A Survey on Fairness in Large Language Models

Yingji Li[1], **Mengnan Du**[2], **Rui Song**[3], **Xin Wang**[3], **Ying Wang**[1,4]
[1]College of Computer Science and Technology, Jilin University, Changchun, China
[2]Department of Data Science, New Jersey Institute of Technology, Newark, USA
[3]School of Artificial Intelligence, Jilin University, Changchun, China
[4]Key Laboratory of Symbolic Computation and Knowledge Engineering
of Ministry of Education, Jilin University, Changchun, China
yingji21@mails.jlu.edu.cn, mengnan.du@njit.edu,
songrui20@mails.jlu.edu.cn, xinwang@jlu.edu.cn, wangying2010@jlu.edu.cn

Abstract

Large language models (LLMs) have shown powerful performance and development prospect 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. First, for medium-scale LLMs, we introduce evaluation metrics and debiasing methods from the perspectives of intrinsic bias and extrinsic bias, respectively. Then, for large-scale 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.

1 Introduction

Large language models (LLMs), such as BERT (Devlin et al. 2019), GPT-3 (Brown et al. 2020), and LLaMA (Touvron et al. 2023a), have shown powerful performance and development prospect 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. Numerous studies have verified that LLMs capture human social biases in unprocessed training data, and biases emerge in encoded embeddings that carry over into downstream tasks (Garg et al. 2018; Sun et al. 2019). Unfair LLM systems make discriminatory, stereotypic, and biased decisions against vulnerable or marginalized demographics, causing undesirable social impacts and potential harms (Blodgett et al. 2020; Kumar et al. 2023).

Social biases in language models are derived primarily from training data collected from human societies. On the one hand, these uncensored corpora contain a lot of harmful information reflecting bias, leading language models to learn stereotyped behaviors (Mehrabi et al. 2022). On the other hand, the labels of different demographic groups in the training data are imbalanced, and the distributional difference can lead to unfair predictions when the model trained under the homogeneity assumption is applied to the heterogeneous real data (Shah, Schwartz, and Hovy 2020). In addition, human factors during language model learning or unanticipated biases in embeddings can cause or even amplify downstream biases (Bansal 2022).

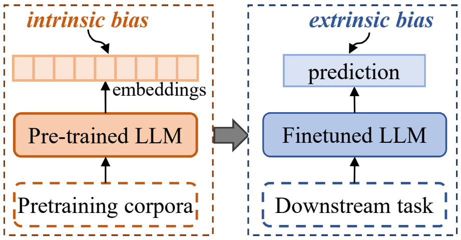

Figure 1: Illustrati of intrinsic and extrinsic bias in the pre-training and fine-tuning training paradigm.

According to the training paradigm, LLMs can be divided into *pre-training and fine-tuning paradigm* as well as *prompting paradigm*. In the pre-training and fine-tuning paradigm, LLMs have less than a billion parameters and are easy to tune, such as BERT and RoBERTA (Liu et al. 2019), which we call *medium-scale LLMs*. Biases in medium-scale LLMs can be roughly understood as two types: intrinsic bias and extrinsic bias (Goldfarb-Tarrant et al. 2021), as shown in Figure 1. Intrinsic bias corresponds to the bias in the embeddings encoded by the LLM and reflects the fairness of model's output representation. Extrinsic bias corresponds to the decision bias of the downstream task and reflects the fairness of model's prediction. In the prompting paradigm, LLMs have more than a billion parameters and are tuned or not tuned based on the prompts, such as GPT-4 (OpenAI 2023) and LLaMA-2 (Touvron et al. 2023b), which we call *large-scale LLMs*. Biases in large-scale LLMs are generally reflected in the model output when given specific prompts.

In this paper, we provide a comprehensive review of related research on fairness in LLMs, where the overall archictecutre is shown in Figure 2. Focusing on medium-scale LLMs under the pre-training and fine-tuning paradigm, we introduce the evaluation metrics in Section 2, and the intrinsic debiasing methods and extrinsic debiasing methods in Section 3 and Section 4, respectively. In Section 5, the fairness of large-scale 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 6.

2 Evaluation Metrics

In this section, we summarize the fairness evaluation metrics for medium-scale LLMs, which are divided into intrinsic metrics and extrinsic metrics. Intrinsic metrics are applied to embeddings, formalizing intrinsic bias by statistically quantifying the associations between targets and concepts. Extrinsic metrics are applied to the output of downstream tasks to characterize extrinsic bias by the performance gap.

2.1 Intrinsic Metrics

Distance-based. Sentence embedding association test (SEAT) (May et al. 2019) adapts the word embeddings association test (WEAT) (Caliskan, Bryson, and Narayanan 2017) to contextual embeddings, which measures the association between two sets of targets (e.g., *male*/*female*) and two sets of attributes (e.g., *family*/*career*) via semantically bleared templates such as "*He/She is a [MASK]*". The cosine distance between the two sets of embeddings is then calculated as the effect size score. The contextualized embedding association test (CEAT) extends WEAT to a dynamic setting by quantifying the distribution of effect sizes for social and cross-bias in contextualized word embeddings (Guo and Caliskan 2021). Given a set of target groups 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.

Probability-based. Discovery of correlations (DisCo) takes the average score of a model's predictions as the measurement (Webster et al. 2020). 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) takes a similar template and measurement (Kurita et al. 2019). 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. StereoSet is a crowd-sourced dataset that measures four stereotype biases, where each sample consists of a context sentence and a set of candidate associations (Nadeem, Bethke, and Reddy 2021). The model chooses among three candidate associations: stereotyped, anti-stereotyped, and irrelevant, and obtains a bias score for each protected group. Similarly, CrowS-Pairs is a dataset containing pairs of stereotyped and anti-stereotyped sentences, which utilizes the pseudo-log-likelihood to compute the perplexity of all tokens conditioned on typical tokens (Nangia et al. 2020). AUL modifies CrowS-Pairs by combining multiple correct predictions instead of testing whether the target token is predicted (Kaneko and Bollegala 2022).

2.2 Extrinsic Metrics

Coreference Resolution. One of the most classical tasks for measuring gender bias is coreference resolution on datasets developed based on the Winograd (Levesque, Davis, and Morgenstern 2012) format. WinoBias is a dataset for the intra-clause coreference resolution task (Zhao et al. 2018), which evaluates the model's ability to associate gender pronouns and occupations in contexts of stereotype and anti-stereotype. The bias score is defined as the difference between the model's assessment of "stereotype" and "anti-stereotype". Similarly, Winogender is also an English coreference resolution dataset based on the Winograd format (Rudinger et al. 2018). 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. In addition, GAP proposes a gender-balanced tagged corpus of 8,908 ambiguous pronname pairs, which can cover more diverse discriminatory pronouns and a more balanced dataset to measure the actual bias of the model more accurately (Webster et al. 2018).

Semantic Similarity. Considering the semantic similarity between sentence pairs allows assessing the associations between gender and occupation, such as STS-B (Cer et al. 2017) and Bias-NLI (Dev et al. 2020). They 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.

Group Fairness. Some representative metrics measure the performance gap of the model for different groups. BOLD is a large-scale fairness benchmark dataset containing natural prompts to evaluate bias across five domains in open-ended English language generation (Dhamala et al. 2021). Given prompts that describe the target population, BOLD measures bias by evaluating the quality of language model generation. Bias-in-Bios 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 (De-Arteaga et al. 2019). 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.

3 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: pre-processing, in-processing, and post-processing (Du et al. 2020).

3.1 Pre-processing

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

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) (Lu et al. 2020; Zmigrod et al. 2019). CDA augments the original corpus with causal intervention, which replaces the

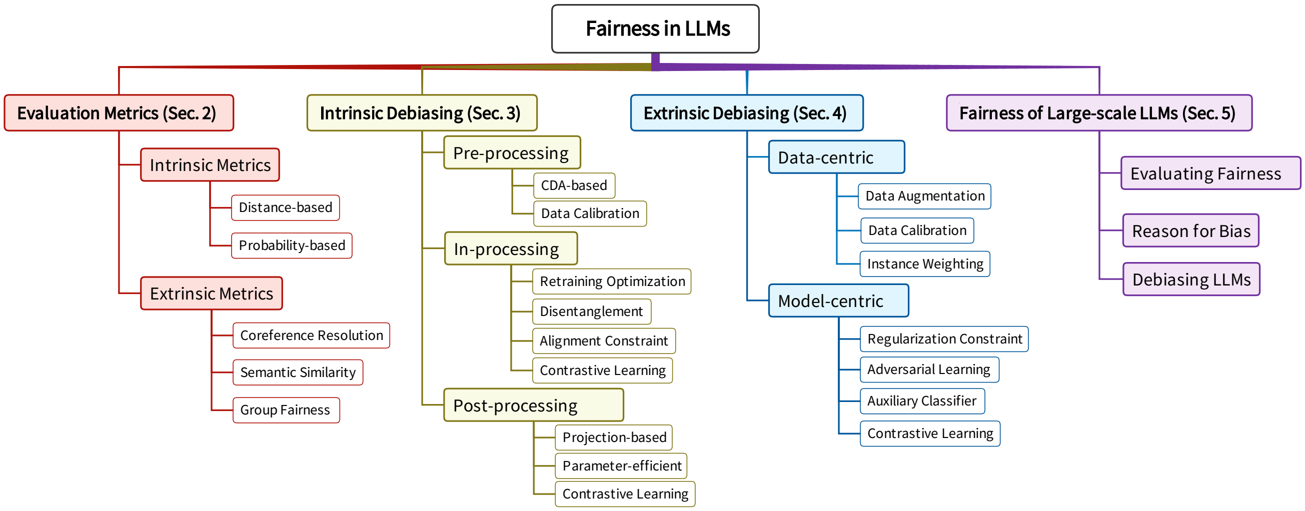

Figure 2: The overall architecture of our survey.

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*). However, it is difficult to completely eliminate bias by simply augmenting the dataset. Therefore, most debiasing methods combine CDA and other debiasing strategies (Stahl, Spliethöver, and Wachsmuth 2022; Xie and Lukasiewicz 2023). They make various improvements based on CDA, but the fundamental idea is to balance the training samples.

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 (Brunet et al. 2019) or programmatically (Ngo et al. 2021), and then delete these subsets to retrain unbiased models. 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 (Amrhein et al. 2023).

3.2 In-processing

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

Retraining Optimization. Retraining models is a direct way to reduce bias, although it can be resource-intensive and difficult to scale. Dropout regularization interrupts the attention mechanism association between words, and can be used to retrain LLMs to reduce gendered correlations (Webster et al. 2020). Bias in the distilled language model can be mitigated using a fair knowledge distillation approach based on counterfactual role reversal, which improves the fairness of the output probabilities of the teacher model to guide a fair student model (Gupta et al. 2022).

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 (Dev et al. 2021; Kaneko and Bollegala 2021). Group-specific subspace projection requires prior group knowledge, some work (Ungless et al. 2022; Omrani et al. 2023) projects representations to stereotype content models (SCM) (Fiske et al. 2002) that rely on theoretical understanding of social stereotypes to define bias subspaces, thus breaking the limitations of prior knowledge.

Alignment Constraint. This mitigation strategy is to constrain models to learn more similar representations by aligning the distributions between different sensitive attributes. Auto-Debias 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 (Guo, Yang, and Abbasi 2022). To mitigate bias in low-resource multilingual models, (Ahn and Oh 2021) proposes to leverage the contextual embeddings of two monolingual BERT and align the less biased one.

Contrastive Learning. 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. MABEL counterfactual augments premises and hypotheses from the natural language inference (NLI) dataset, and then uses a contrastive learning objective on gender-balanced entailment pairs (He et al. 2022). CCPA 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 (Li et al. 2023).

3.3 Post-processing

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

Projection-based. One traditional approach is to remove bias information from representations by linearly separating sensitive and neutral attributes (Dev and Phillips 2019). 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 (Dev et al. 2020; Liang et al. 2020). However, removing only useless information is difficult, and it carries the risk of compromising the original semantics (Garimella et al. 2021).

Parameter-efficient. Parameter-efficient methods are used to address the potentially catastrophic forgetting (Kirkpatrick et al. 2016) 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 adds a popular adapter 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 (Lauscher, Lüken, and Glavas 2021). GEEP injects LLMs with gender equality prompts that are trainable embedding of occupation names (Fatemi et al. 2023). Similarly, LLMs' parameters are fixed while prompts are updated, thus preserving the original useful information.

Contrastive Learning. Contrastive learning can also be used to deal with intrinsic bias in the post-processing. For example, FairFil proposes a neural debiasing method based on the contrastive learning framework (Cheng et al. 2021), 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.

4 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 type: data-centric and model-centric.

4.1 Data-centric Debiasing

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

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 (Dixon et al. 2018). 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 (Pruksachatkun et al. 2021) and counterfactual logit pairing (Garg et al. 2019).

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 (Vanmassenhove, Hardmeier, and Way 2019) and creating a credible gender-balanced adaptation dataset (Saunders and Byrne 2020). In toxic language detection, methods include using transfer learning to reduce bias from a less biased corpus (Park, Shin, and Fung 2018), relabeling samples by dialect and race priming (Sap et al. 2019) or automatically sensing dialects (Zhou et al. 2021), and identifying and removing proxy words associated with identity terms (Panda et al. 2022). 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.

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 (Han, Baldwin, and Cohn 2022). Social bias in text classification is formalized as a selection bias from a non-discriminatory distribution to a discriminatory distribution (Zhang et al. 2020). 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 treats social bias as a special case of the robustness problem caused by shortcut learning (Orgad and Belinkov 2023). 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.

4.2 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.

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. 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 (Wang, Shu, and Culotta 2021). 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. Another plug-and-play debiasing method integrates the training objective of masked language models into downstream classification tasks (Ghanbarzadeh et al. 2023). 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.

Adversarial Learning. The main idea of adversarial learning is to hide sensitive information from the decision function (Ravfogel et al. 2022). 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 (Lahoti et al. 2020). 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 (Han, Baldwin, and Cohn 2021a). Although adversarial debiasing alleviates the bias to a large extent, it still retains important sensitive information in the model encoding and prediction output (Elazar and Goldberg 2018). 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 (Han, Baldwin, and Cohn 2021b).

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 (Ravfogel et al. 2020). 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 (Liu et al. 2021). 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 reselected representations are fed into the original classifier so that the predictions are less disturbed by irrelevant features.

Contrastive Learning. It is cheaper and easier to optimize by combining contrastive learning to mitigate the bias in classifier training (Shen et al. 2021). 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.

5 Fairness of Large-scale LLMs

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

5.1 Evaluating Fairness of Large-scale LLMs

For assessing social bias in large-scale LLMs, the basic strategy is to analyze bias associations in the content generated by the model in response to the input prompts (Cheng, Durmus, and Jurafsky 2023; Ramezani and Xu 2023). This can be performed from different perspectives using a variety of tasks, including prompts completion, dialogue generation, and analogical reasoning, while some work has also developed benchmark datasets to test for social bias.

GPT-3 is declared socially biased and it is validated by prompt completion and co-occurrence tests (Brown et al. 2020). The authors test the association between gender and occupation, and in 83% of 388 occupations prompts are generated with text related to male identifiers. They feed in 800 prompts about gender, race, and religion in a co-occurrence test, and GPT-3's output reflects the presence of social bias in the training data. Other work has shown that GPT-3 has a higher violent bias against Muslims than other religious groups, by leveraging tasks such as prompt completion, story generation, and analogical reasoning to quantify the probability of GPT-3 outputting violent content against Muslim groups (Abid, Farooqi, and Zou 2021).

BBQ is a question answering bias benchmark with nine social bias categories, consisting of 58,492 hand-constructed context examples of ambiguity and disambiguation (Parrish et al. 2022). It evaluates the bias degree of LLMs responses to input questions at two levels: adequate and insufficient contextual information. The test results on UnifiedQA (11B) (Khashabi et al. 2020) 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. Then, BBQ is used to evaluate biases and stereotypes contained in 30 well-known LLMs (Liang et al. 2022). It finds a strong correlation between bias and accuracy in ambiguous contexts for InstructGPT davinci v2 (175B) (Ouyang et al. 2022), T0++ (11B) (Sanh et al. 2022), and TNLG v2 (530B) (Smith et al. 2022), 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.

Recent researches have focused on the fairness evaluation of ChatGPT/GPT-4 (OpenAI 2023). BiasAsker proposes an automated framework for identifying and measuring social biases in conversational AI systems, which identifies absolute and correlated biases in dialogue (Wan et al. 2023). 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. A literature evaluates ChatGPT's fairness performance in high-stakes domains such as education, criminology, finance, and healthcare (Li and Zhang 2023). Considering group fairness and individual fairness, the authors 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.

For a more adequate study, DecodingTrust provides a comprehensive fairness evaluation for ChatGPT and GPT-4, where stereotype bias and fairness are evaluated separately

(Wang et al. 2023). 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.

5.2 What are the Reasons for Model Bias?

Recent large-scale 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-scale LLMs (Santy et al. 2023; Bubeck et al. 2023).

LLaMA-2 (Touvron et al. 2023b) is verified that the bias in its generation is correlated with the frequency of gender pronouns and identity terms in the training data (Touvron et al. 2023b). 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 (Chowdhery et al. 2022). 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 (Bubeck et al. 2023). 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 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 (Santy et al. 2023). 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.

5.3 Debiasing Large-scale LLMs

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

Instruction Fine-tuning. Fine-tuning large-scale 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 (Wei et al. 2022; Chung et al. 2022). Using reinforcement learning from human feedback (RLHF) (Christiano et al. 2017) to instruct fine-tuning is a means of strengthening, the representative work include InstructGPT (Ouyang et al. 2022) and LLaMA-2-chat (Touvron et al. 2023b). InstructGPT 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 (Schulman et al. 2017). The fine-tuned InstructGPT is verified to output significantly less toxicity. However, the results of evaluating bias on modified versions of Winogender (Rudinger et al. 2018) and CrowS-Pairs (Nangia et al. 2020) 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 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.

Prompt Engineering. Prompt engineering has also been used to mitigate the bias of large-scale LLMs in language generation, by designing additional prompts to guide the model to a fairer output without fine-tuning. For example, in the occupation recommendation task, the authors 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 (Bubeck et al. 2023).

6 Discussions

Although the fairness of medium-scale 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-scale 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.

6.1 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 (Jin et al. 2021). But others point out that intrinsic bias and extrinsic bias are not necessarily correlated (Goldfarb-Tarrant et al. 2021; Delobelle et al. 2022), not only in the original setting but even when correcting for metric bias, noise in the dataset, and confounding factors (Cao et al. 2022). Moreover, different metrics are not compatible with each other, making it difficult to guarantee the reliability of the benchmark (Qian et al. 2022). 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.

6.2 Accurately Evaluating Fairness of Large-scale LLMs

Expand methods for quantifying bias. For evaluating the fairness of medium-scale 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-scale LLMs is relatively inadequate. In particular, for many large-scale 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-scale LLMs.

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-scale LLMs development, and they have not been proven to accurately reflect the performance of the model. Although large-scale 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-scale LLMs.

6.3 Further Explore the Reasons for Bias

As we conclude in Section 5.2, some literatures analyze the reasons for the bias in large-scale LLMS through experimental validation, which focus on comparing the associations of pre-training corpora and real-world stereotypes from a data statistical perspective. There are studies that explore the reasons for bias in medium-scale LLMs from other perspectives, such as (Watson, Beekhuizen, and Stevenson 2023) understands how BERT's predicted preferences reflect social attitudes toward gender from the psychological perspective, (Walter et al. 2021) analyzes bias in historical corpora from the political perspective, and (Baldini et al. 2022) explores 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-scale LLMs should also develop more inquiry work to deepen the investigation of reasons for bias from a broader perspective to develop more fair systems.

6.4 Efficiently Debiasing Large-scale LLMs

Improve current debiasing strategies. RLHF-based fine-tuning methods are difficult to generalize in implementation due to their high labor costs and resources. We expect to apply low-cost methods to debias large-scale 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 (Bubeck et al. 2023), creating new opportunities for guiding model's behaviors.

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.

7 Conclusions

We present a comprehensive survey of the fairness problem in LLMs. The social biases mainly come from training data containing harmful information and imbalanced data, and can be divided into intrinsic bias and extrinsic bias. We summarize the fairness researches of LLMs including intrinsic and extrinsic evaluation metrics and debiasing strategies for medium-scale LLMs, as well as fairness evaluation, reasons for bias, and debiasing methods for large-scale LLMs. 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-scale LLMs, which are still in the early stage, practitioners should combine more techniques and build comprehensive and safe language model systems.

References

Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES*, 298–306.

Ahn, J.; and Oh, A. 2021. Mitigating Language-Dependent Ethnic Bias in BERT. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 533–549.

Amrhein, C.; Schottmann, F.; Sennrich, R.; and Läubli, S. 2023. 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*, 4486–4506.

Baldini, I.; Wei, D.; Ramamurthy, K. N.; Singh, M.; and Yurochkin, M. 2022. Your Fairness may Vary: Pretrained Language Model Fairness in Toxic Text Classification. In *Proceedings of the Findings of the Association for Computational Linguistics, ACL*, 2245–2262.

Bansal, R. 2022. A Survey on Bias and Fairness in Natural Language Processing. *CoRR*, abs/2204.09591.

Blodgett, S. L.; Barocas, S.; III, H. D.; and Wallach, H. M. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, 5454–5476.

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems, NeurIPS*.

Brunet, M.; Alkalay-Houlihan, C.; Anderson, A.; and Zemel, R. S. 2019. Understanding the Origins of Bias in Word Embeddings. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, 803–811.

Bubeck, S.; Chandrasekaran, V.; Eldan, R.; Gehrke, J.; Horvitz, E.; Kamar, E.; Lee, P.; Lee, Y. T.; Li, Y.; Lundberg, S.; et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics Derived Automatically from Language Corpora Contain Human-like Biases. *Science*, 356(6334): 183–186.

Cao, Y. T.; Pruksachatkun, Y.; Chang, K.; Gupta, R.; Kumar, V.; Dhamala, J.; and Galstyan, A. 2022. On the Intrinsic and Extrinsic Fairness Evaluation Metrics for Contextualized Language Representations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL*, 561–570.

Cer, D. M.; Diab, M. T.; Agirre, E.; Lopez-Gazpio, I.; and Specia, L. 2017. SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL*, 1–14.

Cheng, M.; Durmus, E.; and Jurafsky, D. 2023. 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*, 1504–1532.

Cheng, P.; Hao, W.; Yuan, S.; Si, S.; and Carin, L. 2021. FairFil: Contrastive Neural Debiasing Method for Pretrained Text Encoders. In *Proceedings of the 9th International Conference on Learning Representations, ICLR*.

Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H. W.; Sutton, C.; Gehrmann, S.; Schuh, P.; Shi, K.; Tsvyashchenko, S.; Maynez, J.; Rao, A.; Barnes, P.; Tay, Y.; Shazeer, N.; Prabhakaran, V.; Reif, E.; Du, N.; Hutchinson, B.; Pope, R.; Bradbury, J.; Austin, J.; Isard, M.; Gur-Ari, G.; Yin, P.; Duke, T.; Levskaya, A.; Ghemawat, S.; Dev, S.; Michalewski, H.; Garcia, X.; Misra, V.; Robinson, K.; Fedus, L.; Zhou, D.; Ippolito, D.; Luan, D.; Lim, H.; Zoph, B.; Spiridonov, A.; Sepassi, R.; Dohan, D.; Agrawal, S.; Omernick, M.; Dai, A. M.; Pillai, T. S.; Pellat, M.; Lewkowycz, A.; Moreira, E.; Child, R.; Polozov, O.; Lee, K.; Zhou, Z.; Wang, X.; Saeta, B.; Diaz, M.; Firat, O.; Catasta, M.; Wei, J.; Meier-Hellstern, K.; Eck, D.; Dean, J.; Petrov, S.; and Fiedel, N. 2022. PaLM: Scaling Language Modeling with Pathways. *CoRR*, abs/2204.02311.

Christiano, P. F.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep Reinforcement Learning from Human Preferences. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems, NeurIPS*, 4299–4307.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V. Y.; Huang, Y.; Dai, A. M.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models. *CoRR*, abs/2210.11416.

De-Arteaga, M.; Romanov, A.; Wallach, H. M.; Chayes, J. T.; Borgs, C.; Chouldechova, A.; Geyik, S. C.; Kenthapadi, K.; and Kalai, A. T. 2019. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT*, 120–128.

Delobelle, P.; Tokpo, E. K.; Calders, T.; and Berendt, B. 2022. Measuring Fairness with Biased Rulers: A Comparative Study on Bias Metrics for Pre-trained Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*, 1693–1706.

Dev, S.; Li, T.; Phillips, J. M.; and Srikumar, V. 2020. On Measuring and Mitigating Biased Inferences of Word Embeddings. In *Proceedings of the 34th Association for the Advancement of Artificial Intelligence, AAAI*, 7659–7666.

Dev, S.; Li, T.; Phillips, J. M.; and Srikumar, V. 2021. 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*, 5034–5050.

Dev, S.; and Phillips, J. M. 2019. Attenuating Bias in Word vectors. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics, AISTATS*, volume 89, 879–887.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*, 4171–4186.

Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.; and Gupta, R. 2021. 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*, 862–872.

Dixon, L.; Li, J.; Sorensen, J.; Thain, N.; and Vasserman, L. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES*, 67–73.

Du, M.; Yang, F.; Zou, N.; and Hu, X. 2020. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 36(4): 25–34.

Elazar, Y.; and Goldberg, Y. 2018. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 11–21.

Fatemi, Z.; Xing, C.; Liu, W.; and Xiong, C. 2023. 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*, 1249–1262.

Fiske, S. T.; Cuddy, A. J.; Glick, P.; and Xu, J. 2002. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Journal of Personality and Social Psychology*, 878–902.

Garg, N.; Schiebinger, L.; Jurafsky, D.; and Zou, J. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proc. Natl. Acad. Sci. USA*, 115(16): E3635–E3644.

Garg, S.; Perot, V.; Limtiaco, N.; Taly, A.; Chi, E. H.; and Beutel, A. 2019. Counterfactual Fairness in Text Classification through Robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES*, 219–226.

Garimella, A.; Amarnath, A.; Kumar, K.; Yalla, A. P.; Natarajan, A.; Chhaya, N.; and Srinivasan, B. V. 2021. 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*, 4534–4545.

Ghanbarzadeh, S.; Huang, Y.; Palangi, H.; Moreno, R. C.; and Khanpour, H. 2023. Gender-tuning: Empowering Finetuning for Debiasing Pre-trained Language Models. In *Proceedings of the findings of the Association for Computational Linguistics: ACL*, 5448–5458.

Goldfarb-Tarrant, S.; Marchant, R.; Sánchez, R. M.; Pandya, M.; and Lopez, A. 2021. 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*, 1926–1940.

Guo, W.; and Caliskan, A. 2021. 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*, 122–133.

Guo, Y.; Yang, Y.; and Abbasi, A. 2022. Auto-Debias: Debiasing Masked Language Models with Automated Biased Prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, ACL*, 1012–1023.

Gupta, U.; Dhamala, J.; Kumar, V.; Verma, A.; Pruksachatkun, Y.; Krishna, S.; Gupta, R.; Chang, K.; Steeg, G. V.; and Galstyan, A. 2022. Mitigating Gender Bias in Distilled Language Models via Counterfactual Role Reversal. In *Proceedings of the Findings of the Association for Computational Linguistics, ACL*, 658–678.

Han, X.; Baldwin, T.; and Cohn, T. 2021a. Decoupling Adversarial Training for Fair NLP. In *Proceedings of the findings of the Association for Computational Linguistics, ACL-IJCNLP*, 471–477.

Han, X.; Baldwin, T.; and Cohn, T. 2021b. Diverse Adversaries for Mitigating Bias in Training. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, 2760–2765.

Han, X.; Baldwin, T.; and Cohn, T. 2022. Balancing out Bias: Achieving Fairness Through Balanced Training. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 11335–11350.

He, J.; Xia, M.; Fellbaum, C.; and Chen, D. 2022. MABEL: Attenuating Gender Bias using Textual Entailment Data. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 9681–9702.

Jin, X.; Barbieri, F.; Kennedy, B.; Davani, A. M.; Neves, L.; and Ren, X. 2021. 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*, 3770–3783.

Kaneko, M.; and Bollegala, D. 2021. Debiasing Pre-trained Contextualised Embeddings. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics, EACL*, 1256–1266.

Kaneko, M.; and Bollegala, D. 2022. Unmasking the Mask - Evaluating Social Biases in Masked Language Models. In *Proceedings of the 36th Association for the Advancement of Artificial Intelligence, AAAI*, 11954–11962.

Khashabi, D.; Min, S.; Khot, T.; Sabharwal, A.; Tafjord, O.; Clark, P.; and Hajishirzi, H. 2020. UNIFIEDQA: Crossing Format Boundaries with a Single QA System. In *Proceedings of the Findings of the 2022 Association for Computational Linguistics, EMNLP*, 1896–1907.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N. C.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; Hassabis, D.; Clopath, C.; Kumaran, D.; and Hadsell, R. 2016. Overcoming Catastrophic Forgetting in Neural Networks. *CoRR*, abs/1612.00796.

Kumar, S.; Balachandran, V.; Njoo, L.; Anastasopoulos, A.; and Tsvetkov, Y. 2023. 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*, 3291–3313.

Kurita, K.; Vyas, N.; Pareek, A.; Black, A. W.; and Tsvetkov, Y. 2019. Measuring Bias in Contextualized Word Representations. *CoRR*, abs/1906.07337.

Lahoti, P.; Beutel, A.; Chen, J.; Lee, K.; Prost, F.; Thain, N.; Wang, X.; and Chi, E. H. 2020. Fairness without Demographics through Adversarially Reweighted Learning. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems, NeurIPS*.

Lauscher, A.; Lüken, T.; and Glavas, G. 2021. Sustainable Modular Debiasing of Language Models. In *Proceedings of the findings of the 2021 Association for Computational Linguistics: EMNLP*, 4782–4797.

Levesque, H. J.; Davis, E.; and Morgenstern, L. 2012. The Winograd Schema Challenge. In *Proceedings of the 30th Principles of Knowledge Representation and Reasoning, KR*.

Li, Y.; Du, M.; Wang, X.; and Wang, Y. 2023. 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*, 14254–14267.

Li, Y.; and Zhang, Y. 2023. Fairness of ChatGPT. *CoRR*, abs/2305.18569.

Liang, P.; Bommasani, R.; Lee, T.; Tsipras, D.; Soylu, D.; Yasunaga, M.; Zhang, Y.; Narayanan, D.; Wu, Y.; Kumar, A.; Newman, B.; Yuan, B.; Yan, B.; Zhang, C.; Cosgrove, C.; Manning, C. D.; Ré, C.; Acosta-Navas, D.; Hudson, D. A.; Zelikman, E.; Durmus, E.; Ladhak, F.; Rong, F.; Ren, H.; Yao, H.; Wang, J.; Santhanam, K.; Orr, L. J.; Zheng, L.; Yüksekgönül, M.; Suzgun, M.; Kim, N.; Guha, N.; Chatterji, N. S.; Khattab, O.; Henderson, P.; Huang, Q.; Chi, R.; Xie, S. M.; Santurkar, S.; Ganguli, S.; Hashimoto, T.; Icard, T.; Zhang, T.; Chaudhary, V.; Wang, W.; Li, X.; Mai, Y.; Zhang, Y.; and Koreeda, Y. 2022. Holistic Evaluation of Language Models. *CoRR*, abs/2211.09110.

Liang, P. P.; Li, I. M.; Zheng, E.; Lim, Y. C.; Salakhutdinov, R.; and Morency, L. 2020. Towards Debiasing Sentence Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, 5502–5515.

Liu, H.; Jin, W.; Karimi, H.; Liu, Z.; and Tang, J. 2021. 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*, 74–85.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Lu, K.; Mardziel, P.; Wu, F.; Amancharla, P.; and Datta, A. 2020. 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*, volume 12300 of *Lecture Notes in Computer Science*, 189–202.

May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*, 622–628.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2022. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6): 115:1–115:35.

Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP*, 5356–5371.

Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 1953–1967.

Ngo, H.; Raterink, C.; Araújo, J. G. M.; Zhang, I.; Chen, C.; Morisot, A.; and Frosst, N. 2021. Mitigating Harm in Language Models with Conditional-likelihood Filtration. *CoRR*, abs/2108.07790.

Omrani, A.; Ziabari, A. S.; Yu, C.; Golazizian, P.; Kennedy, B.; Atari, M.; Ji, H.; and Dehghani, M. 2023. Social-Group-Agnostic Bias Mitigation via the Stereotype Content Model. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, 4123–4139.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Orgad, H.; and Belinkov, Y. 2023. BLIND: Bias Removal With No Demographics. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, 8801–8821.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*.

Panda, S.; Kobren, A.; Wick, M.; and Shen, Q. 2022. 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*, 5073–5085.

Park, J. H.; Shin, J.; and Fung, P. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 2799–2804.

Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. R. 2022. BBQ: A Hand-built Bias Benchmark for Question Answering. In *Proceedings of the findings of the Association for Computational Linguistics, ACL*, 2086–2105.

Pruksachatkun, Y.; Krishna, S.; Dhamala, J.; Gupta, R.; and Chang, K. 2021. 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*, 3320–3331.

Qian, R.; Ross, C.; Fernandes, J.; Smith, E. M.; Kiela, D.; and Williams, A. 2022. Perturbation Augmentation for Fairer NLP. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP*, 9496–9521.

Ramezani, A.; and Xu, Y. 2023. Knowledge of cultural moral norms in large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, 428–446.

Ravfogel, S.; Elazar, Y.; Gonen, H.; Twiton, M.; and Goldberg, Y. 2020. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL*, 7237–7256.

Ravfogel, S.; Twiton, M.; Goldberg, Y.; and Cotterell, R. D. 2022. Linear Adversarial Concept Erasure. In *Proceedings of the 39th International Conference on Machine Learning, ICML*, volume 162, 18400–18421.

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Durme, B. V. 2018. Gender Bias in Coreference Resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*, 8–14.

Sanh, V.; Webson, A.; Raffel, C.; Bach, S. H.; Sutawika, L.; Alyafeai, Z.; Chaffin, A.; Stiegler, A.; Raja, A.; Dey, M.; Bari, M. S.; Xu, C.; Thakker, U.; Sharma, S. S.; Szczechla, E.; Kim, T.; Chhablani, G.; Nayak, N. V.; Datta, D.; Chang, J.; Jiang, M. T.; Wang, H.; Manica, M.; Shen, S.; Yong, Z. X.; Pandey, H.; Bawden, R.; Wang, T.; Neeraj, T.; Rozen, J.; Sharma, A.; Santilli, A.; Févry, T.; Fries, J. A.; Teehan, R.; Scao, T. L.; Biderman, S.; Gao, L.; Wolf, T.; and Rush, A. M. 2022. Multitask Prompted Training Enables Zero-Shot Task Generalization. In *Proceedings of the 10th International Conference on Learning Representations, ICLR*.

Santy, S.; Liang, J. T.; Bras, R. L.; Reinecke, K.; and Sap, M. 2023. NLPositionality: Characterizing Design Biases of Datasets and Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, ACL*, 9080–9102.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, ACL*, 1668–1678.

Saunders, D.; and Byrne, B. 2020. 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*, 7724–7736.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347.

Shah, D.; Schwartz, H. A.; and Hovy, D. 2020. 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*, 5248–5264.

Shen, A.; Han, X.; Cohn, T.; Baldwin, T.; and Frermann, L. 2021. Contrastive learning for fair representations. *arXiv:2109.10645*.

Smith, S.; Patwary, M.; Norick, B.; LeGresley, P.; Rajbhandari, S.; Casper, J.; Liu, Z.; Prabhumoye, S.; Zerveas, G.; Korthikanti, V.; Zheng, E.; Child, R.; Aminabadi, R. Y.; Bernauer, J.; Song, X.; Shoeybi, M.; He, Y.; Houston, M.; Tiwary, S.; and Catanzaro, B. 2022. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. *CoRR*, abs/2201.11990.

Stahl, M.; Spliethöver, M.; and Wachsmuth, H. 2022. To Prefer or to Choose? Generating Agency and Power Counterfactuals Jointly for Gender Bias Mitigation. In *Proceedings of the 55th Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, 39–51.

Sun, T.; Gaut, A.; Tang, S.; Huang, Y.; ElSherief, M.; Zhao, J.; Mirza, D.; Belding, E. M.; Chang, K.; and Wang, W. Y. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL*, 1630–1640.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; Rodriguez, A.; Joulin, A.; Grave, E.; and Lample, G. 2023a. LLaMA: Open and Efficient Foundation Language Models. *CoRR*, abs/2302.13971.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; Bikel, D.; Blecher, L.; Canton-Ferrer, C.; Chen, M.; Cucurull, G.; Esiobu, D.; Fernandes, J.; Fu, J.; Fu, W.; Fuller, B.; Gao, C.; Goswami, V.; Goyal, N.; Hartshorn, A.; Hosseini, S.; Hou, R.; Inan, H.; Kardas, M.; Kerkez, V.; Khabsa, M.; Kloumann, I.; Korenev, A.; Koura, P. S.; Lachaux, M.; Lavril, T.; Lee, J.; Liskovich, D.; Lu, Y.; Mao, Y.; Martinet, X.; Mihaylov, T.; Mishra, P.; Molybog, I.; Nie, Y.; Poulton, A.; Reizenstein, J.; Rungta, R.; Saladi, K.; Schelten, A.; Silva, R.; Smith, E. M.; Subramanian, R.; Tan, X. E.; Tang, B.; Taylor, R.; Williams, A.; Kuan, J. X.; Xu, P.; Yan, Z.; Zarov, I.; Zhang, Y.; Fan, A.; Kambadur, M.; Narang, S.; Rodriguez, A.; Stojnic, R.; Edunov, S.; and Scialom, T. 2023b.

Llama 2: Open Foundation and Fine-Tuned Chat Models. *CoRR*, abs/2307.09288.

Ungless, E. L.; Rafferty, A.; Nag, H.; and Ross, B. 2022. A Robust Bias Mitigation Procedure Based on the Stereotype Content Model. *CoRR*, abs/2210.14552.

Vanmassenhove, E.; Hardmeier, C.; and Way, A. 2019. Getting Gender Right in Neural Machine Translation. *CoRR*, abs/1909.05088.

Walter, T.; Kirschner, C.; Eger, S.; Glavas, G.; Lauscher, A.; and Ponzetto, S. P. 2021. 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*, 51–60.

Wan, Y.; Wang, W.; He, P.; Gu, J.; Bai, H.; and Lyu, M. R. 2023. BiasAsker: Measuring the Bias in Conversational AI System. *CoRR*, abs/2305.12434.

Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; Truong, S. T.; Arora, S.; Mazeika, M.; Hendrycks, D.; Lin, Z.; Cheng, Y.; Koyejo, S.; Song, D.; and Li, B. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. *CoRR*, abs/2306.11698.

Wang, Z.; Shu, K.; and Culotta, A. 2021. Enhancing Model Robustness and Fairness with Causality: A Regularization Approach. In *Proceedings of the 1st Workshop on Causal Inference and NLP*, 33–43.

Watson, J.; Beekhuizen, B.; and Stevenson, S. 2023. 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*, 6790–6809.

Webster, K.; Recasens, M.; Axelrod, V.; and Baldridge, J. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Trans. Assoc. Comput. Linguistics*, 6: 605–617.

Webster, K.; Wang, X.; Tenney, I.; Beutel, A.; Pitler, E.; Pavlick, E.; Chen, J.; and Petrov, S. 2020. Measuring and Reducing Gendered Correlations in Pre-trained Models. *CoRR*, abs/2010.06032.

Wei, J.; Bosma, M.; Zhao, V. Y.; Guu, K.; Yu, A. W.; Lester, B.; Du, N.; Dai, A. M.; and Le, Q. V. 2022. Finetuned Language Models are Zero-Shot Learners. In *Proceedings of the 10th International Conference on Learning Representations, ICLR*.

Xie, Z.; and Lukasiewicz, T. 2023. 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*, 15730–15745.

Zhang, G.; Bai, B.; Zhang, J.; Bai, K.; Zhu, C.; and Zhao, T. 2020. 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*, 4134–4145.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT*, 15–20.

Zhou, X.; Sap, M.; Swayamdipta, S.; Choi, Y.; and Smith, N. 2021. 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*, 3143–3155.

Zmigrod, R.; Mielke, S. J.; Wallach, H. M.; and Cotterell, R. 2019. 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*, 1651–1661.