The Bias Within: Evaluating the Generation and Classification Dichotomy in Large Language Models

Reem Alsharabi Shahad Albalawi Norah bin Mannie Najla Alhamdan Nouf Alamoudi

Guohao Li

Abstract

Large language models (LLMs) have demonstrated impressive generative capabilities but also carry the risk of encoding sociocultural biases. This study explores bias detection in story generation by language models. We use GPT-3.5 to automatically generate a dataset of 800 stories, divided equally between English and Arabic comprising biased and unbiased examples across different categories. Leveraging this dataset, we conduct two experiments: First, we employ GPT-3.5 to classify bias in its own generated stories, revealing limitations in recognizing inherent biases. Second, we fine-tune GPT-2 as a dedicated bias classifier, achieving 97.5% accuracy on unseen data. Our findings highlight the gap between the generation and classification capabilities of language models and underscore the need for task-specific training to enhance bias analysis.

1 Introduction

Large-scale language models (LLMs) have demonstrated remarkable effectiveness across a wide range of tasks, including text generation [9], question answering [7], and story generation [17]. These models derive their proficiency from extensive training on substantial datasets, consistently exhibiting high-quality outputs [8].

However, the promise of LLMs is accompanied by a notable concern: the potential for expressing undesirable representational biases. These biases, often stemming from the reinforcement of societal stereotypes, possess the capacity to propagate and reinforce negative assumptions related to gender, race, religion, and other social constructs [10]. These biases can lead to troubling outcomes, including problematic choices and unfair discrimination in automated systems [3]. Furthermore, social biases and stereotypes negatively affect how people judge different groups, and they play an important role in understanding the language used towards marginalized groups [5].

In light of these challenges, previous work was done to develop bias identification frameworks [15], measure bias in LLMs [12], and classify text via LLMs [16]. However, these studies might struggle to capture nuanced forms of bias, as they often rely on identifying a single type of bias, or categories that may not encompass the full spectrum of biases present in various contexts. Also, most of the existing work focuses on binary scenarios, which restricts the understanding of bias in more complex contexts [14]. Additionally, prior research has not adequately examined using LLMs as bias classifiers for text that is generated by the LLMs themselves. This raises the problem of comprehensively evaluating biases in text generated by LLMs. To address this problem, this paper aims to:

1. Collect a dataset of stories generated by an LLM comprising both biased and unbiased stories across various social bias categories
2. Use the same LLM to classify the bias in the collected stories, evaluating its ability to identify biases in its own generated text

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

3. Fine-tune another LLM and employ it as a classifier to compare its performance

This paper presents a comprehensive exploration of stereotypical bias detection in story generation using LLMs. Namely, we use GPT-3.5 to collect a dataset comprising both biased and unbiased stories, enabling us to evaluate the presence and extent of bias in the generated stories. We collect a substantial number of stories for each category, ensuring a diverse representation of biases. Our dataset encompasses different categories related to social bias, including cultural stereotypes, gender bias, biases about names, nationalities, ages, marital status, profession, and physical appearance. By incorporating these various categories, we aim to offer a comprehensive analysis of bias manifestations within story content. We assign stories in our dataset both a binary bias class and a bias intensity rating on a scale of 1 to 5. This approach allows for a more nuanced analysis of bias within story content and introduces a multi-level classification approach. This multi-faceted evaluation enables us to not only identify the presence or absence of bias but also quantify the degree to which bias is exhibited in the generated stories.

We also leverage the same LLM (GPT-3.5) used for story generation to classify the bias in the collected stories to evaluate its ability to identify bias within these generated stories. This approach allows us to directly assess whether the model can discern and flag instances of bias present in its own generated content.

Moreover, we fine-tune GPT-2 model and employ it for bias detection, aiming to compare its performance against the original model.

This dual approach not only provides a holistic comprehension of the models' proficiency in detecting narrative bias but also serves as a foundation for constructing bias classifiers. These classifiers play a pivotal role in evaluating the credibility of text generated by LLMs, thereby enhancing our capacity to assess the trustworthiness of AI-driven content.

The rest of this paper is organized as follows. Section 2 provides an overview of related work in bias in LLMs, and using LLMs as classifiers. Section 3 details our methodology for data collection and bias classification. We present and analyze our experimental results, and discuss their implications in Section 4. Finally, Section 5 concludes the paper with a summary of findings and suggestions for future research.

2 Related Work

This section delves into the sociocultural biases in language models and addresses challenges in accurate text classification

Bias in Language Models

Mikolov et al., [11] discuss bias in representations learned by language models. The authors noted they risk perpetuating the biases contained in the source data. For example, repeatedly exposing the model to overwhelmingly male representations of certain careers could encode gender biases. They found that subsampling frequent words during the training helped reduce frequency bias by significantly improving representations of uncommon words. However, selecting training data and hyperparameters like vector size can still impact bias, as some configurations may better capture majority views over minority opinions. The findings highlight the importance of carefully choosing datasets, training methods, and hyperparameters to mitigate both frequency bias towards popular words, as well as sociocultural biases in the data.

Moreover, as LLMs advance, the implications of inherent biases have gained increased attention. Recent research [6] has focused on the problems of bias in pre-trained LLMs. These models are prone to incorporating biases from their training data distributions and other factors. Studies have utilized audit methods like classifier probes and word embeddings to detect unwanted statistical associations within the models' learned representations. Debiasing techniques under investigation include data augmentation, adversarial training on debiasing objectives, and constraints during pre-training. However, eliminating biases remains challenging given the open-ended nature of language. More holistic approaches are being explored, such as incorporating representation learning with other techniques like constraint programming. This literature demonstrates the need for ongoing evaluation and mitigation efforts to develop more robust and socially-aware natural language processing systems.

Nadeem et al., [13] developed the Context Association Test (CAT) to technically measure the stereotypical biases exhibited by pre-trained language models, in contrast to evaluating language modeling ability. They crowdsourced a dataset called StereoSet containing 16,995 CATs to empirically test for biases in four domains: gender, profession, race, and religion. Their findings showed that current language models statistically demonstrate strong stereotypical biases. Additionally, they observed a correlation between language modeling performance and the degree of stereotypical bias - highlighting the need to decouple these to achieve unbiased models. It's important to note that the scope of this work does not extend to evaluating the capability of language models to detect bias in text generated by other language models.

Using Language Models as Classifiers

In existing research [2], researchers were able to demonstrate the use of LLMs, specifically ChatGPT, to remove bias from textual data through simplification. They performed sentiment analysis on the original and simplified reviews to ensure sentiment remained the same. The researchers suggest this technique warrants further investigation to develop a capable bias mitigation method for textual data. However, the study also noted several areas for future work.

On the other hand, a recent study [16] shows that LLMs have limitations in classifying text when compared to fine-tuned models. While LLMs have achieved success in various natural language processing (NLP) tasks, they underperform text classification tasks that involve complex linguistic phenomena that require reasoning abilities. Specifically, LLMs struggle with phenomena such as intensification, contrast, and irony due to a lack of reasoning capabilities to adequately address these nuanced language issues. Additionally, the limited number of tokens allowed during in-context learning further hinders LLMs' performance on text classification compared to fully fine-tuned models.

While LLMs have shown strong generative capabilities, being able to produce meaningful output when prompted with a class or topic, their classification performance still lags behind fine-tuned models. Even if an LLM can generate a piece of text that seems to belong to a certain class based on its content and keywords, the model may struggle to reliably classify that same text. This gap between generation and classification performance highlights a key limitation of LLMs. Even after being prompted with class information, the model lacks the reasoning abilities to deeply understand a text's semantic and linguistic features that determine its true class. It can imitate or even surpass human-level writing for a prompted topic, but fall short of carefully analyzing a pre-existing piece of text to assign the proper label.

3 Method

This section shows our methodology for how we collect a dataset with different levels of bias generated by LLM. For the classification of bias, a twofold approach has been undertaken: firstly, we asked the same generator LLM to classify bias in the stories, and secondly, via the fine-tuning of a distinct model.

3.1 Dataset

In this paper, we automated the creation of a dataset containing biased and unbiased stories using OpenAI's GPT-3.5 API [1]. We prompted the model to generate both biased and unbiased stories, for biased stories, we introduced the bias in the prompt, focusing on social and cultural biases. These bias categories encompassed aspects like cultural stereotypes, gender bias, and biases related to names, nationalities, ages, marital status, professions, and physical appearance.

To understand the basis for each bias, the model identified underlying reasons within the generated stories. For evaluating bias intensity, we employed a 1 to 5 rating system, with 1 representing slight bias and 5 denoting high bias. The model provided ratings for each biased story in accordance with this system. This approach enabled a nuanced analysis of bias, introducing a multi-level classification method.

In total, we collected a dataset comprising 800 stories, the dataset was evenly divided into two language groups: English and Arabic, each containing 400 stories, each with 200 stories exhibiting bias and 200 stories devoid of bias. This method allowed us to conduct a comprehensive analysis

of bias within the generated stories. Going beyond binary identification, it furnished a quantifiable measure of the degree of bias present. This multi-faceted approach not only enriches the understanding of bias manifestation in textual content but also equips future research to transcend the confines of binary bias classification.

3.2 Bias Classification

The bias classification strategy encompasses two main avenues: utilizing GPT-3.5 as a classifier and fine-tuning GPT-2 model for dedicated bias classification. This approach streamlines bias assessment and sheds light on the model's capacity to recognize biases it generates. In our evaluation process, accuracy serves as the chosen metric for assessing performance. The ground truth against which we measure accuracy is the label that the model was originally directed to produce stories under.

3.2.1 GPT-3.5 as a Classifier

After collecting the data, we utilized GPT-3.5 to assess bias within the generated stories. We prompted the model to classify each story for bias using OpenAI's GPT-3.5 API [1]. This method holds significance, given that the stories were originally generated by GPT-3.5 itself.

3.2.2 GPT-2 Fine-tuned Classifier

To create a specialized bias classifier, we fine-tuned the GPT-2 model for binary classification, utilizing the implementation described in Cai's work [4]. We compiled our dataset with labeled examples of biased and unbiased stories, ensuring a comprehensive representation of various biases and neutral content. Through iterative training, we meticulously honed the model's ability to discern bias, enabling it to effectively differentiate between the two distinct categories.

We meticulously assessed the model's accuracy using unseen validation and testing sets, employing rigorous evaluation metrics to gauge its performance over successive iterations. This iterative refinement process allowed us to enhance the model's discriminatory capabilities and optimize its precision in identifying bias within textual content.

4 Experiments

In this section, we outline the experimental process for using GPT-3.5 as a classifier, and fine-tuning the GPT-2 model to build a bias classifier. Finally, we evaluate the performance using accuracy as our metric and present a comparative analysis of the two experiments.

4.1 Results of GPT-3.5 as a Classifier

In this experiment, we employed GPT-3.5 to classify the stories in the dataset as biased or unbiased. Each story was presented to the model individually, and the model's classification output was recorded. The overall accuracy of the model in this classification task served as the primary evaluation metric.

Surprisingly, the results of the experiment indicated that GPT-3.5 had an overall classification accuracy of only 53% in detecting bias in stories. This accuracy rate was significantly lower than anticipated, given that the stories within the dataset were initially generated using the same GPT-3.5 model and were designed to showcase explicit stereotypes and biases. The unexpected outcome raises several noteworthy points for discussion.

The observed performance of GPT-3.5 prompts intriguing questions about the model's ability to identify biases within the text. The fact that the same model generated both the biased and unbiased stories in our dataset challenges our assumptions about the model's sensitivity to recognizing biases it itself has generated. This indicates the limitations in the LLMs' comprehension of biases, nuances, and contextual cues that influence the classification process.

This result carries significant implications due to the increasing reliance on LLMs for various tasks, including content generation, information retrieval, and decision support. As society turns to LLMs to assist in complex cognitive tasks, the ability of these models to accurately recognize and address biases becomes paramount. The unexpected performance of GPT-3.5 raises concerns about the reliability of such models in detecting and mitigating biases present in their own output.

Furthermore, the outcome underscores the challenges in developing effective bias detection mechanisms and the importance of thorough testing to reveal potential weaknesses. The discrepancy between the anticipated and actual performance highlights the complexities of bias analysis and the need for continuous improvement in machine learning algorithms.

4.2 Results of Fine-tuned GPT-2 Classifier

In this experiment, we conducted a comprehensive process to fine-tune the GPT-2 model for bias classification and subsequently built a bias classifier. The experiment encompassed various stages including data preprocessing, model architecture, training, and evaluation.

For the data preprocessing phase, we began by fine-tuning the GPT-2 model using the dataset we collected earlier. Stories in the dataset were labeled as either exhibiting bias ("yes") or being unbiased ("no"). This labeling process established the ground truth for training and evaluating our bias classification model.

To ensure the effective utilization of the fine-tuned GPT-2 model, we initiated text tokenization on the input sentences. This involved converting the sentences into tokenized sequences. To adhere to GPT-2's decoder structure, we incorporated left padding to the tokenized sequences, allowing the model to predict subsequent tokens accurately.

The architecture of our bias classification model was built upon the GPT-2 foundation. A pivotal addition to this architecture was the introduction of a linear classifier layer atop GPT-2's 12 decoder layers. This classifier layer was designed with an output dimension matching the number of labels, which in this case was two. Through this configuration, our model was empowered to produce two scores, effectively representing its prediction probabilities for each bias category.

Our labeled dataset was thoughtfully divided into three distinct subsets: a training set consisting of 320 data points, a validation set of 40 data points, and a test set also containing 40 data points.

During the training phase, we employed the cross entropy loss function, which quantified the discrepancy between predicted and actual labels. The optimization of model parameters was executed using the Adam optimizer, with the learning rate initialized at 9e-4. And the model was trained on 30 epochs. Figure 1 demonstrates a desirable trend in the model's training and testing loss. Both the training and validation losses decrease consistently with the progression of training epochs. This pattern indicates that the model is neither overfitting nor underfitting, suggesting a balanced and potentially effective level of generalization.

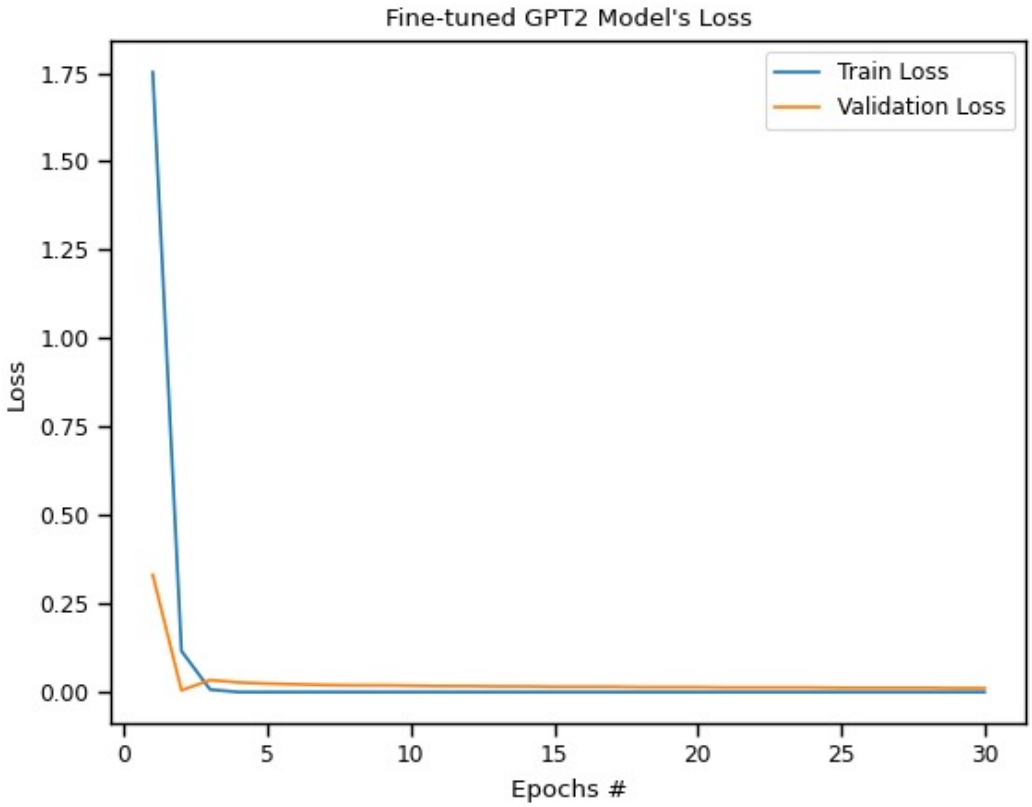

Figure 1. Loss Trends Throughout Training Epochs of Fine-Tuned GPT-2

After training, we evaluated the model's performance using both the validation and test sets. Accuracy emerged as the primary evaluation metric. Impressively, the results were as follows: after 30 training epochs, the model achieved a training accuracy of 99.7%. In the validation phase, the model exhibited a robust accuracy rate of 97.5%. In Figure 2, we illustrate the evolution of training and validation accuracies over the course of training epochs.

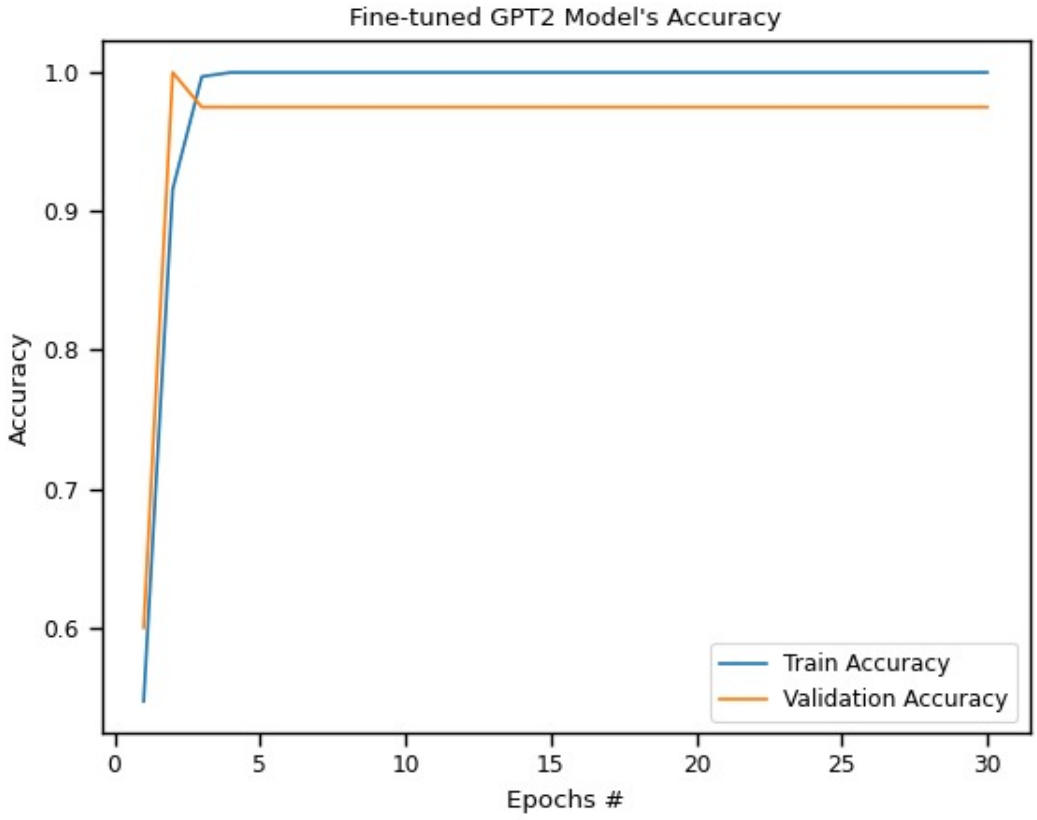

Figure 2. Progression of Training and Validation Accuracy for Fine-Tuned GPT-2

Furthermore, Figure 3 presents the confusion matrix generated from evaluating the model's performance on the test set. The matrix provides a comprehensive visual representation of how the model's predictions align with the actual class labels in the test data. The diagonal of the matrix, where true positives and true negatives are located, provides a visual indicator of the model's accurate predictions. This diagonal pattern is indicative of a good alignment between the model's predictions and the actual class labels.

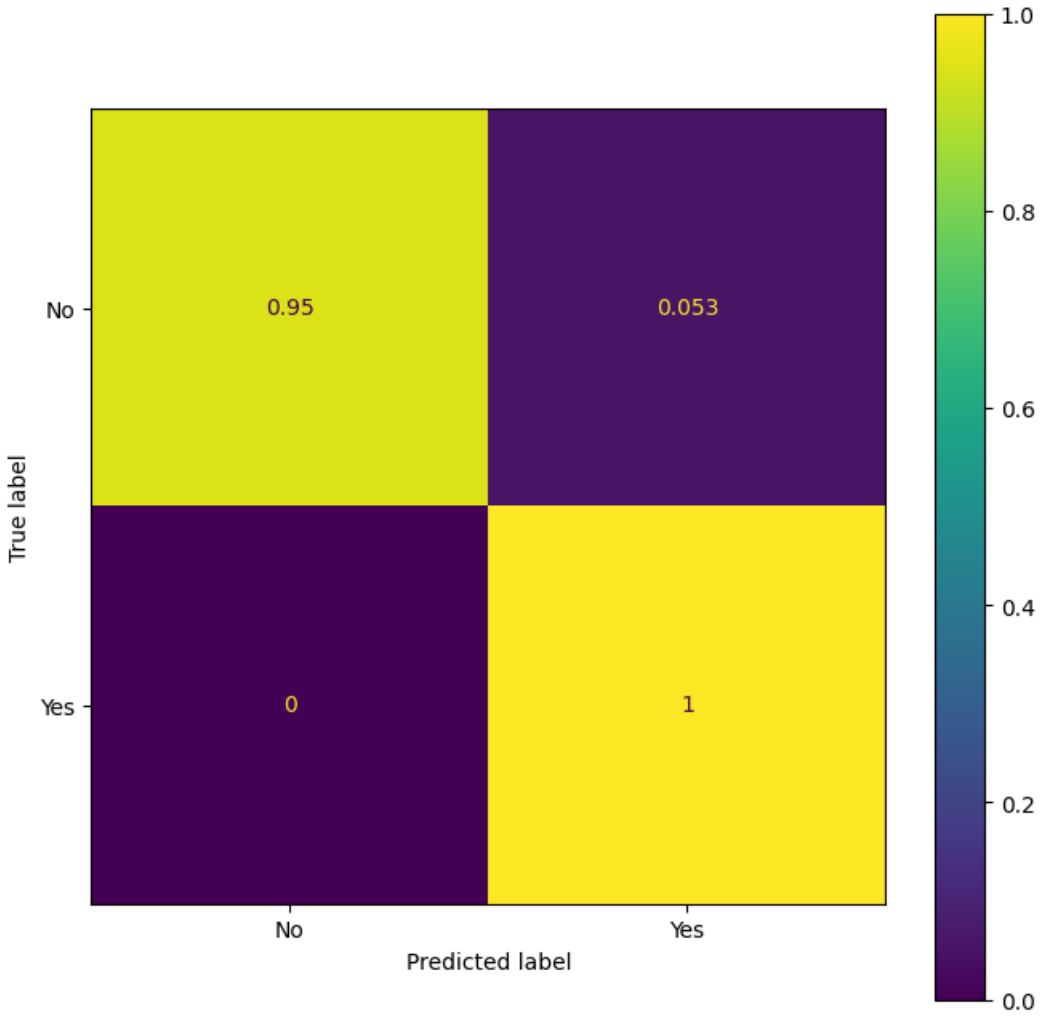

Figure 3. Confusion Matrix for Classification of Test Set using Fine-Tuned GPT-2

4.3 Performance Comparison

In a compelling progression, our experiments show that the fine-tuned GPT-2 classifier outperformed GPT-3.5 performance, even though our dataset was initially generated by the GPT-3.5 architecture. To further understand the gap in the performance, we calculate the accuracy of GPT-3.5 in classifying bias in the same test set used for GPT-2. As mentioned earlier, GPT-2 achieved an impressive 97.5% on the test set. In contrast, GPT-3.5's performance on the same set was notably lower, reaching only 45%. Figure 4, which illustrates a bar graph showcasing the accuracies of both models on the identical test set. The graph provides a clear side-by-side comparison, highlighting the substantial difference in accuracy between GPT-2 and GPT-3.5. Nevertheless, our experiments have provided

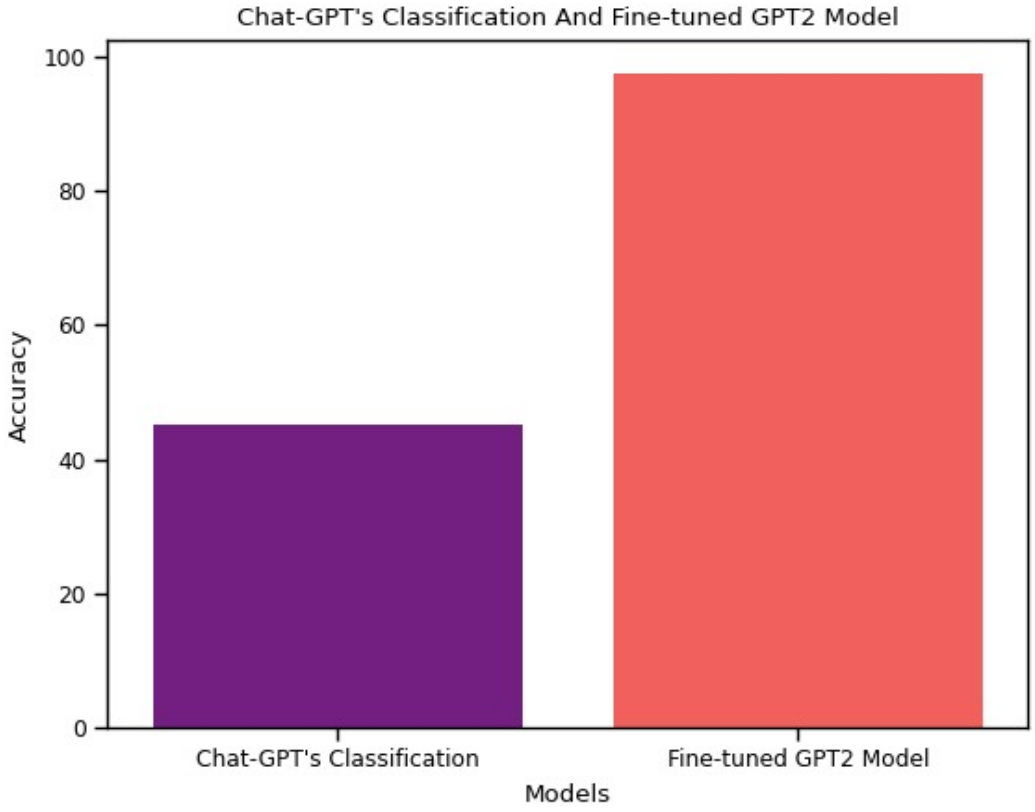

Figure 4. Comparative Accuracy of GPT-2 and GPT-3.5 on Identical Test Set

valuable insights into the bias detection capabilities of GPT-3.5 and the fine-tuned GPT-2 classifier. GPT-3.5's performance in identifying bias within generated content raises questions about its ability to recognize inherent biases, given its role in generating the content. On the other hand, the fine-tuned GPT-2 classifier's success in bias detection highlights the potential of task-specific training to enhance bias recognition. The observed disparity in performance between the two models underscores the importance of customization for effective bias analysis.

5 Conclusion

In this paper, we comprehensively evaluated bias detection in story generation by large language models. We collected a dataset containing binary bias classes and intensity ratings using GPT-3.5 for a nuanced analysis. Our experiments provided insights by using GPT-3.5 to classify its own generated content, achieving only 53% accuracy and raising questions about inherent bias recognition. In contrast, fine-tuning GPT-2 as a dedicated classifier significantly improved performance to 97.5% accuracy. This work revealed notable disparities between language models' generation and classification abilities. Moving forward, customized training approaches hold promise for enhancing bias analysis. Additionally, our collected dataset in English and Arabic that encompasses various levels of

bias across different categories, the dataset can be used for going beyond binary classification toward multi-class classification. Overall, this study underscores the complexities of bias and the importance of thorough evaluation to develop more accountable AI systems.

Limitations This study was limited to binary classification and English language. Future work should expand this to multi-class classification and different languages.

References

[1] Openai platform. https://platform.openai.com/docs/models/gpt-3-5. Accessed: 15-Aug-2023. 3, 4

[2] Charmaine Barker and Dimitar Kazakov. Chatgpt as a text simplification tool to remove bias. 2023. 3

[3] Shikha Bordia and Samuel R. Bowman. Identifying and reducing gender bias in word-level language models. 2019. 1

[4] H. Cai. Gpt2-news-classifier: A streamlit app running gpt-2 language model for text classification, built with pytorch, transformers and aws sagemaker. 2023. GitHub repository. Available at: [https://github.com/haocai1992/GPT2-News-Classifier]. 4

[5] Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. Hate Speech Classifiers Learn Normative Social Stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319, 03 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00550. URL https://doi.org/10.1162/tacl_a_00550. 1

[6] Emilio Ferrara. Should chatgpt be biased? challenges and risks of bias in large language models, 2023. 2

[7] Changwook Jun, Hansol Jang, Myoseop Sim, Hyun Kim, Jooyoung Choi, Kyungkoo Min, and Kyunghoon Bae. ANNA: Enhanced language representation for question answering. In *Proceedings of the 7th Workshop on Representation Learning for NLP*, pages 121–132, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.repl4nlp-1.13. URL https://aclanthology.org/2022.repl4nlp-1.13. 1

[8] Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. Challenges and applications of large language models. 2023. 1

[9] Junyi Li, Tianyi Tang, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. Pretrained language models for text generation: A survey. 2022. 1

[10] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. 139:6565–6576, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liang21a.html. 1

[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. 26, 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf. 2

[12] Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. *CoRR*, abs/2004.09456, 2020. URL https://arxiv.org/abs/2004.09456. 1

[13] Moin Nadeem, Anna Bethke, and Siva Reddy. StereoSet: Measuring stereotypical bias in pretrained language models, August 2021. URL https://aclanthology.org/2021.acl-long.416. 3

[14] Tiago P. Pagano, Rafael B. Loureiro, Fernanda V. N. Lisboa, Rodrigo M. Peixoto, Guilherme A. S. Guimarães, Gustavo O. R. Cruz, Maira M. Araujo, Lucas L. Santos, Marco A. S. Cruz, Ewerton L. S. Oliveira, Ingrid Winkler, and Erick G. S. Nascimento. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing*, 7(1), 2023. ISSN 2504-2289. doi: 10.3390/bdcc7010015. URL https://www.mdpi.com/2504-2289/7/1/15. 1

[15] Shaina Raza, Muskan Garg, Deepak John Reji, Syed Raza Bashir, and Chen Ding. Nbias: A natural language processing framework for bias identification in text, 2023. 1

[16] Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. Text classification via large language models, 2023. 1, 3

[17] Zhuohan Xie, Trevor Cohn, and Jey Han Lau. The next chapter: A study of large language models in storytelling. 2023. 1