ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2.
Available from http://www.i6doc.com/en/.

A Two-Stage Approach for Implicit Bias Detection in Generative Language Models

Jeremy Edwards[1], Renjie Hu[1], Amaury Lendasse[2], Alexander Schlager[3] and Peggy Lindner[1] *

1-University of Houston - Dept. of Information Science Technology
14004 University Boulevard, Sugar Land, TX 77479 - USA

2-Missouri University of Science & Technology
Dept. of Engineering Management and Systems Engineering
223 Engineering Management, 600 W. 14th St., Rolla, MO 65409, USA

3-AIceberg Inc
New York - USA

Abstract. Machine learning and AI are increasingly popular for their impressive task performance. Yet, Natural Language Processing (NLP) models often inadvertently learn harmful biases related to gender and race, leading to skewed predictions. Literature distinguishes between direct and indirect bias. Current research aims to detect and mitigate these biases in machine learning models. This study introduces a two-stage approach to identify both types of gender bias in generative large language models (LLMs), confirming that they can manifest both direct and indirect biases.

1 Introduction

AI models employ mathematical and algorithmic techniques to analyze data and make informed predictions. They are particularly adept at tasks such as text classification, machine translation, and text generation [1, 2]. However, while these models excel in performance optimization, they sometimes inadvertently capture latent biases that can disadvantage specific groups [3]. The emergence of generative language models like OpenAI's ChatGPT, and its competitors Falcon and Gemini, underscores the urgency of addressing the potential for these models to propagate bias. This challenge has become a vital area of research, focusing on both detection and mitigation of bias [4].

The structure of this paper is organized to thoroughly explore the detection of bias in LLM models. Section 2, "Background," lays the foundation by discussing key concepts including the importance of explainability in AI predictions, the prevalence of bias in large language models, approaches to evaluating machine-generated texts, and strategies for bias mitigation. "Methodology," outlined in Section 3, details the approaches and analyses used in this study. Section 4, "Experimental Results," presents the outcomes of the implemented tests and analyses, offering insights into the biases discovered and the effectiveness of mitigation strategies employed. This paper aims to bridge the gap between theoretical approaches and practical outcomes in the realm of AI-driven language processing.

*This research is supported by AIceberg Inc.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

2 Background

Insight into AI predictions is essential for their safe integration into daily life. A notable example involves a model trained to assess pneumonia risk, which learned from real data that asthma patients had lower mortality rates compared to healthy individuals [5]. In practice, though asthma worsens pneumonia recovery, strict medical supervision of asthmatic patients resulted in fewer deaths, leading the AI to produce misleading predictions. Such biases, if undetected in a black box system, are nearly impossible to correct. This underscores the importance of explainability in AI, which often mimics patterns from training data, inadvertently learning and perpetuating existing biases, whether social or political, in various applications like classification and text generation [6].

Racial, gender and political biases can manifest in the outputs of LLMs, including both their natural language responses and word embeddings. Embedders, deep learning models themselves, map words into vector spaces where similar words cluster together [7]. Studies like Feng et al.[8] have explored how fine-tuning LLMs on different data affects these biases. Their findings indicate significant shifts in the models' political stances, impacting tasks like hate speech detection and misinformation identification depending on the political leanings of the fine-tuning data. Continuing this line of inquiry, Jian et al.[9] investigate fine-tuning methods that deliberately bias LLMs to reflect specific community perspectives for social science research.

Various metrics have been developed to assess text generated by LLMs, evaluate the accuracy of word distribution learning and detect biases like gender, racial, and political biases. Probabilistic metrics such as Perplexity [10] and Jensen-Shannon Distance [11] are commonly used to compare the similarity between the word distributions of human-generated and machine-generated texts. For tasks like text generation, classification, and sentiment analysis, texts are broken into n-grams or tokens. These tokens pass through model layers, influencing millions or billions of weights to create "embeddings." These high-dimensional vectors represent natural language and are crucial for downstream tasks, serving as a basis for further analysis.

Research continues to explore methods for mitigating bias in language models, focusing on reducing bias through vector space and statistical techniques. Bolukbasi et al.[12] address bias mitigation by refining vector space representations of words and introducing a gender bias metric based on vector orientation. Bordia and Bowman[13] utilize probabilistic metrics centered around class imbalance to measure bias in numerical data and text. They calculate the probabilities of words appearing in male or female contexts to quantify bias. The approach measures bias in training data and assesses how biases are amplified or mitigated in the outputs of a Long Short-Term Memory (LSTM) model. Their debiasing technique minimizes the projection of embeddings towards defined gender directions, adjusting the intensity of bias reduction via a hyperparameter. This method effectively reduces gender bias in text by neutralizing the association of gendered terms with biased contexts, rather than merely reversing the biases.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

3 Methodology

This study introduces a two-stage methodology to detect bias in open-source, generative language models used as medical chatbots. Initially, the study probes for direct bias by analyzing model responses to specific medical cases. If direct bias is found, a further investigation into indirect bias is conducted. The models, including GPT-2, GPT-Medium, and OPT, are evaluated using medical case vignettes developed by Haider et al. [14], which were originally created to study implicit bias among nurses. Each model is presented with a vignette and a related question, responding on a Likert scale. This study focuses on gender bias, analyzing differences in responses across genders, though the methodology can be adapted to explore other biases.

3.1 Direct Bias Testing

To assess direct bias, the study manipulates clinical vignettes by specifying the gender and race of the patient. A corresponding question is then added to the vignette, serving as a prompt for the tested LLM. The models respond with numerical ratings, which are analyzed to see if there are differences in how male and female patients of the same race are judged. For example, a prompt for the first vignette looks as follows: "A 48-year-old [**race**] [**man/woman**] presents to the emergency department with abdominal pain... It is possible that the [**race**] [**man/woman**] is exaggerating [**his/her**] pain level. On a scale of 1 to 5, where 1 means definitely not exaggerating and 5 means definitely exaggerating, I rate it as...".

3.2 Indirect Bias Testing

If direct bias is detected, a second probing evaluates indirect bias. In this stage, the vignettes include the patient's race and occupation, but gender is not explicitly stated, although gendered pronouns like "him" and "her" are used. An appended question and the original description of the case serve as a prompt for the language model. Responses are again assessed numerically to examine biases. Specifically, this analysis compares responses for patients described with jobs typically dominated by the opposite gender. For instance, responses for male patients are compared against responses for patients described with male pronouns, such as secretaries, a role often female-dominated.

3.3 Hypothesis Testing

Statistical tests were conducted to ascertain whether observed trends and differences in the exploratory data analysis stemmed from random variations or actual biases. For direct bias, a two-sided Mann-Whitney U test from the SciPy Python package assessed the average ratings for case vignettes. To evaluate indirect bias, a one-sided Mann-Whitney U test checked if including a patient's occupation skewed the average ratings in a specific direction. Both tests were performed at the 10% significance level.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

4 Experimental Results

Table 1 shows the results of the direct bias experiments. GPT-2 exhibited a few statistically significant differences in ratings. For question 1 in the first vignette, both black and white male patients received higher ratings. For question 2 in the second vignette, Asian male patients received higher ratings than Asian female patients while White female patients received higher ratings than White male patients. OPT showed statistically significant differences for both questions in vignette 1. Black female patients received consistently higher ratings in response to question 1. For question 2, White female patients received consistently higher ratings. GPT-Medium only showed statistically significant variation for question 2 in the second vignette, with Asian female patients consistently receiving higher ratings than Asian male patients.

Model	Vignette	Question	Asian	Black	Hispanic	White
GPT-Medium	1	1	0.401	0.148	0.252	0.547
		2	0.604	0.94	0.779	0.772
	2	1	0.542	0.256	0.696	0.945
		2	**0.041**	0.294	0.565	0.198
OPT	1	1	0.495	**0.069**	0.797	0.938
		2	0.377	0.977	0.623	**0.053**
	2	1	0.691	0.964	0.438	0.327
		2	0.303	0.124	0.677	0.348
GPT-2	1	1	0.755	**0.092**	0.427	**0.036**
		2	0.352	0.367	0.853	0.737
	2	1	**0.001**	0.686	0.351	**0.005**
		2	0.747	0.233	0.46	0.664

Table 1: Stage 1 results. Direct Bias U-Scores for Vignettes 1 and 2 for all models. Statistically significant results marked as in bold.

Table 2 presents the results of indirect bias testing for GPT-2, OPT, and GPT-Medium. GPT-2 demonstrated indirect bias in the first vignette, where female patients in traditionally male-dominated jobs received ratings more aligned with those of males, and male patients in traditionally female-dominated jobs received ratings akin to those of females. OPT and GPT-Medium each showed indirect bias in one specific question: OPT in the first question of the first vignette and GPT-Medium in the first question of the second vignette.

Direct bias was detected in all models tested, with the favored gender varying by race - suggesting that gender bias in ratings is influenced by racial context. This indicates that direct bias is inherent in large language models (LLMs), reflecting deep learning's tendency to replicate learned patterns without discerning between beneficial and harmful correlations. For example, while demographic

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Model	Vignette	Question	Asian		Black		White	
			M	F	M	F	M	F
GPT-Medium	2	1	**0.014**	0.174	-	-	-	-
OPT	1	1	-	-	0.468	**0.019**	-	-
	2	1	-	-	-	-	0.500	0.715
GPT-2	1	1	-	-	0.475	**0.0004**	**0.043**	0.439
	2	1	0.183	**0.013**	-	-	**0.026**	0.479

Table 2: Stage 2 results. Indirect Bias U-Scores for all models. Statistically significant results marked as in bold.

variables like age, gender, and race might be relevant in medical diagnosis, they should not disproportionately influence a loan default prediction.

Indirect bias was also prevalent across all models. Even without explicit gender information, biases emerged through job titles, which, while not inherently gendered, carried strong gender associations within the models' embeddings. This underscores how LLMs inadvertently emphasize latent relationships that can lead to biased outcomes based on unprovided but inferred attributes such as gender, race, or age. Consequently, simple removal of bias-inducing information is ineffective as LLMs still manifest gender bias through these latent variables.

The study highlights the need to develop mitigation strategies that address direct and indirect biases. Techniques that merely adjust for direct bias, like retraining, are insufficient alone. Effective mitigation must also counteract the indirect biases arising from deep learning's modeling of latent variables, which can perpetuate inequalities even in the absence of explicit biased information. This dual approach is essential for crafting fairer generative LLMs.

5 Conclusions and Further Work

This study introduced a two-stage methodology to detect direct and indirect gender bias in generative language models used as clinical chatbots. The method employs clinical case vignettes, which are modified to vary the patient's race, gender, and occupation, and prompts models to provide numerical ratings. Initially tested for direct bias, models exhibiting such biases were further assessed for indirect bias using the Mann-Whitney U-Test to identify statistically significant differences. This approach revealed both types of biases in open-source models like GPT-2, GPT-Medium, and OPT when evaluating clinical scenarios.

Looking ahead, this methodology should be applied to various use cases across different fields where generative AI is employed, to comprehensively explore gender bias. Moreover, extending the tests to include a broader range of both open and closed source models is crucial. Given the extensive use of these models in industry, evaluating them for both direct and indirect bias is vital to ensure the fairness and justice of systems integrating generative language models.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

References

[1] Holger Schwenk, Daniel Dechelotte, and Jean-Luc Gauvain. Continuous Space Language Models for Statistical Machine Translation. *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 723–730, July 2006.

[2] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Pretrained Language Models for Text Generation: A Survey. *arXiv preprint:2201.05273*, 2022.

[3] Sandra Gabriel Mayson. Bias In, Bias Out. *128 Yale Law Journal 2218*, 2018. University of Georgia School of Law Legal Studies Research Paper No. 2018-35.

[4] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. Mitigating Political Bias in Language Models Through Reinforced Calibration. *arXiv preprint:2104.14795*, 2021.

[5] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1721–1730, Sydney NSW Australia, August 2015. ACM.

[6] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *California Law Review*, 104(3):671–732, 2016.

[7] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644, 2018.

[8] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. *arXiv preprint:2305.08283*, 2023. Publisher: arXiv Version Number: 3.

[9] Hang Jiang, Doug Beeferman, Brandon Roy, and Deb Roy. CommunityLM: Probing Partisan Worldviews from Language Models. *arXiv preprint arXiv:2209.07065*, 2022. Publisher: arXiv Version Number: 1.

[10] Alessio Miaschi, Dominique Brunato, Felice Dell'Orletta, and Giulia Venturi. What Makes My Model Perplexed? A Linguistic Investigation on Neural Language Models Perplexity. In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 40–47, Online, 2021. Association for Computational Linguistics.

[11] Frank Nielsen. On the Jensen-Shannon Symmetrization of Distances Relying on Abstract Means. *Entropy*, 21(5):485, May 2019.

[12] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv preprint:1607.06520*, 2016. Publisher: arXiv Version Number: 1.

[13] Shikha Bordia and Samuel R. Bowman. Identifying and reducing gender bias in word-level language models. In Sudipta Kar, Farah Nadeem, Laura Burdick, Greg Durrett, and Na-Rae Han, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[14] Adil H. Haider, Eric B. Schneider, N. Sriram, Valerie K. Scott, Sandra M. Swoboda, Cheryl K. Zogg, Nitasha Dhiman, Elliott R. Haut, David T. Efron, Peter J. Pronovost, Julie A. Freischlag, Pamela A. Lipsett, Edward E. Cornwell, Ellen J. MacKenzie, and Lisa A. Cooper. Unconscious Race and Class Biases among Registered Nurses: Vignette-Based Study Using Implicit Association Testing. *Journal of the American College of Surgeons*, 220(6):1077–1086e3, June 2015.