looks interesting but the data are not clear & methodology

Analysis of Cultural Biases in Language Models

Omid Reza Heidari, 40267435

October 2024

Introduction

Large Language Model (LLM)-based systems are increasingly used across various fields recently, from medical to industrial and personal to professional applications. Although LLMs enhance efficiency, they are susceptible to biases and vulnerabilities that must be addressed to ensure fair and equitable AI usage [3,7,13]. Cultural bias in LLMs can lead to the reinforcement of stereotypes or the exclusion of minority viewpoints, resulting in unfair outcomes for marginalized groups. Cultural bias refers to the tendency of systems or individuals to favor certain cultural norms or perspectives over others, often leading to unfair treatment or misrepresentation. This bias can manifest in gender disparities, social hierarchies (such as caste), dialectical differences, racial stereotypes, and religious prejudices, affecting how individuals are perceived and treated across various domains of life.

Related Works

Recent studies on various types of cultural biases have proposed numerous methods for identifying these biases [2]. For instance, [4] built upon [6]'s work, revealing that stereotypes related to regional identities are present in datasets like Wikipedia and IndicCorp-en, as well as in language models such as MuRIL and mBERT. Additionally, [11] investigates gender biases in English-Hindi machine translation using a modified TGBI metric.

This metric assesses gender biases in translation systems, taking into account the grammatical aspects specific to Hindi. Furthermore, [5] found that content moderation on Quora inferred Bengali users' nationality and religious identities based on their linguistic behavior, favoring Indian Hindu dialects while marginalizing Bangladeshi Muslim dialects. Regarding racial biases, previous research has applied techniques such as word embedding debiasing [9] and implemented strategies to mitigate biases in hate speech detection [12]. In South Asia, caste-based biases in LLMs often reinforce stereotypes about lower-caste groups, portraying them as less educated or confined to menial roles. Finally, dialect-related biases have been a focus of recent research, with studies such as [1] introducing dialect-aware fine-tuning to improve the handling of dialectal variations, and [13] conducting both qualitative and quantitative evaluations of dialect-based biases in Bengali.

Methodology

Our research will utilize Cultural Alignment Test (Hoftede's CAT) [10] to quantify cultural alignment, which is using Hofstede's cultural dimension framework [8], which offers an explanatory cross-cultural comparison through the latent variable analysis. We will apply our approach to quantitatively evaluate LLMs— namely GPT-4 , GPT-3.5, Claude 3.5, and Gemini 1.5-against the cultural dimensions of regions (which is called racial biases) like the U.S., France, and Iran, using different prompting styles and exploring the effects of language-specific fine-tuning on the models' behavioural tendencies and cultural values.

I have selected GPT-4, GPT-3.5, Claude 3.5, and Gemini 1.5 as the target LLMs for this study due to several compelling reasons. To begin with, these models are freely accessible, making them suitable for academic research without incurring significant costs. They, moreover, have demonstrated exceptional performance across various natural language processing tasks, including those involving nuanced cultural contexts.

I also selected the U.S., France, and Iran for this study because their cultural differences allow us to better observe how LLMs respond to diverse contexts. These countries vary significantly in language, values, and social norms, offering a broad spectrum for evaluating cultural biases. In contrast, selecting similar countries, like the U.S. and Canada, might yield

TABLE 1 Estimated Timeline

13 October	Preparing questionnaires for different countries
18 October	Developing Code to apply questionnaires on differnt LLMs using their API
01 November	Analyzing the results and compareing them with each other
15 November	Preparing for the Final Exam (No work)
22 November	Preparing the poster and writing the final report
29 November	Present the work

similar results, limiting our ability to fully explore the LLMs' adaptability. This diverse selection ensures more meaningful comparisons.

Timeline

References

[1] Fakhraddin Alwajih, Gagan Bhatia, and Muhammad Abdul-Mageed. Dallah: A dialect-aware multimodal large language model for Arabic. In Nizar Habash, Houda Bouamor, Ramy Eskander, Nadi Tomeh, Ibrahim Abu Farha, Ahmed Abdelali, Samia Touileb, Injy Hamed, Yaser Onaizan, Bashar Alhafni, Wissam Antoun, Salam Khalifa, Hatem Haddad, Imed Zitouni, Badr AlKhamissi, Rawan Almatham, and Khalil Mrini, editors, *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 320–336, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[2] MD Mohaymen Ul Anam Amit Agarwal Hitesh Laxmichand Pa-

tel Bhargava Kumar Taki Hasan Rafi Azmine Toushik Wasi, Omid Reza Heidari and Dong-Kyu Chae. A review of human-centric evaluation of cultural bias in indic languages within llms: Rethinking research directions. In *Proceedings of COLING 2025*, 2025.

[3] Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. Measuring implicit bias in explicitly unbiased large language models, 2024.

[4] Shaily Bhatt, Sunipa Dev, Partha Talukdar, Shachi Dave, and Vinodkumar Prabhakaran. Re-contextualizing fairness in NLP: The case of India. In Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, editors, *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, November 2022. Association for Computational Linguistics.

[5] Dipto Das, Carsten Østerlund, and Bryan Semaan. "jol" or "pani"?: How does governance shape a platform's identity? 5(CSCW2), oct 2021.

[6] Thomas A. de Souza. Regional and communal stereotypes of bombay university students. *Indian Journal of Social Work*, 38(1):37–44, 1977.

[7] Justin Edwards, Leigh Clark, and Allison Perrone. Lgbtq-ai? exploring expressions of gender and sexual orientation in chatbots. In *Proceedings of the 3rd Conference on Conversational User Interfaces*, CUI '21, New York, NY, USA, 2021. Association for Computing Machinery.

[8] Geert Hofstede. Dimensionalizing cultures: The hofstede model in context. *Online Readings in Psychology and Culture*, 2(1), 2011.

[9] Thomas Manzini, Lim Yao Chong, Alan W Black, and Yulia Tsvetkov. Black is to criminal as Caucasian is to police: Detecting and removing multiclass bias in word embeddings. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long*

and Short Papers), pages 615–621, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

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

[11] Krithika Ramesh, Gauri Gupta, and Sanjay Singh. Evaluating gender bias in Hindi-English machine translation. In Marta Costa-jussa, Hila Gonen, Christian Hardmeier, and Kellie Webster, editors, *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 16–23, Online, August 2021. Association for Computational Linguistics.

[12] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics.

[13] Azmine Toushik Wasi, Raima Islam, Mst Rafia Islam, Taki Hasan Rafi, and Dong-Kyu Chae. Exploring bengali religious dialect biases in large language models with evaluation perspectives. In *Proceedings of The First Human-centered Evaluation and Auditing of Language Models Workshop at the CHI Conference on Human Factors in Computing Systems*, CHI '24, 2024.