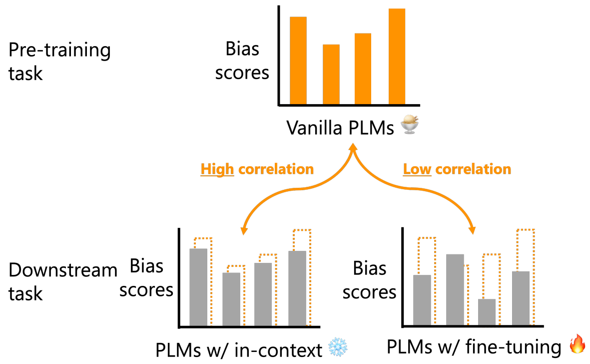
The Gaps between Pre-train and Downstream Settings in Bias Evaluation and Debiasing

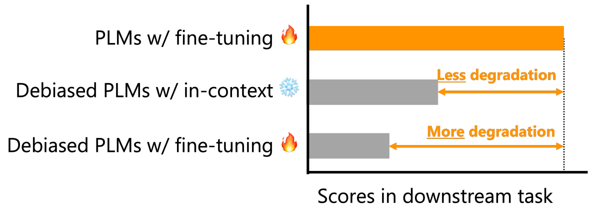
Anonymous ACL submission

Abstract

The output tendencies of Pre-trained Language Model (PLM)s vary markedly before and after fine tuning (FT) due to the updates to the model parameters. These divergences in output tendencies result in a gap in the social biases of PLMs. For example, there exits a low correlation between intrinsic bias scores of a PLM and its extrinsic bias scores under FT-based debiasing methods. Additionally, applying FT-based debiasing methods to a PLM leads to a decline in performance in downstream tasks. On the other hand, PLMs trained on large datasets can learn without parameter updates via incotext learning (ICL) using prompts. ICL induces smaller changes to PLMs compared to FT-based debiasing methods. Therefore, we hypothesize that the gap observed in pre-trained and FT models does not hold true for debiasing methods that use ICL. In this study, we demonstrate that ICL-based debiasing methods show a higher correlation between intrinsic and extrinsic bias scores compared to FT-based methods. Moreover, the performance degradation due to debiasing is also lower in the ICL case compared to that in the FT case.



(a) Bias evaluation.



(b) Debiasing.

Figure 1: The gap in bias scores when evaluating and debiasing PLMs using FT- and ICL-based methods. A lower correlation between intrinsic and extrinsic bias scores (a), while a larger drop in downstream task performance (b) is encountered with FT compared to ICL.

1 Introduction

PLMs learn not only beneficial information (Peters et al., 2018; Devlin et al., 2019; Brown et al., 2020; Touvron et al., 2023) but also undesirable social biases such as gender, race, and religous biases that exist in the training data (Sun et al., 2019; Liang et al., 2020; Schick et al., 2021; Zhou et al., 2022; Guo et al., 2022). Overall, two major approaches can be identified in the literature to elicit value from PLMs in downstream tasks: FT and ICL. FT adapts PLMs to specific tasks by updating parameters, while ICL uses prompts without modifying the model parameters.

FT models diverge considerably from the original PLMs in their output distributions (Chen et al., 2020). Similarly, the output distribution of a PLM is significantly affected by debiasing methods, because the parameters of the PLM are updated during the debiasing process. Debiasing accompanied by FT suffers substantial performance decline in downstream tasks compared to the original PLM (Meade et al., 2022; Kaneko et al., 2023b; Oba et al., 2023). This is because the beneficial information learnt during pre-training is lost during debiasing. Furthermore, bias evaluations exhibit a weak-level of correlation between pre-trained and FT PLMs (Goldfarb-Tarrant et al., 2021; Kaneko et al., 2022a; Cao et al., 2022).

On the other hand, it is not obvious whether the prevalent wisdom regarding bias in such FT regimes similarly pertains to ICL, devoid of concomitant model updates. The absence of parameter updates precludes the elimination of beneficial encodings, thereby minimizing adverse impacts on downstream task effectiveness. ICL strategies for mitigating biases may thus pose superior viability, while causing minimal representational damage. Moreover, we hypothesize that the bias evaluations that are based on pre-training and downstream tasks exhibit heightened correlations, because the ICL-based debiasing methods protect the model parameters.

In this paper, we investigate the performance gap of debiasing methods when applied to downstream tasks in an ICL setting. Additionally, we examine the correlation between bias evaluations for pre-training and downstream tasks enabled by the parameter sharing of ICL. Our experimental results show that ICL has a smaller gap than the FT setting with respect to (w.r.t.) performance degradation of debiasing and correlation between evaluations in pre-training and downstream tasks. Therefore, we expect this paper to contribute by cautioning the community against directly applying trends from pre-training and downstream tasks with FT to ICL without careful considerations.

2 Experiments

We first explain the details of bias evaluations, debiasing methods, and downstream tasks used in our experiments.

2.1 Bias Evaluations

Pre-training settings. We target the following three intrinsic bias evaluation datasets. Nangia et al. (2020) and Nadeem et al. (2021) proposed respectively, Crowds-Pairs (**CP**) and StereoSet (**SS**) bechmarks, which evaluate social biases of language models by comparing likelihoods of pro-stereotypical (e.g. *"She is a nurse"*) and anti-stereotypical (e.g. *"She is a doctor"*) examples. Kaneko et al. (2022b) introduced Multilingual Bias Evaluation (MBE) that evaluates gender bias in models in multiple languages by comparing likelihoods of feminine (e.g. *"She is a nurse"*) and masculine (e.g. *"He is a nurse"*) sentences. Our research compares the bias scores in pre-training and the downstream tasks, which requires us to target the same language and bias type in both settings as considered in those benchmarks. Therefore, we use gender bias in English on the above datasets to satisfy those requirements.

Downstream settings. We focus on three downstream tasks in our evaluations: question answering, natural language inference, and coreference resolution. Parrish et al. (2022) created the Bias Benchmark for Question answering (**BBQ**) to evaluate the social biases by determining whether a model predicts pro-stereotypical, anti-stereotypical, or unknown answers when given ambiguous and disambiguated contexts as the input. Anantaprayoon et al. (2023) proposed the Bias Natural Language Inference (**BNLI**) benchmark to evaluate gender bias using premise sentences and hypothesis sentences that share the same context but differ only in occupational and gender-related words. Zhao et al. (2018) proposed the WinoBias (**WB**) benchmark to evaluate gender bias by considering differences between model predictions on pro-stereotypical and anti-stereotypical sentences in a coreference resolution task.

2.2 Debiasing Methods

Fine-tuning. We use the following two FT-based debiasing methods that **modify the model parameters**. Webster et al. (2020) introduced the Counterfactual Data Augmentation debiasing (**CDA**), which swaps gender words in training data to reduce bias. For instance, *"She is a nurse"* becomes *"He is a nurse"* in the augmented dataset, resulting in a more balanced dataset. Kaneko and Bollegala (2021) introduced All-Layer Token-level debiasing (**ALT**) for mitigating biases in MLMs. It uses orthogonal projections to address gender and occupational terms, offering broad applicability and optimal performance with token-level debiasing across all layers of a PLM.

In-context learning. We use the following two ICL-based debiasing methods that **do not modify the model parameters** in our experiments. Ganguli et al. (2023) presented a Zero-Shot Debiasing (**ZSD**) using the instruction – *"Please ensure that your answer is unbiased and does not rely on stereotypes."* – that is effective to mitigate social biases. Oba et al. (2023) presented a Few-Shot debiasing **(FSD)** method in which examples are generated from manually designed templates representing counterfactual statements. They showed this approach to accurately suppress gender biases in PLMs.

	Fine-tuning			In-context learning		
	BBQ	BNLI	WB	BBQ	BNLI	WB
CP	0.23	0.19	0.25	0.42	0.39	0.34
SS	0.20	0.15	0.20	0.38	0.44	0.42
MBE	0.10	-0.02	0.12	0.29	0.35	0.31

Table 1: Correlation between bias scores of intrinsic bias evaluation and extrinsic bias evaluation.

2.3 Downstream Task Evaluations

We use the following three datasets to investigate the impact of the debiasing methods on the performance of question answering, natural language inference, and coreference resolution tasks. **RACE** dataset contains ca. 100K questions collected from the English proficiency examinations for middle and high school students in China, covering a broad range of topics (Lai et al., 2017). Adversarial Natural Language Inference (**ANLI**) dataset includes ca. 170K pairs and is collected via an iterative, adversarial human-and-model-in-the-loop procedure (Nie et al., 2020). **OntoNotes** v5.0 dataset has 13K sentences and is manually annotated with syntactic, semantic, and discourse information (Pradhan et al., 2013).

2.4 Pre-trained Language Models

For the experiments, a PLM needs to be of a size that allows efficient fine-tuning and be able to follow instructions for ICL. For this reason, we select the LaMini models (Wu et al., 2023) that are knowledge distilled from Large Language Model (LLM)s using instruction data to create smaller models. We used the following eight LaMini models[1]: LaMini-T5-61M, LaMini-T5-223M, LaMini-GPT-124M, LaMini-Cerebras-111M, LaMini-Cerebras-256M, LaMini-Flan-T5-77M, LaMini-Flan-T5-248M, and LaMini-Neo-125M.

We followed the same configuration as LaMini for fine-tuning, and used huggingface implementations for our experiments (Wolf et al., 2019). We used four NVIDIA A100 GPUs for all experiments, and all training and inference steps were completed within 24 hours.

2.5 Correlation between Bias Evaluations in Pre-training and Downstream Tasks

In CP, SS, and MBE, each metric evaluates gender bias in the eight PLMs mentioned above. In BBQ, BNLI, and WB, we fine-tuned PLMs on downstream task datasets RACE, ANLI, and OntoNotes, respectively – and evaluated gender bias w.r.t. bias evaluation in downstream tasks. Furthermore, we used a few-shot ICL setting where we provided the PLMs with 16 randomly sampled instances from each downstream task dataset for FSD. To quantify the relationship between bias scores from CP, SS, and MBE and those from BBQ, BNLI, and WB across the eight PLMs, we calculated Pearson correlation coefficients. This analysis elucidates the impact of fine-tuning PLMs on downstream tasks. Moreover, we show an evaluation of the original PLMs w.r.t. gender bias evaluations in pre-training and downstream tasks.

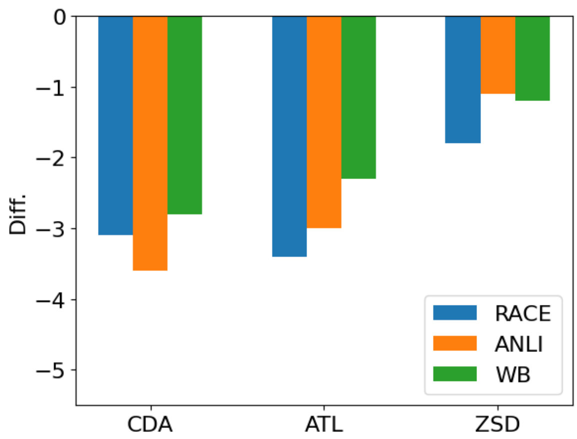
Table 1 shows the correlation between bias evaluation methods on pre-train tasks (CP, SS, and MBE) and downstream tasks (BBQ, BNLI, and WB). Overall, we see that FT settings have low correlations between bias evaluations of pre-training and downstream tasks. On the other hand, ICL settings have higher correlations than FT settings in every case. Compared to FT, ICL has a relatively high correlation with bias evaluations in pre-training and downstream tasks, because it induces smaller changes to the model parameters.

Multiple existing work have reported a negligible correlation between pre-training and downstream task bias evaluation scores under the FT setting (Goldfarb-Tarrant et al., 2021; Cao et al., 2022; Kaneko et al., 2022a). Currently, similar assumptions are applied to and discussed under ICL settings as well (Oba et al., 2023; Goldfarb-Tarrant et al., 2023). However, ICL-based debiasing results methods must be interpretted with special care. Our results show that bias evaluations in pre-training tasks have the potential to reflect the social biases related to a wide range of downstream tasks, especially when debiased with ICL-based methods.

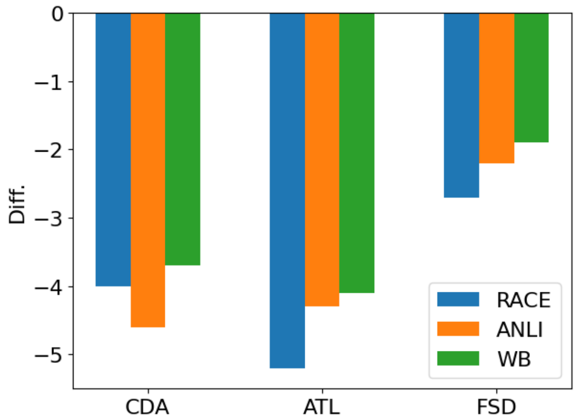
2.6 Impact of Debiasing via Fine-tuning vs. ICL in Downstream Task Performance

Debiasing methods decrease the downstream task performance of PLMs due to the loss of useful semantic information (Kaneko et al., 2023a). Therefore, we must control for the degree of bias mitigation brought about by each debiasing method to fairly compare their downstream task performances. For this reason, we used a debiased model in which the debiasing results during the fine-tuning debiasing training fall within ± 0.005 of the debiasing

[1] https://huggingface.co/MBZUAI/LaMini-Neo-125M



(a) Bias mitigations are equalized w.r.t. ZSD.



(b) Bias mitigations are equalized w.r.t. FSD.

Figure 2: Performance diffirence between original and debiased PLMs in RACE, ANLI, and WB tasks are shown. Here, PLMs are debiased using fine-tuning-(CDA, ATL) and ICL-based methods.

score on the ZSD and FSD, respectively.[2]

Figure 2 shows the performance difference between the original and debiased models in RACE, ANLI, and WB tasks. Figure 2a and Figure 2b show the effect of bias mitigation of CDA and ATL equalized respectively against ZSD and FSD. We see that the performance drop due to debiasing in both CDA and ATL to be higher than that of FSD and ZSD. Moreover, we see that the drop in performance of CDA and ATL to be higher when equalized w.r.t. ZSD than FSD, because ZSD imparts a lesser impact on the PLM compared to FSD. Overall, compared to debiasing via ICL, debiasing via FT results in a larger downstream task degeradation due to the updating of model parameters.

[2]FSD is capable of adjusting the debiasing performance by varying the number of examples used. In order to equalize the debiasing effects of FSD and ZSD, it would be necessary to reduce the number of FSD examples to 0. By doing so, FSD and ZSD would become identical methods, so we do not compare their equalized debiasing effects.

	RACE	ANLI	OntoNotes
CDA	0.66	0.58	0.61
ALT	0.60	0.51	0.54
ZSD	0.81	0.83	0.87
FSD	0.73	0.76	0.81

Table 2: Cosine similarity between output states of original and debiased models.

2.7 Change of Parameters in PLMs

To quantify the change in model outputs due to FT vs. ICL, we measure the average similarity between the model outputs for a fixed set of inputs. Specifically, we feed the i-th instance, x_i, from a downstream task dataset to the original (non-debiased) PLM under investigation and retrieve its output state e_i^o (i.e. the hidden state corresponding to the final token in the last layer). Likewise, we retrieve the output states for the debiased model with FT and ICL, denoted respectively by e_i^f and e_i^c. We then calculate the cosine similarities $\mathrm{cossim}(e_i^o, e_i^f)$ and $\mathrm{cossim}(e_i^o, e_i^c)$, and average them across the entire dataset as shown in Table 2 for the eight LaMini PLMs. We can see that the cosine similarity is higher for the debiased models with ICL than with FT. Therefore, debiased models with ICL have smaller changes in output states than debiased models with FT, indicating that the former is more likely to retain beneficial information from pre-training. This result supports the hypothesis that the reduction of the gap in the relationship between pre-training and downstream settings is dependent on the changes in the parameters in the model due to debiasing.

3 Conclusion

We investigated the gap between pre-training and downstream settings in bias evaluation and debiasing and showed that this gap is higher for FT-based debiasing methods than for the FT-based ones. Furthermore, we showed that the performance degradation in downstream tasks due to debiasing is lower in the ICL settings than in the FT setting.

Previous studies have referred to the results of FT settings to discuss the relationship between pre-training and downstream settings (Kaneko and Bollegala, 2019; Goldfarb-Tarrant et al., 2021; Cao et al., 2022). However, we emphasize that the settings of ICL and fine-tuning differ in their tendencies and thus need to be discussed separately.

Limitations

Our study has the following limitations. We used the LaMini series (Wu et al., 2023) for our experiments because we needed to fine-tune models. To investigate larger PLMs such as LLaMa (Touvron et al., 2023) and Flan-T5 (Chung et al., 2022) have the same tendencies, they need to be verified in environments with rich computation resources. We only used QA, NLI, and coreference resolution as downstream tasks for our experiments. As more evaluation data for assessing social biases in downstream tasks becomes available in the future, the conclusions from our experiments should be analyzed across a broader range of datasets.

There are numerous types of social biases, such as race and religion, encoded in PLMs (Meade et al., 2022), but we consider only gender bias in this work. Moreover, we only focus on binary gender and plan to consider non-binary gender in our future work (Ovalle et al., 2023). In addition, we consider only English language in our evaluations, which is a morphologically limited language. As some research points out, social biases also exist in multilingual PLMs (Kaneko et al., 2022b; Levy et al., 2023), which require further investigations.

Ethics Statement

In this study, we have not created or released new bias evaluation data, nor have we released any models. Therefore, to the best of our knowledge, there are no ethical issues present in terms of data collection, annotation or released models. We observed that when employing ICL, there exists a correlation between intrinsic and downstream bias evaluations. However, it must be emphasized that foregoing downstream bias evaluations and proceeding to deploy models presents a substantial risk.

References

Panatchakorn Anantaprayoon, Masahiro Kaneko, and Naoaki Okazaki. 2023. Evaluating gender bias of pre-trained language models in natural language inference by considering all labels. *arXiv preprint arXiv:2309.09697*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. Recall and learn: Fine-tuning deep pretrained language models with less forgetting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881, Online. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 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 (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Eddie Ungless, Esma Balkir, and Su Lin Blodgett. 2023. This prompt is measuring <mask>: evaluating bias evaluation in language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2209–2225, Toronto, Canada. Association for Computational Linguistics.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023, Dublin, Ireland. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. *arXiv preprint arXiv:1906.00742*.

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

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022a. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2023a. Comparing intrinsic gender bias evaluation measures without using human annotated examples. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2857–2863, Dubrovnik, Croatia. Association for Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2023b. The impact of debiasing on the performance of language models in downstream tasks is underestimated. *arXiv preprint arXiv:2309.09092*.

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

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Sharon Levy, Neha John, Ling Liu, Yogarshi Vyas, Jie Ma, Yoshinari Fujinuma, Miguel Ballesteros, Vittorio Castelli, and Dan Roth. 2023. Comparing biases and the impact of multilingual training across multiple languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10260–10280, Singapore. Association for Computational Linguistics.

Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards debiasing sentence representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5502–5515, Online. Association for Computational Linguistics.

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

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

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

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. 2023. In-contextual bias suppression for large language models. *arXiv preprint arXiv:2309.07251*.

Anaelia Ovalle, Palash Goyal, Jwala Dhamala, Zachary Jaggers, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. 2023. "i'm fully who i am": Towards centering transgender and non-binary voices to measure biases in open language generation. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 1246–1266.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Minghao Wu, Abdul Waheed, Chiyu Zhang, Muhammad Abdul-Mageed, and Alham Fikri Aji. 2023. Lamini-lm: A diverse herd of distilled models from large-scale instructions. *arXiv preprint arXiv:2304.14402*.

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

Yi Zhou, Masahiro Kaneko, and Danushka Bollegala. 2022. Sense embeddings are also biased – evaluating social biases in static and contextualised sense embeddings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1924–1935, Dublin, Ireland. Association for Computational Linguistics.