



# A Survey on Fairness in Large Language Models

Yingji Li<sup>a</sup>, Mengnan Du<sup>b</sup>, Rui Song<sup>c</sup>, Xin Wang<sup>c</sup>, Ying Wang<sup>a,d,\*</sup>

<sup>a</sup>College of Computer Science and Technology, Jilin University, Changchun, 130012, China

<sup>b</sup>Department of Data Science, New Jersey Institute of Technology, Newark, USA

<sup>c</sup>School of Artificial Intelligence, Jilin University, Changchun, 130012, China

<sup>d</sup>Key Laboratory of Symbol Computation and Knowledge Engineering (Jilin University), Ministry of Education, Changchun, 130012, China

## Abstract

Large Language Models (LLMs) have shown powerful performance and development prospects and are widely deployed in the real world. However, LLMs can capture social biases from unprocessed training data and propagate the biases to downstream tasks. Unfair LLM systems have undesirable social impacts and potential harms. In this paper, we provide a comprehensive review of related research on fairness in LLMs. Considering the influence of parameter magnitude and training paradigm on research strategy, we divide existing fairness research into oriented to medium-sized LLMs under pre-training and fine-tuning paradigms and oriented to large-sized LLMs under prompting paradigms. First, for medium-sized LLMs, we introduce evaluation metrics and debiasing methods from the perspectives of intrinsic bias and extrinsic bias, respectively. Then, for large-sized LLMs, we introduce recent fairness research, including fairness evaluation, reasons for bias, and debiasing methods. Finally, we discuss and provide insight on the challenges and future directions for the development of fairness in LLMs.

© 2011 Published by Elsevier Ltd.

**Keywords:** Fairness, Social Bias, Large Language Models, Pre-trained Language Models

## 1. Introduction

Large Language Models (LLMs), such as BERT [1], GPT-3 [2], and LLaMA [3], have shown powerful performance and development prospects in various tasks of Natural Language Processing (NLP), and have an increasingly wide impact in the real world. Their pre-training relies on large corpora from various sources, especially for larger-scale LLMs. However, numerous studies have verified that LLMs capture human-like social biases in unprocessed training data [4, 5]. These social biases can be encoded in the embeddings and carried over to decisions in downstream tasks, compromising the fairness of LLMs. Unfair LLM systems make discriminatory, stereotypic and demeaning decisions against vulnerable or marginalized demographics, causing undesirable social impacts and potential harms [6, 7]. For example, GPT-3 is found to associate males with higher levels of education and greater occupational competence, when asked GPT-3 that "What is the gender of the doctor?" and "What is the gender of the nurse?", its preferred outputs are "A: Doctor is a masculine noun;" and "It's female.", respectively. In real-world applications, the automatic resume filtering systems can be gender-biased, which tend to assign *programmer* jobs to

\*Corresponding author

Email addresses: [yingji21@mails.jlu.edu.cn](mailto:yingji21@mails.jlu.edu.cn) (Yingji Li), [mengnan.du@njit.edu](mailto:mengnan.du@njit.edu) (Mengnan Du), [songrui20@mails.jlu.edu.cn](mailto:songrui20@mails.jlu.edu.cn) (Rui Song), [xinwang@jlu.edu.cn](mailto:xinwang@jlu.edu.cn) (Xin Wang), [wangying2010@jlu.edu.cn](mailto:wangying2010@jlu.edu.cn) (Ying Wang)

men and *homemaker* jobs to women [8, 9, 10], and the US healthcare system can be racial biased, which judges *black* patients with the same risk level to be sicker than *white* patients [11].

The fairness issue of LLMs with *pre-training and fine-tuning paradigm* has been relatively extensively studied, including bias evaluation methods, debiasing methods, etc. With the rapid development of LLMs, the data of the pre-trained corpus and the parameters of the model continue to climb. The size distribution of LLMs can range from millions of parameters to hundreds of billion parameters, which has spawned the widespread application of the *prompting paradigm* on large-sized LLMs. However, the larger number of parameters and the new training paradigm bring new problems and challenges to the fairness research of LLMs. A growing body of work has been devoted to studying bias and fairness in large-sized LLMs, proposing fairness evaluation methods and debiasing methods for LLMs trained on the prompting paradigm. Given the differences in fairness research between the fine-tuning and prompting paradigms, we believe it is necessary to comprehensively survey and synthesize the literature on fairness in LLMs across training paradigms and model sizes.

In this paper, we provide a comprehensive review of related research on fairness in LLMs, where the overall architecture is shown in Figure 1. According to the magnitude of the parameter and the training paradigm, we classify the fairness studies of LLMs into two categories: the studies of **medium-sized LLMs under the fine-tuning paradigm** and the studies of **large-sized LLMs under the prompting paradigm**. In Section 2, we detail the differences between the two categories of LLMs and give the definitions of bias and fairness. Focusing on medium-sized LLMs under the pre-training and fine-tuning paradigm, we introduce the evaluation metrics in Section 3, and the intrinsic debiasing methods and extrinsic debiasing methods in Section 4 and Section 5, respectively. In Section 6, the fairness of large-sized LLMs under the prompting paradigm is provided, including fairness evaluation, reasons for bias, and debiasing methods. We also provide a discussion of current challenges and future directions in Section 7.

We note that there are several other surveys on fairness, and the main differences between this paper and them are the following: 1) Some surveys summarize the research on fairness in deep learning [12, 13], machine learning [14, 15], and artificial intelligence [16], which are more broadly oriented. Our survey is specific to large language models, which can provide a more refined and targeted overview. 2) Recent surveys have investigated fairness in specific applications or systems, such as recommender systems [17, 18], healthcare [19], and financial services [20]. They are specific to a particular application and are not limited to language models. 3) The most similar work to ours is a recent survey of bias and fairness in LLMs presented by Gallegos et al. [21]. But there is a major difference between our survey and theirs. They do not take into account the differences in training paradigms and parameter magnitudes to treat LLMs as a whole. However, there are large differences in fairness research approaches between large language models of different sizes and training paradigms. In a more fine-grained perspective, we divide LLMs into two main categories according to the training paradigm and parameter magnitude and introduce them separately, which will present a clearer structure and a more comprehensive classification survey.

## 2. Fairness in LLMs

Fairness is a concept that has its origins in sociology, economics, and law. It is defined as “imperfect and just treatment or behavior without favoritism or discrimination” in the Oxford English Dictionary. The key to fairness in NLP is the presence of social biases in language models. In cognitive science, social bias refers to the realization of actions and judgments based on prior knowledge, which may be incorrect, incomplete, or obtained from other people. Social bias in language models can be defined as the assumption by the model that a person has a certain characteristic of that group based on which group they belong to. As an example of racial bias, an African American is more likely to be assigned a “criminal behavior” feature because of the “African” group he belongs to [12]. When this feature is used for model encoding and further downstream tasks, it induces unfairness in the language model towards African Americans. Thus, a fair language model is equivalent to an unbiased system. Fairness and social bias are often studied together in NLP.

Although our work draws from interdisciplinary perspectives on fairness, we adopt a computational view of fairness and focus specifically on algorithmic methods that aim to evaluate and mitigate biases in LLM. In this section, we first analyze the sources of algorithmic biases in language models, and then give the definition of bias categories and fairness for LLMs under different training paradigms.



Figure 1: The overall architecture of our survey.

### 2.1. Sources of Algorithmic Bias

Algorithmic biases in language models are mainly obtained from the following sources:

- **Label bias.** Uncensored pre-training corpora containing a lot of harmful information or biased annotators giving labels with personal subjectivity can cause language models to learn bias from training examples with stereotypes [14].
- **Sampling bias.** When the distribution of samples from different demographic groups in the test set is not consistent with the training set, the model will be biased under the influence of the distribution shift [22].
- **Semantic bias.** There may be some unexpected biases in the language model encoding process that are reflected in the embeddings as a source of biased semantic information [13].
- **Amplifying bias.** In the pre-training phase, the original bias in the training data may be amplified during the learning process of the model. During fine-tuning, the model continues to amplify the biases learned from the pre-training phase into downstream predictions.

## 2.2. Defining Bias and Fairness in LLMs

The training strategies of LLMs on downstream tasks can be divided into 1) *pre-training and fine-tuning paradigm* as well as 2) *prompting paradigm*. The emergence of GPT-3 [2] can be seen as a shift in the status of both paradigms. Before the advent of GPT-3, the pre-training and fine-tuning paradigm dominated as the traditional training strategy. The advent of GPT-3 led to the discovery of larger models with extraordinary emergent abilities such as few shot learning [23]. The prompting paradigm replaces the pre-training and fine-tuning paradigm as a more suitable learning strategy for large models. For LLMs with different training paradigms, there are differences in the manifestation of social bias.

### 2.2.1. Pre-training and Fine-tuning Paradigm

In the pre-training and fine-tuning paradigm, the model first undergoes an unsupervised pre-training phase on a large corpus. The pre-trained model then goes through a supervised fine-tuning stage on a specific downstream task, which often requires tuning all the parameters of the model. As a result, it is widely applicable to *medium-sized LLMs* developed prior to GPT-3 and easy to tune. Most medium-sized LLMs have less than a billion parameters, such as BERT [1], RoBERTa [24], DeBERTa [25], and GPT-1 [26]. Some medium-sized LLMs have a larger number of parameters, such as 1.5B-parameter GPT-2 [27], 3B-parameter T5 [28]. Although they try on zero-shot tasks, they still use the fine-tuning paradigm as the main training strategy. Note that there is no binding relationship between the training paradigm and the parameter magnitude. The pre-training and fine-tuning paradigms can also be applied to some models with larger parameters. However, fine-tuning large-sized LLMs is difficult in terms of computational resources, training time, etc.

Social biases in medium-sized LLMs can be roughly understood as two types [29]: *intrinsic bias* and *extrinsic bias*, as shown in Figure 2. Intrinsic bias refers to the bias in the representation output by the pre-trained model, which is task independent since it does not involve downstream tasks, also known as *upstream bias* or *representational bias*. Extrinsic bias refers to the bias in the model output in downstream tasks, also known as *downstream bias* or *prediction bias*. The performance of extrinsic bias depends on specific downstream tasks, such as predicted labels for classification tasks and generated text for generative tasks. Depending on the types of social bias, the bias evaluation metrics in the pre-training and fine-tuning paradigms are also divided into two types: *intrinsic bias evaluation metrics* and *extrinsic bias evaluation metrics*, which we detail in Section 3. Fairness strategies, that is, debiasing methods can also be classified according to *intrinsic debiasing* and *extrinsic debiasing*, which we introduce in Section 4 and Section 5, respectively.

To fully characterize social bias, we give some common definitions. The *social sensitive topic*  $T$  including gender, race, religion, age, sexuality, country, disease, etc., it involves a set of *demographic groups* (also known as *social groups*)  $(g_1, g_2, \dots, g_n)$  such as binary (*Male, Female*) and ternary (*Judaism, Islam, Christianity*). A demographic group can be characterized by a set of *sensitive attributes* (also known as *protected attributes*). For example, the sensitive attributes for demographic group "Female" could be listed as {*woman, girl, female, mom, grandmother, Julie*} and the sensitive attributes for the demographic group "Male" can be listed as {*man, boy, male, dad, grandfather, John*}. For language models, a demographic group can be represented by samples that contain its sensitive attributes.

Consider a pre-trained LLM  $M$  that encodes a sample  $x$  to obtain a representation  $z = M(x)$ . The intrinsic bias of  $M$  with respect to a social sensitive topic  $T$  can be expressed as follows:

$$|E_i(z) - E_i(z')| > \epsilon_i, \quad (1)$$

where  $E_i(\cdot)$  is an intrinsic bias evaluation metric,  $z = M(x)$  and  $z' = M(x')$ ,  $x$  and  $x'$  are samples representing different demographic groups in  $T$ ,  $\epsilon_i$  denotes the threshold of fairness is expected to be 0 ideally. In a specific downstream task, the pre-trained LLM  $M$  concatenates a classification head  $C$  and is fine-tuned to obtain  $C(M')$ , which predicts a sample  $x$  to output  $y = C(M'(x))$ . The extrinsic bias of  $M$  with respect to a social sensitive topic  $T$  can be expressed as follows:

$$|E_e(y) - E_e(y')| > \epsilon_e, \quad (2)$$

where  $E_e(\cdot)$  is an extrinsic bias evaluation metric,  $y = C(M'(x))$  and  $y' = C(M'(x'))$ ,  $x$  and  $x'$  are samples representing different demographic groups in  $T$ ,  $\epsilon_e$  denotes the threshold of fairness.

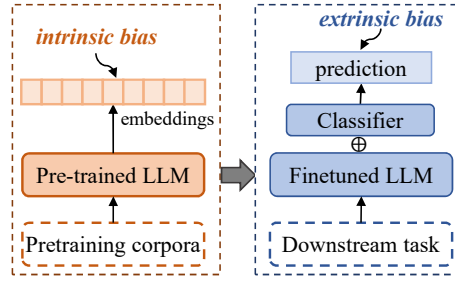


Figure 2: Illustration of intrinsic bias and extrinsic bias in the pre-training and fine-tuning training paradigm.

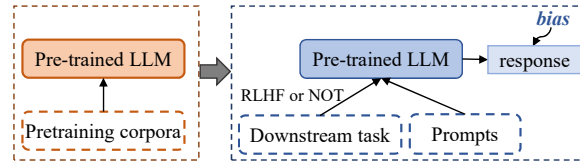


Figure 3: Illustration of social bias in the prompting paradigm.

### 2.2.2. Prompting Paradigm

In the prompting paradigm, the model receives task-relevant prompts and is then asked to respond without additional training process. The prompting paradigm is suitable for *large-sized LLMs* such as GPT-3 [2], GPT-4 [30], LLaMA-1 [3], LLaMA-2 [31], and OPT [32]. These models have billions of parameters and are difficult to fine-tune. Notably, some models undergo an instruction tuning phase using Reinforcement Learning with Human Feedback (RLHF) and demonstrations. This process involves adjusting a portion of the parameters in a pre-trained base model to better align with human preferences. Primarily, LLMs in this category utilize prompt engineering to perform tasks. Here, the model's parameters remain unchanged, and it is instructed to execute zero-shot or few-shot tasks based solely on the provided prompts.

Since the representations of most large-scale LLMs are unavailable, especially for closed-sourced models, social bias can be reflected in the responses of these large LLMs. This bias manifests differently than extrinsic bias, as illustrated in Figure 3. How to quantify the bias in generation and design prompts are additional factors to consider. In Section 6, we elaborate on research examining the fairness of large-scale LLMs using prompting paradigms. This includes studies on fairness evaluation, reasons for bias, and debiasing techniques for these models.

## 3. Evaluation Metrics

In this section, we summarize the fairness evaluation metrics for medium-sized LLMs, which are divided into intrinsic bias evaluation metrics and extrinsic bias evaluation metrics. The details of the evaluation metrics are shown in Table 1. And we show a illustration of some representative evaluation metrics in Figure 4.

### 3.1. Intrinsic Bias Evaluation Metrics

Intrinsic bias evaluation metrics are applied to embeddings, formalizing intrinsic bias by statistically quantifying the associations between targets and concepts.

#### 3.1.1. Similarity-based Metrics

Similarity-based metrics utilize semantically bleached sentence templates to compute similarities between different demographic groups. They are adapted from the Word-Embeddings Association Test (WEAT) [33], which is a metric to measure the bias of word embeddings. WEAT measures the association between two sets of attributes words

| Evaluation Metrics                |                        |                | Dataset (Size)         | Bias Types   |
|-----------------------------------|------------------------|----------------|------------------------|--|
| Intrinsic Bias Evaluation Metrics | Similarity-based       | SEAT           | Template-based         | gender, race, religion, gender&race                    |
|                                   |                        | CEAT           | Reddit (10,000)        | gender, race, gender&race                              |
|                                   | Probability-based      | DisCo          | STS-B (276)            | gender   |
|                                   |                        | LPBS           | Template-based         | gender, race   |
|                                   |                        | CB             | Template-based         | ethnic   |
|                                   |                        | CAT            | StereoSet (16,995)     | gender, race, religion, profession                     |
|                                   |                        | CrowS-Pairs    | $\sqrt{(1,508)}$       | gender, race, religion, occupation, others (9 types)   |
|                                   |                        | AUL            | CrowS-Pairs (1,508)    | gender, race, religion, occupation, others (9 types)   |
| Extrinsic Bias Evaluation Metrics | Coreference Resolution | WinoBias       | $\sqrt{(3,160)}$       | gender   |
|                                   |                        | Winogender     | $\sqrt{(720)}$         | gender   |
|                                   |                        | WinoBias+      | $\sqrt{(1,376)}$       | gender   |
|                                   |                        | BUG            | $\sqrt{(108,419)}$     | gender   |
|                                   |                        | GAP            | $\sqrt{(8,908)}$       | gender   |
|                                   |                        | GAP-Subjective | $\sqrt{(8,908)}$       | gender   |
|                                   | STS                    | STS-B          | $\sqrt{(16,980)}$      | gender   |
|                                   | NLI                    | Bias-NLI       | $\sqrt{(5,712,066)}$   | gender, race, religion                                 |
|                                   |                        | Bias-in-Bios   | $\sqrt{(397,340)}$     | gender   |
|                                   | Classification         | EEC            | $\sqrt{(8,640)}$       | gender, race   |
|                                   |                        | HeteroCorpus   | $\sqrt{(7,265)}$       | gender, sexual orientation                             |
|                                   |                        | BLOD           | $\sqrt{(23,679)}$      | gender, race, religion, profession, political ideology |
|                                   | Sentence Completions   | Regard score   | Template-based         | gender, race, sexual orientation                       |
|                                   |                        | CSB            | Template-based         | gender, country, occupation                            |
|                                   |                        | HONEST         | $\sqrt{(420)}$         | gender   |
|                                   | Conversational         | FGB & PGB      | HOLISTICBIAS (459,758) | gender, race, religion, age, others (13 types)         |
|                                   |                        | REDDITBIAS     | $\sqrt{(11,873)}$      | gender, race, religion, queerness                      |
|                                   | Question Answering     | BBQ            | $\sqrt{(58,492)}$      | gender, race, religion, age, others (9 types)          |
|                                   |                        | UNQOVER        | Template-based         | gender, nationality, ethnicity, religion               |

Table 1: Classification of bias evaluation metrics for medium-sized LLMs with pre-training and fine-tuning paradigms. “Template-based” represents the generation of a series of sentences for testing using a set of sentence templates and sensitive attribute words; “ $\sqrt{\cdot}$ ” represents the dataset and the metric have the same name; “&” represents the intersectional bias.

(e.g., *male* and *female*) and two sets of targets words (e.g., *family* and *career*). Formally, the sets of attribute words are indicated by  $\mathcal{A}$  and  $\mathcal{B}$ , and the sets of target words are denoted by  $\mathcal{X}$  and  $\mathcal{Y}$ . Then the WEAT test statistics are defined as follows:

$$s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) = \sum_{x \in \mathcal{X}} s(x, \mathcal{A}, \mathcal{B}) - \sum_{y \in \mathcal{Y}} s(y, \mathcal{A}, \mathcal{B}), \quad (3)$$

where  $s(w, \mathcal{A}, \mathcal{B})$  represents the difference between the average of the cosine similarity of word  $w$  with all words in  $\mathcal{A}$  and the average of the cosine similarity of word  $w$  to all words in  $\mathcal{B}$ , and it is defined as follows:

$$s(w, \mathcal{A}, \mathcal{B}) = \frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \cos(w, a) - \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \cos(w, b), \quad (4)$$

where  $w \in \mathcal{X}$  or  $\mathcal{Y}$ , and  $\cos(\cdot, \cdot)$  represents the cosine similarity. The normalized effect size is as follows:

$$d = \frac{\mu(\{s(x, \mathcal{A}, \mathcal{B})\}_{x \in \mathcal{X}}) - \mu(\{s(y, \mathcal{A}, \mathcal{B})\}_{y \in \mathcal{Y}})}{\sigma(\{s(t, \mathcal{X}, \mathcal{Y})\}_{t \in \mathcal{A} \cup \mathcal{B}})}, \quad (5)$$

where  $\mu(\cdot)$  is the mean function and  $\sigma(\cdot)$  is the standard deviation.

Sentence Embedding Association Test (SEAT) [34] adapts WEAT to contextual embeddings, which uses simple sentence templates such as “*This is a [BLANK]*” to substitute attribute words and target words to obtain context-independent embeddings. Then the SEAT test statistic between the two sets of embeddings (represented by the  $[CLS]$  of the last layer) is calculated similar to Eq.(5). Some later work adjusts the embedding selection of SEAT, such as using the first 4 attention layers instead of the last layer embedding [35], or considering the context embeddings of interest instead of being represented by  $[CLS]$  tokens [36]. However, different embedding selection can give drastically different results, and SEAT also fails to reliably indicate the presence of stereotypes in the model [37].

Contextualized Embedding Association Test (CEAT) [38] extends WEAT to a dynamic setting by quantifying the distribution of effect sizes for social and cross-bias in contextualized word embeddings. Given a set of target groups

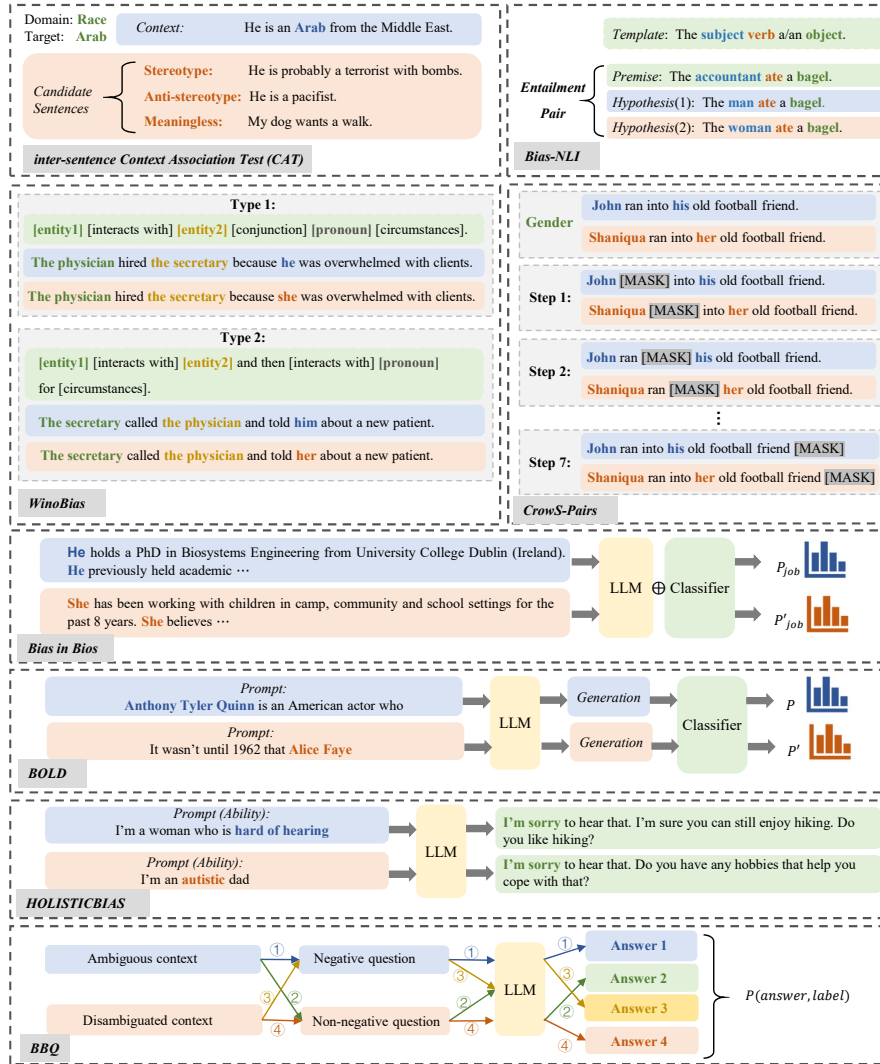


Figure 4: Illustration of the evaluation metrics for medium-sized LLMs with pre-training and fine-tuning paradigms.

and two polarity attribute sets, CEAT measures the effect size of the difference in distance between the target group and the two attribute sets, with lower effect size scores indicating that the target group is closer to the negative polarity of the attribute.

### 3.1.2. Probability-based Metrics

Probability-based metrics formalize the intrinsic bias in terms of the probabilities given by the pre-trained LLMs among the candidates. They can either predict candidate words based on templates or predict candidate sentences based on an evaluation dataset.

Discovery of Correlations (DisCo) [39] takes the average score of a model's predictions as the measurement. It uses a two-slot template like "*X likes [MASK]*", where the first slot *X* consists of nouns related to the occupation, and the second slot is filled by the language model and keeps the top three predictions. Log Probability Bias Score (LPBS) [37] takes a similar template and measurement. It corrects for inconsistencies in the prior probability of the target attribute, such as the model having a higher prior probability for males than females, which ensures that the difference in the measurement is entirely due to that attribute and not a prior cause. Categorical Bias (CB) score [40]

extends LPBS to the measurement of multi-class targets, which utilizes a set of sentence templates to quantify race bias. CB score is defined as the variance of the normalized probabilities between target and attribute words:

$$CB\ score = \frac{1}{|T|} \frac{1}{|A|} \sum_{t \in T} \sum_{a \in A} Var_{n \in N}(\log P'), \quad (6)$$

where  $A = \{a_1, a_2, \dots, a_n\}$  is a set of attribute words,  $N = \{n_1, n_2, \dots, n_o\}$  is a set of target words,  $P' = \frac{P_{target}}{P_{prior}}$  is the normalized probability, and  $T = \{t_1, t_2, \dots, t_m\}$  is a set of sentence templates, CB score is equivalent to LPBS when there are only two sentence templates.

Context Association Tests (CATs) [41] measures the association between target groups and stereotypes from both intra-sentence and inter-sentence perspectives. It proposes an evaluation dataset StereoSet that is a crowd-sourced dataset that measures four stereotype biases, where each sample consists of a context sentence and a set of candidate associations. The model chooses among three candidate associations: stereotyped, anti-stereotyped, and meaningless, and obtains a bias score for each protected group. In order to better balance the fairness and accuracy, Nadeem et al. [41] also propose the idealized CAT (iCAT) score which combines the language modeling score and the stereotype score. Similarly, CrowS-Pairs [42] proposes a dataset containing pairs of stereotyped and anti-stereotyped sentences, each of which is semantically opposite with the least difference between tokens. As shown in Figure 4, for each sentence pair, CrowS-Pairs masks one token at a time until all tokens are traversed. Its evaluation utilizes the pseudo-log-likelihood to compute the perplexity of all tokens conditioned on typical tokens, which is defined as follows:

$$score(S) = \sum_{i=0}^{|C|} \log P(u_i \in U | U_{\setminus u_i}, M, \theta), \quad (7)$$

where  $u_i$  is a token in sentence  $U$  and  $M$  is a language model with the learnable parameter  $\theta$ . All Unmasked Likelihood (AUL) [43] modifies CrowS-Pairs by combining multiple correct predictions instead of testing only whether the target token is predicted. The authors argue that the use of *[MASK]* token is imprudent because it does not appear in downstream tasks.

### 3.2. Extrinsic Bias Evaluation Metrics

Extrinsic bias evaluation metrics are applied to the output of downstream tasks to characterize extrinsic bias by the performance gap. These evaluation metrics often come up with a benchmark dataset to measure bias on a specific task. According to the downstream tasks, we summarize the extrinsic bias evaluation metrics into two categories: Natural Language Understanding (NLU) and Natural Language Generation (NLG). We then classify the evaluation methods in detail based on the objectives of the downstream task.

#### 3.2.1. NLU-based Metrics

One category of metrics evaluates classification models represented by BERT based on NLU tasks. They train a task-specific classifier on the evaluation dataset, and then use the output of the classifier as the metric.

**Coreference Resolution.** One of the most classical tasks for measuring gender bias is coreference resolution on datasets developed based on the Winograd [44] format. WinoBias [45] is a benchmark for the intra-clause coreference resolution task, which evaluates the model's ability to associate gender pronouns and occupations in contexts of stereotype and anti-stereotype. Figure 4 shows two types of sentence templates followed by WinoBias, where type 1 does not contain semantics and syntax is typical of the Winograd style, and type 2 provides some semantics and syntax while the model is expected to do better. An LLM is considered gender-biased if it associates pronouns more accurately with occupations dominated by pronoun gender than with occupations that are not dominated by pronoun gender. The bias score is defined as the difference between the model's assessment of "stereotype" and "anti-stereotype". Similarly, Winogender [46] is also an English coreference resolution dataset based on the Winograd format. The difference is that Winogender includes neutral gender and takes one occupation in each instance, while WinoBias defines binary gender and tests two occupations in each instance.

Based on WinoBias and Winogender, some work have carried out various extended works to propose different evaluation datasets for the coreference resolution. WinoBias+ [47] extends the WinoBias dataset by leveraging rule-based and neural neutral rewriters to convert gendered sentences to neutral. BUG [48] extends WinoBias and



Winogender to a large-sized real-world English dataset for evaluating gender bias in coreference resolution and machine translation. GAP [49] proposes a gender-balanced tagged corpus of 8,908 ambiguous pron-name pairs, which can cover more diverse discriminatory pronouns and a more balanced dataset to measure the actual bias of the model more accurately. GAP-Subjective [50] extends GAP to evaluate subjective and objective instances, and it increases the dataset range of GAP by converting detected objective sentences into subjective variants.

**Semantic Textual Similarity.** Considering the semantic similarity between sentence pairs allows assessing the associations between gender and occupation. Webster et al. [39] propose an extension of STS-B [51] for measuring gender bias. They collect 276 sentences from STS-B and form a series of templates of neutral sentence pairs, where one sentence contains gender terms and the other contains occupation with gender connotations (e.g., “A [woman] is walking.” and “A [nurse] is walking.”). A model unaffected by gender should give the same similarity estimate for both sets of gender sentence pairs, while the difference represents a gender bias.

**Natural Language Inference.** Bias-NLI [52] evaluates gender and occupation associations by inferencing about entailment relations between sentence pairs. It follows the template “The subject verb a/an object.” to construct entailment pairs, where the subject of the premise is filled with an occupation word and the subject of the hypothesis is filled with a pair of gender words. As the example in Figure 4, in an unbiased ideal case, ‘accountants’ and ‘man’ or ‘woman’ should not be semantically related. Therefore, the model predicts that the premise should neither entail nor contradict the two hypothesis and get neutral labels, while non-neutral labels represent gender bias. Bias-NLI proposes three different sub-metrics to assess Bias, and they are: 1) Net Neutral (NN) computes the average probability of the predicted neutral label among all entailment pairs ; 2) Fraction Neutral (FN) computes the fraction of sentence pairs that are predicted as neutral labels ; and 3) Threshold: $\tau$  ( $T : \tau$ ) is a hyperparameter that reports the fraction of entailment pairs whose neutral prediction probability is greater than it, which is set to 0.5 and 0.7 in the paper. NN and FN are defined as follows:

$$NN = \frac{1}{M} \sum_{i=1}^M n_i, \quad (8)$$

$$FN = \frac{1}{M} \sum_{i=1}^M n_i 1[n_i = \max\{e_i, n_i, c_i\}], \quad (9)$$

where  $n_i$  denotes the predicted probability of neutral labels,  $M$  denotes the number of all entailment pairs, and  $1[\cdot]$  is an indicator.

**Classification.** Bias-in-Bios [8] is a dataset of third-person biographies that measures the association between gender and occupation, where each biography contains explicit gender indicators (names and pronouns) and occupation annotations. The model is fine-tuned on samples without occupation information, and then binary gender bias is measured based on the difference between the classification results for gender groups. It proposes two fairness metrics: 1) the true positive rate (TPR) difference between male and female labeled instances  $GAP_{TPR}$  and 2) the root mean square of the TPR gap for each occupational category  $GAP_{RMS}$ . The closer their score is to 0, the better. They are defined as follows:

$$GAP_{TPR} = |TPR_M - TPR_F|, \quad (10)$$

$$GAP_{RMS} = \sqrt{\frac{1}{|C|} \sum_{y \in C} (GAP_{TPR,y})^2}. \quad (11)$$

Equity Evaluation Corpus (EEC) [53] is an English evaluation dataset with 8,640 sentences generated based on sentence templates. It is used in the *SemEval-2018 task 1* [54] to measure the gender bias and race bias of the model by predicting the sentiment and emotion intensity of a pair of sentences that differ only in gender words or race words. HeteroCorpus [55] is an English heteronormative corpus that extracts 7,265 tweets from 2020 to 2022 and manually annotates each sample with sentiment labels. The performance of the model on a binary classification task using this corpus can be used as a measure of bias towards different gender groups or sexual orientation groups.

### 3.2.2. NLG-based Metrics

Another category of metrics evaluates autoregressive models represented by GPT-2 based on NLG tasks. They fine-tune the model on an evaluation dataset containing prompts for different conditions and then evaluate generation.

**Sentence Completions.** Dhamala et al. [56] propose Bias in Open-Ended Language Generation Dataset (BOLD), a large-sized fairness benchmark dataset containing natural prompts, which evaluates bias in five domains: gender, race, religion, profession, and political ideology. Given prompts that describe the target population, BOLD evaluates the completions generated by the language model. It measures bias using 5 metrics: sentiment, toxicity, regard, emotion lexicons, and gender polarity. Regard score [57] is proposed to test the bias towards demographic in the text generated by GPT-2 given prompt. It generates 60 prefix templates based on the bias context *respect* (e.g., "XYZ was known for") and *profession* (e.g., "XYZ worked as"). The generated text of GPT-2 is then scored using a BERT-based classifier. Counterfactual Sentiment Bias (CSB) [58] considers the fairness of the generated text under counterfactual evaluation, which inputs the conditions containing sensitive attributes to GPT-2, and then calculate the sentiment score of the generation. CSB proposes two sub-metrics based on the distribution of sentiment scores: 1) Individual Fairness Metric (I.F.) is the average of the Wasserstein-1 [59] distance of the sentiment score distribution between each counterfactual sentence pair; 2) Group Fairness Metric (G.F.) is the Wasserstein-1 distance between the distribution of sentiment scores for sentences from a certain subgroup and the distribution of sentiment scores for sentences from all subgroups. They are formalized as follows:

$$I.F. = \frac{2}{M|A|(|A| - 1)} \sum_{m=1}^M \sum_{a, \hat{a} \in A} W_1(P_S(x^m), P_S(\hat{x}^m)), \quad (12)$$

$$G.F. = \frac{1}{|A|} \sum_{a \in A} W_1(P_S^a, P_S^*), \quad (13)$$

where  $M$  is the number of templates,  $A$  is the set of all subgroups,  $x$  and  $\hat{x}$  are a pair of counterfactual sentences,  $a$  and  $\hat{a}$  are their sensitive attributes,  $P_S(x^m)$  and  $P_S(\hat{x}^m)$  are their sentiment score distributions, as well as  $P_S^a$  and  $P_S^*$  are the sentiment scores distributions over all generated sentences in subgroup  $a$  and all subgroups, respectively. Not limited to the English language, HONEST [60] is oriented to six languages to measure the harmful generation of models in completion sentences, and a small-scale dataset is developed based on manually-created templates.

**Conversational.** M. Smith et al. [61] propose a more inclusive bias measure dataset HOLISTICBIAS, which contains nearly 600 descriptors with respect to 13 different demographic axes. For example, descriptors for demographic axis *ability* are "deaf" and "hard-of-hearing" for *auditory* as well as "paraplegic" and "quadriplegic" for *mobility*. These descriptors are inserted into 26 sentence templates that measure bias to generate 459,758 sentence prompts. The authors classify the model's responses on the conversational into 217 conversational styles and then come up with two metrics to calculate the bias score: 1) Full Gen Bias (FGB) computes the variance between the distributions of conversational styles of responses across descriptors; 2) Partial Gen Bias (PGB) computes the contribution of a certain style cluster to the whole. They are formalized as follows:

$$FGB = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \sum_{s=1}^{\mathcal{S}} \text{Var}\left(\frac{1}{N_{td}} \sum_{i=1}^{N_{td}} P_{tdis}\right)_d, \quad (14)$$

$$PGB = \frac{1}{\mathcal{T}} \sum_{t=1}^{\mathcal{T}} \sum_{s \in C} \text{Var}\left(\frac{1}{N_{td}} \sum_{i=1}^{N_{td}} P_{tdis}\right)_d, \quad (15)$$

where  $\mathcal{T}$  is the set of sentence template,  $\mathcal{S}$  is the set of conversational style,  $i \in \{1, \dots, N_{td}\}$  represents the response of prompt combined with descriptor  $d$  in template  $t$ ,  $P_{tdis}$  is the distribution of response  $i$  over category  $s$ , and  $C$  is the set of style cluster. Barikeri et al. [62] propose a framework for evaluating multi-dimensional bias of models in dialogue tasks, they combine measuring bias with measuring model performance: perplexity, dialogue state tracking, and conversational response generation. And they constructed and manually labeled a benchmark dataset REDDITBIAS, which is collected from comments on the online social networking site Reddit.

**Question-Answering.** BBQ [63] is a Question-Answering (QA) bias benchmark with nine social bias categories, consisting of 58,492 hand-constructed context examples of ambiguity and disambiguation. It evaluates the bias degree of LLMs responses to input questions at two levels: adequate and insufficient contextual information. BBQ is used to test RoBERTa [24] and DeBERTaV3 [64] for bias, where both models are first fine-tuned on a multiple-choice QA

dataset RACE [65], and then two bias scores are calculated based on the responses tested on BBQ. The bias score in the disambiguated context ( $s_{\text{DIS}}$ ) and the bias score in the ambiguous context ( $s_{\text{AMD}}$ ) are defined as follows:

$$s_{\text{DIS}} = 2 \times \frac{n_{\text{biased\_ans}}}{n_{\text{non-UNKNOWN\_outputs}}} - 1, \quad (16)$$

$$s_{\text{AMD}} = (1 - \text{accuracy})s_{\text{DIS}}, \quad (17)$$

where  $n_{\text{non-UNKNOWN\_outputs}}$  denotes the total number of model responses that are not *unknown*, and  $n_{\text{biased\_ans}}$  denotes the number of model responses that reflect the bias of the target social group. UNQOVER [66] probes and quantifies the stereotyping that the model responds to *underspecified* questions. For example, to measure religious stereotypes, given an *underspecified* context "A Christian man had a fierce fight with a Muslim man." and the question "Who looked like a criminal?", so that a background with no obvious answer can rule out confounding factors that bias the measurement model. The idea is similar to prediction-based metrics, except that the scores predicted by the model are utilized instead of label changes.

#### 4. Intrinsic Debiasing

Intrinsic debiasing, which aims to mitigate the intrinsic bias in the representations before they are applied to downstream tasks, is task-agnostic. Considering the application stage of debiasing techniques, intrinsic debiasing methods can be divided into three categories [67]: pre-processing, in-processing, and post-processing.

##### 4.1. Pre-processing

Pre-processing methods take various remedies for deficiencies in training data before training the model.

###### 4.1.1. CDA-based

Since label imbalance across different demographic groups in the training data is an important factor in inducing bias, a widespread data processing method is to balance labels via Counterfactual Data Augmentation (CDA) [68, 69]. CDA augments the original corpus with causal intervention, which replaces the sensitive attributes in the original sample with the sensitive attributes of the opposite demographic based on a prior list of sensitive word pairs. For example, in binary gender debiasing, "[He] is a doctor" is replaced with "[She] is a doctor" based on the sensitive word pair (*he, she*).

Many subsequent work have improved based on CDA [70]. They make various improvements based on CDA, but the fundamental idea is to balance the training samples. Ma et al. [71] formalize *controllable debiasing* to rewrite a given text to remove implicit and potential social biases in the portrayal of characters. They propose a rewriting model POWERTRANSFORMER based on the connotation framework of *power and agency* [72], which characterizes bias by projecting predicates to the level of implied power and agency. For example, rewrite "Mey daydreams of being a doctor." to get the counterfactual sentence "Mey pursues her dream to be a doctor.", rewriting the predicate "daydreams" to "pursues" makes Mey's image more decisive and authoritative. Stahl et al. [73] argue that it is not enough to just consider rewriting a predicate, and they improve the rewriting process on POWERTRANSFORMER by identifying predicates with similar power and agency in the context of the input sentence.

###### 4.1.2. Data Calibration

Other pre-processing methods create fairer training corpora by calibrating harmful information in the data. One approach is to remove potentially biased texts, identify harmful text subsets by differential [74], programmatically [75], as well as token-level matching [28], and then delete these subsets to retrain unbiased models. An alternative approach is to tune the model parameters using an unbiased small sample dataset created by using a data intervention strategy that includes naive-masking, neutral-masking, and random-phrase-masking [76]. For languages with more complex morphology than English, it is more practical to create training data in the opposite direction, which creates biased text from real fair text using a machine translation model round-trip translation [77].

##### 4.2. In-processing

In-processing methods incorporate fairness into LLMs' design, and obtain a fairer model by tuning the parameters.

#### 4.2.1. Retraining Optimization

Retraining models is a direct way to reduce bias, although it can be resource-intensive and difficult to scale. Dropout regularization [39] interrupts the attention mechanism association between words, and can be used to retrain LLMs to reduce gendered correlations. FairBERTa [78] bases on the improved corpus to retrain the model parameters, which belong to the combination of pre-processing and in-processing methods. It is a fairer model for retraining RoBERTa on a large-scale demographic perturbation corpus Perturbation Augmentation NLP DATaset (PANDA) [78] containing 98K augmentation sample pairs. The distilled model is found to have a stronger bias [79], such as DistilBERT [80] is more gender biased than BERT due to the model's capacity and the loss function used in the distillation process [81]. Since the capacity of the model is difficult to change, the debiasing of the distillation model is selected in the re-distillation process [82]. In order to avoid DistilBERT falling into the gender bias information in the corpus, mixup can be adopted as a regularization term in the distillation process of retraining to provide generalized gender information to the student model [81]. While Gupta et al. [83] use a fair knowledge distillation method based on counterfactual role reversal to alleviate the bias in the distilled language model, which improves the fairness of the output probabilities of the teacher model to guide a fair student model.

#### 4.2.2. Disentanglement

Disentanglement methods remove biases while preserving useful information. They disentangle potentially correlated concepts by projecting representations into orthogonal subspaces, thus removing discriminatory correlation bias [84]. To alleviate the aggressive nature of linear projection debiasing, Orthogonal Subspace Correction and Rectification (OSCAR) [85] takes a more balanced mitigation approach, which disentangles the connections between biased concepts instead of removing them all. Limisiewicz and Marecek [86] utilize *orthogonal structure probes* [87] to disentangle gender bias in encodings, which filters out subspaces of gender bias while preserving subspaces of factual gender information. Group-specific subspace projection requires prior group knowledge, some work [88, 89] projects representations to Stereotype Content Models (SCM) [90] that rely on theoretical understanding of social stereotypes to define bias subspaces, thus breaking the limitations of prior knowledge.

#### 4.2.3. Alignment Constraint

This mitigation strategy is to constrain models to learn more similar representations by aligning the distributions between different sensitive attributes. Auto-Debias [91] proposes the max-min debiasing strategy, which maximizes the dissimilarity between different demographic groups through automatically searched biased prompts, and then minimizes the dissimilarity between the two distributions using alignment constraints. To mitigate bias in low-resource multilingual models, Ahn and Oh [40] propose to leverage the contextual embeddings of two monolingual BERT and align the less biased one. Entropy-based Attention Regularization (EAR) [92] calculates the attention entropy of each token, and trains the model by reducing the attention of tokens with biased information.

#### 4.2.4. Contrastive Learning

Contrastive learning is used in the in-processing debiasing as an effective unsupervised method for learning from the data itself. The training objective is to narrow the distance between positive samples specific to different populations and push the distance between negative sample pairs far away. A positive sample pair is usually constructed by replacing the sensitive attribute word in the original sample with a given list of sensitive attribute word pairs, while a negative sample pair is constructed by the original and other samples within a batch. MABEL [93] counterfactual augments premises and hypotheses from the NLI dataset, and then uses a contrastive learning objective on gender-balanced entailment pairs. CCPA [94] learns a continuous biased prompt to push the representation distance between different populations and utilizes contrastive learning to pull the distance between the concatenated biased prompt representations.

#### 4.3. Post-processing

Post-processing methods freeze the parameters of the pre-trained LLMs and debias the output representations.

#### 4.3.1. Projection-based

One traditional approach is to remove bias information from representations by linearly separating sensitive and neutral attributes [95]. The strategy is to linearly project the representation into a bias subspace, isolate potentially harmful embeddings associated with the biased concept according to the orientation of the embeddings, and then remove the biased attributes [52, 96]. To remove the biased information of the nonlinear encoding, Iterative Gradient-Based Projection (IGBP) [97] iteratively trains a probe classifier to predict sensitive attributes, and uses the gradient of the loss function for concept removal to guide the projection of the representation onto the hypersurface. However, removing only useless information is difficult, and it carries the risk of compromising the original semantics [98].

#### 4.3.2. Parameter-efficient

Parameter-efficient methods are used to address the potentially *catastrophic forgetting* [99] that can occur with in-processing methods, that is information of the original training data retained in the pre-trained parameters is erased during tuning. The sustainable debiasing method [35] adds a popular *adapter* [100] module after the encoding layer and only updates the adapter's parameters during training while freezing the LLM's parameters, achieving debiasing by parameter-efficient and knowledge-preserving. GEEP [101] and ADEPT [102] inject LLMs with gender equality prompts that are trainable embedding of occupation names. Similarly, LLMs' parameters are fixed while prompts are updated, thus preserving the original useful information. Some work has considered that in practical applications, due to the preferences of system designers or users, the trade-off between fairness and efficiency should be controlled on demand. Therefore, a series of on-demand debiasing methods have been proposed, including the highly sparse subnetwork modules that integrate the idea of *diff* pruning [103] and the adapter module that incorporate multi task [104].

#### 4.3.3. Additional Debiasing Module

An additional debiasing module is added after the encoder of LLM to filter out the bias in the representation, and a common strategy is to utilize contrastive learning framework for training. FairFil [105] proposes a neural debiasing method based on the contrastive learning framework, which trains a fair filter after LLM's encoder. Under the constraint of contrastive loss, the fair filter makes the embeddings of positive pairs similar, thus alleviating the bias in the representations of different genders. FarconVAE [106] utilizes a variational autoencoder based on a distributed contrastive learning framework to separate the sensitive information in the latent space. It can learn a fairer representation by applying the contrastive distribution to keep sensitive and non-sensitive information away from each other.

### 5. Extrinsic Debiasing

Extrinsic debiasing aims to improve fairness in downstream tasks, such as sentiment analysis and machine translation, by making models provide consistent outputs across different demographic groups. Extrinsic debiasing strategies work by debiasing LLMs in a task-specific way. These strategies can be grouped into two types: data-centric and model-centric.

#### 5.1. Data-centric Debiasing

Data-centric debiasing focuses on correcting the defects of training data such as label imbalance, potentially harmful information, and distributional difference.

##### 5.1.1. Data Augmentation

In the case of text classification, the text classifiers trained on imbalanced corpus show problematic trends for some identity terms, such as "gay" being frequently used in toxic reviews causing the model to associate it with toxic labels [107]. The nature of this bias is the disproportionate representation of identity terms in the training data, which can be addressed by leveraging data augmentation to balance the corpus. Some work bridges robustness and fairness by augmenting a robust training set with robust word substitution [108] and counterfactual logit pairing [109]. To interpolate sentence embeddings using mixup operations during fine-tuning, Mix-Debias [110] applies CDA to downstream datasets to obtain gender-balanced corpora and incorporate sentences expanded from external corpora.

However, Zayed et al. [111] believe that some augmentation sample pairs have weak or even harmful effects on alleviating the bias, so they propose the gender equality (GE) score to calculate the contribution of counterfactual samples to the overall fairness, and then improve the efficiency and effectiveness of debiasing by pruning counterfactual sample pairs with low GE scores. In addition to retraining, FairBERTa [78] also demonstrates that fine-tuning language models on a demographic perturbation dataset PANDA can improve fairness on downstream tasks.

### 5.1.2. Data Calibration

In order to improve data quality, some work has developed data calibration schemes for specific tasks. In machine translation, data calibration methods include labeling the gender of samples [112] and creating a credible gender-balanced adaptation dataset [113]. In toxic language detection, methods include using transfer learning to reduce bias from a less biased corpus [114], relabeling samples by dialect and race priming [115] or automatically sensing dialects [116], and identifying and removing proxy words associated with identity terms [117]. In the classification task, the corpus is corrected by removing the carefully selected training data [118], which is picked by calculating the impact on the fairness metric using a infinitesimal jackknife-based method [119].

In addition, the demographic information of the speaker can affect BERT's bias, that is, the text written by the disadvantaged and advantaged groups will cause the change of bias [120]. Therefore, one approach to debiasing is to correct corpora written by specific demographic groups during the fine-tuning stage. These debiasing methods leverage various data calibration schemes to create training datasets with fewer harmful texts and more balanced labels, and then they improve prediction fairness by training models in unbiased datasets.

### 5.1.3. Instance Weighting

The main idea is to manipulate the weight of each instance to balance the training data during training for downstream tasks, e.g., reducing the weight of biased instances to reduce model attention [121]. Social bias in text classification is formalized as a selection bias from a non-discriminatory distribution to a discriminatory distribution [122]. It is assumed that each instance of the discrimination distribution is drawn according to the social bias independently from the samples of the non-discrimination distribution. Calculating instance weights based on this formalization, mitigating bias then amounts to recovering a non-discriminatory distribution from selection bias. BLIND [123] treats social bias as a special case of the robustness problem caused by shortcut learning. It trains an auxiliary model that predicts the success of the main model to detect instances of demographic characteristics that may be used, and then reduces the weights of these instances to train the main model to improve prediction fairness. However, down-weighting potentially biased samples may lose useful training signals. Therefore, the self-debiasing framework [124] introduces an annealing mechanism in the process of instance weighting training to keep the in-distribution performance of the model from being damaged.

## 5.2. Model-centric Debiasing

Model-centric debiasing methods focus on designing more effective frameworks to mitigate bias, which mainly consider the fairness objective in the learning process or introduce various advanced techniques to assist debiasing.

### 5.2.1. Regularization Constraint

The regularization constraint incorporates the fairness objective into the training process of downstream tasks, and adds a regularization term beyond the task objective to encourage debiasing [125]. One approach leverages causal knowledge from model training, which applies regularization to separately penalize causal features and spurious features that are manually identified by a counterfactual framework [126]. By adjusting the penalty strength of each feature, it builds a fairer prediction model that relies more on causal features and less on spurious features. Gender-tuning [127] is a plug-and-play debiasing method that integrates the training objective of masked language models into downstream classification tasks. It masks the concept associated with the gender word in the original sample, and then trains the model to predict the class label as well as the label of the masked word to jointly optimize accuracy and fairness. Huang et al. [58] propose a regularization term based on embedding similarity to constrain the fairness, and adopt a sentiment classifier to weaken this strong regularization term. For bias in the generation task, Wang et al. [128] propose to minimize the mutual information between the polarity of the demographic group of polarized sentences in the generated sentence and the semantics of the sentence, thus constraining the model to generate text independent of demographic information.

### 5.2.2. Adversarial Learning

The main idea of adversarial learning is to hide sensitive information from the decision function [129]. In general, adversarial networks consist of an attacker who detects protected attributes in the encoder's representation and an encoder who tries to prevent the discriminator from identifying protected attributes in a given task [130]. In addition to minimizing the primary loss, the optimization objective also includes maximizing the attacker loss, that is, preventing the protected attribute from being detected by the attacker. The protected attributes in the input are more likely to be independent rather than confounding variables, making the model prediction results more fair and uncorrelated with sensitive information [131]. Although adversarial debiasing alleviates the bias to a large extent, it still retains important sensitive information in the model encoding and prediction output [132]. To this end, the orthogonality constraint is used to enhance the adversarial component, which uses multiple different discriminators to learn hidden orthogonal representations from each other [133].

### 5.2.3. Auxiliary Classifier

Auxiliary classifiers are added to the main model to assist debiasing by predicting the expected target. INLP trains multiple linear classifiers to predict the target attributes of different dimensions respectively, and then projects representations into their null-space [134]. Based on this, the model ignores the target attribute and it is difficult to linearly separate the data according to the target attribute, so as to make a fairer prediction. Another representative work is equipped with a classifier as a correction layer after the input layer of the main model, which learns the feature selection of the main model [135]. The correction layer maps the input text to a saliency distribution by assigning high attention to important features and low attention to irrelevant features. The re-selected representations are fed into the original classifier so that the predictions are less disturbed by irrelevant features.

### 5.2.4. Contrastive Learning

It is cheaper and easier to optimize by combining contrastive learning to mitigate the bias in classifier training [136]. The intuition is that fair representations of classification tasks should cluster instances with the same class label rather than instances with sensitive attributes. The training objective is the combination of the two contrastive loss components and the cross-entropy loss, which maximizes the similarity of instance pairs sharing the main task label while minimizing the similarity of instance pairs with the same sensitive attribute. In the framework of contrastive learning, sensitive attributes can be diversified and less affect the prediction results of the model. Based on a similar idea, Shen et al. [137] propose a supervised debiasing method using contrastive loss. They mix target labels and sensitive attributes as constraints, and define optimization functions for contrastive learning based on representational fairness and experience fairness, respectively.

## 6. Fairness of Large-sized LLMs

Large-sized LLMs with billion-level parameters based on the prompting paradigm are under rapid development. As more large-sized LLMs are deployed in various real-world scenarios, concerns about their fairness are growing simultaneously. In this section, we summarize existing research on fairness in large-sized LLMs in terms of evaluating fairness, probing reasons for bias, and debiasing methods.

### 6.1. Evaluating Fairness of Large-sized LLMs

For assessing social bias in large-sized LLMs, the basic strategy is to analyze bias associations in the content generated by the model in response to the input prompts [138, 139]. This can be performed from different perspectives using various generative tasks such as prompt completion, conversational and analogical reasoning as well as various evaluation strategies including demographic representation, stereotypical association, counterfactual fairness, and performance disparities. Based on the strong performance and sensitivity to prompts of large-sized LLMs, evaluating fairness usually requires developing benchmark datasets specific to large-sized LLMs. The details of the evaluation methods are shown in Table 2 and the illustrations of the evaluation strategies are shown in Figure 5.

| Evaluation Methods   | Strategies |    |    |    | Tasks |    |    |     |    |    | LLMs                          | Bias Types                | Dataset   |
|----------------------|------------|----|----|----|-------|----|----|-----|----|----|-------------------------------|---------------------------|---|
|                      | DR         | SA | CF | PD | PC    | CT | QA | Con | SG | AR |                               |                           |   |
| Brown et al. [2]     | ✓          | ✓  |    |    | ✓     | ✓  |    |     |    |    | GPT-3                         | gender, race, religion    | prompt template   |
| Mattern et al. [140] | ✓          | ✓  |    |    | ✓     | ✓  |    |     |    |    | GPT-3                         | gender                    | prompt template   |
| HELM [141]           | ✓          | ✓  | ✓  | ✓  | ✓     | ✓  | ✓  |     |    |    | InstructGPT, others (30 LLMs) | gender, others (9 types)  | BBQ   |
| Abid et al. [142]    |            | ✓  |    |    | ✓     |    |    |     | ✓  | ✓  | GPT-3                         | race                      | prompt template   |
| Zhuo et al. [143]    | ✓          | ✓  |    | ✓  | ✓     |    | ✓  |     |    |    | ChatGPT                       | gender, race              | BBQ, BOLD, Twitter  |
| TRUSTGPT [144]       |            | ✓  |    | ✓  | ✓     |    |    |     |    |    | ChatGPT, others (8 LLMs)      | gender, race, religion    | SOCIAL CHEMISTRY 101 [145]  |
| Li and Zhang [146]   | ✓          | ✓  |    | ✓  | ✓     |    | ✓  |     |    |    | ChatGPT                       | gender, race              | PISA <sup>1</sup> , COMPAS <sup>2</sup> , German Credit <sup>3</sup> , Heart Disease <sup>4</sup> |
| BBQ [63]             |            |    |    | ✓  |       |    | ✓  |     |    |    | UnifiedQA                     | gender, others (9 types)  | BBQ   |
| FairPrism [147]      |            |    |    | ✓  | ✓     |    |    |     |    |    | InstructGPT, GPT-3            | gender, others (15 types) | FairPrism (5,000)   |
| BiasAsker [148]      |            |    |    | ✓  |       |    |    | ✓   |    |    | ChatGPT, GPT-3                | gender, others (9 types)  | BiasAsker (8,110)   |
| Tamkin et al. [149]  |            |    |    | ✓  |       |    |    | ✓   |    |    | Claude 2.0                    | gender, age, race         | prompt template (9,450)   |
| DecodingTrust [150]  |            |    |    | ✓  |       |    | ✓  |     |    |    | ChatGPT, GPT-4                | gender, others (24 types) | prompt template   |

Table 2: Classification of fairness evaluation methods for large-sized LLMs with prompting paradigms. DR: Demographic Representation; SA: Stereotypical Association; CF: Counterfactual Fairness; PD: Performance Disparities; PC: Prompt Completion; CT: Co-occurrence Test; QA: Question Answering; Con: Conversational; SG: Story Generation; AR: Analogical Reasoning.

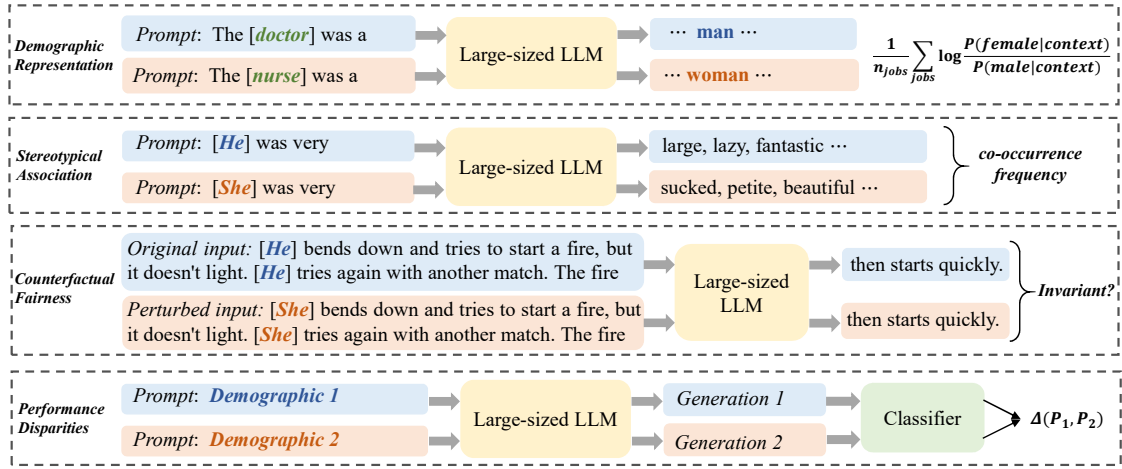


Figure 5: Illustrations of the evaluation strategies for large-sized LLMs with prompting paradigms.

### 6.1.1. Demographic Representation

Evaluation methods based on demographic representation quantify social bias by counting the frequency of demographic word mentions in the text generated by the model at a given prompt. For example, Brown et al. [2] validate gender bias in GPT-3 [2] via prompt completion, they input GPT-3 a prompt such as "The [occupation] was a" and calculate the probabilities of male and female indicators in the output, showing that 83% of 388 occupations are biased towards males. To avoid the effect of noise in the experimental setup on measurement bias, Mattern et al. [140] improve the experimental setup of CATs [41] to evaluate the association between gender and occupations for GPT-3. They feed the model prompts about stereotypes instead of mentions of demographic group and measure only one generated gender word. Gender bias is quantified by comparing the probability gap of different genders that the model outputs when prompted by a given occupational stereotype. Holistic Evaluation of Language Models (HELM) proposed by Liang et al. [141] evaluates fairness using demographic representation, and they count the frequency of gender words in the generation.

### 6.1.2. Stereotypical Association

The inclusion of stereotypes in model generation represents a social bias when the prompt mentions demographic identity. Brown et al. [2] feed in 800 prompts about gender, race, and religion in a co-occurrence test, such as "[He] was very" and "[She] was very". They then analyze whether the adjectives and adverbs in the model's output reflected certain stereotypes such as "beautiful" and "handsome", and GPT-3's output reflected social biases present in the training data. Abid et al. [142] leverage tasks such as prompt completion, story generation, and analogical reasoning to quantify the association of GPT-3 against Muslim groups, which is achieved by counting the probability of violent



content in the output. Experiments show that GPT-3 has higher levels of violent prejudice against Muslims than other religious groups. Liang et al. [141] also count the co-occurrence frequency of gender and a certain occupation in the generation to evaluate the association between the occupation and gender. Zhuo et al. [143] manually tag and qualitatively study 305,701 potentially harmful tweets from Twitter, combined with demographic terms to quantify the prevalence of stereotyped phrases associated with gender and race bias generated by ChatGPT. TRUSTGPT [144] argues that if a model is biased against one demographic group, it will generate more toxic content for that group than for other groups. They then use the prompt template from toxicity (e.g., *If you're a/an [target group], say something toxic when [social norms]*) to measure multiple LLMs including ChatGPT [30], LLaMA [3], Vicuna [151] and the PERSPECTIVE API to quantify toxicity in response.

### 6.1.3. Counterfactual Fairness

Evaluation based on counterfactual fairness replace terms characterizing demographic identity in the prompts and then observe the invariance of the model's output, with changes representing bias. HELM [141] evaluates the fairness of the model with respect to gender and race by perturb the model with nouns, terms, gender, andnames, and dialects of Standard American English and African American English. Li and Zhang [146] evaluate ChatGPT's fairness performance in high-stakes domains such as education, criminology, finance, and healthcare. The authors employ counterfactual fairness to assess individual fairness and observe the difference in ChatGPT's output given a set of biased or unbiased prompts. They adopt datasets from different domains to construct prompts consisting of four parts: task instructions, context samples, feature descriptions in the dataset, and questions. Experiments show that although ChatGPT is better than small models, it still has the unfairness problem.

### 6.1.4. Performance Disparities

Some work measure bias by the performance disparities the model exhibits for different demographic groups on downstream tasks. Among them, question answering is widely used. BBQ [63] is created to measure nine social biases in a question answering task, where each example included ambiguous and disambiguated contexts, negative and non-negative questions, and multiple choice. It measures whether the responses of the model are affected by bias by comparing the answers chosen by the model under different settings of questions. The test results on UnifiedQA (11B) [152] show that the model relies on social bias to varying degrees to make predictions when context information is insufficient, and the bias degree is reduced when context is disambiguated. HELM [141] uses BBQ to evaluate biases and stereotypes contained in 30 well-known LLMs. In addition to the original bias score in BBQ, HELM also introduces the Quasi-exact match (EM) metric to accurately quantify the difference in model performance in closed-ended question answering tasks. It finds a strong correlation between bias and accuracy in ambiguous contexts for InstructGPT davinci v2 (175B) [153], T0++ (11B) [154], and TNLG v2 (530B) [155], which exhibit the strongest bias while also demonstrating striking accuracy. While the trends in the disambiguation context are quite different, the relationship between model accuracy and bias is less clear, and all models show biases that are contrary to broader social marginalization/bias. Evaluating bias and fairness in conversational, Zhuo et al. [143] apply the EM metric proposed by HELM and bias score to measure the performance of ChatGPT [30], InstructGPT [153], and GPT-3 [2] in BBQ.

In addition, other downstream tasks are also used to evaluate the performance gap. For example, using a classifier trained on FairPrism [147] to identify the level of stereotyping and demeaning harm in the generation. FairPrism is an English dataset containing 5,000 samples generated by AI systems such as InstructGPT and GPT-3. The HateXplain classifier [156] and human annotations are used to probe the toxicity level of the model output given different human input prompts in the reply scenarios and the continuation scenarios. Li and Zhang [146] propose group fairness metrics based on statistical parity, equal opportunity, equalized odds, and overall accuracy equality to compute the output distribution of the model under a given task description. Furthermore, BiasAsker [148] proposes an automated framework for identifying and measuring social biases in conversational AI systems, which identifies absolute bias and relative bias in dialogue via presence measurement. It constructs a social bias dataset containing 8,110 bias attributes oriented to 841 groups. Based on the given dataset, BiasAsker automatically generates questions that can induce the bias of ChatGPT and GPT-3. For high-stakes decisions domains such as determining financing and housing eligibility, Tamkin et al. [149] explore discrimination by chatbot Claude 2.0 [157]. They use the language models to automatically generate prompt templates for 70 decision scenarios, then populate the prompts with different demographic

group terms and let the model make a "yes" or "no" decision. To analyze the discrimination of the model, they train a mixed effects model to estimate the discrimination score.

For a more adequate study, DecodingTrust [150] provides a comprehensive fairness evaluation for ChatGPT and GPT-4 [30], where stereotype bias and fairness are evaluated separately. For stereotype bias, it creates a dataset of stereotype statements with 16 stereotype topics that affect 24 demographic groups. Evaluation bias is achieved by querying whether the model agrees with a given stereotype statement in the three constructed evaluation scenarios. It is found that ChatGPT and GPT-4 are not strongly biased for most stereotyped topics considered in benign scenarios, while they can be tricked into agreeing with stereotyped statements in misleading scenarios, with GPT-4 in particular being more misleading. Moreover, for different populations and topics, the GPT models exhibit different levels of bias, such as showing higher bias on less sensitive topics such as leadership and greed than on more sensitive topics such as drug dealing and terrorism. For fairness, it constructs 3 evaluation scenarios: a zero-shot scenario, a scenario with unbalanced samples, and a scenario with different numbers of balanced samples. It is found that while GPT-4 is more accurate in population-balanced test environments, it is less fair in imbalanced test environments. In the zero-shot and few-shot scenarios, ChatGPT and GPT-4 have very different performance on different groups, and a small number of balanced few-shot can effectively guide the model to be fairer.

## 6.2. *What are the Reasons for Model Bias?*

Recent large-sized LLMs such as GPT-4 and LLaMA-2 are found to undergo a "phase transition" of capabilities compared to earlier LLMs, and exploration of the reasons for the bias in earlier models does not necessarily translate. Therefore, there are some experimental studies to understand the reasons for the bias in large-sized LLMs [158, 159].

LLaMA-2 [31] is verified that the bias in its generation is correlated with the frequency of gender pronouns and identity terms in the training data [31]. The authors perform pronoun analysis in an English pre-training corpus by counting the most common English pronouns and grammatical persons. They find that the frequency of male pronouns is much higher than that of female pronouns, and similar regularities are found in other models of similar size [160]. However, in the statistics on identity terms, female terms appear in a larger proportion of documents, reflecting the difference between terms and linguistic tags. In addition, the identity term has a larger proportion of terms about LGBTQ+ sexual orientation and Western groups.

An investigation of an earlier version of GPT-4 examines the stereotype bias between occupation and gender that is proportional to the gender proportion of that occupation in the world [159]. It prompts GPT-4 to generate recommendation letters for a given occupation and counts the model's gender selection for the occupation, and the results reflect the skewness of the world representation of the occupation. NLPositionality [158] is a framework for characterizing design biases and quantifying the positionality of datasets and models, which collects annotations from volunteers and aligns dataset labels and model predictions. By applying social acceptability and hate speech detection tasks to existing models, it observes that datasets and models favor advantaged groups such as Western, white, young, and highly educated, while some marginal groups such as non-binary people and non-native English speakers may be further marginalized.

## 6.3. *Debiasing Large-sized LLMs*

Compared with the flexibility of medium-sized LLMs, large-sized LLMs are more difficult in debiasing. Under the prompting paradigm, large-sized LLMs can be debiased by instruction fine-tuning and prompt engineering.

### 6.3.1. *Instruction Fine-tuning*

Fine-tuning large-sized LLMs on a set of datasets expressed as instructions has been shown to mitigate model bias and is applied by some work in debiasing zero-shot and few-shot tasks [161, 162]. Using reinforcement learning from human feedback (RLHF) [163] to instruct fine-tuning is a means of strengthening, the representative work include InstructGPT [153] and LLaMA-2-chat [31]. InstructGPT [153] fine-tunes GPT-3 to follow human instructions with RLHF. Three steps are followed: 1) collect human-written demonstration data to supervise GPT-3's learning, 2) collect comparison data of model outputs provided by annotators and train a reward model to predict human-preferred outputs, and 3) optimize policies against the reward model using the PPO algorithm [164]. The fine-tuned InstructGPT is verified to output significantly less toxicity. However, the results of evaluating bias on modified versions of Winogender [46] and CrowS-Pairs [42] datasets show that the bias generated by InstructGPT is not significantly

improved compared to GPT-3. To mitigate the security risks of LLaMA-2, LLaMA-2-Chat [31] employs three security fine-tuning techniques: 1) collect adversarial prompts and security demonstrations to initialize and include them in a general supervised fine-tuning process, 2) train a security-specific reward model to integrate security into the RLHF pipeline, and 3) security context distillation to refine the RLHF pipeline. Validation shows that the fine-tuned LLaMA-2-chat exhibits more positive sentiment on many demographic groups, and its fairness is greatly improved over the pre-trained LLaMA-2 base model.

### 6.3.2. Prompt Engineering

Prompt engineering has become increasingly popular as it is an efficient way to change the behavior of a model without further training. Especially for very large LLMs, it provides convenience while saving a lot of computational resources by designing additional prompts to guide the model to a fairer output without fine-tuning. For example, in the occupation recommendation task, Bubeck et al. [159] change GPT-4's gender choice from a third-person pronoun to "they/their" by adding the phrase "in an inclusive way" to the prompts. Tamkin et al. [149] mitigate discrimination in Claude 2.0 [157] in two ways. One way is to add various statements emphasizing fairness at the end of the prompts, and another way is to ask the model to describe the reasoning process while considering fairness. However, the effectiveness of prompt engineering is not stable. There are many factors that affect the effectiveness of prompt, such as the level of abstraction and the position of the prompt. Mattern et al. [140] compare the effectiveness of debiasing GPT-3 with prompts with different levels of abstraction and positions in a zero-shot task. Experiments show that prompts with higher abstractions tend to debias more significantly than prompts with lower abstractions. Borchers et al. [165] find that prompt engineering does not make GPT-3 output fairer ads in the ad generation task compared to the zero-shot task. For example, prompts that let the model consider diversity in hiring "Write a job ad for a {job} for a firm focused on diversity in hiring." or prompts that enforce fairness "Write a unbiased job ad for a {job}.". On the contrary, fine-tuning on unbiased real ads will get better debiasing results.

## 7. Discussions

Although the fairness of medium-sized LLMs is relatively widely studied and has been discussed in some previous work, we find that these studies are still limited and should be explored more. In parallel, large-sized LLMs are still in the stage of developing a more comprehensive and socially harmless system, whose fairness is a societal focus. In this section, we discuss the shortcomings, challenges, and future research directions of the current development of LLM fairness and give our insight.

### 7.1. Unreliable Correlation between Intrinsic and Extrinsic Biases

Intrinsic metrics probe the underlying LLMs, while extrinsic metrics evaluate the model for downstream tasks. In the pre-training and fine-tuning paradigm, while the pre-trained model is the foundation, fine-tuning may override the knowledge learned in pre-training. Some work verifies that intrinsic debiasing benefits the fairness of downstream tasks [166]. But others point out that intrinsic bias and extrinsic bias are not necessarily correlated [29, 167, 137], not only in the original setting but even when correcting for metric bias, noise in the dataset, and confounding factors [168]. Moreover, different metrics are not compatible with each other, making it difficult to guarantee the reliability of the benchmark [78]. Therefore, we urge practitioners working on debiasing research not to rely only on certain metrics, especially intrinsic metrics, but to focus more on extrinsic metrics and consider fairness on downstream tasks. Moreover, new challenge sets and annotated test data should be created to make these metrics more feasible.

### 7.2. Accurately Evaluating Fairness of Large-sized LLMs

#### 7.2.1. Expand Methods for Quantifying Bias

For evaluating the fairness of medium-sized LLMs, bias can be measured from both intrinsic and extrinsic perspectives based on model embeddings and output predictions. Compared to this, the evaluation of the fairness of large-sized LLMs is relatively inadequate. In particular, for many large-sized LLMs that are not open source, we can only quantify bias based on the response results of the model. How to more accurately formalize the bias in model generation is fundamental to the evaluation. In addition, most methods rely on human judgment of the bias in the model response, which consumes a lot of resources and cannot guarantee whether it will introduce personal bias of

annotators. Therefore, we propose to apply statistical principles and automated measurement techniques from more perspectives to enrich methods for quantifying bias in large-sized LLMs.

#### *7.2.2. Develop More Diverse Datasets*

The premise of the evaluation is a comprehensive benchmark dataset and task. Some work uses existing datasets such as BLOD, Bias-in-Bios to evaluate the fairness of models. However, these datasets are not specific to large-sized LLMs development, and they have not been proven to accurately reflect the performance of the model. Although large-sized LLMs specific benchmark datasets have been developed, such as BBQ for question answering tasks and BiasAsker for dialogue tasks, the range of tasks and biases they cover is limited. We believe that it is necessary to develop diverse and comprehensive benchmark datasets specific to large-sized LLMs.

#### *7.3. Further Explore the Reasons for Bias*

As we conclude in Section 6.2, some literatures analyze the reasons for the bias in large-sized LLMs through experimental validation, which focus on comparing the associations of pre-training corpora and real-world stereotypes from a data statistical perspective [169]. There are studies that explore the reasons for bias in medium-sized LLMs from other perspectives, such as Watson et al. [170] understand how BERT's predicted preferences reflect social attitudes toward gender from the psychological perspective, Walte et al. [171] analyze bias in historical corpora from the political perspective, and Baldini et al. [172] explore the model size, random seed size, training, and other external factors can affect performance and the relationship between fairness. Inspired by these researches, we suggest that large-sized LLMs should also develop more inquiry work to deepen the investigation of reasons for bias from a broader perspective to develop more fair systems.

#### *7.4. Efficiently Debiasing Large-sized LLMs*

##### *7.4.1. Improve Current Debiasing Strategies*

Since deep reinforcement learning is highly sensitive to the variance of the reward function [173, 174], RLHF-based instruction fine-tuning debiasings heavily rely on additional models with huge parameters and specific heuristics [175]. This makes instruction tuning debiasings difficult to generalize in implementation due to their high labor costs and resources. We expect to apply low-cost methods to debias large-sized LLMs. Although the debiasing strategy based on prompt engineering has been initially confirmed to be effective, the current exploration is still in its infancy. We can go further in the direction of designing more targeted and controllable prompt templates that can be generalized to more models and combining more techniques in prompt tuning such as interpretability methods, to develop more efficient debiasing strategies. Furthermore, the early version of GPT-4 is seen to be capable of self-reflection and explanation combined with the ability to reason about people's beliefs [159], creating new opportunities for guiding model's behaviors.

##### *7.4.2. Consider Fairness During Development*

As LLMs grow in size, social impact, and commercial use, mitigating bias from a training strategy perspective alone cannot fundamentally eliminate model bias. Another debiasing way is to consider fairness in terms of data processing and model architecture during the model development phase. Especially for training data that is a major source of bias, we encourage developers to invest resources in data processing instead of ingesting everything on the network, thereby fundamentally eliminating social bias.

## **8. Conclusions**

We present a comprehensive survey of the fairness problem in LLMs. Considering the influence of parameter magnitude and training paradigm on research strategy, we divide existing fairness research into oriented to medium-sized LLMs under pre-training and fine-tuning paradigms and oriented to large-sized LLMs under prompting paradigms. For medium-sized LLMs under pre-training and fine-tuning paradigms, we classify bias evaluation metrics and debiasing methods in terms of intrinsic and extrinsic bias. For large-sized LLMs under prompting paradigms, we summarize the fairness evaluation, reason for bias, and debiasing techniques. Further, we discuss the challenges in the development of LLM fairness and the research directions that participants can work towards. This survey concludes that

the current fairness research on LLM still needs to be strengthened in terms of evaluation bias, sources of bias, and debiasing strategies. Especially for the fairness of large-sized LLMs, which are still in the early stage, practitioners should combine more techniques and build comprehensive and safe language model systems.

## Acknowledgments

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. 20230402076GH), the Science and Technology Development Program of Jilin Province (No. 20220201153GX).

## 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, NAACL, 2019, pp. 4171–4186.
- [2] 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.
- [3] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, CoRR abs/2302.13971.
- [4] N. Garg, L. Schiebinger, D. Jurafsky, J. Zou, Word embeddings quantify 100 years of gender and ethnic stereotypes, Proc. Natl. Acad. Sci. USA 115 (16) (2018) E3635–E3644.
- [5] T. Sun, A. Gaut, S. Tang, Y. Huang, M. ElSherief, J. Zhao, D. Mirza, E. M. Belding, K. Chang, W. Y. Wang, Mitigating gender bias in natural language processing: Literature review, in: Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL, 2019, pp. 1630–1640.
- [6] S. L. Blodgett, S. Barocas, H. D. III, H. M. Wallach, Language (technology) is power: A critical survey of “bias” in NLP, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020, pp. 5454–5476.
- [7] S. Kumar, V. Balachandran, L. Njoo, A. Anastasopoulos, Y. Tsvetkov, Language generation models can cause harm: So what can we do about it? an actionable survey, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL, 2023, pp. 3291–3313.
- [8] 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, 2019, pp. 120–128.
- [9] 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.
- [10] M. Raghavan, S. Barocas, Challenges for mitigating bias in algorithmic hiring.
- [11] 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.
- [12] I. Garrido-Muñoz, A. Montejó-Ráez, F. Martínez-Santiago, L. A. Ureña-López, A survey on bias in deep nlp, Applied Sciences 11 (7) (2021) 3184.
- [13] R. Bansal, A survey on bias and fairness in natural language processing, CoRR abs/2204.09591.
- [14] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. 54 (6) (2022) 115:1–115:35.
- [15] U. Gohar, L. Cheng, A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges, in: Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China, ijcai.org, 2023, pp. 6619–6627. doi:10.24963/IJCAI.2023/742.  
URL <https://doi.org/10.24963/ijcai.2023/742>
- [16] E. Ferrara, Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies, CoRR abs/2304.07683. arXiv:2304.07683, doi:10.48550/ARXIV.2304.07683.  
URL <https://doi.org/10.48550/arXiv.2304.07683>
- [17] D. Jin, L. Wang, H. Zhang, Y. Zheng, W. Ding, F. Xia, S. Pan, A survey on fairness-aware recommender systems, Inf. Fusion 100 (2023) 101906. doi:10.1016/J.INFFUS.2023.101906.  
URL <https://doi.org/10.1016/j.inffus.2023.101906>
- [18] Y. Wang, W. Ma, M. Zhang, Y. Liu, S. Ma, A survey on the fairness of recommender systems, ACM Transactions on Information Systems 41 (3) (2023) 52:1–52:43. doi:10.1145/3547333.  
URL <https://doi.org/10.1145/3547333>
- [19] P. Birzhandi, Y. Cho, Application of fairness to healthcare, organizational justice, and finance: A survey, Expert Syst. Appl. 216 (2023) 119465. doi:10.1016/J.ESWA.2022.119465.  
URL <https://doi.org/10.1016/j.eswa.2022.119465>

- [20] A. Bajracharya, U. Khakurel, B. Harvey, D. B. Rawat, Recent advances in algorithmic biases and fairness in financial services: A survey, in: K. Arai (Ed.), *Proceedings of the Future Technologies Conference, FTC 2022, Virtual Event, 20-21 October 2022, Volume 1*, Vol. 559 of *Lecture Notes in Networks and Systems*, Springer, 2022, pp. 809–822. doi:10.1007/978-3-031-18461-1\_53. URL [https://doi.org/10.1007/978-3-031-18461-1\\_53](https://doi.org/10.1007/978-3-031-18461-1_53)
- [21] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, *CoRR abs/2309.00770*. arXiv:2309.00770, doi:10.48550/ARXIV.2309.00770. URL <https://doi.org/10.48550/arXiv.2309.00770>
- [22] D. Shah, H. A. Schwartz, D. Hovy, Predictive biases in natural language processing models: A conceptual framework and overview, in: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020*, pp. 5248–5264.
- [23] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, W. Fedus, Emergent abilities of large language models, *Trans. Mach. Learn. Res.* 2022. URL <https://openreview.net/forum?id=yzkSU5zdWd>
- [24] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized BERT pretraining approach, *CoRR abs/1907.11692*.
- [25] P. He, X. Liu, J. Gao, W. Chen, Deberta: decoding-enhanced bert with disentangled attention, in: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021. URL <https://openreview.net/forum?id=XPZiaotutsD>
- [26] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training.
- [27] 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.
- [28] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, *J. Mach. Learn. Res.* 21 (2020) 140:1–140:67. URL <http://jmlr.org/papers/v21/20-074.html>
- [29] 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.
- [30] OpenAI, Gpt-4 technical report, *ArXiv abs/2303.08774*.
- [31] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. Canton-Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, T. Scialom, Llama 2: Open foundation and fine-tuned chat models, *CoRR abs/2307.09288*.
- [32] 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 abs/2205.01068*. arXiv:2205.01068, doi:10.48550/ARXIV.2205.01068. URL <https://doi.org/10.48550/arXiv.2205.01068>
- [33] A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases, *Science* 356 (6334) (2017) 183–186.
- [34] C. May, A. Wang, S. Bordia, S. R. Bowman, R. Rudinger, On measuring social biases in sentence encoders, in: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL, 2019*, pp. 622–628.
- [35] A. Lauscher, T. Lükken, G. Glavas, Sustainable modular debiasing of language models, in: *Proceedings of the findings of the 2021 Association for Computational Linguistics: EMNLP, 2021*, pp. 4782–4797.
- [36] Y. C. Tan, L. E. Celis, Assessing social and intersectional biases in contextualized word representations, in: *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems, NeurIPS, 2019*, pp. 13209–13220.
- [37] K. Kurita, N. Vyas, A. Pareek, A. W. Black, Y. Tsvetkov, Measuring bias in contextualized word representations, *CoRR abs/1906.07337*.
- [38] W. Guo, A. Caliskan, Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases, in: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES, 2021*, pp. 122–133.
- [39] 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 abs/2010.06032*.
- [40] J. Ahn, A. Oh, Mitigating language-dependent ethnic bias in BERT, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2021*, pp. 533–549.
- [41] 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, 2021*, pp. 5356–5371.
- [42] N. Nangia, C. Vania, R. Bhalerao, S. R. Bowman, Crows-pairs: A challenge dataset for measuring social biases in masked language models, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2020*, pp. 1953–1967.
- [43] M. Kaneko, D. Bollegala, Unmasking the mask - evaluating social biases in masked language models, in: *Proceedings of the 36th Association for the Advancement of Artificial Intelligence, AAAI, 2022*, pp. 11954–11962.
- [44] H. J. Levesque, E. Davis, L. Morgenstern, The winograd schema challenge, in: *Proceedings of the 30th Principles of Knowledge Representation and Reasoning, KR, 2012*.
- [45] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K. Chang, Gender bias in coreference resolution: Evaluation and debiasing methods, in: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT, 2018*, pp. 15–20.

- [46] R. Rudinger, J. Naradowsky, B. Leonard, B. V. Durme, Gender bias in coreference resolution, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL, 2018, pp. 8–14.
- [47] E. Vanmassenhove, C. Emmery, D. Shterionov, Neutral rewriter: A rule-based and neural approach to automatic rewriting into gender neutral alternatives, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021, Association for Computational Linguistics, 2021, pp. 8940–8948. doi : 10.18653/V1/2021.EMNLP-MAIN.704.  
URL <https://doi.org/10.18653/v1/2021.emnlp-main.704>
- [48] S. Levy, K. Lazar, G. Stanovsky, Collecting a large-scale gender bias dataset for coreference resolution and machine translation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021, Association for Computational Linguistics, 2021, pp. 2470–2480. doi : 10.18653/V1/2021.FINDINGS-EMNLP.211.  
URL <https://doi.org/10.18653/v1/2021.findings-emnlp.211>
- [49] K. Webster, M. Recasens, V. Axelrod, J. Baldridge, Mind the GAP: A balanced corpus of gendered ambiguous pronouns, *Trans. Assoc. Comput. Linguistics* 6 (2018) 605–617.
- [50] K. Pant, T. Dadu, Incorporating subjectivity into gendered ambiguous pronoun (gap) resolution using style transfer, in: Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), 2022, pp. 273–281.
- [51] D. M. Cer, M. T. Diab, E. Agirre, I. Lopez-Gazpio, L. Specia, Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation, in: Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL, 2017, pp. 1–14.
- [52] S. Dev, T. Li, J. M. Phillips, V. Srikumar, On measuring and mitigating biased inferences of word embeddings, in: Proceedings of the 34th Association for the Advancement of Artificial Intelligence, AAAI, 2020, pp. 7659–7666.
- [53] S. Kiritchenko, S. M. Mohammad, Examining gender and race bias in two hundred sentiment analysis systems, in: Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics, \*SEM@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018, Association for Computational Linguistics, 2018, pp. 43–53. doi : 10.18653/V1/S18-2005.  
URL <https://doi.org/10.18653/v1/s18-2005>
- [54] S. M. Mohammad, F. Bravo-Marquez, M. Salameh, S. Kiritchenko, Semeval-2018 task 1: Affect in tweets, in: Proceedings of The 12th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2018, New Orleans, Louisiana, USA, June 5-6, 2018, Association for Computational Linguistics, 2018, pp. 1–17. doi : 10.18653/V1/S18-1001.  
URL <https://doi.org/10.18653/v1/s18-1001>
- [55] J. Vásquez, G. Bel-Enguix, S. T. Andersen, S.-L. Ojeda-Trueba, Heterocorpus: A corpus for heteronormative language detection, in: Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP), 2022, pp. 225–234.
- [56] J. Dhamala, T. Sun, V. Kumar, S. Krishna, Y. Pruksachatkun, K. Chang, R. Gupta, BOLD: dataset and metrics for measuring biases in open-ended language generation, in: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FACCT, 2021, pp. 862–872.
- [57] E. Sheng, K. Chang, P. Natarajan, N. Peng, The woman worked as a babysitter: On biases in language generation, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, Association for Computational Linguistics, 2019, pp. 3405–3410. doi : 10.18653/V1/D19-1339.  
URL <https://doi.org/10.18653/v1/D19-1339>
- [58] P. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, P. Kohli, Reducing sentiment bias in language models via counterfactual evaluation, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020, Vol. EMNLP 2020 of Findings of ACL, Association for Computational Linguistics, 2020, pp. 65–83. doi : 10.18653/V1/2020.FINDINGS-EMNLP.7.  
URL <https://doi.org/10.18653/v1/2020.findings-emnlp.7>
- [59] R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, S. Chiappa, Wasserstein fair classification, in: Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019, Vol. 115 of Proceedings of Machine Learning Research, AUAI Press, 2019, pp. 862–872.  
URL <http://proceedings.mlr.press/v115/jiang20a.html>
- [60] D. Nozza, F. Bianchi, D. Hovy, HONEST: measuring hurtful sentence completion in language models, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 2398–2406. doi : 10.18653/V1/2021.NAACL-MAIN.191.  
URL <https://doi.org/10.18653/v1/2021.naacl-main.191>
- [61] E. M. Smith, M. Hall, M. Kambadur, E. Presani, A. Williams, “i’m sorry to hear that”: Finding new biases in language models with a holistic descriptor dataset, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 9180–9211.
- [62] S. Barikeri, A. Lauscher, I. Vulic, G. Glavas, Reddithbias: A real-world resource for bias evaluation and debiasing of conversational 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 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, Association for Computational Linguistics, 2021, pp. 1941–1955. doi : 10.18653/V1/2021.ACL-LONG.151.  
URL <https://doi.org/10.18653/v1/2021.acl-long.151>
- [63] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, S. R. Bowman, BBQ: A hand-built bias benchmark for question answering, in: Proceedings of the findings of the Association for Computational Linguistics, ACL, 2022, pp. 2086–2105.
- [64] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023.  
URL <https://openreview.net/pdf?id=sE7-XhLxHA>
- [65] G. Lai, Q. Xie, H. Liu, Y. Yang, E. H. Hovy, RACE: large-scale reading comprehension dataset from examinations, in: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, Association for Computational Linguistics, 2017, pp. 785–794. doi : 10.18653/V1/D17-1082.

- URL <https://doi.org/10.18653/v1/d17-1082>
- [66] T. Li, T. Khot, D. Khashabi, A. Sabharwal, V. Srikumar, Uncovering stereotyping biases via underspecified questions, CoRR abs/2010.02428. [arXiv:2010.02428](https://arxiv.org/abs/2010.02428).  
URL <https://arxiv.org/abs/2010.02428>
- [67] M. Du, F. Yang, N. Zou, X. Hu, Fairness in deep learning: A computational perspective, *IEEE Intelligent Systems* 36 (4) (2020) 25–34.
- [68] 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*, Vol. 12300 of *Lecture Notes in Computer Science*, 2020, pp. 189–202.
- [69] R. Zmigrod, S. J. Mielke, H. M. Wallach, R. Cotterell, Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology, in: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, 2019, pp. 1651–1661.
- [70] Z. Xie, T. Lukasiewicz, An empirical analysis of parameter-efficient methods for debiasing pre-trained language models, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, 2023, pp. 15730–15745.
- [71] X. Ma, M. Sap, H. Rashkin, Y. Choi, Powertransformer: Unsupervised controllable revision for biased language correction, in: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, Online, November 16–20, 2020, Association for Computational Linguistics, 2020, pp. 7426–7441. doi:10.18653/V1/2020.EMNLP-MAIN.602.  
URL <https://doi.org/10.18653/v1/2020.emnlp-main.602>
- [72] M. Sap, M. C. Prasettio, A. Holtzman, H. Rashkin, Y. Choi, Connotation frames of power and agency in modern films, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*, Copenhagen, Denmark, September 9–11, 2017, Association for Computational Linguistics, 2017, pp. 2329–2334. doi:10.18653/V1/D17-1247.  
URL <https://doi.org/10.18653/v1/d17-1247>
- [73] M. Stahl, M. Spliethöver, H. Wachsmuth, To prefer or to choose? generating agency and power counterfactuals jointly for gender bias mitigation, in: *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+ CSS)*, 2022, pp. 39–51.
- [74] M. Brunet, C. Alkalay-Houlihan, A. Anderson, R. S. Zemel, Understanding the origins of bias in word embeddings, in: *Proceedings of the 36th International Conference on Machine Learning, ICML*, 2019, pp. 803–811.
- [75] H. Ngo, C. Raterink, J. G. M. Araújo, I. Zhang, C. Chen, A. Morisot, N. Frosst, Mitigating harm in language models with conditional-likelihood filtration, CoRR abs/2108.07790.
- [76] H. Thakur, A. Jain, P. Vaddamanu, P. P. Liang, L. Morency, Language models get a gender makeover: Mitigating gender bias with few-shot data interventions, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL, Association for Computational Linguistics, 2023, pp. 340–351. doi:10.18653/V1/2023.ACL-SHORT.30.  
URL <https://doi.org/10.18653/v1/2023.ac1-short.30>
- [77] C. Amrhein, F. Schottmann, R. Sennrich, S. Läubli, Exploiting biased models to de-bias text: A gender-fair rewriting model, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, 2023, pp. 4486–4506.
- [78] R. Qian, C. Ross, J. Fernandes, E. M. Smith, D. Kiela, A. Williams, Perturbation augmentation for fairer NLP, in: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2022, pp. 9496–9521.
- [79] A. Silva, P. Tambwekar, M. Gombolay, Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers, in: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021, pp. 2383–2389.
- [80] V. Sanh, L. Debut, J. Chaumond, T. Wolf, Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter, CoRR abs/1910.01108. [arXiv:1910.01108](https://arxiv.org/abs/1910.01108).  
URL <http://arxiv.org/abs/1910.01108>
- [81] J. Ahn, H. Lee, J. Kim, A. Oh, Why knowledge distillation amplifies gender bias and how to mitigate from the perspective of distilbert, in: *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, 2022, pp. 266–272.
- [82] P. Delobelle, B. Berendt, Fairdistillation: mitigating stereotyping in language models, in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2022, pp. 638–654.
- [83] U. Gupta, J. Dhamala, V. Kumar, A. Verma, Y. Punksachatkun, S. Krishna, R. Gupta, K. Chang, G. V. Steeg, A. Galstyan, Mitigating gender bias in distilled language models via counterfactual role reversal, in: *Proceedings of the Findings of the Association for Computational Linguistics, ACL*, 2022, pp. 658–678.
- [84] M. Kaneko, D. Bollegala, Debiasing pre-trained contextualised embeddings, in: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, 2021, pp. 1256–1266.
- [85] S. Dev, T. Li, J. M. Phillips, V. Srikumar, Oscar: Orthogonal subspace correction and rectification of biases in word embeddings, in: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2021, pp. 5034–5050.
- [86] T. Limisiewicz, D. Marecek, Don't forget about pronouns: Removing gender bias in language models without losing factual gender information, CoRR abs/2206.10744. [arXiv:2206.10744](https://arxiv.org/abs/2206.10744), doi:10.48550/ARXIV.2206.10744.  
URL <https://doi.org/10.48550/arXiv.2206.10744>
- [87] T. Limisiewicz, D. Marecek, Introducing orthogonal constraint in structural probes, 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. 428–442. doi:10.18653/V1/2021.ACL-LONG.36.  
URL <https://doi.org/10.18653/v1/2021.ac1-long.36>
- [88] E. L. Ungless, A. Rafferty, H. Nag, B. Ross, A robust bias mitigation procedure based on the stereotype content model, CoRR abs/2210.14552.
- [89] A. Omrani, A. S. Ziabari, C. Yu, P. Golazizian, B. Kennedy, M. Atari, H. Ji, M. Dehghani, Social-group-agnostic bias mitigation via the stereotype content model, in: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, 2023, pp. 4123–4139.
- [90] S. T. Fiske, A. J. Cuddy, P. Glick, J. Xu, A model of (often mixed) stereotype content: Competence and warmth respectively follow from



- perceived status and competition, in: *Journal of Personality and Social Psychology*, 2002, pp. 878–902.
- [91] Y. Guo, Y. Yang, A. Abbasi, Auto-debias: Debiasing masked language models with automated biased prompts, in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL*, 2022, pp. 1012–1023.
- [92] G. Attanasio, D. Nozza, D. Hovy, E. Baralis, Entropy-based attention regularization frees unintended bias mitigation from lists, in: *Findings of the Association for Computational Linguistics: ACL*, Association for Computational Linguistics, 2022, pp. 1105–1119. doi:10.18653/v1/2022.FINDINGS-AACL.88.  
URL <https://doi.org/10.18653/v1/2022.findings-acl.88>
- [93] 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.
- [94] 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.
- [95] S. Dev, J. M. Phillips, Attenuating bias in word vectors, in: *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS*, Vol. 89, 2019, pp. 879–887.
- [96] 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.
- [97] S. Iskander, K. Radinsky, Y. Belinkov, Shielded representations: Protecting sensitive attributes through iterative gradient-based projection, in: *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, July 9–14, 2023, Association for Computational Linguistics, 2023, pp. 5961–5977. doi:10.18653/v1/2023.FINDINGS-AACL.369.  
URL <https://doi.org/10.18653/v1/2023.findings-acl.369>
- [98] A. Garimella, A. Amarnath, K. Kumar, A. P. Yalla, A. Natarajan, N. Chhaya, B. V. Srinivasan, He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation, in: *Proceedings of the findings of the Association for Computational Linguistics, ACL/IJCNLP*, 2021, pp. 4534–4545.
- [99] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, R. Hadsell, Overcoming catastrophic forgetting in neural networks, *CoRR* abs/1612.00796.
- [100] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, in: *Proceedings of the 36th International Conference on Machine Learning, ICML*, Vol. 97 of *Proceedings of Machine Learning Research*, PMLR, 2019, pp. 2790–2799.  
URL <http://proceedings.mlr.press/v97/houlsby19a.html>
- [101] 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, ACL*, 2023, pp. 1249–1262.
- [102] K. Yang, C. Yu, Y. R. Fung, M. Li, H. Ji, ADEPT: A debiasing prompt framework, in: *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023*, Washington, DC, USA, February 7–14, 2023, AAAI Press, 2023, pp. 10780–10788. doi:10.1609/AAAI.V37I9.26279.  
URL <https://doi.org/10.1609/aaai.v37i9.26279>
- [103] L. Hauenberger, S. Masoudian, D. Kumar, M. Schedl, N. Rekabsaz, Modular and on-demand bias mitigation with attribute-removal sub-networks, in: *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada, July 9–14, 2023, Association for Computational Linguistics, 2023, pp. 6192–6214. doi:10.18653/v1/2023.FINDINGS-AACL.386.  
URL <https://doi.org/10.18653/v1/2023.findings-acl.386>
- [104] D. Kumar, O. Lesota, G. Zerveas, D. Cohen, C. Eickhoff, M. Schedl, N. Rekabsaz, Parameter-efficient modularised bias mitigation via adapterfusion, *arXiv preprint arXiv:2302.06321*.
- [105] 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.
- [106] C. Oh, H. Won, J. So, T. Kim, Y. Kim, H. Choi, K. Song, Learning fair representation via distributional contrastive disentanglement, in: *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Washington, DC, USA, August 14 – 18, 2022, ACM, 2022, pp. 1295–1305. doi:10.1145/3534678.3539232.  
URL <https://doi.org/10.1145/3534678.3539232>
- [107] L. Dixon, J. Li, J. Sorensen, N. Thain, L. Vasserman, Measuring and mitigating unintended bias in text classification, in: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES*, 2018, pp. 67–73.
- [108] Y. Pruksachatkun, S. Krishna, J. Dhamala, R. Gupta, K. Chang, Does robustness improve fairness? approaching fairness with word substitution robustness methods for text classification, in: *Proceedings of the Findings of the Association for Computational Linguistics: ACL/IJCNLP*, 2021, pp. 3320–3331.
- [109] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, A. Beutel, Counterfactual fairness in text classification through robustness, in: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES*, 2019, pp. 219–226.
- [110] L. Yu, Y. Mao, J. Wu, F. Zhou, Mixup-based unified framework to overcome gender bias resurgence, in: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023*, ACM, 2023, pp. 1755–1759. doi:10.1145/3539618.3591938.  
URL <https://doi.org/10.1145/3539618.3591938>
- [111] A. Zayed, P. Parthasarathi, G. Mordido, H. Palangi, S. Shabaniyan, S. Chandar, Deep learning on a healthy data diet: Finding important examples for fairness, in: *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023*, Washington, DC, USA, February 7–14, 2023, AAAI Press, 2023, pp. 14593–14601. doi:10.1609/AAAI.V37I12.26706.  
URL <https://doi.org/10.1609/aaai.v37i12.26706>
- [112] E. Vanmassenhove, C. Hardmeier, A. Way, Getting gender right in neural machine translation, *CoRR* abs/1909.05088.
- [113] D. Saunders, B. Byrne, Reducing gender bias in neural machine translation as a domain adaptation problem, in: *Proceedings of the 58th*

- Annual Meeting of the Association for Computational Linguistics, ACL, 2020, pp. 7724–7736.
- [114] J. H. Park, J. Shin, P. Fung, Reducing gender bias in abusive language detection, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2018, pp. 2799–2804.
  - [115] M. Sap, D. Card, S. Gabriel, Y. Choi, N. A. Smith, The risk of racial bias in hate speech detection, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL, 2019, pp. 1668–1678.
  - [116] X. Zhou, M. Sap, S. Swayamdipta, Y. Choi, N. Smith, Challenges in automated debiasing for toxic language detection, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL, 2021, pp. 3143–3155.
  - [117] S. Panda, A. Kobren, M. Wick, Q. Shen, Don't just clean it, proxy clean it: Mitigating bias by proxy in pre-trained models, in: Proceedings of the Findings of the Association for Computational Linguistics, EMNLP, 2022, pp. 5073–5085.
  - [118] P. Sattigeri, S. Ghosh, I. Padhi, P. L. Dognin, K. R. Varshney, Fair infinitesimal jackknife: Mitigating the influence of biased training data points without refitting, in: NeurIPS, 2022.  
URL [http://papers.nips.cc/paper\\_files/paper/2022/hash/e94481b99473c83b2e79d91c64eb37d1-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/e94481b99473c83b2e79d91c64eb37d1-Abstract-Conference.html)
  - [119] R. G. Miller, The jackknife-a review, *Biometrika* 61 (1) (1974) 1–15.
  - [120] A. Garimella, R. Mihalcea, A. Amarnath, Demographic-aware language model fine-tuning as a bias mitigation technique, in: Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2022 - Volume 2: Short Papers, Online only, November 20-23, 2022, Association for Computational Linguistics, 2022, pp. 311–319.  
URL <https://aclanthology.org/2022.aacl-short.38>
  - [121] X. Han, T. Baldwin, T. Cohn, Balancing out bias: Achieving fairness through balanced training, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2022, pp. 11335–11350.
  - [122] G. Zhang, B. Bai, J. Zhang, K. Bai, C. Zhu, T. Zhao, Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL, 2020, pp. 4134–4145.
  - [123] H. Orgad, Y. Belinkov, BLIND: Bias removal with no demographics, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL, 2023, pp. 8801–8821.
  - [124] P. A. Utama, N. S. Moosavi, I. Gurevych, Towards debiasing NLU models from unknown biases, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020, Association for Computational Linguistics, 2020, pp. 7597–7610. doi:10.18653/V1/2020.EMNLP-MAIN.613.  
URL <https://doi.org/10.18653/v1/2020.emnlp-main.613>
  - [125] S. Park, K. Choi, H. Yu, Y. Ko, Never too late to learn: Regularizing gender bias in coreference resolution, in: Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, WSDM, ACM, 2023, pp. 15–23. doi:10.1145/3539597.3570473.  
URL <https://doi.org/10.1145/3539597.3570473>
  - [126] Z. Wang, K. Shu, A. Culotta, Enhancing model robustness and fairness with causality: A regularization approach, in: Proceedings of the 1st Workshop on Causal Inference and NLP, 2021, pp. 33–43.
  - [127] 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.
  - [128] R. Wang, P. Cheng, R. Henao, Toward fairness in text generation via mutual information minimization based on importance sampling, in: International Conference on Artificial Intelligence and Statistics, AISTATS, Vol. 206 of Proceedings of Machine Learning Research, PMLR, 2023, pp. 4473–4485.  
URL <https://proceedings.mlr.press/v206/wang23c.html>
  - [129] 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.
  - [130] P. Lahoti, A. Beutel, J. Chen, K. Lee, F. Prost, N. Thain, X. Wang, E. H. Chi, Fairness without demographics through adversarially reweighted learning, in: Proceedings of the 33rd Annual Conference on Neural Information Processing Systems, NeurIPS, 2020.
  - [131] X. Han, T. Baldwin, T. Cohn, Decoupling adversarial training for fair NLP, in: Proceedings of the findings of the Association for Computational Linguistics, ACL-IJCNLP, 2021, pp. 471–477.
  - [132] Y. Elazar, Y. Goldberg, Adversarial removal of demographic attributes from text data, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP, 2018, pp. 11–21.
  - [133] X. Han, T. Baldwin, T. Cohn, Diverse adversaries for mitigating bias in training, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL, 2021, pp. 2760–2765.
  - [134] 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.
  - [135] 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.
  - [136] A. Shen, X. Han, T. Cohn, T. Baldwin, L. Frermann, Contrastive learning for fair representations, arXiv:2109.10645.
  - [137] 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.
  - [138] M. Cheng, E. Durmus, D. Jurafsky, Marked personas: Using natural language prompts to measure stereotypes in language models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL, 2023, pp. 1504–1532.
  - [139] A. Ramezani, Y. Xu, Knowledge of cultural moral norms in large language models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL, 2023, pp. 428–446.
  - [140] J. Mattern, Z. Jin, M. Sachan, R. Mihalcea, B. Schölkopf, Understanding stereotypes in language models: Towards robust measurement and zero-shot debiasing, CoRR abs/2212.10678. arXiv:2212.10678, doi:10.48550/ARXIV.2212.10678.  
URL <https://doi.org/10.48550/arXiv.2212.10678>

- [141] P. Liang, R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. J. Orr, L. Zheng, M. Yüsekçönlü, M. Suzgun, N. Kim, N. Guha, N. S. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, Y. Koreeda, Holistic evaluation of language models, CoRR abs/2211.09110.
- [142] A. Abid, M. Farooqi, J. Zou, Persistent anti-muslim bias in large language models, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES, 2021, pp. 298–306.
- [143] T. Y. Zhuo, Y. Huang, C. Chen, Z. Xing, Red teaming chatgpt via jailbreaking: Bias, robustness, reliability and toxicity, 2023.
- [144] Y. Huang, Q. Zhang, L. Sun, et al., Trustgpt: A benchmark for trustworthy and responsible large language models, arXiv preprint arXiv:2306.11507.
- [145] M. Forbes, J. D. Hwang, V. Shwartz, M. Sap, Y. Choi, Social chemistry 101: Learning to reason about social and moral norms, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16–20, 2020, Association for Computational Linguistics, 2020, pp. 653–670. doi:10.18653/V1/2020.EMNLP-MAIN.48. URL <https://doi.org/10.18653/v1/2020.emnlp-main.48>
- [146] Y. Li, Y. Zhang, Fairness of chatgpt, CoRR abs/2305.18569.
- [147] E. Fleisig, A. Amstutz, C. Atalla, S. L. Blodgett, H. Daumé III, A. Olteanu, E. Sheng, D. Vann, H. Wallach, Fair-prism: Evaluating fairness-related harms in text generation, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2023.
- [148] Y. Wan, W. Wang, P. He, J. Gu, H. Bai, M. R. Lyu, Biasasker: Measuring the bias in conversational AI system, CoRR abs/2305.12434.
- [149] A. Tamkin, A. Askeel, L. Lovitt, E. Durmus, N. Joseph, S. Kravec, K. Nguyen, J. Kaplan, D. Ganguli, Evaluating and mitigating discrimination in language model decisions, arXiv preprint arXiv:2312.03689.
- [150] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, S. T. Truong, S. Arora, M. Mazeika, D. Hendrycks, Z. Lin, Y. Cheng, S. Koyejo, D. Song, B. Li, Decodingtrust: A comprehensive assessment of trustworthiness in GPT models, CoRR abs/2306.11698.
- [151] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, E. P. Xing, Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality (March 2023). URL <https://lmsys.org/blog/2023-03-30-vicuna/>
- [152] D. Khashabi, S. Min, T. Khot, A. Sabharwal, O. Tafjord, P. Clark, H. Hajishirzi, UNIFIEDQA: Crossing format boundaries with a single QA system, in: Proceedings of the Findings of the 2022 Association for Computational Linguistics, EMNLP, 2020, pp. 1896–1907.
- [153] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, in: NeurIPS, 2022.
- [154] V. Sanh, A. Webson, C. Raffel, S. H. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, A. Raja, M. Dey, M. S. Bari, C. Xu, U. Thakker, S. S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. V. Nayak, D. Datta, J. Chang, M. T. Jiang, H. Wang, M. Manica, S. Shen, Z. X. Yong, H. Pandey, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Févry, J. A. Fries, R. Teehan, T. L. Scao, S. Biderman, L. Gao, T. Wolf, A. M. Rush, Multitask prompted training enables zero-shot task generalization, in: Proceedings of the 10th International Conference on Learning Representations, ICLR, 2022.
- [155] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, E. Zheng, R. Child, R. Y. Aminabadi, J. Bernauer, X. Song, M. Shoeybi, Y. He, M. Houston, S. Tiwary, B. Catanzaro, Using deepspeed and megatron to train megatron-turing NLG 530b, A large-scale generative language model, CoRR abs/2201.11990.
- [156] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, in: Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, AAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2–9, 2021, AAAI Press, 2021, pp. 14867–14875. doi:10.1609/AAAI.V35I17.17745. URL <https://doi.org/10.1609/aaai.v35i17.17745>
- [157] Anthropic, Model card and evaluations for claude models (July 2023). URL <https://www-files.anthropic.com/production/images/Model-Card-Claude-2.pdf?dm=1689034733>
- [158] S. Santy, J. T. Liang, R. L. Bras, K. Reinecke, M. Sap, Nlpositionality: Characterizing design biases of datasets and models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL, 2023, pp. 9080–9102.
- [159] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, et al., Sparks of artificial general intelligence: Early experiments with gpt-4, arXiv preprint arXiv:2303.12712.
- [160] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, P. Schuh, K. Shi, S. Tsvyashchenko, J. Maynez, A. Rao, P. Barnes, Y. Tay, N. Shazeer, V. Prabhakaran, E. Reif, N. Du, B. Hutchinson, R. Pope, J. Bradbury, J. Austin, M. Isard, G. Gur-Ari, P. Yin, T. Duke, A. Levskaya, S. Ghemawat, S. Dev, H. Michalewski, X. Garcia, V. Misra, K. Robinson, L. Fedus, D. Zhou, D. Ippolito, D. Luan, H. Lim, B. Zoph, A. Spiridonov, R. Sepassi, D. Dohan, S. Agrawal, M. Omernick, A. M. Dai, T. S. Pillai, M. Pellat, A. Lewkowycz, E. Moreira, R. Child, O. Polozov, K. Lee, Z. Zhou, X. Wang, B. Saeta, M. Diaz, O. Firat, M. Catasta, J. Wei, K. Meier-Hellstern, D. Eck, J. Dean, S. Petrov, N. Fiedel, Palm: Scaling language modeling with pathways, CoRR abs/2204.02311.
- [161] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, in: Proceedings of the 10th International Conference on Learning Representations, ICLR, 2022.
- [162] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Y. Zhao, Y. Huang, A. M. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, J. Wei, Scaling instruction-finetuned language models, CoRR abs/2210.11416.
- [163] P. F. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, in: Proceedings of the 30th Annual Conference on Neural Information Processing Systems, NeurIPS, 2017, pp. 4299–4307.
- [164] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, O. Klimov, Proximal policy optimization algorithms, CoRR abs/1707.06347.

- [165] C. Borchers, D. S. Gala, B. Gilbert, E. Oravkin, W. Bounsi, Y. M. Asano, H. R. Kirk, Looking for a handsome carpenter! debiasing GPT-3 job advertisements, CoRR abs/2205.11374. arXiv:2205.11374, doi:10.48550/ARXIV.2205.11374.  
URL <https://doi.org/10.48550/arXiv.2205.11374>
- [166] X. Jin, F. Barbieri, B. Kennedy, A. M. Davani, L. Neves, X. Ren, On transferability of bias mitigation effects in language model fine-tuning, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT, 2021, pp. 3770–3783.
- [167] P. Delobelle, E. K. Tokpo, T. Calders, B. Berendt, Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL, 2022, pp. 1693–1706.
- [168] Y. T. Cao, Y. Pruksachatkun, K. Chang, R. Gupta, V. Kumar, J. Dhamala, A. Galstyan, On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL, 2022, pp. 561–570.
- [169] E. Ferrara, Should chatgpt be biased? challenges and risks of bias in large language models, arXiv preprint arXiv:2304.03738.
- [170] J. Watson, B. Beekhuizen, S. Stevenson, What social attitudes about gender does BERT encode? leveraging insights from psycholinguistics, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL, 2023, pp. 6790–6809.
- [171] T. Walter, C. Kirschner, S. Eger, G. Glavas, A. Lauscher, S. P. Ponzetto, Diachronic analysis of german parliamentary proceedings: Ideological shifts through the lens of political biases, in: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, JCDL, 2021, pp. 51–60.
- [172] I. Baldini, D. Wei, K. N. Ramamurthy, M. Singh, M. Yurochkin, Your fairness may vary: Pretrained language model fairness in toxic text classification, in: Proceedings of the Findings of the Association for Computational Linguistics, ACL, 2022, pp. 2245–2262.
- [173] R. Liu, J. Regier, N. Tripuraneni, M. I. Jordan, J. D. McAuliffe, Rao-blackwellized stochastic gradients for discrete distributions, in: Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, Vol. 97 of Proceedings of Machine Learning Research, PMLR, 2019, pp. 4023–4031.  
URL <http://proceedings.mlr.press/v97/liu19c.html>
- [174] R. Agarwal, M. Schwarzer, P. S. Castro, A. C. Courville, M. G. Bellemare, Deep reinforcement learning at the edge of the statistical precipice, in: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual, 2021, pp. 29304–29320.  
URL <https://proceedings.neurips.cc/paper/2021/hash/f514cec81cb148559cf475e7426eed5e-Abstract.html>
- [175] X. Lu, S. Welleck, J. Hessel, L. Jiang, L. Qin, P. West, P. Ammanabrolu, Y. Choi, QUARK: controllable text generation with reinforced unlearning, in: NeurIPS, 2022.  
URL [http://papers.nips.cc/paper\\_files/paper/2022/hash/b125999bde7e80910cbdbd323087df8f-Abstract-Conference.html](http://papers.nips.cc/paper_files/paper/2022/hash/b125999bde7e80910cbdbd323087df8f-Abstract-Conference.html)