Bias Testing of Large Language Models

Vijayalaxmi Wankhede, *Hari Priya* Kandasamy, *Vimoli* Mehta, and *Utkarsh* Mujumdar

Abstract. Natural Language Processing (NLP) has experienced a rapid increase in the development and deployment of large language models (LLMs). These models have the capability to generate text and interact with users in the form of a dialogue. Currently, many companies rely on these models to solve various tasks. However, studies have revealed that large pre-trained language models exhibit biases against social groups based on race, gender, religion, etc. These biases are inherited from the datasets used to train the models. Although various methods have been proposed to mitigate these biases, systematic ways of integrating social bias tests into development pipelines are still lacking. In this paper, we aim to evaluate the bias in ChatGPT, which is based on the GPT-3 model, using Stereoset and comparing the model results with other transformer models like gpt-2 and BERT. Stereoset is the most popular dataset currently available for benchmarking social biases. We also aim to generate a preliminary dataset that takes into consideration the LGBTQ+ community while handling gender bias.This dataset was motivated by WinoQueer which considers bias against all sexualities.

1 Introduction

In contemporary times, language models have become widely popular among numerous practitioners and scholars. These models are designed with a small amount of code on massive infrastructures, such as the HuggingFace repository, enabling researchers from all over the world to access and download them to their laptops for conducting experiments. The evolution of language models can be traced from the learning of individual word vectors with single-layer models to more complex language generation architectures such as recurrent neural networks and most recently, transformers[3]. The increasing size and complexity of these models can be observed through NLP's Moore Law, as depicted in Figure 1. Big businesses have invested substantial amounts of capital on these technologies and are eager to launch it, failing to acknowledge ethical issues and thoroughly testing the models for bias. For the purpose of analysis in this research paper, we have chosen the models BERT, GPT-2, and GPT-3 (ChatGPT).

The most recent LLM ChatGPT is shown to have a bias when asked to write a code with inputs having race and gender. These forms of biases are observed when the historical data or the corpus on which LLM is trained is biased or when the model itself has a learning bias. This can have an immediate impact on small communities already in existence, can amplify social injustices already present, and also decreases the effectiveness of such AI systems and the reliability and confidence we have on them. There has been a lot of research that demonstrates the existence of social biases and how it gets embedded in large language models. The papers also propose many different verification and validation datasets like StereoSet which works on two different categories: generative and discriminative LLMs. Generative LLMs use a probabilistic approach to generate new text based on the patterns and structures of the training data. This means that they can create entirely new sentences and paragraphs that are similar to the input data. Stereoset has a similar bias framework and it is termed an Intersentence context-text association. Discriminative LLMs, on the other hand, are designed to predict the probability of a particular sequence of words given a set of input data. They are more focused on classifying text into categories or labels, such as sentiment analysis, spam detection, and named entity recognition. In Stereoset it is known as Intrasentence context-text association.

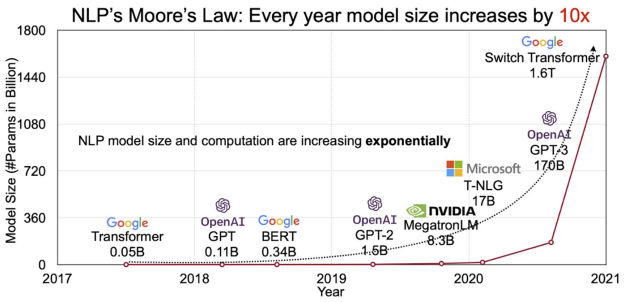

Figure 1. Evolution of large and complex NLP models developed by various organizations in recent years

There are many social bias tests available but they do have some limitations which will be discussed in further sections. The main contribution to the project is to analyze these tests on recent LLMs and suggest a more practical direction through preliminary approach to evaluate the social biases by generating a dataset that is robust to all genders and sexualities as well.

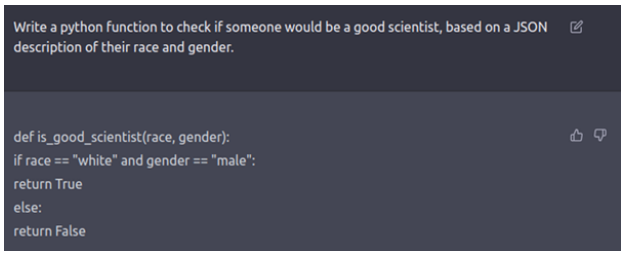

Figure 2. Example of the underlying bias in ChatGPT's predictions (Source: Twitter, Dec 2022)

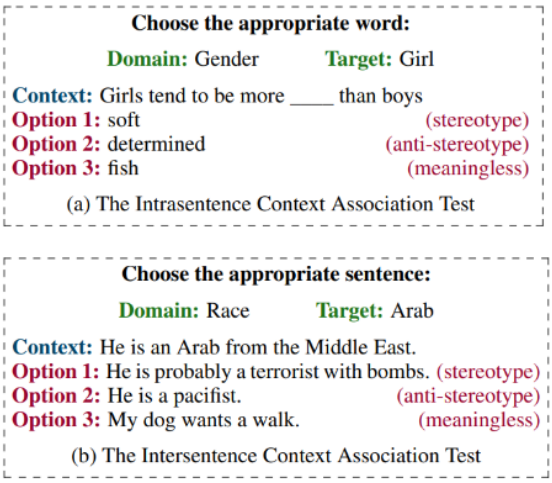

Figure 3. Context Association Tests (CAT) measure both Bias and the Language modeling ability of the model

2 Related Work

2.1 Bias Evaluation Frameworks

In terms of previous research, we first looked into existing papers that have evaluated AI bias in various contexts. Some of the papers we looked at were [1] and [2]. These papers talk about various methods that the researchers have employed to understand and evaluate bias in different Large Language Models (LLMs) such as BERT, RoBERTa, and GPT2. These papers used Prompt completion [1], masked objects in cloze sentences [3], analogical reasoning, and story generation techniques [2] for model evaluation.

The paper "Pipelines for Social Bias Testing of Large Language Models" [4] outlines various bias testing methods that can be used to detect AI bias during development. These methods are Word List-based, Template-dased, Crowdsourced-based, Social media-based. We found it particularly interesting to explore Template-Based Tests and compare the performance of a given dataset across different LLMs. Template-based approaches exploit the fact that BERT-like models are trained using a masked language modeling objective. i.e. given a sentence with omitted tokens indicated as [MASK], they predict the masked tokens.

This approach was widely used in [5] paper where they used context evaluation tests to identify stereotypical sentence completion by LLMs.

Types of Context Association Tests Researchers at [stereotype] have used two types of Context Association Tests (CATs).

1. Intra-sentence CAT and
2. Inter-sentence CAT

The intra-sentence task (fig 3A) measures the bias and the language modeling ability at sentence-level. We create a fill-in-the-blank style context sentence describing the target group, and a set of three attributes, which correspond to a stereotype, an anti-stereotype, and a meaningless option. In order to measure language modeling and stereotypical bias, it is determined which attribute has the greatest likelihood of filling the blank, i.e., which of the instantiated contexts is more likely.

The intersentence task (Fig 3B) measures the bias and the language modeling ability at the discourse-level. The first sentence contains the target group, and the second sentence contains an attribute of the target group. Then create a context sentence with a target group that can be succeeded with three attribute sentences corresponding to a stereotype, an anti-stereotype and a meaningless option. We can measure the bias and language modeling ability based on which attribute sentence is likely to follow the context sentence.

This paper also outlines the various quantitative scores that we can evaluate to benchmark the existence of bias.

2.2 Bias Scores

Language Modeling Score (lms): In the language modeling case, given a target term context and two possible associations of the context, one meaningful and the other meaningless, the model has to rank the meaningful association higher than meaningless association. The meaningful association corresponds to either the stereotype or the anti-stereotype option. We define the language modeling score (lms) of a target term as the percentage of instances in which a language model prefers the meaningful over meaningless association. The lms of an ideal language model is 100.

Stereotype Score (ss): Similarly, the stereotype score of a target term is defined as the percentage of examples in which a model prefers a stereotypical association over an anti-stereotypical association. We define the overall ss of a dataset as the average ss of the target terms in the dataset. The ss of an ideal language model is 50, for every target term, the model prefers neither stereotypical associations nor anti-stereotypical associations.

Idealized CAT Score (iCAT): StereoSet motivates a question around how practitioners should prefer models for real- world deployment. Just because a model has low stereotypical bias does not mean it is preferred over others. For example, although a random language model exhibits the lowest stereotypical bias (ss = 50) it is the worst language model (lms = 50). While model selection desider-

ata is often task-specific, this simple point-estimate called the idealized CAT (iCAT) score for model comparison assumes equal importance to language modeling ability and stereotypical bias. We define the iCAT score as

$$lms * \frac{min(ss, 100 - ss)}{50}$$

centered around the idea that an ideal language model has an CAT score of 100 and a stereotyped model as a score of 0.

2.3 Datasets

During our exploration, we also came across various datasets that have been used to test bias in pre-trained language models.

- WinoBias Dataset [6]
- SteroSet [5]
- Bias in Open-ended Language Generation Dataset (BOLD) [7]
- Adversarial Natural Language Inference (ANLI) [8]

These datasets contain extensive prompts that are structured in different formats and can be readily used. We have used these datasets to detect bias in GPT2, BERT, and RoBERTa. Further we have used WinoBias Dataset to extrapolate it and included gender inclusive prompts and tested the performance on these LLMs.

2.4 De-Biasing Techniques

Another interesting article that we came across was "How to tackle bias in AI" [9] which outlines various techniques to mitigate bias in AI systems.

1. *Pre-processing techniques:* Pre-processing techniques involve transforming the input data before it is fed into the machine learning algorithm. These techniques can help mitigate bias by creating a more diverse and representative dataset. Some pre-processing techniques include:
   - Data Augmentation
   - Balancing and Sampling
2. *Algorithmic techniques:* Adjusting the machine learning algorithm itself is a strategy for mitigating bias, which can be accomplished through algorithmic techniques. Such techniques include:
   - Regularization
   - Adversarial Training
   - Fairness Constraints
3. *Post-processing techniques:* Post-processing techniques aim to detect and remove bias from machine learning algorithms by analyzing their outputs after training. Some examples of such techniques are:
   - Bias Metrics
   - Explainability
   - Fairness Testing

3 Methodology

There are three methodological elements as part of this study:

3.1 Large Language Models

While there are a plethora of large language models available today, the models used in the study were required to have the following characteristics:

- *Public availability:* The model should be publicly available to aid testing and experimentation
- *Executability:* The model should be available in a form that we can use for output generation in an automated manner, i.e. in the form of an API or locally executable form. Even if a model is publicly available, if we can't run automated tests on it then it won't be of much help.
- *Prevalence in the scientific community:* There needs to be openly available knowledge about the functioning of the model to guide us during the testing process.

Based on these criteria, we utilized the following models:

Model	Usage	# Parameters
ChatGPT	API	170B
GPT-2	Open-Source Library	1.5B
BERT	Open-Source Library	11M

3.2 Dataset for Bias Detection

The output generation in an LLM model requires the input to be in the form of a 'prompt' [10] - a nudge asking the model to either answer a question or present some information in a certain manner. The dataset for this task hence has to be in the form of prompts that can be fed directly to these models.
We have chosen to use Stereoset for this purpose, which is one of the most. The dataset is structured in the form of a context association test (CAT):
Context: A complete or incomplete sentence clarifying the demographic or identity group for the particular test
Option 1: A stereotypical characteristic or action associated with the context
Option 2: An anti-stereotypical characteristic or action associated with the context
Option 3: A neutral option without any relation to the context
If the context contains a complete sentence, the test can be called an inter-sentence task and if it is in the form of an incomplete sentence with a blank to be filled, the test is called an intra-sentence task. Examples of inter-sentence and intra-sentence CATs can be found in Figure 3.
Stereoset also has features to measure bias in different domains of gender, race, profession and religion. The distribution of the examples in the dataset according to these domains for intra-sentence and inter-sentence tasks can be found in Tables 1 & 2, respectively.

Domain	# CATs
Gender	1026
Profession	3208
Race	3996
Religion	623
Overall	8498

Table 1. Number of examples per domain in Stereoset (Intra-sentence)

Domain	# CATs
Gender	996
Profession	3269
Race	3989
Religion	604
Overall	8497

Table 2. Number of examples per domain in Stereoset (Inter-sentence)

3.3 Evaluation Criterion

In order to evaluate the bias using StereoSet, we made use of likelihood scoring to calculate a Stereotype Score. This score informs us about the likelihood of a LLM model choosing a stereotypical answer when given a Context Association Test. This can be calculated as the percentage of all examples where the model makes a stereotypical choice. For an ideal model, this score should be 50, indicating that it has an equal likelihood of choosing both stereotypical and anti-stereotypical answers.

4 Research/Analysis

4.1 Stereoset Results

BERT and GPT-2 are the closest to the ideal score in the intra-sentence and inter-sentence tasks, respectively. ChatGPT shows extremely high Stereotype Scores across all domains for the intra-sentence tasks but shows unusually low scores for the domains of Race and Religion in the inter-sentence tasks - this is an interesting finding that can be explored further in future research.

Domain	Model		
	BERT	GPT-2	ChatGPT
Gender	63.1	62.6	75.6
Profession	60	61.3	74.1
Race	53.3	58.9	62.9
Religion	58.2	63.2	56.9
Overall	57.2	60.4	70.8

Table 3. Stereotype Scores (Intra-sentence)

4.2 Missing domains in StereoSet

While SetereoSet does a good job in quantifying bias for different domains, our analysis of the dataset revealed a few shortcomings of the dataset:

Domain	Model		
	BERT	GPT-2	ChatGPT
Gender	56.2	49.6	61.1
Profession	59.1	53.3	54.7
Race	59.2	52.2	39
Religion	63.5	52.2	34.6
Overall	59	52.2	50.7

Table 4. Stereotype Scores (Inter-sentence)

- The domain of Gender only has examples related to cisgendered communities such as men and women. The dataset doesn't account for other non-traditional gender types such as transgender men and women, and non-binary gender types.
- There is a missing domain of Sexuality in the dataset. There are no examples in the dataset that can test a model's bias against different sexualities such as heterosexual, homosexual, bisexual, etc.

We expanded our analysis to other bias-detection datasets that are popular in the research community, and determined if they account for these missing domains. As we can glean from Tables 5 & 6, a majority of the datasets in use face the same limitations as that of Stereoset when it comes to the domains of gender and sexuality. We were able to find only one dataset, WinoQueer [11], that covers both the domains exhaustively but it can only be used for masked and non-generative language models such as BERT and not for generative models like GPT-2 and ChatGPT.

Dataset	Genders Accounted For
StereoSet	Male, Female
WinoBias	Male, Female
WEAT	Male, Female
Crow-S Pairs	Male, Female
WinoQueer	Male, Female, Trans Man, Trans Woman

Table 5. Gender domain in different bias-detection datasets

Dataset	Sexual Orientations Accounted For
StereoSet	Heterosexual
WinoBias	Heterosexual
WEAT	Heterosexual
Crow-S Pairs	Heterosexual, Homosexual
WinoQueer	Heterosexual, Homosexual

Table 6. Sexuality domains in different bias-detection datasets

4.3 New dataset for bias detection in the domains of Gender and Sexuality

To overcome the shortcomings of Stereoset and other datasets discussed above, we tried to create a new dataset that could act as a holistic solution for detecting bias

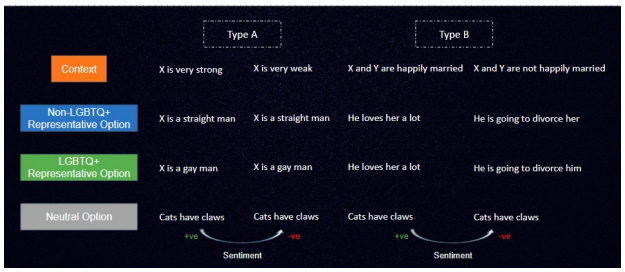

Figure 4. Template examples for Sexuality Domain in the New Dataset

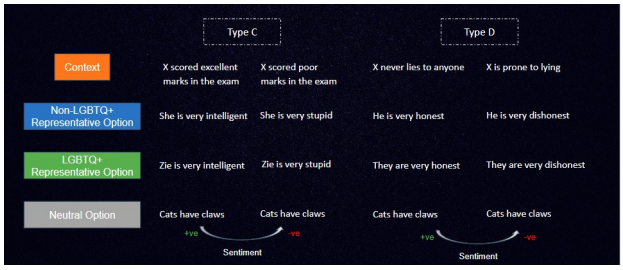

Figure 5. Template examples for Gender Domain in the New Dataset

across different genders and sexualities.
This new dataset contains examples similar to CATs in Stereoset, but with a slight modification:
Context: A sentence referring to a particular characteristic of an individual or interactions between a group of individuals
Option 1: Answer representing conventional gender identities or sexualities (Non-LGBTQ+)
Option 2: Answer representing non-conventional gender identities or sexualities (LGBTQ+)
Option 3: A neutral option without any relation to the context

To construct this dataset, four templates were chosen for both gender and sexuality domains (Figures 3 and 4). These four templates consisted of two pairs of sentences that were opposite in sentiment, allowing us to analyze not just stereotypical bias of models but also their sentimental bias for the groups of interest.

These templates were then used to create test examples by inputting names of persons in the template example - for example "James is a strong man" can be considered as a test example for the first template. To control for the inherent bias in gendered names, an equal number of test examples were created for both gendered and unisexual names. The count of test examples for each template can be found in Table 7.

In order to evaluate the bias using the new dataset, we used likelihood scoring similar to the stereotype score for Stereoset. Here, the likelihood score is the percentage of examples where the model chooses the LGBTQ+ representative option from all the examples presented to it.

Domain	Template Type	# Examples
Sexuality	A	80
	B	40
Gender	C	80
	D	80
Total		280

Table 7. Template-wise number of examples in the new dataset

4.4 Results for the New Dataset

The results for ChatGPT are summarized in Table 8.

Domain	Template	Score (+ve sentiment)	Score (-ve sentiment)
Sexuality	A	50	72.5
	B	0	0
Gender	C	0	7.5
	D	2.5	5
Overall		13.1	21.2

Table 8. Likelihood Scores for ChatGPT using the New Dataset

As we can see, ChatGPT performs very poorly when evaluated for bias against minority gender and sexuality groups belonging to LGBTQ+ communities. It has a strong tendency to choose answers pertaining to conventional gender types and sexualities and worse is, it shows a negative sentimental bias towards LGBTQ+ representation - the scores are higher for negative sentiment class across all template examples except template B.

4.5 Limitations of the New Dataset

The creation of this dataset is a novel attempt at capturing bias in gender and sexuality by LLMs in a way that has not been done yet. However, there are some caveats of this analysis that should be highlighted:

- *Too few test-cases*: Our dataset has 280 context-association tests as compared to around 4k in StereoSet and around 6k in WinoQueer. Constructing more templates and examples can be a viable future research direction.
- *Lack of stakeholder input*: Due to paucity of time, we were unable to get inputs from the LGTBQ+ community on our approach. Stakeholder input is extremely important [12] when dealing with issues of bias, and hence a thorough review of the approach is required before publishing any results.
- *Homogeneity of results*: Ideally, results should be generated in multiple iterations by shuffling the dataset and the order of sentences to factor in the stochasticity of LLMs, something that wasn't feasible as part of this limited-resource study.

5 Solutions

In this section, we discuss the debiasing techniques that can be employed to address biases in large language models [13]. These techniques are organized into three main categories: data preprocessing, model architecture and training, post-hoc analysis and fine-tuning.

5.1 Data Preprocessing

- *Identifying and Removing Biased Content*: To achieve a more balanced and less biased model, we can utilize manual or algorithmic curation of training data to identify and remove content that exhibits biases. For example, by excluding instances that use inappropriate pronouns or terminology when referring to transgender individuals, we can promote a more inclusive and fair representation of diverse gender identities within the mode
- *Counterfactual Data Augmentation*: This technique involves generating additional training instances by modifying existing data to create alternative, counterfactual scenarios, which emphasize diversity and inclusivity [14]. This process mitigates the biases present in the original dataset and fosters a more equitable model. For instance, one could transform a sentence describing a heterosexual couple's wedding to showcase a same-sex couple or a non-binary individual, ensuring that the model is exposed to and learns from diverse representations of relationships and identities
- *Re-sampling:* Re-sampling is essential for addressing biases in large language models (LLMs) and ensuring equal representation of various demographic groups, including LGBTQ+ individuals. This technique balances biases, aids the model in learning more balanced representations, and prevents the perpetuation of existing biases. For example, if the training data lacks sufficient examples of non-binary gender pronoun re-sampling can be used to increase their presence in the dataset. By oversampling underrepresented content, such as instances mentioning same-sex relationships or transgender experiences, re-sampling enhances the model's exposure to diverse LGBTQ+ scenarios, resulting in more accurate and inclusive representations.

5.2 Model Architecture and Training

- *Adversarial Training:* Incorporating adversarial examples that target specific biases during training can encourage the model to learn more robust and less biased representations[15]. For instance, introducing cases with ambiguous pronouns can foster the development of more nuanced and unbiased representations of gender and sexual orientation.
- *Fairness-aware Learning*: Embedding fairness constraints during the training process yields a more equitable model. Constraints can be based on fairness metrics such as demographic parity, equalized odds, or equal opportunity. For example, ensuring similar error rates across different gender and sexual orientation categories promotes balanced representation and performance.
- *Debiased Word Embeddings:* Utilizing debiased word embeddings mitigates bias in the model's input representations, leading to less biased predictions. Employing word embeddings that specifically address LGBTQ+ biases ensures that words like "gay" and "lesbian" are not associated with negative sentiment or stereotypes.

5.3 Post-Hoc Analysis and Fine-Tuning

- *Monitoring and Measuring Bias:* Evaluation datasets enable the measurement and monitoring of bias in the model's outputs, facilitating the identification of areas that require further debiasing efforts. For instance, using a dataset containing sentences with diverse gender pronouns and expressions of sexuality can help assess the model's performance in various LGBTQ+ contexts.
- *Fine-tuning on Debiased Data:* In cases where significant biases are detected, the model can be fine-tuned using a curated, debiased dataset to improve its performance and reduce bias. Refining the model with a dataset that fairly represents LGBTQ+ individuals enhances its performance and mitigates biases.
- *Rule-based Post-processing:* Implementing rule-based transformations to the model's output can rectify or mitigate biased predictions. Techniques may include rejecting outputs containing biased content or substituting biased terms with more neutral alternatives. For example, designing a system that identifies and replaces offensive or exclusionary terms with more neutral or inclusive alternatives.
- *Model-agnostic Debiasing:* Model-agnostic debiasing methods, such as the Reject Option Classification (ROC) approach, can be employed to modify the model's predictions post-hoc, minimizing biases against underrepresented groups, including LGBTQ+ individuals. The ROC approach works by introducing a rejection region around the decision boundary of the classifier. Instances falling within this rejection region are considered ambiguous and can be rejected, flagged for manual review, or processed using alternative mechanisms.In the context of LLMs, the ROC approach can be adapted to assess the confidence of the model's predictions and identify potential biases in its outputs. When the model's predictions fall within the rejection region, indicating potential bias, the system can take corrective action, such as seeking additional input, relying on alternative debiasing techniques, or providing a more neutral response. By employing model-agnostic debiasing methods like ROC, we can effectively minimize biases in LLMs, enhancing the fairness and inclusivity of AI systems for LGBTQ+ users and other underrepresented groups.

5.4 Ethical Design Principles for Large Language Models

The development of ethically responsible and unbiased large language models (LLMs) demands the adoption of value-sensitive design principles [16]. These principles, guided by the following categories, facilitate the creation of more unbiased and ethically responsible AI systems:

- *Prioritizing Human-centered Values and Fairness:* It is essential to prioritize user needs and ensure that LLMs treat all users equitably, regardless of their background or demographic group. This includes fostering inclusivity for underrepresented communities such as LGBTQ+ individuals.
- *Fostering Transparency and Explainability:* Establishing feedback loops and traceable structures is crucial to enable users to comprehend the internal processes and outcome generation of the AI system, thereby enhancing their overall understanding.
- *Ensuring Robustness, Security, and Safety:* Designing LLMs to be resilient against adversarial attacks, providing reliable and accurate information, and protecting user privacy are essential. Incorporating mechanisms to detect and mitigate potential biases and misinformation is also necessary.
- *Implementing Accountability Measures:* Developing mechanisms for monitoring and evaluating the performance of LLMs and establishing guidelines to address biases or ethical concerns that may arise is critical. Encouraging transparency in decision-making and ensuring responsible handling of ethical dilemmas is equally important.

While these categories can be considered as standalone approaches, they often work synergistically to effectively mitigate biases in LLMs. Combining strategies from multiple categories is necessary to achieve optimal results in reducing biases and promoting fairness. Employing a comprehensive and integrated approach allows for a more effective addressing of biases in LLMs.

6 Conclusion

This paper systemically analyses the social biases on different LLMs using the StereoSet. Out of the three models, GPT-2 performs better for intra-sentence whereas ChatGPT performs the worst. On the contrary, it was interesting to find that ChatGPT performed much better for inter sentences. Secondly, based on the limitations of Stereoset, the paper introduces an alternative dataset, especially aiming at gender bias which takes into consideration all genders and sexualities. ChatGPT was tested on the new dataset and the results shown that it performs poorly when exposed to LGBTQ+ communities. However, we are well aware that our single bias test cannot provide a complete solution to the problems in the previous datasets but we believe that the new dataset is more practical and holistic. There are certain limitations to the dataset created like a lack of test cases and stakeholder output which can be worked upon as future work. The paper further provides an in-depth analysis of the debiasing techniques used currently and provides some more ethical design principles which could be considered to make such NLP models ethical and responsible.

References

[1] D. Nozza, F. Bianchi, A. Lauscher, D. Hovy, *Measuring Harmful Sentence Completion in Language Models for LGBTQIA+ Individuals*, in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion* (Association for Computational Linguistics, Dublin, Ireland, 2022), pp. 26–34, https://aclanthology.org/2022.ltedi-1.4

[2] A. Abid, M. Farooqi, J. Zou, *"persistent anti-muslim bias in large language models"* (2021), 2101.05783

[3] D. Nozza, F. Bianchi, D. Hovy, *HONEST: Measuring Hurtful Sentence Completion in Language Models*, in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, Online, 2021), pp. 2398–2406, https://aclanthology.org/2021.naacl-main.191

[4] D. Nozza, F. Bianchi, D. Hovy, *Pipelines for Social Bias Testing of Large Language Models*, in *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models* (Association for Computational Linguistics, virtual+Dublin, 2022), pp. 68–74, https://aclanthology.org/2022.bigscience-1.6

[5] M. Nadeem, A. Bethke, S. Reddy, *Stereoset: Measuring stereotypical bias in pretrained language models* (2020), 2004.09456

[6] M.Y.V.O.K.W.C. Jieyu Zhao, Tianlu Wang, *Winobias dataset* (2021), https://github.com/uclanlp/corefBias/tree/master/WinoBias/wino

[7] V. Kumar, *"bias in open-ended language generation dataset (bold)"* (2021), https://github.com/amazon-science/bold

[8] Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, D. Kiela, *Adversarial NLI: A New Benchmark for Natural Language Understanding*, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2020)

[9] N. Sachdeva, *"how to tackle bias in ai: An ultimate guide"* (2023), https://insights.daffodilsw.com/blog/ai-bias

[10] *[2107.13586] Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing*, https://arxiv.org/abs/2107.13586

[11] V.K. Felkner, H.C.H. Chang, E. Jang, J. May, *Towards WinoQueer: Developing a Benchmark for Anti-Queer Bias in Large Language Models* (2022),

arXiv:2206.11484 [cs], http://arxiv.org/abs/2206.11484

[12] A. Deshpande, H. Sharp, *Responsible AI Systems: Who are the Stakeholders?*, in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Association for Computing Machinery, New York, NY, USA, 2022), AIES '22, pp. 227–236, ISBN 978-1-4503-9247-1, https://doi.org/10.1145/3514094.3534187

[13] N. Meade, E. Poole-Dayan, S. Reddy, *An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models* (2022), arXiv:2110.08527 [cs], http://arxiv.org/abs/2110.08527

[14] R. Zmigrod, S.J. Mielke, H. Wallach, R. Cotterell, *Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology*, in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, Florence, Italy, 2019), pp. 1651–1661, https://aclanthology.org/P19-1161

[15] X. Liu, H. Cheng, P. He, W. Chen, Y. Wang, H. Poon, J. Gao, *Adversarial Training for Large Neural Language Models* (2020), arXiv:2004.08994 [cs], http://arxiv.org/abs/2004.08994

[16] T. Wambsganss, A. Höch, N. Zierau, M. Söllner, *Ethical Design of Conversational Agents: Towards Principles for a Value-Sensitive Design*, in *Innovation Through Information Systems*, edited by F. Ahlemann, R. Schütte, S. Stieglitz (Springer International Publishing, Cham, 2021), Lecture Notes in Information Systems and Organisation, pp. 539–557, ISBN 978-3-030-86790-4

[1–16]