computers

MDPI

Article

Mitigating Large Language Model Bias: Automated Dataset Augmentation and Prejudice Quantification

Devam Mondal *,† and Carlo Lipizzi *,†

Center for Complex Systems and Enterprises, Stevens Institute of Technology, Hoboken, NJ 07030, USA

* Correspondence: dmondal@stevens.edu (D.M.); clipizzi@stevens.edu (C.L.)

† These authors contributed equally to this work.

Abstract: Despite the growing capabilities of large language models, concerns exist about the biases they develop. In this paper, we propose a novel, automated mechanism for debiasing through specified dataset augmentation in the lens of bias producers that can be useful in a variety of industries, especially ones that are "restricted" and have limited data. We consider that bias can occur due to intrinsic model architecture and dataset quality. The two aspects are evaluated using two different metrics we created. We show that our dataset augmentation algorithm reduces bias as measured by our metrics. Our code can be found on an online GitHub repository.

Keywords: natural language processing; large language models; dataset augmentation; computational social science

check for updates

Citation: Mondal, D.; Lipizzi, C. Mitigating Large Language Model Bias: Automated Dataset Augmentation and Prejudice Quantification. *Computers* **2024**, *13*, 141. https://doi.org/10.3390/computers13060141

Academic Editors: Yorghos Voutos, Akrivi Krouska, Christos Troussas, Phivos Mylonas and Cleo Sgouropoulou

Received: 30 April 2024
Revised: 28 May 2024
Accepted: 30 May 2024
Published: 4 June 2024

1. Introduction

In recent years, large language models (LLMs) have revolutionized the field of natural language processing, enabling remarkable advancements in tasks such as text generation, translation, and sentiment analysis. These models, driven by their immense size and training on vast textual corpora, have exhibited impressive capabilities in understanding and generating human-like text. However, beneath the surface of their remarkable achievements lies a profound challenge—the omnipresent issue of bias, defined as the "systematic error that arises as a result of a given choice of the texts used to train language models" by Navigli et al. Bias in LLMs, often derived from the biases present in the data they are trained on and the inherent architecture of the model, has raised ethical concerns and has the potential to reinforce harmful stereotypes and misinformation.

This paper aims to address mechanisms to reduce this bias through a comprehensive approach that utilizes both dataset augmentation and metric creation. Our algorithm and metrics can be used to quantify and mitigate large language model bias in various industries, ranging from education to agriculture. However, our approach proves most effective when mitigating bias in "restricted industries", more specifically, industries where data are limited due to confidentiality or availability of information. Examples of restricted industries include the defense, medical, and financial fields. Our research takes on a two-fold approach because of the documented increase in LLM perplexity and greater embedding of prejudice when trained on small datasets or datasets with biased content.

First, we explore automated dataset augmentation to mitigate bias, using the concept of a bias producer to describe broad creators of bias, such as ethnicity or sexuality, and biasers that serve as specific examples. These bias producers can be elements of generic bias (such as gender or race) or be industry-specific. Such an augmentation algorithm is autonomous and, therefore, reduces the implicit bias introduced by annotators and other human means of defining bias. It can also be extended to other languages when provided with the correct embeddings. We also define two new metrics for quantifying bias about both datasets and models, the db-index and mb-index, respectively. These metrics provide a crucial feedback loop for researchers and developers to monitor, analyze, and minimize LLMs' bias.

We use the dataset augmentation procedure to create more equitable datasets and then fine-tune LLMs on these datasets. We use the aforementioned metrics to demonstrate the effectiveness of our algorithm in reducing biased responses in these LLMs.

2. Literature Review

As mentioned previously, in recent years, the proliferation of large language models (LLMs) has revolutionized natural language processing tasks, increasing efficiency in various use cases. However, concerns about biases embedded within these models from the external corpus and human knowledge base have prompted significant research efforts to categorize and mitigate biases. The existing literature surrounding this topic can be clustered into five groups:

2.1. Cluster 1: Types and Examples of Bias in the Realm of LLMs

A large amount of literature has laid the foundation for categorizing bias in the realm of an LLM environment. Hovy et al. established five sources of bias in natural language processing, asserting that the process of data selection, dataset annotation, input representations (word embeddings), model architecture, and research design can instill prejudice within an LLM [1]. Moreover, Navigli et al. attribute dataset selection and quality to being the single most significant "producer" of prejudice, with unbalanced topics, outdated text in corpora, and narrow-minded dataset creators instilling bias within LLMs [2]. Navigli et al. also define attributes LLMs exhibit bias against, such as age, culture, nationality, and religion, providing examples for each generated by GPT-2, GPT-3, and the BLOOM transformer models [2].

However, the major uncovered topic in this cluster is a method of quantifying the bias. Categorizing the biases and recognizing their source provides a qualitative framework to address them but does not enable a quantitative method of treating them.

2.2. Cluster 2: Bias in the Application of LLMs in Restricted Industries

A large amount of literature illustrates the various biases of LLMs when applied to "restricted industries". Here, we define "restricted industries" with data that is unique in nature and confidential. Li et al., for instance, explore the various biases LLMs exhibit in the medical field, with many of these models being trained primarily on English corpora from developed countries, therefore biasing understanding of disease towards high-income nations [3]. Moreover, Mikhailov explores various biases of LLMs in the military, including the possibilities of offensive hallucinations [4].

Much like Cluster 1, the challenge in this cluster lies in a lack of a quantitative approach to measure the amount of bias for a restricted industry LLM or dataset. Without this framework, tackling the bias with an algorithm would not be possible.

2.3. Cluster 3: Dataset Bias in the Realm of LLMs

Literature also provides insight into how datasets that are misrepresentative, poor in quality, or rely on subjective assessments for creation can instill bias within LLMs. Wiegand et al. demonstrated the issues with the Waseem dataset regarding detecting abusive language in social media, with the LLMs becoming biased towards sports [5]. As the majority of abusive tweets in the Waseem dataset were disproportionately related to sports, the LLM associated abuse with words such as commentators and announcers. Moreover, the dataset's tweets were skewed towards three authors, with this authorial bias becoming embedded within the LLM. Additionally, Geva et al. found that datasets reliant on annotators caused LLMs to develop bias, being able to "pick up" annotators that produced large numbers of samples (demonstrated by better model performance when annotator ID is supplied to the model) [6]. Yet, these LLMs, reliant on datasets with annotator subjectivity, fail to generalize to new examples created by new annotators.

The challenge in this cluster is a reliance on human annotators. As mentioned, these annotators introduce subjectivity when labeling text for arbitrary concepts, introducing more unintended bias in fine-tuning large language models [7]. Novel mechanisms that aim to remediate dataset bias must do so autonomously, without human intervention.

2.4. Cluster 4: Inherent Bias in LLM Architectures

In addition to the dataset aspect of bias, much literature describes how specific LLM architectures, particularly the long short-term memory (LSTM) and Transformer, can exhibit bias towards particular characteristics of the human knowledge base. For example, White et al. demonstrated through the creation of artificial languages how LSTMs do not have any preference for word order, yet Transformer architectures prefer head-final languages [8].

The challenge in this cluster is, once again, quantification. A framework to measure this bias does not exist.

2.5. Cluster 5: Addressing and Remediating Bias

In response, a large amount of literature aims to mitigate bias associated with LLMs. For example, Lee et al. aimed to reduce LLM social bias against certain Korean demographic groups by creating KOSBI, a high-quality dataset with contexts (generated through rejection sampling) and sentences from the contexts [9]. Both were then annotated as safe or unsafe. Dixon et al. aimed to reduce bias by mining additional corpora from an unbiased source (Wikipedia), then created a 'pinned' metric to measure fairness based on area-under-curve (AUC) [10]. Renaldi et al. explored debiasing through domain adaptation, more specifically through fine-tuning, parameters freezing, and attention matrix training, using metrics like StereoSet (Nadeem et al.) and GLUE (Wang et al.) to measure bias and LLM quality, respectively [11–13]. Guo et al. proposed a distribution alignment loss (DAL) to mitigate bias, first generating biased prompts, then using the DAL to reduce Jensen-Shannon divergence (JSD) between distributions for a masked token when other critical parts of the prompt are changed [14]. Huang et al. suggested eliminating bias by reducing the Wasserstein-1 distance between sentiment distributions of different token collections (each of a different demographic) in a phrase [15]. This process, named counterfactual evaluation, could be done by embedding regularization (where cosine similarity between two token collections would be reduced) or sentiment regularization.

The gaps in this cluster are primarily focused on the dataset. Most approaches aim to correct intrinsic aspects of the LLM itself, and approaches utilizing datasets to debias rely on annotators, which may introduce inherent bias.

Therefore, in this paper, we address the gaps in the applications of Cluster 5 in Cluster 2. More specifically, we propose a novel debiasing mechanism aimed at LLMs in various industries but most effective for "restricted industries" through automated dataset augmentation. Additionally, we propose a novel quantitative measure of model bias by taking performance into account, as well as another quantitative measure that assesses dataset bias.

3. Approach

In this section, we hope to provide background and motivation for the metrics and dataset augmentation algorithm we created to reduce bias in LLMs.

3.1. Dataset Augmentation

To reduce bias in LLMs, we use the concept of a *bias producer*, a social or industry "lens" initially containing a set of words known as *biasers*. Formally, if β is a bias producer, at the end of our process, there will be a biased set b, where $b_1, b_2, b_3 \ldots \in b$ are all examples of β.

For example, given the bias producer of "ethnicity", the biaser set contains examples of ethnicities, such as "Nigerian", "African–American", and "Indian". Determining the bias producer and the number of biasers is up to the user, industry, and use case of the LLM.

Each entry in the dataset is then swept for examples of a bias producer using named-entity recognition. When the first biaser is met, the entry is recopied, and the biaser is changed with another set member. This process repeats $|b| - 1$ times, with all elements of b filling in the biaser. This mechanism allows the dataset to broaden in size without reliance on external data. Such an approach that eliminates reliance on external, unbiased data is especially beneficial for "restricted industries" where data are confidential.

After this, each entry undergoes content morphism, where each entry is upshifted through contextual word embedding sentence augmentation and downshifted to a summary to better capture human language. Both are then added to the dataset.

For example, consider dataset d with the entry "Indians are helpful in math". We define the bias producer β as "ethnicity" and establish a two-element biaser set (b) of $\{$"*American*", "*Swedish*"$\}$. Analyzing the dataset's first entry with named-entity recognition shows that "Indian" is an example of a bias producer ("ethnicity"). A new entry (Entry 1) is created where the first element of the biaser set ("American") replaces "Indian". Another entry (Entry 2), which is the summarization of Entry 1, is created. This summarization is made using the spacY library and a BART model, more specifically through spacY's abstractive summarization function. Another entry (Entry 3), an augmented version of Entry 1, is created. Entries 1, 2, and 3 are then added to the dataset. The above process is repeated for all other members of the biaser set.

Please note that this method eliminates the need for annotation, which can introduce subjectivity and bias. Furthermore, unlike most augmentation processes, such a process is targeted, addressing a single source of bias through the phrasing of the bias producer. There may be subjectivity in choosing the bias producer and biasers, but they must be subjective because they depend on the LLM's usage and industry. Additionally, bias evolves due to evolution in cultural norms; bias producers and biasers should therefore be dynamic and open to change.

3.2. LLM Bias Classification

To assess the performance of models after being fine-tuned on an augmented debiased dataset, we propose a new metric called the mb-index. Because LLMs for the same industry may be fine-tuned on different datasets of varying sizes, this normalized metric for dataset size provides an "index of bias for performance per data entry trained on". Formally, given a dataset d, perplexity $p(d)$, and stereotype score $s(d)$, mb-index is defined as:

$$\frac{p(d) * s(d)}{|d|} \tag{1}$$

The stereotype score, a new metric, is a result derived from an extension of the Intersentence Context Association Test Nadeem et al. proposed in conjunction with the StereoSet score [12]. However, rather than the LLM "picking" the best answer to the context provided in a multiple-choice setting, it generates a 30-character continuation of the context, defined as I.

Given three choices, one reinforcing a stereotype (A), the other reinforcing the anti-stereotype (B), and the third being a nonsensical sentence (C), the cosine similarity between the embedding-based vectorized version of I and the embedding-based vectorized version of each option is calculated. The greatest similarity is then used to classify the generated text as stereotypical, anti-stereotypical, or nonsensical. This process is continued through each entry of the StereoSet dataset.

From this definition, we assert that the stereotype score is the proportion of continuations classified as stereotypical for all continuations not marked as nonsensical. Defining T_A as the total number of stereotypical continuations (I_A) and T_B as the total number of anti-stereotypical continuations (I_B), the stereotype score is:

$$\frac{T_A}{T_A + T_B} \tag{2}$$

For an ideal LLM, the stereotype score should be near 0, as the model's responses should not be similar to stereotypical responses. It is fine if the model's responses are similar to the anti-stereotypical responses, as the training procedure promotes the model to "think anti-stereotypically", where various qualities can be attributed to a lens.

A "good" mb-index is close to 0, where the stereotype score is ideal, and perplexity is ideally minimized to 0.

3.3. Dataset Bias Classification

However, bias is not just limited to the model but also a given dataset. Thus, we propose another metric called the db-index. This metric quantifies the bias present in a dataset using cosine similarity. Cosine similarity is a numerical value that shows how "close" two vectors are in a space. Given vectors A and B, it is defined as:

$$\frac{A \cdot B}{||A||||B||} \tag{3}$$

Given a target dataset (or portion/cluster of) d_t and a comparison dataset d_c (containing biased and abusive language), a random entry $e_c \in d_c$ is picked. Then, cosine similarity, $d_{\cos\theta}$, between the vector of the comparison entry and each entry $e_d \in d_t$ is calculated:

$$d_{\cos\theta} = \sum_{i=1}^{|d_t|} \frac{e_{d,i} \cdot e_c}{||e_{d,i}||||e_c||} \tag{4}$$

To obtain a dataset's db-index (db), we cluster the dataset and find each cluster's db-index (db_c) by dividing the cluster's $d_{\cos\theta}$ by the cluster size. We then find the total dataset's db-index by averaging each cluster's db-index.

More specifically, we first convert each entry in the corpus into an embedding vector. We then segment the corpus into semantically homogeneous groups using k-means clustering. Because the initial number of clusters is difficult to set, we run clustering with an arbitrary value of 4 clusters. Then, we conduct hyperparameter tuning through grid search to optimize the number of semantically homogeneous clusters. Then, k-means clustering is undertaken again. Figure 1 below shows this process:

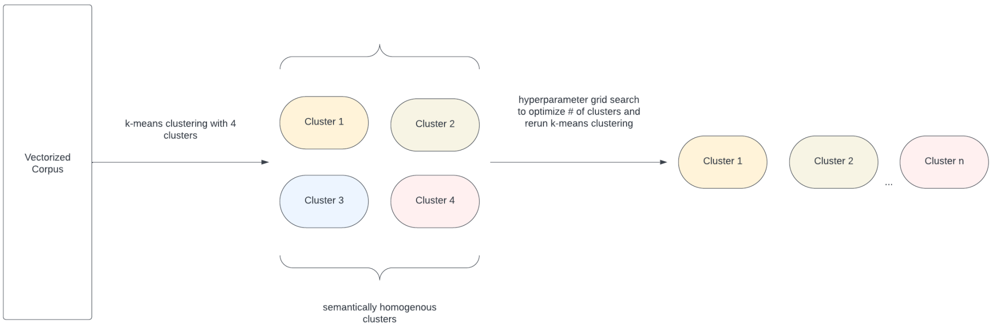
Vectorized Corpus
k-means clustering with 4 clusters
Cluster 1
Cluster 2
Cluster 3
Cluster 4
semantically homogenous clusters
hyperparameter grid search to optimize # of clusters and rerun k-means clustering
Cluster 1
Cluster 2
...
Cluster n

Figure 1. Obtaining semantically homogeneous clusters through k-means clustering and hyperparameter grid search.

$d_{\cos\theta}$ is then found for each cluster c using the above algorithm. $d_{\cos\theta}$ is then divided by the cluster's size $|c|$ to yield the cluster's db-index.

$$db_c = \frac{d_{\cos\theta}}{|c|} \tag{5}$$

The total dataset db-index is then found by averaging all the clusters' db-indices.

4. Materials, Methods, and Results

We first began with the dataset augmentation procedure. Given that the augmentation method is most effective when addressing data in "restricted industries", we sought to augment a dataset of government reports about military and other classified content. This dataset initially contains around 17,500 entries. To imitate a situation with a lack of available data, our dataset augmentation method focused on two small subsets of the dataset: Sample A, containing ten elements, and Sample B, containing 50 elements. Such dataset sizes mimic situations in the aforementioned industries where there is a lack of substantial data due to its unique and confidential nature.

We then conducted dataset augmentation with the bias producer β of ethnicity, the biaser set containing 20 different races generated through a LangChain process with Chat-GPT3.5 content generation. The above process, detailed in Part 1 of Approach, was used twice. First, Sample A was augmented to produce a dataset of 1641 elements in size. Second, Sample B was augmented to produce a dataset of size 4248 elements. Figure 2 provides a flow chart of this entire process:

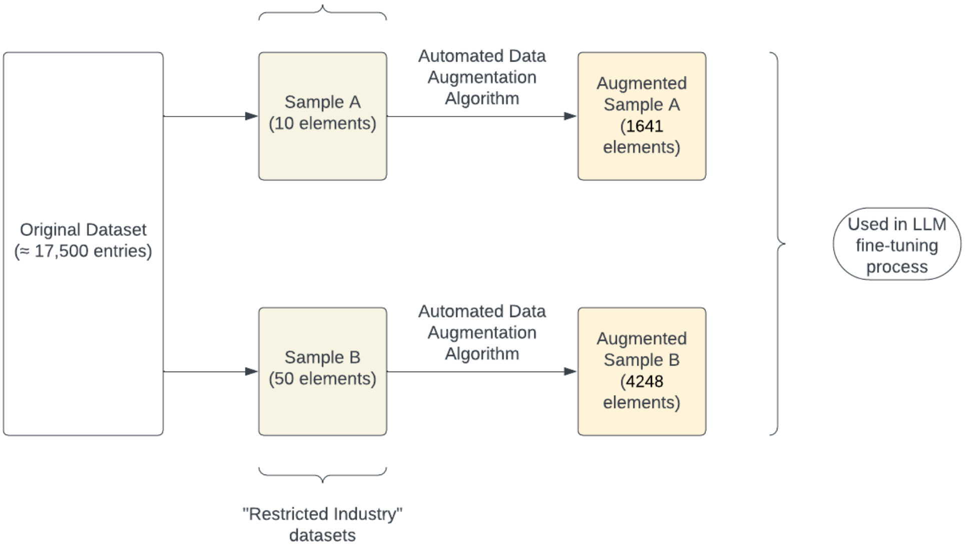

Figure 2. Sourcing of datasets used for subsequent LLM fine-tuning. Note the four datasets (two augmented and two directly taken from the original dataset).

We then calculated the db-index of both Sample A and Sample B and their augmented counterparts concerning implicit bias five times each and then averaged the results to yield the final db-index. Table 1 shows the results.

Table 1. The db-indices of the four datasets (calculated for the two augmented and two normal datasets).

Dataset	db-Index
Sample A	0.56
Sample A (augmented)	0.49
Sample B	0.71
Sample B (augmented)	0.65

Next, four LLMs were chosen to be fine-tuned on the augmented data and original data. All four LLMs were Meta AI's LLaMa 13b Chat models with HuggingFace formatted weights and biases. Each LLM was fine-tuned on different data with the following:

- LLM A was fine-tuned on a subsection of the original dataset containing the same number of samples as augmented Sample A.

- LLM B was fine-tuned on a subsection of the original dataset containing the same number of samples as augmented Sample B.
- LLM C was fine-tuned on augmented Sample A.
- LLM D was fine-tuned on augmented Sample B.

All LLMs were fine-tuned in a causal language modeling process (next-token prediction) and through QLoRA (Quantized Low-Rank Adaptation). After fine-tuning, each LLM's perplexity, stereotype score, and mb-index were calculated using the original dataset as the reference. Additionally, we normalize all mb-indices through min-max normalization to enable easier comparison. Table 2 below shows the results:

Table 2. Performance (perplexity) and bias metrics (stereotype score, mb-index, and normalized mb-index) for the four LLMs.

LLM	Perplexity	Stereotype Score	mb-Index	Normalized mb-Index
A *	6.4660	0.55	2.16×10^{-3}	1
B **	6.2920	0.52	7.65×10^{-4}	0.15
C ***	4.9290	0.45	1.36×10^{-3}	0.51
D ****	4.9290	0.45	5.24×10^{-4}	0

* Fine-tuned on the subsection of the original dataset containing the same number of samples as augmented Sample A. ** Fine-tuned on the subsection of the original dataset containing the same number of samples as augmented Sample B. *** Fine-tuned on augmented Sample A. **** Fine-tuned on augmented Sample B.

5. Discussion

As seen in Table 1, the automated dataset algorithm can reduce the db-index of a dataset. The augmented datasets substantially decreased the db-index compared to their original counterparts.

As seen in Table 2, LLMs C and D, fine-tuned on the augmented datasets, have less perplexity than LLMs A and B. This suggests that augmented datasets created through the algorithm mentioned above can increase LLM performance.

Additionally, the stereotype scores for LLMs C and D are also less compared to LLMs A and B, suggesting that the dataset augmentation mechanism reliant on a bias producer "lens" that substitutes in members of a biaser set is effective at removing LLM tendency towards stereotypical responses.

Therefore, because LLMs C and D have lower perplexities and stereotype scores, they are better performing and less biased in the quantitative measures described above due to a lower mb-index. Even though the mb-indices, which are absolute rather than relative values, are small, they are significant and can be compared because all the LLMs were fine-tuned on the same dataset.

6. Limitations and Further Research

We pick a random entry from the comparison dataset to compute the db-index for the approach. However, this may produce inaccurate results if specific data entries in the comparison dataset are significantly more biased than others. Therefore, a feasible solution that is interesting to explore could be creating a "distribution of bias" for the comparison dataset and then avoiding outliers when picking the random entry.

The datasets used to calculate the db-index were on a scale of tens of thousands due to the limits on the available public data. It would be beneficial to see db-indices being produced for datasets containing millions of records to assess the efficiency of our algorithm.

Additionally, k-means clustering was used to create semantically homogeneous groups when calculating db-index. Given that the embedding vectors used had 300 dimensions, and k-means clustering utilizes Euclidean distance, there exists a possibility that the groups created are not accurate due to the curse of dimensionality. It would be interesting to see other clustering methods used or k-means clustering after the embeddings undergo principal component analysis to reduce dimension size and resulting performance.

Additionally, the LLMs fine-tuned were medium-sized (13 billion parameters). It would be beneficial to see larger LLMs (70 billion parameters or more) being fine-tuned on datasets augmented through our bias-reducing approach and their mb-index performance.

Also, it would be interesting to explore bias reduction in other non-English languages using this framework, given the correct embeddings.

7. Conclusions

A pressing matter in the ever-evolving field of natural language processing is the bias present in large language models. In this paper, we outline a mechanism to tackle bias caused by training and fine-tuning data within large language models through an automated augmentation algorithm based on bias producers. We also provide ways to quantify bias inherent to datasets and large language models through the db-index and mb-index accordingly. It is important to note that our mechanism should be able to extrapolate to other languages and not just be restricted to English. The only requirement is that the correct embeddings are provided.

We hope to continue democratizing our work in this paper by creating an online platform where natural language processing enthusiasts and professionals can see the bias within their large language models and datasets before implementing them in their systems.

Author Contributions: Conceptualization, D.M. and C.L.; methodology, D.M. and C.L.; software, D.M. and C.L.; validation, D.M. and C.L.; formal analysis, D.M. and C.L.; investigation, D.M. and C.L.; resources, D.M. and C.L.; data curation, D.M. and C.L.; writing—original draft preparation, D.M. and C.L.; writing—review and editing, D.M. and C.L.; visualization, D.M. and C.L.; supervision, C.L.; project administration, C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Data were found on HuggingFace (government reports dataset).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hovy, D.; Prabhumoye, S. Five sources of bias in natural language processing. *Lang Linguist. Compass* **2021**, *15*, e12432 [CrossRef]
2. Navigli, R.; Conia, S.; Ross, B. Biases in Large Language Models: Origins, Inventory, and Discussion. *J. Data Inf. Qual.* **2023**, *15*, 1–21. [CrossRef]
3. Li, H.; John, M.; Purkayastha, S.; Celi, L.; Trivedi, H.; Gichoya, J. Ethics of large language models in medicine and medical research. *Lancet Digit. Health* **2023** , *5*, e333–e335. [CrossRef]
4. Mikhailov, D. Optimizing National Security Strategies through LLM-Driven Artificial Intelligence Integration. *arXiv* **2023**, arXiv:2305.13927.
5. Wiegand, M.; Ruppenhofer, J.; Kleinbauer, T. Detection of Abusive Language: The Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*; Association for Computational Linguistics: Minneapolis, MN, USA, 2019; pp. 602–608.
6. Geva, M.; Goldberg, Y.; Berant, J. Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets. *arXiv* **2019**, arXiv:1908.07898.
7. Batanović, V.; Cvetanović, M.; Nikolić, B. A versatile framework for resource-limited sentiment articulation, annotation, and analysis of short texts. *PLoS ONE* **2020**, *15*, e0242050. [CrossRef]
8. White, J.; Cotterell, R. Examining the Inductive Bias of Neural Language Models with Artificial Languages. *arXiv* **2021**, arXiv:2106.01044.
9. Lee, H.; Hong, S.; Park, J.; Kim, T.; Kim, G.; Ha, J.W. KoSBI: A Dataset for Mitigating Social Bias Risks Towards Safer Large Language Model Applications. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track), Toronto, ON, Canada, 9–14 July 2023; p. 6026.
10. Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; Vasserman, L. Measuring and Mitigating Unintended Bias in Text Classification. In Proceedings of the AIES '18: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, LA, USA, 2–3 February 2018; pp. 67–73.
11. Renaldi, L.; Ruzzetti, E.; Venditti, D.; Dario, O.; Zanzotto, F. A Trip Towards Fairness: Bias and De-Biasing in Large Language Models. *arXiv* **2023**, arXiv:2305.13862.

12. Nadeem, M.; Bethke, A.; Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Virtual, 1–6 August 2021; Volume 1, pp. 5356–5371.
13. Wang, A.; Singh, A.; Michael, J.; Levy, O.; Bowman, S. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium, 1 November 2018; pp. 353–355.
14. Gao, Y.; Yang, Y.; Abbasi, A. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 1012–1023.
15. Huang, P.S.; Zhang, H.; Jiang, R.; Stanforth, R.; Welbl, J.; Rae, J.; Yogatama, D.; Kohli, P. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP, Punta Cana, Dominican Republic, 16–20 November 2021; pp. 65–83.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.