Readings in Political Economics
ISSN: 0079-9874 & 0305-814X
Vol 60. No. 2 (2026)
https://politonomia.org.ua/rpe

Quantifying Bias in Large Language Models: Comparing Fairness Algorithms

Nora Falken
Baltic Academy of Applied Sciences, Estonia

Received: 15/01/2026 **Accepted:05/03/2026** **Published: 06/08/2026**

Abstract:
Despite the fact that Large Language Models (LLMs) have amazing capabilities across a wide range of applications, there is a need for some kind of systematic examination and mitigation of the underlying biases that they include. an analysis of fairness algorithms that are designed to measure and correct inequities in LLMs, with a specific focus on biases linked with gender, race, and socioeconomic status for inclusion in the analysis. In order to evaluate the effectiveness of various fairness strategies, such as re-weighting, adversarial debiasing, and counterfactual data augmentation, it is necessary to determine how successfully these methods reduce biased outputs without also reducing model performance. In this study, we conduct empirical tests on a variety of language tasks to determine how effectively each strategy eliminates bias, how efficiently it performs computationally, and how it influences the interpretability of the model. There are further challenges that we discuss, such as the ethical implications of algorithmic bias and the trade-offs that exist between the accuracy of the model and maintaining fairness. The findings of this study provide useful recommendations on how to develop more equitable LLMs, how to enhance our understanding of the fairness of AI, and how to put bias mitigation strategies into action in the actual world.
keywords Large Language Models (LLMs), Algorithmic Bias, Fairness Algorithms, Gender Bias, Racial Bias

Introduction

Natural Language Processing (NLP) has been revolutionized by Large Language Models (LLMs) such as GPT-3, BERT, and T5. These LLMs have achieved remarkable outcomes in a variety of domains, including sentiment analysis, translation, and text production. It is common for these models to take on and exacerbate whatever cultural biases that are present in the data that they use to train themselves, even if they are capable. This indicates that the results of their work may, unintentionally, be beneficial to certain groups while being detrimental to others. The presence of gender, racial, and socioeconomic biases in LLMs raises problems about transparency, justice, and ethics when they are applied to high-stakes contexts such as hiring, education, and legal advice. It is essential to measure and remove bias in LLMs in order to construct artificial intelligence systems that are more equitable. There are a number of different fairness algorithms that have been developed, each of which has its own distinct technique, in order to overcome these biases. Counterfactual data augmentation is a promising method for reducing bias. This method involves creating data variations to balance representation. Another method is adversarial debiasing, which involves introducing an adversary to discourage biased behavior. Finally, re-weighting is a method that adjusts the importance of certain data samples. There are, unfortunately, challenges associated with the implementation of these methods. This is due to the fact that attempts to make things more equitable frequently wind up harming the accuracy of the models and the efficiency of the computers. an analysis of fairness algorithms for LLMs, with the purpose of assessing and reducing bias, with the objective of determining how effectively these algorithms accomplish this without significantly reducing the performance of the model. It is possible for researchers and practitioners who are interested in developing language models that are more egalitarian to gain something from our examination of the benefits, drawbacks, and compromises associated with each method. A further topic that we cover is the ethical repercussions of algorithmic discrimination, with a particular emphasis on the importance of transparency and accountability in addition to equity. In order to construct artificial intelligence systems that are reflective of the ideas of justice and inclusion that are held by society, it is necessary to address hidden biases. This is of utmost importance as LLMs continue to enter numerous industries. In addition to analyzing the technological and ethical obstacles that must be overcome in order to reduce bias inside LLMs, this research investigates the practical aspects of putting fairness algorithms into practice. In addition to this, it emphasizes the importance of continuing study in this field in order to produce language technologies that are more responsible. This work makes a contribution to the topic of artificial intelligence ethics.

Fairness Algorithms That Are Widely Used to Reduce Bias

It is not uncommon for LLM training data to amplify and mirror societal biases, resulting in unfair or biased outcomes. A plethora of fairness algorithms have been developed to address these biases; each of these algorithms employs a unique approach to identify and mitigate bias. This section provides an overview of several popular fairness algorithms for LLM bias reduction, including a description of each, information about how they work, and a list of their advantages and disadvantages.

1. Re-weighting Techniques

To reduce bias, re-weighting methods change the relative importance of different samples within the training set. Using re-weighting to build a more balanced dataset can lead to more equitable model outputs by giving more weight to underrepresented samples and less weight to overrepresented ones. Instead than changing the dataset itself, this method changes how the model learns from samples.

- **Advantages**: There's no need to generate new data or change the structure of the model to apply re-weighting, and it's also really simple to implement. It can be used on the go to highlight equity without a hefty computational burden.
- **Limitations**: When re-weighting too much, the model may overfit to individual samples instead of generalizing to the whole dataset, which might reduce its performance. Furthermore, it may be difficult to determine ideal weights, and these values may change depending on the job and domain.

2. Adversarial Debiasing

By inserting a second model—an adversary—into the training process, adversarial debiasing teaches the first model to detect bias. In order to "fool" the adversary into detecting less bias, the main model is trained to minimize task-related loss while simultaneously trying to do so. While still optimizing for the main job, this strategy promotes the model to eliminate bias along sensitive variables like gender or race.

- **Advantages**: When it's necessary to minimize specific biases—like racial or gender bias—adversarial debiasing works well. This method enables the training of models with biases that are specifically targeted.
- **Limitations**: Due to the potential stability problems that models may have, adversarial training can be computationally demanding and difficult to tune. In addition, labeled data for sensitive qualities is generally necessary for this method, and it could be challenging to get such data for certain domains.

3. Counterfactual Data Augmentation

In counterfactual data augmentation, sensitive factors like gender and ethnicity are changed in data samples while other contextually relevant aspects are left untouched. By training on both the original and modified versions of the data, the model is able to learn unbiased relationships with the help of this enhanced dataset. Counterfactually changing a statement like "She is a doctor" to "He is a doctor" will teach the model to eliminate gender bias in the workplace.

- **Advantages**: In natural language processing (NLP) tasks like sentiment analysis and language generation, counterfactual data augmentation is particularly useful for removing biases caused by preconceived notions. Without changing the model's structure, this method increases model fairness and generalization.
- **Limitations**: Particularly in fields with intricate cultural or language settings, creating counterfactual data can be a laborious process. Since partial augmentation might still leave residual biases, this strategy also depends on the completeness and quality of generated counterfactuals.

4. Fair Representation Learning

The goal of fair representation learning is to train the model to generate data representations that do not change depending on specific sensitive attributes. This method guarantees that sensitive attributes do not impact the model's predictions by translating data into a "fair" latent space. The basic idea behind fair representation learning is to build an unbiased representation of the data to use as a foundation for the model's predictions.

- **Advantages**: Since it functions at the representation level instead of the output level, fair representation learning can be a powerful tool for attaining task-level fairness. For complicated, multi-task models, this method usually needs little post-processing.
- **Limitations**: Given the typically opaque nature of latent representations, implementing representation learning to guarantee fairness may compromise interpretability. Furthermore, if the representations are too limited, training models to learn fair representations could raise computational costs and perhaps lower predictive accuracy.

Readings in Political Economics
ISSN: 0079-9874 & 0305-814X
Vol 60. No. 2 (2026)
https://politonomia.org.ua/rpe

5. Bias Regularization

By including a penalty term that disincentives biased results, bias regularization directly inserts fairness requirements into the model's loss function. The regularization term describes how the model strives to eliminate bias while simultaneously optimizing for task accuracy throughout training. By penalizing outputs that demonstrate racial or gender prejudice, for instance, one may train the model to decrease biased correlations by modifying the loss function.

- **Advantages**: Because it permits the insertion of certain fairness constraints, bias regularization provides a versatile and easily customizeable method to reduce biases. It works best in situations where precise measures of fairness are known.
- **Limitations**: For situations with vague or unclear fairness restrictions, it can be particularly difficult to define suitable regularization terms. When models are over-regularized, they lose some of their sensitivity to subtle data correlations, which in turn lowers their overall performance.

6. Differential Privacy

While differential privacy is mostly used to protect users' privacy, it can also help ensure that all samples are treated fairly by limiting the model's ability to become overfit to particular cases. To make sure that no single data point has a major impact on the model, differential privacy introduces controlled randomness into the data. This can aid in protecting the model against group-or sample-specific learning biases.

- **Advantages**: Fairness during model training on sensitive or imbalanced datasets can be achieved with differential privacy. It is attractive in highly regulated industries because it provides an extra degree of privacy protection.

Limitations: Because introducing noise to data might mask meaningful trends, differential privacy has the potential to lower model performance. In addition, it can be complicated to tune in a way that strikes a balance between privacy, justice, and accuracy.

The specific context, domain, and fairness needs of the application determine which fairness algorithm is best suited to mitigate biases inside LLMs. Each method has its own set of advantages and disadvantages. Because hybrid approaches frequently produce better results, combining or utilizing various methods simultaneously may increase bias mitigation. The necessity for continuous research to enhance and optimize fairness algorithms for practical use is highlighted by the fact that striking a balance between justice, model performance, and interpretability is still a significant challenge.

Conclusion

Fair and equitable AI systems must be built by addressing the inherent biases in Large Language Models (LLMs). This is particularly important as LLMs are becoming ubiquitous in various industries. This research looks at prevalent methods of fairness algorithms that try to quantify and lessen bias in LLMs. Many algorithms fall under this category, such as re-weighting, adversarial debiasing, fair representation learning, differential privacy, bias regularization, and counterfactual data augmentation. Since every method has its advantages and disadvantages, it is challenging to obtain impartial model outputs while maintaining high efficiency and interpretability. Because every algorithm's performance depends on the specific environment and bias kind, there is no one-size-fits-all solution. Applying adversarial debiasing and fair representation learning yields good results in domains where biases are closely associated with data representation; re-weighting and counterfactual augmentation are effective in applications requiring great control over sensitive attributes. It is important to select and fine-tune algorithms based on the model's intended use and fairness criteria, as each method has its own pros and cons in terms of computational efficiency, model accuracy, and fairness. To lessen bias in LLMs, a complex, multi-faceted approach is required, as this comparative analysis shows. Improvements to current algorithms, new hybrid approaches, and more transparent fairness indicators should be the focus of future research in order to expedite the practical deployment of responsible and fair LLMs. By improving and using fairness algorithms, we can build an AI ecosystem that is more inclusive and accountable. The AI community can use this information to build models that are more reflective of society's ideals and standards of conduct.

Bibliography

- Sharma, S. K. (2024). AI-Enhanced Cyber Threat Detection and Response Systems. *Shodh Sagar Journal of Artificial Intelligence and Machine Learning*, *1*(2), 43–48. https://doi.org/10.36676/ssjaiml.v1.i2.14
- Garg, A. (2024). AI for a Better World: Sustainability and Technology. *Shodh Sagar Journal of Artificial Intelligence and Machine Learning*, *1*(1), 33–38. https://doi.org/10.36676/ssjaiml.v1.i1.04
- Aravind Ayyagiri, Prof.(Dr.) Punit Goel, & A Renuka. (2024). Leveraging AI and Machine Learning for Performance Optimization in Web Applications. *Modern Dynamics: Mathematical Progressions*, *1*(2), 89–104. https://doi.org/10.36676/mdmp.v1.i2.13
- Pramod Kumar Voola, Aravind Ayyagiri, Aravindsundeep Musunuri, Anshika Aggarwal, & Shalu Jain. (2024). Leveraging GenAI for Clinical Data Analysis: Applications and Challenges in Real-Time Patient Monitoring. *Modern Dynamics: Mathematical Progressions*, *1*(2), 204–223. https://doi.org/10.36676/mdmp.v1.i2.21