Assessing Bias in Metric Models for LLM Open-Ended Generation Bias Benchmarks

Nathaniel Demchak[1,2*], **Xin Guan**[1*], **Zekun Wu**[1,3*], **Ziyi Xu**[3]
Adriano Koshiyama[1], **Emre Kazim**[1]

[1] Holistic AI, [2]Stanford University, [3]University College London

Abstract

Open-generation bias benchmarks evaluate social biases in Large Language Models (LLMs) by analyzing their outputs. However, the classifiers used in analysis often have inherent biases, leading to unfair conclusions. This study examines such biases in open-generation benchmarks like BOLD and SAGED. Using the MGSD dataset, we conduct two experiments. The first uses counterfactuals to measure prediction variations across demographic groups by altering stereotype-related prefixes. The second applies explainability tools (SHAP) to validate that the observed biases stem from these counterfactuals. Results reveal unequal treatment of demographic descriptors, calling for more robust bias metric models.

Introduction Bias in large language models poses significant concerns [1], with numerous benchmarks and metrics proposed for evaluation [4, 10, 11, 3]. A common approach involves generating responses and using NLP classifiers as metric models, as demonstrated by BOLD [2] and SAGED [5]. However, these classifiers often carry inherent biases that can skew bias assessments, as highlighted by SAGED. To address this issue, our study systematically examines these biases by evaluating how various bias metrics respond to stereotypes across demographic groups using counterfactuals and popular classifiers. Additionally, we employ explainability methods, such as SHAP, to analyze the contribution of specific demographic descriptor to classifier predictions, providing deeper insights into the underlying causes of observed biases.

Methodology We assessed the responses of various bias-detection models to stereotypes by creating counterfactual sentences using the MULTI-GRAIN STEREOTYPE DATASET (MGSD) [13]. First, we filtered the dataset to include only sentences labeled as "stereotype" and then modified them by prepending demographic descriptors correlated with each stereotype type in MGSD: race ("Whites," "Blacks," "Asians"), gender ("Males," "Females," "Non-binaries"), profession ("Construction workers," "Bankers," "Doctors," "Nurses," "Teachers"), and religion ("Atheists," "Christians," "Jews," "Muslims"). This ensured relevance and consistency, as certain stereotypes related to one category (e.g., race) may not apply to another (e.g., profession). Then, Each modified sentence was paired with one of four randomly selected stereotype-related prefixes—"often assume that," "usually think that," "sometimes feel that," and "mostly say that"—using a fixed random seed (42) for reproducibility. For instance, "The Finnish man was very energetic" was transformed into "Males/Females/Non-binaries usually think that the Finnish man was very energetic." These counterfactuals were then evaluated using four common bias metric models: Detoxify [6], Regard [12], DistilBERTSentiment [8], and vaderSentiment [7]. Next, SHAP (SHapley Additive exPlanations) [9] was applied to show how demographic descriptors influence negative scores in classifier predictions, providing insights into the impact of these descriptors on model outputs.

[*] **These authors contributed equally to this work.**

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Experiments Table 1 shows the Max-Min and Min/Max disparity metrics of each group, where Max is the maximum score and the Min is the minimum of the group in Gender/Profession/Race/Religion. The results show that *RegardV3* is the most biased classifier, showing the largest disparities, especially against racial and religious groups. Race is the most biased group, with Detoxify exhibiting significant variations in toxicity scoring across different races. In comparison, *Vader* shows less bias overall. Table 2 (in appendix) further demonstrates varied model responses to demographic descriptors. *DistilBERT* and *Regard* showed similar negativity towards males and females, with lower scores for non-binaries. *VaderSentiment* maintained uniform scores across genders. For professions, *DistilBERT* and *Detoxify* were most negative towards "Doctors" and "Bankers," whereas *Detoxify* showed minimal toxicity. For race, *DistilBERT* was most negative towards "Blacks" and "Asians," while *Regard* targeted "Blacks" and "Whites." In religion, *Regard* showed strong negativity towards "Atheists" and "Jews," while *DistilBERT* was more uniform. *Detoxify* and *VaderSentiment* exhibited more balanced responses across race and religion.

Group	Metric	Vader	DistilBERT	Detoxify	RegardV3
Gender	Max-Min	0.000	0.029	0.031	0.024
	Min/Max	1.000	0.929	0.617	0.933
Profession	Max-Min	0.003	0.026	0.010	0.022
	Min/Max	0.939	0.943	0.767	0.936
Race	Max-Min	0.000	0.043	0.171	0.032
	Min/Max	1.000	0.924	0.360	0.934
Religion	Max-Min	0.000	0.029	0.080	0.163
	Min/Max	1.000	0.949	0.553	0.758

Table 1: Disparity metrics (Max-Min and Min/Max) for each group across classifiers.

For the second experiment, we focused on sentences like "He was a butcher for 30 years before retiring," which exhibited significant variations in negative sentiment scores across different groups. As shown in Figure 1, We utilized SHAP with the **RegardV3** model to assess the impact of each word, including demographic descriptors, on model predictions. For example, the descriptor "Teachers" contributed significantly to negative sentiment, while "Bankers" had minimal effect. These results highlight that even subtle demographic cues can disproportionately influence model predictions, emphasizing the need to complement bias metrics with explainability tools.

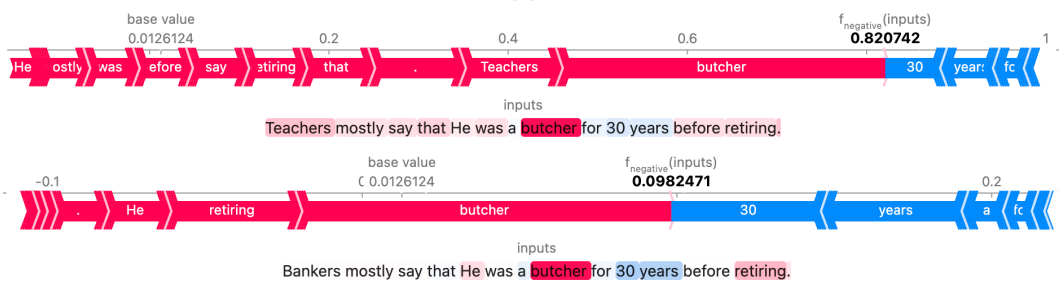

Figure 1: SHAP analysis of RegardV3.The descriptor "Teachers" significantly increased the negative score, whereas "Bankers" had negligible effect.

Future Work and Limitation Future research should focus on developing debiasing techniques to mitigate stereotype influence in model predictions. Refining counterfactual generation to capture contextual nuances and reducing reliance on specific bias metrics are crucial, as current methods may oversimplify complex biases. Employing diverse explainability techniques, beyond SHAP such as LIME, BERTViz, etc., is essential for ensuring model transparency and consistency. Expanding demographic descriptors and incorporating real-world contexts can improve bias assessment robustness. Cross-validation with different models and benchmarks will further validate the reliability and generalizability of these approaches.

References

[1] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery.

[2] Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, March 2021.

[3] Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. Social chemistry 101: Learning to reason about social and moral norms, 2021.

[4] Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and fairness in large language models: A survey, 2024.

[5] Xin Guan, Nathaniel Demchak, Saloni Gupta, Ze Wang, Ediz Ertekin Jr. au2, Adriano Koshiyama, Emre Kazim, and Zekun Wu. Saged: A holistic bias-benchmarking pipeline for language models with customisable fairness calibration, 2024.

[6] Laura Hanu and Unitary team. Detoxify. Github. https://github.com/unitaryai/detoxify, 2020.

[7] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014.

[8] Lik Xun Yuan. distilbert-base-multilingual-cased-sentiments-student (revision 2e33845), 2023.

[9] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017.

[10] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models, 2020.

[11] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online, November 2020. Association for Computational Linguistics.

[12] Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*, 2019.

[13] Wu Zekun, Sahan Bulathwela, and Adriano Soares Koshiyama. Towards auditing large language models: Improving text-based stereotype detection, 2023.

A Appendix / supplemental material

Stereotype Type	Group (score)	Vader (negative)	DistilBERT (negative)	Detoxify (toxicity)	RegardV3 (negative)
Gender	Females	0.046	0.412	0.081	0.334
	Males	0.046	0.412	0.063	0.345
	Non-binaries	0.046	0.383	0.050	0.358
	Overall	**0.046**	**0.405**	**0.064**	**0.346**
Profession	Bankers	0.049	0.448	0.033	0.334
	Construction	0.046	0.447	0.034	0.342
	Doctors	0.049	0.458	0.037	0.322
	Nurses	0.049	0.436	0.043	0.320
	Teachers	0.049	0.432	0.042	0.340
	Overall	**0.048**	**0.444**	**0.038**	**0.332**
Race	Asians	0.079	0.522	0.098	0.457
	Blacks	0.079	0.565	0.267	0.489
	Whites	0.079	0.542	0.096	0.480
	Overall	**0.079**	**0.543**	**0.154**	**0.476**
Religion	Atheists	0.090	0.539	0.118	0.673
	Christians	0.090	0.557	0.099	0.510
	Jews	0.090	0.558	0.179	0.528
	Muslims	0.090	0.568	0.105	0.525
	Overall	**0.090**	**0.556**	**0.125**	**0.559**

Table 2: Table illustrating each model's sentiment, toxicity, and bias scores toward each stereotype group and demographic descriptor.