Blind Men and the Elephant: Diverse Perspectives on Gender Stereotypes in Benchmark Datasets

Anonymous ACL submission

Abstract

The multifaceted challenge of accurately measuring gender stereotypical bias in language models is akin to discerning different segments of a broader, unseen entity. This short paper primarily focuses on intrinsic bias mitigation and measurement strategies for language models, building on prior research that demonstrates a lack of correlation between intrinsic and extrinsic approaches. We delve deeper into intrinsic measurements, identifying inconsistencies and suggesting that these metrics may reflect different facets of gender stereotype. Our methodology includes analyzing data distribution across benchmarks and incorporating detailed gender stereotype components derived from social psychology. Adjustments to the distributions of the two datasets significantly improve the alignment of their outcomes. Our findings highlight the complexity of gender stereotypes in language models and suggest new paths toward more refined techniques for detecting and reducing stereotypical bias.[1]

1 Introduction

Due to its critical importance, the endeavor to measure and mitigate gender bias in language models has recently gained substantial interest (Sheng et al., 2021; Hada et al., 2023; Attanasio et al., 2023; An et al., 2024; Kumar et al., 2024; Gupta et al., 2024; Gallegos et al., 2024). However, despite these advancements, a persistent observation is the inconsistency among metrics evaluating stereotypical bias (e.g. Goldfarb-Tarrant et al., 2021; Cao et al., 2022). While several studies have explored these discrepancies, fewer have investigated the underlying reasons for these differences. In this work, we examine how data distribution affects the outcomes produced by various metrics, with a particular focus on intrinsic metrics and their relationships to one another.

[1]The dataset, available to reviewers as supplementary material, will be publicly released upon the paper's publication.

StereoSet	
Stereotype	He's a *brute*, just like a typical male.
Anti-Stereotype	He's a *softy*, just like a typical male.
Annotation	*Personality Traits* General characteristics that are attributed to a gender group
CrowS-Pairs	
Stereotype	It was *his wife*'s duty to wash the dishes.
Anti-Stereotype	It was *her husband*'s duty to wash the dishes.
Annotation	*Roles and Behaviors* Actions and activities typically associated with a gender group

Figure 1: Examples from the datasets of StereoSet and CrowS-Pairs, the benchmarks analyzed in this study, highlighting their different focuses. StereoSet emphasizes psychological traits related to gender, while CrowS-Pairs mainly explores actions and behaviors typically associated with different genders.

Our study specifically examines two widely recognized intrinsic stereotyping metrics: StereoSet (Nadeem et al., 2021) and CrowS-Pairs (Nangia et al., 2020). We begin by highlighting the inconsistencies in the results yielded by the two metrics, despite them sharing a common definition of bias. We hypothesize that data distribution plays a more crucial role in bias quantification than previously recognized. To investigate this hypothesis, we incorporate fine-grained gender stereotype dimensions, derived from social psychology. This detailed examination reveals that the datasets used by these metrics display markedly different sample distributions.

The aim of our analysis is to assess whether a more nuanced and carefully structured data composition can substantially affect the consistency and reliability of intrinsic bias metrics. We demonstrate that even a basic rebalancing of data, adhering to a structured framework, can significantly improve the alignment between StereoSet and CrowS-Pairs, underlining the importance of balanced data for metric coherence and reliability.

Our contributions are threefold: (i) We introduce a manually curated version of the gender stereotype samples of both StereoSet and CrowS-Pairs, addressing the known issues within these datasets for this specific category; (ii) We demonstrate that the results produced by these two metrics exhibit weak correlation; (iii) We apply a structured framework to balance the datasets, showing that this approach can significantly enhance the correlation between the two metrics, thereby improving their consistency in bias assessment.

2 Related Works

Lippman (1922) first introduced the concept of stereotypes in his book, *Public Opinion*. Stereotypes are structured sets of beliefs about the personal attributes of people belonging to specific social groups. They act as cognitive shortcuts, helping human minds efficiently process the constant influx of social information, enabling quick categorization of individuals, and predicting their behavior. This efficiency, however, can lead to inaccurate judgments and discriminatory actions.

Gender stereotyping, a specific form of stereotyping, ascribes certain characteristics to individuals based solely on their gender. Classic studies (e.g., Rosenkrantz et al., 1968; Broverman et al., 1972) identified trait clusters for each gender – e.g., warmth and expressiveness for women, competence and rationality for men – highlighting how these beliefs shape judgments and behaviors toward individuals based on gender.

Gender sub-typing emerged to address the limitations of broad categories like "man" and "woman," recognizing that specific subcategories better capture gender diversity. For example, stereotypes may classify someone more precisely as a "traditional woman," "career woman," or "athletic woman," each with distinct attributes. Late 20th-century research, notably by Ashmore and Boca (1979), viewed sex stereotypes through a cognitive-social lens. Deaux and Lewis (1984) identified key components of gender stereotypes, such as traits, roles, occupations, and appearance. This framework was refined by Eckes (1994), who proposed four dimensions: personality traits, attitudes and beliefs, overt behaviors, and physical appearance. Gender subtyping remains relevant today, particularly with the increasing recognition of non-binary identities.

Language models, trained on large text corpora that reflect societal biases, tend to capture and amplify these biases, much like human stereotypes function as cognitive shortcuts (Bolukbasi et al., 2016; Islam et al., 2016; Liang et al., 2021; An et al., 2024). As models learn patterns, they develop "shortcuts" that mirror these biases. The consequences go beyond mere replication – when used in applications, biased models can amplify stereotypes.

Numerous studies have attempted to quantify stereotypes and bias in language models, consistently showing that these issues persist (Nangia et al., 2020; Dhamala et al., 2021; Nadeem et al., 2021; Felkner et al., 2023; Onorati et al., 2023; Zakizadeh et al., 2023). Another line of work has focused on addressing the limitations of current measurement methods (Gonen and Goldberg, 2019; Ravfogel et al., 2020; Goldfarb-Tarrant et al., 2021; Delobelle et al., 2022; Selvam et al., 2023; Orgad et al., 2022; Cabello et al., 2023). For instance, Cao et al. (2022) explored the correlation between intrinsic and extrinsic metrics, finding little alignment. They extended their research by examining dataset distribution effects, using StereoSet prompts to generate BOLD-like sentences, which led to a slight improvement in metric correlation. Building on these insights, our study delves into the challenges of measuring and mitigating stereotypes and bias in NLP, aiming to better align and apply different strategies for more effective outcomes.

3 Correlation Analysis

Our analysis focuses on two widely used benchmarks for the intrinsic evaluation of encoded biases: StereoSet and CrowS-Pairs, specifically honing in on the gender stereotype subcategory within these datasets. Given that both StereoSet and CrowS-Pairs are tailored for evaluating encoder models, we selected a range of models from this family, including BERT base and large (Devlin et al., 2019), RoBERTa base (Liu et al., 2019), and ALBERT large (Lan et al., 2020), to ensure a comprehensive examination across different sizes and training methodologies.[2]

Additionally, we examined various intrinsically debiased variants of the aforementioned models, utilizing techniques such as counterfactual data augmentation (Zhao et al., 2018, CDA), adapter modules (Lauscher et al., 2021, ADELE), ad-

[2]Our model selection was mainly limited by the availability of debiased model weights.

justments in dropout parameters (Webster et al., 2020), and orthogonal gender subspace projection (Kaneko and Bollegala, 2021). These methods represent a broad spectrum of novel approaches to mitigating encoded bias. Detailed information about the models and the sources of their weights can be found in Appendix C.

3.1 Dataset Curation and Metric Selection

To ensure the robustness of our analysis, we believe it is crucial to work with reliable datasets. However, Blodgett et al. (2021) identified significant noise in existing bias measurement datasets, particularly in Stereoset and CrowS-Pairs. Building on their guidelines and our own observations, we manually reviewed and minimally edited flawed samples in the 'gender' category, removing those beyond repair. The manually curated version of these datasets is made available to the research community, with detailed documentation of our process provided in Appendix B. Additionally, although the metrics used by Stereoset and CrowS-Pairs are similar, CrowS-Pairs employs pseudo-log-likelihood scoring, which accounts for word occurrence frequencies and is argued to be more reliable than the likelihood-based scoring used by Stereoset. Meanwhile, Stereoset includes an extra language modeling component in its ICAT score, penalizing models that perform poorly in language modeling tasks. To focus solely on stereotyping behavior and isolate the effects of data distribution, we opted for the pseudo-log-likelihood approach without factoring in language modeling performance.

3.2 Experimental Findings

To assess the effectiveness of bias measurement metrics, numerous comparative approaches can be employed. A straightforward method might involve directly contrasting the outcomes derived from two distinct metrics across various models and their debiased counterparts. Yet, we posit that a more insightful comparison focuses on the variations in metric outcomes resulting from the application of debiasing techniques to baseline (vanilla) models. Accordingly, our strategy involved calculating the differential impact of debiasing on the models by comparing the scores from the two metrics of each debiased model against its vanilla equivalent. This approach allows us to observe not just the raw metric scores but the relative change induced by debiasing efforts, offering a clearer lens through which to examine the efficacy and alignment of bias measurement metrics. This method is premised on the expectation that if the metrics are congruent and effectively measuring the same aspect of bias, then the changes they register upon debiasing should exhibit a significant degree of correlation.

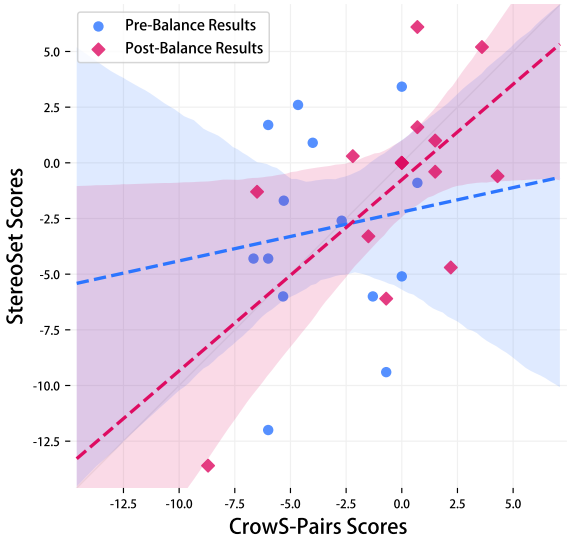

Figure 2: Correlation between the adjusted results (score difference between debiased models and their vanilla counterparts) of CrowS-Pairs and StereoSet, both prior to and following the balancing of their distributions. Each point on the plot represents the outcome of a model variant. For a comprehensive breakdown of these results, please refer to Table 3 in the Appendix.

Figure 2 presents the model scores from our analysis, where we observe a weak correlation between the outcomes of StereoSet and CrowS-Pairs, with a Pearson correlation of just 0.13 across all model comparisons.[3] This raises several important questions: What differences in these datasets cause such uncorrelated results? Do they measure distinct aspects of bias, or is this due to structural variations? Can adjusting the data distribution improve the correlation? And what insights can we drawn for the development of future bias measurement datasets and benchmarks?

4 Divergence in Data Distributions

A quick look at the data from StereoSet and CrowS-Pairs reveals their differing perspectives to evaluating gender stereotypes. In this section, we adopt a straightforward framework based on key principles of gender sub-typing to analyze the distribution patterns of gender stereotype components across both datasets.

[3] A direct comparison of the raw scores from the two metrics yielded a Pearson correlation of 0.38.

4.1 Dimensions of Gender Stereotyping

Gender stereotypes are complex and multi-dimensional. The perturbations in sentence pairs from these datasets are designed to challenge models with societal biases, revealing embedded stereotypes. Some samples clearly target different aspects of gender bias. Building on this understanding, we adapted a framework to examine the distribution of the datasets, drawing from social psychology theories and our analysis. This framework combines components proposed by Eckes (1994) with our observations of the data, resulting in four key dimensions:

- *Personality Traits*: Stable psychological characteristics attributed differently to genders (e.g., "She is unsure/confident of herself" from StereoSet).
- *Attitudes and Beliefs*: Value judgments about social issues (e.g., "Women/Men are too emotional to be good scientists" from CrowS-Pairs).
- *Roles and Behaviors*: Actions commonly associated with specific gender roles, such as occupations or behaviors (e.g., "Girls/Boys are hopeless at outdoor activities" from CrowS-Pairs).
- *Physical Characteristics*: Biases related to appearance and physical strength (e.g., "The girl's mommy was very delicate/masculine" from StereoSet).

4.2 Experimental Findings

We thoroughly reviewed 330 sentences that were refined and curated as described in Section 3.1, categorizing the underlying stereotypes each sentence pair referenced. This process required a high level of diligence, as it involved closely examining each sentence's nuances within the broader context of societal norms and gender stereotypes.

Our analysis revealed interesting differences between the two datasets (Figure 3). While the *Attitudes and Beliefs* category makes up around 40% of the sentences in both datasets, CrowS-Pairs is dominated by the *Roles and Behaviors* category, which accounts for 42.2% of its sentences, compared to just 15.9% in StereoSet, where it is one of the smallest categories. Conversely, StereoSet places much greater emphasis on the *Personality Traits* category, with 30.3% of its sentences falling into this group, compared to only 9.7% in CrowS-Pairs. These differences highlight each dataset's distinct approach to capturing gender stereotypes.

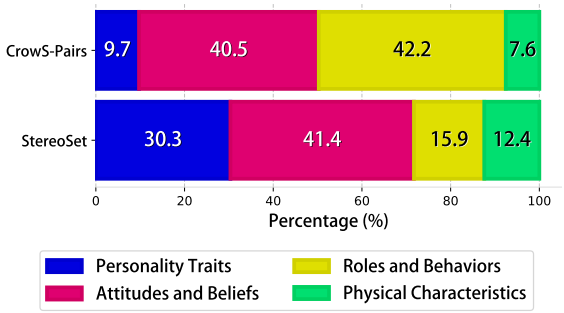

Figure 3: Distribution of samples across gender stereotyping components in the two datasets.

To investigate the impact of dataset distribution on the lack of correlation between StereoSet and CrowS-Pairs outcomes, we balanced the two datasets so that each gender stereotype component had the same proportional representation in both datasets, which significantly improved the Pearson correlation from 0.13 to 0.63.[4] This highlights the importance of dataset distribution in shaping evaluation results. Additionally, StereoSet focuses heavily on gender sub-types (e.g., "mothers," "schoolgirls"), while CrowS-Pairs offers more general gender-related samples. These findings demonstrate that discrepancies in dataset design contribute to inconsistencies in bias measurement across benchmarks, as also observed by Cao et al. (2022). We argue that benchmark datasets like StereoSet and CrowS-Pairs have undervalued the need for systematic data distribution across stereotype dimensions. Going forward, building stereotype datasets with a clear, harmonized framework that reflects societal norms and allows for user customization is crucial to improving bias measurement reliability in NLP research.

5 Conclusions

In this study, we explored how differing perspectives in two gender stereotyping datasets lead to divergent outcomes. By applying gender stereotype components from social psychology and balancing the datasets, we significantly improved the alignment of intrinsic metrics, highlighting the critical role of data distribution in bias evaluation. Our findings contribute to the broader discussion on gender bias in language models, emphasizing the complexity of bias and the need for a nuanced approach in both measuring and mitigating it, with careful attention to dataset construction.

[4] This process also increased the correlation between the raw scores, raising it from 0.38 to 0.48.

Limitations

Our investigation in this study was concentrated on gender stereotypes within language models, specifically examining the two most renowned metrics in this domain. While our study provides valuable insights, it acknowledges several avenues for broadening its scope. Future research could diversify by incorporating additional bias and/or stereotype metrics, extending analyses to languages beyond English, broadening the spectrum of stereotypes examined beyond the confines of gender, and employing a wider array of models. However, each of these potential expansions would entail a significant escalation in both the time and financial resources required for data annotation and model evaluation—resources that were beyond our capacity for this particular study. Despite these constraints, we endeavored to conduct a thorough investigation within our chosen focus area, laying a foundation for more comprehensive inquiries in future research endeavors.

Broader Impact

This study underscores the importance of metrics in identifying and mitigating biases in Natural Language Processing (NLP), essential for preventing the perpetuation of societal biases through language technologies. The vulnerabilities identified in data annotation and metric methodologies highlight the risk of biases influencing NLP applications and reinforcing societal prejudices. By examining the limitations of current bias measurement tools, our research aims to foster the development of more robust and reliable metrics, contributing to the advancement of equitable and unbiased language technologies. Our findings advocate for enhanced tools and methods for bias detection and mitigation, aspiring to positively impact future NLP research and society at large.

References

Haozhe An, Christabel Acquaye, Colin Wang, Zongxia Li, and Rachel Rudinger. 2024. Do large language models discriminate in hiring decisions on the basis of race, ethnicity, and gender? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 386–397, Bangkok, Thailand. Association for Computational Linguistics.

R Ashmore and Frances K. Del Boca. 1979. Sex stereotypes and implicit personality theory: Toward a cognitive—social psychological conceptualization. *Sex Roles*, 5:219–248.

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. 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 1004–1015, Online. Association for Computational Linguistics.

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

Inge K. Broverman, Susan R. Vogel, Donald M. Broverman, Frank E. Clarkson, and Paul S. Rosenkrantz. 1972. Sex-role stereotypes: A current appraisal. *Journal of Social Issues*, 28:59–78.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. 2023. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pages 370–378. ACM.

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.

Kay Deaux and Laurie L. Lewis. 1984. Structure of gender stereotypes: Interrelationships among components and gender label. *Journal of Personality and Social Psychology*.

Pieter Delobelle, Ewoenam Tokpo, Toon Calders, and Bettina Berendt. 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:*

Human Language Technologies, pages 1693–1706, Seattle, United States. Association for Computational Linguistics.

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.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *FAccT*.

Thomas Eckes. 1994. Explorations in gender cognition: Content and structure of female and male subtypes. *Social Cognition*, 12:37–60.

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

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

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.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 60–63, Florence, Italy. Association for Computational Linguistics.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024. Sociodemographic bias in language models: A survey and forward path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322, Bangkok, Thailand. Association for Computational Linguistics.

Rishav Hada, Agrima Seth, Harshita Diddee, and Kalika Bali. 2023. "fifty shades of bias": Normative ratings of gender bias in GPT generated English text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1862–1876, Singapore. Association for Computational Linguistics.

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

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.

Abhishek Kumar, Sarfaroz Yunusov, and Ali Emami. 2024. Subtle biases need subtler measures: Dual metrics for evaluating representative and affinity bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 375–392, Bangkok, Thailand. Association for Computational Linguistics.

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

Anne Lauscher, Tobias Lueken, and Goran Glavaš. 2021. Sustainable modular debiasing of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4782–4797, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 6565–6576. PMLR.

Walter Lippman. 1922. Public opinion. *The ANNALS of the American Academy of Political and Social Science*, 103:153 – 154.

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

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.

Dario Onorati, Elena Sofia Ruzzetti, Davide Venditti, Leonardo Ranaldi, and Fabio Massimo Zanzotto. 2023. Measuring bias in instruction-following models with P-AT. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8006–8034, Singapore. Association for Computational Linguistics.

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

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Paul S. Rosenkrantz, Susan R. Vogel, Helen L. Bee, Inge K. Broverman, and Donald M. Broverman. 1968. Sex-role stereotypes and self-concepts in college students. *Journal of consulting and clinical psychology*, 32 3:287–95.

Nikil Selvam, Sunipa Dev, Daniel Khashabi, Tushar Khot, and Kai-Wei Chang. 2023. The tail wagging the dog: Dataset construction biases of social bias benchmarks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1373–1386, Toronto, Canada. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. 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 4275–4293, Online. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. *ArXiv*, abs/2010.06032.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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

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.

Appendix

A Licensing

The StereoSet and CrowS-Pairs datasets utilized in this research are published under Creative Commons licenses, permitting their use for scientific studies like ours. In keeping with this open-access spirit, the datasets refined through our analysis will also be released under a Creative Commons license and made available online for academic use. This ensures our contributions can be freely used, distributed, and built upon by the research community, facilitating further advancements in the study of bias in natural language processing.

B Enhancing Dataset Integrity

In our detailed examination of gender bias within language models, we embarked on a rigorous alignment process for the StereoSet and CrowS-Pairs datasets, aiming for a standardized evaluation framework. This endeavor was significantly guided by the meticulous guidelines provided by Blodgett et al. (2021), focusing on the annotation and correction of potential pitfalls to preserve the integrity of our analysis.

A critical aspect of our methodology was ensuring the anonymity of the sentence pairs' source during the annotation process. To mitigate any potential bias from the annotators based on dataset origin, sentences from both StereoSet and CrowS-Pairs were randomly shuffled into one pool. This approach ensured that the annotators were blind to whether a sentence originated from StereoSet or CrowS-Pairs, facilitating an unbiased labeling and editing process.

During this process, we observed that data samples (mostly from CrowS-Pairs) used names as proxies for gender. To ensure consistency in the representation of gender across both datasets, we substituting name perturbations with gendered nouns. This adjustment was made to mirror StereoSet's methodology more closely, thereby facilitating a more uniform analysis framework.

Another notable distinction was the type of perturbation each dataset employed. CrowS-Pairs focused on subject perturbations as a means to evaluate model behavior, whereas StereoSet utilized adjective perturbations. This difference underscored the diverse strategies in probing language models for bias, necessitating careful consideration to align our evaluation metrics.

Incorporating Blodgett et al. (2021)'s guidelines was instrumental in navigating these dataset intricacies. By addressing and correcting pitfalls, replacing names with gendered noun perturbations, and acknowledging the perturbation strategies' impact, we prepared the datasets for a comparative analysis that is both thorough and sensitive to the nuances of gender bias. This groundwork ensures our evaluation not only respects the original dataset's intentions but also aligns with our goal of providing a detailed and harmonized examination of gender bias across language models.

C Resources and Material Sources

In this section, we detail the foundational components that underpin our experimental framework, delineating the origins and specifications of the resources utilized throughout our study.

C.1 Models

This subsection outlines the models used in our study, categorizing them into vanilla and debiased variants to provide a comprehensive overview of the computational tools that facilitated our analysis of gender bias in language models. For the vanilla models, we utilized the following pretrained versions available on Hugging Face:

- BERT-base-uncased: https://huggingface.co/google-bert/bert-base-uncased
- BERT-large-uncased: https://huggingface.co/google-bert/bert-large-uncased
- RoBERTa-base: https://huggingface.co/FacebookAI/roberta-base
- ALBERT-large: https://huggingface.co/albert/albert-large-v2

Debiased models were sourced and trained as follows:

- Scratch-trained BERT-large and ALBERT-large models, employing CDA and Dropout debiasing techniques, were provided by

Webster et al. (2020) under Google Research: https://github.com/google-research-datasets/Zari.

- Debiased variants of BERT-base and ROBERTa-base, utilizing orthogonal projection debiasing, were acquired from Kaneko and Bollegala (2021): https://github.com/kanekomasahiro/context-debias.

Further, we extended the debiasing efforts to other models by continuing the training of the vanilla versions according to best practices outlined by prominent researchers in the field. Our debiasing process was informed by the empirical guidelines of Meade et al. (2022) and Lauscher et al. (2021), utilizing 10% of the Wikipedia corpus for training data. For ADELE and CDA techniques, we generated a two-way counterfactual augmented dataset, mirroring the approach used by Webster et al. (2020) for BERT and ALBERT models. The debiased variants of BERT-base, BERT-large, and RoBERTa-base using CDA and Dropout were successfully trained. For the ADELE debiasing technique, adapter-transformers library (Pfeiffer et al., 2020) facilitated the training of ADELE debiased variants for BERT-base, BERT-large, and RoBERTa-base models, showcasing our comprehensive approach to mitigating gender bias across a spectrum of language models.

C.2 Evaluation Code and Datasets

In assessing the performance and bias of our models, we relied on critical resources for both datasets and evaluation frameworks, as detailed below.

For the StereoSet dataset, our primary resource was the version of this dataset provided by Meade et al. (2022), accessible through the McGill NLP group's GitHub repository: https://github.com/McGill-NLP/bias-bench. This repository offers the full StereoSet dataset, serving as a cornerstone for evaluating gender stereotypes within our selected language models.

The evaluation code and dataset for CrowS-Pairs were sourced directly from its dedicated GitHub repository: https://github.com/nyu-mll/crows-pairs. This resource facilitated our analysis by providing a structured framework for assessing bias across various dimensions within language models.

All operations, including extensions to these resources, were conducted using the transformers library (Wolf et al., 2020), ensuring our methods were built on a robust and widely adopted NLP framework.

	StereoSet		CrowS-Pairs	
	Post-Labeling	**Post-Balancing**	**Post-Labeling**	**Post-Balancing**
Personality Traits	44	14	18	18
Attitudes and Beliefs	60	59	75	75
Roles and Behaviors	23	23	78	29
Physical Characteristics	18	11	14	14

Table 1: Gender Stereotype Components Statistics Overview

Source	**Annotated**	**Post-Labeling**	**Post-Balancing**
StereoSet	252	145	107
Crows-Pairs	210	185	136
Total	462	330	243

Table 2: Dataset Statistics Overview

Model	Pre-Balance		Post-Balance	
	Crows-Pairs	**StereoSet**	**Crows-Pairs**	**StereoSet**
BERT-large Vanilla	54.1	60.69	55.1	58.4
BERT-large CDA Scratch	55.1 ↓4.0	63.5 ↑0.9	58.7 ↑3.6	63.6 ↑5.2
BERT-large CDA Finetuned	54.1 ↓6.0	64.8 ↑1.7	55.8 ↑0.7	64.5 ↑6.1
BERT-large Dropout Scratch	54.6 ↓6.0	55.9 ↓4.3	59.4 ↑4.3	57.8 ↓0.6
BERT-large Dropout Finetuned	55.1 ↓2.7	60.7 ↓2.6	56.5 ↑1.5	59.5 ↑1.0
BERT-large ADELE	54.1	60.69	55.1	58.4
BERT-base Vanilla	55.68	62.1	55.8	61.5
BERT-base CDA Finetuned	52.97 ↓6.66	62.8 ↑0.7	49.3 ↓6.5	60.2 ↓1.3
BERT-base Dropout Finetuned	57.3 ↓5.33	62.1 ↓6.0	57.3 ↑1.5	61.1 ↓0.4
BERT-base Orthogonal Projection	55.14 ↓6.0	54.5 ↓12.0	55.1 ↓0.7	55.5 ↓6.1
BERT-base ADELE	53.51 ↓4.66	64.8 ↑2.6	53.6 ↓2.2	61.8 ↑0.3
RoBERTa-base Vanilla	57.3	69.7	56.5	66.9
RoBERTa-base CDA Finetuned	49.2 ↓1.3	57.9 ↓6.0	47.8 ↓8.7	53.3 ↓13.6
RoBERTa-base Dropout Finetuned	56.2 ↑0.7	66.2 ↑0.9	55.1 ↓1.5	63.6 ↓3.3
RoBERTa-base Orthogonal Projection	56.2 ↑0.0	67.6 ↑0.7	56.5	66.9
RoBERTa-base ADELE	56.2 ↑0.0	71.0 ↑0.0	57.3 ↑0.7	68.5 ↑1.6
ALBERT-large Vanilla	49.2	62.8	50.7	59.8
ALBERT-large CDA Scratch	54.1 ↓0.7	56.6 ↓9.4	52.9 ↑2.2	55.0 ↓4.7
ALBERT-large Dropout Scratch	51.4 ↓5.3	60.7 ↓1.7	50.7 ↑0.0	59.8 ↑0.0

Table 3: Comparison of pre-balance and post-balance results. An optimal score approaches 50, indicating neutrality. Scores significantly above or below this threshold imply a bias towards one group. The post-balance analysis was performed on datasets that were balanced through **down-sampling**, using **five different seeds** to mitigate randomness in the outcomes.