Identifying and Adapting Transformer-Components Responsible for Gender Bias in an English Language Model

Abhijith Chintam[1][2], **Rahel Beloch**[1]
[1]Master AI, University of Amsterdam, [2]Pegasystems, Amsterdam, The Netherlands
archintam@gmail.com, mail@rahelbeloch.de

Willem Zuidema, Michael Hanna[*], **Oskar van der Wal**[*]
Institute for Logic, Language & Computation, University of Amsterdam
{w.h.zuidema, m.w.hanna, o.d.vanderwal}@uva.nl

Abstract

Language models (LMs) exhibit and amplify many types of undesirable biases learned from the training data, including gender bias. However, we lack tools for effectively and efficiently changing this behavior without hurting general language modeling performance. In this paper, we study three methods for identifying causal relations between LM components and particular output: causal mediation analysis, automated circuit discovery and our novel, efficient method called DiffMask+ based on differential masking. We apply the methods to GPT-2 small and the problem of gender bias, and use the discovered sets of components to perform parameter-efficient fine-tuning for bias mitigation. Our results show significant overlap in the identified components (despite huge differences in the computational requirements of the methods) as well as success in mitigating gender bias, with less damage to general language modeling compared to full model fine-tuning. However, our work also underscores the difficulty of defining and measuring bias, and the sensitivity of causal discovery procedures to dataset choice. We hope our work can contribute to more attention for dataset development, and lead to more effective mitigation strategies for other types of bias.

1 Introduction

Modern neural language models exhibit social biases, such as biases based on gender, religion, ethnicity and other *protected attributes*. These biases may lead to real harms when used in downstream applications (e.g. Hovy and Spruit, 2016; Weidinger et al., 2021). Detecting and mitigating biases in language models has therefore become an important area of research.

Early detection methods relied on lists of words to measure associations with e.g., specific genders (e.g. Caliskan et al., 2017). Most current detection methods work with curated sets of sentence pairs or triplets, and measure differences in sentence probabilities or anaphora resolution probabilities (e.g. May et al., 2019; Nadeem et al., 2021; Nangia et al., 2020; Basta et al., 2019). Proposed mitigation strategies include targeted changes to the training data (e.g., CDA; Lu et al., 2020), training procedure (e.g., adversarial learning; Zhang et al., 2018), model parameters (e.g., INLP; Ravfogel et al., 2020), or language generation procedure (e.g., "self-debiasing"; Schick et al., 2021).

Despite this work, we still lack a proper understanding of how to best measure biases (how do we guarantee the representativeness for real-world harm of a set of sentence pairs, or of a linguistic phenomenon such as anaphora resolution?), how biases are implemented in the language model internals (is there a unified locus, or is, e.g., gender bias the aggregate effect of many independent model decisions?), and what techniques are effective at reducing undesirable downstream behavior (e.g., is data curation more or less effective than filtering output? Is intervening in the model internals feasible?). Empirically, success in detecting and mitigating biases depends on many factors, including the choice of embeddings, training regimes, data sets and model choices (Blodgett et al., 2020, 2021; Talat et al., 2022; Delobelle et al., 2022; Barrett et al., 2019; Van Der Wal et al., 2022).

The "black-box" nature of LMs makes it difficult to identify and interpret how bias manifests and propagates in them, especially relying solely on correlational methods. The starting point for the current paper is the intuition that if, instead, it were possible to find *causal* relationships between the model's internal representations and its downstream bias, we could more effectively measure and intervene on these undesirable behaviors.

We therefore turn to a recent series of papers on interpretability methods that focus on causal discovery. In Section 2 we discuss three such methods,

[*] Shared senior authorship.

Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 379–394
December 7, 2023. ©2023 Association for Computational Linguistics

of which we adapt one (DiffMask) for our needs in Section 3. Our new method is more efficient than other causal methods, which is especially relevant when applied to large language models (LLMs). In Section 3 we also report results from these three methods when applied to GPT2-small and the problem of gender bias, and find that they discover largely overlapping sets of components, despite huge differences in computation requirements. In Section 4 we use the identified components to adapt GPT-2 small, using parameter-efficient fine-tuning procedures. We demonstrate how gender bias in LMs can be reduced with minimal effect to their language modelling performance by making targeted interventions to their components. However, we also recognize the limitations of operationalizing gender bias as we do, using minimal pairs of contrasting sentences—which simplify gender as a *binary* construct and may not work so well for other languages than English—and call for future research to develop reliable and validated bias measures (see van der Wal et al., 2023).

2 Related Work

Where and how LMs implement output behaviors—from high-level phenomena like gender stereotypes, to lower-level ones like subject-verb agreement—is an active field of study. In providing an overview of related work, we focus on causal methods for locating mechanisms in section 2.1, as non-causal methods can yield misleading conclusions (Ravichander et al., 2021; Elazar et al., 2021). Further, we review previous work on targeted changes to Language models and their behavior in section 2.2

2.1 Locating Mechanisms in Language Models

Causal methods study model processing by intervening in (altering) model processing, and observing the changes in model behavior caused by these interventions. They aim to address the shortcomings in observational methods by ensuring a causal link between mechanisms found in model internals, and model behavior.

Many such techniques determine which representations or components are important to model processing by ablating them. Ablations can range from zeroing out neurons (Lakretz et al., 2019; Mohebbi et al., 2023), to replacing them with a baseline (De Cao et al., 2021a; Bau et al., 2018), or replacing them with another example's activation (Vig et al., 2020; Geiger et al., 2021). All of these techniques return unstructured sets of important components without specifying their interaction.

In recent years, the *circuits* abstraction of transformer models (Elhage et al., 2021) has become popular. This framework views transformer models as computational graphs, and aims to find subgraphs responsible for certain tasks. This technique has been used to find circuits for indirect object detection and the greater-than operation in GPT-2 (Wang et al., 2023; Hanna et al., 2023), as well as to study larger models (Lieberum et al., 2023); it has also been automated (Conmy et al., 2023).

Note that although causal methods can provide a higher degree of confidence in localizing mechanisms, they are not foolproof. For example, Meng et al. (2023) propose causal tracing, a method for locating fact storage in LMs; they then edit GPT-2 XL's factual knowledge by performing edits at relevant locations. However, recent work has showed that although edits may be successful, the localization found by causal tracing is not predictive of edit success (Hase et al., 2023). So, even causal localizations should be assessed thoroughly.

2.2 Targeted Changes to Language Models and Their Behavior

One way to mitigate bias in LMs is to change their parameters or internal representations; however, making large changes can be computationally expensive and have unintended side-effects on model behavior. Past work has studied how to make targeted changes to LMs that avoid these pitfalls. We only discuss works on intervening in the model's representations and parameter-efficient fine-tuning on curated datasets, but other bias mitigation strategies exist as well (see e.g., Meade et al., 2022).

Model Interventions One line of research focuses on removing undesirable concepts from a LM's representations directly. Early methods like *hard-debias* based on principal component analysis (Bolukbasi et al., 2016) and *iterated null-space projection* (INLP, Ravfogel et al., 2020) identify and remove linear representations of gender (bias) from embedding spaces; while others make targeted changes to the activations of LMs (De Cao et al., 2021b; Belrose et al., 2023) or edit the components directly (Meng et al., 2022, 2023).

Altering activations at run-time is one promising way to mitigate (gender) bias in LMs. LEACE (Belrose et al., 2023), for example, convincingly removes linearly-encoded gender information from activations. Similarly, De Cao et al. (2021b) use an approach called *differentiable masking* (DiffMask) to identify small neuron subsets responsible for bias and intervene on them for reducing bias.

However, a downside of these activation-altering methods is that they require an intervention on the activations at each inference step. Moreover, it is not obvious which model activations we should run these on; for instance, it is unlikely that we want to remove gender information from every input token.

Parameter-Efficient Fine-tuning Another approach that avoids some of the pitfalls of changing the LM's representations directly, is to fine-tune on a carefully constructed dataset. Previous work has shown the importance of considering the training data in understanding the biases learned by LMs (e.g., Zhao et al., 2018; Zmigrod et al., 2019; Bordia and Bowman, 2019; Lu et al., 2020; Bender et al., 2021; Sellam et al., 2022; Van Der Wal et al., 2022; Biderman et al., 2023). Given this, fine-tuning on curated datasets is a promising strategy for mitigating gender bias in LMs (Solaiman and Dennison, 2021; Levy et al., 2021; Gira et al., 2022; Kirtane and Anand, 2022). Falling within this paradigm is *parameter-efficient* fine-tuning, where only some of the model parameters are updated—this may not only be computationally more efficient, but even yield better results (Lauscher et al., 2021; Gira et al., 2022; Xie and Lukasiewicz, 2023).

Our work is most similar to Gira et al. (2022), who also use parameter-efficient fine-tuning for debiasing GPT-2 small. However, we study the effect of fine-tuning individual attention heads, while they focus on embedding layers, LayerNorm parameters, adding linear input/output transformation parameters, and a combination thereof. Moreover, Gira et al. do not adhere to any specific strategy when selecting the components to fine-tune. In contrast, our method provides a principled approach to identify the components that are causally important for the task at hand and then fine-tune them.

Xie and Lukasiewicz's (2023) work is also related to ours. They verify the effectiveness of parameter-efficient bias mitigation techniques like adapter tuning (Houlsby et al., 2019) and prefix tuning (Li and Liang, 2021) on various types of LMs and biases. These methods introduce extra tuneable parameters instead of directly tuning the model parameters themselves.

Our approach could mitigate gender bias to an extent with minimal degradation in language modelling performance, similar to the results of Xie and Lukasiewicz (2023) and Gira et al. (2022). However, making a direct comparison is challenging due to differences in evaluation criteria and employed datasets. Gira et al. (2022) exclusively assess their method on StereoSet (Nadeem et al., 2021), whereas we have evaluated our approach on multiple benchmarks, as discussed in Section 4.2. Xie and Lukasiewicz (2023) evaluate their fine-tuning methods using similar benchmarks as ours, but they employ the older CrowS-Pairs (Nangia et al., 2020) dataset for stereotype score and WikiText2 (Merity et al., 2016) for perplexity. We use a newer, improved version of CrowS-Pairs (Névéol et al., 2022) and the much larger WikiText-103 (Merity et al., 2016) instead.

3 Locating Gender Bias

In this section, we investigate the question: where in a given LM is gender bias introduced? We study this in GPT-2 small (Radford et al., 2019), an English-language, auto-regressive pre-trained transformer LM.[1] Its small size—12 transformer layers, with 12 attention heads and 1 multi-layer perceptron (MLP) each—makes it a good object of close studies like we perform. We seek to identify the subset of the 144 attention heads that introduce gender bias into the last position of GPT-2's input, where GPT-2 produces next-token predictions. We identify these heads in the context of inputs that lead to gender-biased next-tokens from GPT-2.

This study thus focuses on attention heads. Though prior work has emphasized the role of MLPs in gender bias and memorization (Vig et al., 2020; Geva et al., 2022; Meng et al., 2023), we argue that attention heads are also an interesting subject of analysis. Unless the final word of the input contains gender information that causes the production of biased next-tokens, this information must be introduced from other positions via attention heads.

To determine where GPT-2 small introduces gender bias into its output, we use three methods: causal mediation analysis (CMA), automated circuit discovery, and our own novel method that combines the first approach with differential masking. We then compare the results of these three methods.

[1] The code for our experiments can be found here: `https://github.com/iabhijith/bias-causal-analysis`

3.1 Methodology

All methods we use rely on a core technique as outlined in Vig et al. (2020): swapping model component activations during a forward pass on one input, with activations taken from the model when run on another input which induces an opposite behaviour in the model. For this purpose, we use the *Professions dataset* from Vig et al. (2020), which contains templated sentences designed to elicit gender bias. The sentences in the dataset take the form "The {profession} said that". GPT-2's continuations on these sentences tend to be stereotypical—if the profession is *nurse*, GPT-2 outputs *she*, while if it is *doctor*, GPT-2 outputs *he*.

For each sentence in the dataset we generate a corresponding counterfactual sentence with the profession word replaced by anti-stereotypical gender-specific word. If the normal sentence's profession is female-stereotyped, its corresponding counterfactual sentence is "The *man* said that"; for male-stereotyped professions, the counterfactual contains *woman*. These sentences are designed to maximize the change in model behavior with respect to the predicted pronoun; this makes it easier to identify important components. The dataset contains sentences generated from 17 templates and 299 professions resulting in 5083 sentences in total. For all methods that follow, we intervene on the last position of the sentence.

3.1.1 Causal Mediation Analysis

Vig et al. (2020) were the first to use CMA (Pearl, 2014) to locate gender bias in GPT-2; we adopt their methods as a baseline. CMA relies on a simple hypothesis: if a component is important to the model's behavior on a task, swapping its output activation with another will change model behavior. More formally, let $\mathbf{x}$ and $\tilde{\mathbf{x}}$ be normal and counterfactual inputs respectively, and let i be the index of the component (attention head or MLP) under investigation. We first run the model on $\mathbf{x}$, and observe its output distribution $p(y|\mathbf{x})$, Then, we run the model on $\tilde{\mathbf{x}}$ and save $\tilde{\mathbf{h}}_i$, the counterfactual output of component i. Then we run the model on $\mathbf{x}$ again, but replace $\mathbf{h}_i$ with $\tilde{\mathbf{h}}_i$ during the forward pass. This yields an altered model output distribution $\tilde{p}(y|\mathbf{x})$. Vig et al. (2020) measure how important a component i is to a model behaviour b using Natural Indirect Effect (NIE), the expected proportional difference in model behavior after intervening on component i. If b_{null} is the original behaviour of the model and $b_{i,intv}$ is the behaviour of the model after intervening on component i, then NIE can be evaluated as shown in Equation (1):

$$\text{NIE}(i, b) = \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\in\mathcal{D}}\left[\frac{b_{i,intv}}{b_{null}} - 1\right] \tag{1}$$

Vig et al. (2020) use the definition in eq. (2) to measure biased behaviour in a LM. It is the ratio of the probabilities assigned by the model to an anti-stereotypical continuation as against a stereotypical continuation given a context. In case of Professions dataset (Vig et al., 2020), it is the ratio of probability assigned to anti-stereotypical pronoun versus the probability assigned to stereotypical pronoun.

$$b(\mathbf{x}) = \frac{p(y = \text{anti-stereo}|\mathbf{x})}{p(y = \text{stereo}|\mathbf{x})} \tag{2}$$

The aforementioned technique analyzes individual components; Vig et al. propose two methods to gather a *set* of important components. Using the top-k strategy, they evaluate every component, and select the k components that cause the most change in model behavior. Using the k-greedy strategy, they evaluate all components, and add the most impactful one. Then, they evaluate each component again, ablating both it *and* their set; they once again add the most impactful component. They repeat the latter step until they have a set of size k.

3.1.2 Circuit Discovery

The circuits framework, which views models as computational graphs, provides a related technique for identifying mechanisms in LMs. While Vig et al.'s CMA approach generates a component set (nodes) relevant to a task, the circuits approach generates a set of edges, resulting in a detailed subgraph. However, the underlying methodology is similar to CMA: we ablate edges via swaps, and see which edges hurt performance once ablated. Though our fine-tuning techniques only target nodes (not edges), comparing CMA and circuits localisations of bias could still be insightful.

We use Conmy et al.'s (2023) automated circuit discovery code (ACDC) to identify model components relevant to (gender) bias. This technique iteratively tests model edges, removing those that can be ablated without changing task performance. We use ACDC on the same professions dataset as CMA, and measure task performance as the difference in probability assigned to stereotypical and non-stereotypical pronoun continuations.

3.1.3 Differentiable Masking With CMA

We finally propose our own method for localizing relevant LM components that combines two approaches: Vig et al.'s (2020) CMA and De Cao et al.'s (2021a) differentiable masking (DiffMask). Our method is motivated by a notable challenge with CMA, namely, how to select the best size-k subset of model components that contributes to bias. Vig et al.'s two strategies for this (top-k and k-greedy as discussed in Section 3.1.1) both have downsides. A top-k strategy assumes that components' importance is independent, while a k-greedy strategy is expensive, requiring k evaluations of all components' importance. A full sweep of the search space would be combinatorially expensive.

This combinatorial search problem can be reformulated as an optimization problem using a differentiable relaxation (Louizos et al., 2018; Bastings et al., 2019; De Cao et al., 2021a,b; Schlichtkrull et al., 2021). DiffMask, proposed by De Cao et al. (2021b) precisely apply the reformulation to learn an almost-binary differentiable stochastic mask over a model's components, indicating which are important, and which are not. Unimportant components are those whose outputs can be ablated without changing model behavior.

We adapt DiffMask in two ways, and label our variant DiffMask+. First, instead of using surrogate models that instantiate distribution per input, we directly learn a distribution for the stochastic mask. This change is crucial because it helps us identify a single, generalizable set of components responsible for bias in the language model across the entire dataset, which is essential for downstream fine-tuning. Second, instead of learning interventions to ablate a component's activations, we use corresponding activations generated from the counterfactual sentences.

Besides these changes, training and inference with this mask proceed as in De Cao et al. (2021b). At every time step, we run a forward pass of the model on an example from the *Professions dataset*. We stochastically replace component outputs with corresponding counterfactual outputs, according to the mask; components with higher mask weights are replaced to a greater degree. We train the mask to induce the largest change in gendered pronoun prediction possible, while minimizing both the number of non-zero mask entries, and the magnitude of overall changes made to the model's output distribution. This procedure yields a mask over our components, whose expected values lie in $[0, 1]$; higher values indicate more important components. For more details, see Appendix B.

3.2 Experiments

We use the three methods discussed above to discover the components that cause gender bias in GPT-2 small. For CMA and DiffMask+, we limit our analysis to attention heads. All experiments were implemented using the TransformerLens[2] library (Nanda and Bloom, 2022). For CMA, we used Vig et al.'s top-k strategy and selected only the top 10 heads as the NIE quickly diminishes beyond this point. Similarly, for DiffMask+, we chose the 10 heads with the highest expected mask value at the end of training. To find our circuit, we ran ACDC, finding a whole circuit containing attention heads and other components as shown in Figure 4 in Appendix A. For hyperparameters and training details, see Appendix C.

3.3 Results

Figure 1 shows the attention heads selected using each method. For ACDC, we show only the attention heads from the full circuit. All methods find attention heads located mostly in the final layers of the model; this contrasts with Vig et al. (2020), who find heads in middle layers. This may be due to the fact that Vig et al. (2020) mainly assess gender bias in co-reference resolution in their attention intervention experiments and accordingly use the WinoBias (Zhao et al., 2018) and Winogender (Rudinger et al., 2018) datasets. The results suggest that the dataset used for discovery influences the components picked by these methods.

The Venn diagram in Figure 1 shows the overlap of heads across methods. We observe a significant overlap: 5 of the top 10 heads are shared by all three methods. Attention heads selected using CMA and ACDC have more overlap and as observed in the mitigation results in Section 4.3 the two methods perform similarly on different metrics. The fact that DiffMask+ yields 4 heads that are not shared might be due to its objective: DiffMask+ attempts to maximally change gendered pronoun prediction *while still minimally changing the distribution overall*. This latter constraint is absent from the other two methods.

We also note that the selected heads are located in the later half of the model. We hypothesize that

[2] https://github.com/neelnanda-io/TransformerLens

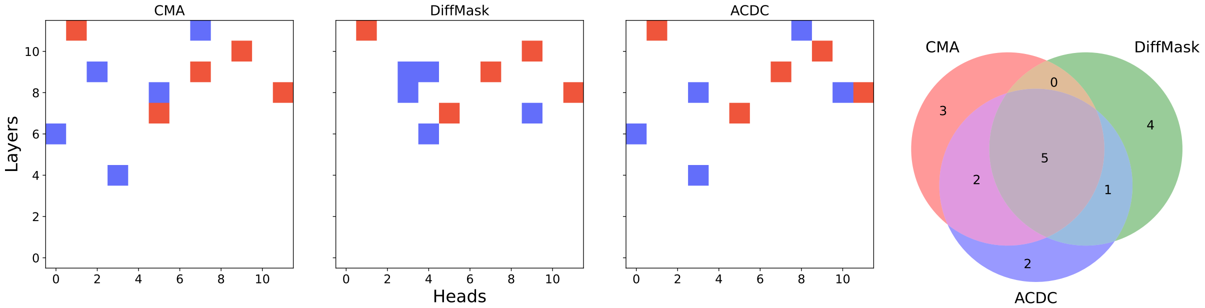

Figure 1: Top 10 attention heads selected using CMA, DiffMask+ and ACDC. Overlapping heads are shown in red. The Venn diagram shows the overlap counts between all combinations of the sets.

this may be because these heads are transferring gender information from the profession position to the end position of the sentence. Although earlier heads can also attend to gender tokens, prior work suggests that entities are enriched by lower-layer MLPs before information is extracted from them by later attention heads (Geva et al., 2023).

4 Mitigating Gender Bias

Having identified components responsible for gender bias in GPT-2 small, we test whether this information can be used to mitigate the bias. To this end, we fine-tune the model on a dataset carefully curated to be gender balanced—this has been shown to lead to a reduction in gender bias (Gira et al., 2022). We compare the effectiveness of fine-tuning only the components found in the previous section to various baselines, both fine-tuned and not.

4.1 Fine-tuning Dataset and Models

We test the effectiveness of parameter-efficient fine-tuning with the identified GPT-2 components at mitigating gender bias. We fine-tune on the BUG dataset[3] (Levy et al., 2021), which contains annotated natural sentences containing one or more gendered pronouns. We use the balanced version of BUG, which has an equal number of masculine and feminine pronouns, to counteract GPT-2's gender bias in pronouns. For each model in Table 1, we fine-tune only the specified subset of GPT-2's parameters and compare our methods to the not fine-tuned GPT-2 model, our **baseline**. Appendix D contains fine-tuning details.

4.2 Metrics

We use several metrics and baselines to evaluate the effectiveness of the bias mitigation under the different conditions. To measure gender bias, we use WinoBias (Zhao et al., 2018) and the gender bias subset of CrowS-Pairs by Névéol et al. (2022). We also measure model performance on the original *Professions dataset* using which important components were found. To ensure that fine-tuning did not harm models' general language modeling abilities, we also measure these, via WikiText perplexity (Merity et al., 2016) and accuracy on BLiMP (Warstadt et al., 2020). All metrics, except for the perplexity, are defined as the ratio of times that the model prefers the correct/anti-stereotypical over the incorrect/stereotypical variant. Given a dataset $\mathcal{D}$ with pairs of stereotypical and anti-stereotypical sentences $(\mathbf{x}, \tilde{\mathbf{x}})$, the Stereotype Score is defined as follows.

$$\mathrm{SS} = \frac{1}{|\mathcal{D}|} \sum_{(\mathbf{x},\tilde{\mathbf{x}})\in\mathcal{D}} \mathbb{I}_{p(\mathbf{x})>p(\tilde{\mathbf{x}})} \tag{3}$$

WinoBias We measure the models' gender bias using WinoBias. Even if this dataset with its small linguistic variety might not exactly reflect real-world biased language (Lior and Stanovsky, 2023),

Table 1: All fine-tuned models and corresponding components selected for fine-tuning in Section 4. DM means our proposed method DiffMask+.

Model Name	Selected Components
Full Model	Entire model.
Random Attn Heads	Set of 10 randomly selected attention heads *not* found by CMA, ACDC or DM.
All Attn Layers	All attention layers including the attention projection.
Last 4 Attn Layers	Last 4 attention layers.
ACDC	MLPs, attention heads, and embedding layers found by ACDC.
ACDC Attn Heads	Attention heads from the ACDC circuit.
CMA Attn Heads	Top 10 attention heads found by CMA.
DM Attn Heads	Top 10 attention heads found by DiffMask+.

[3] https://github.com/SLAB-NLP/BUG

it is widely used as its simplicity allows for controlled experiments. We measure models' gender bias using WinoBias' type 2 dataset[4] (Zhao et al., 2018). This dataset consists of sentences containing two occupation terms and one gendered pronoun; models must determine which occupation the pronoun refers to. In type 2 examples, the sentence's syntax always determines the correct occupation (regardless of the pronoun's gender). For each sentence there is one pro- and one anti-stereotypical version, which differ only in the gender of the pronoun used. We consider a model biased if it consistently assigns higher probability to the pro-stereotypical sentence. We record the proportion of examples where the model assigns higher probability to the pro-stereotypical version. Note that our metric differs from the original metric, which was formulated in terms of co-reference resolution accuracy.

CrowS-Pairs The gender bias subset of CrowS-Pairs measures gender bias in LMs, construed more broadly than occupation-gender associations. It consists of minimal pairs, a more and a less stereotypical sentence. We consider a systematic preference for more stereotypical sentences (by comparing perplexities) to indicate a biased model. As in WinoBias, the bias is measured as the proportion of examples where the model prefers the stereotypical sentence. In our experiments, we use an updated version from Névéol et al. (2022) where potential validity issues (including those identified by Blodgett et al. (2021)) have been addressed.

Professions We use the *Professions dataset*, with which we found bias-relevant components, to assess gender bias in the fine-tuned models. For every sentence in the dataset, we measure the probability assigned to the pro-/anti-stereotypical continuations (either *he* or *she*, depending on the example). We measure the proportion of examples where the pro-stereotypical continuation is more probable.

BLiMP We evaluate our models' linguistic abilities using BLiMP. BLiMP consists of a number of datasets, each of which targets a specific linguistic phenomenon. Each dataset contains examples, each of which is a minimal sentence pair: one sentence is correct and the other incorrect, with respect to the targeted phenomenon. The model should systematically assign a higher probability to the correct sentence. We report accuracy on BLiMP as a whole, as well as on the *Gender Anaphor Agreement* (AGA) and *Subject Verb Agreement* (SVA) subtasks. We do this to understand the effect of our fine-tuning on these specific linguistic phenomena, where gender is only relevant for one of these tasks.

WikiText We evaluate our models' general language modeling performance by computing their perplexity on the test split of the WikiText-103 corpus[5] (4358 examples) (Merity et al., 2016), which consists of "Good" and "Featured" Wikipedia articles. Higher perplexity might indicate that fine-tuning hurt general language modeling abilities.

4.3 Results

Table 2 presents the average bias evaluation results for CrowS-Pairs, WinoBias, and Professions, as well as for the perplexity and BLiMP metrics.

Bias Metrics We find that all types of fine-tuning improve performance on the *Professions dataset* (details in the appendix; Figure 5). This suggests that the fine-tuning procedure successfully changed model behavior. However, not all types of fine-tuning are equal: fine-tuning strategies that targeted late attention heads yielded models with lower stereotyping and variance than those that targeted other components, spread throughout the model.

Similarly, the CrowS-Pairs results in Figure 2 show that models where only the attention heads discovered using the three methods from Section 3 were fine-tuned, achieve the best results in terms of gender bias reduction. In contrast, fine-tuning random attention heads yields no reduction in gender bias. The DM Attention Heads model in particular significantly reduces bias with an average stereotype score as defined in eq. (3) from 0.58 of the baseline to 0.55. Additionally, the scores of DM Attention Heads model have low variance while fine-tuning all attention layers, the full model, or ACDC components yields high-variance results.

Evaluation on WinoBias yields contrasting results (Table 2). Fine-tuning the attention heads only marginally reduced the gender bias on average. Surprisingly, fine-tuning the last 4 attention layers achieved the best reduction in gender bias.

At first glance, the CrowS-Pairs and WinoBias results are mixed. Fine-tuning the full model, last

[4] We choose not to discuss the results for the type 1 dataset because we do not test an actual co-reference resolution task, but rather compute the perplexities of continuing with one or the other gendered pronoun.

[5] https://huggingface.co/datasets/wikitext

Table 2: Effect comparison of the different fine-tuning interventions. Reported are perplexity (PPL, measured on WikiText), three measures of linguistic adequacy (full BLiMP as well as subject-verb and anaphora agreement portions of BLiMP), and the gender bias measures from CrowS-Pairs, WinoBias, and the Professions benchmarks/datasets. The cells show the % improvement (positive is better as indicated by ↑) w.r.t. the original GPT-2 before fine-tuning, averaged over 5 seeds (absolute scores are in Appendix E). * indicates $p < 0.05$ for two-sided one sample *t*-test, where the original GPT-2 performance serves as the population mean.

		perplexity ↑	*linguistic adequacy* ↑			*gender bias measures* ↑		
		PPL	BLiMP	SV	AGA	CrowS.	WinoB.	Prof.
baselines	full model	-44.2	-3.9*	-2.9*	1.2*	-1.4	4.6*	2.3
	random attn heads	-17.0	-3.0*	-0.9*	0.2	-0.2	1.9	1.3
broad interventions	all attn layers	-19.1	-2.0*	-1.4*	1.4*	-0.6	0.1	0.6*
	last 4 attn layers	-12.6	-3.4*	0.4*	-1.2*	-0.2	4.2*	3.2*
	acdc	-38.8	-4.6*	-1.8*	0.6	-0.9	3.3*	3.1
narrow interventions	acdc attn heads	-16.6	-3.0*	0.3*	0.2	3.5*	1.4	1.9*
	cma attn heads	-16.6	-3.0*	0.3*	0.2	3.5*	1.4	1.9*
	dm attn heads	-17.5	-2.4*	0.2	-0.0	4.8*	0.9	2.9*

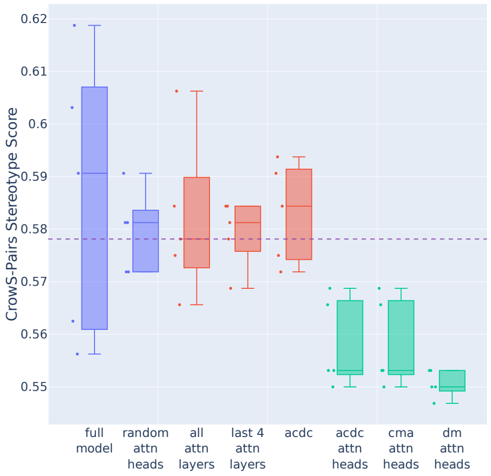

Figure 2: CrowS-Pairs results (here: lower is better). Purple models are baselines; the dotted line shows the non-fine-tuned GPT-2 performance.

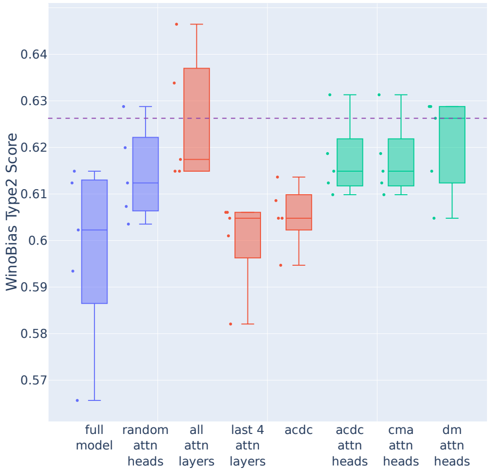

Figure 3: WinoBias Type2 Stereotype Score (here: lower is better). Purple models are baselines; the dotted line shows the non-fine-tuned GPT-2 performance.

4 attention layers, or ACDC components yields the most improvement on WinoBias, but these models score badly on CrowS-Pairs. However, the reverse is not true: the models that improved most on CrowS-Pairs also improved on WinoBias—although not consistently (Figure 3). We postulate many potential explanations for the divergent outcomes seen between WinoBias and CrowS-Pairs. First, WinoBias could simply be rewarding models that perform randomly or poorly at co-reference resolution, although good overall BLiMP AGA scores suggest this is not the case. Second, gender bias in co-reference resolution might stem from a component set distinct from the ones we discovered. This is supported by Vig et al.'s findings, which revealed a distinct set of attention heads that contribute to gender bias in co-reference resolution. Finally, this might be linked to how the bias measures are operationalized, which we will come back to in Section 5.

WikiText & BLiMP Both the perplexity measured on WikiText and accuracies on BLiMP inform us about the general language modeling capability before and after fine-tuning. For WikiText, we observe that fine-tuning more parameters—as when we fine-tune the full model or ACDC circuit—hurts the perplexity more; the fully fine-tuned model performs the worst, increasing perplexity to 34.16 from 23.69. In contrast, targeted fine-tuning of attention heads increases perplexity by a much lower margin. This trade-off motivates finding a minimal component set to fine-tune, in order to mitigate bias while maintaining general language modeling ability.

All fine-tuned models attain lower performance on BLiMP overall than the pre-trained baseline; as in the WikiText case, the more components fine-tuned, the more performance drops. However, examining the performance on agreement subtasks reveals more nuance. On SVA, fine-tuning only the top-10 attention heads found using the methods from Section 3 improved performance by a small margin. On AGA, almost all fine-tuned models attained scores on par with the baseline. So, while fine-tuning small sets of attention heads hurt BLiMP performance overall, the maintained performance on SVA and AGA suggest that agreement ability, gender-related or not, are not hurt.

5 Discussion & Conclusions

With this work, we provide an exploratory study of the identification and mitigation of gender bias in GPT-2. Our three different methods identify model components relevant to gender bias—according to our results, they largely agree on the most relevant attention heads: most of the heads responsible for gender bias are found mainly in the last four attention layers. We then intervene on each method's found components to mitigate the gender bias but maintain language modeling performance. We find that language modeling performance deteriorates only minimally for our 'narrow' interventions, but deteriorates more in conditions where a larger amount of components/parameters are adapted by fine-tuning.

Regarding computational efficiency, we find that the circuits approach is computationally inefficient compared to the other methods. For explanatory and exploratory work, like ours, circuits are very useful and can yield fine-grained insights into the model mechanisms. However, if resource efficiency is a high priority, we suggest using other methods than (automatic) circuit discovery. One key contribution of this paper is a new and very efficient method, DiffMask+, which finds a minimal set of attention heads for fine-tuning, while being computationally less prohibitive than methods such as automatic circuit discovery.

Limitations Have we reached our goal of reducing bias, using computational efficient methods? Considering the measured gender bias, we successfully reduced the bias on two out of three datasets. This is encouraging, but our results also reveal some inconsistencies between different ways of measuring bias. This is not unexpected; in fact, much previous work has highlighted many issues that put the validity and reliability of current bias measures into question (e.g., Blodgett et al., 2021; Talat et al., 2022; Dev et al., 2022). Bias measures may target very different manifestations of the bias of interest (van der Wal et al., 2023). We therefore attribute the observed inconsistencies to the implicit versus explicit gender bias in different datasets, which could be represented differently in model components, and thus also targeted differently by fine-tuning.

Despite these challenges, we tried to address some of these concerns by using multiple different bias metrics and testing the consistency of these across different seeds. We believe that the success of our approach is heavily contingent upon the datasets employed for both component identification and the subsequent fine-tuning of the chosen components. For example, using template-based datasets such as WinoBias or Professions could reduce the identified components' generalizability, as components that contribute to one form of gender bias may not contribute to another. The same applies to the fine-tuning stage as well. Using a dataset with limited variability in structure might result in only partial mitigation of the behavior. We therefore conclude that for even better bias reduction, it is essential to use and develop datasets that are diverse and representative of the behaviour being studied.

Future work For a wider picture of how our findings integrate in bias identification and mitigation studies, we would like to compare our approaches to other promising methods in the literature like concept erasure at the activation level (e.g., LEACE; Belrose et al., 2023) and changes to the language generation procedure (e.g., "self-debiasing"; Schick et al., 2021). Future work should also test whether these mitigation strategies generalize to different conditions, for example, language models larger than GPT-2 small. Lastly, we also stress the importance of developing methodologies for operationalizing other forms of bias than binary gender in English, and to overcome difficulties we currently face when using contrastive sets and existing bias benchmarks.

6 Acknowledgements

OW's contributions are financed by the Dutch Research Council (NWO) as part of project 406.DI.19.059.

References

Maria Barrett, Yova Kementchedjhieva, Yanai Elazar, Desmond Elliott, and Anders Søgaard. 2019. Adversarial removal of demographic attributes revisited. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6330–6335, Hong Kong, China. Association for Computational Linguistics.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, page 33–39, Florence, Italy. Association for Computational Linguistics.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable neural predictions with differentiable binary variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James R. Glass. 2018. Identifying and controlling important neurons in neural machine translation. *ArXiv*, abs/1811.01157.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form. *arXiv preprint arXiv:2306.03819*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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)*, pages 1004–1015, Online. Association for Computational Linguistics.

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.

Shikha Bordia and Samuel Bowman. 2019. Identifying and reducing gender bias in word-level language models. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 7–15.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. ArXiv:1608.07187 [cs].

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. (arXiv:2304.14997). ArXiv:2304.14997 [cs].

Nicola De Cao, Michael Schlichtkrull, Wilker Aziz, and Ivan Titov. 2021a. How do decisions emerge across layers in neural models? interpretation with differentiable masking. (arXiv:2004.14992). ArXiv:2004.14992 [cs, stat].

Nicola De Cao, Leon Schmid, Dieuwke Hupkes, and Ivan Titov. 2021b. Sparse interventions in language models with differentiable masking. (arXiv:2112.06837). ArXiv:2112.06837 [cs].

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 1693–1706, Seattle, United States. Association for Computational Linguistics.

Sunipa Dev, Emily Sheng, Jieyu Zhao, Aubrie Amstutz, Jiao Sun, Yu Hou, Mattie Sanseverino, Jiin Kim, Akihiro Nishi, Nanyun Peng, et al. 2022. On measures of biases and harms in nlp. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 246–267.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac

Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. (arXiv:2106.02997). ArXiv:2106.02997 [cs].

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 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.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does gpt-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. 2023. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Neeraja Kirtane and Tanvi Anand. 2022. Mitigating gender stereotypes in Hindi and Marathi. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 145–150, Seattle, Washington. Association for Computational Linguistics.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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)*, pages 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shahar Levy, Koren Lazar, and Gabriel Stanovsky. 2021. Collecting a large-scale gender bias dataset for coreference resolution and machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2470–2480, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *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)*, abs/2101.00190.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. 2023. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in chinchilla.

Gili Lior and Gabriel Stanovsky. 2023. Comparing humans and models on a similar scale: Towards cognitive gender bias evaluation in coreference resolution.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Christos Louizos, Max Welling, and Diederik P. Kingma. 2018. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. *Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday*, pages 189–202.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. 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)*, page 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2023. Locating and editing factual associations in gpt. (arXiv:2202.05262). ArXiv:2202.05262 [cs].

Kevin Meng, Arnab Sen Sharma, Alex Andonian, Yonatan Belinkov, and David Bau. 2022. Mass-editing memory in a transformer. *arXiv preprint arXiv:2210.07229*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *ArXiv*, abs/1609.07843.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. 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)*, page 5356–5371, Online. Association for Computational Linguistics.

Neel Nanda and Joseph Bloom. 2022. Transformerlens.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. French crows-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8521–8531.

Judea Pearl. 2014. Interpretation and identification of causal mediation. *Psychological methods*, 19.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Michael Sejr Schlichtkrull, Nicola De Cao, and Ivan Titov. 2021. Interpreting graph neural networks for {nlp} with differentiable edge masking. In *International Conference on Learning Representations*.

Thibault Sellam, Steve Yadlowsky, Ian Tenney, Jason Wei, Naomi Saphra, Alexander D'Amour, Tal Linzen, Jasmijn Bastings, Iulia Raluca Turc, Jacob Eisenstein, Dipanjan Das, and Ellie Pavlick. 2022. The multiberts: BERT reproductions for robustness analysis. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. In *Advances in Neural Information Processing Systems*, volume 34, pages 5861–5873. Curran Associates, Inc.

Zeerak Talat, Aurélie Névéol, Stella Biderman, Miruna Clinciu, Manan Dey, Shayne Longpre, Sasha Luccioni, Maraim Masoud, Margaret Mitchell, Dragomir Radev, et al. 2022. You reap what you sow: On the challenges of bias evaluation under multilingual settings. In *Proceedings of BigScience Episode# 5–Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 26–41.

Oskar van der Wal, Dominik Bachmann, Alina Leidinger, Leendert van Maanen, Willem Zuidema, and Katrin Schulz. 2023. Undesirable biases in nlp: Averting a crisis of measurement. *arXiv preprint arXiv:2211.13709*.

Oskar Van Der Wal, Jaap Jumelet, Katrin Schulz, and Willem Zuidema. 2022. The birth of bias: A case study on the evolution of gender bias in an English language model. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 75–75, Seattle, Washington. Association for Computational Linguistics.

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. In *Advances in Neural Information Processing Systems*, volume 33, page 12388–12401. Curran Associates, Inc.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: A circuit for indirect object identification in gpt-2 small.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Laura Weidinger, John F. J. Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zachary Kenton, Sande Minnich Brown, William T. Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William S. Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2021. Ethical and social risks of harm from language models. *ArXiv*, abs/2112.04359.

Zhongbin Xie and Thomas Lukasiewicz. 2023. An empirical analysis of parameter-efficient methods for debiasing pre-trained language models. (arXiv:2306.04067). ArXiv:2306.04067 [cs].

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Ran Zmigrod, Sabrina J Mielke, Hanna Wallach, and Ryan Cotterell. 2019. Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661.

A Circuit Discovery

The circuit discovered in GPT-2 small model using professions datset is shown in Figure 4

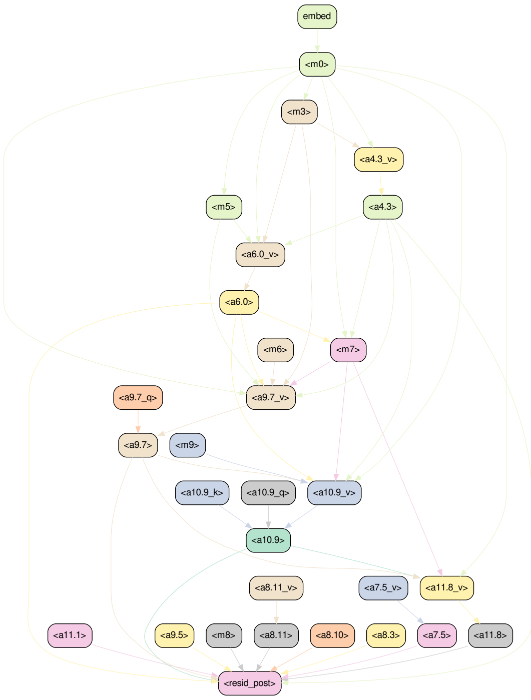

Figure 4: Circuit discovered in the GPT-2 small model using Professions dataset.

B DiffMask+ Implementation Details

During inference, DiffMask+ works as follows. We have two inputs—our normal input $\mathbf{x}$ and our counterfactual input $\tilde{\mathbf{x}}$—as well as a k-dimensional binary mask $\mathbf{m} \in \{0, 1\}^k$; for GPT-2 small, the number of components k is 144 as we choose to only select attention heads. We run forward passes on both inputs, recording each component's output on the normal dataset $(\mathbf{h}_1, \ldots, \mathbf{h}_k)$ and the counterfactual dataset $(\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_k)$. Finally, we run the model once more on the normal input, applying the mask: we replace each original component output $\mathbf{h}_i$ with the potentially masked output $\mathbf{h}'_i = (1 - m_i) \cdot \mathbf{h}_i + m_i \cdot \tilde{\mathbf{h}}_i$[6]. If our mask captures which components are important, our masked model should behave as if it were receiving the counterfactual input.

DiffMask+'s training setup is slightly different. We cannot learn a purely binary mask, as that would not be differentiable. Instead, we learn a

[6] We can apply our mask either at every time step, or at only the final time step.

parameterization of a hard concrete distribution (Louizos et al., 2018), a type of distribution that falls in $[0,1]$ and assigns non-zero probability to both 0 and 1. This distribution is parameterized by a location vector $\mathbf{z} \in [0,1]^k$, and can be sampled to produce a mask $\mathbf{m} \in [0,1]^k$. When it comes time to mask the model, we simply sample a mask from the distribution $p_{\mathbf{z}}(\mathbf{m})$; note that this mask may no longer be strictly binary. However, we can generate a deterministic and truly binary mask for use at inference time in expectation (mask set to 0 if expected value < 0.5, and 1 otherwise).

With this setup, we can train our mask; we begin by initializing the location vector to $[0.5]^k$. We then train it on our dataset $\mathcal{D}$, optimizing a loss adapted from De Cao et al. (2021b) which is composed of three individual loss terms. The first, targets our task of interest—gender bias. If the original input would lead to a prediction of stereotypical pronoun y_o, e.g. "she", and corresponding anti-stereotypical pronoun is y_c, e.g. "he", we minimize $\tilde{p}(y_o|\mathbf{x})/\tilde{p}(y_c|\mathbf{x})$ where $\tilde{p}$ is the intervened or masked model's output distribution. This is minimized when the anti-stereotypical prediction is much more likely than the original stereotypical prediction, i.e. when the relevant model components are intervened with the corresponding counterfactual output.

The second loss term is the expected number of non-zero elements in our sampled mask; we want our mask to be sparse. Ideally, this would be a hard constraint, where the number of non-zero elements is $\leq \alpha$ for a chosen α; we will instead use a Lagrangian relaxation of this constraint. The third term is the KL divergence between the unmasked model's output distribution $p(y|\mathbf{x})$ and masked model's output distribution $\tilde{p}(y|\mathbf{x})$; we want our masking to minimally change model output, besides task-relevant output. Formally, and much like De Cao et al. (2021b), we optimize:

$$\begin{aligned}\max_{\lambda}\min_{\mathbf{z}} & \sum_{\mathbf{x},y_o,y_c\in\mathcal{D}} \frac{\tilde{p}(y_o|\mathbf{x})}{\tilde{p}(y_c|\mathbf{x})} \\ & + \lambda\left(\sum_{i=1}^{k}\mathbb{E}_{p_{z_i}(m_i)}[m_i \neq 0] - \alpha\right) \\ & + \beta D_{KL}(p(y|\mathbf{x})||\tilde{p}(y|\mathbf{x}))\end{aligned} \tag{4}$$

Here, α and β are hyperparameters regulating sparsity and KL-divergence weight, respectively; $\lambda \in \mathbb{R}_{\geq 0}$ is our Lagrangian multiplier. Optimizing this loss should produce a mask that captures the components relevant to gender bias, while being maximally sparse, and still mostly preserving the model's output distribution.

C Component Discovery Hyperparameters

We optimized the DiffMask loss using Adam (Kingma and Ba, 2014) for 200 epochs on the professions dataset with a learning rate 10^{-3} and a constant schedule. We choose the sparsity hyperparameter $\alpha = 10$ for selecting 10 attention heads and the KL-Divergence weight $\beta = 1$ as proposed in De Cao et al. (2021b). At the end of the training, we choose the top-10 heads with the highest expected value of the location parameter of the stochastic mask.

For the ACDC experiment, we chose a threshold of 0.01, eliminating edges if ablating them caused a change in performance of less than 0.01, as measured by our pronoun probability difference metric.

D Fine-tuning experiment

In Section 4, we fine-tune each model for a maximum of 20 epochs using AdamW optimizer (Loshchilov and Hutter, 2017) with an initial learning rate 10^{-4} and a linear schedule. We optimize Cross Entropy Loss. The BUG balanced dataset contains 25844 sentences, which we split into gender-balanced training and validation sets, containing 90% and 10% of the data respectively. We use the validation loss both for selecting the best model and early stopping with a patience of 10 epochs.

E Additional Results

Table 3 shows all results of fine-tuned models and baselines rounded to up to 2 decimals. Figure 5 shows the stereotype scores of different models evaluated on the Professions dataset. Figure 6 shows the perplexity of different models evaluated on WikiText-103. Figure 7 shows the BLiMP overall results measured over 5 different iterations. Similarly, Figure 8 and Figure 9 shows the AGA and SVA results respectively.

Table 3: Comparison of the effect of the different fine-tuning interventions. Reported are perplexity (PPL, measured on WikiText), three measures of linguistic adequacy (full BLiMP, and subject-verb and anaphora agreement portions of BLiMP), as well as the gender biases measures from CrowS-Pairs, WinoBias, and the Professions benchmarks/datasets.

		perplexity	*linguistic adequacy*			*gender bias measures*		
		PPL	BLiMP	SV	AGA	CrowS.	WinoB.	Prof.
baseline	original gpt2	23.69	0.80	0.90	0.95	0.58	0.63	0.84
	full model	34.16	0.77	0.87	0.97	0.59	0.60	0.82
	random attn heads	27.72	0.77	0.89	0.96	0.58	0.61	0.83
broad interventions	all attn layers	28.22	0.78	0.89	0.97	0.58	0.63	0.83
	last 4 attn layers	26.67	0.77	0.90	0.94	0.58	0.60	0.81
	acdc	32.89	0.76	0.88	0.96	0.58	0.61	0.81
narrow interventions	acdc attn heads	27.62	0.77	0.90	0.96	0.56	0.62	0.82
	cma attn heads	27.62	0.77	0.90	0.96	0.56	0.62	0.82
	dm attn heads	27.84	0.78	0.90	0.95	0.55	0.62	0.81

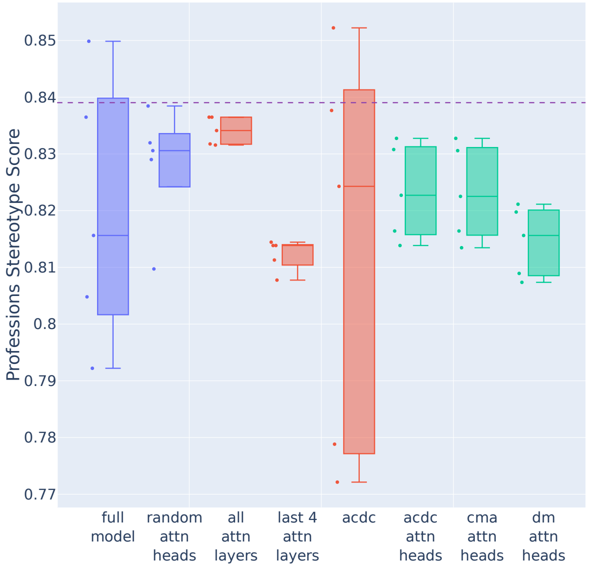

Figure 5: Professions Stereotype Score (here: lower is better). Purple models are baselines; the dotted line shows the non-fine-tuned GPT-2 performance.

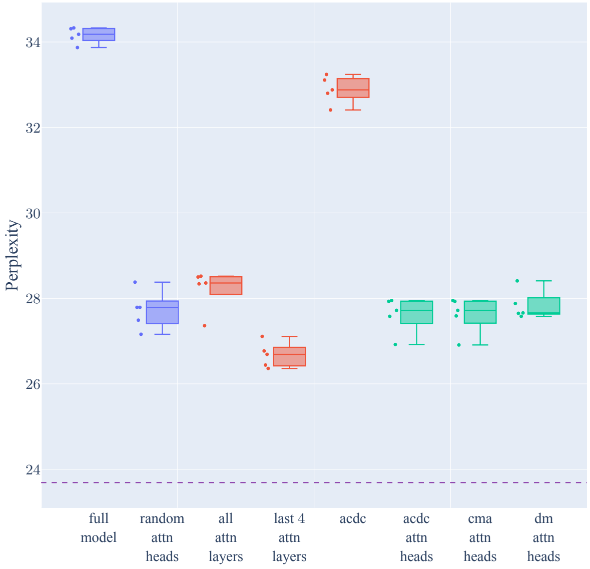

Figure 6: Test perplexity (lower is better) on WikiText-103. Purple models are baselines; the dotted line shows the non-fine-tuned GPT-2 performance.

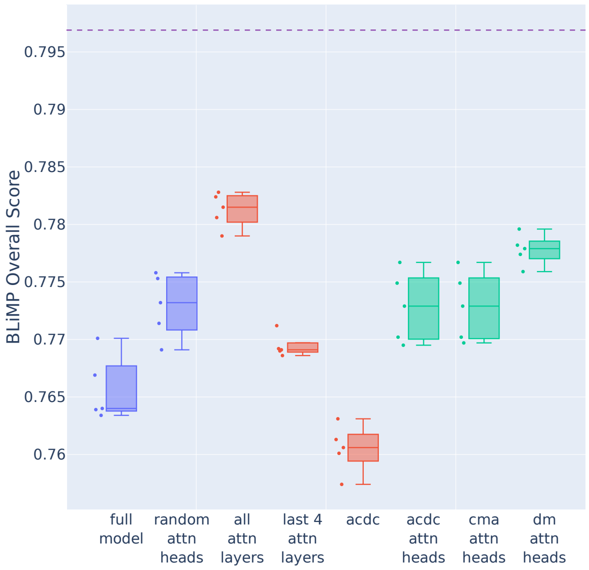

Figure 7: BLiMP Overall results (higher is better). Purple models are baselines; the dotted line shows the non-fine-tuned GPT-2 performance.

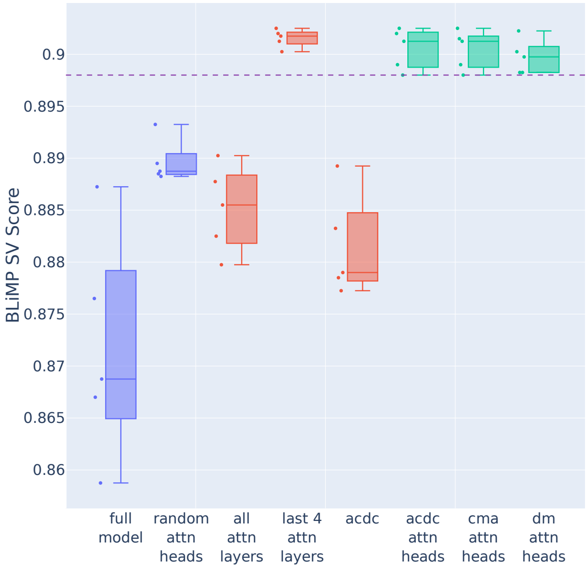

Figure 9: BLiMP Subject Verb Agreement results (higher is better). Purple models are baselines; the dotted line shows the non-fine-tuned GPT-2 performance.

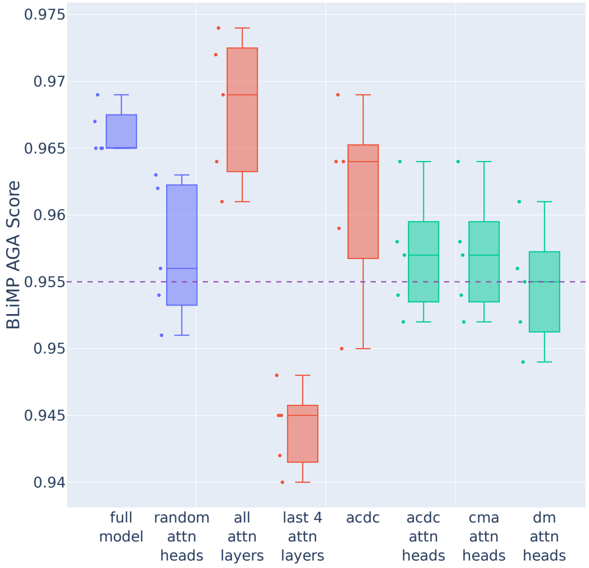

Figure 8: BLiMP Anaphor Gender Agreement results (higher is better). Purple models are baselines; the dotted line shows the non-fine-tuned GPT-2 performance.