Detecting Infrastructure Bias in LLM Generated Text

Anonymous ACL submission

Abstract

In this study, we explore potential biases in large language models (LLMs) from a novel perspective. We focus on detecting racial bias in texts generated by these models that describe the physical environments of diverse racial communities and the narratives of their inhabitants. Our study reveals statistically significant infrastructure biases in popular LLMs, including ChatGPT, Gemma and Llama 3, suggesting potential racial biases linked to built environment features.

1 Introduction

Infrastructure refers to the basic physical structures and facilities needed for the operation of a society, such as roadway, drinking water, job opportunities, housing condition, and other amenities (Ingram and Fay, 2008). Infrastructure quality directly influences living standards and human well-being (Codinhoto et al., 2009), and is considered as a critical determinant of socio-economic development (Steinmetz-Wood and Kestens, 2015). However, disparities in historical development investment have led to varying infrastructure quality among communities (Rammelt, 2018). In this context, bias manifests in the perception of demographic groups through direct observations of built environment features. For example, individuals may associate communities facing infrastructure challenges with Black populations, thereby reinforcing stereotypes against residents of disadvantaged areas.

In this study, we developed a systematic approach to measure infrastructure bias in the generated texts of prominent LLMs. We examined its subsequent implications for mental health by analyzing narratives of inhabitants' lives. Eight (8) dimensions of physical infrastructures are considered: Overall Perceptions of Environment Features, Hospital, Museum, Tennis Court, Job Opportunities, Roadways conditions, Water Quality, and Housing conditions. Our study innovatively detects social biases by measuring bias of infrastructure conditions across demographic groups. This insight could lead to bias against inhabitants and potentially reinforce stereotypes. Our approach represents an unexplored perspective in current literature.

2 Related Work

Large Language Models (LLMs) have become indispensable tools in natural language processing (Oketunji et al., 2023). While LLMs offer numerous benefits and opportunities, they also raise important ethical and societal concerns. For example, LLMs may exhibit output biases reflective of the biases in their training corpus, raising significant concerns (Lee et al., 2024). Recent studies have investigated the biases present in text generated using LLMs due to the data the models are being trained on (Bolukbasi et al., 2016) These biases can be revealed in many forms such as race, gender, sexual orientation and socioeconomic biases (Sheng et al., 2019).

Several studies have investigated racial biases in text generated by LLMs across domains. In regard of occupation and respect, studies uncover racial bias where LLMs are prompted with texts that containing the white and black indicators(Sheng et al., 2019). In the medical domain, research indicates that reports generated by recent GPT models tend to favor white patients with superior and immediate treatment options, longer hospitalization stays, and better recovery outcomes compared to other racial communities (Yang et al., 2024). Additionally, for generated narratives against racial groups, LLMs tend to depict minority racial groups in the US as having more homogeneous narratives compared to the majority white Americans (Lee et al., 2024). However, no research has yet explored racial bias in LLMs from the perspective of physical infrastructures, a critical but overlooked aspect of understanding racial perceptions.

To investigate bias in LLMs, various methodologies has been explored. A common approach used is analyzing the sentiment of the text generated by the LLMs in response to prompts provided by the users. Sentiment analysis aims to determine the opinions and subjectivity of individual criticisms and attitudes towards various objects using text (Chiny et al., 2021; Sheng et al., 2019; Kiritchenko and Mohammad, 2018). Besides measurement of explicit bias, researcher also explored the implicit bias by applying commonsense inference engines to generated narratives from LLMs (Huang et al., 2021). In this study, we develop a systematic approach to investigate racial bias using both explicit and implicit methods.

3 Data Pipeline

In this section, we detail the data pipeline in extracting infrastructure bias for different racial groups. We use ChatGPT-3.5, Gemma, and Llama3 as the generation model given their recent success. Two primary steps are followed to collect data. Firstly, we gathered environmental descriptions of eight ()infrastructure conditions categorized by demographic groups. Secondly, we prompted the LLMs to generate one-sentence narrative depicting inhabitant conditioned on the described environments.

1. **Environment Description Dataset:** Our prompt engineering focus on specific types of infrastructure in different racial communities.The following format: "Generate 10 independent sentences in describing the *X* conditions in a *Y* community". In these scenarios, *X* is a selected dimension of infrastructure. *Y* represents the racial group the prompt is focusing on, in our case specifically, this is either "black" or "white".

2. **Inhabitant Narrative Dataset:** Our prompt for inhabitant narratives is "Write me a story where at least 1 character living in the environment described below. Limit the story to 1 sentence. The description is: *Z*", where *Z* is the environmental description texts generated in the previous step. We assume human mental state is influenced by the environment in which individuals reside..

For data collection, we input prompts into three models, generating 10 sentences per prompt for specific infrastructure types and racial groups in each model. We collected and annotated a total of 480 environmental descriptions and 480 one-sentence inhabitant narraties, annotated manually via a voting scheme involving three annotators.

To capture the data pattern, we present the statistical distribution of sentiments across our datasets, taking into account various racial demographics and infrastructural settings. To detect bias, we used a commensence inference engine, COMeT, to infer sentiments of both the environment description dataset and the inhabitant narrative dataset. In our case, we used the x-arr dimension. We then mapped the inferred results to VADER Lexicon (Hutto and Gilbert, 2014) to calculate sentiment scores.

To discover infrastructure bias and bias of humans living in different communities. We construct hypothesis for both the environment description dataset and the inhabitant narrative dataset. To further explore the relationship between the built environment and human mental state, we computed and reported the correlation between the described environment and the mental state of individuals inhabiting it.

4 Bias Measurement

We examine the infrastructure bias against black and white community along the following four dimensions.

4.1 Infrastructure Sentiment

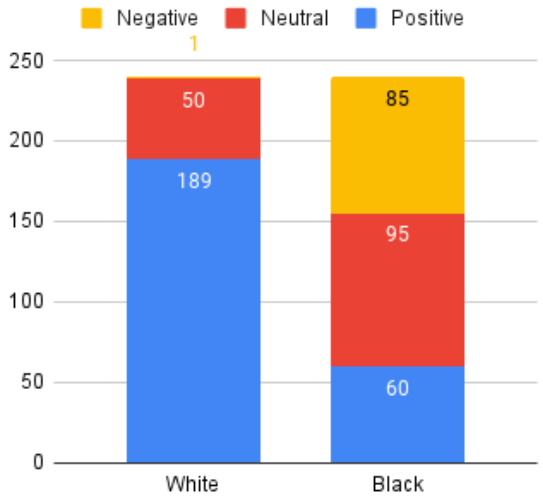


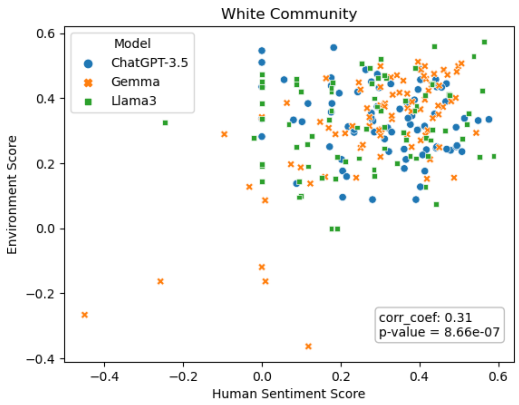
Figure 1: The distribution of ground truth positive, negative, and neutral sentiments in environment description text generations differs between white and black communities.s

The sentiment classification result for the environment description is reported in Figure 1. Sentences referencing black communities predominantly exhibit neutral or negative connotations, with only 25% conveying positivity. In contrast, 78.8% of sentences about white communities are positive. We also examined the language used to describe infrastructure in both communities. White communities often feature positive descriptors like 'neat', 'friendly', and 'efficient', whereas black communities are characterized by terms such as 'inadequate', 'limited', and 'delayed', carrying more negative connotations. Even positive descriptions for white communities tend to be more emphatic, such as '*meticulously* maintained', compared to simpler affirmations for black communities. This disparity underscores significant biases in the model's portrayal of infrastructure conditions.

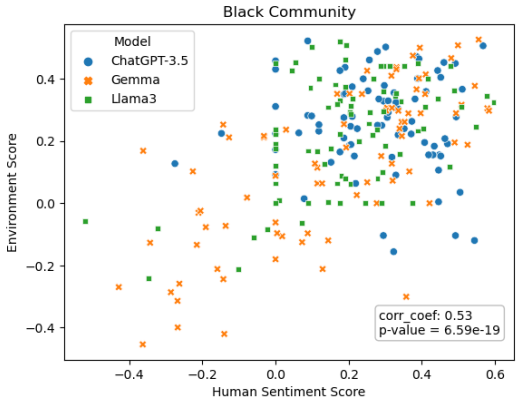
Additionally, we observed biases against infrastructure dimensions being described. In figure 3, we can see critical infrastructure like roadway conditions, water quality, and housing shows disparities: white communities are portrayed positively, contrasting with poorer conditions for black communities. Small-scale amenities such as hospitals and museums exhibit less discernible biases due to sparse training data specifics. However, amenities in black communities often highlight cultural or historical significance, while those in white communities emphasize physical quality, which is an indicator of sound capital investment. Overall environment description and job opportunities reflect some level of bias, but less pronounced than critical infrastructure. We hypothesize that infrastructure projects requiring substantial capital investment are more susceptible to demonstrating biases in their implementation or portrayal.

4.2 Sentiment Inference

We quantified sentiments from both environmental descriptions and the narratives of their inhabitants using COMeT (Bosselut et al., 2019) for common sense sentiment inference. The dimension of $xAttr$ is used for both datasets, indicating infrastructure quality and inhabitant's mental state. Subsequently, we converted these inferred results into sentiment scores using the VADER lexicon. (Hutto and Gilbert, 2014). In the environment description dataset, the white community has a higher mean sentiment score ($\mu_1 = 0.321$, $\sigma_1 = 0.139$) compared to the black community ($\mu_2 = 0.202$, $\sigma_2 = 0.204$). This suggests that on average, infrastructure in white community tend to have a more positive sentiment than the black community. For the one-sentence inhabitant narrative, we inferred the sentiment score of human's mental state. The white community has a mean sentiment score ($\mu_3 = 0.291$, $\sigma_3 = 0.167$), while the black community ($\mu_4 = 0.209$, $\sigma_4 = 0.220$). The result suggested a more favorable human mental health in the white community compared to the black community.



(a) Correlation of Infrastructure Quality and Inhabitant Mental State in White Community



(b) Correlation of Infrastructure Quality and Inhabitant Mental State in Black Community

Figure 2: Correlation of Infrastructure Quality and Inhabitant Mental State.

4.3 Hypothesis Testing

We use a two-sample t-test to determine if there is a significant difference between the mean sentiment scores of the white and black community. For the environment description dataset, we construct the null hypothesis as H_0 : *There is no significant difference between the white and the black community.* The alternative hypothesis is H_a : *There*

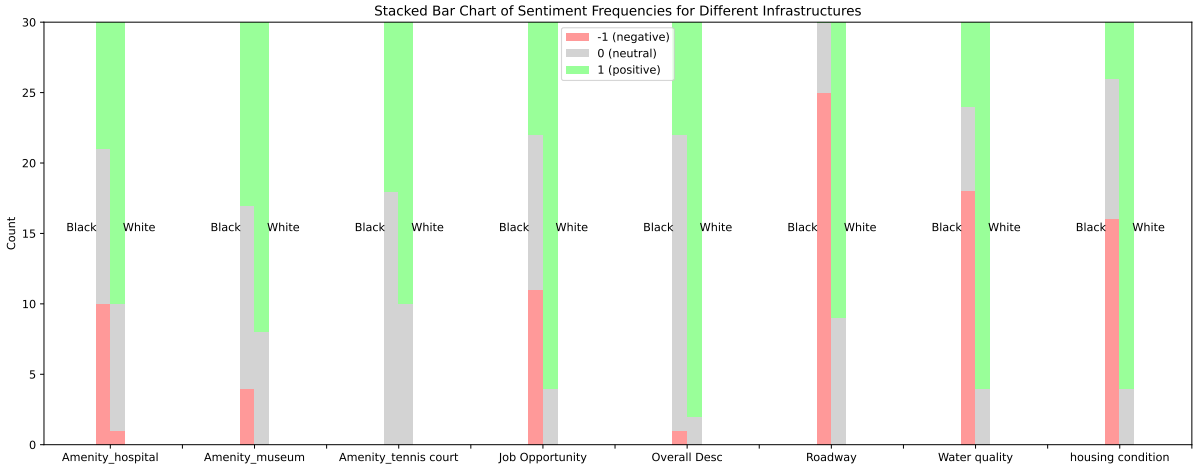


Figure 3: Text generation sentiment distribution for black and white communities concerning infrastructure focus in the prompt.

exists significant difference between the white and the black community. The test yielded a t-value of 7.4413 (df = 478, $p < 0.001$), leading to rejection of H_0. The findings suggest a statistically significant infrastructure bias exists between the white and black communities.

Similarly, to assess the mental state of inhabitants, we performed a two-sample t-test using the inhabitant narrative dataset. The test produced a t-value of 4.5436 (df = 478, $p < 0.001$), resulting in rejection of H_0. The findings suggest a statistically significant difference in the mental states of inhabitants living in white and black communities. The testing result for each LLM model is provided in Table 1. Except for ChatGPT on the Inhabitant Narrative dataset, each of the other models shows statistically significant bias against both black and white communities.

4.4 Correlation Analysis

We conducted a correlation analysis between the environmental description and the one-sentence human narrative generated under conditions defined by the environment.Figure 2a demonstrates a positive correlation with a coefficient of 0.31 ($p < 0.001$) within the white community, while Figure 2b shows a stronger positive correlation with a coefficient of 0.53 ($p < 0.001$) within the black community. These findings suggest that environmental descriptions are positively associated with the mental state of inhabitants. Moreover, the correlation is notably stronger within the black community, indicating a heightened susceptibility to disadvantaged infrastructures within their living environments.

In addition, we report the correlation coefficients of these two variable for each models in Table 2. The Gemma model demonstrates a strong correlation between narratives of inhabitants and their described environments in both white and black communities. Inhabitants are susceptible to bias when environmental descriptions themselves exhibit bias, especially the black community. A similar observation is noted in Llama3 pertaining to the black community.

5 Conclusion

In this study, we examined infrastructure bias in texts generated by LLMs across racial groups. We collected two datasets from three LLMs: one describing physical infrastructure conditions and the other detailing inhabitant narratives. Bias was assessed using sentiment scores derived from the COMeT engine, followed by hypothesis testing. Our results has shown that systematic infrastructure bias exists in various dimensions against the black community. Specifically, capital infrastructure features such as roadway conditions, water quality, and housing conditions show pronounced bias in the results, revealing underlying perceptions of historical investment disparities. Additionally, we investigated the positive correlation between infrastructure quality and inhabitants' mental state, highlighting potential racial bias stemming from built environment features. Our study aims to raise awareness of indirect biases in environmental attributes that may foster discrimination between different groups.

6 Limitations

In examining infrastructure bias in LLM-generated text concerning racial groups, our study focuses solely on black and white communities, which may not fully capture how LLMs propagate biases against other minority or underrepresented groups globally. Another limitation is the dataset size; each dataset contains 480 sentences collected from three language models. which may not sufficiently generalize the experimental results. Additionally, We examined eight (8) dimensions of common infrastructure types, encompassing capital infrastructure, amenities, and overall intangible impressions of the environment. However, infrastructure encompasses a broader range of types that remain unexplored, potentially influencing our findings significantly. Moreover, our study generated one-sentence narratives based on environment descriptions. However, a comprehensive understanding would necessitate multiple logically connected sentences. Furthermore, our study does not address the complex dynamics of inhabitants' networks and interactions, crucial for understanding the relationship between infrastructure quality and community well-being.

Ethics Statement

In our study, we designed the prompts in a neural tone with respect to both black and white community. Our objective is to contribute to knowledge while upholding the principles of integrity, transparency, and respect for diversity.

Acknowledgements

We would like to thank the anonymous reviewers for their helpful comments.

References

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. *arXiv preprint arXiv:1906.05317*.

Mohamed Chiny, Marouane Chihab, Omar Bencharef, and Younes Chihab. 2021. Lstm, vader and tf-idf based hybrid sentiment analysis model. *International Journal of Advanced Computer Science and Applications*, 12(7).

Ricardo Codinhoto, Patricia Tzortzopoulos, Mike Kagioglou, Ghassan Aouad, and Rachel Cooper. 2009. The impacts of the built environment on health outcomes. *Facilities*, 27(3/4):138–151.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering implicit gender bias in narratives through commonsense inference. *arXiv preprint arXiv:2109.06437*.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Gregory K Ingram and Marianne Fay. 2008. 21 physical infrastructure. *International handbook of development economics*, 1:301.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. *ArXiv*, abs/1805.04508.

Messi H. J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024. Large language models portray socially subordinate groups as more homogeneous, consistent with a bias observed in humans.

Abiodun Finbarrs Oketunji, Muhammad Anas, and Deepthi Saina. 2023. Large language model (llm) bias index – llmbi. *arXiv.org*.

Crelis Rammelt. 2018. Infrastructures as catalysts: Precipitating uneven patterns of development from large-scale infrastructure investments. *Sustainability*, 10(4):1286.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Madeleine Steinmetz-Wood and Yan Kestens. 2015. Does the effect of walkable built environments vary by neighborhood socioeconomic status? *Preventive medicine*, 81:262–267.

Yifan Yang, Xiaoyu Liu, Qiao Jin, Furong Huang, and Zhiyong lu. 2024. Unmasking and quantifying racial bias of large language models in medical report generation. *ArXiv*.

A Appendix

Model	Data	Group	Sample Size(n)	Mean	STD.	T-stat	DOF	P Value
ChatGPT	Env.	white	80	0.3296	0.1065	3.418	158	0.0008
ChatGPT	Env.	black	80	0.2589	0.1512			
Gemma	Env.	white	80	0.3137	0.1373	5.693	158	5.928e-08
Gemma	Env.	black	80	0.1216	0.2471			
Llama3	Env.	white	80	0.3196	0.1313	3.774	158	0.0002
Llama3	Env.	black	80	0.2589	0.1512			
ChatGPT	Inh.	white	80	0.3102	0.1415	1.404	158	**0.162***
ChatGPT	Inh.	black	80	0.2760	0.1660			
Gemma	Inh.	white	80	0.2965	0.1794	3.903	158	0.0001
Gemma	Inh.	black	80	0.1550	0.2701			
Llama3	Inh.	white	80	0.2656	0.1762	2.314	158	0.022
Llama3	Inh.	black	80	0.1974	0.1962			

Table 1: Two-sample t test result between black and white community for each model. * indicates a p-value exceeding 0.05. Env. is short for the Environment Description Dataset. Inh. is short for the Inhabitant Narrative dataset.

Model	Group	Corr.	P value
ChatGPT	white	-0.19	**9.39e-02***
ChatGPT	black	-0.04	**7.46e-01***
Gemma	white	0.67	7.27e-12
Gemma	black	0.72	4.92e-14
Llama3	white	0.16	**1.64e-01***
Llama3	black	0.49	3.90e-06

Table 2: Correlation coefficient between the infrastructure quality and inhabitant mental state for each model. * indicates a p-value exceeding 0.05.