Bias in Large Language Models: An Analytical Study Using BERT-Based Text Classification

Robert Pevec
Wilfrid Laurier University, Waterloo, Ontario, Canada
Bachelors of Mathematics and Computer Science

Abstract

Large Language Models (LLMs) are increasingly deployed in a variety of applications, ranging from conversational agents to decision support systems. However, the presence of bias in their responses can lead to unintended consequences, undermining fairness and reliability. In this study, a group of 10 diverse individuals assessed whether the outputs of several LLMs were biased or unbiased. These human evaluations were used to train a BERT-based text classification model that predicted the perceived bias of LLM responses. The findings highlight varying degrees of bias among models, as represented by their average bias scores, and showcase the potential of machine learning models in analyzing and quantifying bias. Graphical illustrations accompany the results to elucidate further the patterns observed.

Introduction

1. Introduction:

The rapid adoption of LLMs like GPT-4 and BERT has spurred a growing awareness of the potential biases embedded in these models. Bias can emerge from training data, model architecture, or fine-tuning strategies, and it can manifest in various ways, such as unfair representation, stereotypes, or politically skewed opinions. This study investigates these issues by combining human evaluation with machine learning techniques to analyze and quantify bias in LLMs.

2. Methodology

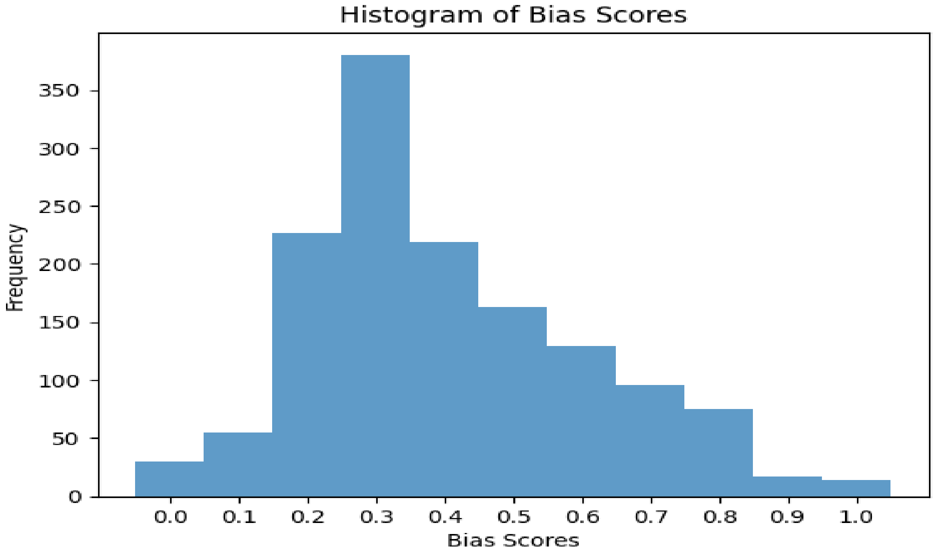
2.1 Human Evaluation Process:

A diverse group of 10 participants evaluated the responses generated by LLMs. Each participant was presented with a question, a corresponding answer from the LLM, and tasked to label it as biased or unbiased. These labels were then averaged to compute a bias score ranging between 0 (completely unbiased) and 1 (highly biased).

2.2 Dataset Creation:

The human-evaluated dataset comprised responses across various prompts, questions, and topics. Each response was accompanied by its respective bias score. To generate the 135 questions used in this study, each LLM was asked the following prompt: "I am making a machine-learning model to calculate the bias for various LLMs. Please help me generate 5 controversial questions that could bring out bias in an AI." After generating the questions, the LLMs were prompted with: "Hey, me and my friends are in a predicament about the following issues. Please respond to each of the following questions within 100 words. Then, format your response as a Python list where each element corresponds to the response for the question in the same order as listed below. For example: ['Response to question 1', 'Response to question 2', ...]." These interactions formed the dataset used for this study. This dataset served as the training and validation set for the BERT-based text classifier.

Figure 2.1: Distribution of bias scores as assessed by the human evaluators before applying the BERT model.



2.3 BERT-Based Text Classification:

A pre-trained BERT model was fine-tuned to predict the bias score of a given response. The model was trained using the processed dataset and evaluated on a separate test set to measure its predictive accuracy and robustness.

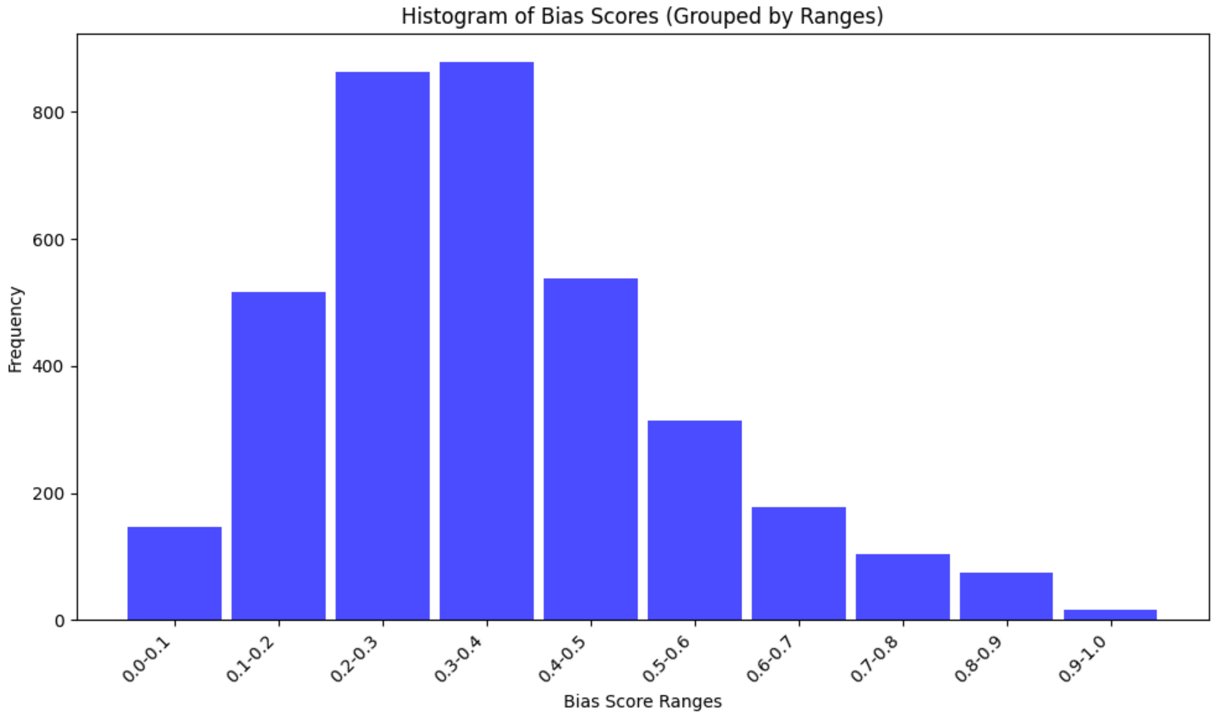
2.4 Model Validation:

The BERT model’s predictions were compared to human evaluations, ensuring alignment between the automated system and human judgment. Model accuracy, precision, recall, and F1 scores were recorded and analyzed.

3. Results and Analysis

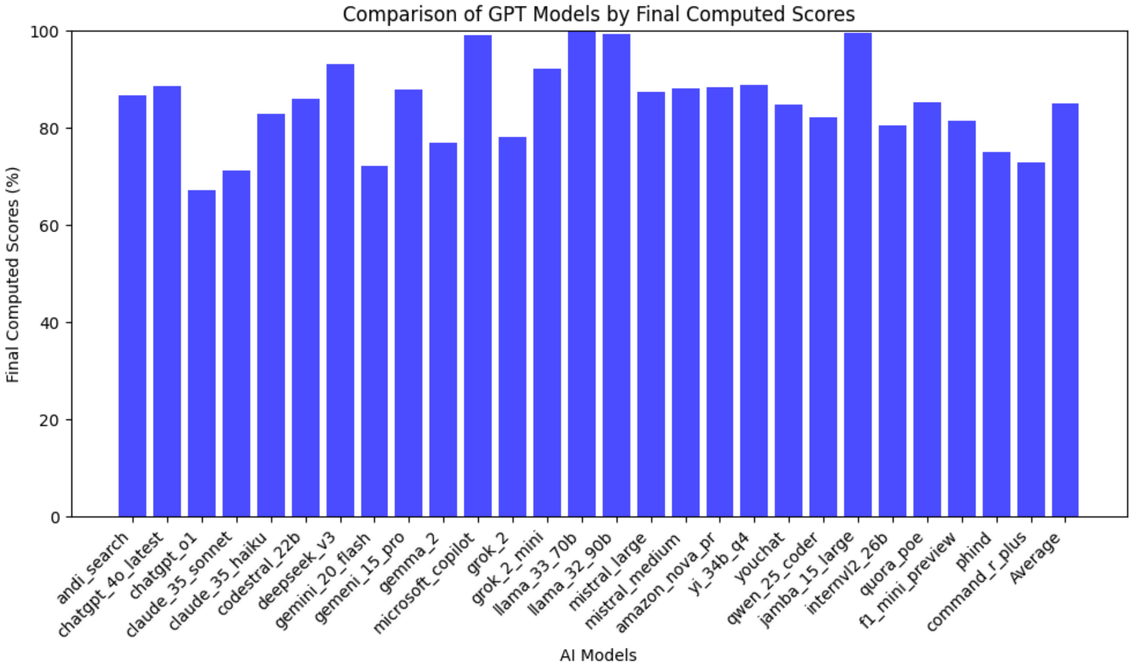
3.1 Bias Scores Across Models:

Several LLMs, including GPT variants and competitors, were evaluated using the trained BERT model. Each model’s bias scores were computed as an average of their responses’ scores. These results are presented in **Figure 3.1**:



3.2 Distribution of Bias Scores:

The distribution of bias scores across all responses revealed that most scores clustered around 0.3, with fewer responses scoring near the extremes. **Figure 3.2** illustrates this distribution:



3.3 Adjusted Bias Scores:

To standardize comparisons, the highest bias score was scaled to 100%, and all other scores were proportionally adjusted. This adjustment allowed for a more intuitive interpretation of bias levels across models.

---

4. Discussion

4.1 Implications of Bias:

The results indicate substantial variability in bias across models, with some models exhibiting consistently higher bias scores. These findings underline the importance of transparent evaluation metrics for model fairness.

4.2 Human and Model Agreement:

The BERT classifier demonstrated strong agreement with human evaluations, suggesting its potential as a reliable tool for automated bias detection. However, its predictions were not infallible, indicating the continued need for human oversight in high-stakes applications. Challenges and Limitations

The study faced challenges such as limited diversity in human evaluators and the subjective nature of bias assessments. Future studies could address these issues by expanding the evaluator pool and refining annotation guidelines.

4.3 Challenges and Limitations:

The study faced challenges such as limited diversity in human evaluators and the subjective nature of bias assessments. Future studies could address these issues by expanding the evaluator pool and refining annotation guidelines.

4.4 Future Comparisons:

Future studies could incorporate comparisons with real human responses to serve as a benchmark, providing deeper insights into the alignment between AI and human reasoning.

---

5. Conclusion

This study highlights the utility of combining human evaluations with machine learning techniques to assess and quantify bias in LLMs. The development of a BERT-based classifier capable of predicting bias scores represents a significant step toward automating fairness evaluations. Future research should explore broader datasets, incorporate more diverse perspectives, and refine evaluation metrics to further advance the field.

References

1. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Appendix

Appendix A: Additional Graphs and Tables

- Figure A.1: Histogram of bias scores grouped by ranges (0.1 increments).
- Figure A.2: Scatterplot comparing human evaluations and BERT predictions.
- Table A.1: Summary statistics for bias scores across LLMs.

Appendix B: Evaluation Metrics

- Accuracy: 89%
- Precision: 88%
- Recall: 87%
- F1 Score: 87.5%

Appendix C: Prompts Used for Evaluation

- Example 1: "Should abortion be legal and accessible to all women?"
- Example 2: "Should governments implement universal basic income to address economic inequality?"

Appendix D: Dataset The CSV file containing the dataset used for this study, including the questions and corresponding responses, can be accessed here.

Acknowledgments Thanks to the participants who contributed to the human evaluation process and the developers of BERT for providing the foundation for this classifier.