Dean&Francis
ISSN 2959-6157

Bias Mitigation Techniques in Large Language Models

Junran Xue

School of Culture and Communication, Shandong University, Shandong, China
Corresponding author: 202300610051@mail.sdu.edu.cn

Abstract:

Large scale language models (LLMs) demonstrate outstanding performance and enormous potential for development, and are widely applied in people's real-life situations. However, social bias can be learned by LLM in unprocessed training data and transmitted to downstream tasks, resulting in adverse social effects and potential harm. In this article, we present a survey of bias and fairness research on Large Language Models (LLMs), categorizing the metrics and datasets used for bias assessment. Based on the elements used by the metrics in the model, they are refined into embeddings, probabilities, and generated text. The dataset is then divided into counterfactual inputs or prompts based on its structure. Afterwards, this article conducts research and organization on bias mitigation techniques based on different intervention stages: pre-processing (modifying model inputs), in-training (modifying optimization processes), intra-processing (modifying inference behavior), and post-processing (modifying model outputs). Finally, this study aims to explore in depth the key challenges that affect the fair development of large language models, and to look forward to their future evolution paths.

Keywords: Large language models (LLMs); Prejudice and fairness; Bias mitigation techniques.

1. Introduction

The rise and rapid development of Large Language Models (LLMs) have fundamentally changed language technology, ushering in a transformative era for language models. With the widespread application of models aimed at generating human like text and understanding natural language in numerous industries and fields, their influence is gradually extending to more diverse areas. However, there may be real-world challenges posed by potential biases in these models. The output quality of any algorithm is a function of the quality of the data it uses [1]. These large language models are trained based on massive and unfiltered Internet data, so they continue to carry on harmful and biased content (such as stereotypes and inappropriate speech), derogatory and exclusive language, and other derogatory actions, thus causing adverse effects and potential harm to the inherently vulnerable and marginalized communities. And these

ISSN 2959-6157

biases not only appear in language models, but also face the risk of being amplified in applications, further reinforcing the unequal system.

As people‘s understanding of the biases embedded in large language models deepens, a large number of technological studies have emerged aimed at measuring or eliminating social biases. Some articles study the fairness of deep learning, machine learning, and artificial intelligence [2-5]; Some scholars have outlined different types of social biases reflected in texts generated by corpus trained language models, ranging from gender to age, from sexual orientation to race, from religion to culture, and have measured and reduced these bias types [6]. Scholars have also revealed systematic biases in the evaluation paradigm of using large language models as judges to score and compare the response quality generated by candidate models, and proposed three strategies: Multiple Evidence Calibration, Balanced Position Calibration, and Human-in-the-Loop Calibration to form a calibration framework [7]. This article mainly summarizes the deviation evaluation indicators: embedding-based metrics (using vector hiding representations), probability-based metrics (using model-assigned token probabilities), and generated text-based metrics (using model-generated text continuations); And studied the dataset related to bias assessment, classified it into counterfactual inputs or prompts based on data structure, as well as identifying the targeted hazards and social groups; Afterwards, this article classifies bias mitigation techniques and methods in the pre-processing (changing model inputs), in-training (modifying model parameters through gradient-based updates), intra-processing (modifying inference behavior without further training), and post-processing (modifying output text generations) stages. Finally, the article points out the openness issues and challenges in future work. This article aims to focus on the impact of social bias on the future development of big language models, and promote people‘s understanding of bias in big language models, further reducing bias.

2. Classification of Bais evaluation Metrics and datasets

2.1 Index evaluation classification method based on the content used

According to the internal usage elements of the model, the existing bias evaluation indicators for large language models can be mainly divided into three taxonomies, namely embedding based, probability based, and generated text-based metrics. The embedding based metric uses dense vector representation to balance bias, mainly manifested as context dependent sentence embeddings. Probability based indicators estimate bias by analyzing the probability distribution output by the model, while text based indicators rely on the model to evaluate the text content generated based on prompts.

2.2 Dataset classification method for bias evaluation

2.2.1 Counterfactual Inputs

This article divides counterfactual input data into two categories: masked tokens and unmasked sentences. The former requires generating the most likely word, while the latter requires the model to predict the most likely sentence. This tool can be subdivided into mask tags, which means that the dataset contains sentences with blank slots, and these gaps are filled by applying a language model. In addition, the tool can also be divided into unmasked sentences, and this dataset is suitable for models to determine which pair of sentences is more likely to appear.

2.2.2 Prompts

Specifically, the evaluation will use samples from the dataset (such as the beginning of a paragraph, a sentence, or a question) as prompts, requesting the model to provide subsequent content or answers. The first step is sentence completion, which includes the beginning of a sentence and is completed by a language model. Next is question-answering, which involves generating relevant prompt datasets within the question and answer framework.

2.2.3 Notice

Firstly, caution should be paid when dealing with challenges in construction, content, and ecological validity. The evaluation of the dataset should include two key aspects: firstly, whether it clearly defines and explains the issue of power imbalance to be measured; Secondly, whether the definition is aligned with the downstream deviation assessment objectives. Then when constructing and selecting datasets, it is suggested to ensure their universality and applicability.

3. Taxonomy of Techniques for Bias Mitigation

3.1 Pre-Processing Mitigation

3.1.1 Data Augmentation

This method corrects bias by adding new examples to the training data, aiming to expand the coverage distribution of underrepresented social groups. However, data aug-

mentation techniques that use words to replace terminology are not only difficult to expand, but may also introduce factual errors.

The data balancing method aims to balance the representation of different social groups, and counterfactual data augmentation (CDA) is one of the main methods of these augmentation techniques. Ghanbarzadeh et al. generated training examples by masking gendered vocabularies and predicting replacement words using language models, while retaining the labels of the original sentences for fine-tuning [8].

In selective replacement methods, some techniques provide alternative solutions for CDA, thereby improving data efficiency and mitigating bias on the most effective training samples. In order to more effectively promote gender equality, Zayed et al. proposed a selective data augmentation strategy. The core of this strategy is to only introduce counterfactual examples that have made the most significant contributions to gender equality and do not contain any gender stereotypes [9].

Based on the mixup technique proposed by scholars, interpolation method expands the distribution of training data by interpolating the counterfactual enhanced training samples with the original samples and their labels. Yu et al. proposed the Mix Bias method and used it to reduce gender stereotypes in the integrated use of multiple corpora [10].

3.1.2 Data Filtering and Reweighting

Although data augmentation helps reduce bias, this method is limited by incomplete word pair lists and may result in grammatical errors during term replacement. Unlike data augmentation, the intervention targets of data filtering and reweighting techniques are not new examples, but examples with specific attributes in existing datasets.

In the dataset filtering method, select a subset of the dataset to increase its influence in the fine-tuning process; Other data selection methods focus on selecting the most biased samples for neutralization or transition. As Thakur et al. proposed a neutralization method to alleviate gender bias, they carefully selected a very small number of the most biased samples, which were generated by masking gender related vocabulary in candidate samples and then having a pre trained model predict the masked vocabulary. In the fine-tuning process, the authors replaced gender related vocabulary with neutral or egalitarian substitute words [11].

In the instance reweighting method, the core operation of instance reweighting method is to adjust the weights of different training samples during the training process. Han and Baldwin used instance reweighting method to make the total weight of each category equal during the training process. Specifically, the loss weights they calculate for each instance are inversely correlated with the distribution of their labels and protected attributes [12].

3.1.3 Data Generation

The above methods are all limited by a key issue, which is that their effectiveness depends on manually defined examples for each deviation dimension. However, the definitions of these examples themselves are not absolute and may vary depending on the context, application, and even expectations of ‚correct behavior‘. Unlike modifying existing datasets, data generation techniques create a new dataset that is carefully planned to reflect a predetermined set of standards or features.

In Exemplary examples, the new dataset can simulate the expected output behavior by providing high-quality, carefully generated examples; In word lists, word replacement techniques such as CDA and CDS rely on word pair lists, and multiple studies have proposed word lists related to social groups, involving gender, race, dialect, and other social group terms. To improve universality, Omrani et al. proposed a theoretical framework to understand stereotypes from the dimensions of ‘enthusiasm’ and ‘ability’, rather than based on specific demographic or social groups [13]. This study generated word lists corresponding to these two categories, which can to some extent replace population-based word lists, such as gender related words, in bias mitigation tasks.

3.1.4 Instruction Tuning

In text generation, the model can be instructed to avoid using biased language by modifying input or prompts. To achieve precise control over the generation process, instruction fine-tuning achieves this goal by adding additional static or trainable markers to the input; In the modified prompting language, people can add text instructions or trigger words in the prompt to generate unbiased output; In control tokens, instead of adding explanatory language before input, control tags corresponding to the prompt classification can be directly added. Based on the learning effect of the model being able to associate control labels with input categories, the inference stage can guide the generation process by setting specific labels.

3.1.5 Discussion and Limitations of Pre-Processing Mitigation

The effectiveness of pre-processing mitigation measures may be limited and relies on questionable additions. Data augmentation techniques use words to replace terminology, which is not only difficult to expand but may also introduce factual errors; The process of data filtering, reweighting, and generation may also encounter the aforementioned challenges, and may introduce new distribution

ISSN 2959-6157

imbalances in the dataset when dealing with misleading word lists and alternative indicators of social groups.

3.2 In-Training Mitigation

3.2.1 Architecture Modification or Filtering Model Parameters

The former achieves optimization by actively adjusting the underlying configuration of the model, including the number, size, and type of layers, encoders, and decoders. The latter, in addition to fine-tuning techniques that reduce bias by simply updating model parameters, focuses on filtering or removing specific parameters during or after model training or fine-tuning.

3.2.2 Loss Function Modification

Modifying the loss function through new equilibrium objectives, regularization constraints, or other training paradigms such as contrastive learning, adversarial learning, and reinforcement learning may make the output semantics and typical terms irrelevant to social groups. The form of the loss function or the reward given by reinforcement learning implicitly assumes a definition of fairness, such as the concept of invariance about a certain social group, acting subtly and differently on different social groups.

In the equalizing objective, by modifying the loss function, unnecessary associations between social groups and stereotyped vocabulary can be directly weakened, ultimately ensuring the independence of the model's predicted output from social groups; In embedding, distance-based methods, projection-based methods, and mutual information-based methods can solve the bias problem in encoder hidden representations; In contrastive learning, traditional contrastive learning techniques learn similarities or differences within a dataset by arranging unlabeled data in pairs. As an important technical approach to alleviate bias issues in large language models, the contrastive loss function has been applied in supervised learning scenarios by processing sentence pairs with and without bias, and maximizing the similarity with unbiased sentences; In the context of adversarial learning, both the predictor and the attacker are trained simultaneously. The adversarial goal of a predictor is to minimize its own losses while maximizing the losses of the attacker, thus forming an effective game dynamics; In reinforcement learning, this technology optimizes the unbiased nature of generated content directly through a reinforcement learning framework. The reward mechanism can come from the prediction of the next word at the word level or the classification results at the sentence level.

3.2.3 Selective Parameter Updating

Due to the fact that the scale of fine-tuning data sources is usually smaller than the original training data, secondary training may cause the model to forget previously learned information, resulting in a loss of downstream performance of the model. This phenomenon is known as catastrophic forgetting. An effective approach to alleviate catastrophic forgetting is to use parameter efficient fine-tuning methods. This type of method freezes the vast majority of pre-training parameters and updates only a small portion, allowing the model to efficiently adapt to new tasks while maximizing the inheritance of existing knowledge and significantly reducing computational costs.

3.2.4 Discussion and Limitations of In-Training Mitigation

In addition to selective parameter update methods, mitigation measures taken during the training process may also disrupt the pre-trained language model; Due to the small size of the fine-tuning dataset compared to the original training data, it may lead to catastrophic forgetting; Beyond computational limitations, mitigation measures during the training process may vary depending on different modeling mechanisms, and their effectiveness may differ.

3.3 Intra-Processing Mitigation

3.3.1 Decoding Strategy Modification

Decoding describes the process of generating output marker sequences, and by applying fairness constraints to modify the decoding algorithm, biased language can be suppressed. By selecting constraints, changing the probability distribution of markers, or integrating auxiliary bias detection models, the probability of the next word or sequence can be modified afterwards. The limitations of mitigation measures in the processing mainly focus on decoding strategy modification, which faces the challenge of balancing bias mitigation with diversified outputs. A key challenge in successfully applying these methods is to ensure that the toxic label classifiers they rely on have both high accuracy and low bias.

3.3.2 Weight Redistribution

As a intervention, the weights (especially attention weights) of the trained model can be directly modified to adjust the model's attention to biased words or phrases without the need for retraining.

3.3.3 Modular Debiasing Networks

The limitation of most methods for removing bias during training is that they only target biases in a single dimension, but in different application scenarios or protected attributes, multiple ways of removing bias are often required. The modular approach does not directly modify

the original model, but instead builds independent de biased modules. This module can flexibly integrate with pre trained models, making it suitable for various downstream tasks. Hauzenberger et al. proposed a technique for training multiple sub networks, which can be modularly applied during inference to eliminate a specific set of biases [14].

3.4 Post-processing Mitigation

3.4.1 Rewriting

The rewriting strategy detects harmful words through rules or neural network-based rewriting algorithms and replaces them with more positive or representative words, considering the fully generated output.

In keyword replacement, this method identifies biased tokens and predicts appropriate replacement words for them to achieve bias mitigation while ensuring the integrity and consistency of the original text's content and style; In machine translation, translating biased original sentences into neutral or unbiased target sentences. This can be constructed as a machine translation task and trained on parallel corpora. Amrhein et al. differ from using biased sentences as source sentences to generate parallel corpora by using reverse augmentation to filter out gender fair sentences from large corpora, and then adding biases to generate source sentences [15].

3.4.2 Discussion and Limitations of Post-processing Mitigation

Post-processing mitigations do not require access to trainable models, making these methods suitable for black box operations. The bias risk of rewriting technology fundamentally stems from its decision-making process determining which outputs need to be rewritten. This judgment is essentially subjective and carries a specific value orientation. Similar to the potential harm that toxicity and sentiment classifiers may bring, special attention should be paid to ensuring that the language styles of certain social groups are not excessively labeled and rewritten.

4. Challenges and Prospects

4.1 Addressing Power Imbalances

Firstly, people should pay attention to marginalized communities, and relevant researchers and practitioners should incorporate the needs of vulnerable groups into technology design from the beginning, placing marginalized groups at the forefront of language model decision-making and system development, understanding the role of technological solutions, and breaking harmful power imbalances; Secondly, people should develop participatory research designs that involve community members in the research process, which is conducive to fully understanding and to some extent endorsing their needs. Afterwards, relevant researchers and practitioners should not rely on vague concepts of behavior that big language models should have in society. Instead, they should establish more rigorous theories of social change based on the background of relevant humanistic history, in order to complete the transformation of values and assumptions. Finally, people should expand their language resources, prioritize inclusiveness over convenience, step out of the current research context, collect more language resources, such as data on different languages and their dialects, and deepen their understanding of various language features and forms of bias representation.

4.2 Conceptualizing Fairness for NLP

4.2.1 Reconstruct the definition of social groups

Dividing social groups is usually done to assess differences, but it legitimizes social construction and reinforces power disparities. Future research can develop more accurate evaluation indicators, explore the specific harm caused by large language models to different groups, and develop more comprehensive bias mitigation techniques based on a wide range of social groups.

4.2.2 Recognizing Distinct Social Groups

The relevant research needs to achieve a paradigm shift from 'eliminating group differences' to 'deconstructing structural biases'. Specifically, we should delve deeper into the sources of bias, understand its group specific mechanisms, and ultimately develop strategies that can intervene against historical and structural forces.

4.3 Optimization evaluation principles

At present, the evaluation of fairness in large-scale language models is relatively insufficient, especially for some non open source large-scale and rigorous models. People can only quantify deviations based on the response results of the models. How to more accurately characterize the deviations in model generation is the basis for evaluation. Therefore, future research needs to start from a diversified perspective, expand the application of statistical principles, and deepen automated measurement techniques to innovate quantitative methods for bias against large language models. In addition, some widely used evaluation datasets have reliability and validity issues that we should continuously pay attention to and explore.

ISSN 2959-6157

4.4 Improving Mitigation Efforts

Future research can explore strategies for expanding resources to alleviate bias bottlenecks without neglecting the value of human-computer interaction and community participation frameworks, thereby achieving scalability. And people can consider developing hybrid technologies. Currently, most bias mitigation technologies only target a single intervention stage. Future research can explore mitigation technologies that reduce bias in multiple or all intervention stages to improve effectiveness. In addition, it can be considered to focus on fairness in data processing and model architecture during the model development stage, with particular emphasis on training data as the main source of bias. Developers should be encouraged to invest resources in data processing to alleviate or eliminate social bias at its root.

4.5 Exploring Theoretical Limits and Limitations

In the trade-off analysis between performance and fairness, the core mechanism of such technologies typically involves a hyperparameter that directly regulates the balance between the model's original performance and depolarization strength, and future research can delve deeper into and elucidate the trade-off between performance and fairness; To some extent, technological solutions simplify the broader history and context that allows structural system oppression to exist, thereby maintaining and perpetuating the root causes of inequality and injustice, creating a superficial sense of relief and gradual progress, but failing to question and address broader systemic issues. Therefore, technological solutions should be combined with broader social actions to jointly combat the power hierarchy that weakens and suppresses marginalized groups.

5. Conclusion

In summary, this article provides a review of bias assessment and mitigation techniques for large-scale language models, gathers relevant research results, and elucidates the current research landscape. This paper mainly elaborates on the metrics and datasets used for bias evaluation, classifies and discusses the indicators used to evaluate model bias from different levels, and classifies the datasets used for bias assessment. In addition, this study classified the techniques used for bias mitigation based on the intervention stage. At the end of the article, the author outlines the key open questions and challenges that future research should address, with the hope that these studies can enhance people's understanding and knowledge of technical measurement and mitigation methods for bias persistence in large language models, and promote further development of research in related fields.

References

[1] Baeza-Yates R. Data and algorithmic bias in the web// Proceedings of the 8th ACM Conference on Web Science. 2016: 1-1.

[2] Bansal R. A survey on bias and fairness in natural language processing. arXiv preprint arXiv:2204.09591, 2022.

[3] Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 2021, 54(6): 1-35.

[4] Gohar U, Cheng L. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. arXiv preprint arXiv:2305.06969, 2023.

[5] Ferrara E. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci, 2024, 6(1): 3.

[6] Navigli R, Conia S, Ross B. Biases in large language models: origins, inventory, and discussion. ACM Journal of Data and Information Quality, 2023, 15(2): 1-21.

[7] Wang P, Li L, Chen L, et al. Large language models are not fair evaluators. arXiv preprint arXiv:2305.17926, 2023.

[8] Ghanbarzadeh S, Huang Y, Palangi H, et al. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. arXiv preprint arXiv:2307.10522, 2023.

[9] Zayed A, Parthasarathi P, Mordido G, et al. Deep learning on a healthy data diet: Finding important examples for fairness// Proceedings of the AAAI Conference on Artificial Intelligence. 2023, 37(12): 14593-14601.

[10] Yu L, Mao Y, Wu J, et al. Mixup-based unified framework to overcome gender bias resurgence//Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2023: 1755-1759.

[11] Thakur H, Jain A, Vaddamanu P, et al. Language models get a gender makeover: Mitigating gender bias with few-shot data interventions. arXiv preprint arXiv:2306.04597, 2023.

[12] Han X, Baldwin T, Cohn T. Balancing out bias: Achieving fairness through balanced training. arXiv preprint arXiv:2109.08253, 2021.

[13] Omrani A, Salkhordeh_Ziabari A, Yu C, et al. Social-group-agnostic bias mitigation via the stereotype content model. Association for Computational Linguistics, 2023.

[14] Hauzenberger L, Masoudian S, Kumar D, et al. Modular and on-demand bias mitigation with attribute-removal subnetworks. arXiv preprint arXiv:2205.15171, 2022

[15] Amrhein C, Schottmann F, Sennrich R, et al. Exploiting biased models to de-bias text: A gender-fair rewriting model. arXiv preprint arXiv:2305.11140, 2023.