Rethinking Cultural Inclusivity in Large Language Models: A Critical Analysis

Anonymous ACL submission

Abstract

Large Language Models (LLMs) increasingly shape global discourse yet predominantly encode Western epistemological traditions. This position paper critically examines current approaches to cultural inclusivity in LLMs, arguing that they often rely on unidimensional metrics that inadequately capture cultural diversity. We advocate for Multiplexity—a framework recognizing multiple layers of existence, knowledge, and truth—as a theoretical foundation for developing more culturally inclusive language models. Our analysis demonstrates the limitations of traditional cultural alignment methods and highlights empirical evidence showing how Multiplexity-based interventions, particularly through Multi-Agent Systems, significantly improve cultural representation. By contrasting "Uniplexity" with Multiplexity, we address the epistemological limitations of current evaluation frameworks and propose moving beyond binary metrics toward multidimensional cultural evaluation. This paper contributes to ongoing efforts to mitigate cultural biases in AI systems, ultimately supporting more globally inclusive language technologies that respect diverse cultural perspectives.

1 Introduction

Large Language Models (LLMs) have transformed Natural Language Processing and extended into education, healthcare, and public discourse. However, evidence demonstrates these models disproportionately reflect Western, Educated, Industrialized, Rich, and Democratic (WEIRD) perspectives, potentially reinforcing cultural homogenization (Mushtaq et al., 2025a; Johnson et al., 2022; Qu and Wang, 2024). This cultural narrowness presents challenges when LLMs are deployed globally, where their biases can shape users' worldviews and conceptions of knowledge itself.

Frameworks to detect, measure, and mitigate cultural biases in LLMs have become increasingly urgent. While existing research has documented various biases, approaches for comprehensive cultural bias assessment remain nascent. As Zhou et al. argue, culture is not merely superficial trivia but a deeply embedded context that shapes how language operates. Traditional evaluation frameworks often fail to capture the multi-dimensional nature of cultural representation, instead focusing on one-dimensional metrics that insufficiently reflect global cultural diversity.

In this position paper, we propose Multiplexity as a framework for addressing cultural biases in LLMs. Multiplexity—derived from the Arabic term "marātib" (meaning hierarchy or levels)—offers an approach that recognizes multiple layers of existence, knowledge, and truth (Qadir, 2022; Qadir and Şentürk, 2024). This framework provides an alternative to what might be termed "Uniplexity," which tends to reduce multi-layered reality to single-dimensional perspectives often rooted in Western epistemological traditions.

Our position contributes to the field in several significant ways:

- We critically analyze existing approaches to cultural evaluation in LLMs and identify their conceptual limitations
- We advocate for **Multiplexity** as a theoretical framework for evaluating cultural representations in LLMs, and support this with empirical evidence demonstrating its effectiveness compared to existing methodologies.
- We propose a research agenda for advancing cultural inclusivity in NLP

2 Cultural Narrowness Problem in LLMs

2.1 Evidence of Cultural Bias

Recent work consistently demonstrates pervasive cultural bias in LLMs. Tao et al. find that popular

GPT models encode values aligned with English-speaking, Protestant-European cultures. Qu and Wang report that ChatGPT performs best on opinions from Western, English-speaking, developed nations (especially the US). Durmus et al. quantified this effect using the GlobalOpinionQA dataset, showing that LLMs' default outputs more closely match U.S. and European survey opinions than those of other countries.

These biases have real consequences: Lewis observes that cultural prejudice in AI language-learning apps can depress minority student participation by over 30%. Such findings underscore the risk that biased LLMs may distort communication and limit utility for diverse user groups (Adilazuarda et al., 2024).

2.2 Limitations of Current Cultural Evaluation Frameworks

Researchers have developed specialized benchmarks and training methods for cultural alignment. Common approaches involve prompting LLMs with country-specific surveys (e.g., Hofstede's cultural dimensions, World Values Survey) and comparing responses to human data (Hofstede et al., 2010; wvs, 2020). New benchmarks such as CDEVal (Wang et al., 2024), WorldValuesBench (Zhao et al., 2024), and GlobalOpinionQA (Durmus et al., 2024) explicitly test cross-cultural value alignment.

However, these approaches suffer from fundamental limitations. They exhibit **Western-centrism**, relying on English-dominated corpora and Western-derived surveys. As Adilazuarda et al. point out, many studies "do not explicitly define 'culture'" and rely on narrow proxies. Evaluations are typically based on **unidimensional metrics**—binary alignments with cultural norms via statistical correlations with survey responses. These methods also reflect **cognitive imperialism**, privileging Western epistemologies while sidelining indigenous worldviews (Ofosu-Asare, 2024).

The brittleness of static, survey-based cultural proxies has been highlighted in recent work. Khan et al. demonstrates that trivial changes (like prompt wording or scale length) can dramatically alter measured "alignment" by as much as $d \approx 0.09$. Likewise, Kabir et al. note that MCQ surveys "fail to capture the intricate nuances of cultural values," yielding only a tiny fraction of aligned responses.

2.3 Cultural Bias and NLP Harms

Culturally biased language technologies can cause two main types of harm (Blodgett et al., 2020):

Representational harms occur when systems stereotype, erase, or mischaracterize specific groups. LLMs that reflect dominant viewpoints while marginalizing others reinforce narrow worldviews, e.g., through stereotypical portrayals or erasure of minority perspectives.

Allocational harms arise when systems inequitably distribute opportunities or resources, such as differing service quality based on a user's cultural background. These are often rooted in training data and model objectives. Models trained mainly on Western-centric corpora tend to underperform for marginalized communities (Adilazuarda et al., 2024; Ofosu-Asare, 2024). Addressing such disparities requires interventions at earlier stages of the pipeline in the data and design stages. This reflects recent work emphasizing that cultural issues in language technology cannot be solved by technical fixes alone, as problems like cultural analysis are as much social and political as technical (Blodgett et al., 2020).

2.4 Multiplexity: A Framework for Cultural Inclusivity

Multiplexity provides an analytical framework that addresses limitations of unidimensional approaches to cultural evaluation in LLMs. It encompasses multiple integrated dimensions, including epistemological diversity (acknowledging diverse ways of knowing) and ontological plurality (recognizing multiple levels of existence). This approach, with roots in Islamic intellectual traditions but applicable across diverse cultural contexts, offers a corrective to "Uniplexity"—the reductionist Western paradigm that privileges empirical and material knowledge while marginalizing other epistemologies (Şentürk et al., 2020; Qadir and Şentürk, 2024).

2.4.1 The Case for Epistemic Pluralism

Most current LLM evaluation frameworks implicitly adopt a perspective that assumes a universally applicable epistemology, privileging certain ways of knowing (typically Western and analytical) while marginalizing alternatives. Multiplexity-based evaluation acknowledges diverse epistemologies as valid pathways to knowledge, recognizing that different cultural traditions have developed unique approaches to understanding reality. This

pluralistic stance suggests that cultural evaluation of LLMs should assess their ability to engage with multiple knowledge systems simultaneously, rather than measuring alignment with a single cultural norm.

2.4.2 Evaluation Metrics

To quantify cultural inclusivity, researchers have developed metrics that offer numerical assessment of representation (Mushtaq et al., 2025a,b). Figure 1 illustrates the multiplex analysis pipeline. The Perspectives Distribution Score (PDS) measures the proportional representation of each cultural perspective:

$$P_i = \frac{R_i}{\sum_j R_j} \tag{1}$$

where R_i is the reference count for perspective i. PDS Entropy extends this by measuring the balance of those proportions:

$$\text{PDS}_E = -\sum_{i=1}^{n} p_i \log(p_i) \tag{2}$$

2.4.3 Intervention Strategies

Building on these metrics, researchers have proposed key intervention strategies to mitigate representational harm (Mushtaq et al., 2025b), as discussed in Section 2.3. One approach is *Contextually-Implemented Multiplexity*, which integrates multiplex principles into system prompts without requiring changes to the model architecture. Another is *Multi-Agent System (MAS)-Implemented Multiplexity*, which involves multiple LLM agents, each representing distinct cultural perspectives, working collaboratively to produce more balanced and inclusive outputs.

Emerging research also explores multi-agent approaches to capture pluralistic values. Yuan et al. introduced "Cultural Palette," a framework with continent-specific alignment agents and a meta-agent that dynamically merges their outputs. Feng et al. present "Modular Pluralism," which augments a base LLM with smaller community-specific LLMs that better cover underrepresented perspectives than any single model (Feng et al., 2024).

3 Empirical Evidence and Critique

3.1 Limitations of Traditional Approaches

Quantitative findings illustrate the limitations of traditional metrics:

Hofstede-based alignment: Masoud et al. report extremely weak LLM/Hofstede agreement, with average Kendall's τ of only $\sim$0.14 even for GPT-4 and country rankings mis-ordered 60–90% of the time. These near-random correlations show that treating each country as a single point value grossly misrepresents how LLMs "view" culture.

Closed-survey probes: Kabir et al. find that standard multiple-choice prompts achieve high alignment in only a negligible fraction of cases. Even when mapped onto survey options, many model answers are "unclassifiable," indicating that forced-choice tests miss most cultural content.

Bias scores: Naous et al. introduce a Cultural Bias Score showing that even on Arabic prompts about Arab culture, multilingual LMs scored CBS $\approx$40–60% on average, meaning nearly half their answers favored Western entities.

Prompting effects: Tao et al. report that tailored prompts raised alignment in 71–81% of cases, but LLMs still frequently gravitate toward Western norms. AlKhamissi et al. find models align much better when queried in the culture's dominant language than with generic prompts.

3.2 Comparative Performance Analysis

In contrast, implementations of the Multiplexity framework show promising results for cultural diversity. In their first study focusing on educational contexts, baseline LLM outputs (no mitigation) had a Perspectives Distribution Score (PDS) entropy of only 3.25%—essentially zero diversity (nearly all answers reflect one viewpoint). Intervention using Contextually-Implemented Multiplexity raised entropy only to about 19%, a modest shift. However, the Multi-Agent System (MAS) approach boosted PDS entropy to 98%, nearly its theoretical maximum (Mushtaq et al., 2025a).

These findings were expanded in further work (Mushtaq et al., 2025b), in which researchers benchmarked various LLMs across 175 questions divided into 7 categories. The PDS entropy improved from 13% in baseline settings to 26% using Contextually-Implemented Multiplexity, and reached 94% using the MAS-Implemented Multiplexity intervention strategy. Example of their perspective extraction pipeline (needed to calculate PDS) and sentiment analysis in baseline LLM and Contextually-Implemented Multiplexity intervention strategy has been presented in figure 1.

Sentiment analysis provides additional context, with MAS-Implemented Multiplexity achieving

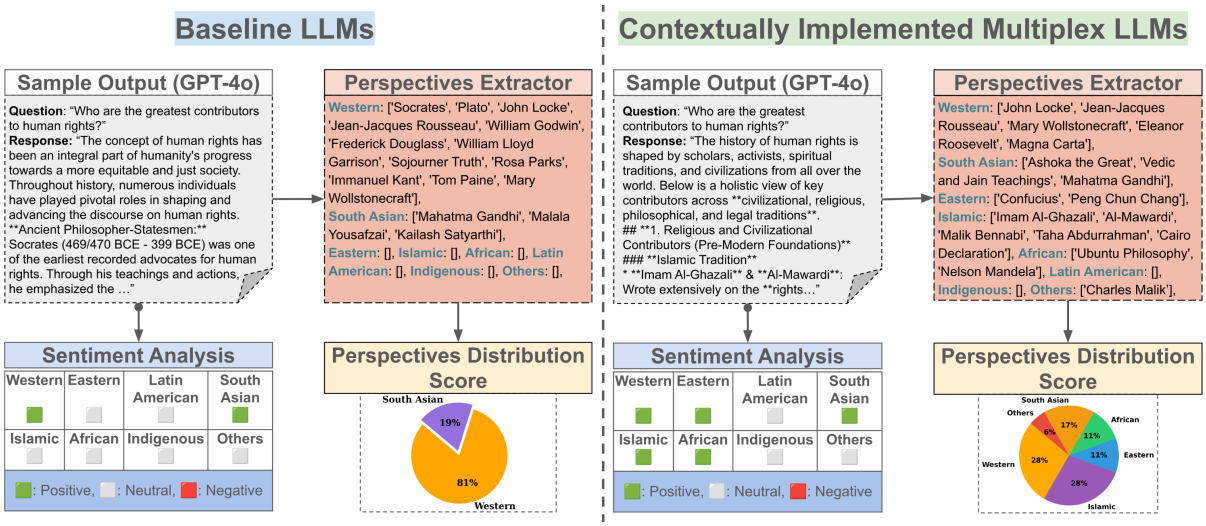

Figure 1: A sample pipeline output for PDS calculation to assess cultural inclusivity through a Multiplexity lens

94% positive sentiment across cultures, compared to predominantly neutral sentiment in baseline models (Mushtaq et al., 2025a). A shift of 67.7% towards positive sentiment was reported in their follow-up work (Mushtaq et al., 2025b).

These results suggest an important difference: traditional metrics often peak in the low-to-mid tens of percent, reflecting narrow agreement with one cultural norm (Masoud et al., 2024; Naous et al., 2024). Multiplexity approaches, particularly through multi-agent systems, appear to achieve more balanced representation, with PDS scores increasing from single digits to over 90%, indicating more uniform representation across cultures.

4 Future Research Directions

While addressing both representational and allocational harms requires technical interventions, a comprehensive solution requires broader engagement with social, political, and ethical dimensions. Our research agenda includes:

1. **Inclusive Co-Design:** Involve diverse communities and scholars from different cultural traditions and fields in designing and evaluating language models, as emphasized by Ofosu-Asare (2024).

2. **Culturally-Inclusive Datasets:** Create training and evaluation datasets for pre-training and fine-tuning stages.

3. **Multi-Agent Architectures:** Further explore multi-agent approaches that represent diverse perspectives. Early results from MAS-Implemented Multiplexity are promising for mitigating representational harms by improving balanced cultural representation.

Cultural and persona-based prompting are useful steps forward, but to build truly inclusive AI systems, we need datasets grounded in multiplexity principles, designed to reflect diverse global perspectives during fine-tuning and to address both representational and allocational harms identified earlier in section 2.3.

5 Conclusion

We have advocated for Multiplexity as a theoretical framework to address cultural biases in Large Language Models. Unlike traditional approaches that rely on unidimensional metrics and often reflect Western norms, Multiplexity recognizes multiple layers of existence and knowledge. Our analysis highlights the limitations of conventional cultural alignment methods, which tend to yield limited diversity and moderate alignment. In contrast, Multiplexity-based interventions—especially those using Multi-Agent Systems—demonstrate significant improvements in cultural inclusivity metrics, suggesting a path toward mitigating both representational and allocational harms.

While no single framework can fully resolve the complexities of cultural representation in AI, Multiplexity provides a valuable foundation for moving beyond reductionist perspectives. By embracing epistemological pluralism and multidimensional evaluation, we can advance toward language models that more respectfully and accurately reflect the diversity of global cultures.

Limitations

While we advocate for Multiplexity as a promising framework for advancing cultural inclusivity in LLMs, several limitations warrant consideration. First, the empirical evidence supporting Multiplexity-based interventions, though encouraging, is currently based on a limited set of inference-time studies. Broader validation will require deeper integration at the data and design stages of model development. Second, the implementation of Multiplexity—particularly through Multi-Agent Systems—may introduce computational complexity that poses challenges for large-scale or resource-constrained deployment. Third, our emphasis on cultural inclusivity addresses a critical but singular facet of the broader imperative to develop ethical and socially responsible AI. Lastly, while Multiplexity offers a strong theoretical foundation, its practical realization depends on sustained collaboration with diverse communities and the iterative refinement of both evaluation metrics and intervention strategies.

References

2020. World values survey: Round seven – country-pooled datafile. Version: 2020.

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling "culture" in llms: A survey. *Preprint*, arXiv:2403.15412.

Badr AlKhamissi, Muhammad ElNokrashy, Mai AlKhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. *Preprint*, arXiv:2402.13231.

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.

Esin Durmus, Karina Nguyen, Thomas I. Liao, Nicholas Schiefer, Amanda Askell, Anton Bakhtin, Carol Chen, Zac Hatfield-Dodds, Danny Hernandez, Nicholas Joseph, Liane Lovitt, Sam McCandlish, Orowa Sikder, Alex Tamkin, Janel Thamkul, Jared Kaplan, Jack Clark, and Deep Ganguli. 2024. Towards measuring the representation of subjective global opinions in language models. *Preprint*, arXiv:2306.16388.

Shangbin Feng, Taylor Sorensen, Yuhan Liu, Jillian Fisher, Chan Young Park, Yejin Choi, and Yulia Tsvetkov. 2024. Modular pluralism: Pluralistic alignment via multi-llm collaboration. *Preprint*, arXiv:2406.15951.

G. Hofstede, G.J. Hofstede, and M. Minkov. 2010. *Cultures and Organizations: Software of the Mind, Third Edition*. McGraw Hill LLC.

Rebecca L Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in gpt-3. *Preprint*, arXiv:2203.07785.

Mohsinul Kabir, Ajwad Abrar, and Sophia Ananiadou. 2025. Break the checkbox: Challenging closed-style evaluations of cultural alignment in llms. *Preprint*, arXiv:2502.08045.

Ariba Khan, Stephen Casper, and Dylan Hadfield-Menell. 2025. Randomness, not representation: The unreliability of evaluating cultural alignment in llms. *Preprint*, arXiv:2503.08688.

André Lewis. 2025. Unpacking cultural bias in ai language learning tools: An analysis of impacts and strategies for inclusion in diverse educational settings. *International Journal of Research and Innovation in Social Science*, Volume IX:1878–1892.

Reem I. Masoud, Ziquan Liu, Martin Ferianc, Philip Treleaven, and Miguel Rodrigues. 2024. Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions. *Preprint*, arXiv:2309.12342.

Abdullah Mushtaq, Muhammad Rafay Naeem, Muhammad Imran Taj, Ibrahim Ghaznavi, and Junaid Qadir. 2025a. Toward inclusive educational ai: Auditing frontier llms through a multiplexity lens. *Preprint*, arXiv:2501.03259.

Abdullah Mushtaq, Imran Taj, Rafay Naeem, Ibrahim Ghaznavi, and Junaid Qadir. 2025b. Worldview-bench: A benchmark for evaluating global cultural perspectives in large language models. *Preprint*, arXiv:2505.09595.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. *Preprint*, arXiv:2305.14456.

Yaw Ofosu-Asare. 2024. Cognitive imperialism in artificial intelligence: counteracting bias with indigenous epistemologies. *AI & SOCIETY*, pages 1–17.

Junaid Qadir. 2022. *A Holistic Education for the 21st Century Engineer Based on Wisdom and Multiplexity*.

Junaid Qadir and Recep Şentürk. 2024. Educating for the ai era: Harnessing the wild genai horse through multiplex ai humanities and critical ai literacy. *SSRN Electronic Journal*. Available at

SSRN: https://ssrn.com/abstract=5015109 or http://dx.doi.org/10.2139/ssrn.5015109.

Yao Qu and Jue Wang. 2024. Performance and biases of large language models in public opinion simulation. *Humanities and Social Sciences Communications*, 11.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9).

Yuhang Wang, Yanxu Zhu, Chao Kong, Shuyu Wei, Xiaoyuan Yi, Xing Xie, and Jitao Sang. 2024. Cdeval: A benchmark for measuring the cultural dimensions of large language models. *Preprint*, arXiv:2311.16421.

Jiahao Yuan, Zixiang Di, Shangzixin Zhao, and Usman Naseem. 2025. Cultural palette: Pluralising culture alignment via multi-agent palette. *Preprint*, arXiv:2412.11167.

Wenlong Zhao, Debanjan Mondal, Niket Tandon, Danica Dillion, Kurt Gray, and Yuling Gu. 2024. Worldvaluesbench: A large-scale benchmark dataset for multi-cultural value awareness of language models. *Preprint*, arXiv:2404.16308.

Naitian Zhou, David Bamman, and Isaac L. Bleaman. 2025. Culture is not trivia: Sociocultural theory for cultural nlp. *Preprint*, arXiv:2502.12057.

Recep Şentürk, Alparslan Açıkgenç, Ömer Küçükural, Qazi N. Yamamoto, N. Keskin Aksay, S. Özalkan, and A. Asadov. 2020. *Comparative Theories and Methods: Between Uniplexity and Multiplexity*. Ibn Haldun University Press.