Blind Men and the Elephant: Diverse Perspectives of Gender Bias in Stereotype Benchmarks

Anonymous ACL submission

Abstract

The multifaceted challenge of accurately measuring gender bias in language models is akin to discerning different segments of a broader, unseen entity. This short paper mainly focuses on intrinsic bias mitigation and measurement strategies for language models, building on prior research that demonstrate a lack of correlation between intrinsic and extrinsic approaches. We delve deeper into the realm of intrinsic measurements, identifying inconsistencies and positing that these metrics might reflect diverse facets of gender bias. Our methodology encompasses an analysis of data distribution across benchmarks coupled with the implementation of an intricate gender bias categorization derived from social psychology. Adjustments made to the distributions of the two datasets lead to significant enhancement in the alignment of their outcomes. Our findings not only underscore the complexity inherent in gender bias in language models but also forge new paths toward more refined techniques for bias detection and reduction. [1]

1 Introduction

Due to its critical importance, the endeavor to measure and reduce gender bias in language models has recently gained substantial interest (Silva et al., 2021; Kaneko and Bollegala, 2021; Lauscher et al., 2021; Zhao et al., 2018; Webster et al., 2020). However, despite these advancements, a persistent observation is the inconsistency among metrics evaluating stereotypical bias (Goldfarb-Tarrant et al., 2021; Orgad et al., 2022). This recurring observation highlights a crucial gap in comprehending the nuances of bias dynamics within language models. Cao et al. (2022) suggested that issues such as misalignment in data distribution might contribute to the inconsistent conclusions from various metrics, though their investigation in this area was somewhat limited. In this work, we aim to expand on this area, with a particular focus on intrinsic metrics and their interrelationships.

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: Representative samples from StereoSet and CrowS-Pairs (the two datasets examined in this work) that highlight the distinct thematic focus of each dataset. StereoSet predominantly features sentences related to *personality traits*, i.e., psychological characteristics associated with different genders. In contrast, CrowS-Pairs primarily focuses on *roles and behaviors*, i.e., observable, temporally consistent actions and patterns.

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. Building on this, we put forth the hypothesis that the influence of data distribution on bias quantification may be more critical than previously considered. To investigate this hypothesis, we incorporate fine-grained gender stereotype subcategories, 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

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

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.

2 Related Works

2.1 Gender Bias

The investigation into stereotypes, including those based on gender, mainly originates from the field of social psychology (Lippman, 1922). Early research broadly categorized gender into "men" and "women," focusing on the terms and concepts linked to these categories (Broverman et al., 1972). As research evolved, there was a shift towards recognizing more nuanced subcategories of gender, indicating a more sophisticated grasp of gender stereotypes (Eckes, 1994). A pivotal study by Deaux and Lewis (1984) explored the various facets of gender stereotyping, analyzing the interaction of these aspects within the larger framework of societal perceptions of gender.

This evolution towards a more refined understanding of gender stereotypes in social psychology mirrors a similar progression in the field of Natural Language Processing (NLP). Initial gender bias studies in NLP mirrored the broad categorizations of traditional social psychology (Islam et al., 2016; Bolukbasi et al., 2016). Recently, however, there has been a shift towards addressing more specific gender subtypes and complexities (Felkner et al., 2023).

This study builds on these foundations, integrating insights from social psychology into NLP to deepen our understanding of gender bias in language models. By bridging these fields, we aim to refine bias evaluation metrics in NLP, recognizing gender stereotype as a complex, multifaceted phenomenon.

2.2 Consistency of Bias Metrics

The techniques to mitigate and measure bias in NLP models are generally categorized into two main approaches: *intrinsic* and *extrinsic*. Intrinsic methods directly engage with the language modeling task to tackle bias (Nangia et al., 2020; Nadeem et al., 2021), while extrinsic methods focus on downstream tasks, often targeting the outputs of the classifiers built on top of a language model (De-Arteaga et al., 2019).

One might expect that addressing bias through either intrinsic or extrinsic methods would improve a model's fairness across various evaluation metrics. However, recent studies have begun to reveal that this assumption may not hold true. Research led by Goldfarb-Tarrant et al. (2021) unveiled a surprising disconnect: intrinsic debiasing techniques and their measurable impacts on bias, as captured by intrinsic bias metrics like WEAT (Islam et al., 2016), do not align with the biases manifesting in practical applications such as co-reference resolution and hate speech detection. This observation was further supported by Orgad et al. (2022), who discovered a similar lack of correlation but in the reverse direction. Their work demonstrated that even when bias is mitigated extrinsically in tasks like occupation classification, it does not always reflect in intrinsic bias metrics such as CEAT (Guo and Caliskan, 2021). Cao et al. (2022) added to this discourse by suggesting that intrinsic and extrinsic measures often operate independently, without any significant correlation in their outcomes. They proposed that aligning the definitions of bias, protected groups, and evaluation datasets could be key to bridging this gap.

Building on these insights, our study delves into the intricacies of intrinsic bias mitigation and measurement within NLP. We aim to investigate the relationships–or lack thereof–between different intrinsic bias measurement and mitigation strategies, hoping to shed light on how these techniques can be more effectively aligned and applied.

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

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

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), adjustments 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 Refinement

Acknowledging the critical role of dataset integrity in our analysis, we implemented measures to reduce noise and other confounding factors in the evaluation datasets, drawing on recommendations from Blodgett et al. (2021). More details about this process are provided in Appendix B. Moreover, to rule out the impact stemming from differences in metrics, we standardized the evaluation setting across the two metrics. Notably, StereoSet incorporates a language modeling score in its final assessment, penalizing models that perform poorly in language modeling objectives. However, CrowS-Pairs employs a pseudo-log-likelihood calculation, argued to be more reliable due to its incorporation of word occurrence frequencies. For our analysis, we opted for the pseudo-likelihood calculation, focusing solely on stereotyping behavior without considering other model attributes.

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.

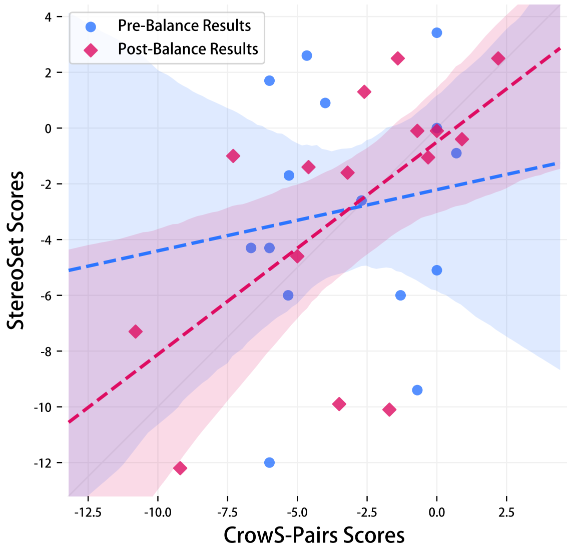

Figure 2: Correlation between the adjusted results (difference between debiased models and their vanilla counterparts) of CrowS-Pairs and StereoSet, both prior to and following the balancing of their distributions. Each point on the plot represents the outcome of a model variant. For a comprehensive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

Following our dataset refinement efforts, we observed that the outcomes from StereoSet and CrowS-Pairs exhibit a lack of correlation. A Pearson correlation analysis yielded a mere 0.13 across all model comparisons (see results in Figure 2). This finding opens a Pandora's box of questions regarding the nature and effectiveness of these evaluation metrics. The most pressing questions include: Why do these metrics, even after extensive adjustments, fail to correlate? Considering their shared goal of measuring stereotypes in language models, what causes this disconnect? And what are the broader implications of this lack of correlation?

4 Divergence in Dataset Distributions

We hypothesized that the divergent perspectives of StereoSet and CrowS-Pairs, reflected in their sample distributions, significantly contribute to the lack of correlation seen in our preliminary findings. Inspired by core principles from social psychology, we developed a framework focused on essential gender aspects to examine the distribution patterns across the two datasets.

4.1 Dimensions of Gender Stereotyping

Recognizing gender stereotypes as complex, multifaceted constructs highlights their significant yet nuanced impact on shaping perceptions. Inspired by this understanding, we introduce a framework designed to examine gender bias within NLP datasets, integrating key social psychology theories with our analytical insights. Our framework merges the categories proposed by Deaux and Lewis (1984) and Eckes (1994) with observations from our comprehensive analysis of bias in datasets. This synthesis results in four distinct, identifiable dimensions:

- ***Personality Traits***: Stable, individual psychological characteristics that attributed differently to genders.
- ***Attitudes and Beliefs*****:** Value judgments and beliefs about various social issues and targets.
- ***Roles and Behaviors***: Actions and activities commonly associated with specific gender groups, including occupations, roles, overt behaviors, and behavioral preferences.
- ***Physical Characteristics***: Biases related to appearance and physical strength.

4.2 Experimental Findings

In the StereoSet and CrowS-Pairs datasets, sentence pairs with perturbations were specifically designed to challenge models with societal stereotypes, thereby uncovering embedded biases. Our approach posited that these stereotypes fall into one of the four gender stereotype components we defined. We conducted a thorough review of 266 sentences that were refined and enhanced as described in Section 3.1, to assess their congruence with our gender stereotype framework. This evaluation process demanded a high degree of diligence, necessitating a deep dive into each sentence's implications within the complex matrix of societal norms and stereotypical representations.

Figure 3 shows the distribution of instances in the two datasets across the four dimensions. Our analysis highlighted the significant gap between the compositions of the two datasets: StereoSet predominantly explores stereotypes related to personality traits with more than half of the instances belonging to this dimension (9.3% of instances in CrowS-Pairs lie within this category). On the contrary, CrowS-Pairs focuses on roles and behaviors with nearly half of its instances (compared to 12.8% for StereoSet).

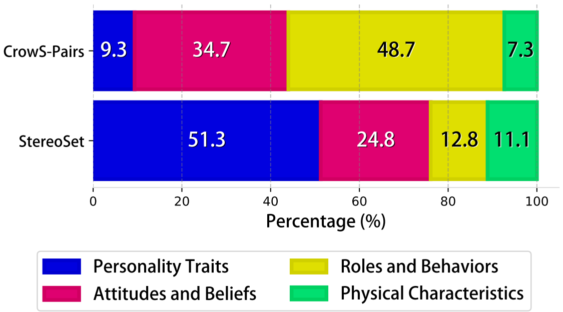

Figure 3: Distribution of samples across gender stereotyping components in the two datasets.

To explore the influence of dataset distribution on the lack of correlation between StereoSet and CrowS-Pairs outcomes, we balanced the datasets for equitable representation across gender stereotype components and re-evaluated the models. This process, illustrated in Figure 2, significantly increased the Pearson correlation from 0.13 to 0.59, confirming that disparities in dataset distribution are pivotal in determining the outcomes of evaluation metrics. This finding not only supports Cao et al. (2022)'s observations about the critical need for dataset alignment in bias measurement but also emphasizes that without aligned datasets, expecting correlated results between different metrics becomes untenable. Our study highlights the essential role of dataset harmonization in achieving reliable bias measurement across metrics, proposing a unified approach to enhance the integrity of bias research in NLP.

5 Conclusions

In this focused study, we examined how different perspectives of two gender stereotyping datasets can lead to significantly divergent outcomes. The application of gender stereotype components from social psychology to balance these datasets significantly boosted the alignment of the corresponding intrinsic metrics, emphasizing the critical role of dataset composition in bias evaluation. Our findings enrich the overarching discourse on gender bias in language models, underscoring that bias is a complex, multifaceted issue. It necessitates a sophisticated approach to accurately measure and effectively mitigate, highlighting the intricate interplay between dataset construction and bias evaluation.

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

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 Raymond Vogel, Donald M. Broverman, Frank E. Clarkson, and Paul S. Rosenkrantz. 1972. Sex-role stereotypes: A current appraisal. *Journal of Social Issues*.

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.

Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 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* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 120–128. ACM.

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

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.

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.

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.

Wei Guo and Aylin Caliskan. 2021. Detecting emergent intersectional biases: Contextualized word embeddings contain a distribution of human-like biases. In *AIES '21: AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, USA, May 19-21, 2021*, pages 122–133. ACM.

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.

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.

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.

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.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a comprehensive understanding and accurate evaluation of societal biases in pre-trained transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed H. Chi, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models. Technical report.

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.

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 CrowS-Pairs predominantly used names as proxies for gender, differing from StereoSet's approach, which relied on gendered words. To ensure consistency in the representation of gender across both datasets, we adapted the CrowS-Pairs sentences, 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 in CrowS-Pairs, 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	60	11	14	14
Attitudes and Beliefs	29	29	52	37
Roles and Behaviors	15	15	73	19
Physical Characteristics	13	9	11	11

Table 1: Gender Stereotype Components Statistics Overview

Source	**Annotated**	**Post-Labeling**	**Post-Balancing**
StereoSet	225	117	81
Crows-Pairs	184	149	64
Total	409	266	145

Table 2: Dataset Statistics Overview

Model	Pre-Balance		Post-Balance	
	Crows-Pairs	StereoSet	Crows-Pairs	StereoSet
BERT-large Vanilla	60.0	58.1	61.0	58.9
BERT-large CDA Scratch	56.0 ↓4.0	59.0 ↑0.9	59.7 ↓1.4	61.4 ↑2.5
BERT-large CDA Finetuned	54.0 ↓6.0	59.8 ↑1.7	58.4 ↓2.6	60.2 ↑1.3
BERT-large Dropout Scratch	54.0 ↓6.0	53.9 ↓4.3	56.4 ↓4.6	57.5 ↓1.4
BERT-large Dropout Finetuned	57.3 ↓2.7	55.6 ↓2.6	60.3 ↓0.7	58.8 ↓0.1
BERT-large ADELE	60.0	61.5 ↑3.42	60.7 ↓0.3	57.8 ↓1.1
BERT-base Vanilla	61.3	65.0	66.2	67.2
BERT-base CDA Finetuned	54.7 ↓6.7	60.7 ↓4.3	55.4 ↓10.8	59.9 ↓7.3
BERT-base Dropout Finetuned	56.0 ↓5.3	59.0 ↓6.0	61.1 ↓5.0	62.7 ↓4.6
BERT-base Orthogonal Projection	55.3 ↓6.0	53.0 ↓12.0	57.0 ↓9.2	55.0 ↓12.2
BERT-base ADELE	56.67 ↓4.7	67.5 ↑2.6	58.822 ↓7.3	66.2 ↓1.0
RoBERTa-base Vanilla	56.7	65.0	58.7	62.6
RoBERTa-base CDA Finetuned	55.3 ↓1.3	59.0 ↓6.0	55.5 ↓3.2	61.0 ↓1.6
RoBERTa-base Dropout Finetuned	57.3 ↑0.7	64.1 ↓0.9	60.9 ↑2.2	65.2 ↑2.5
RoBERTa-base Orthogonal Projection	56.7	59.8 ↓5.1	59.6 ↑0.9	62.2 ↓0.4
RoBERTa-base ADELE	56.7	65.0	58.7 ↑2.0	62.5 ↓0.1
ALBERT-large Vanilla	55.3	61.5	56.4	64.6
ALBERT-large CDA Scratch	54.7 ↓0.7	52.1 ↓9.4	54.7 ↓1.7	54.5 ↓10.1
ALBERT-large Dropout Scratch	50.0 ↓5.3	59.8 ↓1.7	52.9 ↓3.5	54.7 ↓9.9

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.