Gender Encoding Patterns in Pretrained Language Model Representations

Mahdi Zakizadeh
TeIAS, Khatam University, Iran
m.zakizadeh@khatam.ac.ir

Mohammad Taher Pilehvar
Cardiff University, UK
pilehvarmt@cardiff.ac.uk

Abstract

Gender bias in pretrained language models (PLMs) poses significant social and ethical challenges. Despite growing awareness, there is a lack of comprehensive investigation into how different models internally represent and propagate such biases. This study adopts an information-theoretic approach to analyze how gender biases are encoded within various encoder-based architectures. We focus on three key aspects: identifying how models encode gender information and biases, examining the impact of bias mitigation techniques and fine-tuning on the encoded biases and their effectiveness, and exploring how model design differences influence the encoding of biases. Through rigorous and systematic investigation, our findings reveal a consistent pattern of gender encoding across diverse models. Surprisingly, debiasing techniques often exhibit limited efficacy, sometimes inadvertently increasing the encoded bias in internal representations while reducing bias in model output distributions. This highlights a disconnect between mitigating bias in output distributions and addressing its internal representations. This work provides valuable guidance for advancing bias mitigation strategies and fostering the development of more equitable language models. [1]

1 Introduction

Pretrained language models (PLMs) have revolutionized natural language processing (NLP) by enabling a wide range of applications (Devlin et al., 2019; Radford et al., 2019; Brown et al., 2020; Touvron et al., 2023; Jiang et al., 2023; Dubey et al., 2024). These models, trained on vast amounts of data, capture intricate patterns and knowledge, including gender-related information. However, alongside their impressive capabilities, PLMs also encode harmful biases that raise significant ethical concerns (Silva et al., 2021; Field et al., 2021; Ferrara, 2023). These biases can perpetuate stereotypes, misrepresent individuals and groups, and lead to unfair treatment in various applications, thereby impacting social justice and equity (e.g. Park et al., 2018; Kiritchenko and Mohammad, 2018; Chen et al., 2024; Levy et al., 2024).

Understanding how PLMs encode and propagate gender information is critical for developing effective bias mitigation strategies. This challenge grows increasingly urgent with the widespread adoption of retrieval-augmented generation (RAG) techniques, which rely on encoder-derived representations to retrieve contextually relevant documents (Wu et al., 2025). If gender biases are deeply embedded in these encoder-derived representations, RAG pipelines risk amplifying societal biases at an unprecedented scale by retrieving and propagating stereotypical or discriminatory content.

Despite extensive research on bias in language models, much of the focus has been on identifying and measuring bias rather than comprehensively analyzing how it is embedded within the model's internal representations. Previous studies have explored bias in transformer-based models, developing metrics to quantify bias (Islam et al., 2016; May et al., 2019; Nangia et al., 2020; Nadeem et al., 2021; Felkner et al., 2023), implementing techniques to reduce it (Zhao et al., 2018a; Lauscher et al., 2021; Kaneko and Bollegala, 2021; Webster et al., 2020; Schick et al., 2021), and investigating its underlying causes (Bolukbasi et al., 2016; Kaneko et al., 2022). However, there remains a limited understanding of the mechanisms through which biases are encoded and how different training and fine-tuning processes influence these biases within model weights.

To address this gap, we use an information-theoretic approach, specifically Minimum Description Length (MDL) probing proposed by Voita and Titov (2020), to explore how gender bias is encoded in various encoder-based architectures. By exam-

[1] The code utilized in this study is available at https://github.com/mzakizadeh/Gender-Encoding-Patterns

ining different layers of PLMs, we identify where biases emerge and how fine-tuning and debiasing techniques impact these representations.

Our work is inspired by Mendelson and Belinkov (2021) who studied the impact of debiasing techniques used to reduce the model's reliance on spurious correlations between data and labels in natural language inference on model's representations. In summary, our contributions are twofold:

- We pinpoint the specific parts of encoder-based PLMs responsible for encoding gender information, highlighting critical layers where bias is most pronounced.
- We assess the effect of various debiasing methods, demonstrating that pretrained debiasing objectives outperform post-hoc mitigation approaches in reducing encoded bias.

2 Related Works

In this section, we review some of the related studies on gender bias in language models, bias mitigation and measurement methods, and probing techniques and their use in bias evaluation.

2.1 Bias in Language Models

Early investigations into gender bias in language models unveiled that static embeddings not only encode but also amplify human-like biases within their representations (Islam et al., 2016; Bolukbasi et al., 2016). Subsequently, various studies have proposed methods to manipulate the embedding space or learning algorithms to mitigate bias in such models (Bolukbasi et al., 2016; Zhao et al., 2018b). However, as Gonen and Goldberg (2019) demonstrated, these techniques only provide superficial solutions, as biased information is not entirely removed from the model's embedding space.

The introduction of contextualized word embeddings, such as BERT (Devlin et al., 2019), posed new challenges, as manipulating representation space became more intricate compared to static embeddings. Contextualized language models have been shown to exhibit bias against demographic groups, including gender (Zhao et al., 2019; Silva et al., 2021).

Despite these advancements, a comprehensive comparative analysis between various bias mitigation methods remained lacking. This gap was addressed by Meade et al. (2022), who conducted an empirical investigation into the effectiveness of multiple debiasing techniques. Through their experimentation, they selected diverse debiasing approaches, continued pretraining models with these techniques, and demonstrated their efficacy using prominent bias mitigation metrics. Additionally, they assessed the impact of these techniques on downstream performance, measuring model performance on the General Language Understanding Evaluation (GLUE; Wang et al., 2019) test set. As the results indicated that the debiasing techniques did not significantly compromise downstream performance, they hypothesized that these methods might not negatively affect model representations. However, they did not provide concrete evidence to support their claims. This highlights the need for further research and analysis to thoroughly understand the implications and effectiveness of different debiasing techniques in the context of language models.

While earlier studies have explored the presence of gender bias in static and contextualized embeddings, they primarily focused on identifying and quantifying bias or testing basic mitigation strategies. Our study takes a different approach by investigating how biases are encoded within the internal representations of language models. This deeper exploration helps uncover where and how bias manifests, providing insights into mitigating these issues more effectively.

2.2 Probing Techniques and Bias Evaluation

Probing is a valuable technique for determining the knowledge characteristics captured by language models. With advancements in methods for interpreting model behavior, probing has gained traction in the research community. The introduction of Minimum Description Length probing (MDL probing; Voita and Titov, 2020), has enabled researchers to explore the knowledge encoded in language model representations in more depth. MDL probing has been utilized to assess biases in model representations, as demonstrated by Mendelson and Belinkov (2021) and Orgad et al. (2022).

Intriguingly, Mendelson and Belinkov (2021) found that debiasing methods intended to make models robust against spurious correlations in datasets, inadvertently led to an increase in biased information in model representations. On the other hand, Orgad et al. (2022) employed MDL as a metric for assessing bias and demonstrated its stronger correlation with extrinsic bias metrics used in conjunction with extrinsic bias mitigation techniques

compared to other intrinsic bias measurement methods.

Building on the advancements of probing techniques, particularly the use of structured methods to interpret model behaviors, our work delves into the mechanisms by which gender biases are encoded. By systematically evaluating model layers, we aim to understand how different mitigation and fine-tuning strategies influence the internal representations of bias, extending the applications of probing techniques to new depths.

2.3 Knowledge Localization and Bias

Knowledge localization has emerged as a critical area of study in NLP, focusing on identifying subnets within language models that are responsible for specific tasks, domains, or linguistic properties (Hendy et al., 2022; Panigrahi et al., 2023; Song et al., 2024; Choenni et al., 2023). These techniques have been extended to explore gender bias, pinpointing the internal components of models that encode bias.

For example, Chintam et al. (2023) employed causal inference methods, including techniques such as causal mediation analysis and differential masking, to identify attention heads responsible for biased behaviors in transformer models. Their work highlighted the ability to localize gender bias and proposed parameter-efficient fine-tuning strategies to mitigate it. Similarly, Lutz et al. (2024) introduced local contrastive editing, a technique leveraging unstructured pruning to precisely localize individual model weights responsible for encoding gender stereotypes. This method enabled them to edit these weights efficiently, mitigating bias without significant degradation of model performance.

Although our research aligns with prior efforts in localizing bias within pretrained language models, we introduce a distinct methodological perspective. Furthermore, by broadening the scope of experimentation across diverse models and mitigation strategies, we aim to comprehensively explore how and where gender bias is encoded. Our analysis reinforces previous findings about bias concentration in specific model layers, while also paving the way for targeted and efficient intervention techniques.

3 Background

Probing datasets are typically defined as $D = \{X, Y_p\}$, where X represents the input data, and Y_p represents the linguistic property or knowledge we are seeking to extract from the language model. The usage of language models involves two distinct stages. In the first stage, the language model, denoted as $f_\theta : X \rightarrow Z$, transforms the input X into a latent space Z, where X denotes the textual input, Z represents the latent representation of the text, and θ encompasses the model's weights. This latent space captures complex linguistic features and representations that encode the underlying information within the input text. Subsequently, in the second stage, a classifier, denoted as $g_\sigma : Z \rightarrow Y$, is employed to map the latent space Z to the corresponding label space Y. The classifier is denoted by g_σ, with σ encompassing its parameters. This two-stage approach facilitates the language model's ability to learn intricate language structures and encode relevant knowledge, while the classifier enables the extraction and utilization of this knowledge for various downstream tasks and analyses.

Traditionally, probing classifiers attempted to train on frozen language model weights, ensuring that the transformation from X to Z remains unchanged during training. Subsequently, the classifier learns how to map the latent space Z to the target property space Y_p. If the classifier can effortlessly learn this transformation with a limited amount of data, it was concluded that the language model possesses the relevant linguistic information (Belinkov, 2022). However, such traditional probing approaches have been shown to exhibit limitations. These methods can yield unreliable results as they tend to classify representations of random data similarly to those of actual data, indicating their inadequacy in capturing variations in representations (Zhang and Bowman, 2018). As a consequence, the outcomes of these traditional probing methods are highly dependent on hyperparameter choices and might not reliably reflect the true linguistic properties encoded within the language model representations. To address these issues and obtain more robust probing results, recent advancements have introduced innovative techniques, such as the Minimum Description Length (MDL) probing approach proposed by Voita and Titov (2020).

In MDL probing, the objective is not solely to assess the accuracy of the shallow classifier but also to measure the effort required to extract the targeted linguistic information from the model representations. Formally, they establish that a code exists to losslessly compress the labels using

Shannon-Huffman code such that $L_p(y_{1,z}|x_{1,z}) = -\sum_{i=1}^{z} log_2 p(y_i|x_i)$. Note that this is the cross-entropy loss. Furthermore, they define the uniform code length as $L^{\text{unif}}(y_{i,z}|x_{i,z}) = z log_2(C)$ where C is the number of classes in our task.

Given a model $P_\theta(y|x)$ with learnable parameters θ, they choose blocks $1 = n_0 < n_1 < ... < n_s = N$ and encode data by these blocks. The model starts by transmitting the data using the uniform code length for the first chunk. The model is then trained to predict labels y from the data x, and also used to predict the labels. The next block is transmitted using this trained new model. This process continues until the entire dataset is covered. Online code length is calculated as follows:

$$L^{\text{online}}\ (y_{1:z} \mid x_{1:z}) = z_1 \log_2 C - \sum_{i=1}^{S-1} \log_2 p_{\theta_i}\left(y_{n_i+1:n_{i+1}} \mid x_{n_i+1:n_{i+1}}\right) \quad (1)$$

Note that this encourages the model to perform well with smaller blocks, as if the model performs well in compressing the data in the block n_i, the compression will be increased for the subsequent block n_{i+1}.

Having calculated the code lengths, they compare the cross-entropy loss against the uniform code length to find the final compression. Formally, compression ($\mathcal{C}$) is defined as the ratio $\frac{L^{\text{online}}}{L^{\text{unif}}}$, quantifying how much the model compresses gender information relative to a uniform baseline.

4 Methodology

For this study, we focus on gender information as the knowledge property being probed. We will employ MDL probing to evaluate this phenomenon.

Models. Our experiments analyze the representations generated by a diverse range of models. We primarily focus on BERT (Devlin et al., 2019), ALBERT (Lan et al., 2020), and RoBERTa (Liu et al., 2019), which are widely used architectures in NLP, and we explore different variations and sizes of these models. Additionally, we examine with a newer model architecture called JINA Embeddings (Günther et al., 2023), which is popular in retrieval-augmented generation (RAG) pipelines. This model architecture offers a promising alternative due to the long context size and competitive performance, as claimed by the authors. By comparing these models, we aim to identify common patterns in how they encode gender information and assess their performance in mitigating biases.

Probing Dataset. We use the Bias in Bios dataset (De-Arteaga et al., 2019), which consists of 396,347 biographies. In this dataset, the gender of each individual is provided as a label alongside their occupation. This allows us to explore how gender information is encoded in language models when analyzing these biographies. In the Bias in Bios dataset, each data point is structured as a triplet $\{X, Y, Y_p\}$, where X represents a biography, Y denotes the true occupation label from one of 28 possible categories, and Y_p indicates the gender of the person featured in the biography.

Bias Definition and Implications. We formally define bias in terms of gender information encoding using the MDL probing framework. Let $f_\theta : X \to Z$ represent a language model with parameters θ that transforms input text X into latent representations Z. Let $f_{\theta_{\text{rand}}}$ be the same model architecture but with randomly initialized weights θ_{rand}. We denote the compression of gender information from these representations using online code length as $\mathcal{C}_\theta$ and $\mathcal{C}_{\theta_{\text{rand}}}$ respectively.

A model f_θ exhibits gender bias at layer l if the gender information can be extracted with significantly higher compression compared to a randomly initialized model with the same architecture:

$$\mathcal{C}_{\theta^l} - \mathcal{C}_{\theta^l_{\text{rand}}} > \delta \quad (2)$$

where θ^l and θ^l_{rand} represent the model parameters at layer l for the trained and randomly initialized models respectively, and $\delta > 0$ is a threshold determining the significance of the difference.

If a model encodes significant gender information, it could use this in decision-making, which is problematic for tasks like Bias in Bios, where we aim to predict occupations without relying on gender. This issue extends to retrieval tasks, such as systems finding resumes for job positions, where gender should not influence results. If retrieval models use gender information, they could reinforce biases that propagate through LLM workflows, leading to unfair outcomes and reinforcing stereotypes. Addressing this bias is essential for creating fairer and more ethical systems.

5 Gender Encoding Analysis

Building upon the framework outlined in the previous sections, we conducted our main experiment to

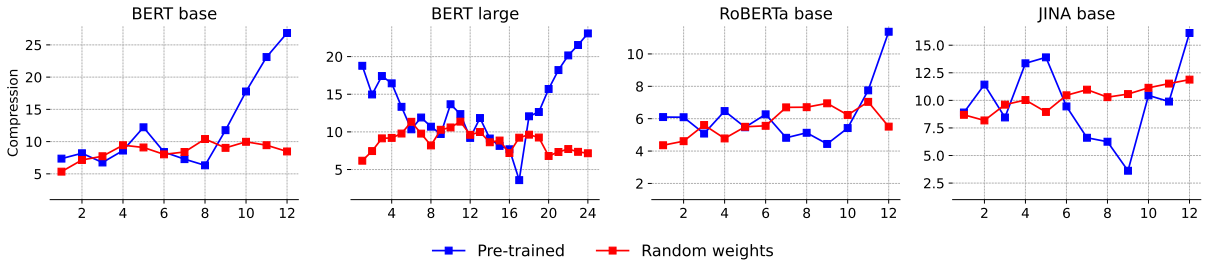

Figure 1: Gender information compression across different layers for various encoder models.

investigate whether there is a consistent pattern in how different encoder models encode gender information within their representations. Our primary goal was to determine if various models, despite architectural differences, exhibit similar behaviors in the way they handle gender-related information across their layers.

We experimented with a diverse range of encoder models to ensure the robustness of our findings. The main models discussed in this part are BERT-base, BERT-large, RoBERTa-base, and the base version of JINA Embeddings; however, we also saw similar results with ALBERT and a small version of JINA Embeddings. The results from these models are not included here due to space constraints.

Using the MDL probing method, we measured the amount of gender information that can be compressed from the representations at each layer of these models. Figure 1 illustrates the compression rates of gender information across different layers for the selected models. For each layer of the model, we also included a random baseline, which involves calculating compression for each layer of a model initialized with the same architecture but random weights. This baseline serves as a control to determine whether the observed compression is due to meaningful encoding of gender information or merely random noise.

Analyzing the results, we observed that models start with varying amounts of encoded gender information in their initial layers: while smaller models, like BERT base, do not exhibit gender information compression in their initial layers, larger models, such as BERT large, show high compression right from the first layer.

A consistent pattern emerges across all models. Initially, the models seem to reduce the gender information signal within their representations. This reduction continues up to a certain layer, typically close to the final layers. At this critical point, the compression rate of the random baseline representations becomes notably higher than that of the actual model's representations. Beyond this point, the models begin to reconstruct the gender information within their representations. By the final layer, all models demonstrate the highest amount of compression of gender information compared to any other layer. This indicates that, after initially suppressing the gender signal, the models ultimately encode it strongly in their final representations.

This pattern suggests a two-phase process in how encoder models handle gender information: (i) In the early layers, models may abstract away from specific attributes like gender, focusing instead on general linguistic features. (ii) In the later layers, models reintroduce and amplify gender-related information, potentially utilizing it for downstream tasks but also risking the propagation of bias. These insights underscore the pervasive nature of bias in language models and the need for targeted strategies to mitigate it, particularly in the layers where gender information is reintroduced.

6 Impact of Bias Mitigation

Bias mitigation in language models seeks to address both overt biases in model outputs and the subtler, systemic biases embedded within the model's internal representations. Effective techniques should suppress these encoded biases while maintaining model utility. In this section, we investigate the impact of various debiasing methods on compression values, used as a measure of encoded gender information, and evaluate their effectiveness across different experimental setups and models.

6.1 Experimental Settings

The experiments assess the performance of four debiasing methods applied to encoder-based language models, including BERT (base and large) and RoBERTa base. We begin by validating the correct implementation of the debiased variations of these models using a series of intrinsic benchmarks,

Model	Technique Name	CrowS-Pairs	StereoSet	DiFair (GNS)
BERT-base	Vanilla	58.02	62.02	63.91
	CDA	51.15 ↓6.87	72.98 ↑10.96	86.44 ↑22.53
	Dropout	57.25 ↓0.77	66.45 ↑4.43	68.59 ↑4.68
	Orthogonal Projection	53.44 ↓4.58	66.00 ↑3.98	60.46 ↓3.45
	ADELE	54.20 ↓3.82	64.76 ↑2.74	80.21 ↑16.30
RoBERTa-base	Vanilla	54.96	66.50	73.38
	CDA	51.15 ↓3.81	63.59 ↓2.91	82.58 ↑9.20
	Dropout	53.44 ↓1.52	69.26 ↑2.76	78.90 ↑5.52
	Orthogonal Projection	51.53 ↓3.43	69.19 ↑2.69	80.27 ↑6.89
	ADELE	49.62 ↓5.34	65.88 ↓0.62	70.67 ↓2.71
BERT-large	Vanilla	55.34	63.99	58.70
	Pretrained CDA	53.82 ↓1.52	70.59 ↑6.60	84.26 ↑25.56
	Pretrained Dropout	46.56 ↓8.78	54.95 ↓9.04	91.09 ↑32.39
	Post-Hoc CDA	56.87 ↑1.53	69.14 ↑5.15	84.56 ↑25.86
	Post-Hoc Dropout	57.63 ↑2.29	67.45 ↑3.46	64.03 ↑5.33

Table 1: Evaluation of debiasing on model weights for three benchmarks. "Metric Score" from CrowS-Pairs aims for 50; deviations suggest gender bias. "ICAT Score" and "Gender Neutrality Score" aim for 100 on StereoSet and DiFair, respectively.

as all debiasing techniques evaluated are intrinsic in nature. Specifically, we employ the CrowS-Pairs (Nangia et al., 2020), StereoSet (Nadeem et al., 2021), and DiFair (Zakizadeh et al., 2023) benchmarks. The evaluation results for these benchmarks are summarized in Table 1. The findings indicate that all debiased models demonstrate effectiveness, with at least two benchmarks showing improved fairness metrics compared to their vanilla counterpart.

Overview of Debiasing Techniques We employed four distinct debiasing strategies to assess the impact of debiasing on model representations. Counterfactual Data Augmentation (CDA; Zhao et al., 2018a) replaces gendered terms with neutral counterparts and retrains the model on the augmented data, effectively neutralizing biased associations. Adapter-Based Debiasing (ADELE; Lauscher et al., 2021) uses CDA-augmented data to train modular adapters that reduce bias without retraining the entire model. Dropout applies higher dropout rates during training, hypothesizing that enhanced regularization can reduce encoded biases (Webster et al., 2020). Finally, Orthogonal Projection (Kaneko and Bollegala, 2021) removes gender-related components from intermediate representations through linear projections, offering a lightweight post-hoc solution. Among the described bias mitigation techniques, ADELE and Orthogonal Projection are inherently post-hoc methods. Conversely, CDA and Dropout may be implemented at any stage, either during the post-hoc phase or from the onset of training.

Debiasing Effectiveness Based on our experiments in the previous section, gender-related information predominantly concentrates in the initial and final layers of the examined models. Given our formal definition of gender bias, we can precisely define the effectiveness of a debiasing method. Let $f_{\theta_{\text{debias}}}$ represent a model after applying a debiasing technique, with θ_{debias} denoting its parameters, and f_θ the original vanilla model with parameters θ. An ideal debiasing method is considered effective if it satisfies:

$$\mathcal{C}_{\theta^l_{\text{debias}}} \leq \min(\mathcal{C}_{\theta^l}, \mathcal{C}_{\theta^l_{\text{rand}}} + \delta) \tag{3}$$

where θ^l_{debias}, θ^l, and θ^l_{rand} represent the parameters at layer l for the debiased model, vanilla model, and randomly initialized model respectively, L denotes the total number of layers, and $\delta \geq 0$ is our bias significance threshold.

In simple terms, a debiasing method is effective if, across all layers, it reduces the compression of gender information below both the vanilla model and the threshold established by the random baseline. This indicates successful elimination of the gender signal from the representations throughout the entire model architecture. Conversely, if a method fails to satisfy this criterion at any layer, it indicates that the debiasing approach is ineffective or even counterproductive in terms of compression.

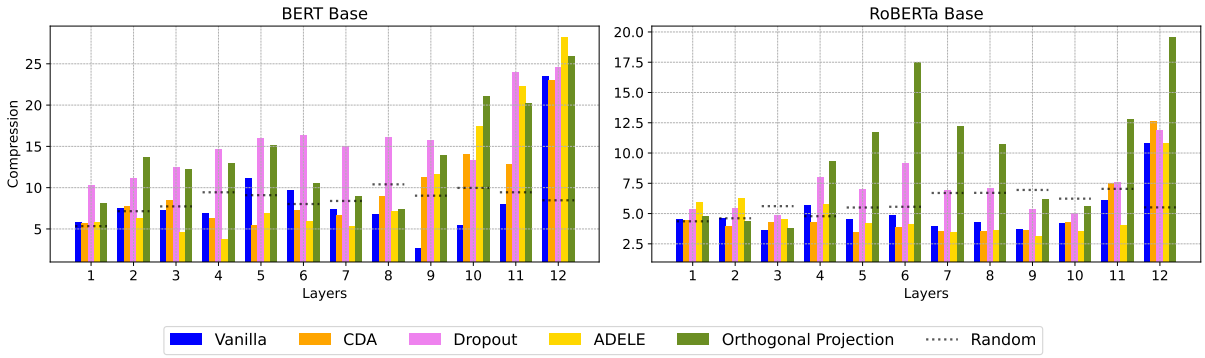

Figure 2: Effect of various bias mitigation procedures on gender information compression across different layers of base models.

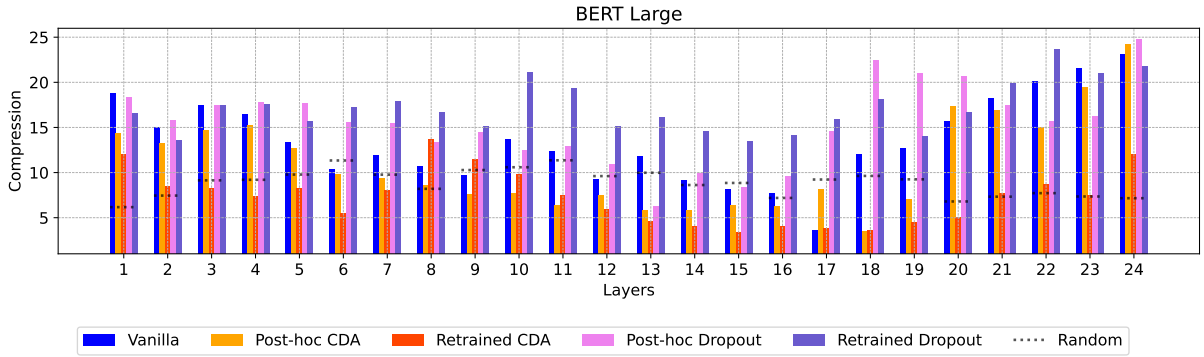

Figure 3: Effect of various bias mitigation procedures on gender information compression across different layers of BERT-Large.

6.2 Results and Analysis

The results of these experiments are presented in Figures 1 and 3. Our analysis reveals that, with the exception of training-time CDA, the remaining methods were ineffective in reducing bias in the models. Some methods, such as ADELE and training-time Dropout, show mixed results, suggesting that their effectiveness may be influenced by factors such as model architecture and training parameters. In the following discussion, we will elaborate on these observations in detail.

Layer-Wise Trends in Compression Compression values exhibited a consistent pattern across all models. In the lower layers, gender information was minimally compressible, suggesting that these layers encode relatively little bias. However, in the final layers, compression values increased sharply, indicating that gender information becomes more concentrated and accessible as representations become more abstract.

Impact of Training-Time Debiasing Training-time CDA on BERT-large demonstrated the most substantial reduction in final-layer compression. The compression value in the final layer decreased from 23.08 in the vanilla model to 11.98 after retraining with CDA, confirming its effectiveness in suppressing gender information throughout the model. Similarly, training-time dropout resulted in a lower final-layer compression compared to the vanilla model, though its effect was less pronounced than CDA.

Effectiveness of Post-Hoc Methods Post-hoc CDA and dropout, applied across all models, were generally less effective in mitigating gender encoding. In BERT-large, post-hoc CDA failed to achieve the same level of suppression as training-time CDA, resulting in a final-layer compression of 20.34. Dropout exhibited inconsistent behavior across models; in some cases, it preserved or even amplified gender information. For instance, in BERT-base, the final-layer compression increased from 23.47 (vanilla) to 24.63 with post-hoc dropout, indicating that this method does not reliably suppress bias.

Comparison Across Model Architectures RoBERTa-base consistently displayed lower

compression values than BERT-based models, suggesting that its architecture inherently encodes less gender-related information. This observation aligns with its performance on intrinsic bias benchmarks, where it demonstrated reduced sensitivity to gendered associations. Comparing BERT-base and BERT-large also indicates that larger models tend to store more gender information in their representations, which also aligns with the results obtained from the intrinsic bias benchmarks. This suggests that as model capacity increases, so does its ability to encode and retain gendered associations, reinforcing the need for targeted mitigation strategies in larger models.

While all debiasing methods contributed to reducing gender encoding to some extent, none completely eliminated it across all layers. Training-time CDA proved the most effective strategy, whereas post-hoc methods showed limited success, particularly in mitigating gender encoding in the final layers. These findings indicate that bias is deeply ingrained in model representations and that effective mitigation requires intervention during training rather than post-hoc adjustments.

For practical applications where reducing gender encoding is a priority, retraining with targeted debiasing objectives remains the most reliable approach. Future work could explore hybrid strategies that combine training-time and post-hoc techniques to enhance bias suppression without requiring full retraining.

7 Impact of Fine-tuning

While encoder models are widely used in retrieval systems, their representations are typically fine-tuned for downstream tasks such as classification. Understanding how this process influences gender bias encoded in model representations is critical, as fine-tuning may alter or amplify existing biases. In this section, we investigate how fine-tuning affects gender-related information stored in model layers and evaluate its implications for bias mitigation.

7.1 Experimental Settings

We fine-tuned three encoder models – BERT-base, BERT-large, and RoBERTa-base – on the BiosBias dataset. The task involves predicting an individual's occupation from their biography, framed as a 28-class classification problem. Models were trained for 5 epochs using a learning rate of 2×10^{-5}. To isolate the impact of fine-tuning, we compared the fine-tuned models against two baselines: (i) their original pretrained versions and (ii) "randomized" counterparts initialized with untrained weights but fine-tuned on the same task. Layer-wise MDL probing was applied to all models to measure gender information compression before and after fine-tuning.

7.2 Results and Analysis

The experimental results, presented in Figure 4, reveal several noteworthy patterns in how fine-tuning affects gender information encoding.

Reduced Gender Information Fine-tuning consistently led to a substantial reduction in gender information compression across all models. This reduction was particularly pronounced in the final layers, where the original models had shown the highest concentration of gender information.

Below-Random Compression In many cases, the compression values of fine-tuned models fell below those of their random baselines. Notably, even the random baselines of fine-tuned models showed lower compression compared to their pretrained counterparts. This suggests that task-specific fine-tuning may actively suppress the encoding of gender information in favor of task-relevant features.

Shift in Representational Focus The dramatic reduction in gender information compression indicates that fine-tuning redirects the model's internal representations toward task-specific features and away from demographic attributes like gender. This finding suggests that much of the bias observed in fine-tuned models may originate from the classification head rather than from biases encoded in the underlying representations.

These findings carry significant implications for bias mitigation in language models. The observation that fine-tuning naturally reduces encoded gender information while potentially concentrating bias in the classification layer explains the limited impact of intrinsic debiasing methods on extrinsic bias metrics (Orgad et al., 2022; Cao et al., 2022). While task-specific fine-tuning may serve as an implicit form of representation-level bias mitigation, our results suggest that future debiasing efforts should focus more on the classification components added during fine-tuning rather than the encoder representations alone.

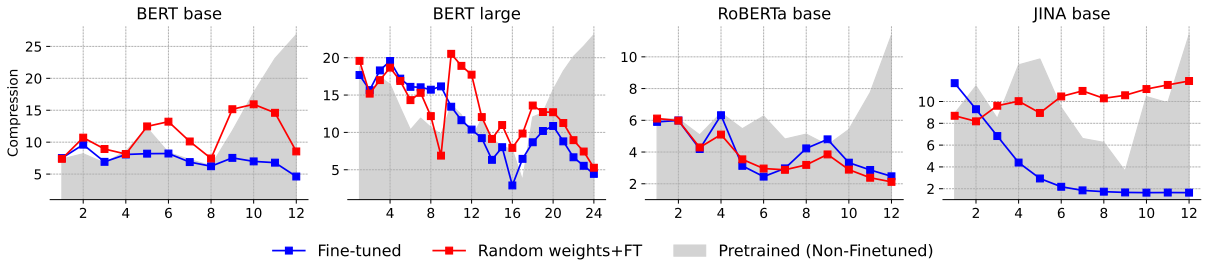

Figure 4: Gender information compression across different layers for the fine-tuned encoder models. The pretrained compression values correspond to the blue line shown in Figure 1.

8 Conclusions

Our analysis reveals that pretrained language models follow a consistent pattern of gender encoding: early layers suppress gender signals, while later layers amplify them, embedding bias deeply into abstract representations. Current debiasing techniques, particularly post-hoc interventions, show limited efficacy in altering these internal mechanisms. Task-specific fine-tuning reduces encoded gender information but risks concentrating residual bias in downstream classifiers, underscoring the need for holistic mitigation strategies that target both representations and decision layers. Collectively, these findings challenge conventional debiasing paradigms, advocating for proactive integration of fairness objectives during pretraining and architecture-aware interventions targeting bias propagation pathways.

Broader Impacts

Our results have significant implications for the design and deployment of language models. First, they underscore the inadequacy of post-hoc debiasing methods, urging researchers to integrate fairness objectives directly into pretraining. Second, the localization of bias in later layers suggests targeted interventions, such as modifying specific layers or attention heads, could offer efficient mitigation pathways. Finally, practitioners must recognize that reducing bias in representations does not guarantee fairness in downstream applications; rigorous evaluation of classifiers and datasets remains essential. These insights advocate for a paradigm shift toward inherently fair model architectures and training frameworks.

Limitations

While this work provides critical insights, several limitations warrant consideration. First, our analysis focuses on gender bias in English-language biographies, leaving broader sociocultural and intersectional biases unexplored. Second, the study centers on encoder-based models; future work should validate findings in decoder-based architectures and multimodal systems. Lastly, the interplay between task-specific fine-tuning and bias propagation requires deeper exploration across diverse applications. Addressing these gaps will advance our understanding of bias dynamics and mitigation in increasingly complex language technologies.

References

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4349–4357.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Yang Trista Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570,

Dublin, Ireland. Association for Computational Linguistics.

Yuen Chen, Vethavikashini Chithrra Raghuram, Justus Mattern, Rada Mihalcea, and Zhijing Jin. 2024. Causally testing gender bias in LLMs: A case study on occupational bias. In *Causality and Large Models @NeurIPS 2024*.

Abhijith Chintam, Rahel Beloch, Willem Zuidema, Michael Hanna, and Oskar van der Wal. 2023. Identifying and adapting transformer-components responsible for gender bias in an English language model. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 379–394, Singapore. Association for Computational Linguistics.

Rochelle Choenni, Dan Garrette, and Ekaterina Shutova. 2023. Cross-lingual transfer with language-specific subnetworks for low-resource dependency parsing. *Computational Linguistics*, pages 613–641.

Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 120–128. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, et al. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Virginia Felkner, Ho-Chun Herbert Chang, Eugene Jang, and Jonathan May. 2023. WinoQueer: A community-in-the-loop benchmark for anti-LGBTQ+ bias in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9126–9140, Toronto, Canada. Association for Computational Linguistics.

Emilio Ferrara. 2023. Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. *CoRR*, abs/2304.07683.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. 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 1905–1925, Online. Association for Computational Linguistics.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. 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 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, Maximilian Werk, Nan Wang, and Han Xiao. 2023. Jina embeddings 2: 8192-token general-purpose text embeddings for long documents. *CoRR*, abs/2310.19923.

Amr Hendy, Mohamed Abdelghaffar, Mohamed Afify, and Ahmed Y. Tawfik. 2022. Domain specific subnetwork for multi-domain neural machine translation. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 351–356, Online only. Association for Computational Linguistics.

Aylin Caliskan Islam, Joanna J. Bryson, and Arvind Narayanan. 2016. Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, abs/1608.07187.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Masahiro Kaneko and Danushka Bollegala. 2021. Debiasing pre-trained contextualised embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1256–1266, Online. Association for Computational Linguistics.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2740–2750, Seattle, United States. Association for Computational Linguistics.

Svetlana Kiritchenko and Saif Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

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.

Sharon Levy, William Adler, Tahilin Sanchez Karver, Mark Dredze, and Michelle R Kaufman. 2024. Gender bias in decision-making with large language models: A study of relationship conflicts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5777–5800, Miami, Florida, USA. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Marlene Lutz, Rochelle Choenni, Markus Strohmaier, and Anne Lauscher. 2024. Local contrastive editing of gender stereotypes. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21474–21493, Miami, Florida, USA. Association for Computational Linguistics.

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)*, pages 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.

Michael Mendelson and Yonatan Belinkov. 2021. Debiasing methods in natural language understanding make bias more accessible. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1545–1557, Online and Punta Cana, Dominican Republic. 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)*, pages 5356–5371, Online. Association for Computational Linguistics.

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.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. Association for Computational Linguistics.

Abhishek Panigrahi, Nikunj Saunshi, Haoyu Zhao, and Sanjeev Arora. 2023. Task-specific skill localization in fine-tuned language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 27011–27033. PMLR.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

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

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.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Ran Song, Shizhu He, Shuting Jiang, Yantuan Xian, Shengxiang Gao, Kang Liu, and Zhengtao Yu. 2024.

Does large language model contain task-specific neurons? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7101–7113, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed H. Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. Technical report.

Xuyang Wu, Shuowei Li, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2025. Does RAG introduce unfairness in LLMs? evaluating fairness in retrieval-augmented generation systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10021–10036, Abu Dhabi, UAE. Association for Computational Linguistics.

Mahdi Zakizadeh, Kaveh Miandoab, and Mohammad Pilehvar. 2023. DiFair: A benchmark for disentangled assessment of gender knowledge and bias. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1897–1914, Singapore. Association for Computational Linguistics.

Kelly Zhang and Samuel Bowman. 2018. Language modeling teaches you more than translation does: Lessons learned through auxiliary syntactic task analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 359–361, Brussels, Belgium. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. 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 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. 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.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.