Artificial Intelligence 348 (2025) 104401

Contents lists available at ScienceDirect

Artificial Intelligence

journal homepage: www.elsevier.com/locate/artint

BATED: Learning fair representation for Pre-trained Language Models via biased teacher-guided disentanglement

Yingji Li [a], Mengnan Du [b], Rui Song [a], Mu Liu [a], Ying Wang [a,c,*]

[a] College of Computer Science and Technology, Jilin University, Changchun, 130012, China
[b] Department of Data Science, New Jersey Institute of Technology, Newark, USA
[c] Key Laboratory of Symbol Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun, 130012, China

ARTICLE INFO

Keywords:
Pre-trained Language Models
Fairness
Social bias
Feature disentanglement
Causal contrastive learning

ABSTRACT

With the rapid development of Pre-trained Language Models (PLMs) and their widespread deployment in various real-world applications, social biases of PLMs have attracted increasing attention, especially the fairness of downstream tasks, which potentially affects the development and stability of society. Among existing debiasing methods, intrinsic debiasing methods are not necessarily effective when applied to downstream tasks, and the downstream fine-tuning process may introduce new biases or catastrophic forgetting. Most extrinsic debiasing methods rely on sensitive attribute words as prior knowledge to supervise debiasing training. However, it is difficult to collect sensitive attribute information of real data due to privacy and regulation. Moreover, limited sensitive attribute words may lead to inadequate debiasing training. To this end, this paper proposes a debiasing method to learn fair representation for PLMs via **BiA**sed **TE**acher-guided **D**isentanglement (called **BATED**). Specific to downstream tasks, BATED performs debiasing training under the guidance of a biased teacher model rather than relying on sensitive attribute information of the training data. First, we leverage causal contrastive learning to train a task-agnostic general biased teacher model. We then employ Variational Auto-Encoder (VAE) to disentangle the PLM-encoded representation into the fair representation and the biased representation. The Biased representation is further decoupled via biased teacher-guided disentanglement, while the fair representation learn downstream tasks. Therefore, BATED guarantees the performance of downstream tasks while improving the fairness. Experimental results on seven PLMs testing three downstream tasks demonstrate that BATED outperforms the state-of-the-art overall in terms of fairness and performance on downstream tasks.

1. Introduction

Pre-trained Language Models (PLMs), such as BERT [1], GPT-2 [2], and Large Language Models (LLMs) like LLaMA-3 [3,4] and GPT-4 [5], have been widely adopted across various Natural Language Processing (NLP) tasks due to their powerful language modeling capabilities. However, PLMs have been shown to suffer from social biases that carry over from the representation into decisions on downstream tasks [6,7], and may even compromise real-world NLP systems. For example, the automatic resume filtering system [8,9]

* Corresponding author at: College of Computer Science and Technology, Jilin University, Changchun, 130012, China.
E-mail addresses: yingji21@mails.jlu.edu.cn (Y. Li), mengnan.du@njit.edu (M. Du), songrui@jlu.edu.cn (R. Song), liumu23@mails.jlu.edu.cn (M. Liu), wangying2010@jlu.edu.cn (Y. Wang).

https://doi.org/10.1016/j.artint.2025.104401
Received 25 June 2024; Received in revised form 29 April 2025; Accepted 5 August 2025

Available online 8 August 2025

may have the gender bias, which is more inclined to assign "*male*" applicants to "*doctors*" and "*female*" applicants to "*nurses*" for the same resumes. The US healthcare system [10,11] is found to be racially biased in that for "*white*" patients and "*black*" patients with the same level of risk, it calculates a higher prevalence for "*black*" patients. The application of PLMs with social biases will further aggravate the adverse social situation of vulnerable groups who are discriminated against and marginalized, with potential social harm and unpredictable impact. Therefore, a growing number of researchers have focused on mitigating the social biases of PLMs to improve the fairness of decisions in downstream tasks.

The social biases of PLMs can be roughly divided into two main categories: intrinsic bias and extrinsic bias [12]. Intrinsic bias, also known as upstream bias or representation bias, refers to harmfulness in the representation encoded by PLMs. Extrinsic bias, also known as downstream bias or decision bias, quantifies the fairness of PLMs predictions on downstream tasks. In recent years, there are proposed many debiasing methods to mitigate the social biases of PLMs. Corresponding to the type of biases, debiasing methods can be divided into intrinsic debiasing methods and extrinsic debiasing methods [12]. Intrinsic debiasing methods are task-agnostic and performed before downstream tasks, aiming to improve fairness of PLM-encoded representations. Extrinsic debiasing methods are task-specific and performed during fine-tuning of downstream tasks, aiming to improve the fairness of decisions made by PLMs.

Intrinsic debiasing methods mitigate the intrinsic bias by debiasing the representation of the PLMs' encoder output. A representative approach is Counterfactual Data Augmentation (CDA) [13,14], which uses sensitive attribute words pairs (e.g., "*male*" and "*female*" specific to gender groups, "*white*" and "*black*" specific to race groups) to match and replace the original samples and then retrain PLMs with augmented sample pairs. FairFil [15], MABEL [16], and CCPA [17] combine CDA and contrastive learning [18] to debias by approximating the representation between augmented sample pairs. In addition, methods such as Auto-Debias [19] searching for biased prompts, BNS [20] masking biased neurons, and DeepSoftDebias [21] utilizing residual neural networks all target intrinsic bias. Although intrinsic debiasing methods are universal because they do not restrict downstream tasks, there are some drawbacks when applied to downstream tasks. On the one hand, it is not certain that there is a necessary relationship between intrinsic and extrinsic bias [22,23]. On the other hand, fine-tuning on downstream tasks may induce catastrophic forgetting [24] or introduce additional new biases [25]. Therefore, the debiasing performance of PLMs trained with intrinsic debiasing methods will be weakened in downstream tasks.

Extrinsic debiasing designs mitigation strategies for specific downstream tasks, which are usually based on adversarial training [26, 27], orthogonal projection [28], and regularization constraints [29]. Most of them take sensitive attribute words as prior knowledge to supervise the debiasing training. However, due to the limitations of privacy and regulation, it may be difficult to collect sensitive attribute information of real data. In addition, the hand-crafted sensitive attribute words are limited and difficult to cover all the training data, resulting in inadequate debiasing training.

Motivated by the above challenges, this paper proposes a debiasing method to learn fair representation for PLMs via **Bi**ased **TE**acher-guided **D**isentanglement (called **BATED**). Specific to downstream tasks, BATED performs debiasing training under the guidance of a biased teacher model rather than relying on sensitive attribute information of the training data. Specifically, we first train a task-agnostic general biased teacher model using causal contrastive learning [30]. In the debiasing training phase, we use a Variational Auto-Encoder (VAE) [31,32] to disentangle the PLM-encoded representation into a fair representation and a biased representation. The biased teacher model is then used to encourage the distillation of the biased representation. Downstream task loss is used to constrain the fair representation to learn task-related information. Extensive experiments show that BATED improves decision fairness while maintaining the performance of PLMs in downstream tasks. Our contributions are summarized as follows.

- We propose a debiasing method that uses a biased teacher model to guide the VAE for feature disentanglement. It can mitigate the social biases of PLMs in the debiasing training process without relying on the sensitive attribute information of the data, which makes the debiasing method not limited by data privacy and expands the scope of application.
- We propose a method to train a task-agnostic general biased teacher model using causal contrastive learning, and design a feature distillation loss that utilizes a biased teacher model to guide biased feature disentanglement.
- Extensive debiasing experiments are conducted on seven PLMs: BERT, DistilBERT, ELECTRA, OPT, GPT-2, Qwen-2.5, and LLaMA-3.2 on three downstream tasks. The experimental results demonstrate that our proposed debiasing method BATED generally outperforms the state-of-the-art overall in terms of fairness and performance on downstream tasks.

2. Related work

In research related to PLMs, the concept of bias spans multiple domains, such as model bias under out-of-distribution (OOD) generalization scenarios [33] and other forms of inductive or task-specific biases. In this paper, we focus specifically on social biases related to model fairness, which concern the equitable treatment of different demographic or identity groups in the predictions of PLMs. In this section, we introduce recent debiasing methods to mitigate the social biases of PLMs from two perspectives: intrinsic debiasing methods and extrinsic debiasing methods.

2.1. Intrinsic debiasing

Intrinsic debiasing methods correspond to intrinsic biases, which are task-agnostic and performed in upstream tasks before downstream tasks. They mitigate the representational harm by applying debiasing strategies to the representation output by the PLMs' encoder. Intrinsic debiasing methods often employ unsupervised learning to train models. For example, Sent-Debias [34] projects the sentence representation of PLMs into the bias space and then removes the bias by orthogonality. DeepSoftDebias [21] enhances

soft debiasing by replacing the transformation matrix with a residual neural network, allowing for more expressive modeling of input-output mappings. FairFil [15], MABEL [16], and CCPA [17] mitigate the social biases in PLM's representation by employing unsupervised contrastive learning to approximate the representation between the augmented pairs. There are also strategies that employ unsupervised learning to retrain PLMs, CDA [14,35,36] augments the training corpus, Dropout [13] adds a dropout term to the architecture, and BNS [20] to find and mask biased neurons using a modified integrated gradient interpretability algorithm. All these strategies are essentially retraining the model parameters. In addition, other methods are also included Auto-Debias [19] and CD^3 [37], they propose strategies that employ discrete prompts to amplify the biases of the original samples before debiasing training.

These debiasing methods are general for downstream tasks, but there are some drawbacks. The first point is that intrinsic debiasing methods require additional training phases and training data before downstream tasks. Second of all, some work has found that there is no necessary link between intrinsic and extrinsic bias, so mitigating intrinsic bias does not mean mitigating extrinsic bias [22,23]. Moreover, when models trained with intrinsic debiasing are applied to downstream tasks, the fine-tuning process may suffer from catastrophic forgetting [24] or introduce new biases, thereby weakening debiasing performance [25]. In this paper, we focus on extrinsic debiasing methods, which directly debiasing downstream tasks to avoid the negative impact of the fine-tuning process on intrinsic debiasing methods. Furthermore, to obtain general bias information, we train a biased teacher model with intrinsic bias to guide the disentanglement of biased representation under an unsupervised contrastive learning framework.

2.2. Extrinsic debiasing

Extrinsic debiasing methods correspond to extrinsic biases, which are task-specific and debiasing during fine-tuning on downstream tasks. They maintain the fairness of prediction results by synchronously debiasing as the model learns task information. Debiasing with adversarial learning is an intuitive strategy to mitigate bias by leveraging an encoder to prevent the discriminator from recognizing the protected attributes [26,38]. R-LACE [27] adopts a debiasing strategy based on adversarial learning, which formulates the problem as a minimax game and adopts convex relaxation to solve. CAGD [39] collaborates contrastive learning and adversarial training to achieve gender-neutral representations of the GPT family of models. Some work employs regularization constraints in downstream tasks [40]. As an example, Gender-tuning [29] perturbs the gender words in the original samples and uses them as the training objective of the Masked Language Model, which is then integrated into the fine-tuning process of the downstream tasks. And INLP [28] proposes to utilize a repeatedly trained linear classifier to predict the target space and then remove the bias by projecting the representation into the null space of the target space. In addition, supervised contrastive learning can also be used for downstream debiasing [41,42]. Recently many debiasing techniques have been proposed specifically for LLMs, which improve the fairness of responses through techniques such as direct preference optimization [43], multi-task learning [44], and multi-agent frameworks [45].

Most of extrinsic debiasing methods take sensitive attribute words as prior knowledge to supervise the debiasing training. However, due to the limitations of privacy and regulation, it may be difficult to collect sensitive attribute information of real data. Even if the sensitive information is not restricted, labeling the sensitive attribute of the training data still requires a lot of labor cost. Therefore, many debiasing methods cannot be applied when the sensitive attribute information in the task dataset is not available. In addition, the hand-crafted sensitive attribute words are limited and difficult to cover all the training data, resulting in inadequate debiasing. This paper proposes a debiasing method that performs debiasing training without providing sensitive attribute information for downstream tasks. Our method utilizes the biased teacher model as a supervisor for debiasing training, thus eliminating the need for sensitive information annotations in the task dataset.

3. Preliminaries

Specific to demographic groups, a *social sensitive topic* $\mathcal{T} = \{T^1, T^2, \cdots, T^N\}$ contains N bias directions, each of which corresponds to a *social subgroup*, and each social subgroup can be represented by a set of *sensitive attribute words* $(t_1, t_2, \cdots, t_m)$. In this paper, we consider the binary gender group $Gender = \{Male, Famale\}$,[1] for which examples of sensitive attribute words are (*"he","his","man","father","boy"*) and (*"she","her","woman","mother","girl"*).

Given a PLM M, its encoder is denoted by $E(\cdot)$ and its classification head on a certain downstream task D is denoted by $C(\cdot)$. For a sample x in the task dataset $\mathcal{X}$, M encodes x to obtain the representation $\mathbf{h} = E(x)$, and the classification head C further outputs the prediction $\hat{y} = C(\mathbf{h})$.

Our objective is to train a VAE that disentangle the representation $\mathbf{h}$ output by the encoder $E(\cdot)$ of the PLM into a fair representation $\mathbf{z}$ and a biased representation $\mathbf{b}$, and then feed the fair representation $\mathbf{z}$ into the task classification head C. Finally, the fair prediction $\hat{y}$ is obtained, that is, the task prediction results are not distinguished by the male subgroup samples and the female subgroup samples.

4. Methodology

In this section, we propose the debiasing method BATED to learn the fair representation for PLMs in downstream tasks via biased teacher-guided disentanglement. BATED consists of two stages: 1) training the biased teacher model; and 2) debiasing training via

[1] In view of the wide range of debiasing researches and benchmark datasets, this paper chooses gender groups as the study case. It's important to note that gender in the real world may not be binary.

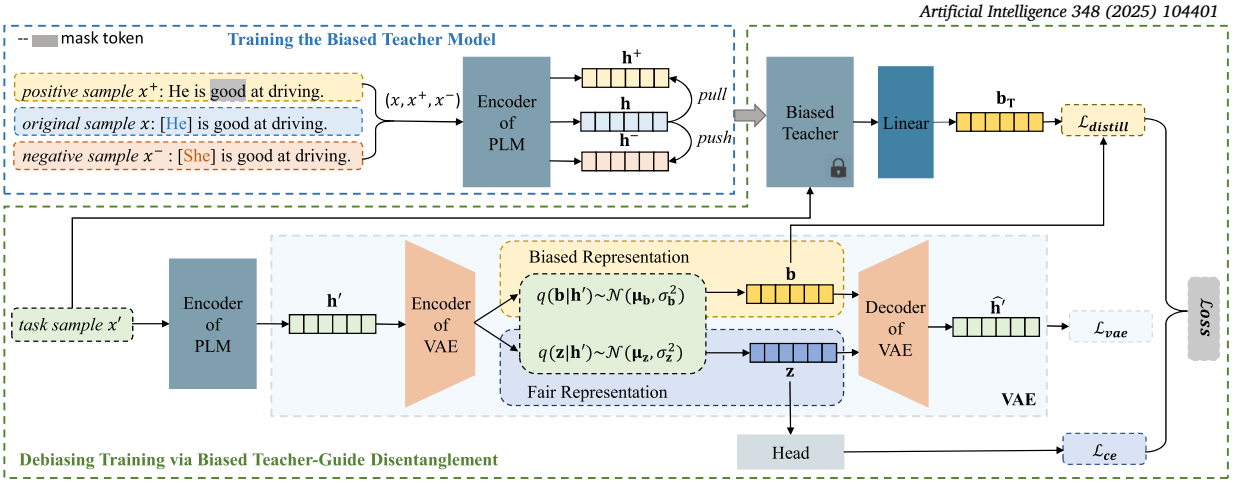

Fig. 1. The Framework of BATED. In the stage of training biased teacher model, we utilize causal contrastive learning to train the encoder of PLM to obtain a task-agnostic general biased teacher model. In the stage of debiasing training via biased teacher-guided disentanglement, we leverage VAE to disentangle the PLM-encoded representation $\mathbf{h}'$ into a fair representation $\mathbf{z}$ and a biased representation $\mathbf{b}$. The fair representation $\mathbf{z}$ performs downstream tasks, while the biased representation $\mathbf{b}$ is encouraged to fit the biased teacher model. In this process, the parameters of the biased teacher model are fixed and a linear layer is employed to align the dimensions of the representation $\mathbf{b}_T$ output by the biased teacher model to the biased representation $\mathbf{b}$.

biased teacher-guided disentanglement. The framework of BATED is shown in Fig. 1. In the first stage, we propose to exploit causal contrastive learning to train the PLM's encoder as a biased teacher model, amplifying the intrinsic gender bias of the PLM. This process is independent of downstream tasks and involves training a model on unlabeled data to learn general biases. In the second stage, we employ a VAE consisting of an encoder and a decoder to disentangle the PLM-encoded representation into a fair representation and a biased representation. The fair representation is supervised by the downstream task, and the biased representation is supervised by the biased teacher model. In this process, the parameters of the biased teacher model are fixed and the parameters of PLM and VAE are updated.

4.1. Training the biased teacher model

It is not always easy to obtain sensitive attribute information in the dataset of downstream tasks, especially in some privacy-preserving fields, and labeling the sensitive information of the sample requires a lot of resources. In addition, the limited hand-crafted sensitive attribute words are difficult to match to all the training data, resulting in inadequate debiasing training. Therefore, we propose a debiasing method that can address the challenge of unavailability of sensitive information of task datasets, which uses a biased teacher model to guide disentanglement without relying on sensitive attributes in the extrinsic debiasing training process. To this end, we propose to train a task-agnostic general biased teacher model, which serves as a supervision signal in debiasing training to replace the sensitive attribute labels in the task dataset.

In this stage, we aim to train the encoder of PLM under an unsupervised learning framework to learn the generic gender bias from the training data, amplifying the intrinsic bias of PLM. Inspired by C2L [30], we design a strategy for training a biased teacher model using causal contrastive learning. Specifically, we utilize CDA to construct the causal triplet [30], where the original and augmented samples are treated as a pair of negative samples, while the original and perturbed samples are treated as a pair of positive samples. We use causal triples for contrastive learning [46,47] to learn the biased representation by pushing the representation distance between negative sample pairs and narrowing the representation distance between positive sample pairs.

4.1.1. Construct the causal triplet

First, we preprocess the training corpus with predefined sensitive attribute words to obtain gender-related training data. Specific to gender demographic groups, the social sensitive topic $Gender = \{Male, Famale\}$ has two bias directions. Given a set of sensitive attribute word pairs $\{(t_1, \hat{t}_1), (t_2, \hat{t}_2), \cdots, (t_m, \hat{t}_m)\}$ corresponding to each bias direction and an original training corpus $\mathcal{X}$, we use each sensitive attribute word t to match each sample in the training corpus $\mathcal{X}$, and then keep the samples that contain sensitive attribute words and delete the samples that do not. The filtered corpus is used as a training dataset $\mathcal{X}_b$ for training the biased teacher model, where each sample contains either sensitive attribute words for male subgroup or sensitive attribute words for female subgroup. This process essentially filters the training data to emphasize demographic-sensitive contexts, thereby enabling the model to focus on learning the representational difference across demographic subgroups.

We then construct causal triples by applying CDA to the training dataset $\mathcal{X}_b$, each of which contains a pair of positive and a pair of negative samples. Specifically, for each original sample $x \in \mathcal{X}_b$ containing the sensitive attribute word t, we perform CDA on x to obtain a negative sample x^- by replacing t with the opposite word t'. The original sample x and the negative sample x^- represent two different gender subgroups, respectively. To build a positive sample x^+, we randomly mask each token in x with probability ϵ. As a result, we construct the causal triple (x, x^+, x^-) of the original sample x.

4.1.2. Causal contrastive learning

Given the encoder $E_b(\cdot)$ of the PLM M_b, we input the causal triple (x, x^+, x^-) and output the encoded representation triple $(\mathbf{h}, \mathbf{h}^+, \mathbf{h}^-)$, where $\mathbf{h} = E(x)$, $\mathbf{h}^+ = E(x^+)$, and $\mathbf{h}^- = E(x^-)$. We use the contrastive learning framework for biased training, where the training objective is to narrow the representation distance between positive samples and push the representation distance between negative samples. Following C2L [30], the contrastive loss adopts the margin-based ranking loss [48], defined as follows:

$$\mathcal{L}_{bias} = max(0, \triangle m + s(\mathbf{h}, \mathbf{h}^+) - s(\mathbf{h}, \mathbf{h}^-)), \tag{1}$$

where $\triangle m$ is a hyper-parameter representing the margin value and $s(\cdot, \cdot)$ is a distance function between representation which we set as cosine distance. Under the causal contrastive learning framework, negative sample pairs encourage PLM to learn representation that are inconsistent for different gender groups, while positive sample pairs constrain PLM to learn the original semantics. Thus, the biased teacher model tries to retain language modeling ability while learning gender bias. The model captures general gender bias information from the training data itself during unsupervised learning to amplify its intrinsic bias. This phase is independent of the downstream tasks, and once trained, the trained biased teacher model can be used to guide arbitrary downstream tasks.

4.2. Debiasing training via biased teacher-guided disentanglement

In the debiasing training phase, we aim to improve the decision fairness of PLMs in specific downstream tasks. Inspired by some work on textual feature disentanglement [49,50], we exploit a VAE [31,32] to disentangle the representation encoded by PLM into a fair representation and a biased representation. The fair representation is supervised by labels of downstream tasks to learn task-relevant semantic information. In the case that sensitive attribute information in the task dataset is not available, we utilize the biased teacher model trained in the first stage as a supervisor to guide the further decoupling of the biased representation. In this process, we fix the parameters of the biased teacher model and adopt a linear layer to transform the dimensions of the representation output by the teacher model into the dimensions of the biased representation. We update the parameters of the encoder of PLM, VAE, classification head of the task, and the linear layer.

4.2.1. Feature disentanglement via VAE

We introduce two continuous latent variables to capture task information and bias information respectively. To separate task information and bias information from each other, we propose to utilize VAE to disentangle the original representation encoded by PLM into a fair representation containing task information and a biased representation containing bias information. Because VAE uses variational inference principles, it attempts to approximate the true posterior distribution of latent variables by minimizing the Kullback–Leibler divergence between the learned and prior distributions. This allows us to decompose the original representation into interpretable subspaces: task-relevant and bias-related representations.

Specifically, given a dataset $\mathcal{X}'$ of the downstream task and a PLM M_d to debias, we encode the sample $x' \in \mathcal{X}'$ with the encoder $E_d(\cdot)$ to obtain the representation $\mathbf{h}' = E_d(x')$. We employ a VAE to disentangle the representation $\mathbf{h}'$, the VAE's encoder $q(\cdot)$ that splits into two heads encoding $\mathbf{h}'$ into two latent representations: the fair representation $\mathbf{z}$ and the biased representation $\mathbf{b}$, the decoder $p(\cdot)$ decodes the two latent representations into the reconstructed representation $\widehat{\mathbf{h}'}$. According to the independent multivariate Gaussian assumption of VAE [32], each dimension of a multi-dimensional Gaussian distribution is conditionally independent, and each dimension obeys a one-dimensional Gaussian distribution and does not affect each other. Thus, the probability of the input representation $\mathbf{h}'$ can be computed as:

$$\begin{aligned} p(\mathbf{h}') &= \int p(\mathbf{z}, \mathbf{b}) p(\mathbf{h}' | \mathbf{z}, \mathbf{b}) \, \mathrm{d}\mathbf{z} \, \mathrm{d}\mathbf{b} \\ &= \int p(\mathbf{z}) p(\mathbf{b}) p(\mathbf{h}' | \mathbf{z}, \mathbf{b}) \, \mathrm{d}\mathbf{z} \, \mathrm{d}\mathbf{b}, \end{aligned} \tag{2}$$

where $p(\mathbf{z})$ and $p(\mathbf{b})$ are priors and are independent multivariate Gaussian $\mathcal{N}(\mathbf{0}, \mathrm{I})$, and $p(\mathbf{h}' | \mathbf{z}, \mathbf{b})$ is given by the decoder. Under the conditional independence assumption [51], the training loss function of the VAE is defined to maximize the evidence lower bound (ELBO):

$$\begin{aligned} \log p(\mathbf{h}') &\geq \mathrm{ELBO} \\ &= \mathbb{E}_{q(\mathbf{z}|\mathbf{h}')q(\mathbf{b}|\mathbf{h}')}[\log p(\mathbf{h}' | \mathbf{z}, \mathbf{b})] - \mathrm{KL}(q(\mathbf{z}|\mathbf{h}') || p(\mathbf{z})) - \mathrm{KL}(q(\mathbf{b}|\mathbf{h}') || p(\mathbf{b})) \\ &= \mathcal{L}_{vae}, \end{aligned} \tag{3}$$

where $\mathrm{KL}(\cdot)$ is the Kullback-Leibler divergence function, $q(\mathbf{z}|\mathbf{h}')$ and $q(\mathbf{b}|\mathbf{h}')$ are the posterior of two latent representations which are assumed to be independent and taking the form $\mathcal{N}(\mathbf{u}_\mathbf{z}, \sigma_\mathbf{z}^2)$ and $\mathcal{N}(\mathbf{u}_\mathbf{b}, \sigma_\mathbf{b}^2)$, respectively. Here, $\mathbf{u}_\mathbf{z}$, $\sigma_\mathbf{z}$, $\mathbf{u}_\mathbf{b}$, $\sigma_\mathbf{b}$ are encoded by the encoder $q(\cdot)$. Then $\mathbf{z}$ and $\mathbf{b}$ are obtained by sampling from the posterior distribution using reparameterization trick [31], and they are fed into the decoder $p(\cdot)$ to reconstruct the input representation $\mathbf{h}'$. In Eq. (3), the first term corresponds to the process by which the decoder $p(\cdot)$ decodes $\mathbf{z}$ and $\mathbf{b}$ to obtain the reconstructed representation $\widehat{\mathbf{h}'}$ to fit the input representation $\mathbf{h}'$, and the last two terms correspond to the disentanglement of the fair representation and biased representation.

4.2.2. Distillation of biased representation

To encourage the decoupling of biased representation, we employ the biased teacher model trained in the first stage to supervise feature distillation of biased representation. The biased teacher model is task-agnostic and can exhibit general gender bias when it is applied to a certain downstream task. In the process of feature distillation, the biased representation can learn the bias information in the task data under the guidance of the biased teacher model.

Formally, for each sample $x' \in \mathcal{X}'$, the biased teacher model $E_b(\cdot)$ encodes it to obtain the representation $\mathbf{b_T}$. We feed the biased representation output by the teacher model into a linear layer to align with the dimensions of the biased representation. Following the work [52], we adopt a feature whitening strategy to normalize the biased representation $\mathbf{b}$ and the teacher representation $\mathbf{b_T}$, ensuring that they are comparable in scale and distribution. Specifically, we implement this whitening using a non-parametric layer normalization operator, which standardizes the input by removing the mean and scaling by the standard deviation across feature dimensions, without applying learnable affine transformations (i.e., no scaling or bias). Formally, the whitening operation $\zeta(\cdot)$ applied to a vector $\mathbf{x} \in \mathbb{R}^d$ is given by:

$$\zeta(\mathbf{x}) = \frac{\mathbf{x} - \mu(\mathbf{x})}{\sigma(\mathbf{x})}. \tag{4}$$

This normalization aligns the distributions of $\mathbf{b}$ and $\mathbf{b_T}$, making them comparable. To distill information from the teacher, we apply a smooth l_1 loss between the whitened features:

$$\mathcal{L}_{distill} = \begin{cases} \frac{1}{2}(\zeta(\mathbf{b}) - \zeta(\mathbf{b_T}))^2/\lambda, & |\zeta(\mathbf{b}) - \zeta(\mathbf{b_T})| \le \lambda, \\ |\zeta(\mathbf{b}) - \zeta(\mathbf{b_T})| - \frac{1}{2}\lambda, & otherwise, \end{cases} \tag{5}$$

where λ is a hyperparameter (default value 2.0) that controls the transition point between quadratic and linear behavior. This loss function behaves quadratically for small discrepancies, encouraging precise alignment between features, while switching to a linear form for larger discrepancies to reduce sensitivity to outliers and stabilize training. Under this distillation constraint, the biased representation $\mathbf{b}$ is progressively encouraged to align with the teacher's fair representation $\mathbf{b_T}$. This process effectively promotes disentanglement between fair and biased representations, enabling the model to retain task-relevant information while suppressing undesired bias.

4.2.3. Debiasing training

After separating the biased representation $\mathbf{b}$ from the input representation $\mathbf{h}'$, the fair representation $\mathbf{z}$ should be gender insensitive and not degrade performance on downstream tasks. Therefore, we use the fair representation $\mathbf{z}$ to learn the downstream task. The task objective with cross-entropy loss is defined as:

$$\mathcal{L}_{ce} = \text{CrossEntropy}(C(\mathbf{z})), \tag{6}$$

where $C(\cdot)$ is the classification head of the task.

The overall objective of debiasing training consists of three loss functions as:

$$\mathcal{L}oss = \mathcal{L}_{ce} + \alpha\mathcal{L}_{distill} + \gamma\mathcal{L}_{vae}, \tag{7}$$

where α is the hyper-parameter that controls the weights of the feature distillation and γ is the hyper-parameter that controls the weights of the VAE. After debiasing training, we jointly optimize PLM and VAE under the guidance of biased teacher to disentangle the fair representation that do not contain bias information and can predict downstream tasks from the representation encoded by PLM.

5. Experiments

In this section, in order to verify the effectiveness of our proposed debiasing method, we conduct a comprehensive experimental analysis of BATED by answering the following four Research Questions (RQ).

RQ1: How effective is applying BATED to debias PLMs in downstream tasks?
RQ2: How effective is the training strategy for causal contrastive learning?
RQ3: How each module in BATED contributes to the performance of the model?
RQ4: How well does BATED generalize for debiasing out-of-domain tasks?

5.1. Experimental setup

5.1.1. Baselines

We select eight general and state-of-the-art methods for debiasing PLMs as baselines, and they are introduced as follows:

- **INLP** [28] is a debiasing method based on separating sensitive information. It iteratively trains a linear classifier to predict the debiasing target, and then projects the neural representation into the null space of the debiasing target to separate the sensitive information.

Table 1
Implementation details of downstream tasks.

Setting / Task	Dataset Size	Training Set	Validation Set	Test Set	Learning Rate	Batch Size	Max Length	Hyper-parameter	
								α	γ
SC	163,095	114,167 (70%)	24,464 (15%)	24,464 (15%)	$3e-5$	128	128	0.5	0.02
Bios	251,101	163,216 (65%)	62,775 (25%)	25,110 (10%)	$5e-5$	128	128	0.1	0.02
UBM	1,459,516	1,399,662 (70%)	29,927 (15%)	29,927 (15%)	$1e-5$	128	128	0.1	0.005

- **Sent-Debias** [34] proposes a debiasing method based on projected sentence representation, which uses sentence templates to contextualize the priori sensitive attribute words, and then estimates and eliminates the projection of the sentence representation in the bias subspace.
- **FairFil** [15] proposes the debiasing method of training a fair filter, which uses CDA to augment the original sample to a pair of positive samples, and then trains an additional neural network under an unsupervised contrastive learning framework to filter the bias in the PLM-encoded representation.
- **Auto-Debias** [19] proposes uses the max-min Jensen–Shannon divergence to debias. Based on the given sensitive attribute word list and stereotype word list, Auto-Debias aligns the model's stereotype of different groups by concatenating the discrete prompts that can amplify the bias to the sensitive attribute words.
- **MABEL** [16] generalize CDA to the natural language inference datasets. It generates augmented positive samples of "premise-hypothesis" sentence pairs containing sensitive attribute words, leveraging contrastive learning to approximate the representations between the positive sample sentence pairs to mitigate bias.
- **CCPA** [17] solves the possible overfitting in the debiasing process. It uses CDA to augment the original samples and trains a neural network to learn continuous prompts that can push away the representation between the augmented sample pairs, thus increasing the difficulty of the contrastive debiasing process.
- **BNS** [20] utilizes the integrated gap gradient to locate neurons in the language model that can be attributed to social biases, and then suppresses the bias neuron by setting the bias neuron to zero, which achieves the debiasing of the model by shifting the bias distribution.
- **DeepSoftDebias** [21] adjusts the soft debiasing method by replacing the transformation matrix with a neural network composed of residual blocks. It handles more complex functional mappings between input and output embeddings by exploiting the ability to represent sequences of transformation matrices.

5.1.2. Datasets

Training the biased teacher model is an unsupervised learning process. Therefore, in the stage of training the biased teacher model, we adopt five real-world datasets widely used in prior work [34,15,17] as original corpus, which are Stanford Sentiment Treebank [53], POM [54], WikiText-2 [55], Reddit [56] and MELD [57]. From the original corpus, we match more than 150,000 samples containing sensitive attributes to form the training dataset.

The debiasing training phase is specific to downstream tasks. We choose three general classification tasks for experiments, and their dataset size details are shown in Table 1, which are described as follows:

- **Sentiment Classification (SC)** task aggregates samples from five sentiment classification datasets: mr [58], foods [59], IMDB [60], Product Sentiment Classification,[2] and Twitter Sentiment Analysis.[3] Among them, mr, foods, and IMDB are binary classification datasets, Product Sentiment Classification and Twitter Sentiment Analysis are three-classification datasets. We only select positive and negative samples in Product Sentiment Classification and Twitter Sentiment Analysis to unify the SC task. The dataset of SC does not contain gender annotations, and to evaluate the effectiveness of the debiasing method, we utilize the sensitive attribute words to match the dataset and use them as gender annotations of the samples. The matched dataset contains 163,095 samples, which are divided into the train-validation-test set according to the ratio of 70%:15%:15%.
- **Bias-in-Bios (Bios)** [61] is a task in which a third person biography predicts occupation. We capture more than 250,000 samples from 28 occupational categories as the overall dataset, where each sample is annotated with a binary gender label. Consistent with the setup of previous work [61], we split the train-validation-test set with a 65%:25%:10% ratio.
- **Unintended Bias Metrics (UBM)** [62] is a binary toxicity detection tasks. Its dataset contains more than 1.4 million sample with 450,000 labeled identity group samples, and we split it into train-validation-test set with a 70%:15%:15% ratio. Samples labeled with toxicity greater than 0.5 are considered toxic, and the threshold of the metric AEGs is set to 0.5.

5.1.3. Implementation details

We adopt seven PLMs as base models for debiasing experiments, including **BERT** [1], **DistilBERT** [63], **ELECTRA** [64], **OPT** [65], **GPT-2** [2], **QWEN-2.5** [66,67] and **LLaMA-3.2** [3,4]. The selection criteria for base models are based on three key considerations: 1) practical prevalence in real-world applications, 2) diversity in parameter scales, and 3) representation of different architectural paradigms. Specifically, BERT for its widespread adoption, DistilBERT as a representative of compact models, ELECTRA and OPT to

[2] https://www.kaggle.com/datasets/akash14/product-sentiment-classification/data.
[3] https://www.kaggle.com/datasets/saurabhshahane/twitter-sentiment-dataset.

Table 2
Implementation details for training the biased teacher models.

Setting \ Model	BERT	DistilBERT	ELECTRA	OPT	GPT-2	QWEN-2.5	LLaMA-3.2
Learning Rate	$1e-5$	$1e-6$	$1e-6$	$1e-7$	$1e-6$	$1e-6$	$1e-6$
ϵ	15%	15%	15%	15%	15%	15%	15%
$\triangle m$	2	2	2	2	2	2	2
Batch Size	128	128	128	128	128	128	128
Max Length	128	128	128	128	128	128	128

cover different architectural approaches, along with three LLMs (GPT-2, QWEN-2.5 and LLaMA-3.2) to examine the scalability of our method across varying model sizes. This comprehensive selection enables us to thoroughly evaluate the generalization capability of our debiasing approach. In each debiasing experiment, the biased teacher model is consistent with the base model. All experiments use checkpoints as `bert-base-uncased`, `distilbert-base-uncased`, `google/electra-base-generator`, `fa-cebook/opt-350m`, `gpt2-xl`, `Qwen/Qwen2.5-1.5B`, and `meta-llama/Llama-3.2-3B` implemented based on Huggingface Transformer [68].

In the stage of training the biased teacher model, the probability of randomly masking each word ϵ is 15% [1], $\triangle m$ is set to 2, and the training is stopped after one epoch. The batch size is set to 128 and the maximum sentence length is set to 128. To prevent the biased teacher model from converging to suboptimal solutions, we select learning rates based on both the parameter scale of each model and the stability of loss convergence during fine-tuning. The learning rate of training biased BERT is $1e-5$, the learning rate of training biased DistilBERT, biased ELECTRA, biased GPT-2, biased QWEN-2.5, and biased LLamA-3.2 is $1e-6$, and the learning rate of training biased OPT is $1e-7$. Following the previous work [69,34,15,37], we choose the sensitive attribute words specific to gender groups as: {FEMALE, MALE} = {(woman, man), (girl, boy), (female, male), (she, he), (mother, father), (daughter, son), (gal, guy), (her, his), (herself, himself), (Mary, John)}, alongside plural forms. Implementation details for training the biased teacher models are summarized in Table 2.

In the stage of debiasing training via the biased teacher model, VAE employs an encoder and a decoder, which are both composed of two linear layers and an activation layer. For each downstream task, the debiasing model is trained exclusively on the respective training set of each dataset, followed by validated on the validation set, with final performance evaluation conducted on a held-out test set. For the three downstream tasks, the batch size is set to 128, the maximum sentence length is set to 128, the hyper-parameter α is chosen from [1, 0.5, 0.2, 0.1, 0.05], and the hyper-parameter γ is chosen from [1, 0.5, 0.2, 0.05, 0.02, 0.005, 0.002]. The ranges of α and γ are selected based on preliminary empirical studies. These ranges are designed to span a balanced spectrum of influence between the bias mitigation term and the main task loss. We explore the hyper-parameters in Section 5.5.2, and the results reported in other experiments are under the optimal hyper-parameter settings provided in Table 1. We report the average and standard deviation of the results from running 5 experiments.

To ensure fairness against the baselines, all baseline methods adopt their original experimental setup for debiasing training, and then align with the BATED setup during the fine-tuning of the downstream tasks. We re-run INLP and Sent-Debias with the code provided by [70], re-run FairFil, Auto-Debias, CCPA, BNS, and DeepSoftDebias with the codes provided by their authors, and adopt MABEL with the checkpoint uploaded by its authors.

The resources required to run the experiment are determined by the number of parameters of the PLM. All experiments can be performed on a single instance of an NVIDIA RTX 3080 Ti GPU card. GPT-2, QWEN-2.5, and LLaMA-3.2 need to adopt the Low-Rank Adaptation [71] technique in the PEFT library[4] to reduce the number of parameters for tuning, otherwise the batch size needs to be adjusted to 4.

5.2. Evaluation metrics

In order to comprehensively evaluate our proposed BATED, we employ multiple fairness metrics including four intrinsic bias metrics and two extrinsic bias metrics to evaluate the fairness of the model, and four task metrics: **Accuracy (Acc.)**, **F1**, **Precision**, and **Recall** to evaluate the performance of the model in downstream tasks.

5.2.1. Intrinsic bias metrics

Intrinsic bias metrics formalize the intrinsic bias of a model by measuring the representations encoded by the model, and they do not require downstream tasks. Since our proposed strategy for training biased teacher models is an unsupervised learning process, we utilize four general intrinsic bias metrics: SEAT [72], CrowS-Pairs [73], StereoSet [74], and GBES to evaluate biased teacher models in Section 5.4.

Sentence Encoder Association Test (SEAT) [72] fills two sets of sensitive attribute words $\mathcal{A}$ and $\mathcal{B}$ and two sets of stereotype words $\mathcal{X}$ and $\mathcal{Y}$ into the bleached sentence template and measures the effect size between them:

$$d = \frac{\mu(\{s(x,\mathcal{A},\mathcal{B})\}) - \mu(\{s(y,\mathcal{A},\mathcal{B})\})}{\sigma(\{s(t,\mathcal{X},\mathcal{Y})\}_{t\in\mathcal{A}\cup\mathcal{B}})}, \tag{8}$$

where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, $\mu(\cdot)$ is the mean function and $\sigma(\cdot)$ is the standard deviation. And $s(w,\mathcal{A},\mathcal{B})$ is the bias degree defined as:

[4] https://huggingface.co/docs/peft/package_reference/lora.

$$s(w, \mathcal{A}, \mathcal{B}) = \mu(cos(w, a)) - \mu(cos(w, b)). \tag{9}$$

The gender subsets in SEAT are *6*, *6b*, *7*, *7b*, *8* and *8b*. We report the absolute value of the effect size for each subset and the average over all subsets (*Avg.*) in Table 6. Values closer to 0 represent less bias, and values further away from 0 represent more bias.

Crowdsourced Stereotype Pairs (CrowS-Pairs) [73] proposes a crowdsourced dataset containing semantically opposite stereotyped and anti-stereotyped sentences. It computes the perplexity of all tokens conditioned on typical tokens using pseudo-log-likelihood, defined as follows:

$$score(S) = \sum_{i=0}^{|C|} \log P(u_i \in U | U_{\backslash u_i}, M, \theta), \tag{10}$$

where u_i is a token in sentence U and M is a language model with the learnable parameter θ. We report scores for stereotyped and anti-stereotyped sentences denoted as *Sereo* and *Anti-Stereo*, respectively, which have values close to 50, indicating that the model is less biased.

StereoSet [74] is a crowdsourced dataset measuring four social biases, where each sample consists of a contextual sentence and a set of candidate associations. The model chooses among three candidate associations: stereotypic, anti-stereotypic, and meaningless, and obtains a stereotypic score for each protected group. The percentage of the model choosing stereotype over anti-stereotype represents the bias degree of the model, denoted as *SS Score*, where closer to 50 represents less bias.

Gender Bias Evaluation Set (GBES)[5] is used to evaluate the binary gender bias of PLMs under different stereotyped occupations. It contains 1,585 sentence pairs about stereotype occupations constructed by converting base sentences from WinoMT [75] into sentences for male subgroup and sentences for female subgroup, respectively. Bias is assessed by calculating the similarity of sentence pair embeddings. We adopt cosine similarity as the evaluation function and report the average of cosine similarity of all sentence pairs, whose closer to 1 indicates less gender bias of the model.

5.2.2. Extrinsic bias metrics

Extrinsic bias metrics measure the model's decision fairness in downstream tasks. Based on the fairness criterion demographic parity [76], we employ two extrinsic bias metrics to evaluate the performance of debiasing methods in three downstream tasks in Section 5.3. SC task and Bios task are measured by four sub-metrics TPR_M, TPR_F, TPG_GAP, and TPG_RMS proposed by Bios [61], and UBM task is measured by five sub-metrics AUC_sub, AUC_bpsn, AUC_bnsp, AEG_neg, and AEG_pos proposed by UBM [62].

Bios proposes fairness metrics for calculating bias using the true positive rate. TPR_M represents the true positive rate for the male samples, TPR_F represents the true positive rate for the female samples, and the closer they are to 1.0, the lower the bias. TPR_GAP represents the gap between the true positive rates of the male and female samples, and TPR_RMS represents the root mean square of the gap between the true positive rates of the male and female samples in each category, where closer they are to 0 indicates lower bias. They are defined as follows:

$$\begin{aligned} \text{TPR}_\text{GAP} &= |\text{TPR}_\text{M} - \text{TPR}_\text{F}|, \\ \text{TPR}_\text{RMS} &= \sqrt{\frac{1}{|Y|} \sum_{y \in Y} (\text{TPR}_{\text{GAP,y}})^2}. \end{aligned} \tag{11}$$

UBM proposes five fairness sub-metrics including three sub-metrics based on the Area Under the Receiver Operating Characteristic Curve (AUC) AUC_sub, AUC_bpsn, and AUC_bnsp, which are closer to 1.0 to indicate better fairness, and two sub-metrics based on Average Equality Gap (AEGs) AEG_neg and AEG_pos, which are closer to 0 to indicate better fairness. AUCs metrics are calculated based on the negative/positive mis-ordering between identity subgroups and backgrounds:

$$\begin{aligned} \text{AUC}_\text{sub} &= \text{AUC}_\text{sub}(D_g^- + D_g^+), \\ \text{AUC}_\text{bpsn} &= \text{AUC}_\text{bpsn}(D^+ + D_g^-), \\ \text{AUC}_\text{bnsp} &= \text{AUC}_\text{bnsp}(D^- + D_g^+), \end{aligned} \tag{12}$$

where D^+ and D^- represent the positive and negative examples of the background set, D_g^+ and D_g^- represent the positive and negative examples of the identity subgroup.

AEGs metrics are built on the equality gap metric to calculate the true positive rate difference between subgroups and backgrounds for a specific threshold, which describes more subtle differences in distribution. They are defined as follows:

$$\begin{aligned} \text{AEG}_\text{neg} &= \frac{1}{2} - \frac{\text{MWU}(\text{D}_\text{g}^+, \text{D}^+)}{|D_g^+||D^+|}, \\ \text{AEG}_\text{pos} &= \frac{1}{2} - \frac{\text{MWU}(\text{D}_\text{g}^-, \text{D}^-)}{|D_g^-||D^-|}, \end{aligned} \tag{13}$$

where $\text{MWU}(\cdot, \cdot)$ is the Mann-Whitney U test statistic.

[5] https://huggingface.co/datasets/flax-sentence-embeddings/Gender_Bias_Evaluation_Set.

Table 3
Debiasing results on the SC task. The best result is indicated in **bold**. The suboptimal result is indicated in underline. * represent statistically significant compared to other results (base on Mann-Whitney U test; $\rho < 0.05$). MABEL does not provide code to run on GPT-2 and LLaMA-3.2.

Metric / Model	Fairness metrics (%)				Task metrics (%)			
	TPR_M↑	TPR_F↑	TPR_{GAP}↓	TPR_{RMS}↓	Acc.↑	F1↑	Precision↑	Recall↑
BERT	94.08±0.04	95.55±0.08	1.47±0.21	1.32±0.15	95.15±0.04	94.45±0.05	94.31±0.10	94.45±0.12
+INLP	95.07 ± 0.05	94.35±0.05	0.72±0.18	1.48±0.12	94.96±0.05	94.28±0.04	94.25±0.09	94.31±0.11
+Sent-Debias	94.02±0.04	94.83±0.10	0.81±0.16	0.78 ± 0.16	94.94±0.06	94.24±0.06	94.09±0.07	94.40±0.09
+FairFil	94.53±0.10	95.43±0.06	0.90±0.20	1.02±0.17	95.03±0.05	94.34±0.04	94.20±0.09	94.48±0.08
+Auto-Debias	94.85±0.08	94.11±0.11	0.74±0.13	1.65±0.20	94.96±0.06	94.29±0.06	94.33±0.09	94.25±0.09
+MABEL	94.39±0.05	95.19±0.08	0.81±0.16	1.23±0.16	94.91±0.03	94.17±0.03	93.72±0.12	94.67±0.07
+CCPA	94.82±0.06	95.67 ± 0.11	0.85±0.14	0.87±0.21	94.92±0.04	94.22±0.04	94.12±0.08	94.32±0.11
+BNS	94.88±0.10	95.56±0.07	0.68 ± 0.22	0.98±0.16	95.13±0.04	94.30±0.02	94.37 ± 0.07	94.68 ± 0.09
+DeepSoftDebias	94.71±0.06	95.49±0.06	0.78±0.17	0.90±0.15	95.18 ± 0.05	94.48 ± 0.04	94.22±0.08	94.50±0.09
+BATED	**95.08±0.09***	**95.73±0.08***	**0.65±0.13***	**0.53±0.16***	**95.24±0.04**	**94.58±0.04**	**94.46±0.07**	**94.70±0.08**
GPT-2	88.72±0.06	90.34 ± 0.03	1.63±0.11	2.27±0.35	**90.49±0.03**	**88.96±0.07**	88.11±0.04	**90.02±0.06**
+INLP	88.63±0.05	89.92±0.04	1.29 ± 0.09	2.15±0.30	90.32±0.04	88.74±0.06	88.03±0.03	89.05±0.07
+Sent-Debias	88.30±0.07	89.65±0.06	1.35±0.12	1.49±0.28	90.28±0.05	88.65±0.05	87.92±0.04	89.18±0.08
+FairFil	88.43±0.06	90.12±0.05	1.69±0.10	1.75±0.32	90.37±0.04	88.82±0.05	87.98±0.05	89.11±0.07
+Auto-Debias	88.12±0.08	89.45±0.07	1.33±0.14	2.42±0.28	90.31±0.06	88.71±0.06	88.05±0.04	89.10±0.09
+MABEL	-	-	-	-	-	-	-	-
+CCPA	88.60±0.07	90.22±0.08	1.62±0.13	1.59±0.36	90.26±0.05	88.63±0.05	87.94±0.06	89.16±0.08
+BNS	88.66±0.09	90.28±0.06	1.62±0.19	1.35 ± 0.33	90.47 ± 0.04	88.78±0.03	88.09 ± 0.06	88.60±0.08
+DeepSoftDebias	88.78 ± 0.07	90.22±0.05	1.54±0.15	1.42±0.32	90.32±0.05	88.86 ± 0.05	87.94±0.07	88.88±0.08
+BATED	**89.14±0.05***	**90.36±0.05***	**1.22±0.09***	**0.92±0.22***	90.38±0.04	88.58±0.05	**88.14±0.03**	89.20 ± 0.07
LLaMA-3.2	94.71±0.05	96.27±0.07	1.56±0.08	2.30±0.15	96.02 ± 0.02	95.47±0.04	95.28±0.06	**95.66±0.05**
+INLP	94.93±0.04	96.51±0.06	1.59±0.07	2.90±0.12	96.01±0.03	95.47±0.04	95.34±0.05	95.59±0.06
+Sent-Debias	95.26±0.03	96.51±0.05	1.26±0.06	1.89±0.10	96.02±0.02	**95.48±0.03**	95.34 ± 0.04	95.43±0.05
+FairFil	95.18±0.04	96.53 ± 0.06	1.35±0.07	2.54±0.11	95.97±0.03	95.40±0.04	95.17±0.05	95.63 ± 0.06
+Auto-Debias	95.26±0.03	96.46±0.05	1.21 ± 0.06	2.17±0.10	96.00±0.02	95.45±0.03	95.32±0.04	95.59±0.05
+MABEL	-	-	-	-	-	-	-	-
+CCPA	95.09±0.04	96.41±0.06	1.32±0.07	2.17±0.11	96.00±0.03	95.44±0.04	95.30±0.05	95.59±0.06
+BNS	94.71±0.05	96.03±0.08	1.32±0.08	1.71 ± 0.12	96.01±0.03	95.45±0.04	95.31±0.05	95.60±0.06
+DeepSoftDebias	95.29 ± 0.03	96.51±0.05	1.22±0.06	1.93±0.10	96.00±0.02	95.44±0.03	95.30±0.04	95.59±0.05
+BATED	**95.53±0.02***	**96.61±0.04***	**1.09±0.05***	**1.25±0.08***	**96.03±0.02**	95.47 ± 0.03	**95.37±0.04**	95.61±0.05
DistilBERT	93.88±0.07	94.83±0.03	0.95±0.13	1.81±0.20	94.54±0.03	93.82±0.02	**93.90±0.04**	93.74±0.06
+BATED	**94.50±0.04**	**95.07±0.03**	**0.58±0.10**	**0.75±0.15**	**94.59±0.03**	**93.86±0.03**	93.85±0.03	**93.87±0.05**
ELECTRA	94.20±0.07	93.87±0.08	0.33±0.08	2.29±0.20	94.52±0.03	93.79±0.02	**93.84±0.04**	93.75±0.08
+BATED	**94.53±0.08**	**94.59±0.10**	**0.06±0.02**	**0.17±0.08**	**94.64±0.03**	**93.86±0.03**	93.44±0.05	**94.31±0.06**
OPT	93.08±0.04	92.09±0.06	0.99±0.10	3.00±0.17	**93.41±0.02**	**92.49±0.02**	92.29±0.04	92.69±0.05
+BATED	**93.36±0.05**	**92.67±0.07**	**0.67±0.06**	**2.15±0.12**	93.36±0.02	92.40±0.03	**92.51±0.03**	**92.80±0.06**
QWEN-2.5	93.63±0.08	94.75±0.07	1.12±0.15	1.53±0.13	**95.10±0.03**	94.42±0.07	94.21±0.05	**94.65±0.10**
+BATED	**94.53±0.05**	**95.19±0.06**	**0.66±0.09**	**0.56±0.08**	95.09±0.04	**94.43±0.06**	**94.25±0.05**	94.61±0.10

5.3. Debiasing performance analysis

5.3.1. Results of debiasing experiment

To answer research question **RQ1**, we test the debiasing ability of BATED against BERT, DistilBERT, ELECTRA, OPT, GPT-2, QWEN-2.5, and LLaMA-3.2 on three downstream tasks and compare with eight baseline methods. Meanwhile, we report the experimental results of the original BERT, DistilBERT, ELECTRA, OPT, GPT-2, QWEN-2.5, and LLaMA-3.2 as a reference. Table 3, Table 4, and Table 5 show the debiasing results on SC task, Bios task, and UBM task, respectively. We make the following main observations.

❶ *BATED can greatly improve the fairness of the original PLMs in downstream tasks.* In terms of fairness metrics, BATED alleviates the extrinsic biases of the seven original PLMs on three downstream tasks. According to the four fairness sub-metrics in Tables 3 and 4, our proposed debiasing method performs outstanding on both SC task and Bios task. The decrease of TPR_{GAP} and TPR_{RMS} indicates that BATED can reduce the prediction gap between different gender groups, while the increase of TPR_M and TPR_F indicates that BATED can not only reduce the bias but also improve the prediction performance for each group. The five fairness sub-metrics of Table 5 measure extrinsic biases from two dimensions, AUC_{sub}, AUC_{bpsn}, and AUC_{bnsp} are evaluated from a macro perspective, AEG_{neg} and AEG_{pos} are evaluated from a more nuanced perspective. Encouragingly, compared to the original PLMs, BATED achieves the best results on all the five fairness sub-metrics, both for the male and female subgroups.

❷ *Compared with the baseline methods, BATED has significant and stable debiasing performance.* Tables 3 to 5 show the comparison of applying BATED and baseline methods to debias BERT, GPT-2, and LLaMA-3.2 on each of the three tasks, respectively. Baseline methods can alleviate the bias of the original PLMs to varying degrees, but their debiasing ability is limited and it is difficult to meet the requirements of all measurement dimensions. For example, INLP performs well on TPR_M and TPR_{GAP} but poorly on TPR_F

Table 4
Debiasing results on the Bios task. The best result is indicated in **bold**. The suboptimal result is indicated in underline. * represent statistically significant compared to other results (base on Mann-Whitney U test; $\rho < 0.05$). MABEL does not provide code to run on GPT-2 and LLaMA-3.2.

Metric / Model	Fairness metrics (%)				Task metrics (%)			
	TPR_M↑	TPR_F↑	TPR_{GAP}↓	TPR_{RMS}↓	Acc.↑	F1↑	Precision↑	Recall↑
BERT	84.33±0.05	85.69±0.03	1.37±0.17	14.87±0.19	85.26±0.05	80.21±0.06	79.97±0.03	81.20±0.02
+INLP	84.94±0.04	85.86±0.03	0.92±0.10	14.04±0.13	85.36±0.05	80.44±0.05	**80.33±0.05**	81.21±0.03
+Sent-Debias	84.69±0.06	85.59±0.02	0.90±0.12	15.52±0.17	85.36±0.04	80.16±0.04	79.09±0.06	81.93±0.03
+FairFil	84.79±0.05	85.62±0.04	0.84 ± 0.10	13.99±0.16	85.24±0.03	79.95±0.06	78.65±0.10	81.80±0.05
+Auto-Debias	84.90±0.05	85.88±0.03	0.98±0.13	13.91±0.15	85.34±0.04	80.00±0.07	78.42±0.07	82.26±0.03
+MABEL	85.01 ± 0.05	85.90±0.03	0.89±0.07	13.93±0.14	85.34±0.03	80.11±0.05	78.99±0.06	81.84±0.05
+CCPA	84.88±0.04	85.76±0.04	0.89±0.09	13.78±0.16	85.37 ± 0.04	**80.50±0.06**	79.35±0.05	82.26±0.06
+BNS	84.86±0.04	85.81±0.04	0.94±0.08	13.76 ± 0.15	85.26±0.03	80.31±0.04	80.03 ± 0.07	82.33 ± 0.03
+DeepSoftDebias	85.00±0.06	85.92 ± 0.03	0.92±0.05	14.11±0.17	85.30±0.03	80.06±0.07	78.87±0.09	81.96±0.03
+BATED	**85.17±0.05***	**85.98±0.04***	**0.81±0.07***	**13.68±0.12***	**85.39±0.03**	80.45 ± 0.04	79.90±0.06	**82.41±0.04**
GPT-2	80.17±0.07	81.07 ± 0.05	0.90±0.08	16.20±0.22	**80.77±0.02**	72.31±0.03	69.63±0.12	**76.88±0.23**
+INLP	79.54±0.06	80.45±0.05	0.91±0.05	15.70±0.17	80.32±0.03	72.11±0.04	68.43±0.18	74.93±0.19
+Sent-Debias	79.37±0.04	80.21±0.04	0.84±0.10	16.11±0.16	80.21±0.04	71.69±0.05	67.28±0.35	74.51±0.16
+FairFil	79.67±0.09	80.38±0.09	0.71±0.06	15.69±0.14	80.37±0.05	71.45±0.06	68.57±0.19	75.13±0.25
+Auto-Debias	79.89±0.08	80.87±0.07	0.98±0.13	15.04±0.17	80.19±0.03	71.67±0.05	69.22±0.17	75.20±0.17
+MABEL	-	-	-	-	-	-	-	-
+CCPA	80.05±0.05	80.76±0.06	0.71±0.08	15.08±0.16	80.46±0.05	70.28±0.05	68.17±0.19	75.47 ± 0.26
+BNS	80.30±0.06	81.02±0.05	0.72±0.09	15.32±0.24	80.40±0.05	72.41 ± 0.06	69.79±0.16	75.33±0.21
+DeepSoftDebias	80.31 ± 0.07	80.97±0.05	0.66 ± 0.09	14.89 ± 0.25	80.50±0.06	71.89±0.08	69.88 ± 0.15	74.98±0.18
+BATED	**80.47±0.05***	**81.07±0.03***	**0.61±0.06***	**14.77±0.14***	80.54 ± 0.04	**72.47±0.06**	**70.57±0.14**	75.24±0.18
LLaMA-3.2	84.15±0.07	84.95±0.08	0.80±0.11	16.32±0.34	84.48±0.04	80.49 ± 0.10	79.66±0.22	**82.46±0.19**
+INLP	84.14±0.06	84.88±0.07	0.74±0.10	15.20±0.30	84.40±0.04	80.13±0.09	79.46±0.20	82.04±0.18
+Sent-Debias	83.69±0.08	84.40±0.09	0.71±0.12	15.24±0.33	83.93±0.05	78.83±0.11	77.18±0.23	81.45±0.21
+FairFil	84.08±0.07	84.81±0.08	0.73±0.11	16.14±0.35	**84.57±0.03**	79.46±0.10	78.14±0.21	82.04±0.19
+Auto-Debias	83.72±0.09	83.64±0.10	0.09±0.15	14.69±0.32	83.68±0.06	78.76±0.12	77.75±0.24	81.99±0.22
+MABEL	-	-	-	-	-	-	-	-
+CCPA	84.15±0.06	84.80±0.07	0.65±0.10	15.49±0.31	84.37±0.04	80.40±0.09	79.34±0.20	81.46±0.18
+BNS	84.22±0.06	84.82±0.07	0.60 ± 0.09	14.88±0.30	84.21±0.04	80.02±0.09	79.96 ± 0.19	81.88±0.17
+DeepSoftDebias	84.30 ± 0.05	84.96 ± 0.06	0.66±0.09	14.52 ± 0.28	84.47±0.03	80.54 ± 0.08	79.43±0.18	81.45±0.16
+BATED	**84.48±0.06***	**85.01±0.05***	**0.53±0.07***	**14.29±0.29***	84.48 ± 0.03	**80.72±0.11**	80.00±0.19	82.39 ± 0.20
DistilBERT	84.32±0.03	85.47±0.03	1.16±0.10	15.45±0.12	85.12±0.03	**79.97±0.04**	78.96±0.02	**81.86±0.04**
+BATED	**84.74±0.04**	**85.57±0.03**	**0.83±0.06**	**14.44±0.10**	**85.19±0.02**	79.75±0.02	**78.97±0.02**	81.50±0.05
ELECTRA	83.31±0.03	84.31±0.05	1.00±0.17	16.24±0.21	84.08±0.01	**78.39±0.03**	**77.39±0.03**	79.92±0.02
+BATED	**83.92±0.04**	**84.47±0.04**	**0.55±0.09**	**14.27±0.12**	**84.12±0.02**	77.74±0.04	76.17±0.02	**80.61±0.04**
OPT	82.03±0.03	83.08±0.01	1.05±0.06	15.62±0.23	**82.70±0.02**	**76.10±0.05**	75.07±0.04	78.20±0.03
+BATED	**82.30±0.03**	**83.14±0.02**	**0.84±0.04**	**13.65±0.17**	82.68±0.03	75.30±0.04	**75.84±0.03**	**78.88±0.04**
QWEN-2.5	85.25±0.04	86.53±0.05	1.28±0.07	15.21±0.17	**86.52±0.03**	**81.92±0.08**	80.58±0.19	83.82±0.11
+BATED	**85.81±0.03**	**86.54±0.06**	**0.73±0.07**	**12.56±0.13**	86.14±0.04	81.65±0.09	**80.79±0.14**	**83.91±0.12**

and TPR_{RMS} on BERT in the SC tasks, and does not perform well on the Bios and UBM tasks. Similar situations arise for other debiasing methods. Compared to the baselines, BATED achieves the best and significant results in all fairness metrics for all tasks, which indicates that our proposed debiasing method has the strongest debiasing ability and stable performance.

❸ *BATED does not compromise the language modeling capabilities of PLMs.* Observing the results for the task metrics reported in the three Tables, it can be found that our proposed debiasing method achieves comparable accuracy scores to the original PLMs overall, and this is true for all seven PLMs. In many cases, BATED even exceeds the scores of the task metrics of the original PLMs, especially in the SC task. The results of the task metrics verify that BATED can preserve the original language modeling abilities of PLMs while alleviating biases.

❹ *BATED achieves an excellent trade-off between fairness and accuracy with respect to other debiasing methods.* In debiasing research, the trade-off between fairness and accuracy is a major challenge, and it is not yet best defined. As shown in Tables 3 through 5, baseline methods either mitigate bias at the expense of task performance or maintain task performance with weak debiasing effects. In contrast, BATED can effectively debias PLMs in both binary and multi-class classification tasks and preserve or even enhance the learning ability of the model for downstream tasks. Therefore, we can conclude that BATED is able to trade off fairness and accuracy.

5.3.2. Case study

To more clearly demonstrate the effectiveness of our debiasing method, we present several sample outputs from different downstream tasks, as shown in Fig. 2. For each task, we select four representative examples and highlight the gender-related segments in the text. We report both the original BERT predictions and the outputs after applying BATED. These examples offer an intuitive illustration of BATED's effectiveness in mitigating gender bias.

Table 5
Debiasing results on the UBM task. 'f' and 'm' represent the results for the female and male subgroups, respectively. The best result is indicated in **bold**. The suboptimal result is indicated in underline. * represent statistically significant compared to other results (base on Mann-Whitney U test; $\rho < 0.05$). MABEL does not provide code to run on GPT-2 and LLaMA-3.2.

Metric / Model	Fairness metrics (%)					Task metrics (%)			
	AUC_{sub}↑ (f/m)	AUC_{bpsn}↑ (f/m)	AUC_{bnsp}↑ (f/m)	AEG_{neg}↓ (f/m)	AEG_{pos}↓ (f/m)	Acc.↑	F1↑	Precision↑	Recall↑
BERT	91.90±0.02/91.12±0.02	92.91±0.03/92.78±0.04	96.12±0.04/95.86±0.03	20.24±0.10/20.00±0.08	7.69±0.05/8.05±0.07	95.40±0.01	82.88±0.05	80.06±0.12	86.46±0.07
+INLP	91.73±0.03/91.11±0.03	93.03±0.02/92.34±0.03	95.83±0.03/95.95±0.04	19.87±0.12/20.67±0.10	7.94±0.07/8.25±0.05	95.39±0.02	82.48±0.06	79.07±0.14	**87.09±0.08**
+Sent-Debias	92.03±0.04/91.32±0.02	92.91±0.02/92.48±0.04	96.23±0.05/96.19 ± 0.03	22.37±0.15/21.68±0.07	7.49±0.05/8.41±0.06	95.42±0.01	83.25 ± 0.05	80.89±0.11	86.10±0.12
+FairFil	91.86±0.02/91.01±0.04	92.33±0.04/92.01±0.04	96.30 ± 0.04/96.11±0.03	19.89±0.13/19.12±0.08	7.57±0.04/7.96±0.05	95.39±0.01	83.19±0.07	**80.96±0.12**	85.87±0.17
+Auto-Debias	92.05±0.03/91.18±0.03	93.01±0.03/92.52±0.03	96.11±0.06/96.00±0.04	20.99±0.14/20.92±0.11	7.39±0.05/8.24±0.06	95.38±0.02	82.82±0.10	80.00±0.15	86.39±0.08
+MABEL	92.05±0.03/91.20±0.02	92.95±0.04/93.00±0.05	96.22±0.03/95.95±0.06	21.02±0.13/20.14±0.07	7.61±0.04/8.13±0.03	95.44 ± 0.01	82.90±0.08	79.85±0.16	86.85±0.06
+CCPA	92.06±0.02/91.20±0.05	93.00±0.02/92.65±0.04	96.21±0.04/96.02±0.03	20.47±0.11/20.48±0.09	6.82±0.06/8.07±0.08	95.39±0.02	82.98±0.07	80.37±0.14	86.21±0.09
+BNS	92.02±0.03/91.20±0.03	93.08±0.04/92.57±0.04	96.20±0.05/96.17±0.03	19.65 ± 0.15/19.04 ± 0.14	6.66 ± 0.06/7.99 ± 0.07	95.40±0.02	82.90±0.08	80.43±0.14	86.88 ± 0.07
+DeepSoftDebias	92.07 ± 0.02/91.35 ± 0.03	93.11 ± 0.05/93.04 ± 0.04	96.10±0.04/96.07±0.05	20.11±0.12/19.79±0.11	7.33±0.04/8.27±0.06	95.41±0.03	82.89±0.06	80.56±0.12	86.70±0.05
+BATED	**92.11±0.03*/91.41±0.03***	**93.40±0.03*/93.13±0.04***	**96.33±0.05*/96.35±0.04***	**17.55±0.10*/18.24±0.09***	**5.87±0.05*/7.92±0.05***	**95.47±0.02**	**83.28±0.06**	80.92 ± 0.13	86.14±0.10
GPT-2	89.39±0.05/88.99±0.05	90.01±0.06/89.26±0.05	94.72±0.02/95.04±0.03	19.93±0.23/20.38±0.18	8.63±0.16/8.40±0.20	**95.02±0.02**	**80.19±0.03**	75.88±0.07	86.81±0.05
+INLP	89.25±0.06/88.95±0.06	90.15±0.05/89.12±0.06	94.65±0.03/95.10±0.04	19.45±0.25/20.85±0.20	8.78±0.15/8.55±0.18	94.99±0.03	79.85±0.05	75.12±0.09	87.05 ± 0.06
+Sent-Debias	89.52±0.05/89.10±0.05	90.03±0.06/89.35±0.05	94.85±0.04/95.15±0.03	21.25±0.22/21.05±0.17	8.35±0.17/8.65±0.19	94.96±0.03	80.05±0.04	76.15±0.08	86.45±0.08
+FairFil	89.43±0.06/88.85±0.06	89.75±0.07/88.95±0.06	94.90±0.03/95.08±0.04	19.55±0.26/19.25±0.21	8.45±0.16/8.35±0.17	94.90±0.02	80.10 ± 0.05	76.25 ± 0.10	86.20±0.10
+Auto-Debias	89.50±0.05/89.05±0.05	90.20±0.05/89.40±0.05	94.80±0.05/95.07±0.04	20.35±0.24/20.65±0.19	8.15±0.18/8.50±0.20	94.88±0.03	79.95±0.06	75.75±0.09	86.75±0.07
+MABEL	-	-	-	-	-	-	-	-	-
+CCPA	89.55±0.05/89.08±0.06	90.18±0.05/89.52±0.05	94.83±0.04/95.09±0.04	19.85±0.23/20.15±0.18	7.95±0.17/8.25±0.19	94.94±0.04	80.05±0.05	75.95±0.08	86.50±0.08
+BNS	89.48±0.05/89.20 ± 0.05	90.25±0.06/89.65 ± 0.05	94.82 ± 0.05/95.13 ± 0.04	18.20 ± 0.27/19.15±0.22	7.85 ± 0.17/8.15 ± 0.18	95.00±0.02	80.00±0.06	76.10±0.08	87.00±0.06
+DeepSoftDebias	89.58 ± 0.04/89.15±0.05	90.30 ± 0.04/89.60±0.05	94.75±0.05/95.05±0.05	18.75±0.25/18.95 ± 0.20	8.05±0.16/8.45±0.19	95.01 ± 0.03	79.98±0.05	76.20±0.08	86.95±0.06
+BATED	**89.72±0.04*/89.46±0.05***	**90.38±0.03*/89.98±0.03***	**94.79±0.05*/95.11±0.05***	**16.12±0.18*/18.02±0.14***	**6.98±0.12*/6.45±0.13***	95.00±0.02	80.09±0.03	**76.48±0.05**	**87.08±0.05**
LLaMA-3.2	92.39±0.04/91.55 ± 0.05	92.91±0.05/92.40±0.04	96.22±0.03/96.21±0.03	21.57±0.20/21.36±0.18	8.34±0.14/7.83±0.13	95.58±0.02	83.11±0.05	79.51±0.08	88.02 ± 0.06
+INLP	92.25±0.05/91.44±0.05	92.70±0.06/92.15±0.05	96.15±0.04/96.10±0.04	21.12±0.21/21.08±0.20	7.55±0.13/7.20±0.14	95.36±0.02	82.95±0.06	79.33±0.08	**88.15±0.07**
+Sent-Debias	92.31±0.05/91.50±0.05	92.83±0.05/92.25±0.05	96.18±0.03/96.13±0.03	21.78±0.20/21.45±0.18	7.12±0.15/6.98±0.14	95.47±0.02	83.05±0.05	80.65 ± 0.09	86.85±0.07
+FairFil	92.29±0.04/91.42±0.05	92.65±0.06/92.10±0.06	96.20±0.04/96.17±0.03	21.44±0.21/21.10±0.20	6.74±0.14/6.50 ± 0.13	95.35±0.02	83.00±0.05	79.58±0.08	86.90±0.06
+Auto-Debias	92.35±0.05/91.48±0.05	92.89±0.05/92.33±0.05	96.21±0.04/96.20±0.03	21.26±0.19/21.05±0.17	7.05±0.14/6.75±0.13	95.45±0.02	83.08±0.06	80.49±0.08	87.00±0.06
+MABEL	-	-	-	-	-	-	-	-	-
+CCPA	92.36±0.05/91.53±0.05	92.85±0.05/92.28±0.05	96.19±0.03/96.18±0.03	21.30±0.20/21.20±0.19	6.32 ± 0.14/6.60±0.13	95.56±0.04	83.06±0.05	79.46±0.08	86.98±0.07
+BNS	92.38±0.05/91.51±0.05	92.97 ± 0.05/92.72 ± 0.05	96.23±0.04/96.19±0.04	20.15 ± 0.21/20.00 ± 0.19	7.94±0.15/7.65±0.14	95.53±0.03	83.10±0.06	79.61±0.08	87.01±0.06
+DeepSoftDebias	92.40 ± 0.04/91.54±0.05	92.89±0.05/92.66±0.05	96.24 ± 0.03/96.22 ± 0.03	20.18±0.20/20.12±0.18	7.85±0.14/7.55±0.13	95.54±0.03	83.12 ± 0.05	79.66±0.07	87.43±0.06
+BATED	**92.71±0.03*/91.94±0.04***	**93.97±0.03*/93.42±0.03***	**96.81±0.03*/96.74±0.03***	**19.47±0.19*/19.66±0.18***	**4.69±0.11*/5.37±0.12***	95.56 ± 0.02	**83.63±0.04**	**82.43±0.06**	86.95±0.07
DistilBERT	91.71±0.04/90.88±0.03	93.00±0.02/92.65±0.04	95.81±0.08/95.88±0.06	21.95±0.17/21.19±0.07	7.89±0.19/8.73±0.06	95.41±0.02	**82.76±0.05**	**79.66±0.11**	86.80±0.21
+BATED	**91.95±0.03/90.96±0.03**	**93.12±0.03/92.91±0.05**	**96.29±0.07/96.10±0.05**	**19.39±0.14/19.51±0.04**	**6.46±0.11/7.88±0.06**	**95.43±0.02**	82.29±0.04	78.44±0.13	**87.77±0.22**
ELECTRA	91.52±0.05/90.62±0.07	92.64±0.08/92.27±0.05	95.85±0.08/95.68±0.03	21.81±0.21/21.07±0.17	8.45±0.12/9.25±0.15	95.35±0.01	**82.66±0.10**	**79.81±0.25**	86.29±0.36
+BATED	**91.80±0.04/91.01±0.06**	**93.22±0.08/92.85±0.06**	**96.13±0.06/95.99±0.04**	**19.61±0.18/19.13±0.14**	**7.48±0.15/8.86±0.10**	**95.38±0.02**	82.30±0.11	78.71±0.22	**87.22±0.35**
OPT	91.22±0.01/90.75±0.04	92.21±0.05/91.59±0.13	95.80±0.09/96.08±0.06	21.53±0.33/20.35±0.28	8.33±0.27/9.12±0.31	95.35±0.03	82.19±0.05	78.63±0.18	87.08±0.22
+BATED	**91.24±0.01/90.93±0.03**	**93.17±0.04/92.35±0.12**	**95.87±0.08/96.13±0.07**	**19.41±0.21/18.78±0.19**	**7.71±0.29/8.11±0.38**	**95.35±0.02**	**82.65±0.06**	**80.62±0.16**	**87.32±0.25**
QWEN-2.5	91.07±0.04/90.31±0.05	91.41±0.05/90.73±0.05	95.98±0.04/96.01±0.03	20.85±0.22/21.46±0.20	7.11±0.14/7.50±0.16	**95.37±0.02**	82.45±0.06	79.15±0.09	**86.86±0.06**
+BATED	**91.88±0.03/91.12±0.04**	**92.45±0.03/91.85±0.04**	**96.37±0.03/96.30±0.03**	**19.15±0.20/20.09±0.18**	**5.22±0.12/5.61±0.13**	95.23±0.02	**82.62±0.05**	**80.43±0.07**	85.25±0.07

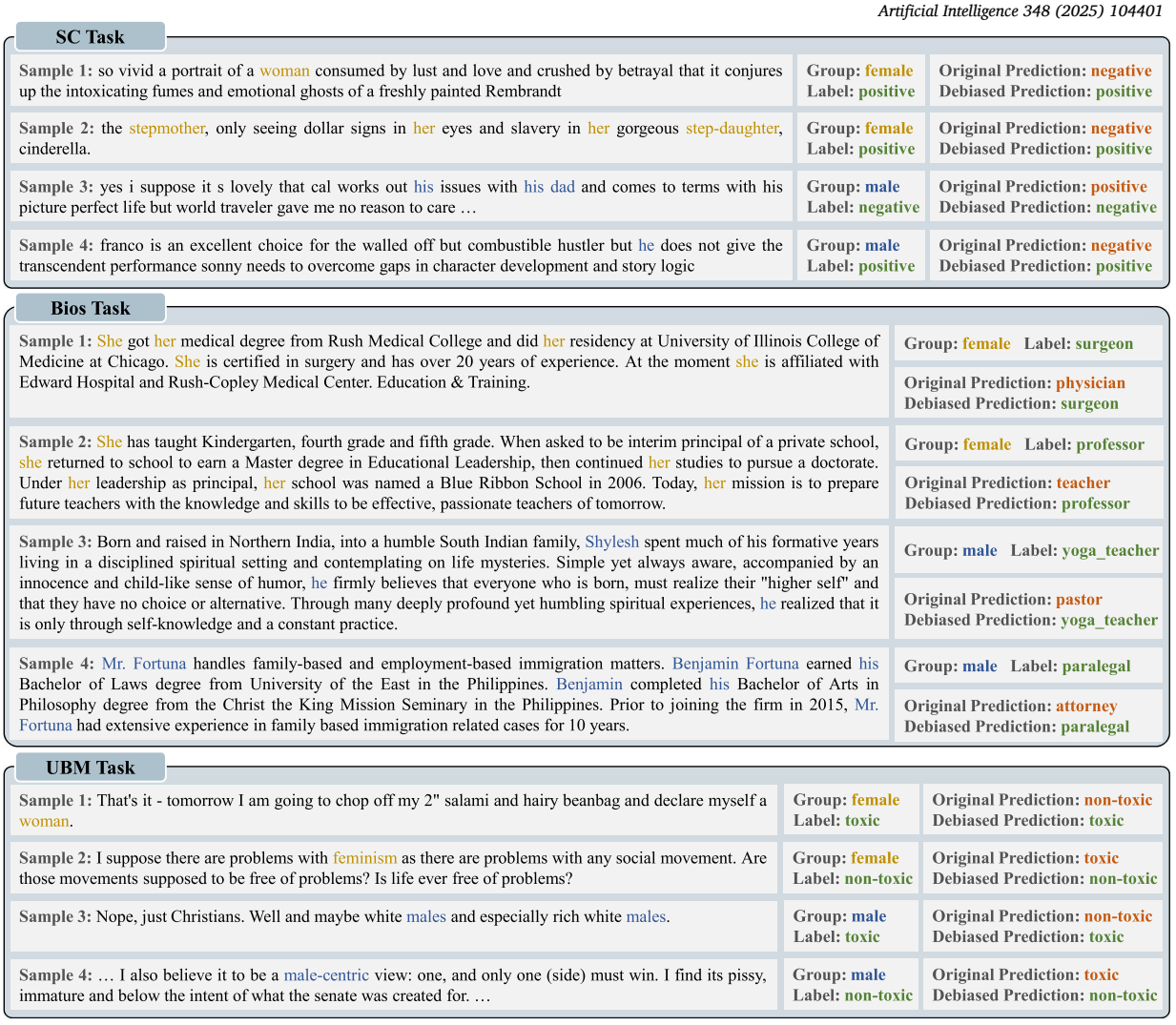

Fig. 2. Example illustration of BERT's original and debiased prediction results on three downstream tasks.

In the SC task, gender cues such as "*woman*", "*stepmother*", or "*his dad*" led the original model to biased sentiment predictions, for instance, associating female characters with negative sentiment. After applying BATED, the model's predictions align more closely with the semantic content, demonstrating improved fairness. In the Bios task, gender markers like "*she*" and "*Mr. Fortuna*" skew the original model toward stereotypical associations, classifying women as "*physician*" instead of "*surgeon*", or men as "*attorney*" rather than "*paralegal*". With BATED, such biased mappings are corrected, reflecting a more equitable recognition of professional roles regardless of gender. In the UBM task, gender terms often trigger toxic predictions from the original model. For example, sentences mentioning "*woman*" or "*feminism*" are misclassified as toxic even in neutral or non-hostile contexts. BATED mitigates these effects, reducing false positives while preserving the original semantic intent. Together, these results not only support BATED's quantitative performance improvements, but also underscore its real-world effectiveness in reducing gender bias and enhancing model fairness and reliability.

Overall, the experiments of applying BATED to debiasing seven PLMs on three downstream tasks demonstrate that it can effectively mitigate bias in downstream tasks. BATED can achieve a trade-off between fairness and accuracy, which can improve the fairness of PLM's decision without affecting the modeling abilities of PLMs on downstream tasks. Thus, the first research question **RQ1** is answered.

5.4. Analysis of biased teacher model

To answer research question **RQ2**, We evaluate the intrinsic biases of the trained biased teacher models using the four intrinsic bias metrics introduced in Section 5.2.1. The experimental results of seven PLMs are shown in Table 6. We conclude with the following observations.

Table 6
Intrinsic bias test results for the biased teacher models on SEAT, CrowS-Pairs, StereoSet, and GBES. SEAT is reported as absolute values. ◊: the closer to 50, the less bias. More biased results are shown in **bold**.

Metric / Model	SEAT							CrowS-Pairs◊		SS Score◊	GBES ↑
	6 ↓	6b ↓	7 ↓	7b ↓	8 ↓	8b ↓	Avg. ↓	Stereo	Anti-Stereo		
BERT	0.477	0.108	0.253	0.254	0.399	**0.636**	0.354	57.86	56.31	60.28	0.961
$BERT_{bias}$	**1.917**	**1.992**	**0.355**	**0.349**	**0.433**	0.357	**0.900**	**38.36**	**75.73**	**61.36**	**-0.874**
DistilBERT	1.380	0.446	0.179	**1.242**	**0.837**	**1.217**	0.883	59.75	52.43	60.63	0.970
$DistilBERT_{bias}$	**1.991**	**1.993**	**0.451**	0.541	0.517	0.499	**0.999**	**37.74**	**70.87**	**65.49**	**-0.865**
ELECTRA	0.844	0.298	0.506	0.830	0.323	**1.206**	0.668	48.43	51.46	54.99	0.992
$ELECTRA_{bias}$	**1.976**	**1.981**	**0.825**	**0.896**	**0.588**	0.575	**1.140**	**43.40**	**59.22**	**56.26**	**-0.437**
OPT	1.216	0.218	0.134	**1.255**	0.335	**0.993**	0.692	**64.14**	56.31	66.24	0.981
OPT_{bias}	**1.547**	**0.988**	**0.963**	0.990	**0.598**	0.680	**0.961**	44.65	**68.93**	**67.32**	**-0.126**
GPT-2	1.080	0.304	0.559	0.941	0.473	1.202	0.759	**65.41**	51.46	**68.70**	0.976
$GPT\text{-}2_{bias}$	**1.163**	**0.434**	**1.090**	**1.286**	**0.668**	**1.336**	**0.996**	58.49	**59.22**	68.26	**0.590**
QWEN-2.5	0.448	0.216	0.452	0.560	0.412	0.573	0.444	56.60	57.28	69.02	0.998
$QWEN\text{-}2.5_{bias}$	**0.532**	**0.248**	**0.564**	**0.623**	**0.446**	**0.646**	**0.510**	**59.75**	57.28	**69.81**	**0.586**
LLaMA-3.2	1.343	0.374	0.436	1.529	0.905	1.430	1.003	66.67	54.37	71.08	0.988
$LLaMA\text{-}3.2_{bias}$	**1.472**	**0.434**	**0.465**	**1.591**	**0.940**	**1.441**	**1.057**	**67.30**	**56.37**	**71.44**	**0.461**

❶ *The biased teacher model successfully learns generic bias information.* Observing the results of the four intrinsic bias metrics of Table 6, it can be found that the biased teacher model captures more bias information than the original PLM, for all seven PLMs. From the results of SEAT, the biased teacher models obtain stronger bias than the original PLMs on most of the subsets, and their average effect sizes are both significantly elevated, which represents more intrinsic bias. Similarly, CrowS-Pairs, StereoSet, and GBES also verify the success of the training of the teacher models. In particular, $BERT_{bias}$, $DistilBERT_{bias}$, $ELECTRA_{bias}$, $QWEN\text{-}2.5_{bias}$, and $LLaMA\text{-}3.2_{bias}$ achieve more biased scores than the original BERT, DistilBERT, ELECTRA, QWEN-2.5, and LLaMA-3.2 on all metrics. Although OPT_{bias} and $GPT\text{-}2_{bias}$ do not have a significant increase in bias due to the high bias of the original OPT and GPT-2, it can still be concluded that they successfully capture the bias information when combined with the three intrinsic metrics. Therefore, the experimental results of testing biased teacher models show that the proposed strategy of leveraging causal contrastive learning can train PLMs to successfully capture social biases in representations.

❷ *It is effective to utilize the biased teacher model to guide the disentanglement of biased representations and fair representations.* In combination with Table 6 and Table 3 to Table 5, Table 6 indicates that the teacher models are trained to have generic bias information, and Table 3 to Table 5 indicate that BATED is able to significantly debias PLMs in downstream tasks. By jointly analyzing the test results of the debiasing experiment and the biased teacher model, it can be concluded that the biased teacher model can effectively guide VAE to disentangle the fair representations and the biased representations, so as to improve the decision-making fairness of PLMs.

Overall, by evaluating the intrinsic biases of the trained biased teacher models, we verify that the proposed training strategy for causal contrastive learning enables the model to capture generic gender bias information in the training corpus. The learned bias information can further guide the biased representations to fit so that they can be distilled from the original representations to obtain an unbiased and fair representations. This addresses the second research question **RQ2** raised at the beginning of this section.

5.5. Analysis of modules in BATED

To answer research question **RQ3**, we conduct ablation analysis and hyper-parameter analysis to comprehensively explore the role of each module in BATED on model performance.

5.5.1. Ablation analysis

We conduct a set of ablation experiments on three downstream tasks, using BERT and LLaMA-3.2 as base models, respectively. Corresponding to each module of BATED, we employ three variants of BATED as follows:

- $BATED_{nT}$ represents that there is no biased teacher model guiding the distillation of biased representations during debiasing training, but only VAE is adopted for disentanglement.
- $BATED_{nVAE}$ represents using two linear transformations instead of VAE to predict the fair representations and biased representations.
- $BATED_{rand}$ represents that the teacher model is trained using random masked tokens instead of causal contrastive learning.

Similar to Section 5.3, we adopt two extrinsic bias metrics to evaluate the model's decision fairness on downstream tasks, where TPR_M, TPR_F, TPR_{GAP}, and TPR_{RMS} are used to evaluate SC task and Bios task, and AUC_{sub}, AUC_{bpsn}, and AUC_{bnsp}, AEG_{neg} and

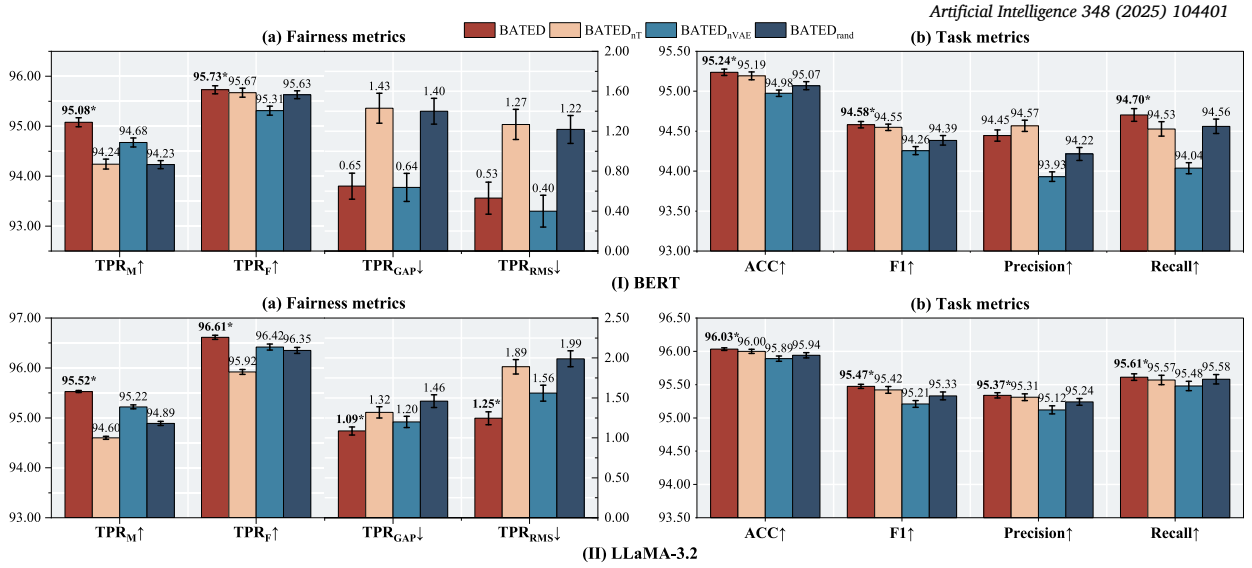

Fig. 3. Ablation results on the SC task. Data in **bold** with '*' represent that BATED is statistically significant compared to the other data (base on Mann-Whitney U test; $\rho < 0.05$).

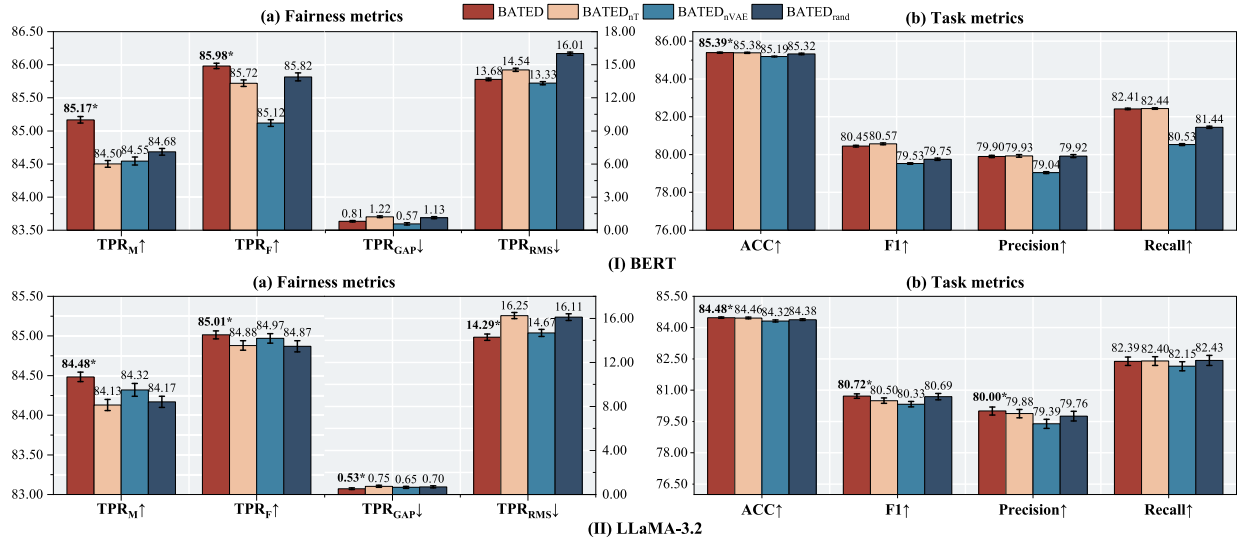

Fig. 4. Ablation results on the Bios task. Data in **bold** with '*' represent that BATED is statistically significant compared to the other data (base on Mann-Whitney U test; $\rho < 0.05$).

AEG$_{pos}$ are used to evaluate UBM task. And four task metrics Acc, F1, Precision, and Recall are used to evaluate the task performance of the model. The results of ablation experiments are shown in Figs. 3, 4, and 5, and we make a few main observations.

❶ *Each module contributes to the debiasing performance of BATED.* The results presented in Figs. 3, 4, and 5 demonstrate that removing any individual component leads to a noticeable degradation in decision fairness across all downstream tasks. Specifically, for each fairness metric, the complete BATED model consistently achieves the lowest bias scores, whereas its ablated variants BATED$_{nT}$, BATED$_{nVAE}$, and BATED$_{rand}$ exhibit significantly more biased results. These results indicate that both the use of a biased teacher model and the VAE for representation disentanglement, as well as the application of causal contrastive learning to train the biased teacher model, contribute significantly to BATED's overall debiasing capability.

❷ *The biased teacher model has a severe impact on debiasing performance and a slight impact on task performance.* Among the three variants, it can be found that BATED$_{nT}$ has the most obvious decrease in fairness, while it is basically constant in accuracy. The reason is that BATED$_{nT}$ fails to disentangle the fair representation and biased representation due to not using a biased teacher model, while VAE ensures the learning ability of PLM in the tasks. Therefore, BATED$_{nT}$ achieves the same task accuracy as BATED but the bias is not mitigated. Similarly, BATED$_{rand}$, which utilizes a randomly trained teacher model without any bias-specific training, exhibits debiasing performance comparable to BATED$_{nT}$, while maintaining similar task accuracy. This similarity highlights a critical insight: simply introducing a teacher model without explicitly encoding bias-related knowledge is insufficient for effective disentanglement.

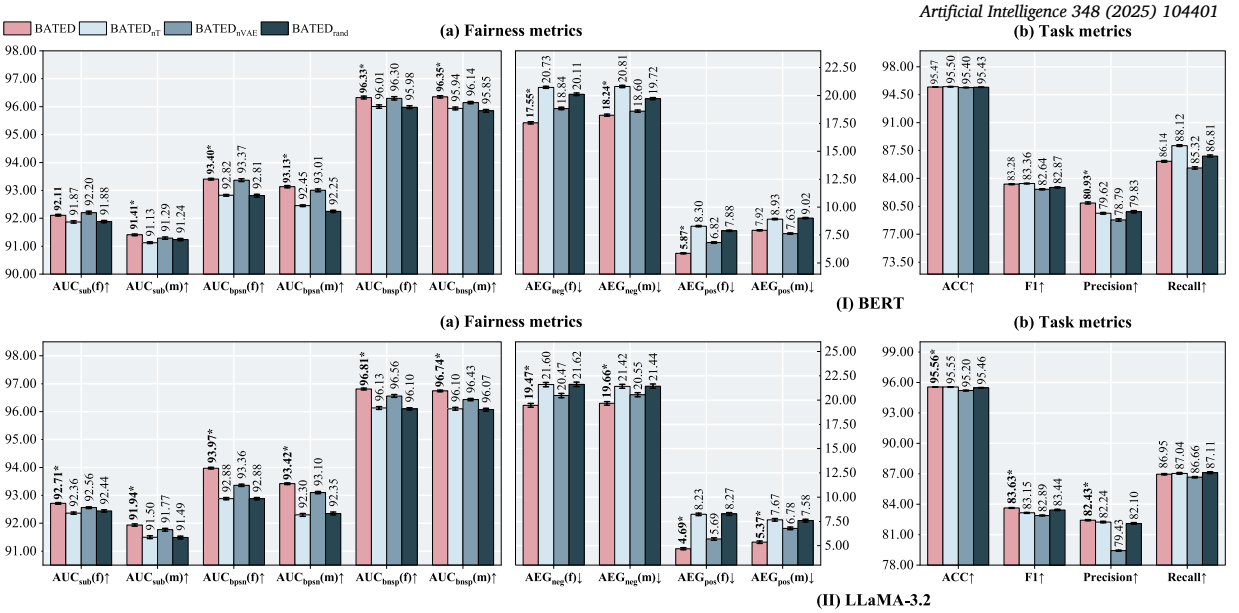

Fig. 5. Ablation results on the UBM task. Data in **bold** with '*' represent that BATED is statistically significant compared to the other data (base on Mann-Whitney U test; $\rho < 0.05$).

The random teacher lacks structured inductive bias and meaningful semantic alignment with the actual bias patterns present in the data, thereby failing to provide useful gradient signals during the distillation of biased representations. In BATED, the teacher model is trained with causal contrastive learning, which explicitly encourages the model to capture and emphasize the causal effect of protected attributes on predictions. This targeted training guides the student model to extract biased components that correlate with sensitive attributes. In contrast, the random teacher has no such structure or objective, its representations are essentially noise from the perspective of the bias dimensions. As a result, the student model receives weak or misleading supervision during the distillation phase, impairing its ability to isolate and neutralize bias-related features. Therefore, the performance of BATED$_{rand}$ reinforces the necessity of training the biased teacher model with a principled objective, such as causal contrastive learning. Doing so ensures that the teacher captures the systematic influence of biases, enabling the overall framework to effectively disentangle and mitigate those biases in the final representations. By explicitly modeling biases, the teacher model plays a critical role in promoting meaningful representation disentanglement and improving fairness outcomes.

❸ *The VAE plays a crucial role in supporting the task performance of BATED while also contributing to its debiasing capability.* Compared to BATED$_{nT}$, BATED$_{nVAE}$ exhibits the most significant decline in task-related metrics, along with a noticeable drop in fairness performance. This indicates that although BATED$_{nVAE}$ is able to mitigate bias to some extent, presumably due to the guidance from the biased teacher model, it struggles to maintain strong language modeling ability in downstream tasks in the absence of the VAE component. The underlying reason lies in the VAE's role in structuring the latent space. The VAE facilitates the disentanglement of fair and biased representations by enforcing a probabilistic regularization and reconstruction constraint. This not only aids in separating sensitive information from task-relevant features but also preserves the semantic coherence necessary for effective downstream task performance. Without this constraint, BATED$_{nVAE}$ lacks a principled mechanism to organize latent representations, which results in less informative and less generalizable features for downstream classifiers. Furthermore, although the biased teacher can still guide the extraction of biased components in BATED$_{nVAE}$, the absence of the VAE leads to poor alignment between the fair and biased subspaces. This misalignment weakens the model's ability to isolate bias accurately and utilize fair representations effectively for predictions. Consequently, BATED$_{nVAE}$ not only fails to achieve optimal fairness but also suffers from degraded task performance due to the collapse of meaningful feature learning. In summary, the VAE serves as a foundational component in BATED, ensuring effective bias disentanglement and robust representation learning. Its removal leads to a failure in performance, highlighting its essential role in balancing fairness and task effectiveness.

In conclusion, for BATED, the biased teacher model ensures that the model can mitigate the bias, and the VAE constraint model obtains stronger language modeling ability to assist in further improving the fairness.

5.5.2. Hyper-parameter analysis

To further explore the contribution of each module to BATED in more depth, we conduct hyper-parameter experiments to analyze how the values of the two hyper-parameters α and γ of the $\mathcal{L}oss$ (Eq. (7)) affect the model performance. We take the value of α from [1, 0.5, 0.2, 0.1, 0.05] and the value of γ from [1, 0.5, 0.2, 0.05, 0.02, 0.005, 0.002], and employ BATED trained by the loss of each hyper-parameter combination to debias BERT and LLaMA-3.2 in downstream tasks. Figs. 6, 7, and 8 correspond to SC tasks, Bios tasks, and UBM tasks, respectively. We choose to report TPR$_{GAP}$ and TPR$_{RMS}$ to show the fairness of the SC task and Bios task, average AUC$_{AVG}$ of AUC$_{sub}$, AUC$_{bpsn}$ and AUC$_{bnsp}$ and average AEG$_{AVG}$ of AEG$_{neg}$ and AEG$_{pos}$ to show the fairness of the

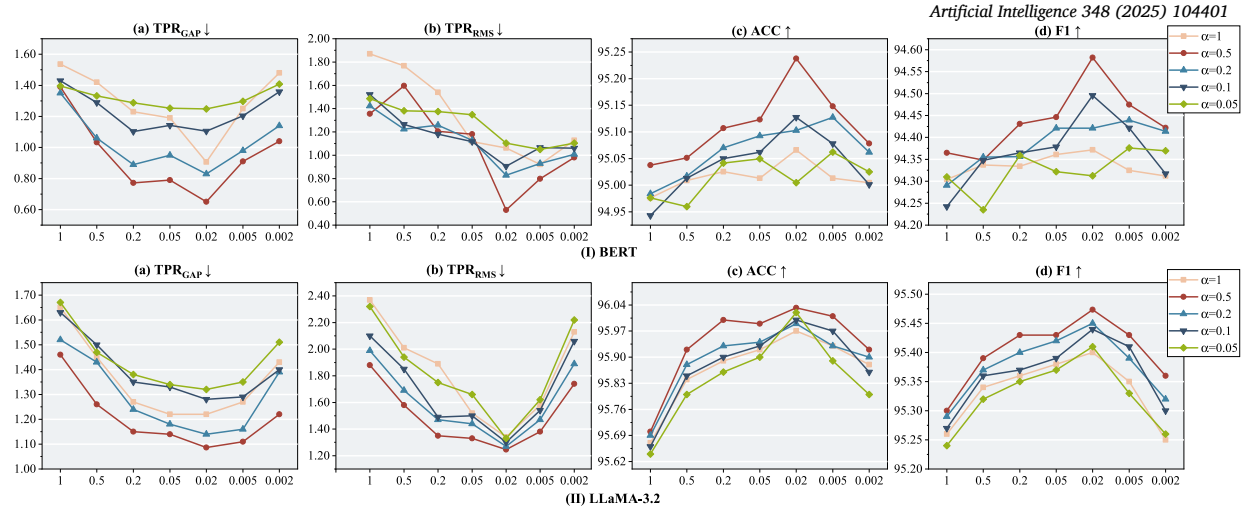

Fig. 6. Experimental results for hyper-parameter analysis on the SC task. The horizontal axis is the value of the hyper-parameter γ, and the vertical axis is the value of the hyper-parameter α. Subfigures (a) and (b) represent fairness metrics, Subfigures (c) and (d) represent task metrics.

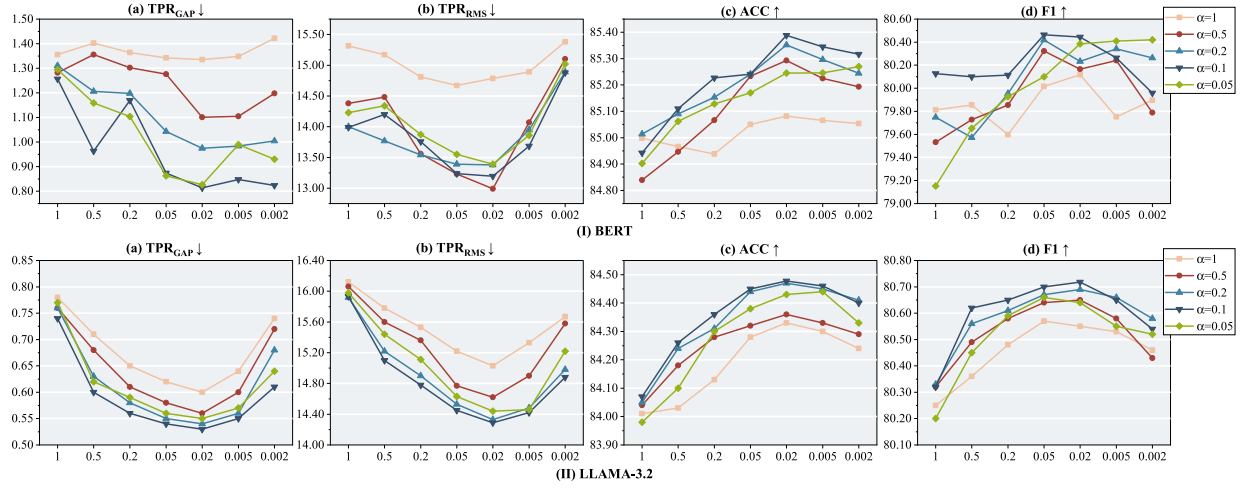

Fig. 7. Experimental results for hyper-parameter analysis on the Bios task. The horizontal axis is the value of the hyper-parameter γ, and the vertical axis is the value of the hyper-parameter α. Subfigures (a) and (b) represent fairness metrics, Subfigures (c) and (d) represent task metrics.

UBM task. In addition, when the hyper-parameters α and γ are fixed to the optimal values, respectively, the change in debiasing performance caused by the value of another hyper-parameter is shown in Figs. 9 and 10. Here, TPR_{M} and TPR_{F} are reported, and the gap between the two represents TPR_{GAP}. Since the TPR metric is not adopted for the UBM task, we only report the results for the SC task and Bios task. By observing Figs. 6 to 10, we obtain the following findings.

❶ *Balancing different modules is conducive to improve the fairness and accuracy of BATED simultaneously.* In the overall objective function $\mathcal{L}oss$, the hyper-parameter α controls the influence of the biased teacher model in guiding the distillation of biased representations, while γ regulates the strength of the VAE in disentangling fair and biased representations. In the case where the task training objective is dominant, the weights between the two modules need to be reasonably balanced. Figs. 6, 7, and 8 demonstrate the trends of fairness and accuracy of the BATED under different combinations of α and γ for each downstream task. By observation and summary, it can be found that the optimal combination of hyper-parameters for each task is $\alpha = 0.5$, $\gamma = 0.02$ for SC task, $\alpha = 0.1$, $\gamma = 0.02$ for Bios task, and $\alpha = 0.1$, $\gamma = 0.005$ for UBM task, as shown in Table 1. Under these settings, BATED achieves both reduced bias and improved task performance. In contrast, while alternative combinations still outperform the original BERT model, they fall short of achieving optimal fairness and accuracy simultaneously. These findings highlight the importance of appropriately balancing the module contributions to maximize both debiasing effectiveness and task performance in BATED.

❷ *Overly strong or weak distillation of biased representations can limit both the debiasing effectiveness and overall task performance of BATED.* It can be seen from Figs. 6, 7, and 8 that for each value of γ, BATED generally obtains the worst scores for both the fairness metrics and the task metrics for each task when $\alpha = 1$. In addition, Figs. 9(a) and 10(a) also shows that TPR_{M}, TPR_{F}, and the difference between them, TPR_{GAP}, all have the worst scores when γ takes the optimal value. This suggests that distilling biased representations too strongly does not lead to performance gains. On the one hand, forcibly distilling the biased representation to fit the biased

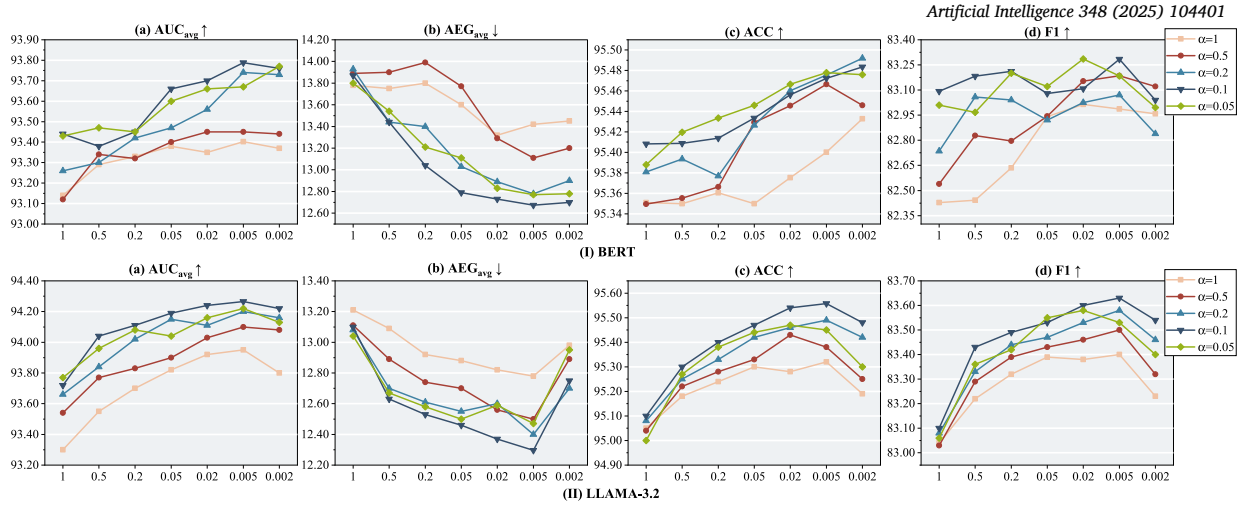

Fig. 8. Experimental results for hyper-parameter analysis on the UBM task. The horizontal axis is the value of the hyper-parameter γ, and the vertical axis is the value of the hyper-parameter α. Subfigures (a) and (b) represent fairness metrics, Subfigures (c) and (d) represent task metrics.

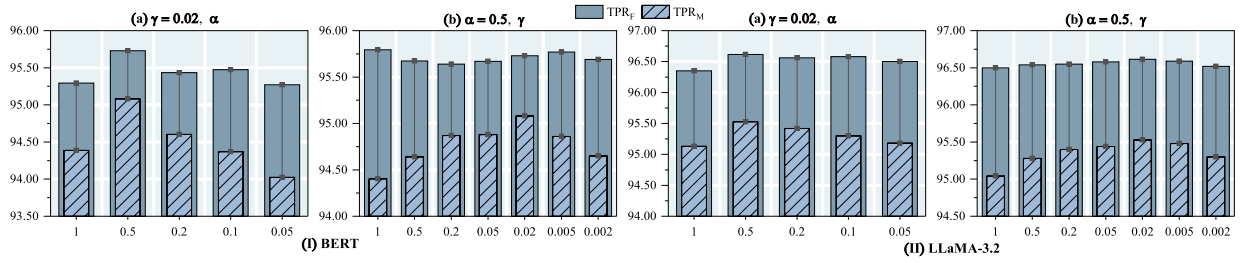

Fig. 9. Results of hyper-parameter experiments on the SC task. Subfigure (a) shows the impact of hyper-parameter α on fairness when γ is fixed to the optimal value of 0.02. Subfigure (b) shows the impact of hyper-parameter γ on fairness when hyper-parameter α is fixed to the optimal value of 0.5. The gap between TPR_F and TPR_M represents the TPR_{GAP}.

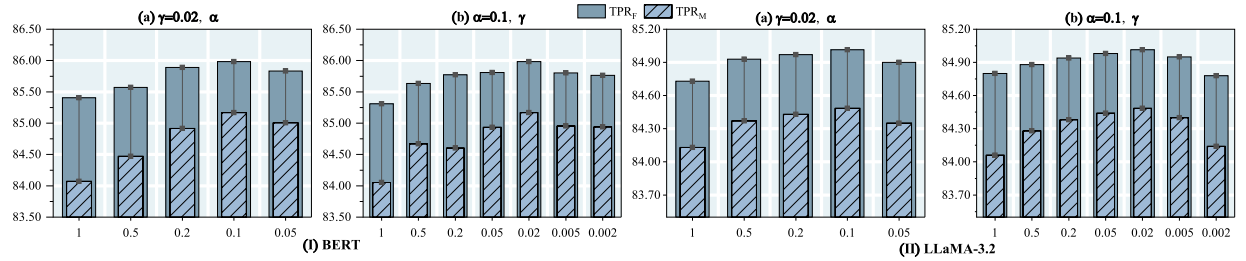

Fig. 10. Results of hyper-parameter experiments on the Bios task. Subfigure (a) shows the impact of hyper-parameter α on fairness when γ is fixed to the optimal value of 0.02. Subfigure (b) shows the impact of hyper-parameter γ on fairness when hyper-parameter α is fixed to the optimal value of 0.1. The gap between TPR_F and TPR_M represents the TPR_{GAP}.

teacher model may cause the biased representation to learn redundant knowledge and introduce additional useless information. On the other hand, the fast fitting of the biased representation will affect the disentanglement process of VAE and indirectly damage the convergence process of the model. As α decreases, both fairness and accuracy improve, but this improvement is not unbounded. Each task shows an inflection point beyond which further reduction in α leads to diminishing or negative returns. When α is too small, the distillation signal becomes too weak, impairing the learning of biased representations and ultimately hindering BATED's overall performance.

❸ *It is able to maximize the performance of BATED by reasonably constraining the disentanglement between fair and biased representations.* The value of the hyper-parameter γ represents the weight of the disentanglement objective of the VAE in the overall training objective. The horizontal axis of each subfigure of Figs. 6 to 8 embodies the trend of the performance of BATED as γ changes. Among them, as γ decreases, the fairness metrics of SC and Bios tasks show a trend of first decreasing, achieving the lowest bias at the best value of γ, and then increasing. In the task metrics, it shows a trend of first increasing, achieving the highest accuracy at the best γ value, and then decreasing. For the UBM task, the fairness and accuracy also show a similar trend. These results show that when VAE is used to

Table 7
Debiasing results on the Amazon task. All models are pre-trained on the SC task and then tested on the Amazon task. The best result is indicated in **bold**. The suboptimal result is indicated in underline. * represent statistically significant compared to other results (base on Mann-Whitney U test; $\rho < 0.05$).

Metric / Model	Fairness metrics (%)				Task metrics (%)			
	TPR_M↑	TPR_F↑	TPR_{GAP}↓	TPR_{RMS}↓	Acc.↑	F1↑	Precision↑	Recall↑
BERT	81.54±0.15	83.12±0.18	1.58±0.12	2.00±0.20	82.24±0.10	**71.18±0.25**	68.28±0.30	**81.72±0.12**
+INLP	81.92±0.12	83.25±0.15	1.33±0.10	1.88±0.18	82.30±0.08	71.05 ± 0.22	68.45±0.25	81.68 ± 0.10
+Sent-Debias	82.15±0.10	83.51±0.14	1.36±0.09	1.70±0.16	82.32±0.07	70.95±0.20	68.66±0.22	81.65±0.09
+FairFil	82.08±0.11	83.35±0.15	1.27 ± 0.09	1.71±0.17	82.28±0.08	71.00±0.21	68.53±0.24	81.67±0.10
+Auto-Debias	82.12±0.09	83.45±0.13	1.33±0.08	1.69±0.15	82.33±0.07	70.90±0.19	68.67±0.21	81.64±0.08
+MABEL	82.25±0.08	83.53±0.12	1.28±0.08	1.60±0.14	82.34±0.06	70.88±0.18	68.71±0.20	81.63±0.08
+CCPA	82.30±0.07	83.65±0.11	1.35±0.07	1.58±0.13	82.35 ± 0.06	70.85±0.17	68.75±0.19	81.62±0.07
+BNS	82.20±0.06	83.55±0.10	1.35±0.07	1.53±0.12	82.33±0.05	70.80±0.16	68.86 ± 0.18	81.61±0.07
+DeepSoftDebias	82.42 ± 0.05	83.83 ± 0.09	1.41±0.06	1.45 ± 0.11	82.31±0.05	70.75±0.15	68.83±0.17	81.60±0.06
+BATED	**82.66±0.04***	**83.91±0.08***	**1.25±0.05***	**1.34±0.10***	**82.35±0.04**	70.98±0.14	**69.03±0.15**	81.66±0.05
LLaMA-3.2	83.08±0.14	85.49±0.16	2.41±0.15	3.99±0.22	84.15±0.12	**73.41±0.24**	69.76±0.28	88.58 ± 0.10
+INLP	83.35±0.12	85.60±0.14	2.25±0.13	3.75±0.20	84.20±0.10	73.30 ± 0.22	70.05±0.25	88.45±0.09
+Sent-Debias	83.58±0.10	85.75±0.13	2.17±0.12	3.50±0.18	84.25±0.09	73.20±0.20	70.30±0.23	88.35±0.08
+FairFil	83.62±0.11	85.78 ± 0.13	2.16±0.12	3.45±0.19	84.23±0.09	73.18±0.21	70.35±0.24	88.32±0.08
+Auto-Debias	83.75±0.09	85.75±0.12	2.00±0.11	3.30±0.17	84.28±0.08	73.15±0.19	70.45±0.22	88.28±0.07
+MABEL	83.82±0.08	85.70±0.11	1.88±0.10	3.20±0.16	84.30±0.07	73.10±0.18	70.55±0.21	88.25±0.07
+CCPA	83.90±0.07	85.60±0.10	1.70±0.09	3.10±0.15	84.32±0.07	73.05±0.17	70.65±0.20	**88.60±0.06**
+BNS	83.98±0.06	85.66±0.09	1.68±0.08	3.00±0.14	84.34 ± 0.06	73.00±0.16	70.75±0.19	88.15±0.06
+DeepSoftDebias	84.10 ± 0.05	85.77±0.08	1.67 ± 0.07	2.80 ± 0.13	**84.36±0.05**	72.95±0.15	70.85 ± 0.18	88.10±0.05
+BATED	**84.22±0.04***	**85.81±0.07***	**1.59±0.05***	**2.27±0.10***	84.33±0.04	73.28±0.14	**71.09±0.15**	88.24±0.05

disentend the fair representations and the biased representations, an appropriate constrained disentanglement objective is beneficial to the effectiveness of debiasing training. We analyze this because excessive separation of biased information from the original representations will inevitably delete useful information, while too weak separation will lead to the legacy of bias. In addition, in Figs. 9(b) and 10(b), the variation of TPR_M, TPR_F, and TPR_{GAP} with the value of γ when α takes the optimal value further validates our analysis.

Through ablation experiments and hyperparameter experiments, we exhaustively explore and analyze each module of BATED. We find that the biased teacher model and VAE play a key role in both debiasing performance and task performance of BATED, which answers the third research question **RQ3** posed in this section.

5.6. Generalization performance analysis

To answer research question **RQ4**, we investigate the generalization performance of BATED on out-of-domain tasks. Specifically, we select an additional sentiment classification dataset, Amazon,[6] and directly apply the debiasing model trained on the SC task to this new domain to evaluate its generalization capabilities. The Amazon dataset consists of customer reviews, from which we extract only the samples labeled as positive or negative to ensure consistency with the binary classification setup of the SC task. After applying predefined sensitive attribute word filtering, we identified 5,727 gender-related samples, which are used in the testing experiment. We conduct testing experiments on BERT and LLaMA-3.2, while reporting the results of the generalization experiments for all baseline methods, as shown in Table 7. The conclusion of our observations is as follows.

❶ *BATED generalizes effectively to unseen domains while maintaining strong debiasing performance.* The proposed BATED framework demonstrates remarkable generalizability, achieving statistically significant improvements in fairness metrics across both model architectures. Specifically, when applied to BERT, BATED reduces the gender prediction disparity TPR_{GAP} from 1.58% to 1.25% (21% relative reduction) and decreases the root mean square bias TPR_{RMS} from 2.00% to 1.34% (33% relative reduction). Similar improvements are observed for LLaMA-3.2, with TPR_{GAP} decreasing from 2.41% to 1.59% (34% reduction) and TPR_{RMS} from 3.99% to 2.27% (43% reduction). These consistent improvements across metrics and model architectures suggest that BATED learns generalizable bias mitigation patterns rather than task-specific corrections.

❷ *BATED outperforms baselines in cross-domain fairness preservation.* Comparative analysis reveals BATED's superior generalization performance relative to existing debiasing approaches. While baseline methods exhibit varying degrees of performance degradation when applied to the out-of-domain Amazon task, BATED maintains its debiasing effectiveness. For instance, INLP shows limited transfer capability, achieving only marginal TPR_{GAP} reduction (1.33% vs BATED's 1.25% for BERT). DeepSoftDebias demonstrates partial generalization but with significant performance trade-offs (F1 score decrease of 0.43% compared to BATED). Other methods, such as Sent-Debias and FairFil, exhibit inconsistent performance across fairness metrics. This comparative advantage suggests that BATED's multi-component architecture is particularly effective at learning transferable debiasing transformations.

❸ *Task performance remains stable during cross-domain transfer.* BATED maintains strong task performance during domain transfer, with accuracy remaining within 0.01% of baseline performance for both BERT (82.35% vs 82.24%) and LLaMA-3.2 (84.33% vs

[6] https://www.kaggle.com/datasets/bittlingmayer/amazonreviews.

84.15%). This performance preservation, coupled with significant bias reduction, suggests that BATED's debiasing transformations selectively target bias-related features while preserving task-relevant information - a critical requirement for practical deployment.

❹ *BATED's scalability to larger models extends to cross-domain settings.* The framework demonstrates consistent generalization across model scales, effectively reducing bias in both the moderate-sized BERT architecture and the substantially larger LLaMA-3.2 model. This scalability is particularly noteworthy given the documented challenges in debiasing larger language models, suggesting that BATED's approach may be broadly applicable across the current generation of PLMs.

Overall, the experiments demonstrate that BATED effectively generalizes to the out-of-domain task while maintaining strong debiasing performance and preserving model accuracy. The framework significantly reduces gender bias without task-specific adaptation, outperforming all baselines. Thus, **RQ4** is answered affirmatively.

6. Conclusions and future work

In this paper, we propose a debiasing method to mitigate the social biases of PLMs in downstream tasks, which learns fair representation via the biased teacher-guided disentanglement. We employ VAE to disenttangle PLM-encoded representation into the fair representation and the biased representation, and leverage the biased teacher model to guide further decoupling of biased representation, while the fair representation is constrained by task loss. Furthermore, we design strategies that leverage causal contrastive learning to train model to capture general social biases. Our debiasing method does not need to obtain sensitive attribute information of task dataset in the debiasing phase, thus avoiding the limitations brought by privacy protection and regulation. BATED preserves PLM's language modeling capabilities on downstream tasks while improving decision fairness. Experiments on three downstream tasks on seven PLMs validate that BATED outperforms state-of-the-art methods on an overall evaluation of fairness and task performance.

While the proposed method is theoretically extensible to multilingual debiasing applications, its experimental validation in this paper is limited to English-language tasks due to limitations in available benchmarks and datasets. Existing research has revealed divergent bias patterns between multilingual and monolingual models. For instance, Parra [77] observed that multilingual variants of BERT, DistilBERT, and RoBERTa generally exhibit weaker gender stereotypes compared to their monolingual model. Future work may explore debiasing approaches for other linguistic units, depending on the availability of multilingual benchmark datasets and fairness evaluation metrics. Frameworks such as the MBE proposed by Kaneko et al. [78] could provide foundational evaluation benchmarks. However, the development of a more comprehensive assessment system remains an open challenge, which we defer to future research.

CRediT authorship contribution statement

Yingji Li: Writing – original draft, Methodology, Formal analysis, Conceptualization, Validation, Investigation, Data curation. **Mengnan Du:** Supervision, Methodology. **Rui Song:** Methodology, Formal analysis. **Mu Liu:** Formal analysis, Validation, Writing – review & editing, Investigation, Visualization. **Ying Wang:** Project administration, Supervision, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We express gratitude to the anonymous reviewers for their hard work and kind comments. The work was supported in part by the National Natural Science Foundation of China (No. 62272191, No. 62372211), the International Science and Technology Cooperation Program of Jilin Province (No. 20240402067GH), the China Postdoctoral Science Foundation Funded Project (No. 2024M761122).

Data availability

Data will be made available on request.

References

[1] J. Devlin, M. Chang, K. Lee, K. Toutanova, 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, NAACL, Association for Computational Linguistics, 2019, pp. 4171–4186.
[2] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (8) (2019) 9.
[3] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Proceedings of the 33rd Annual Conference on Neural Information Processing Systems, NeurIPS, 2020.
[4] AI@Meta, Llama 3 model card, https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
[5] OpenAI, GPT-4 technical report, CoRR, arXiv:2303.08774, https://doi.org/10.48550/ARXIV.2303.08774.
[6] A. Caliskan, J.J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, Science 356 (6334) (2017) 183–186.

[7] P. Chen, X. Guo, Y. Li, X. Zhang, Z. Feng, Mitigating language bias of lmms in social intelligence understanding with virtual counterfactual calibration, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2024, pp. 1300–1310, https://aclanthology.org/2024.emnlp-main.77.
[8] K.V. Deshpande, S. Pan, J.R. Foulds, Mitigating demographic bias in ai-based resume filtering, in: Proc. 28th UMAP Adjun. - Adjun. Publ. ACM Conf. User Model., Adapt. Pers., ACM, 2020, pp. 268–275.
[9] L. Ding, Y. Hu, N. Denier, E. Shi, J. Zhang, Q. Hu, K.D. Hughes, L. Kong, B. Jiang, Probing social bias in labor market text generation by chatgpt: a masked language model approach, in: Proceedings of the 38th Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems 2024, NeurIPS, 2024, http://papers.nips.cc/paper_files/paper/2024/hash/fce2d8a485746f76aac7b5650db2679d-Abstract-Conference.html.
[10] Z. Obermeyer, B. Powers, C. Vogeli, S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, Science 366 (6464) (2019) 447–453.
[11] M. Moukheiber, L. Moukheiber, D. Moukheiber, H. Lee, Unmasking societal biases in respiratory support for ICU patients through social determinants of health, in: Proceedings of the 33rd International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3–9, 2024, 2024, pp. 7421–7429, https://www.ijcai.org/proceedings/2024/821.
[12] Y. Li, M. Du, R. Song, X. Wang, Y. Wang, A survey on fairness in large language models, CoRR, arXiv:2308.10149, https://doi.org/10.48550/ARXIV.2308.10149.
[13] K. Webster, X. Wang, I. Tenney, A. Beutel, E. Pitler, E. Pavlick, J. Chen, S. Petrov, Measuring and reducing gendered correlations in pre-trained models, CoRR, arXiv:2010.06032.
[14] K. Lu, P. Mardziel, F. Wu, P. Amancharla, A. Datta, Gender bias in neural natural language processing, in: Proceedings of the Logic, Language, and Security - Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday, in: Lecture Notes in Computer Science, vol. 12300, 2020, pp. 189–202.
[15] P. Cheng, W. Hao, S. Yuan, S. Si, L. Carin, Fairfil: contrastive neural debiasing method for pretrained text encoders, in: Proceedings of the 9th International Conference on Learning Representations, ICLR, 2021.
[16] J. He, M. Xia, C. Fellbaum, D. Chen, MABEL: attenuating gender bias using textual entailment data, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2022, pp. 9681–9702.
[17] Y. Li, M. Du, X. Wang, Y. Wang, Prompt tuning pushes farther, contrastive learning pulls closer: a two-stage approach to mitigate social biases, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL, 2023, pp. 14254–14267.
[18] Z. Yang, Y. Cheng, Y. Liu, M. Sun, Reducing word omission errors in neural machine translation: a contrastive learning approach, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL, Association for Computational Linguistics, 2019, pp. 6191–6196, https://doi.org/10.18653/v1/p19-1623.
[19] Y. Guo, Y. Yang, A. Abbasi, Auto-debias: debiasing masked language models with automated biased prompts, in: Proc. 60th Annu. Meeting Assoc. Comput. Linguistics, 2022, pp. 1012–1023.
[20] Y. Liu, Y. Liu, X. Chen, P. Chen, D. Zan, M. Kan, T. Ho, The devil is in the neurons: interpreting and mitigating social biases in language models, in: Proceedings of the 12th International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7–11, 2024, 2024, https://openreview.net/forum?id=SQGUDc9tC8.
[21] A. Rakshit, S. Singh, S. Keshari, A.G. Chowdhury, V. Jain, A. Chadha, From prejudice to parity: a new approach to debiasing large language model word embeddings, in: Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025, 2025, pp. 6718–6747, https://aclanthology.org/2025.coling-main.450/.
[22] S. Goldfarb-Tarrant, R. Marchant, R.M. Sánchez, M. Pandya, A. Lopez, Intrinsic bias metrics do not correlate with application bias, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP, 2021, pp. 1926–1940.
[23] T. Katô, Y. Miyao, Analyzing correlations between intrinsic and extrinsic bias metrics of static word embeddings with their measuring biases aligned, CoRR, arXiv:2409.09260, https://doi.org/10.48550/ARXIV.2409.09260.
[24] Z. Fatemi, C. Xing, W. Liu, C. Xiong, Improving gender fairness of pre-trained language models without catastrophic forgetting, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, vol. 2, Short Papers, ACL 2023, Toronto, Canada, July 9–14, 2023, Association for Computational Linguistics, 2023, pp. 1249–1262.
[25] R.B. Loureiro, T.P. Pagano, F.V.N. Lisboa, L.F.S. Nascimento, E.L.S. de Oliveira, I. Winkler, E.G.S. Nascimento, Correlation-based methods for representative fairness metric selection: an empirical study on efficiency and caveats in model evaluation, Expert Syst. Appl. 268 (2025) 126344, https://doi.org/10.1016/J.ESWA.2024.126344.
[26] X. Han, T. Baldwin, T. Cohn, Decoupling adversarial training for fair NLP, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, vol. ACL/IJCNLP 2021 of Findings of ACL, August 1–6, 2021, Association for Computational Linguistics, 2021, pp. 471–477, https://doi.org/10.18653/V1/2021.FINDINGS-ACL.41.
[27] S. Ravfogel, M. Twiton, Y. Goldberg, R.D. Cotterell, Linear adversarial concept erasure, in: Proceedings of the 39th International Conference on Machine Learning, ICML, vol. 162, 2022, pp. 18400–18421.
[28] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, Y. Goldberg, Null it out: guarding protected attributes by iterative nullspace projection, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020, pp. 7237–7256.
[29] S. Ghanbarzadeh, Y. Huang, H. Palangi, R.C. Moreno, H. Khanpour, Gender-tuning: empowering fine-tuning for debiasing pre-trained language models, in: Proceedings of the Findings of the Association for Computational Linguistics: ACL, 2023, pp. 5448–5458.
[30] S. Choi, M. Jeong, H. Han, S. Hwang, C2L: causally contrastive learning for robust text classification, in: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, the Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, AAAI Press, 2022, pp. 10526–10534, https://doi.org/10.1609/AAAI.V36I10.21296.
[31] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, in: 2nd International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings, Banff, AB, Canada, April 14–16, 2014, 2014, http://arxiv.org/abs/1312.6114.
[32] Y. Bao, H. Zhou, S. Huang, L. Li, L. Mou, O. Vechtomova, X. Dai, J. Chen, Generating sentences from disentangled syntactic and semantic spaces, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, vol. 1, Long Papers, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Association for Computational Linguistics, 2019, pp. 6008–6019, https://doi.org/10.18653/V1/P19-1602.
[33] A. Liusie, Y. Fathullah, M.J.F. Gales, Teacher-student training for debiasing: general permutation debiasing for large language models, in: Proceedings of the Findings of the Association for Computational Linguistics, ACL, 2024, pp. 1376–1387, https://doi.org/10.18653/V1/2024.FINDINGS-ACL.81.
[34] P.P. Liang, I.M. Li, E. Zheng, Y.C. Lim, R. Salakhutdinov, L. Morency, Towards debiasing sentence representations, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020, pp. 5502–5515.
[35] R. Zmigrod, S.J. Mielke, H.M. Wallach, R. Cotterell, Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology, in: Proc. 57th Annu. Meeting Assoc. Comput. Linguistics, 2019, pp. 1651–1661.
[36] M.L. Olson, R. Khanna, L. Neal, F. Li, W. Wong, Counterfactual state explanations for reinforcement learning agents via generative deep learning, Artif. Intell. 295 (2021) 103455.
[37] Y. Li, M. Du, R. Song, X. Wang, M. Sun, Y. Wang, Mitigating social biases of pre-trained language models via contrastive self-debiasing with double data augmentation, Artif. Intell. 332 (2024) 104143, https://doi.org/10.1016/j.artint.2024.104143, https://www.sciencedirect.com/science/article/pii/S0004370224000791.
[38] M. Standen, J. Kim, C. Szabo, Adversarial machine learning attacks and defences in multi-agent reinforcement learning, ACM Comput. Surv. 57 (5) (2025) 124, https://doi.org/10.1145/3708320.

[39] N. Torres, Contrastive adversarial gender debiasing, Nat. Lang. Process. J. 8 (2024) 100092, https://doi.org/10.1016/J.NLP.2024.100092.
[40] H. Liu, W. Jin, H. Karimi, Z. Liu, J. Tang, The authors matter: understanding and mitigating implicit bias in deep text classification, in: Proceedings of the Findings of the Association for Computational Linguistics, ACL/IJCNLP, 2021, pp. 74–85.
[41] A. Shen, X. Han, T. Cohn, T. Baldwin, L. Frermann, Does representational fairness imply empirical fairness?, in: Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, 2022, pp. 81–95.
[42] A. Shen, X. Han, T. Cohn, T. Baldwin, L. Frermann, Contrastive learning for fair representations, arXiv:2109.10645.
[43] Z. Xu, W. Chen, Y. Tang, X. Li, C. Hu, Z. Chu, K. Ren, Z. Zheng, Z. Lu, Mitigating social bias in large language models: a multi-objective approach within a multi-agent framework, in: AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA, AAAI Press, 2025, pp. 25579–25587, https://doi.org/10.1609/AAAI.V39I24.34748.
[44] M. Spliethöver, S.N. Menon, H. Wachsmuth, Disentangling dialect from social bias via multitask learning to improve fairness, in: Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and Virtual Meeting, August 11–16, 2024, 2024, pp. 9294–9313, https://doi.org/10.18653/V1/2024.FINDINGS-ACL.553.
[45] O. Dige, D. Arneja, T.F. Yau, Q. Zhang, M. Bolandraftar, X. Zhu, F.K. Khattak, Can machine unlearning reduce social bias in language models?, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: EMNLP 2024 - Industry Track, Miami, Florida, USA, November 12–16, 2024, 2024, pp. 954–969, https://aclanthology.org/2024.emnlp-industry.71, 2024.
[46] R.D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, Y. Bengio, Learning deep representations by mutual information estimation and maximization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019, 2019, OpenReview.net, https://openreview.net/forum?id=Bklr3j0cKX.
[47] J. Li, X. Zhou, Curegraph: contrastive multi-modal graph representation learning for urban living circle health profiling and prediction, Artif. Intell. 340 (2025) 104278, https://doi.org/10.1016/J.ARTINT.2024.104278.
[48] Z. Zhang, Z. Zhao, Z. Lin, J. Zhu, X. He, Counterfactual contrastive learning for weakly-supervised vision-language grounding, in: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6–12, 2020, Virtual, 2020, https://proceedings.neurips.cc/paper/2020/hash/d27b95cac4c27feb850aaa4070cc4675-Abstract.html.
[49] E. Creager, D. Madras, J. Jacobsen, M.A. Weis, K. Swersky, T. Pitassi, R.S. Zemel, Flexibly fair representation learning by disentanglement, in: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9–15 June 2019, Long Beach, California, USA, in: Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 1436–1445, http://proceedings.mlr.press/v97/creager19a.html.
[50] R. Song, F. Giunchiglia, Y. Li, M. Tian, H. Xu, TACIT: a target-agnostic feature disentanglement framework for cross-domain text classification, in: Proceedings of the 38th AAAI Conference on Artificial Intelligence, 2024, pp. 18999–19007, https://doi.org/10.1609/aaai.v38i17.29866.
[51] C.P. Burgess, I. Higgins, A. Pal, L. Matthey, N. Watters, G. Desjardins, A. Lerchner, Understanding disentangling in β-vae, CoRR, arXiv:1804.03599, http://arxiv.org/abs/1804.03599.
[52] Y. Wei, H. Hu, Z. Xie, Z. Zhang, Y. Cao, J. Bao, D. Chen, B. Guo, Contrastive learning rivals masked image modeling in fine-tuning via feature distillation, CoRR, arXiv:2205.14141, https://doi.org/10.48550/ARXIV.2205.14141.
[53] R. Socher, A. Perelygin, J. Wu, J. Chuang, C.D. Manning, A.Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP, ACL, 2013, pp. 1631–1642, https://aclanthology.org/D13-1170/.
[54] S. Park, H.S. Shim, M. Chatterjee, K. Sagae, L. Morency, Computational analysis of persuasiveness in social multimedia: a novel dataset and multimodal prediction approach, in: Proceedings of the 16th International Conference on Multimodal Interaction, ICMI, ACM, 2014, pp. 50–57, https://doi.org/10.1145/2663204.2663260.
[55] S. Merity, C. Xiong, J. Bradbury, R. Socher, Pointer sentinel mixture models, in: Proceedings of the 5th International Conference on Learning Representations, ICLR, 2017, OpenReview.net, https://openreview.net/forum?id=Byj72udxe.
[56] M. Völske, M. Potthast, S. Syed, B. Stein, Tl;dr: mining reddit to learn automatic summarization, in: Proceedings of the Workshop on New Frontiers in Summarization, NFiS@EMNLP, Association for Computational Linguistics, 2017, pp. 59–63, https://doi.org/10.18653/v1/w17-4508.
[57] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, R. Mihalcea, MELD: a multimodal multi-party dataset for emotion recognition in conversations, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL, Association for Computational Linguistics, 2019, pp. 527–536, https://doi.org/10.18653/v1/p19-1050.
[58] B. Pang, L. Lee, Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales, in: ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25–30 June 2005, University of Michigan, USA, The Association for Computer Linguistics, 2005, pp. 115–124, https://aclanthology.org/P05-1015/.
[59] J.J. McAuley, J. Leskovec, From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews, in: 22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13–17, 2013, International World Wide Web Conferences Steering Committee / ACM, 2013, pp. 897–908, https://doi.org/10.1145/2488388.2488466.
[60] A.L. Maas, R.E. Daly, P.T. Pham, D. Huang, A.Y. Ng, C. Potts, Learning word vectors for sentiment analysis, in: The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19–24 June, 2011, Portland, Oregon, USA, The Association for Computer Linguistics, 2011, pp. 142–150, https://aclanthology.org/P11-1015/.
[61] M. De-Arteaga, A. Romanov, H.M. Wallach, J.T. Chayes, C. Borgs, A. Chouldechova, S.C. Geyik, K. Kenthapadi, A.T. Kalai, 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, ACM, 2019, pp. 120–128, https://doi.org/10.1145/3287560.3287572.
[62] D. Borkan, L. Dixon, J. Sorensen, N. Thain, L. Vasserman, Nuanced metrics for measuring unintended bias with real data for text classification, in: Web Conf. - Companion World Wide Web Conf., WWW, ACM, 2019, pp. 491–500.
[63] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, CoRR, arXiv:1910.01108, http://arxiv.org/abs/1910.01108.
[64] K. Clark, M. Luong, Q.V. Le, C.D. Manning, ELECTRA: pre-training text encoders as discriminators rather than generators, in: Proceedings of the 8th International Conference on Learning Representations, ICLR, 2020, OpenReview.net, https://openreview.net/forum?id=r1xMH1BtvB.
[65] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M.T. Diab, X. Li, X.V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P.S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, OPT: open pre-trained transformer language models, CoRR, arXiv:2205.01068, https://doi.org/10.48550/ARXIV.2205.01068.
[66] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Fan, Qwen2 technical report, arXiv preprint, arXiv:2407.10671.
[67] Q. Team, Qwen2.5: a party of foundation models, https://qwenlm.github.io/blog/qwen2.5/, September 2024.
[68] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T.L. Scao, S. Gugger, M. Drame, Q. Lhoest, A.M. Rush, Transformers: state-of-the-art natural language processing, in: Proc. Conf. Empir. Methods Natural Lang. Process., 2020, pp. 38–45.
[69] T. Bolukbasi, K. Chang, J.Y. Zou, V. Saligrama, A.T. Kalai, Man is to computer programmer as woman is to homemaker? Debiasing word embeddings, in: Proc. 30th Adv. Neural Inf. Proces. Syst., 2016, pp. 4349–4357.

[70] N. Meade, E. Poole-Dayan, S. Reddy, An empirical survey of the effectiveness of debiasing techniques for pre-trained language models, in: Proc. 60th Annu. Meeting Assoc. Comput. Linguistics, 2022, pp. 1878–1898.
[71] Y. Yu, C.-H.H. Yang, J. Kolehmainen, P.G. Shivakumar, Y. Gu, S.R.R. Ren, Q. Luo, A. Gourav, I.-F. Chen, Y.-C. Liu, et al., Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition, in: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2023, pp. 1–8.
[72] C. May, A. Wang, S. Bordia, S.R. Bowman, R. Rudinger, On measuring social biases in sentence encoders, in: Proc. Conf. N. Am. Chapter Assoc. Comput. Linguist.: Hum. Lang. Technol., 2019, pp. 622–628.
[73] N. Nangia, C. Vania, R. Bhalerao, S.R. Bowman, Crows-pairs: a challenge dataset for measuring social biases in masked language models, in: Proc. Conf. Empir. Methods Natural Lang. Process., 2020, pp. 1953–1967.
[74] M. Nadeem, A. Bethke, S. Reddy, 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, ACL/IJCNLP, Association for Computational Linguistics, 2021, pp. 5356–5371, https://doi.org/10.18653/v1/2021.acl-long.416.
[75] G. Stanovsky, N.A. Smith, L. Zettlemoyer, Evaluating gender bias in machine translation, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, Long Papers, ACL 2019, Florence, Italy, July 28 - August 2, 2019, 2019, pp. 1679–1684, https://doi.org/10.18653/V1/P19-1164.
[76] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, R.S. Zemel, Fairness through awareness, in: Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8–10, 2012, ACM, 2012, pp. 214–226.
[77] I. Parra, Unmasked: quantifying gender biases in masked language models through linguistically informed job market prompts, in: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2024: Student Research Workshop, St. Julian's, Malta, March 21-22, 2024, 2024, pp. 61–70, https://aclanthology.org/2024.eacl-srw.6.
[78] M. Kaneko, A. Imankulova, D. Bollegala, N. Okazaki, 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, NAACL 2022, Seattle, WA, United States, July 10–15, 2022, 2022, pp. 2740–2750, https://doi.org/10.18653/v1/2022.naacl-main.197.