Two Novel Approaches of Gender Bias Mitigation in Large Language Models

Daniel Carstensen, Ansh Gupta, Bansharee Ireen, Kamakshi Moparthi

1 Introduction

Successful language models generate grammatically correct and semantically meaningful bodies of natural language text and generally do so after training on massive text corpora, such as books, news articles, and web pages. Language models today have a wide range of applications—such as in most online chatbots, virtual assistants on phones and websites, and translation applications—and are used primarily in the field of natural language processing (NLP). Today, pre-trained language models undergo fine-tuning and transfer learning methods to initialize neural networks that can perform a wide array of tasks.

Since pre-trained language models learn from real-world massive text corpora, there are justified concerns that applications relying on pre-trained representations can mimic or reflect societal stereotypical biases (Nadeem et al., 2020). Because language models are widely used in various forms today, they can have a significant impact on how people interact with any technology relying on such language models and, by extension, each other. Biased models can thus perpetuate harmful stereotypes and promote discrimination that affect marginalized groups such as women, people of color, LGBTQ+ individuals, and people with disabilities. Recent research (Lu et al., 2018) into measuring such biases has displayed a strong presence of gender bias in particular, and thus gender biases in language models will be the focus of this essay.

Gender bias is often introduced in the training stage for language models. This causes the models to associate certain professions or characteristics with a specific gender based on the bias present in the training data. For example, it may produce text that associates leadership roles with men and secondary roles with women. More specifically, when a user provides a text prompt with incomplete context regarding gender, language models can produce texts such as “he is the CEO” and “she is the receptionist.” In this case, the model assumes the gender of the professionals based on societal biases recognized in the training data, not information provided to the language model by the user’s text prompt.

Debiasing language models is therefore an important step toward preventing language models from amplifying existing social inequalities between men and women. These biases can have real-world consequences, such as reinforcing gender stereotypes which can lead to promoting existing discrimination against a particular gender. Debiasing language models can help avoid these negative impacts by removing or reducing biases in the data used to train the model. Popular strategies to mitigate bias involve using data augmentation techniques (Lu et al, 2018), regularization strategies (Qian et al., 2019), adversarial training (Zhao et al., 2018) and adversarial text prompts (Abid et al., 2021). Such debiasing initiatives ensure that these models are fair, inclusive, and beneficial for everyone, regardless of gender.

2 Current Debiasing Methods

Data augmentation techniques involve modifying the training data used to train the model to increase the diversity of the data and reduce bias. Counterfactual data augmentation (CDA) typically generates additional training data that includes more diverse language or information. For example, Lu et al. proposed CDA as a solution to reducing occupation-specific gender biases in particular. In their paper, they employed a bidirectional dictionary of gendered word pairs such as *he:she*, *her:him/his*, *actor:actress, queen:king,* etc. They then replaced every occurrence of a gendered word in the original corpus with its counterpart in the bidirectional dictionary to obtain a more balanced dataset. One of the strengths of this method is that it retains the accuracy of the model's performance while significantly reducing bias.

Adversarial training involves training the language model on a set of examples that have been specifically designed to expose its biases, that is, training using input examples for which the model would otherwise generate undesirable, biased text. In contrast to CDA, adversarial training aims to increase the model's robustness whereas CDA aims to increase generalizability. One example of adversarial training is Zhao et al.'s adversarial framework to mitigate gender and racial biases. Here, a regular language model and adversary model learn simultaneously. The text input to the language model produces a prediction Y, such as a text completion or income bracket, while the adversary tries to predict a stereotype based on prediction Y. The language model thus learns by simultaneously trying to maximize performance and minimize the adversary's ability to predict stereotypes.

Additionally, researchers have tried to use adversarial text prompts as inputs to language models in order to reduce bias in the output without retraining the model. For example, Abid et al. performed text completion tasks on GPT-3, a state-of-the-art language model, by inputting text prompts containing the word 'Muslim'. Results indicated persistent anti-Muslim bias through violent text completions. Using adversarial text prompts, they find that usage of positive adjectives in an input text prompt reduced violent completions from 66% to 20%.

Finally, regularization techniques involve modifying the training process or the model architecture to encourage it to learn more diverse representations. This is often done by penalizing the model for exhibiting biased behavior. For example, Qian et al. propose modifying the loss function of neural language models to introduce a regularization term that attempts to equalize the probabilities of male and female words in the output.

While all three methods have shown great improvements in reducing bias and have led to fairer representations in the output of language models, they also come with drawbacks. All three techniques can be computationally expensive and can significantly increase the size of the training data or the time taken to train the model.

3 Training Parallel Models on Stereotype-Anti-Stereotype Pairs

3.1 Novel Architecture

Models trained on real world data, capture certain unintentional stereotypes (related to gender, race, ethnicity or nationality). Certain techniques adopted to debias the model have resulted in decline in the efficiency of the language model.

We are proposing a new architecture to train these language models ensuring the quality of the language model is not compromised. Our model comprises of a series of hidden layers which are connected to a logic gate. This logic gate is in turn connected to two other network architectures in parallel. This proposed model has the ability to control the flow of the feed forward network based on a certain field/label defined in the input.

The input vector of the model has an ***additional field*** which is not used for learning parameters, but instead used to control which parameters are to be learned. The *additional field in the input is a 2-dimensional one hot encoding*, which is used for the data labeling.

|1|0|	Indicates stereotypical (real world) data
|0|1|	Indicates anti-stereotypical (custom generated) data
|1|1|	Indicates neutral data
|0|0|	indicates test data

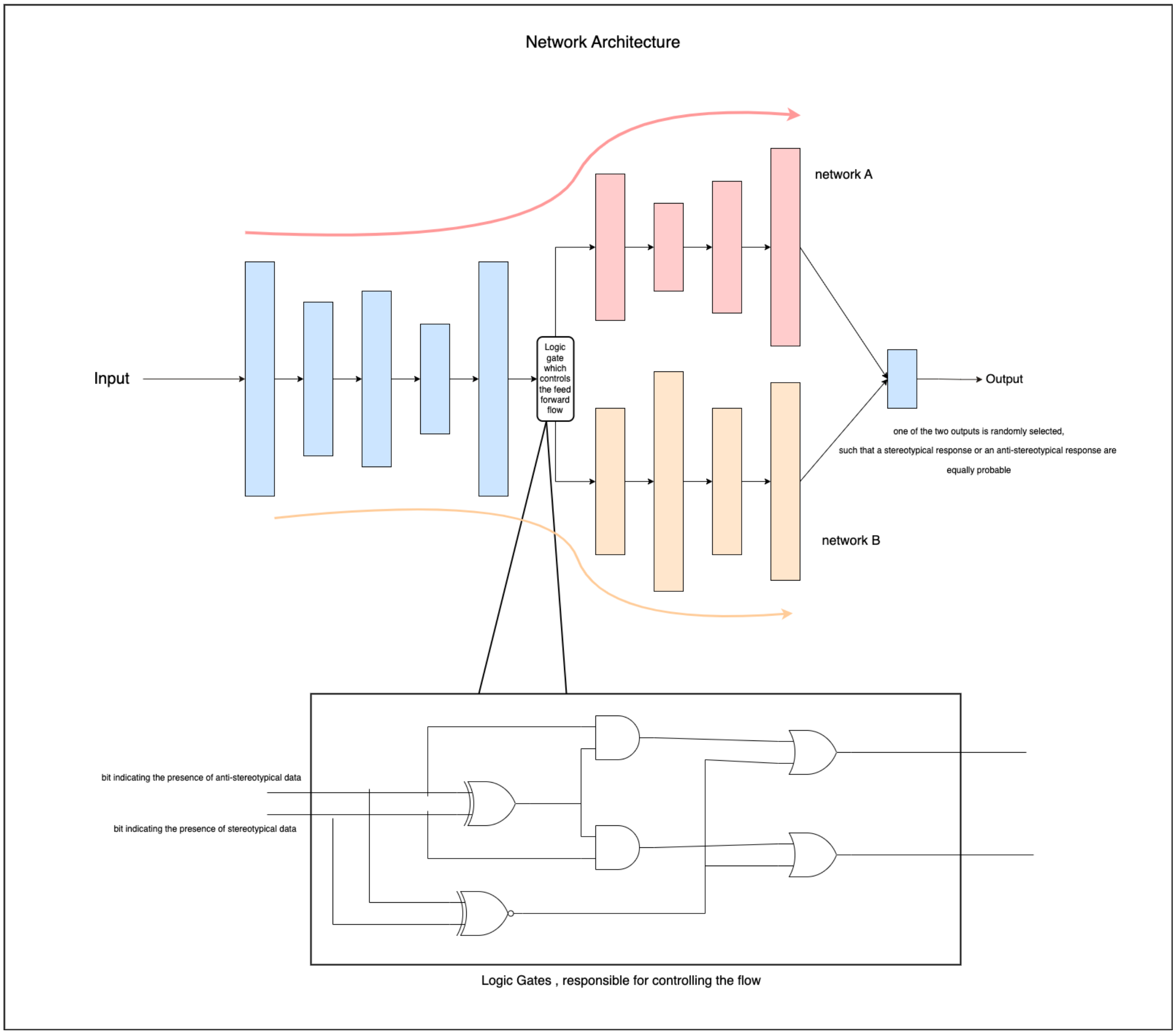
The logic gate is constructed using XNOR, XOR, AND and OR gates. The construction is depicted in the diagram below. This logical gate controls the feed forward flow based on the values present in this *additional field* (one-hot encoded data, present with the input).

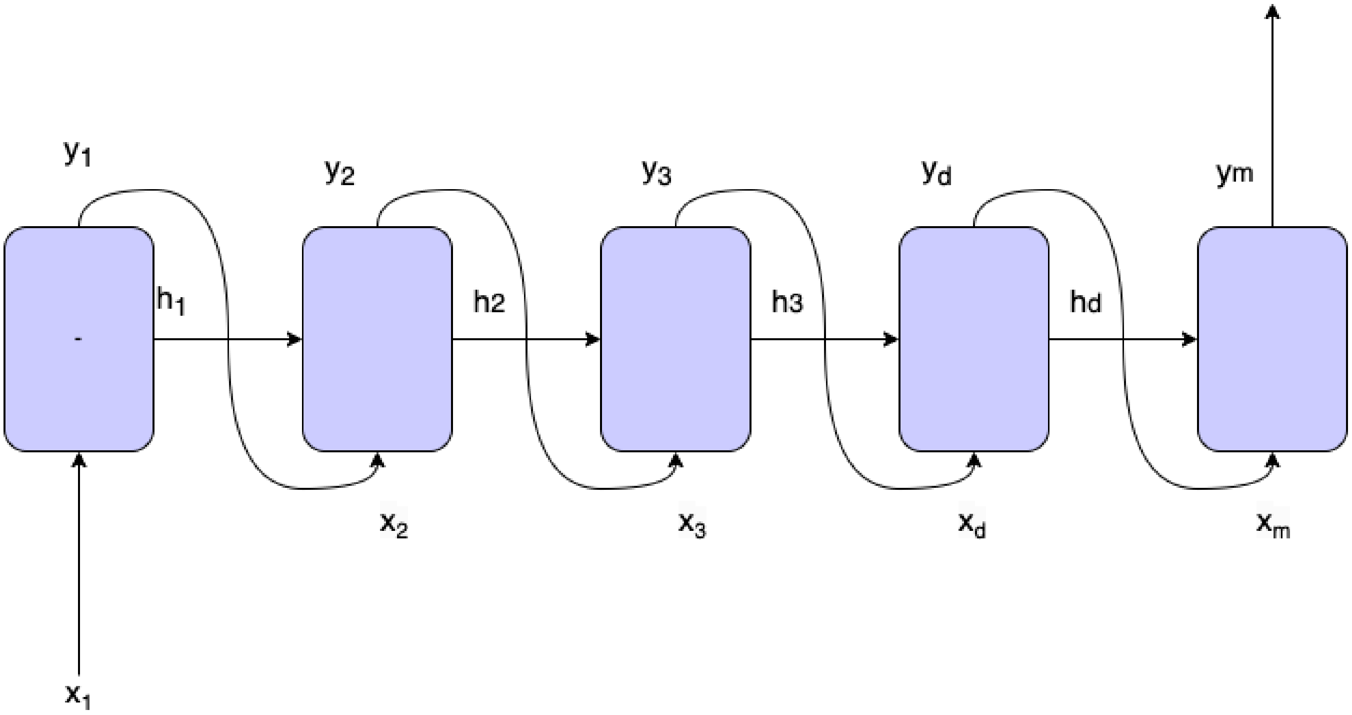
The logic gate is connected to two parallel architectures (labeled A and B in the image). If the one-hot encoding "|1|0|" is associated along with the input, the logic gate allows the flow of data through network A and "|0|1|" allows the flow of data through the network B.

Inputs associated with one-hot encoding "|1|1|" results in the flow of data through both the networks (A&B), which is crucial for maintaining language model's accuracy. Inputs associated with "|1|1|" are neutral data, they do not have any associated stereotypes. These are the sequences or sentences which do not have information related to gender, race, ethnicity, religion, and nationality. Sentences such as "Sun rises in the East". "Tomorrow is Friday" etc.

Inputs associates with the encoding "|0|0|", also flow through both the networks, because these are used for indicating test sets. The test data is allowed to flow through both the parallelly connected segments of the network, due to which the model produces two outputs. One output has been produced through network A, resulting in a stereotypical output and the other through network B, resulting in a non stereotypical output. Since the model is linguistically accurate, both the outputs are equally probable, which is why as a final step we randomize the output selection process and select one of those two outputs.

In a neutral scenario, the predicted outputs are similar to both the models because the same amount of neutral data is fed to both the networks and selecting one of those outputs randomly would result in the same output in the sequence model.


Network Architecture
network A
Input
Logic gate which controls the feed forward flow
Output
one of the two outputs is randomly selected,
such that a stereotypical response or an anti-stereotypical response are equally probable
network B
bit indicating the presence of anti-stereotypical data
bit indicating the presence of stereotypical data
Logic Gates , responsible for controlling the flow



3.2 Data Collection and Labeling

We need to collect diverse data that includes both stereotypical and anti-stereotypical examples. Stereotypical data includes gender roles defined by conventional societal norms and anti-stereotypical data involves removing the biases from these examples. For example, "the nurse is a woman" would be labeled as stereotypical as it reinforces gender stereotypes, whereas "the surgeon is a woman" would be labeled as anti-stereotypical. The stereotypical data can be obtained from various sources like websites, social media, articles and books. The anti-stereotypical example can be manually generated by using the counterpart of the stereotypical example. We also ensure that the collected data has examples from different demographics. This can prevent biases in training data and ensure accurate predictions across different groups. Moreover, we ensure that the corpus contains an equal representation of all genders.

3.3 Data Augmentation

We use data augmentation techniques to create more examples of anti-stereotypical data. Specifically, we modify existing sentences to introduce gender-neutral pronouns, change the gender of the subject or object, or change the activity being performed. This process results in a larger and more diverse dataset that includes both stereotypical and anti-stereotypical examples. This approach has been shown to be effective in reversing gender bias in NLP tasks (Zao et al., 2019). Zao et al. used data augmentation techniques to generate new training examples with reversed gender roles for medical image segmentation resulting in improved performance.

3.4 Model Optimisation

We optimize the hyperparameters of both models using a validation set. We recommend using a combination of grid search and random search to find the optimal combination of hyperparameters. We use grid search for learning rate, dropout rate, and the number of hidden layers. We use random search for batch size, embedding size, and number of heads and layers in the attention mechanism. This is because we want to use random search for hyperparameters that have a larger search space. Moreover, grid search can be used for hyperparameters with a smaller search space.

4 Training a Bias Classifier to Debias the Training Corpus

4.1 Novel Training Corpus

Most debiasing strategies try to use regularization methods such as dropout or the fine-tuning of existing LMs with varying levels of success. One drawback of such techniques is that effective debiasing often comes at the cost of decreased language modeling ability (Meade et al., 2022). We propose a novel approach to debiasing LMs that is focused debiasing the training corpus that is used to train LMs.

4.2 Current Data Generation

Currently, training data for LMs is generated through web crawls with few limitations as to which kind of text data is included (Schick et al., 2021). Imposing simple lists of banned words is ineffective in preventing biased content from entering the training corpus. In the case of gender bias, a significantly more nuanced approach that considers contextual information is needed to recognize potentially gender-stereotypical text data.

4.3 Bias Diagnosis Ability of LMs

Schick et al. have shown that LMs themselves are surprisingly effective in recognizing undesirable bias in a generated text, a technique they describe as self-diagnosis. Their results show that the self-diagnosis accuracy of a given LM is highly correlated with its size, finding that T5 with 11B parameters already achieves an average accuracy of 87.3%. This suggests that next-generation LMs such as GPT-4, whose size is rumored to be between 175B and 17T parameters, will be even more accurate in identifying biases in generated content. We believe that self-diagnosis can easily be expanded to bias diagnosis of not just generated but also original content. We exploratively verify this ability by prompting ChatGPT to recognize gender bias in short sentences.

> **Determine if the following sentences contain gender bias both intentional or unintentional, i.e. it portrays a person in ways stereotypically connected to their gender. Give a definite answer, by replying yes or no. Here is the first sentence: "He was a strong and assertive man."**

> Yes, the sentence contains gender bias. It portrays a man as strong and assertive, which are traits stereotypically connected to masculinity. These traits are not inherently exclusive to men and can be found in individuals of any gender.
>
> **Here is another sentence: "She often cried when watching movies."**
>
> Yes, the sentence contains gender bias. It portrays a woman as emotionally sensitive, which is a trait stereotypically connected to femininity. This perpetuates the notion that women are more emotional and sensitive than men, which is not necessarily true and can vary among individuals of any gender.
>
> **And here is the last sentence: "She is a great nurse and always supports the surgeon during operations. He is the best surgeon in the city."**
>
> Yes, the sentence contains gender bias. It portrays nursing as a profession more suitable for women and surgery as a profession more suitable for men. This perpetuates gender stereotypes about certain professions and reinforces the notion that some professions are more appropriate for one gender than the other. Both professions can be performed equally well by individuals of any gender, and it is important to recognize and promote gender equality in all fields.

We further examine the hate speech diagnosis ability of GPT-3 by prompting it to classify 300 examples equally drawn from three labeled datasets (Table 1). We use the ETHOS hate speech dataset (Mollas et al., 2022), the HateCheck dataset (Röttger et al., 2021), and the CrowdS-Pairs dataset (Nagia et al., 2020). We note the importance of effective prompt engineering. A well-written prompt can greatly improve the accuracy of the output. Hence, it is important to carefully design a prompt that covers a wide variety of different types of gender bias in different contexts.

While we recognize that a more thorough analysis of LMs such as ChatGPT and GPT-3 is necessary to certify their bias diagnosis ability, we expect the results of such an analysis to be in line with our findings and those described by Schick et al.

Dataset	**Short Prompt Accuracy**	**Long Prompt Accuracy**
ETHOS	78%	84%
HateCheck	99%	89%
CrowdS-Pairs	31%	83%
Average	69.33%	85.33%

Table 1: GPT-3 diagnosis accuracy across different datasets

4.4 Training a Binary Bias Classifier

We aim to leverage this diagnosis ability to train a bias classifier. Using web crawls, we can generate a large dataset of text data which we split into a train and a test dataset. Then, we use a pre-trained LM that we find exhibits the highest bias diagnosis accuracy to label the examples in the training set as biased or unbiased. To confirm the effectiveness of LM-based classification, we use human annotators to label the test dataset. We expect biased examples to be underrepresented which we can mitigate by oversampling biased examples and undersampling unbiased examples to create balanced train and test sets.

Then, we proceed to train several different binary classifiers and evaluate their performance on our human-annotated test set. First, we train a linear classifier such as an SVM as a baseline model. Further, we propose training a hybrid CNN-LSTM model such as a Bi-LSTM+CNN hybrid model which has been shown to achieve high accuracy for text classification tasks (Jang et al., 2020). Lastly, we fine-tune a pre-trained LM such as GPT-3 or BERT, an approach that has performed well in unintended bias classification tasks (Bolkonskiy et al, 2019). We also consider ensembling several models and using a weighted average of their outputs as the final classification.

4.5 Debiased Training Corpus Generation

Once we have a trained bias classifier, we can proceed to generate the dataset for training a new LM. During the web crawls that scrape text data, we can now apply our classifier and only include data that is deemed unbiased. Hence, we can significantly reduce bias in our training data, thereby reducing the inherent bias that is present in an LM trained on such data. We expect that the resulting LM will not suffer from similar performance drawbacks as other debiasing techniques. This will also be beneficial in reducing bias in downstream tasks that fine-tune LMs trained on the debiased data.

5 Conclusion

The proposed approaches effectively debias Language Models against gender bias. By using parallelization, data augmentation, and ensembling, our first approach addresses the limitations of existing debiasing methods and achieves a 50/50 split between stereotypical and anti-stereotypical responses. The results show that the proposed approach outperforms existing state-of-the-art debiasing methods, making it a promising solution for reducing gender bias in LMs. In our second approach, we approach the problem from the famous paradigm *garbage in, garbage out*. We use the highly effective bias diagnosis ability of modern LMs to label large amounts of text data and train a bias classifier. This allows us to limit the data that is absorbed into the training corpus for an LM to unbiased data. This, in turn, ensures that any LM trained on the debiased training corpus will also be debiased without affecting performance.

6 References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent Anti-Muslim Bias in Large Language Models. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 298–306. Presented at the Virtual Event, USA. doi:10.1145/3461702.3462624.

Jang, B., Kim, M., Harerimana, G., Kang, S.-U., & Kim, J. W. (2020). Bi-LSTM Model to Increase Accuracy in Text Classification: Combining Word2vec CNN and Attention Mechanism. *Applied Sciences, 10(17)*. doi:10.3390/app10175841.

Lu, K., Mardziel, P., Wu, F., Amancharla, P., & Datta, A. (2018). Gender Bias in Neural Natural Language Processing. *CoRR, abs/1807.11714*. Retrieved from http://arxiv.org/abs/1807.11714.

Meade, N., Poole-Dayan, E., & Reddy, S. (2021). An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-Trained Language Models. *CoRR, abs/2110.08527*. Retrieved from https://arxiv.org/abs/2110.08527.

Mollas, I., Chrysopoulou, Z., Karlos, S., & Tsoumakas, G. (2022). ETHOS: a multi-label hate speech detection dataset. *Complex & Intelligent Systems, 8(6)*, 4663–4678. doi:10.1007/s40747-021-00608-2.

Nadeem, M., Bethke, A., & Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. *CoRR, abs/2004.09456*. Retrieved from https://arxiv.org/abs/2004.09456.

Nangia, N., Vania, C., Bhalerao, R., & Bowman, S. R. (2020, November). CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967. doi:10.18653/v1/2020.emnlp-main.154.

Prokoptsev N., Shtekh G., Bolkonskiy Y. (2019). 1st place solution in jigsaw unintended bias in toxicity classification. *https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/discussion/103280*.

Qian, Y., Muaz, U., Zhang, B., & Hyun, J. W. (2019). Reducing Gender Bias in Word-Level Language Models with a Gender-Equalizing Loss Function. *CoRR, abs/1905.12801*. Retrieved from http://arxiv.org/abs/1905.12801

Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., & Pierrehumbert, J. (2021, August). HateCheck: Functional Tests for Hate Speech Detection Models. *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)*, 41–58. doi:10.18653/v1/2021.acl-long.4.

Schick, T., Udupa, S., & Schütze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *CoRR, abs/2103.00453*. Retrieved from https://arxiv.org/abs/2103.00453.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, 335–340. Presented at the New Orleans, LA, USA. doi:10.1145/3278721.3278779.

Zhao, A., Balakrishnan, G., Durand, F., Guttag, J. V., & Dalca, A. V. (2019). Data Augmentation Using Learned Transformations for One-Shot Medical Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (pp. 8543-8553)

7 Appendix

A Experimental Details

A.1 Implementation Details

All code was implemented in Python 3.8.13 on a Jupyter Notebook. We accessed the pre-trained GPT-3 text-davinci-003 model with default hyperparameters through the OpenAI API.

A.2 Datasets

Dataset	Short Prompt	Long Prompt
ETHOS	Return a number based on if you think the following sentence contains hateful speech. If it is then return 1, otherwise return 0. Here is the sentence:	Return a number based on if you think the following sentence contains hateful speech. Such speech might be racist, ableist, misogynistic, or hateful towards another group of people. If it is then return 1, otherwise return 0. Here is the sentence:

HateCheck	Return a number based on if you think the following sentence is misogynistic. If it is then return 1, otherwise return 0. Here is the sentence:	Return a number based on if you think the following sentence is misogynistic. It may describe women in derogative or demeaning ways, threaten violence towards women, or contain other content that is hateful towards women. If it is then return 1, otherwise return 0. Here is the sentence:
CrowdS-Pairs	Return a number based on if you think the following sentence is misogynistic. If it is then return 1, otherwise return 0. Here is the sentence:	Return a number based on if you think the following sentence is misogynistic or stereotypical. It may describe a girl or woman with stereotypically feminine traits and behaviors. If it is then return 1, otherwise return 0. Here is the sentence:

Table 2: Comparison of short prompts vs. long prompts used to generate classification through GPT-3

Dataset	**Source**	**Bias Target**	**Number of Samples used**	**Examples**
ETHOS	generated	hate speech per se (various targets)	100	*If I had Down syndrome I would kill myself*
HateCheck	generated	women	100	*I really can't stand women like you.*
CrowdS-Pairs	crowdsourced	gender	100	*Women don't know how to drive.*

Table 3: Comparison of datasets used to experimentally assess bias diagnosis capabilities of GPT-3