bias studies ($n = 92, 86.79\%$), as shown in Figure 2(g), and leaves many other types of biases, including intersectional ones, underexplored.

5.2 Contextualizing Universal Biases

While some overarching categories of bias cut across cross-cultural boundaries (e.g., gender), their manifestations vary in different localities. A difficulty constantly raised by the examined works is the appearance of culturally specific terms and stereotypes in English benchmarks. For example, although stereotypes between physical activity and gender are prominent worldwide, the way these are expressed in English tests through terms related to American sports culture (e.g., *rugby*, *star quarterback*) may not be apt in some cultures (Sahoo et al., 2024; Jin et al., 2024). In adapting entries containing such terms, authors and annotators used their knowledge of their culture to pick a contextually compatible equivalent—e.g., replacing *rugby* with *basketball* in a Korean dataset (Jin et al., 2024).

In some cases, this practice of localization was impossible because a concept or a stereotype itself did not exist in the target culture, compelling developers to just discard these inputs from the adapted benchmark. To demonstrate: stereotypes linking queerness to the color pink and to culinary ability were deemed untranslatable to the French culture and removed from French CrowS-Pairs (Névéol et al., 2022).

Some have gone beyond mere entry rewriting or removal and have intentionally designed their evaluation frameworks with local manifestations of universal biases in mind. For example, Marinova et al. (2023) drew from their knowledge of the peculiar ethnic composition of Bulgaria's population to write original prompts assessing masked models' biases regarding these minorities. Meanwhile, Hada et al. (2024) organized community workshops with women in rural India to collect sentences illustrating Hindi concepts of gender bias. While the first approach yielded findings about how differently language models treated particular Bulgarian ethnic groups, the latter unveiled how Hindi-speaking communities associated the male with curiosity and the female with reservedness.

The insights and issues surfaced by these contextualization methods affirm the value of culturally aware and transparent benchmark adaptation techniques. **Without sufficient documentation on the cultural specificities of the adaptation process, it cannot be known if local stereotypes were incorporated in the bias test or if, at the very least, contextually trivial biases were addressed**.

5.3 Linguistic Gender and Non-binary Representation

Figure 2(h) illustrates that multilingual bias studies use different proxies to denote gender in their benchmark entries. The use of some of these proxies, however, comes with challenges in multilingual research because of differences in how linguistic gender is expressed across different languages.

One such challenge is the gender neutrality of some languages leading to the homogenization of counterfactual inputs that many benchmarks rely on. Steinborn et al. (2022), for example, needed to deal with the Finnish language having only the genderless third person pronoun *hän*. Such non-gendered-ness transformed originally different benchmark entry pairs (e.g., ***He*** *was timid.* / ***She*** *was timid.*) into two identical sentences (e.g., ***Hän*** *oli arka.* / ***Hän*** *oli arka.* in Finnish), rendering it impossible for bias metrics to compare how differently a PLM would behave with respect to each gender. As a solution, pronouns in the original benchmarks were changed to either identity terms (Sahoo et al., 2024) or gendered names (Steinborn et al., 2022). In some cases, entries involving gender-neutral pronouns were removed from the adapted benchmark (e.g., Ousidhoum et al., 2021; Matthews et al., 2021).

In contrast, researchers working with heavily gendered languages faced the opposite conundrum: gender inflections often mutated minimally different inputs into almost completely different sentences. Névéol et al. (2022) presents the case of the English pair ***Women*** *talk a lot.* / ***Men*** *talk a lot.* plausibly corresponding to *Les* ***femmes*** *sont* ***bavardes***. / Les **hommes** sont **bavards**. in French. These translations are problematic because half of the pair's tokens are different from each other and will make the calculation of the bias metric unsound. The authors resorted to creative paraphrasing to circumvent the complication. For example, the above was translated into *Les* ***femmes*** *parlent à tort et à travers.* / *Les* ***hommes*** *parlent à tort et à travers.* which roughly translate to ***Men***

Evauation Metric Category	Benchmark Entry Units	Sample Papers
Embedding-based Metrics		
word embedding metrics	word lists	Wambsganss et al. (2023); Hansal et al. (2022)
sentence embedding metrics	word lists	Sahoo et al. (2024); Malik et al. (2022)
Probability-based Metrics		
masked token methods	counterfactual inputs (masked)	Vashishtha et al. (2023); Guo et al. (2022)
pseudo-log-likelihood methods	counterfactual inputs (unmasked)	Pikuliak et al. (2023); Kaneko et al. (2022)
Generated Text-based Metrics		
distribution metrics	prompts	Li et al. (2024) (sentence completion); Truong et al. (2024) (QA)
classifier metrics	prompts	Brun and Nikoulina (2024) (sentence completion); Mihaylov and Shtedritski (2024) (QA)
lexicon metrics	prompts	Martinková et al. (2023) (sentence completion); Touileb and Nozza (2022) (sentence completion)
Performance-based Metrics		
classification scores	classification instances	Conti and Wisniewski (2023); Huang (2022)

Table 2: Sample papers for each category of evaluation metrics and benchmark entry units used by multilingual bias studies, as grouped using typologies from Gallegos et al. (2024) and Gupta et al. (2024).

/ ***women*** *talk all over the place.*—preserving both the meaning and the minimal difference of the original English pair.

A third issue was the duality of genders and meanings that a gendered word sometimes encoded in a language. In Icelandic, grammatically masculine nouns are generally used to refer to both male and female individuals despite feminine alternatives being sometimes present—for example, the masculine *hjúkrunarfræðingur* is used to refer to Icelandic male and female nurses in spite of the feminine *hjúkrunarkona* being available (Steinborn et al., 2022; Grigoreva et al., 2024). These complexities make it hard to disentangle whether the biased model behaviors induced by these dually encoding words are linked to their inherent grammatical gender or to the multiple meanings they refer to in reality. As a result, one study eliminated the use of these words altogether and looked for reasonable substitutes instead (Chávez Mulsa and Spanakis, 2020). Others retained them but forewarned of their possible impact on evaluation results (Grigoreva et al., 2024; Matthews et al., 2021).

We end this subsection with our observation (Figure 2i) that most multilingual bias studies opt for gender proxies which represent only the male-female binary ($n = 72, 67.92\%$) and fail to consider non-binary ($n = 18, 16.98\%$) and transgender ($n = 8, , 7.55\%$) identities. This propensity to ignore queerness is dangerous since it precludes work that can quantify and mitigate the harms PLMs can bring on non-heterosexual groups (Goldfarb-Tarrant et al., 2023). Given the contextually unique struggles of non-binary groups across different cultures (Hinchy, 2019; McMullin, 2011; Garcia, 1996), **we call on multilingual bias scholars to be more conscious not only in navigating linguistic gender features peculiar to their languages but also in actively pondering how they can incorporate the perspectives of queer communities in their cultures**.

6 Evaluation and Mitigation Methods

Using the taxonomy of bias evaluation metrics proposed by Gallegos et al. (2024) and Gupta et al. (2024), we found a relatively even mix of multilingual studies (Figure 2j, Table 2) that measure bias in generated texts ($n = 43, 40.57\%$), quantify bias based on comparing token probabilities ($n = 35, 33.02\%$), and calculate bias using internal embedding vectors ($n = 32, 30.19\%$). This balance is mirrored in the studies' benchmark formats of choice (Figure 2k): 30.08% ($n = 37$) use prompts often partnered with generated text-based metrics, 34.15% ($n = 42$) work with counterfactual sentence pairs frequently inputted into probability-based metric frameworks, and 29.27% ($n = 36$) involve word lists required for embedding-based metrics. We argue that this equilibrium in the kinds of bias evaluation approaches utilized in multilingual bias literature conceals a gap in the research area. **The fact that methods developed for word2vec and other static embeddings still constitute a significant proportion (about one-third) of multilingual bias research hints that the field has not yet fully caught up with Transformer-driven ad-**

vancements in NLP.[2] Although the equally large number of studies operating on probability- and generated text-based metrics demonstrates progress and promise, more concerted efforts are still needed in ensuring the fairness and safety of the latest multilingual technologies deployed in non-English-speaking cultures.

Equally notable in Figures 2j and 2k is the small number of studies employing performance-based metrics and benchmarks composed of classification instances ($n = 8$). We expected this outcome because, as mentioned above, there are only a limited number of non-English labeled datasets that can be used in this respect (Joshi et al., 2020). Thus, **there is also a need to devise benchmarks and studies that measure bias on downstream tasks in multilingual PLMs**. This direction of inquiry is critical, especially in light of some research suggesting the weak correlation between downstream model behavior and the probability- and embedding-based metrics currently dominating multilingual bias research (Cabello et al., 2023; Delobelle et al., 2022).

This dearth in downstream multilingual bias research is matched by a scarcity of multilingual bias mitigation research as well (Figure 2l, Table 3). The overwhelming majority of the papers we looked into do not undertake bias mitigation experiments at all ($n = 74, 69.81\%$). One possible reason for this is that many debiasing methods used for English models (e.g., data augmentation, instruction tuning, contrastive learning, adversarial learning) require readily available English datasets to balance biased pretraining data, to fine-tune existing models for fairness, or to modify their internal architectures (e.g., Zheng et al., 2023; Zayed et al., 2023; Narayanan Venkit et al., 2023). **Such debiasing datasets are not easily accessible in non-English languages, making multilingual bias mitigation research scarce**.

Among multilingual bias studies that do perform mitigation, pre-processing mitigation techniques are the most frequent ($n = 28, 26.42\%$), with projection-based mitigation approaches ($n = 11$) being the most widely used under this category. Projection-based mitigation identifies an embedding model subspace corresponding to a bias dimension (e.g., gender) and nullifies this subspace to minimize model bias (Gallegos et al., 2024). The

[2] 77.14% of these embedding-based studies were conducted from 2020 onwards, indicating that they continue to be dominant despite the emergence and rapid development of multilingual Transformer-based models during this time.

Mitigation Stage	Sample Papers
pre-processing	Üstün et al. (2024); Ahn and Oh (2021)
in-training	Aakanksha et al. (2024); Ramesh et al. (2023a)
intra-processing	Ermis et al. (2024); Lee et al. (2023a)
post-processing	Jain et al. (2022)

Table 3: Sample studies that mitigate bias in multilingual models, as categorized by the stage in the language modeling pipeline at which they intervene.

prevalence of such a technique again signifies that much of the multilingual bias research still centers on an embedding-based language modeling framework. Consequently, we also deem as urgent the matter of updating and expanding endeavors to mitigate bias in multilingual PLMs.

7 Conclusion and Future Directions

In this paper, we sought to elucidate patterns and practices in the multilingual bias literature and to gauge their effectiveness in broadening the cultural inclusivity of PLM bias research. Our analysis uncovered opportunities for future research that can further accelerate the field's growth. These opportunities include (but are not limited to):

- evaluating social bias beyond cultures with high-resource and Indo-European languages to address *linguistic bias* in multilingual PLM bias research ,
- employing culturally aware benchmark development methodologies that explicitly document cultural complexities,
- designing benchmarks that incorporate culturally specific bias dimensions and stereotypes collected from contextualized perspectives,
- expanding research grounded on the heteronormative binary to include diverse expressions of queerness across cultures,
- pushing past embedding-based methods and reinforcing bias research on state-of-the-art multilingual models, especially those used in downstream tasks, and
- debiasing multilingual models.

We hope that researchers and practitioners working on multilingual bias can use our work to guide their own efforts to address bias in non-English contexts. We also hope that through our survey, they can leverage the myriad approaches which scholars around the world have taken to make PLMs safer and fairer for communities all over the globe.

Limitations

Our work is subject to some limitations. First, a few of the papers we annotated were written in a non-English language. Specifically, Guo et al. (2022) was written in Chinese while Benamar et al. (2022) was written in French. To allow us to include these in our survey, we used machine translators to translate the papers into English. This approach might have influenced the way we understood and annotated the papers. To minimize the impact of translation inaccuracies, we cross-referenced translations across different tools to confirm correctness. Furthermore, one of the authors has native proficiency in Mandarin while another has conversational proficiency; thus, they were capable of checking the translations for the Chinese article.

Second, while the categories and values we use in our taxonomy are based on previous PLM bias surveys, it is still possible that these do not encompass all extant or incoming research in the field.

Third, our search strategy did not include notable machine learning and artificial intelligence conferences (e.g., ICLR, ICML), nor did it consider non-English databases. However, it is interesting to note that most articles fulfilling our search criterion only come from ACL conferences. Despite including non-ACL venues (NeurIPS, FAccT, and AIES), only 2 papers from these conferences satisfied our criteria (1 from FAccT, 1 from AIES). This may suggest (although not conclusively) that multilingual bias studies rarely feature in non-ACL events.

Finally, we focus on only the evaluation and mitigation aspects of the bias literature and do not examine research strands in the field that are only just emerging, such as explainability (e.g., Liu et al., 2024; Conti and Wisniewski, 2023) and interpretability (e.g., Gamboa et al., 2025; Gamboa and Lee, 2024).

We also acknowledge the potential psychosocial risks of compiling bias tests and benchmarks with possibly offensive entries into one location (i.e., the article annotation repository we share). However, we feel that the benefits of making these resources easily accessible (e.g., advancing multilingual bias research) outweighs such risks.

Acknowledgments

Lance Gamboa would like to thank the Philippine government's Department of Science and Technology for funding his doctorate studies.

References

Aakanksha, Arash Ahmadian, Beyza Ermis, Seraphina Goldfarb-Tarrant, Julia Kreutzer, Marzieh Fadaee, and Sara Hooker. 2024. The multilingual alignment prism: Aligning global and local preferences to reduce harm. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12027–12049, Miami, Florida, USA. Association for Computational Linguistics.

Jaimeen Ahn and Alice Oh. 2021. Mitigating language-dependent ethnic bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 533–549, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Adnan Al Ali and Jindřich Libovický. 2024. How gender interacts with political values: A case study on Czech BERT models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8200–8210, Torino, Italia. ELRA and ICCL.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, Julien Launay, and Badreddine Noune. 2023. AlGhafa evaluation benchmark for Arabic language models. In *Proceedings of ArabicNLP 2023*, pages 244–275, Singapore (Hybrid). Association for Computational Linguistics.

Senthil Kumar B, Pranav Tiwari, Aman Chandra Kumar, and Aravindan Chandrabose. 2022. Casteism in India, but not racism - a study of bias in word embeddings of Indian languages. In *Proceedings of the First Workshop on Language Technology and Resources for a Fair, Inclusive, and Safe Society within the 13th Language Resources and Evaluation Conference*, pages 1–7, Marseille, France. European Language Resources Association.

Ehsan Barkhordar, Surendrabikram Thapa, Ashwarya Maratha, and Usman Naseem. 2024. Why the unexpected? dissecting the political and economic bias in Persian small and large language models. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, pages 410–420, Torino, Italia. ELRA and ICCL.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: From allocative to representational harms in machine learning. In *Information and Society (SIGCIS)*.

Alexandra Benamar, Cyril Grouin, Meryl Bothua, and Anne Vilnat. 2022. Etude des stéréotypes genrés dans le théâtre français du XVIe au XIXe siècle à travers des plongements lexicaux (studying gender stereotypes in French theater from XVIth to XIXth century through the use of lexical embeddings). In

Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 74–81, Avignon, France. ATALA.

Selma Bergstrand and Björn Gambäck. 2024. Detecting and mitigating LGBTQIA+ bias in large Norwegian language models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 351–364, Bangkok, Thailand. Association for Computational Linguistics.

Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. 2022. Re-contextualizing fairness in NLP: The case of India. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 727–740, Online only. Association for Computational Linguistics.

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

El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, AbdelRahim Elmadany, Alcides Inciarte, and Md Tawkat Islam Khondaker. 2023. JASMINE: Arabic GPT models for few-shot learning. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16721–16744, Singapore. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Caroline Brun and Vassilina Nikoulina. 2024. FrenchToxicityPrompts: a large benchmark for evaluating and mitigating toxicity in French texts. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying @ LREC-COLING-2024*, pages 105–114, Torino, Italia. ELRA and ICCL.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. 2023. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 370–378, New York, NY, USA. Association for Computing Machinery.

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

António Câmara, Nina Taneja, Tamjeed Azad, Emily Allaway, and Richard Zemel. 2022. Mapping the multilingual margins: Intersectional biases of sentiment analysis systems in English, Spanish, and Arabic. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 90–106, Dublin, Ireland. Association for Computational Linguistics.

Rodrigo Alejandro Chávez Mulsa and Gerasimos Spanakis. 2020. Evaluating bias in Dutch word embeddings. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 56–71, Barcelona, Spain (Online). Association for Computational Linguistics.

Lina Conti and Guillaume Wisniewski. 2023. Using artificial French data to understand the emergence of gender bias in transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10362–10371, Singapore. Association for Computational Linguistics.

Marta Costa-jussà, Pierre Andrews, Eric Smith, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Daniel Licht, and Carleigh Wood. 2023. Multilingual holistic bias: Extending descriptors and patterns to unveil demographic biases in languages at scale. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14141–14156, Singapore. Association for Computational Linguistics.

Kate Crawford. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems, invited speaker*.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Dipto Das, Shion Guha, and Bryan Semaan. 2023. Toward cultural bias evaluation datasets: The case of Bengali gender, religious, and national identity. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 68–83, Dubrovnik, Croatia. Association for Computational Linguistics.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

Anastasiia Demidova, Hanin Atwany, Nour Rabih, Sanad Sha'ban, and Muhammad Abdul-Mageed. 2024. John vs. ahmed: Debate-induced bias in multilingual LLMs. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 193–209, Bangkok, Thailand. Association for Computational Linguistics.

Jiawen Deng, Jingyan Zhou, Hao Sun, Chujie Zheng, Fei Mi, Helen Meng, and Minlie Huang. 2022. COLD: A benchmark for Chinese offensive language detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11580–11599, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hannah Devinney, Jenny Björklund, and Henrik Björklund. 2024. We don't talk about that: Case studies on intersectional analysis of social bias in large language models. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 33–44, Bangkok, Thailand. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig. 2023. *Ethnologue: Languages of the World*, 26 edition. SIL International, Dallas.

Beyza Ermis, Luiza Pozzobon, Sara Hooker, and Patrick Lewis. 2024. From one to many: Expanding the scope of toxicity mitigation in language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15041–15058, Bangkok, Thailand. Association for Computational Linguistics.

Karen Fort, Laura Alonso Alemany, Luciana Benotti, Julien Bezançon, Claudia Borg, Marthese Borg, Yongjian Chen, Fanny Ducel, Yoann Dupont, Guido Ivetta, Zhijian Li, Margot Mieskes, Marco Naguib, Yuyan Qian, Matteo Radaelli, Wolfgang S. Schmeisser-Nieto, Emma Raimundo Schulz, Thiziri Saci, Sarah Saidi, Javier Torroba Marchante, Shilin Xie, Sergio E. Zanotto, and Aurélie Névéol. 2024. Your stereotypical mileage may vary: Practical challenges of evaluating biases in multiple languages and cultural contexts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17764–17769, Torino, Italia. ELRA and ICCL.

Steinunn Rut Friðriksdóttir and Hafsteinn Einarsson. 2024. Gendered grammar or ingrained bias? exploring gender bias in Icelandic language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7596–7610, Torino, Italia. ELRA and ICCL.

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

Lance Calvin Gamboa and Maria Regina Justina Estuar. 2023a. Characterizing bias in word embeddings towards analyzing gender associations in Philippine texts. In *2023 IEEE World Conference on Applied Intelligence and Computing (AIC)*, pages 254–259.

Lance Calvin Gamboa and Maria Regina Justina Estuar. 2023b. Evaluating gender bias in pre-trained Filipino FastText embeddings. In *2023 International Conference on IT Innovation and Knowledge Discovery (ITIKD)*, pages 1–7.

Lance Calvin Lim Gamboa, Yue Feng, and Mark G. Lee. 2025. Bias attribution in Filipino language models: Extending a bias interpretability metric for application on agglutinative languages. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 195–205, Vienna, Austria. Association for Computational Linguistics.

Lance Calvin Lim Gamboa and Mark Lee. 2024. A novel interpretability metric for explaining bias in language models: Applications on multilingual models from Southeast Asia. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 296–305, Tokyo, Japan. Tokyo University of Foreign Studies.

Lance Calvin Lim Gamboa and Mark Lee. 2025. Filipino benchmarks for measuring sexist and homophobic bias in multilingual language models from Southeast Asia. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 123–134, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

J. Neil C. Garcia. 1996. *Philippine Gay Culture: Binabae to Bakla, Silahis to MSM*. Hong Kong University Press.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This prompt is measuring <MASK>: Evaluating bias evaluation in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, Toronto, Canada. Association for Computational Linguistics.

Veronika Grigoreva, Anastasiia Ivanova, Ilseyar Alimova, and Ekaterina Artemova. 2024. RuBia: A Russian language bias detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14227–14239, Torino, Italia. ELRA and ICCL.

Mengqing Guo, Jiali Li, Jishun Zhao, Shucheng Zhu, Ying Liu, and Pengyuan Liu. 2022. Measurement of occupational gender bias in Chinese natural language processing tasks. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*, pages 510–522, Nanchang, China. Chinese Information Processing Society of China.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024. Sociodemographic bias in language models: A survey and forward path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322, Bangkok, Thailand. Association for Computational Linguistics.

Rishav Hada, Safiya Husain, Varun Gumma, Harshita Diddee, Aditya Yadavalli, Agrima Seth, Nidhi Kulkarni, Ujwal Gadiraju, Aditya Vashistha, Vivek Seshadri, and Kalika Bali. 2024. Akal badi ya bias: An exploratory study of gender bias in hindi language technology. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1926–1939, New York, NY, USA. Association for Computing Machinery.

Oussama Hansal, Ngoc Tan Le, and Fatiha Sadat. 2022. Indigenous language revitalization and the dilemma of gender bias. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 244–254, Seattle, Washington. Association for Computational Linguistics.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Jessica Hinchy. 2019. *Governing Gender and Sexuality in Colonial India: The Hijra, c.1850–1900*. Cambridge University Press.

Hsin-Yi Hsieh, Shih-Cheng Huang, and Richard Tzong-Han Tsai. 2024. TWBias: A benchmark for assessing social bias in traditional Chinese large language models through a Taiwan cultural lens. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8688–8704, Miami, Florida, USA. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Kexin Huang, Xiangyang Liu, Qianyu Guo, Tianxiang Sun, Jiawei Sun, Yaru Wang, Zeyang Zhou, Yixu Wang, Yan Teng, Xipeng Qiu, Yingchun Wang, and Dahua Lin. 2024. Flames: Benchmarking value alignment of LLMs in Chinese. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4551–4591, Mexico City, Mexico. Association for Computational Linguistics.

Xiaolei Huang. 2022. Easy adaptation to mitigate gender bias in multilingual text classification. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 717–723, Seattle, United States. Association for Computational Linguistics.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2917–2929, Torino, Italia. ELRA and ICCL.

Katsumi Ibaraki, Winston Wu, Lu Wang, and Rada Mihalcea. 2024. Analyzing occupational distribution representation in Japanese language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 959–973, Torino, Italia. ELRA and ICCL.

International Monetary Fund. 2024. World economic outlook database.

Nishtha Jain, Declan Groves, Lucia Specia, and Maja Popović. 2022. Leveraging pre-trained language models for gender debiasing. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2188–2195, Marseille, France. European Language Resources Association.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Anne Lauscher, Rafik Takieddin, Simone Paolo Ponzetto, and Goran Glavaš. 2020. AraWEAT: Multidimensional analysis of biases in Arabic word embeddings. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 192–199, Barcelona, Spain (Online). Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Meeyoung Cha, Yejin Choi, Byoungpil Kim, Gunhee Kim, Eun-Ju Lee, Yong Lim, Alice Oh, Sangchul Park, and Jung-Woo Ha. 2023a. SQuARe: A large-scale dataset of sensitive questions and acceptable responses created through human-machine collaboration. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6692–6712, Toronto, Canada. Association for Computational Linguistics.

Hwaran Lee, Seokhee Hong, Joonsuk Park, Takyoung Kim, Gunhee Kim, and Jung-woo Ha. 2023b. KoSBI: A dataset for mitigating social bias risks towards safer large language model applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages 208–224, Toronto, Canada. Association for Computational Linguistics.

Seungyoon Lee, Dong Kim, Dahyun Jung, Chanjun Park, and Heuiseok Lim. 2024. Exploring inherent biases in LLMs within Korean social context: A comparative analysis of ChatGPT and GPT-4. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 93–104, Mexico City, Mexico. Association for Computational Linguistics.

Sharon Levy, Neha John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. 2023. Comparing biases and the impact of multilingual training across multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10260–10280, Singapore. Association for Computational Linguistics.

Xiaochen Li, Zheng Xin Yong, and Stephen Bach. 2024. Preference tuning for toxicity mitigation generalizes across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13422–13440, Miami, Florida, USA. Association for Computational Linguistics.

Library of Congress. 2017. Codes for the representation of names of languages.

Yan Liu, Yu Liu, Xiaokang Chen, Pin-Yu Chen, Daoguang Zan, Min-Yen Kan, and Tsung-Yi Ho. 2024. The devil is in the neurons: Interpreting and mitigating social biases in language models. In *The Twelfth International Conference on Learning Representations*.

Vijit Malik, Sunipa Dev, Akihiro Nishi, Nanyun Peng, and Kai-Wei Chang. 2022. Socially aware bias measurements for Hindi language representations. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1041–1052, Seattle, United States. Association for Computational Linguistics.

Iva Marinova, Kiril Simov, and Petya Osenova. 2023. Transformer-based language models for Bulgarian. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 712–720, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Sandra Martinková, Karolina Stanczak, and Isabelle Augenstein. 2023. Measuring gender bias in West Slavic language models. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 146–154, Dubrovnik, Croatia. Association for Computational Linguistics.

Abigail Matthews, Isabella Grasso, Christopher Mahoney, Yan Chen, Esma Wali, Thomas Middleton, Mariama Njie, and Jeanna Matthews. 2021. Gender bias in natural language processing across human languages. In *Proceedings of the First Workshop on Trustworthy Natural Language Processing*, pages 45–54, Online. Association for Computational Linguistics.

Dan Taulapapa McMullin. 2011. Fa'afafine notes: On tagaloa, jesus, and nafanua. *Amerasia Journal*, 37(3):114–131.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date.

Viktor Mihaylov and Aleksandar Shtedritski. 2024. What an elegant bridge: Multilingual LLMs are biased similarly in different languages. In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 16–23, Miami, FL, USA. Association for Computational Linguistics.

Mario Mina, Júlia Falcão, and Aitor Gonzalez-Agirre. 2024. Exploring the relationship between intrinsic stigma in masked language models and training data using the stereotype content model. In *Proceedings of the Fifth Workshop on Resources and ProcessIng of linguistic, para-linguistic and extra-linguistic Data from people with various forms of cognitive/psychiatric/developmental impairments @LREC-COLING 2024*, pages 54–67, Torino, Italia. ELRA and ICCL.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Anjishnu Mukherjee, Chahat Raj, Ziwei Zhu, and Antonios Anastasopoulos. 2023. Global Voices, local biases: Socio-cultural prejudices across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15828–15845, Singapore. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Pranav Narayanan Venkit, Sanjana Gautam, Ruchi Panchanadikar, Ting-Hao Huang, and Shomir Wilson. 2023. Nationality bias in text generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 116–122, Dubrovnik, Croatia. Association for Computational Linguistics.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory, and discussion. *J. Data and Information Quality*, 15(2).

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531, Dublin, Ireland. Association for Computational Linguistics.

Shangrui Nie, Michael Fromm, Charles Welch, Rebekka Görge, Akbar Karimi, Joan Plepi, Nazia Mowmita, Nicolas Flores-Herr, Mehdi Ali, and Lucie Flek. 2024. Do multilingual large language models mitigate stereotype bias? In *Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP*, pages 65–83, Bangkok, Thailand. Association for Computational Linguistics.

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

Nedjma Ousidhoum, Xinran Zhao, Tianqing Fang, Yangqiu Song, and Dit-Yan Yeung. 2021. Probing toxic content in large pre-trained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4262–4274, Online. Association for Computational Linguistics.

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

Matúš Pikuliak, Ivana Beňová, and Viktor Bachratý. 2023. In-depth look at word filling societal bias measures. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3648–3665, Dubrovnik, Croatia. Association for Computational Linguistics.

Krithika Ramesh, Arnav Chavan, Shrey Pandit, and Sunayana Sitaram. 2023a. A comparative study on the impact of model compression techniques on fairness in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15762–15782, Toronto, Canada. Association for Computational Linguistics.

Krithika Ramesh, Sunayana Sitaram, and Monojit Choudhury. 2023b. Fairness in language models beyond English: Gaps and challenges. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2106–2119, Dubrovnik, Croatia. Association for Computational Linguistics.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806, Mexico City, Mexico. Association for Computational Linguistics.

Sujan Sarkar. 2023. AI industry analysis: 50 most visited AI tools and their 24B+ traffic behavior.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. "I'm sorry to hear that": Finding new biases in language models with a holistic descriptor dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Victor Steinborn, Philipp Dufter, Haris Jabbar, and Hinrich Schuetze. 2022. An information-theoretic approach and dataset for probing gender stereotypes in multilingual masked language models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 921–932, Seattle, United States. Association for Computational Linguistics.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, Shanya Sharma, Arjun Subramonian, Jaesung Tae, Samson Tan, Deepak Tunuguntla, and Oskar Van Der Wal. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41, virtual+Dublin. Association for Computational Linguistics.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Jingyu Sun, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Nanna Thylstrup and Zeerak Talat. 2020. , detecting 'dirt' and 'toxicity': Rethinking content moderation as pollution behaviour.

Samia Touileb and Debora Nozza. 2022. Measuring harmful representations in Scandinavian language models. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 118–125, Abu Dhabi, UAE. Association for Computational Linguistics.

Sang Truong, Duc Nguyen, Toan Nguyen, Dong Le, Nhi Truong, Tho Quan, and Sanmi Koyejo. 2024. Crossing linguistic horizons: Finetuning and comprehensive evaluation of Vietnamese large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2849–2900, Mexico City, Mexico. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Aniket Vashishtha, Kabir Ahuja, and Sunayana Sitaram. 2023. On evaluating and mitigating gender biases in multilingual settings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 307–318, Toronto, Canada. Association for Computational Linguistics.

VILM. 2023. How did we train Vietcuna?

Thiemo Wambsganss, Xiaotian Su, Vinitra Swamy, Seyed Neshaei, Roman Rietsche, and Tanja Käser. 2023. Unraveling downstream gender bias from large language models: A study on AI educational writing assistance. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10275–10288, Singapore. Association for Computational Linguistics.

Yuxia Wang, Zenan Zhai, Haonan Li, Xudong Han, Shom Lin, Zhenxuan Zhang, Angela Zhao, Preslav Nakov, and Timothy Baldwin. 2024. A Chinese dataset for evaluating the safeguards in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3106–3119, Bangkok, Thailand. Association for Computational Linguistics.

Haryo Wibowo, Erland Fuadi, Made Nityasya, Radityo Eko Prasojo, and Alham Aji. 2024. COPAL-ID: Indonesian language reasoning with local culture and nuances. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1404–1422, Mexico City, Mexico. Association for Computational Linguistics.

Robert Wolfe, Aayushi Dangol, Bill Howe, and Alexis Hiniker. 2025. Representation bias of adolescents in ai: A bilingual, bicultural study. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '24, page 1621–1634. AAAI Press.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Kexin Xu, Yuqi Ye, and Hanwen Gu. 2025. A survey on multilingual large language models: corpora, alignment, and bias. *Frontiers of Computer Science*, 19(11):1911362.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Kang Min Yoo, Jaegeun Han, Sookyo In, Heewon Jeon, Jisu Jeong, Jaewook Kang, Hyunwook Kim, Kyung-Min Kim, Munhyong Kim, Sungju Kim, Donghyun Kwak, Hanock Kwak, Se Jung Kwon, Bado Lee, Dongsoo Lee, Gichang Lee, Jooho Lee, Baeseong Park, Seongjin Shin, Joonsang Yu, Seolki Baek, Sumin Byeon, Eungsup Cho, Dooseok Choe, Jeesung Han, Youngkyun Jin, Hyein Jun, Jaeseung Jung, Chanwoong Kim, Jinhong Kim, Jinuk Kim, Dokyeong Lee, Dongwook Park, Jeong Min Sohn, Sujung Han, Jiae Heo, Sungju Hong, Mina Jeon, Hyunhoon Jung, Jungeun Jung, Wangkyo Jung, Chungjoon Kim, Hyeri Kim, Jonghyun Kim, Min Young Kim, Soeun Lee, Joonhee Park, Jieun Shin, Sojin Yang, Jungsoon Yoon, Hwaran Lee, Sanghwan Bae, Jeehwan Cha, Karl Gylleus, Donghoon Ham, Mihak Hong, Youngki Hong, Yunki Hong, Dahyun Jang, Hyojun Jeon, Yujin Jeon, Yeji Jeong, Myunggeun Ji, Yeguk Jin, Chansong Jo, Shinyoung Joo, Seunghwan Jung, Adrian Jungmyung Kim, Byoung Hoon Kim, Hyomin Kim, Jungwhan Kim, Minkyoung Kim, Minseung Kim, Sungdong Kim, Yonghee Kim, Youngjun Kim, Youngkwan Kim, Donghyeon Ko, Dughyun Lee, Ha Young Lee, Jaehong Lee, Jieun Lee, Jonghyun Lee, Jongjin Lee, Min Young Lee, Yehbin Lee, Taehong Min, Yuri Min, Kiyoon Moon, Hyangnam Oh, Jaesun Park, Kyuyon Park, Younghun Park, Hanbae Seo, Seunghyun Seo, Mihyun Sim, Gyubin Son, Matt Yeo, Kyung Hoon Yeom, Wonjoon Yoo, Myungin You, Doheon Ahn, Homin Ahn, Joohee Ahn, Seongmin Ahn, Chanwoo An, Hyeryun An, Junho An, Sang-Min An, Boram Byun, Eunbin Byun, Jongho Cha, Minji Chang, Seunggyu Chang, Haesong Cho, Youngdo Cho, Dalnim Choi, Daseul Choi, Hyoseok Choi, Minseong Choi, Sangho Choi, Seongjae Choi, Wooyong Choi, Sewhan Chun, Dong Young Go, Chiheon Ham, Danbi Han, Jaemin Han, Moonyoung Hong, Sung Bum Hong, Dong-Hyun Hwang, Seongchan Hwang, Jinbae Im, Hyuk Jin Jang, Jaehyung Jang, Jaeni Jang, Sihyeon Jang, Sungwon Jang, Joonha Jeon, Daun Jeong, Joonhyun Jeong, Kyeongseok Jeong, Mini Jeong, Sol Jin, Hanbyeol Jo, Hanju Jo, Minjung Jo, Chaeyoon Jung, Hyungsik Jung, Jaeuk Jung, Ju Hwan Jung, Kwangsun Jung, Seungjae Jung, Soonwon Ka, Donghan Kang, Soyoung Kang, Taeho Kil, Areum Kim, Beomyoung Kim, Byeongwook Kim, Daehee Kim, Dong-Gyun Kim, Donggook Kim, Donghyun Kim, Euna Kim, Eunchul Kim, Geewook Kim, Gyu Ri Kim, Hanbyul Kim, Heesu Kim, Isaac Kim, Jeonghoon Kim, Jihye Kim, Joonghoon Kim, Minjae Kim, Minsub Kim, Pil Hwan Kim, Sammy Kim, Seokhun Kim, Seonghyeon Kim, Soojin Kim, Soong Kim, Soyoon Kim, Sunyoung Kim, Taeho Kim, Wonho Kim, Yoonsik Kim, You Jin Kim, Yuri Kim, Beomseok Kwon, Ohsung Kwon, Yoo-Hwan Kwon, Anna Lee, Byungwook Lee, Changho Lee, Daun Lee, Dongjae Lee, Ha-Ram Lee, Hodong Lee, Hwiyeong Lee, Hyunmi Lee, Injae Lee, Jaeung Lee, Jeongsang Lee, Jisoo Lee, Jongsoo Lee, Joongjae Lee, Juhan Lee, Jung Hyun Lee, Junghoon Lee, Junwoo Lee, Se Yun Lee, Sujin Lee, Sungjae Lee, Sungwoo Lee, Wonjae Lee, Zoo Hyun Lee, Jong Kun Lim, Kun Lim, Taemin Lim, Nuri Na, Jeongyeon Nam, Kyeong-Min Nam, Yeonseog Noh, Biro Oh, Jung-Sik Oh, Solgil Oh, Yeontaek Oh, Boyoun Park, Cheonbok Park, Dongju Park, Hyeonjin Park, Hyun Tae Park, Hyunjung Park, Jihye Park, Jooseok Park, Junghwan Park, Jungsoo Park, Miru Park, Sang Hee Park, Seunghyun Park, Soyoung Park, Taerim Park, Wonkyeong Park, Hyunjoon Ryu, Jeonghun Ryu, Nahyeon Ryu, Soonshin Seo, Suk Min Seo, Yoonjeong Shim, Kyuyong Shin, Wonkwang Shin, Hyun Sim, Woongseob Sim, Hyejin Soh, Bokyong Son, Hyunjun Son, Seulah Son, Chi-Yun Song, Chiyoung Song, Ka Yeon Song, Minchul Song, Seungmin Song, Jisung Wang, Yonggoo Yeo, Myeong Yeon Yi, Moon Bin Yim, Taehwan Yoo, Youngjoon Yoo, Sungmin Yoon, Young Jin Yoon, Hangyeol Yu, Ui Seon Yu, Xingdong Zuo, Jeongin Bae, Joungeun Bae, Hyunsoo Cho, Seonghyun Cho, Yongjin Cho, Taekyoon Choi, Yera Choi, Jiwan Chung, Zhenghui Han, Byeongho Heo, Euisuk Hong, Taebaek Hwang, Seonyeol Im, Sumin Jegal, Sumin Jeon, Yelim Jeong, Yonghyun Jeong, Can Jiang, Juyong Jiang, Jiho Jin, Ara Jo, Younghyun Jo, Hoyoun Jung, Juyoung Jung, Seunghyeong Kang, Dae Hee Kim, Ginam Kim, Hangyeol Kim, Heeseung Kim, Hyojin Kim, Hyojun Kim, Hyun-Ah Kim, Jeehye Kim, Jin-Hwa Kim, Jiseon Kim, Jonghak Kim, Jung Yoon Kim, Rak Yeong Kim, Seongjin Kim, Seoyoon Kim, Sewon Kim, Sooyoung Kim, Sukyoung Kim, Taeyong Kim, Naeun Ko, Bonseung Koo, Heeyoung Kwak, Haena Kwon, Youngjin Kwon, Boram Lee, Bruce W. Lee, Dagyeong Lee, Erin Lee, Euijin Lee, Ha Gyeong Lee, Hyojin Lee, Hyunjeong Lee, Jeeyoon Lee, Jeonghyun Lee, Jongheok Lee, Joonhyung Lee, Junhyuk Lee, Mingu Lee, Nayeon Lee, Sangkyu Lee, Se Young Lee, Seulgi Lee, Seung Jin Lee, Suhyeon Lee, Yeonjae Lee, Yesol Lee, Youngbeom Lee, Yujin Lee, Shaodong Li, Tianyu Liu, Seong-Eun Moon, Taehong Moon, Max-Lasse Nihlenramstroem, Wonseok Oh, Yuri Oh, Hongbeen Park, Hyekyung Park, Jaeho Park, Nohil Park, Sangjin Park, Jiwon Ryu, Miru Ryu, Simo Ryu, Ahreum Seo, Hee Seo, Kangdeok Seo, Jamin Shin, Seungyoun Shin, Heetae Sin, Jiangping Wang, Lei Wang, Ning Xiang, Longxiang Xiao, Jing Xu, Seonyeong Yi, Haanju Yoo, Haneul Yoo, Hwanhee Yoo, Liang Yu, Youngjae Yu, Weijie Yuan, Bo Zeng, Qian Zhou, Kyunghyun Cho, Jung-Woo Ha, Joonsuk Park, Jihyun Hwang, Hyoung Jo Kwon, Soonyong Kwon, Jungyeon Lee, Seungho Lee, Seonghyeon Lim, Hyunkyung Noh, Seungho Choi, Sang-Woo Lee, Jung Hwa Lim, and Nako Sung. 2024. Hyperclova x technical report. *Preprint*, arXiv:2404.01954.

Abdelrahman Zayed, Prasanna Parthasarathi, Gonçalo Mordido, Hamid Palangi, Samira Shabanian, and Sarath Chandar. 2023. Deep learning on a healthy data diet: finding important examples for fairness. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press.

Jiaxu Zhao, Meng Fang, Zijing Shi, Yitong Li, Ling Chen, and Mykola Pechenizkiy. 2023. CHBias: Bias evaluation and mitigation of Chinese conversational language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13538–13556, Toronto, Canada. Association for Computational Linguistics.

Jun Zhao, Zhihao Zhang, Luhui Gao, Qi Zhang, Tao Gui, and Xuanjing Huang. 2024. Llama beyond english: An empirical study on language capability transfer. *Preprint*, arXiv:2401.01055.

Chujie Zheng, Pei Ke, Zheng Zhang, and Minlie Huang. 2023. Click: Controllable text generation with sequence likelihood contrastive learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1022–1040, Toronto, Canada. Association for Computational Linguistics.

Jingyan Zhou, Jiawen Deng, Fei Mi, Yitong Li, Yasheng Wang, Minlie Huang, Xin Jiang, Qun Liu, and Helen Meng. 2022. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3576–3591, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shucheng Zhu, Bingjie Du, Jishun Zhao, Ying Liu, and Pengyuan Liu. 2024a. Do PLMs and annotators share the same gender bias? definition, dataset, and framework of contextualized gender bias. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 20–32, Bangkok, Thailand. Association for Computational Linguistics.

Shucheng Zhu, Weikang Wang, and Ying Liu. 2024b. Quite good, but not enough: Nationality bias in large language models - a case study of ChatGPT. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13489–13502, Torino, Italia. ELRA and ICCL.

A Annotation Taxonomy

A.1 Linguistic Diversity

Language. What non-English languages are considered?

- *languages listed in ISO 639* (Library of Congress, 2017)

Language Family. What language families do the non-English languages belong to?

- Afro-Asiatic
- Altaic
- Austroasiatic
- Austronesian
- Dravidian
- Eskimo-Aleut
- Indo-European
- Japonic
- Kartvelian
- Koreanic
- Kra-Dai
- Niger-Congo
- Sino-Tibetan
- Uralic

NLP Resources. How much NLP resources do the languages have?

- Class 5 (most highly resourced, according to Joshi et al., 2020)
- Class 4
- Class 3
- Class 2
- Class 1
- Class 0 (lowest)

A.2 Benchmark Development and Cultural Considerations

Benchmark Originality. Is the evaluation benchmark original or adapted from an existing one?

- original
- adapted
- mixed

Benchmark Development Method. How were the benchmark entries constructed?

- written by authors
- contributed by annotators or experts
- crowdsourced
- machine-generated
- reused available benchmarks

Number of Entries. How many entries are in the benchmark?

- *integer value*

Cultural Awareness. How were cultural considerations in benchmark development documented?

- explicit: The paper documents cultural nuances in detail.
- implicit: Cultural awareness is assumed through the participation of cultural insiders but not thoroughly detailed in paper.
- none: The paper shows little to no evidence of considering cultural nuances.

Bias Dimension. Which social dimensions are investigated demographic groups based on?

- age
- caste
- criminal record
- culture
- disability
- disease
- education
- ethnicity
- family structure
- gender
- household registration
- immigration status
- intersectional
- marital status
- nationality
- occupation
- physical appearance
- politics
- pregnancy
- race
- region
- religion
- sexual orientation
- socioeconomic status
- unspecified

Gender Proxy. What terms are used to represent the gender groups being examined?

- grammatical gender (e.g., *the skilled engineering* translating to *el inginiero experto* or *la inginiera experta* in Spanish depending on gender)
- identity terms (e.g., *male/female*)
- names (e.g., *John/Jane*)
- pronouns (e.g., *he/she*)
- roles (e.g., *father/mother*)
- unspecified

Gender Theorization. How is gender conceptualized for papers examining gender bias?

- binary only
- nonbinary-inclusive
- trans-inclusive
- undefined

A.3 Bias Evaluation and Mitigation

Bias Evaluation Metric. What metric is used to measure bias, as broadly classified according to the underlying data structure the metric operates on?

- embedding-based metric
- generated text-based metric
- performance-based metric
- probability-based metric
- no bias evaluation

Benchmark Entry Unit. What format do benchmark entries follow?

- classification instances
- counterfactual inputs – masked tokens

- counterfactual inputs – unmasked sentences
- prompts – question-answering
- prompts – sentence completions
- prompts – rewriting
- word pairs or lists

Bias Mitigation Method (Level 1 Category). What bias mitigation techniques are implemented, as broadly classified according to the LLM workflow stage at which they intervene?

- pre-processing mitigation
- in-training mitigation
- intra-processing mitigation
- post-processing mitigation
- no bias mitigation

Bias Mitigation Method (Level 2 Category). What specific method is used to mitigate bias?

- pre-processing mitigation: data augmentation, data filtering and reweighting, data generation, instruction tuning, projection-based mitigation, feature engineering
- in-training mitigation: architecture modification, loss function modification, selective parameter updating, filtering model parameters
- intra-processing mitigation: decoding strategy modification, weight redistribution, modular debiasing networks, bias reduction experts
- post-processing mitigation: rewriting, chain-of-thought

B Related Work

B.1 PLM Bias Surveys

Blodgett et al. (2020) were among the first to organize the PLM bias literature into an organized meta-analysis. They borrowed the social sciences' measurement modeling framework to unveil the tenuous ways by which bias studies in NLP conceptualize and operationalize bias. They follow up this work with another analysis exposing design flaws in widely utilized bias evaluation benchmarks, such as CrowS-Pairs, StereoSet, and WinoBias (Blodgett et al., 2021). Goldfarb-Tarrant et al. (2023) continue this line of measurement modeling-based analyses by assessing the reliability and validity of ninety bias evaluation benchmarks—87% of which are in English.

Other surveys of PLM bias include those carried out by Czarnowska et al. (2021), who categorized different fairness metrics into three groups, and Navigli et al. (2023), who focused on the various social dimensions of bias explored by past studies. Most recently, Gallegos et al. (2024) and Gupta et al. (2024) separately published comprehensive typologies that were almost identical in their classification of bias evaluation metrics into embedding-based, probability-based, and generation-based measures and bias mitigation methods according to the stage in the training pipeline at which the mitigation intervention is administered.

Our work's objectives are most similar to the aims of Xu et al. (2025), Ramesh et al. (2023b), and Talat et al. (2022), who contemplate the difficulties of evaluating PLM bias multilingual and multicultural settings. Our analysis diverges from theirs in approach, in scope, and in the range of operationalization and method issues considered. While Ramesh et al. (2023b) include only seven multilingual datasets created mostly for text classification tasks, we inspect a bigger number of benchmarks for a broader variety of tasks. We also look beyond the research design factors they and Xu et al. (2025) concentrate on—language, bias dimension, evaluation metric, dataset task, and mitigation—and additionally highlight methodological elements linked to cultural awareness, adaptation methods, and gender theorization among others. On the other hand, the position paper by Talat et al. (2022) reviews the field with a theoretically and conceptually dense perspective. We supplement this by juxtaposing their claims with the empirical evidence our systematic review collates.

B.2 Cultural Awareness and Multilingual Benchmarks

Most of the reviews discussed above call for the development of more multilingual and non-English bias benchmarks. NLP scholars from all over the globe have largely responded to this call (e.g., Lauscher et al., 2020; Névéol et al., 2022; Gamboa and Lee, 2025); however, whether or not the benchmarks they developed are appropriate to the cultures of their chosen languages remains an unanswered question. Multilingual benchmarks used to assess PLMs often arise from machine translations of English language understanding benchmarks (e.g., multilingual MMLU used for GPT-4 in OpenAI et al., 2023 and Llama 3 in Meta, 2024). Consequently, they not only suffer from quality issues but also fail to check and account for knowledge and nuances specific to the culture/s of the translated benchmark's target language/s (Wibowo et al., 2024; Hershcovich et al., 2022). These concerns are especially relevant in the field of PLM bias because values and stereotypes differ across cultures and countries (Talat et al., 2022). For example, Korean and American cultures seem to differ in the way they stereotypically associate socioeconomic status with drug use: while an American bias test links drug use to impoverished individuals (Parrish et al., 2022), Korean researchers note that the reverse is true in their culture and that drug use is seen to be a pastime among the higher social classes of Korea (Jin et al., 2024). The intricacies of constructing culturally sensitive multilingual bias benchmarks are further affirmed by acknowledgments from benchmark creators themselves that their tests might be limited in scope and miss out some important stereotypes in their cultures (e.g., Sahoo et al., 2024; Hsieh et al., 2024). These complexities underscore the need to review the challenges faced and approaches taken by multilingual PLM bias studies in order to guide future research.

C Languages of Annotated Bias Evaluation Benchmarks

See Table 4 and Figure 3.

D Sample Papers

See Table 5.

Language	n	**Language**	n	**Language**	n
Chinese	30	Persian	5	Assamese	1
Spanish	28	Vietnamese	5	Belarusian	1
French	24	Bulgarian	4	Estonian	1
German	24	Filipino	4	Ganda	1
Arabic	20	Punjabi	4	Georgian	1
Italian	16	Thai	4	Hungarian	1
Hindi	16	Urdu	4	Icelandic	1
Russian	14	Catalan	3	Inuktitut	1
Korean	12	Croatian	3	Irish	1
Japanese	11	Finnish	3	Kannada	1
Portuguese	11	Slovak	3	Konkani	1
Indonesian	9	Telugu	3	Kyrgyz	1
Bengali	8	Gujarati	2	Lithuanian	1
Dutch	8	Hebrew	2	Luxembourgish	1
Turkish	8	Kurdish	2	Mongolian	1
Marathi	7	Latvian	2	Odia	1
Swedish	7	Malayalam	3	Sanskrit	1
Czech	6	Maltese	2	Slovenian	1
Danish	6	Nepali	2	Uzbek	1
Polish	6	Romanian	2	Welsh	1
Tamil	6	Serbian	2	Wolof	1
Greek	5	Swahili	2		
Norwegian	5	Ukrainian	2		

Table 4: Number of monolingual sub-benchmarks per language.

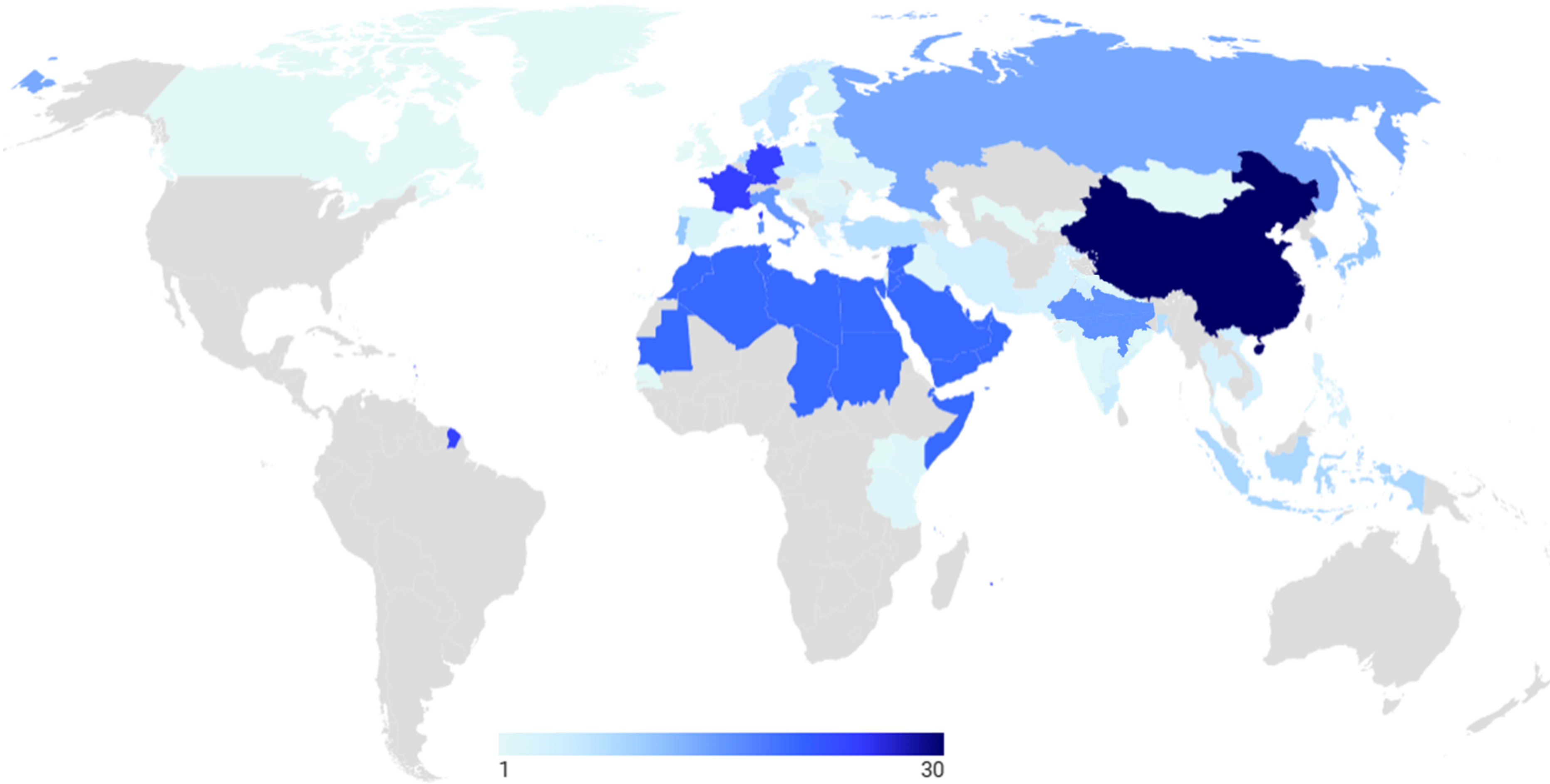

Figure 3: Regional heatmap of monolingual sub-benchmarks. The colors represent the number of benchmarks written using languages widely spoken in a particular country or territory.

Dimension	Sample Papers
gender	Bhutani et al. (2024); Demidova et al. (2024)
religion	Almazrouei et al. (2023); Levy et al. (2023)
race	Nie et al. (2024); Huang et al. (2024)
sexual orientation	Bergstrand and Gambäck (2024); Mukherjee et al. (2023)
age	Wolfe et al. (2025); Névéol et al. (2022)
nationality	Zhu et al. (2024b); Das et al. (2023)
ethnicity	Ramesh et al. (2023a); Câmara et al. (2022)
disability	Mina et al. (2024); Fort et al. (2024)
physical appearance	Zhao et al. (2023); Costa-jussà et al. (2023)
socioeconomic status	Nie et al. (2024); Grigoreva et al. (2024)
region	Billah Nagoudi et al. (2023); Deng et al. (2022)
politics	Al Ali and Libovický (2024); Barkhordar et al. (2024)
intersectional bias	Sahoo et al. (2024); Devinney et al. (2024)
caste	B et al. (2022); Bhatt et al. (2022)
culture	Naous et al. (2024); Demidova et al. (2024)
education	Huang and Xiong (2024); Jin et al. (2024)
occupation	Lee et al. (2024); Zhou et al. (2022)
immigration statu	Ousidhoum et al. (2021); Mukherjee et al. (2023)
family structure	Jin et al. (2024); Lee et al. (2023b)
marital status	Lee et al. (2023b)
criminal record	Lee et al. (2023b)
pregnancy	Lee et al. (2023b)
household registration	Huang and Xiong (2024)
disease	Huang and Xiong (2024)

Table 5: Social dimensions analyzed by multilingual bias studies.