ChatGPT exhibits bias towards developed countries over developing ones, as indicated by a sentiment analysis approach

Georgios P. Georgiou[1 2]

[1]Department of Languages and Literature, University of Nicosia, Nicosia, Cyprus

[2]Director of the Phonetic Lab

georgiou.georg@unic.ac.cy

Abstract

This study analyzes how ChatGPT characterizes developed and developing countries using a sentiment analysis framework. We selected 10 countries with the highest Human Development Index (HDI) and 10 countries with the lowest. The sentiment analysis provided scores indicating the degree of positivity in the descriptions of these countries provided by ChatGPT. The results revealed that ChatGPT generally expressed positive sentiments about all countries. However, strong evidence emerged showing that countries with high HDI received more positive sentiments compared to those with low HDI. These findings highlight the bias of the model in describing developed versus developing countries. Ultimately, the study highlights the importance of adjusting large language models to ensure fairer representations of countries.

Keywords: ChatGPT, developing countries, developed countries, sentiment analysis

1. Introduction

Large Language Models (LLMs) have transformed the landscape of Natural Language Processing (NLP). The capability of LLMs like ChatGPT in producing high-quality text closely mimicking human language is well-documented (Adeshola & Adepoju, 2023; Chukwuere 2024). Nevertheless, LLMs inadvertently mirror and perpetuate biases present in their training data (Caliskan, 2017). While content filtering techniques have been employed to mitigate harmful outputs (Markov et al., 2023), biases can persist within the model itself (Ray, 2023). Deploying biased models in real-world applications can have detrimental consequences, as demonstrated by incidents such as those involving Artificial Intelligence (AI) healthcare predictions (Obermeyer et al., 2019).

LLMs demonstrate various biases associated with gender (Gross, 2023), language (Georgiou, 2024), religion (Abid et al., 2021), politics (Rozado, 2023), and nationality (Venkit et al., 2023) among others. Zhou et al. (2024) investigated the nationality bias of ChatGPT using a sample of 195 countries, with descriptions provided in both English and Chinese. The authors evaluated the output using vocabulary richness, sentiment, and offensiveness metrics. Evaluations of language have also been conducted by humans and ChatGPT. The findings indicated that although the generated content was largely positive, ChatGPT produced negative content when given prompts with negative connotations. Although the model viewed its output as neutral, it consistently demonstrated self-awareness of nationality bias when evaluated using the same pair-wise comparison

annotation method employed by human annotators.

The examination of country bias in AI-generated language has received minimal scientific attention. In a relevant study, Boussidan et al. (2023) explored the biases of ChatGPT regarding various countries around the globe. The authors followed a sentiment analysis approach, prompting the model to assign a positivity score to each country. Prompts were provided in four different languages, namely French, English, Russian and Arabic. The findings revealed that North American and European countries received higher scores, whereas African countries received the lowest. South American and Asian countries typically fell in the middle range. The results also denoted variations across languages. When prompted in French, African countries, particularly those colonized by France, tended to receive more negative scores. In contrast, when prompted in English, the model assigned positive scores to Commonwealth nations like India and Australia. However, the sentiment analysis in the above study relied on the scores assigned to these countries by ChatGPT, using a scale developed by the authors. The conclusion drawn is that ChatGPT is biased towards specific countries. For instance, Salinas et al. (2023) reported that when prompted to select 20 nationalities, the model chose exclusively from Western countries, omitting African nations.

This study aims to fill a research gap by investigating the sentiments found in ChatGPT-generated language regarding developed versus developing countries. We uniquely employ a sentiment analysis framework, which indicates the sentiment scores of the generated texts about the countries under investigation. These scores were derived from an embedded online dictionary, which assessed the positivity of each word in the generated language. The comparison between developed and developing countries is performed through statistical modeling. Since the LLM is sensitive to biases based on previous research, we hypothesize that developed countries will exhibit higher sentiment scores than developing countries. By analyzing sentiment scores in AI-generated texts, the study aims to reveal potential biases, promoting fairer and more accurate representations. This effort supports ethical AI development, enhances trustworthiness in AI systems, and ensures informed decision-making.

2. Methodology

2.1. Procedure

We used ChatGPT-3.5 to generate the texts. We employed a prompt designed to elicit unbiased thoughts about specific countries. Specifically, we presented the following prompt to ChatGPT: “Please provide us with any thoughts about [name of the country] within 10 sentences”. The sample consisted of 20 countries selected according to their Human Development Index (HDI). HDI is a composite statistic that combines life expectancy, education (measured by both the average years of schooling completed and the expected years of schooling at the start of education), and per capita income. This index categorizes countries into four levels of human development. Higher HDI scores correspond to longer lifespans, higher education levels, and greater gross national income per capita adjusted for purchasing power parity. HDI is employed by the United Nations Development Program’s Human Development Report Office to assess and compare the development progress of countries (World Health Organization, 2024).

The selected countries were retrieved from the latest Human Development Report 2023-24 and include data from 2022 (Conceição, 2024). These countries were Switzerland, Norway, Iceland, Hong Kong, Denmark, Sweden, Ireland, Germany, Singapore, and Netherlands as well as Sierra Leone, Burkina Faso, Yemen, Burundi, Mali, Niger, Chad, Central African Republic, South Sudan, and Somalia. The first 10 countries had the highest HDI in the report (0.946 – 0.967, Standard Deviation (*SD*) = 0.007), while the other 10 countries had the lowest HDI (0.38 – 0.424, *SD* = 0.02).

2.2. Analysis

The sentiment analysis was conducted with the use of the *SentimentAnalysis* package in R (R Core Team, 2024). Sentiments were extracted utilizing the QDAP dictionary from the *qdapDictionaries* package (Rinker, 2021). The value of particular words ranges from -1 (highly negative) to 1 (highly positive). Scores close to zero indicate neutral sentiment.

We used a Bayesian regression model via the *brms* package (Bürkner et al., 2023) in R to analyze our data. This is because of its potential to handle small sample data (Georgiou, 2023). The dependent variable included the sentiment `SCORE` measured between -1 and 1. `HDI` (low/high) was modeled as the fixed factor, while `COUNTRY` was treated as a random factor. Weakly informative priors were used, given the lack of predefined assumptions about the data parameters (Georgiou & Giannakou, 2024). These priors followed a student's *t*-distribution with 3 degrees of freedom, a mean of 0, and an *SD* of 2.5 (Georgiou & Kaskampa, 2024). The Evidence Ratio (ER) was used to assess the likelihood of the test hypotheses compared to their alternatives. We adhered to Jeffreys's (1961) approach, considering an Evidence Ratio (ER) of 10 or higher as strong evidence in favor of a hypothesis, and an ER of 0.1 or lower as strong evidence against a hypothesis.

3. Results

The sentiment analysis indicated positive average sentiment scores (i.e., > 0) for all countries under investigation. However, according to the descriptive statistics, the language related to the high HDI group had more positive sentiments compared to the low HDI group. Figure 1 shows the sentiment scores of both high and low HDI countries together with their *SD*s. The scores ranged between -0.29 – 0.57 for high HDI and -0.5 – 0.5 for low HDI. Figure 2 illustrates the sentiment scores and the *SD*s for each country with the high and low HDI.

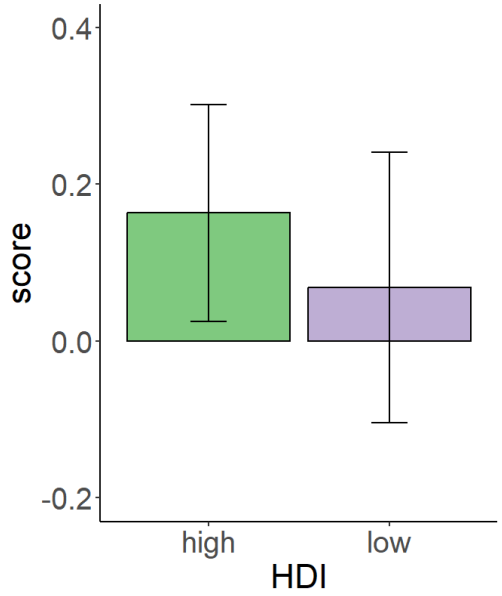

Figure 1: Sentiment scores for countries with high and low HDI

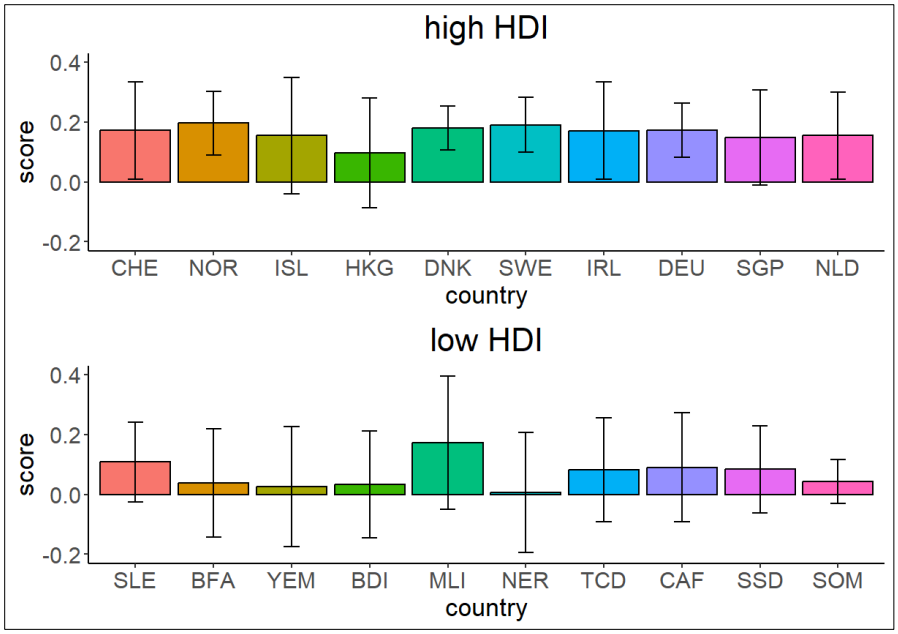

Figure 2: Sentiment scores for each country with high and low HDI

We utilized a Bayesian regression model to assess whether ChatGPT exhibited sentiment differences between countries with high HDI and those with low HDI. According to the analysis, the Credible Interval (CI) for high HDI suggests that there is a 95% probability that the true value of the sentiment score lies between 0.13 and 0.20. As the values do not cross zero, there is strong evidence that the true value of high HDI was greater than zero, indicating in this case positive sentiments for these countries. Similarly, the CI for low HDI

indicates a 95% probability that the true value of the parameter lies between 0.04 and 0.10. This provides strong evidence that the low HDI is significantly greater than zero, implying that these countries are associated with positive sentiments. Subsequent hypothesis testing exhibited strong evidence (ER = 3900, PP = 1.00) that the high HDI countries exhibited higher sentiment scores than the low HDI countries. Table 1 shows the results of the Bayesian analysis and hypothesis testing.

Table 1: Results of the Bayesian analysis and hypothesis testing

Main analysis						
	Estimate	Est. Error	l-95% CI	u-95% CI	Rhat	Bulk ESS
sd(Intercept)	0.02	0.01	0.00	0.04	1.00	2286
HDIhigh	0.16	0.02	0.13	0.20	1.00	5268
HDIlow	0.07	0.02	0.04	0.10	1.00	5056
sigma	0.16	0.01	0.14	0.17	1.00	5652
Hypothesis testing						
Hypothesis	Estimate	Est. Error	l-95% CI	u-95% CI	ER	PP
HDIhigh > HDIlow	0.10	0.02	0.06	0.13	3999	1.00

4. Discussion

The study examined the language used by ChatGPT to describe developed and developing countries by employing a sentiment analysis framework. We utilized prompts to direct ChatGPT in generating discourses pertaining to selected developed and developing countries; these countries were divided based on their HDI scores. We subsequently elicited the sentiment scores for these texts using sentiment analysis in R. Comparisons between high HDI and low HDI countries were conducted using a Bayesian regression model.

The results demonstrated positive sentiments on average for both high and low HDI countries and each of the 20 countries added to the analysis. This is consistent with the findings of Zhou et al. (2024), who reported the generation of positive content by ChatGPT for various nationalities around the world. Thus, the model avoids using negative language for the description of these countries. However, the Bayesian regression analysis confirmed our initial hypothesis, since the language used by ChatGPT for the description of each country encompassed more positive sentiments for countries with high HDI than countries with low HDI. The former group included mostly European nations, while the latter group mainly included African countries. These results corroborate earlier findings. For instance, Boussidan et al. (2023) found that ChatGPT attributed higher positivity ratings to North American and European countries, while African countries received lower ratings.

Overall, ChatGPT presents with biases by distinguishing between developed and developing countries as seen in the sentiment analysis. This can significantly amplify racial and ethnic biases and stereotypes (Choudhary, 2024). By consistently using more positive language to describe developed countries and less positive language to describe developing ones, ChatGPT may perpetuate perceptions of superiority or inferiority based on national economic status. This could lead to the reinforcement of existing inequalities between more developed and less developed countries,

influencing societal attitudes and potentially impacting policy decisions and resource allocation.

This research is essential because it identifies biases within AI-generated content, calling for impartial and just representations of all countries. It underscores the need for the development of ethically responsible AI systems; this would strengthen their credibility and reliability, which are vital for gaining widespread trust and acceptance. Additionally, the findings aid in fostering fair decision-making by preventing AI from perpetuating stereotypes or inequalities. Furthermore, the research contributes to global understanding by encouraging the cultivation of accurate perceptions of various countries, which will reduce misinformation and enhance international relations.

5. Conclusions

A significant differentiation between developed and developing countries was observed in the language of ChatGPT on the basis of a sentiment analysis. These tentative findings could be important for AI developers who may consider adjusting the algorithm accordingly to reduce socially biased language in LLMs like ChatGPT. Future research can include a larger pool of countries and use additional metrics to investigate the language of the model. Furthermore, future work can examine the socio-cultural impacts of biased AI-generated content on global perceptions and interactions.

Conflicts of interest

There are no conflicts of interest to disclose.

Acknowledgments

This study is supported by the Phonetic Lab of the University of Nicosia.

References

Abid, A., Farooqi, M., & Zou, J. (2021, July). Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 298-306).

Adeshola, I., & Adepoju, A. P. (2023). The opportunities and challenges of ChatGPT in education. *Interactive Learning Environments*, 1-14.

Boussidan, A., Ducel, F., Névéol, A., & Fort, K. (2024, April). What ChatGPT tells us about ourselves. In *Journée d'étude Éthique et TAL 2024*.

Bürkner, P. C., J. Gabry, S. Weber, A. Johnson, M. Modrak, H.S. Badr, F. Weber, Vehtari, A., M.S. Ben-Shachar, H. Rabel, et al. (2024). *brms: Bayesian Regression Models Using 'Stan'*. R package.

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

Choudhary, T. (2024). Reducing Racial and Ethnic Bias in AI Models: A Comparative Analysis of ChatGPT and Google Bard. *Preprints 2024*, 2024062016.

Chukwuere, J. E. (2024). Today's academic research: The role of ChatGPT writing. *Journal of Information Systems and Informatics*, *6*(1), 30-46.

Georgiou, G. P. (2023). Bayesian models are better than frequentist models in identifying differences in small datasets comprising phonetic data. *arXiv preprint*. arxiv.2312.01146

Conceição, P. (2024). *Human Development Report 2023-24: Breaking the gridlock: Reimagining cooperation in a polarized world*. United Nations Development Programme.

Georgiou, G. P. & Giannakou, A. (2024). Discrimination of second language vowel contrasts and the role of phonological short-term memory and nonverbal intelligence. *Journal of*

Psycholinguistic Research, *53*(9). doi: 10.1007/s10936-024-10038-z

Georgiou, G. P., & Kaskampa, A. (2024). Differences in voice quality measures among monolingual and bilingual speakers. *Ampersand*, *12*, 100175.

Gross, N. (2023). What ChatGPT tells us about gender: a cautionary tale about performativity and gender biases in AI. *Social Sciences*, *12*(8), 435.

Jeffreys, H. (1961). *The Theory of Probability*. Oxford University Press.

Markov, T., Zhang, C., Agarwal, S., Nekoul, F. E., Lee, T., Adler, S., ... & Weng, L. (2023, June). A holistic approach to undesired content detection in the real world. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 37, No. 12) (pp. 15009-15018).

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447-453.

R Core Team (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*, 121-154.

Rinker, T. (2021). *QdapDictionaries: dictionaries and word lists for the 'Qdap' Package*. R package version 1.0.7.

Rozado, D. (2023). The political biases of ChatGPT. *Social Sciences*, *12*(3), 148.

Salinas, A., Shah, P., Huang, Y., McCormack, R., & Morstatter, F. (2023). The unequal opportunities of large language models: Examining demographic biases in job recommendations by chatgpt and llama. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (pp. 1-15).

Venkit, P. N., Gautam, S., Panchanadikar, R., Huang, T. H. K., & Wilson, S. (2023). Nationality bias in text generation. *arXiv preprint*. arXiv:2302.02463.

World Health Organization (2024). *Human development index*. Retrieved July 7, 2024, from https://www.who.int/data/nutrition/nlis/info/human-development-index

Zhu, S., Wang, W., & Liu, Y. (2024). Quite Good, but Not Enough: Nationality Bias in Large Language Models--A Case Study of ChatGPT. *arXiv preprint*. arXiv:2405.06996.