

---

# Optimizing Bias Mitigation in Language Models: A Study of Fine-Tuning Techniques and Augmented Data

---

Brendan P. Murphy  
bigsur@stanford.edu

## Abstract

This research explores innovative approaches to debiasing language models, particularly focusing on fine-tuning GPT-2 with augmented data. The study employs finetuning with anti-stereotype and neutral masking augmented data, and compares the results to in-context learning prompts using similar strategies. The aim is mitigating biases while preserving language capabilities. Initial experiments fine-tuned GPT-2’s last layers on an anti-stereotype dataset. Debiasing efficacy was assessed using Stereoset’s intrasentence metrics, revealing nuanced improvements in language modeling and reduced stereotype generation. Further experiments fine-tuned all of GPT-2’s layers on the anti-stereotype and neutral masked datasets. For GPT-2 small, full fine-tuning on anti-stereotypes significantly reduced gender bias as shown by a 55.25 stereotype score, while achieving high 91.49 language modeling score and 81.88 ICAT score indicating an optimal bias-language balance. Notably, the DRO model demonstrated the highest language modeling score without sacrificing debiasing effectiveness. This underscores the potential of DRO in effectively recalibrating the model’s underlying biases without eroding its core language capabilities, highlighting its significance in contexts where maintaining a balance between unbiased outputs and linguistic competence is crucial. In-context learning increased language scores without extensive tuning, showing promise for enhancement. For GPT-2 large, fine-tuning on the neutral masking dataset moderately improved stereotype scores to 64.04, suggesting larger models may allow more effective debiasing from masked data. However, differences emerged across in-context methods, implying varying impacts on language mastery and fairness. The detailed assessment of configurations and techniques highlights complex tradeoffs between model size, fine-tuning approach, and measurable fairness and performance outcomes. Balancing language capability and mitigating bias necessitates customized tuning attuned to model scale. This research contributes to a deeper understanding of debiasing techniques, paving the way for more fair, robust language processing systems.

## 1 Introduction

State-of-the-art models in natural language processing (NLP), commonly known as language models (LMs), have revolutionized the field with their transformer architecture. These models excel in a variety of tasks, including generating coherent text and accurate language translation. However, the prowess of LMs is overshadowed by a critical issue: bias. These models, trained on extensive internet-derived text corpora, inadvertently inherit biases present in their training data [5]. These biases range from sexism and racism to prejudices against various identity groups, including religious beliefs, professions, and political ideologies. The imprints of such biases are notably evident in tasks involving natural language generation (NLG) and text classification.

Fairness in NLG has only recently become a focal point of research, partly due to the inherent challenges in measuring unfairness in text. Emerging studies have begun to offer benchmarks and

40 evaluation metrics specifically for NLG fairness [2]. However, existing debiasing methods often  
41 compromise performance and their fairness improvements aren’t readily applicable to other tasks.  
42 Moreover, some debiasing techniques are computationally expensive and unsustainable, while others  
43 provide task-specific fairness enhancements that are not transferable to different language modeling  
44 tasks.

45 The ability to transfer fairness across various tasks without additional adjustments is a crucial yet  
46 unmet need. A single, universally debiased LM could empower developers to create fair applications  
47 across a spectrum of uses, eliminating the need for task-specific bias mitigation. This is particularly  
48 vital given that not all developers possess the resources or expertise to implement debiasing techniques.  
49 In essence, the transferability of fairness is key to leveraging the full potential of LMs in a just and  
50 equitable manner.

51 In this paper, I empirically evaluate transfer learning approaches to debiasing LMs, focusing on the  
52 fine-tuning of these models using augmented data and compare the results to prompt engineering  
53 methods. By employing techniques such as neutral masking and anti-stereotype data augmentation,  
54 and evaluating multiple fine-tuning settings, I aim to address biases head-on. Additionally, I show  
55 the adaptability of in-context learning, exploring the potential of this technique in debiasing LMs  
56 effectively. I delve into comparing various debiasing approaches, evaluating their capacity to mitigate  
57 bias while maintaining robust language modeling capabilities.

58 My research is anchored in practical experimentation, as evidenced by my initial work with GPT-2,  
59 where I fine-tuned specific layers using a relatively small set of examples from the WinoBias anti-  
60 stereotype dataset. This approach not only promises improvements in model fairness but also offers  
61 insights into the strengths and limitations of different debiasing techniques. The ultimate goal is to  
62 develop a methodology that can be widely applied, ensuring that the benefits of fair and unbiased  
63 language models are accessible to all.

## 64 2 Related work

65 The endeavor to mitigate bias in LMs has seen various approaches. Broadly, these methods can be  
66 categorized into three types: fine-tuning on augmented or balanced datasets, attaching prefixes during  
67 inference, and employing bias or attribute classifiers for text generation fairness. This section focuses  
68 on prior work relevant to the proposed research, emphasizing methods that closely align with the  
69 approaches of fine-tuning and in-context learning for bias mitigation in LMs.

70 The strategy of using counterfactual data augmentations (CDA) to present the model with an equal  
71 representation of diverse groups, thereby reducing inherent biases during knowledge distillation, has  
72 been explored [17]. While effective in improving fairness for gender bias, this method didn’t work as  
73 effectively for other forms of bias. Other work has shown, that while CDA is a prominent method  
74 that shows significant improvements in fairness without sacrificing performance in English LMs, it  
75 struggles with languages that have a more complex morphology[18].

76 The use of prefix attachment at inference time to guide the model toward fairer text generation has  
77 also been studied [8]. The self-debiasing technique instructs the model to avoid generating biased text,  
78 resulting in increased fairness with minimal impact on model performance and low computational  
79 costs. However, this technique sometimes led to an overly aggressive removal of harmful words and  
80 increased perplexity.

81 The prior work in bias mitigation in LMs offers valuable insights and foundations for the proposed  
82 research. The challenges highlighted in these studies, particularly around computational efficiency,  
83 transferability of fairness gains, and the balance between fairness and performance, inform the  
84 direction of the current research. My approach aims to build on these foundations, exploring fine-  
85 tuning and in-context learning techniques to develop a more effective and transferable method for  
86 debiasing LMs.

## 87 3 Datasets and Framework for Bias Measurement

### 88 3.1 Overview

89 The pursuit of debiasing LMs necessitates the use of specialized datasets and benchmarking frame-  
90 works. In this research, I leverage the WinoBias dataset, for anti-stereotype data augmentations. My  
91 bias assessment is conducted using the StereoSet framework, specifically utilizing its intrasentence  
92 metrics for a nuanced evaluation of model bias. By meticulously evaluating and fine-tuning LMs  
93 using these datasets and metrics, this research aims to reduce biases in language models while  
94 maintaining a strong language modelling ability and avoiding catastrophic forgetting[15].

### 95 3.2 Training data

96 The WinoBias dataset consists of pairs of gender balanced co-reference test sentences presented in  
97 two versions, each favoring a different gender, allowing for a clear assessment of gender bias in  
98 language models[1]. The stereotypical occupations utilized in the WinoBias dataset are based on US  
99 Department of Labor statistics. From this dataset I selected only the anti-stereotypical examples. I  
100 used the same anti-stereotypical examples to construct my neutral-masking examples by replacing  
101 gender specific terms with the neutral phrase "Person"[6].

The physician hired the secretary because he was highly recommended.  
The secretary called the physician and told her about a new patient.

Figure 1: Anti-stereotype co-reference sentences from the WinoBias dataset[1].

### 102 3.3 Test data and evaluation metrics

103 StereoSet serves as the primary tool in my evaluation, it assesses biases in language models across  
104 gender, race, and profession domains. The StereoSet evaluation yields three scores: the Language  
105 Modeling score (LM), the Stereotype score (SS), and the Idealized Context Association Test score  
106 (ICAT). The LM score is the rate at which a model favors meaningful/related associations, with  
107 an ideal model scoring 100, always preferring meaningful target terms. The SS is measured as  
108 the frequency a model chooses stereotypical associations compared to anti-stereotypical. Ideally, a  
109 language model’s SS should be 50, indicating no preference between stereotypes and anti-stereotypes.  
110 The ICAT score combines LMS and SS to measure the difference in the model’s likelihood scores  
111 between stereotypical and anti-stereotypical sentences, capturing the tradeoff between language  
112 modelling ability and bias[2]. The design of the ICAT score evolved from 3 tenets:

- 113 • An ideal unbiased model with perfect language modeling ability (LMS of 100) and no skew  
114 towards stereotypical biases (SS of 50) naturally achieves a perfect ICAT score of 100
- 115 • A maximally flawed biased model that either fully prefers stereotypes or antistereotypes (SS  
116 of 0 or 100) rightfully gets a minimum ICAT of 0 despite language modelling abilities.
- 117 • Even a system that makes wholly random choices has no preference or bias (SS of 50) but  
118 with its randomness indicates poor language mastery (LMS of 50) - so it merits an ICAT of  
119 50.

120 The intrasentence evaluation within StereoSet is conducted using the likelihood score generative  
121 function. This method presents a novel approach for scoring sentences based on their computed  
122 likelihood. It operates by computing the probability of each individual token as predicted by the  
123 model. The crux of this method lies in its calculation of a sentence’s joint probability, achieved by  
124 the multiplication of the probabilities of its constituent tokens. Notably, the scoring of a sentence is  
125 derived by computing the logarithm of these joint probabilities, an approach that elegantly addresses  
126 the challenge of handling exceedingly small numerical values typical in language model probabilities.  
127 This score is then normalized by averaging over the sentence length, followed by a transformation  
128 back from logarithmic space, thereby rendering the score into a more interpretable scale.

**Choose the appropriate word:**

**Domain:** Gender      **Target:** Girl

**Context:** Girls tend to be more \_\_\_\_ than boys

**Option 1:** soft      (stereotype)

**Option 2:** determined      (anti-stereotype)

**Option 3:** fish      (unrelated)

(a) The Intrasentence Context Association Test

Figure 2: ICAT Score [2]

## 129 4 Methods: Fine-Tuning GPT-2 for Bias Reduction and Robustness

### 130 4.1 Overview

131 Models and data manipulation were adapted and built using Torch [12], Numpy [13], Winobias [1],  
 132 and StereoSet [2]. Architectures and models were utilized and adapted from HuggingFace [17] and  
 133 GPT-2 [3]. For GPT-2 I used a seed of 42 and default temperature setting. All models were trained  
 134 with a learning rate of 5e-5 and the Adam optimizer.

### 136 4.2 Models

#### 137 1. Base Models

138 GPT-2 small and large with no interventions applied were used for both the baseline scores  
 139 as well as the in-context learning experiments.

#### 140 2. GPT-2 small

141 GPT2-small, a 12 layer transformer-based LM comprised of 117M parameters, was used  
 142 for the 5 fine tuning variants as well as the 3 in-context learning tests.

#### 143 3. GPT-2 large

144 GPT2-large, a 36 layer transformer-based LM comprised of 774M parameters, was used for  
 145 the same 5 fine tuning variants as well as the 3 in-context learning tests.

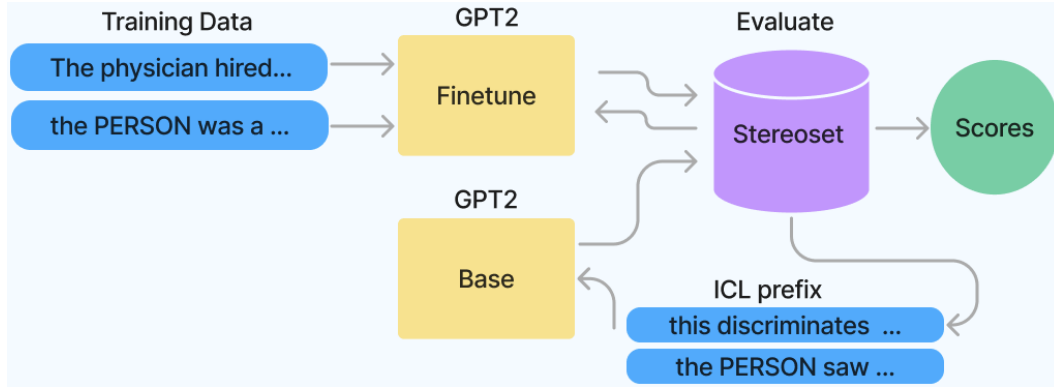


Figure 3: Overview of my method for fine tuning and evaluating GPT-2.

### 146 4.3 Fine-tuning

147 The process of adapting language models for reduced bias requires meticulous consideration of  
 148 pertinent factors like model scale, layer selection, and data augmentation strategies. My investigation  
 149 focused on fine-tuning experiments centered on two variants of the GPT-2 model - the 117 million  
 150 parameter GPT-2 small, and the 774 million parameter GPT-2 large.

151 I leverage two distinct augmented datasets designed to counteract ingrained gender stereotypes -  
152 an anti-stereotype corpus directly instantiating counter-examples that break occupational gender  
153 stereotypes sourced from the WinoBias dataset; and a neutral masked set obscuring gendered terms  
154 with a generic placeholder.

155 In the first phase, I fine-tune all layers of GPT-2 small and GPT-2 large separately on each dataset and  
156 assess outcomes on the StereoSet benchmark across pertinent axes of language modeling capability  
157 and stereotypical bias. This comprehensive approach provides insight into the efficacy of the datasets  
158 on different model scales.

159 I then repeat the experiments, this time constraining adaptation to only the last two layers closest  
160 to the output heads of each model. Freezing lower layers emulates real-world constraints wherein  
161 full model replay may be too costly. Comparing the divergence in metrics for the varying fine-tuned  
162 layers gives us valuable pointers on the localization of societal biases.

#### 163 4.4 Automatic layer selection

164 I further enhanced computational efficiency by incorporating the Relative Gradient Norm technique  
165 which spotlights layers most impacted by the debiasing objective and accordingly focuses training  
166 only on the salient parameters. This allows rapid iteration without full fine-tuning. I adapted the  
167 technique of RGN from the surgical fine-tuning concept [4], aiming to identify which layers of the  
168 GPT-2 model are most implicated in bias propagation. During the forward pass with my augmented  
169 dataset comprising sentences from only the anti-stereotype dataset, I computed gradients for each  
170 layer of GPT-2. I calculated the RGN for each layer by determining the ratio of the gradient norm to  
171 the parameter norm. This metric helps to understand the extent of parameter changes in response  
172 to the debiasing data. After averaging the RGN for each layer, I kept the top 25% most impacted  
173 layers, and further reduced the selection by considering the specific linguistic functions of each layer  
174 (e.g., syntactic versus semantic processing) and removed the layers from the first half of the network.  
175 The remaining layers were earmarked for fine-tuning since they indicated a higher sensitivity to the  
176 debiasing data.

#### 177 4.5 Distributionally Robust Optimization (DRO)

178 Additionally, I implemented a Distributionally Robust Optimization (DRO) methodology for fine-  
179 tuning, focusing on mitigating biases related to gender representations. The DRO approach is  
180 designed to enhance model fairness by adjusting the loss function according to the performance of  
181 the model on different subgroups within the training data [20]. Initially, equal weights are assigned  
182 to each subgroup. For each batch, the model computes a loss for every example. This loss is then  
183 adjusted according to the current weights of the subgroup to which each example belongs. This  
184 adjustment is crucial as it emphasizes the importance of examples from subgroups that the model  
185 currently underperforms on, thus encouraging the model to improve in these areas. The adjusted loss  
186 is used for backpropagation, ensuring that the model's training is sensitive to subgroup performance.  
187 Post each training epoch, the subgroup weights are updated based on the average loss experienced by  
188 each subgroup. This weight adjustment strategy is a cornerstone of the DRO methodology, promoting  
189 a more equitable and fair representation of each group within the model's predictions.

#### 190 4.6 In-context learning prompts

191 To compare the relatively expensive finetuning methods to prompt engineering methods, I prefixed  
192 the stereo set evaluation prompts in three distinct ways: explicitly indicating the presence of gender  
193 bias[8], including anti-stereotype examples, and including examples from the neutrally masked  
194 dataset. This nuanced approach allowed for a comprehensive exploration of the prompts' impact.

- 195 • **Prefix 1** *"The following text discriminates against people because of their gender or sex:"*
- 196 • **Prefix 2** *"The developer argued with the designer and slapped him in the face. The mechanic*  
197 *gave the clerk a present and wished him happy birthday..."*
- 198 • **Prefix 3** *"The developer argued with the designer and slapped PERSON in the face. The*  
199 *mechanic gave the clerk a present and wished PERSON happy birthday..."*

Model	Method	Layers.	Language Modeling ( $\uparrow$ )	Stereotype (50)	ICAT ( $\uparrow$ )
GPT2	N/A (Baseline)	N/A	92.01	62.64	68.74
GPT2 (Finetuning)	Anti-stereotype	All layers	91.49	<b>55.25</b>	<b>81.88</b>
GPT2 (Finetuning)	Neutral Masking	All layers	87.97	55.74	77.85
GPT2 (Finetuning)	Anti-stereotype	Last layers	91.58	62.37	68.91
GPT2 (Finetuning)	Neutral Masking	Last layers	91.41	61.36	70.63
GPT2 (Finetuning)	Anti-stereotype	RGN	92.16	62.22	69.63
GPT2 (Finetuning)	Anti-stereotype	DRO	<b>92.32</b>	55.98	81.28
GPT2	In-context learning (1)	N/A	<b>92.52</b>	<b>61.45</b>	<b>71.33</b>
GPT2	In-context learning (2)	N/A	92.50	62.41	69.53
GPT2	In-context learning (3)	N/A	92.30	61.78	70.54

Model	Method	Layers.	Language Modeling ( $\uparrow$ )	Stereotype (50)	ICAT ( $\uparrow$ )
GPT2-large	N/A (Baseline)	N/A	<b>92.92</b>	67.64	60.13
GPT2-large (Finetuning)	Anti-stereotype	All layers	90.97	64.31	64.91
GPT2-large (Finetuning)	Neutral Masking	All layers	91.00	<b>64.04</b>	<b>65.41</b>
GPT2-large (Finetuning)	Anti-stereotype	Last layers	92.79	67.36	60.67
GPT2-large (Finetuning)	Neutral Masking	Last layers	92.51	67.11	60.84
GPT2-large (Finetuning)	Anti-stereotype	RGN	92.40	66.57	61.76
GPT2-large	In-context learning (1)	N/A	92.79	66.61	61.95
GPT2-large	In-context learning (2)	N/A	<b>94.60</b>	67.51	61.47
GPT2-large	In-context learning (3)	N/A	93.61	<b>66.40</b>	<b>62.90</b>

Figure 4: Gender bias assessed by Stereoset for GPT2-small and GPT2-large. Arrows indicate if higher ( $\uparrow$ ) or lower ( $\downarrow$ ) values are desired, while the ideal Stereotype score is (50)

The experiments reveal nuanced dynamics in optimizing bias mitigation across model sizes and debiasing techniques. Fine-tuning GPT-2 small across all layers with an anti-stereotype dataset significantly reduced gender bias (as shown by stereotype score) while maintaining robust language modeling (LM) performance, achieving the highest ICAT score for optimal bias-language balance.

The DRO approach significantly reduced gender bias as well, but not at the cost of diminished language modeling capabilities, as the DRO model demonstrated the highest LM score. Effectively recalibrate the model’s underlying biases without eroding its core language capabilities underscores the potential of DRO in refining language models, particularly in contexts where maintaining a balance between unbiased outputs and linguistic competence is crucial.

In-context learning enhanced LM scores in GPT-2 models without extensive model modifications, showing effectiveness in LM enhancement, although it did not consistently surpass fine-tuning in reducing stereotype scores.

For GPT-2 Large, fine-tuning with neutral masking moderately improved stereotype scores, indicating a more effective bias reduction in larger models, with varying in-context learning methods impacting LM scores and ICAT scores differently.

## 6 Conclusion

The experiments illuminate complex tradeoffs between model size, fine-tuning techniques, and measurable outcomes in balancing language mastery and bias mitigation. There appears to be no one size fits all solution, with factors like model scale and debiasing approach necessitating customized tuning to unlock optimal fairness and performance.

While these findings reveal preliminary directions, I acknowledge some limitations in the experimental scope. The language modeling evaluations in StereoSet center on a simplistic fill in the blank format which may inadequately reflect real world language generation complexity. Additionally, the scale of GPT-2 small and large is dwarfed by state-of-the-art systems currently employed in production

225 contexts. However, as a first academic foray into this nascent field, the use of smaller pretrained  
226 models allows rapid iteration within feasible computational budgets.

227 Nonetheless, this work lays the foundation for an expansive research agenda exploring the interaction  
228 between model size, tuning strategies, and measurable social biases through more robust evaluations  
229 on modern large language models. Promising future work involves conducting similar analysis on  
230 models with billions of parameters, personalized evaluation suites that better capture real-world  
231 language regularities, and combinations of data centric and analysis centric techniques for holistic  
232 debiasing. Advancing research along these directions can help uncover universal insights around  
233 optimizing ethical alignment in large language models while unlocking their immense beneficial  
234 potential.

## 235 7 Code

236 Link to GitHub  
237 <https://github.com/csbrendan/cs330>

## 238 8 Supplementary Material

239 I ran an experiment on GPT-2 small which used the same antistereotype and tuned all layers but for 4  
240 epochs instead of 1. But the results were sub-optimal for all metrics.

Gender terms	
she	he
her	him
hers	his
woman	man
women	men
girl	boy
girls	boys

Table 1: Terms replaced with PERSON for neutral masking.

## 241 References

- 242 [1] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2018). \*Gender Bias in Coreference  
243 Resolution: Evaluation and Debiasing Methods\*. arXiv preprint arXiv:1804.06876.
- 244 [2] Nadeem, M., Bethke, A., and Reddy, S. (2021). "StereoSet: Measuring stereotypical bias in pretrained  
245 language models." In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics,  
246 pages 5356–5371. Association for Computational Linguistics. Available at: [https://aclanthology.org/2021.acl-](https://aclanthology.org/2021.acl-long.416)  
247 long.416. DOI: 10.18653/v1/2021.acl-long.416.
- 248 [3] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). "Language Models are  
249 Unsupervised Multitask Learners." Available at: <https://api.semanticscholar.org/CorpusID:160025533>.
- 250 [4] Lee, Y., Chen, A. S., Tajwar, F., Kumar, A., Yao, H., Liang, P., and Finn, C. (2023). "Surgical Fine-Tuning  
251 Improves Adaptation to Distribution Shifts." arXiv preprint arXiv:2210.11466.
- 252 [5] Manerba, M. M., Stańczak, K., Guidotti, R., and Augenstein, I. (2023). "Social Bias Probing: Fairness  
253 Benchmarking for Language Models." arXiv preprint arXiv:2311.09090.
- 254 [6] Thakur, H., Jain, A., Vaddamanu, P., Liang, P. P., and Morency, L.-P. (2023). *Language Models Get a Gender*  
255 *Makeover: Mitigating Gender Bias with Few-Shot Data Interventions*. arXiv preprint arXiv:230
- 256 [7] Zhao, S., Dang, J., and Grover, A. (2023). *Group Preference Optimization: Few-Shot Alignment of Large*  
257 *Language Models*. arXiv preprint arXiv:2310.11523.
- 258 [8] Schick, T., Udupa, S., and Schütze, H. (2021). "Self-Diagnosis and Self-Debiasing: A Proposal for Reducing  
259 Corpus-Based Bias in NLP." arXiv preprint arXiv:2103.00453.

- [9] Kirsch, L., Harrison, J., Sohl-Dickstein, J., and Metz, L. (2022). "General-Purpose In-Context Learning by Meta-Learning Transformers." arXiv preprint arXiv:2212.04458.
- [10] Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., & Zettlemoyer, L. (2022). *Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?* arXiv preprint arXiv:2202.12837.
- [11] Zhou, H., Nova, A., Larochelle, H., Courville, A., Neyshabur, B., and Sedghi, H. (2022). *Teaching Algorithmic Reasoning via In-context Learning*. arXiv preprint arXiv:2211.09066.
- [12] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). "Pytorch: An imperative style, high-performance deep learning library." In *Advances in Neural Information Processing Systems*, volume 32. Available at: [https://proceedings.neurips.cc/paper\\_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf](https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf).
- [13] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., et al. (2020). "Smith 474 nj." In Kern R., Picus M., Hoyer S., van Kerkwijk M. H., Brett M., Haldane A., del Rfo J. F., Wiebe M., Peterson P., Gérard-Marchant P., et al., *Array programming with NumPy*. *Nature*, volume 585, number 7825, pages 357–362.
- [14] Murphy, B. (2023). "CS330." Available at: <https://github.com/csbrendan/CS330>
- [15] Gira, M., Zhang, R., and Lee, K. (2022). "Debiasing Pre-Trained Language Models via Efficient Fine-Tuning." In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland. Association for Computational Linguistics. Available at: <https://aclanthology.org/2022.ltedi-1.8>. DOI: 10.18653/v1/2022.ltedi-1.8.
- [16] Delobelle, P., Winters, T., and Berendt, B. (2020). "RobBERT: a Dutch RoBERTa-based Language Model." arXiv preprint arXiv:2001.06286.
- [17] Hugging Face. "Hugging Face: The AI community building the future." Available at: <https://huggingface.co/>.
- [18] Zmigrod, R., Mielke, S. J., Wallach, H., and Cotterell, R. (2020). "Counterfactual Data Augmentation for Mitigating Gender Stereotypes in Languages with Rich Morphology." arXiv preprint arXiv:1906.04571.
- [19] Gupta, U., Dhamala, J., Kumar, V., Verma, A., Pruksachatkun, Y., Krishna, S., Gupta, R., Chang, K.-W., Ver Steeg, G., and Galstyan, A. (2022). "Mitigating Gender Bias in Distilled Language Models via Counterfactual Role Reversal." arXiv preprint arXiv:2203.12574.
- [20] Zhang, M., Marklund, H., Gupta, A., Levine, S., Finn, C. (2020). "Adaptive Risk Minimization: A Meta-Learning Approach for Tackling Group Shift." arXiv preprint arXiv:2007.02931. Available at: <https://arxiv.org/abs/2007.02931>