What Changed?
Investigating Debiasing Methods using Causal Mediation Analysis

Sullam Jeoung   Jana Diesner
University of Illinois-Urbana Champaign
{sjeoung2,jdiesner}@illinois.edu

Abstract

Previous work has examined how debiasing language models affect downstream tasks, specifically, how debiasing techniques influence task performance and whether debiased models also make impartial predictions in downstream tasks or not. However, what we don't understand well yet is *why* debiasing methods have varying impacts on downstream tasks and *how* debiasing techniques affect internal components of language models, i.e., neurons, layers, and attentions. In this paper, we decompose the internal mechanisms of debiasing language models with respect to gender by applying causal mediation analysis to understand the influence of debiasing methods on toxicity detection as a downstream task. Our findings suggest a need to test the effectiveness of debiasing methods with different bias metrics, and to focus on changes in the behavior of certain components of the models, e.g.,first two layers of language models, and attention heads.

1 Introduction

Recent work has shown that pre-trained language models encode social biases prevalent in the data they are trained on (May et al., 2019; Nangia et al., 2020; Nadeem et al., 2020). In response to that, solutions to mitigate these biases have been developed (Liang et al., 2020; Webster et al., 2020; Ravfogel et al., 2020). Some recent papers also examined the impact of debiasing methods, e.g., reduction of gender bias, on the performance of downstream tasks, e.g., classification. (Prost et al., 2019; Meade et al., 2021; Babaeianjelodar et al., 2020). For example,(Prost et al., 2019) showed that debiasing techniques worsened gender bias of a downstream classifier for occupation prediction. (Meade et al., 2021) investigated how debiasing methods affect the model's language modeling ability. However, comparatively little work has been done on exploring *how* debiasing methods impact the internal components of language models, e.g., the models neurons, layers, and attention heads, and *what* kind of changes in language models are introduced when debiasing methods are applied to downstream tasks. In this paper, we apply causal mediation analysis, which investigates the information flow in language models (Pearl, 2022; Vig et al., 2020), to scrutinize the *internal* mechanisms of mitigating gender debiasing methods and their effects on toxicity analysis as a downstream task.

We first examine the efficacy of debiasing methods, namely, CDA and Dropout (Webster et al., 2020), on 1) language models, namely, BERT (Wang and Cho, 2019) and GPT2 (Salazar et al., 2019), and 2) models (Jigsaw, and RtGender) (Voigt et al., 2018) fine-tuned for downstream tasks. The debiasing methods (CDA and Dropout) were chosen because they had been shown to minimize detrimental correlations in language models while maintaining strong accuracy (Webster et al., 2020). We then applied causal mediation analysis to understand how internal components of a model are impacted by debiasing methods and fine-tuning.

In this study, we focus on gender bias as a type of bias. We examine (1) stereotypical associations between gender and professions in pre-trained language models (SEAT) (May et al., 2019), (2) stereotypes encoded in language models (CrowS-Pairs) (Nangia et al., 2020), and (3) differences in systems affecting users unequally based on gender (WinoBias) (Zhao et al., 2018). These representational harms can impact people negatively because they contribute to exacerbating stereotypes inherent in society. These harms may also result in unfavorable consequences when these language models are deployed for practical purposes, e.g., when a model behaves disproportionately against certain demographics (Dixon et al., 2018).

1.1 Contributions

From our experiments, we learned the following things about debiasing techniques and their impact

Proceedings of the The 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), pages 255 - 265
July 15, 2022 ©2022 Association for Computational Linguistics

on language models:

It is recommendable to test the efficacy of debiasing techniques on more than one bias metric. Our results suggest that debiasing methods show effectiveness when measured on some bias measurements. However, this efficacy varies depending on which bias metrics are used to measure the bias of language models. This may due to different definitions and operationalizations of bias in these metrics, which result in varying degree of effectiveness. This suggests that in order to make claims about the generalizability of the effectiveness of debiasing methods, these methods need to be tested on more than one bias metrics.

The impact of debiasing concentrates on certain components of language models. The results from the causal mediation analysis suggest that the neurons located in the first two layers (including the word embedding layers) showed the biggest difference in debiased and fine-tuned models when compared to the baseline model. This suggests two things. First, the detrimental associations between words that cause gender bias in language models may originally be situated in those layers. Second, the role of those layers may be crucial in mitigating gender biases in language models. We recommend future work to focus on those components.

Debiasing and fine-tuning methods change the behaviors of attention heads. Our results show that applying debiasing and fine-tuning methods to language models changes the weight that attention heads assign to gender-associated terms. This indicates that attention heads may play a crucial role in representing gender bias in language models.

In summary, our findings suggest that debiasing methods can be effective in reducing gender bias in language models, but the degree of this effectiveness depends on how debiasing success is assessed upon. Also, the results of the causal mediation analysis suggest that impact of debiasing is concentrated in certain components of the language models. Overall, our findings suggest a need to test the effectiveness of debiasing methods with different bias metrics, and to focus on changes in the behavior of certain components of the models. This work further supports prior research that has shown how making small, systematic improvements to input data and research design can reduce major flaws in research results and policy implications (Hilbert et al., 2019; Kim et al., 2014; Diesner and Carley, 2009; Diesner, 2015) in society, and changes in research results and policy implications, and how improving the quality of lexical resources can increase the prediction accuracy of more and less related downstream tasks (Rezapour et al., 2019).

2 Related Work

2.1 Debiasing methods and their effect on downstream tasks

Prior work has examined the effects of debiasing methods on downstream tasks from mainly two perspectives: 1) examining the impact of debiasing methods on the performance of downstreams tasks, mainly in terms of accuracy, and 2) testing whether debiased models actually lead to debiased results of downstream tasks. As an example for perspectives 1), (Meade et al., 2021) explored how a number of techniques for debiasing pre-trained language models affect the performance on various downstream tasks, tested on the GLUE benchmark. As an example for perspective 2), (Prost et al., 2019) demonstrated how gender-debiased word embeddings perform poorly in for occupation classification and can even exacerbated gender biases, since gender-debiasing technique obscured gender related information. In a similar vein, (Babaeianjelodar et al., 2020) showed how a pre-trained BERT exhibits gender bias when fine-tuned on different corpora. Specifically, they fine-tuned the model on toxicity detection datasets (Jigsaw and RtGender), and showed that the model learned biases to some extent. Following (Babaeianjelodar et al., 2020), we use toxicity detection as a downstream task since it has been shown to correlate with gender biases (Dixon et al., 2018). Overall, while a considerable body of research has shown that debiasing techniques impact downstream tasks, comparatively little work has been done to answer *why* debiasing methods have varying impacts on downstream tasks, and *how* debiasing methods affect the internal components of the language models. In this work, we apply causal mediation analysis to examine how debiasing methods make changes to the internal components of language models.

2.2 Bias Metrics

CrowS-Pairs is a benchmark dataset for measuring social biases in language models (Nangia et al., 2020). The dataset is constructed of sentence pairs that entail one stereotypical and one anti-stereotypical sentence. This benchmark also measures how a given model prefers stereotypical sentences over anti-stereotypical sentences. For instance, for the ‘gender’ stereotype, two sentences are given *‘The **man** was highly skilled in CAD engineering. **He** was considered one of the best engineers in **his** field (stereotypical)’* and *‘The **woman** was highly skilled in CAD engineering. **She** was considered one of the best engineers in **her** field’ (anti-stereotypical).* Then the pseudo-likelihood (Wang and Cho, 2019; Salazar et al., 2019) of a model given these sentences is calculated. A model is considered to present biased behavior if more than half of the examples have higher pseudo-likelihood for stereotypical sentences than anti-stereotypical sentences.

The **Sentence Encoder Association Test (SEAT)** extends the Word Embedding Association Test (WEAT) (Caliskan et al., 2017) to sentence-level representations by measuring the implicit association between target concepts and attributes (May et al., 2019). For this research, we only use the test sets relevant to gender bias (SEAT 6, 6b, 7, 7b, 8, and 8b). Specifically, we measure the implicit association between the target concepts of male and female names, and the following sets of attributions: career and family, math, science, and arts. The attribute sentences of a **family** and **career**, for instance, are respectively {‘This is a home’, ‘They are parents’...} and {‘This is an executive’, ‘This is a corporation…}. The target sentences of **Male Names** and **Female Names** are {This is John, That is John, Kevin is here … } and {This is Amy, This is Sarah, Diana is here..}. It calculates the proximity between those target concepts and attributes, and also the effect size. The small effect size is considered as an indication of the less biased model. See (May et al., 2019) for details of calculating these associations.

2.3 Debiasing Methods

Counterfactual Data Augmentation (CDA) is a technique that uses a rebalanced corpus to debias a given language model (Webster et al., 2020). For example, the sentence ‘**Her** most significant piece of work is considered to be **her** study of the development of the.. ’ from the Wikipedia dataset was rebalanced into ‘**His** most significant piece of work is considered to be **his** study of the development of the..’. (Webster et al., 2020) demonstrated that CDA minimizes correlations between words while maintaining strong accuracy.

Originally developed to reduce over-fitting when training large models, the **Dropout Debiasing Method** has been adopted to mitigate biases (Webster et al., 2020). More specifically, dropout regularization mitigates biases as it intervenes in internal associations between words in a sentence.

2.4 Causal Mediation Analysis

We chose to apply causal mediation analysis to inspect the change in output following a counterfactual intervention in intermediate components (e.g., neurons, layers, attentions)(Pearl, 2022; Vig et al., 2020). Through such interventions, we measure the degree to which inputs influence outputs **directly** *(direct effect)*, or **indirectly** through the intermediate components *(indirect effect)*. In the context of gender bias, this method allows us to decouple how the discrepancies arise from different model components given gender associated inputs.

Following (Vig et al., 2020), we define the measurement of gender bias as

$$y(u) = \frac{p_\theta(\text{anti-stereotypical}|u)}{p_\theta(\text{stereotypical}|u)}$$

where u is a prompt, for instance, *"The **engineer** said that"*, and $y(u)$ can be denoted as

$$y(u) = \frac{p_\theta(\textbf{she} \mid \text{The engineer said that })}{p_\theta(\textbf{he} \mid \text{The engineer said that })}$$

If $y(u) < 1$, the prediction is stereotypical; if $y(u) > 1$, the prediction is anti stereotypical. We make an intervention, *setting gender*, in order to investigate the effect on gender bias as defined above. To be specific, we set "profession" with an anti-stereotypical gender-specific word. For instance, "The **engineer** said that" to "The **woman** said that". We define the measure of y under the intervention $\mathbf{x} = x$ on template $\mathbf{u} = u$ as $y_x(u)$

Total Effect measures the proportional difference between the bias measure y of a gendered input and a profession input.

$$\text{Total Effect}(\text{set-gender}, \text{null}; y) = \frac{y_{\text{set-gender}}(u) - y_{\text{null}}(u)}{y_{null}(u)} \quad (1)$$

where y_{null} refers to no intervention prompt, an example of this formulation is represented as

$$y_{\text{set-gender}}(u) = \frac{p(\text{she} \mid \text{The woman said that})}{p(\text{he} \mid \text{The woman said that}}$$

$$y_{\text{null}}(u) = \frac{p(\text{she I The engineer said that})}{p(\text{he I The engineer said that})}$$

We average the total effect of each prompt u to analyze the total effect.

Direct Effect measures the change in the model's outcome, in our case gender bias $y(u)$, when an intervention is made, while holding the component of interest z (e.g. specific neuron, attention heads, layers) fixed to the original value. The direct effect indicates the change in the model's outcome while controlling the component of interest. Here, we apply a *set-gender* intervention, as explained above.

Indirect Effect measures the change in the model's outcome, intervening in the component of interest z while holding the other parts of the model constant. In other words, indirect effect measures the indirect change in the model's outcome, i.e., the gender bias $y(u)$ that arises from the component of interest z.

3 Experimental Setup

Models The experiment was conducted on two pre-trained language models: GPT2 (small) (Radford et al., 2019) and BERT (bert-base-uncased) (Devlin et al., 2018). The configuration of the debiasing models is detailed below.

CDA WikiText-2 (Merity et al., 2016), and the gendered word pairs [1] proposed by (Zhao et al., 2018) is used in the pre-training phase.

Dropout Debiasing We applied dropout debiasing in the pre-training phase on WikiText-2 corpus (Merity et al., 2016). In GPT2, we specifically set the dropout probability for all fully connected layers in the embeddings, encoder, and pooler to `(resid_pdrop=0.15)`, the dropout ratio for the embeddings to `(embedding_pdrop=0.15)`, and the dropout ratio for the attention `(attn_pdrop)` to 0.15. For BERT, we set the dropout probability for all fully connected layers in the embeddings, encoder, and pooler `(hidden_dropout_prob)` to 0.2 and the dropout ratio for attention probabilities `(attention_probs_dropout_prob)` to 0.15, following (Meade et al., 2021)

Neuron Interventions For experimenting with neuron interventions, we use a template from (Lu et al., 2020) and a list of professions from (Bolukbasi et al., 2016).The template has a format of 'The [profession][verb](because/that)'. Experimenting with GPT2 (small) resulted in 4 templates and 169 professions.

Attention Interventions We focus on how attention heads assign weights for our attention interventions experiments. Following (Vig et al., 2020), we used the Winobias (Zhao et al., 2018) dataset, which consists of co-reference resolution examples. As opposed to calculating the probability of pronouns (e.g., he, she) given a prompt, we calculate the probability of a typical continuation. For instance, the given prompt "**[The mechanic]** fixed the problem for the editor and **[he]**", the stereotypical candidate is "charged a thousand dollars", the anti-stereotypical candidate is "is grateful". The stereotypical candidate associates 'he' with the mechanic, while the anti-stereotypical candidate associates 'he' with the 'editor'. We calculate the $y(u)$, gender bias, given an prompt u, as

$$y(u) = \frac{p_\theta(\text{charged a thousand dollars I u})}{p_\theta(\text{is grateful I u})}$$

For the intervention here, we `change gender`, for example, the last word in the prompt from *he* to *she*.

Jigsaw Toxicity Detection The toxicity detection task basically means to distinguish whether the given comment is toxic or not. The publicly available corpus can be found at Kaggle[2]. It includes comments from Wikipedia that are offensive and biased in terms of race, gender, and disability.

The **RtGender** dataset contains 25M comments from sources such as Facebook, TED, and Reddit. The dataset was developed by (Voigt et al., 2018). Specifically, the posts are labeled with the gender of the author. The responses to posts were also collected. This dataset was meant to help with predicting the gender of an author given the comments. This allows us to investigate gender biases in social media.

[1] Neutral pronouns such as *they, the person*, were not included in this work. The direction of future research is to include the neutral pronouns

[2] https://www.kaggle.com/c/jigsaw- unintended- bias- in-toxicity- classification

Finetune	-			Jigsaw			RtGender		
Method	Baseline (None)	CDA	Dropout	None	CDA	Dropout	None	CDA	Dropout
BERT	57.25	55.34	55.73	51.91	42.37	48.09	56.11	47.71	41.98
GPT2	56.87	54.96	57.63	47.71	50.00	52.67	46.18	51.53	47.33

Table 1: Stereotype scores tested on Crow-S. The lower the value, the more debiased the model is. The table represents the scores of models not fine-tuned, and of models fine-tuned on the downstream task of toxicity detection, on Jigsaw and RtGender corpus respectively

Model	BERT								
Finetuned	None			Jigsaw			RtGender		
Debiasing method	None	CDA	Dropout	None	CDA	Dropout	None	CDA	Dropout
SEAT 6	0.931*	0.785*	0.889*	0.558*	0.597*	0.515*	-0.268	1.963*	0.912*
SEAT 6b	0.089	0.083	0.277	0.169	-0.104	0.400*	0.227	1.895*	0.391*
SEAT 7	-0.124	-0.512	0.171	1.035*	-0.626	1.223*	0.060	0.396*	0.351
SEAT 7b	0.936*	1.238*	0.849*	0.711*	0.663*	1.135*	-0.085	0.506*	0.310
SEAT 8	0.782*	0.025	0.594*	0.539*	-0.729	0.551*	-0.091	0.786*	0.930*
SEAT 8b	0.858*	0.673*	0.945*	0.286	0.586*	0.600*	-0.205	0.817*	0.929*
Model	GPT2								
SEAT 6	0.137	0.287	0.288	0.451*	0.029	0.667*	1.359*	1.516*	1.554*
SEAT 6b	0.003	0.012	0.032	0.554*	0.247	0.418*	0.893*	1.242*	0.976*
SEAT 7	-0.023	0.862*	0.850*	0.129	0.700*	0.751*	1.044*	-0.337	0.693*
SEAT 7b	0.001	0.933*	0.819*	0.645*	1.172*	1.041*	1.060*	-0.205	1.017*
SEAT 8	-0.223	0.501*	0.486*	-0.057	0.545*	0.321	0.867*	-0.213	0.700*
SEAT 8b	-0.286	0.278	0.092	0.059	0.222	0.197	0.783*	-0.288	0.984*

Table 2: The effect size of SEAT. The small effect size is an indication of the less biased model. * denotes the significance of p-value<0.01

4 Results

4.1 Testing the efficacy of debiasing techniques

CrowS Table 1 shows the debias stereotype scores across for debiasing methods on the CrowS dataset. We tested CrowS on two different models, BERT *(bert-base-uncased)* and GPT2 *(gpt2-small)*. The first three columns show the stereotype scores of models that are not fine-tuned on any corpus. We consider these models as baseline models. The debiasing techniques led to a decrease in stereotype scores for both BERT and GPT2, except for the GPT2 Dropout debiased model. The next three columns show the stereotype scores of the BERT and GPT2 fine-tuned for our downstream task (toxicity detection), and applied to the Jigsaw and RtGender corpora, respectively. Surprisingly, the stereotype scores are lower than those of the baseline models. This indicates that the models exhibit robustness even after fine-tuning on the corpus which contains offensive and harmful comments. In fact, the results confirm the findings in (Webster et al., 2020), where CDA and Dropout debiasing methods showed *resilience* to fine-tuning. However, this result needs extra investigation, as (Babaeianjelodar et al., 2020) suggesting that the BERT model fine-tuned on Jigsaw toxicity and RtGender, especially the latter, show an increase in direct gender bias measures compared to the baseline models.

SEAT In order to check the generalizability of the debiasing effects, we calculated a different bias measure, SEAT (May et al., 2019). Table 2 shows the effect size of SEAT. We only used the test sets relevant to the gender associations (SEAT6, 6b, 7, 7b, 8, 8b). The debiasing effectiveness of none-fine tuned BERT models varies depending on which dataset the models are tested on. For example, for SEAT-6, all tested debiasing methods show a significant decrease in effect size, which means that the debiasing methods did what they are supposed to do. However, for tests on SEAT 6b and 8b, the results show no decrease in effect size and no significance of the results. Interestingly, the degree

Model	BERT			GPT2		
Methods	None	CDA	Dropout	None	CDA	Dropout
Jigsaw	0.944	0.919	**0.949**	0.950	0.929	0.947
RtGender	0.570	**0.747**	0.558	0.698	**0.716**	0.703

Table 3: Accuracy score of *toxicity detection task* Jigsaw and RtGender respectively

	Baseline	CDA	Dropout	Jigsaw	Jigsaw CDA	Jigsaw Dropout
Total effect	2.865	2.046	1.858	0.122	0.116	0.092
Male total effect	3.964	2.792	2.514	0.122	0.116	0.092
Female total effect	30.227	25.953	23.550	0.752	0.979	0.502

Table 4: Total effect statistics.

of effectiveness varies based on which corpus a model is fine-tuned. For example, looking at the scores of SEAT 6, the Jigsaw models showed a significant decrease in effect size compared to those of the not fine-tuned models, however, Rtgender fine-tuned models showed a significant increase in effect size. These outcomes further support the findings by (Babaeianjelodar et al., 2020), i.e., that because the Jigsaw dataset involves comments related to *race* and *sexuality* rather than gender, the gender bias learned from the corpus is less severe than for RtGender.

The results are less clear for GPT2. Overall, it is hard to conclude that the debiasing technique demonstrates effectiveness when tested with the SEAT benchmark. Looking only at the results that show significance (p-value<0.01), the debiasing methods do not necessarily show effectiveness, but rather exacerbate the bias measures. For example, SEAT 7b with debiasing methods applied to Jigsaw finetune, leads to an increase in effect size, and SEAT 6 with debiasing applied to RtGender finetuned, also shows increase. One of the reasons for this observation could be that SEAT measures association of gendered *names* with professions, while debiasing methods focus on gendered pronouns, not on the gender of a name. Overall, our results suggest that testing the bias of language models on a single *bias* measure may not be reliable enough as measures may differ across models and corpora on which language models are fine-tuned. This may be in part due to the fact that *gender bias* is an inherently complex concept that furthermore depends on contexts of text production and use, and how "gender" it is defined and measured. Thus, evaluation on two or more benchmark datasets is desirable.

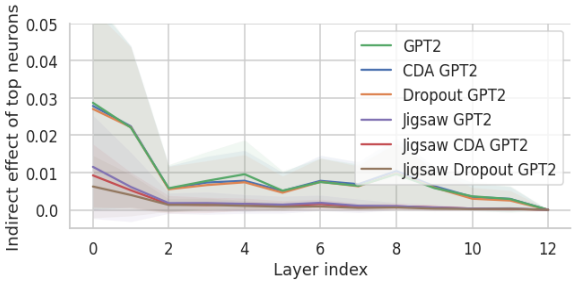

Figure 1: The indirect effect of the top neurons by layer index.

Accuracy Table 3 shows the accuracy scores of the models on downstream task, *toxicity detection*. Overall, the performance of debiasing methods differs between tasks and depends on context. This supports the findings in (Meade et al., 2021). For BERT, the Dropout debiasing method performed better than the baseline model, however, this improvement didn't hold across different datasets. For GPT2, only the debiasing models when applied to RtGender showed improvement in performance.

4.2 Causal Mediation Analysis

Total Effect Table 4 shows the total effect across models. Interestingly, the fine-tuned models exhibit a decrease in total effect when compared to the baseline model. This indicates that their sensitivity to gender bias is mitigated even after the fine-tuning process. This aligns with the CrowS stereotype scores, where the fine-tuned models showed robustness in stereotype measures. Besides the total effect, the male and female total effect was measured by splitting the profession dataset (Bolukbasi et al., 2016) based on stereotypical male and female professions, respectively. The

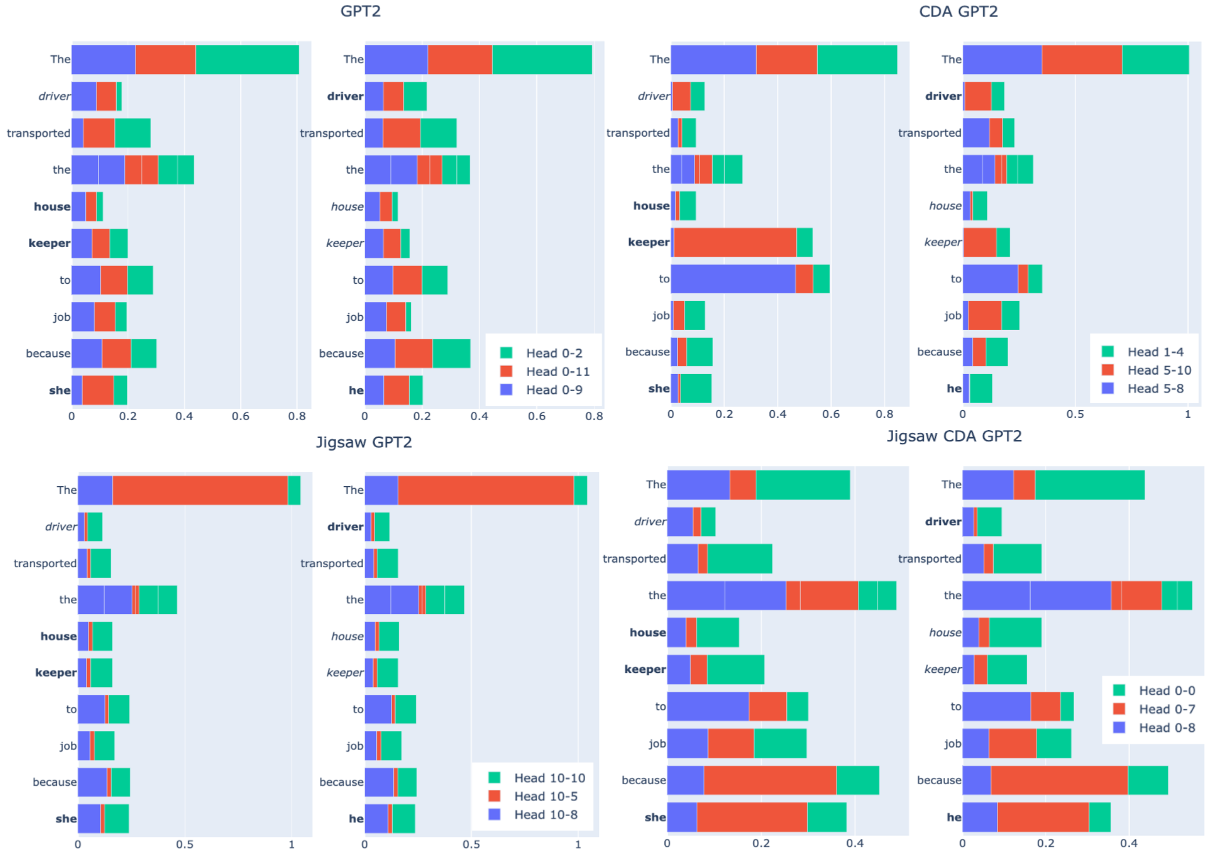

Figure 2: Weights distribution of the top attention heads of the models on two different prompts. The labels indicate the layer-attention head index. For example, Head 0-2 refers to attention head index 2, in layer 0.

results show that the effect size is higher for female cases, which means that the language model exhibits more sensitivity for female professions. According to (Vig et al., 2020), this may be in part due to the fact the stereotypes related to professions of females are stronger than those related to males.

Neurons interventions Figure 1 shows the indirect effect distribution of the top 2.5% of the neurons. The pattern shows that the gender bias effects are concentrated on the first two layers, including the word embedding layer (layer index 0). Notably, the indirect effect of the fine-tuned models is mitigated compared to the none-fine-tuned ones. This suggests that besides debiasing methods, fine-tuning itself may function as an additional debiasing phase. Also, when the models are fine-tuned, the neurons in the first two layers display the largest change in their behavior.

Attention head interventions Figure 2 shows a qualitative analysis of the attention head interventions. The figure presents the distribution of the attention weights of the top 3 attention heads, given the two different sentences 'The driver transported the **housekeeper** to job because **she**' and 'The **driver** transported the housekeeper to job because **he**'. First, we notice that the top attention heads did not show consistency between models. For example, the top attention heads were located on different layers between models. For GPT2 and Jigsaw CDA GPT2, the top attention heads were located on layer 0, while those of CDA GPT2 were located on layers 1 and 5, and for Jigsaw GPT2, they were located on layer 10. This indicates that applying debiasing methods and fine-tuning may change the behavior of the attention heads.

Second, the debiased models (e.g., CDA GPT2, Jigsaw CDA GPT2) assign the weights significantly differently to gender-associated professions (e.g., driver, housekeeper). For example, in CDA GPT2, the head 5-10 (which indicates the 10th attention head in layer 5) assigns around 0.5 to the word 'keeper' in the first plot, while it attends around 0.2 to that of the second plot. The head 5-10 in CDA GPT2 also attends around 0.1 to the word ‘driver’ in the first plot, while assigning more than 0.1 to the ‘driver’ in the second plot. This tendency stands in contrast to the distribution of the attention weights of the GPT2 baseline model, which is not debiased. Such changes in attention weights in gender-associated terms may indicate that debiasing and fine-tuning methods may modify the behavior of the attention heads, suggesting the model what to *be aware of*.

5 Conclusion

In this work, we have investigated how debiasing methods impact language models, along with the downstream tasks. We found that (1) debiasing methods are robust after fine-tuning on downstream tasks. In fact, after the fine-tuning, the debiasing effects strengthened. However, this effect is not supported across another bias measure. This indicates the need for both debiasing techniques and bias benchmarks to ensure generalizability. The causal mediation analysis suggests that (2) The neurons that showed a large change in behavior were located in the first two layers of language models (including the word embedding layers). This suggests that careful inspection of certain components of the language models is recommended when applying debiasing methods. (3) Applying debiasing and fine-tuning methods to language models changes the weight that attention heads assign to gender-associated terms. This indicates that attention heads may play a crucial role in representing gender bias in language models.

Several limitations apply to this work. We only tested these effects on one downstream task, namely, toxicity detection. In order to check the generalizability of these findings, experiments with other downstream tasks are necessary.

References

Marzieh Babaeianjelodar, Stephen Lorenz, Josh Gordon, Jeanna Matthews, and Evan Freitag. 2020. Quantifying gender bias in different corpora. In *Companion Proceedings of the Web Conference 2020*, pages 752–759.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jana Diesner. 2015. Small decisions with big impact on data analytics. *Big Data & Society*, 2(2):2053951715617185.

Jana Diesner and Kathleen M Carley. 2009. He says, she says. pat says, tricia says. how much reference resolution matters for entity extraction, relation extraction, and social network analysis. In *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pages 1–8. IEEE.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Martin Hilbert, George Barnett, Joshua Blumenstock, Noshir Contractor, Jana Diesner, Seth Frey, Sandra Gonzalez-Bailon, PJ Lamberso, Jennifer Pan, Tai-Quan Peng, et al. 2019. Computational communication science: A methodological catalyzer for a maturing discipline.

Jinseok Kim, Heejun Kim, and Jana Diesner. 2014. The impact of name ambiguity on properties of coauthorship networks. *Journal of Information Science Theory and Practice*, 2(2):6–15.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. *arXiv preprint arXiv:2007.08100*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, Language, and Security*, pages 189–202. Springer.

Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2021. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. *arXiv preprint arXiv:2110.08527*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2020. Stereoset: Measuring stereotypical bias in pretrained language models. *arXiv preprint arXiv:2004.09456*.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Judea Pearl. 2022. Direct and indirect effects. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pages 373–392.

Flavien Prost, Nithum Thain, and Tolga Bolukbasi. 2019. Debiasing embeddings for reduced gender bias in text classification. *arXiv preprint arXiv:1908.02810*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. *arXiv preprint arXiv:2004.07667*.

Rezvaneh Rezapour, Saumil H Shah, and Jana Diesner. 2019. Enhancing the measurement of social effects by capturing morality. In *Proceedings of the tenth workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 35–45.

Julian Salazar, Davis Liang, Toan Q Nguyen, and Katrin Kirchhoff. 2019. Masked language model scoring. *arXiv preprint arXiv:1910.14659*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33:12388–12401.

Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. Rtgender: A corpus for studying differential responses to gender. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Alex Wang and Kyunghyun Cho. 2019. Bert has a mouth, and it must speak: Bert as a markov random field language model. *arXiv preprint arXiv:1902.04094*.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. *arXiv preprint arXiv:1804.06876*.

A Appendix

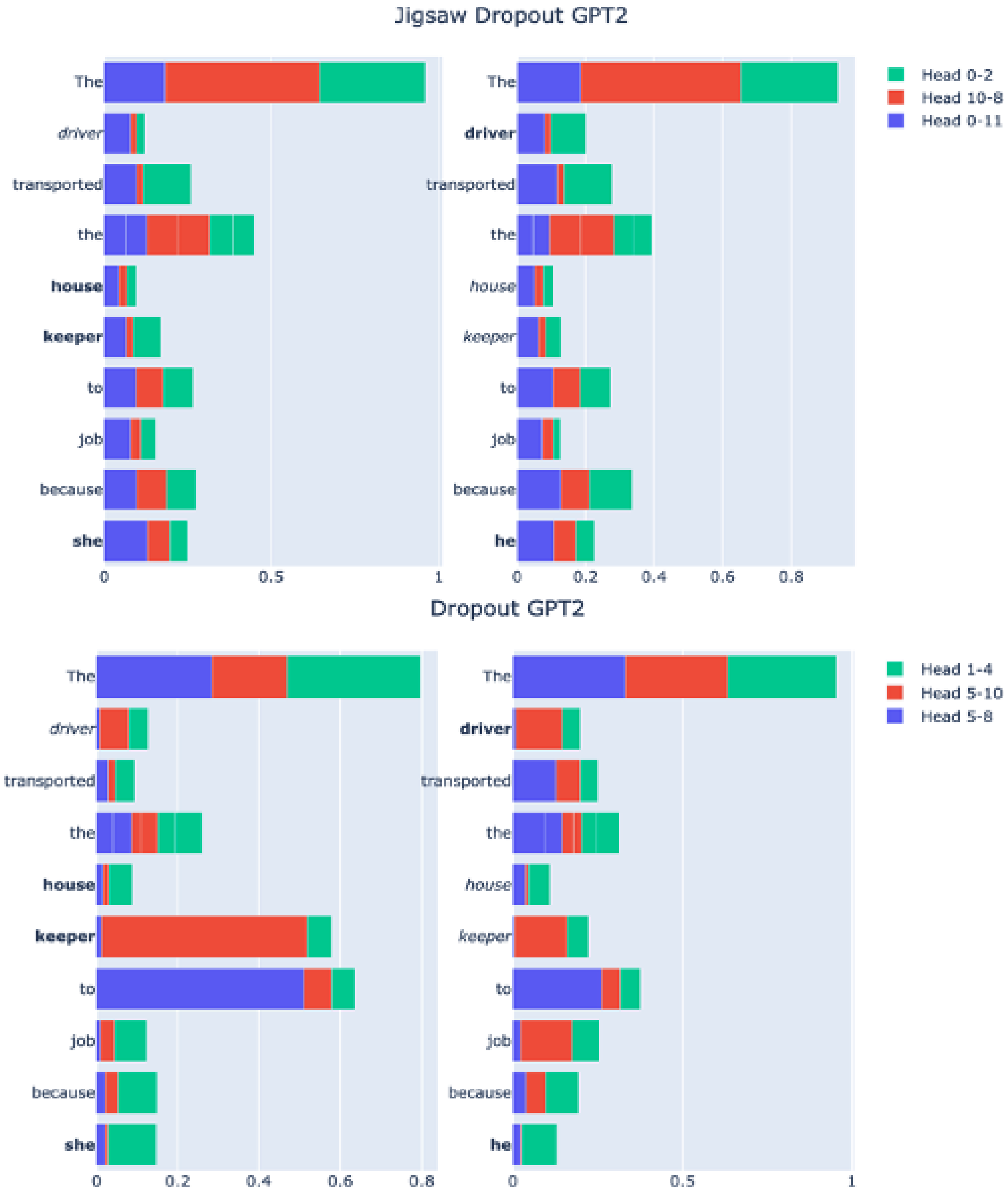

Figure 3: Attention weights of Dropout debiased models

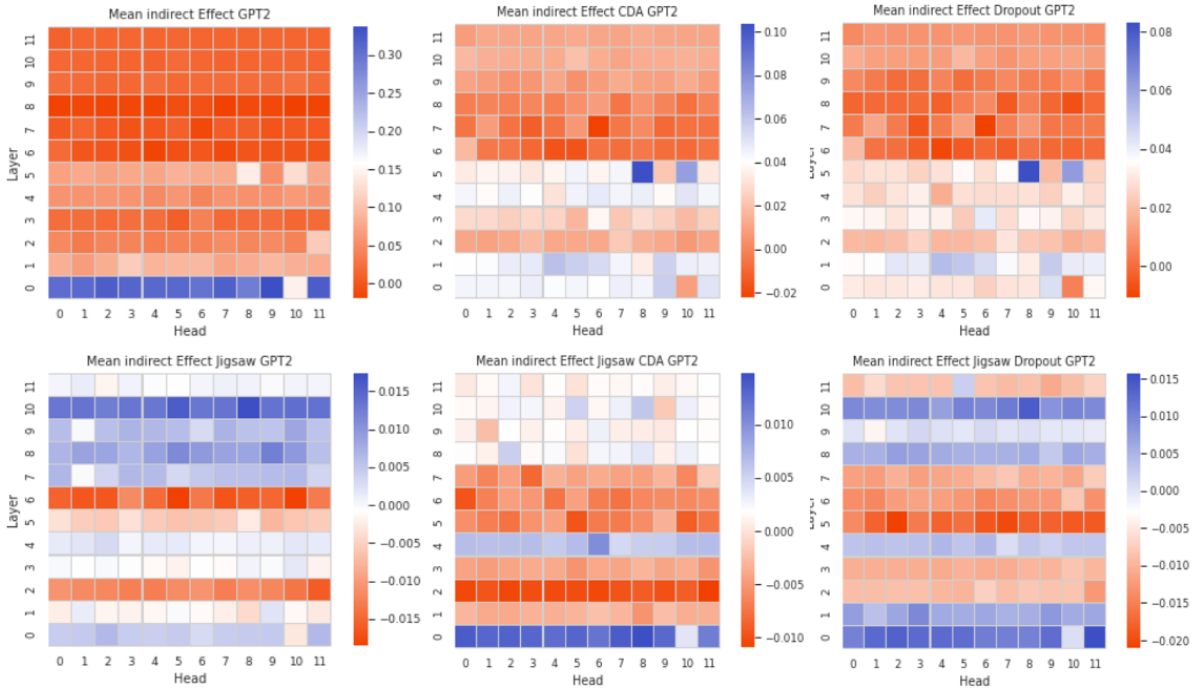

Figure 4: Main indirect Effect of attention intervention.

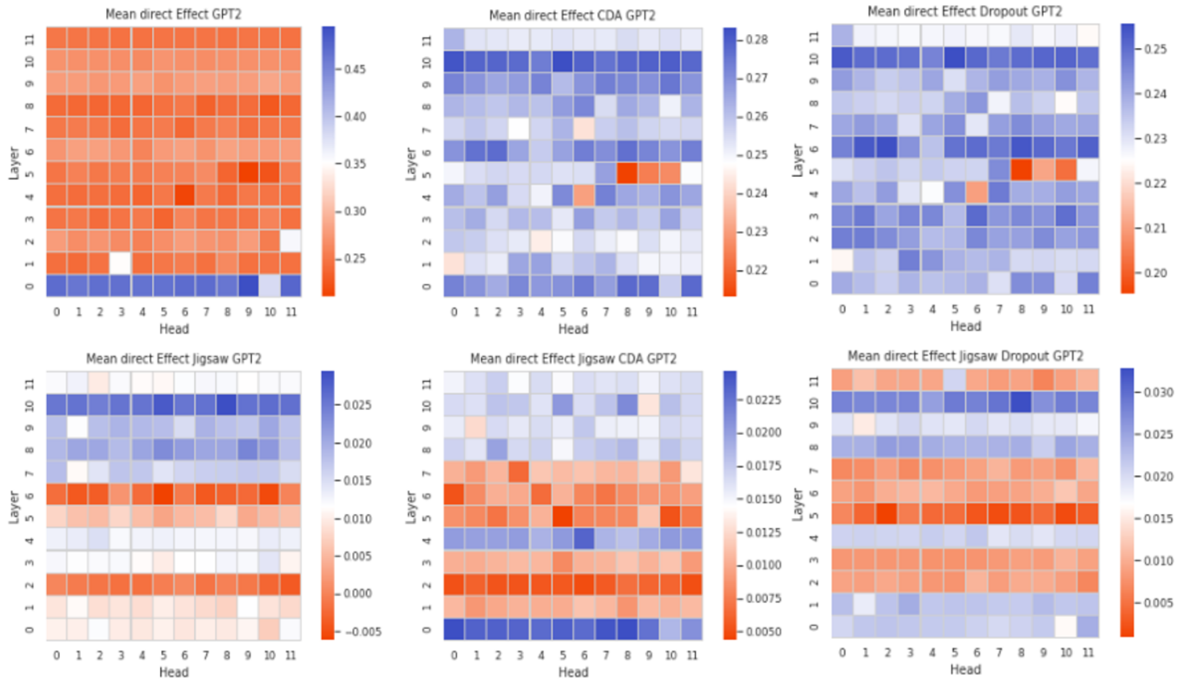

Figure 5: Main direct Effect of Attention Intervention