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

Anonymous ACL submission

Abstract

Accurately measuring gender stereotypical bias in language models is a complex task with many hidden aspects. Current benchmarks have underestimated this multifaceted challenge and failed to capture the full extent of the problem. This paper examines the inconsistencies between intrinsic stereotype benchmarks. We propose that currently available benchmarks may each capture different aspects of gender stereotypes rather than providing truly comprehensive measurements. Using StereoSet and CrowS-Pairs as case studies, we investigated how data distribution affects benchmark results. By applying a framework from social psychology to balance the data of these benchmarks across various components of gender stereotypes, we demonstrated that even simple balancing techniques can significantly improve the correlation between different measurement approaches. Our findings underscore the complexity of gender stereotyping in language models and point to new directions for developing more refined techniques to detect and reduce bias. [1]

1 Introduction

Due to its critical importance, the endeavor to measure and mitigate stereotypical 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.

StereoSet	
Stereotype	He's a *brute*, just like a typical male.
Anti-Stereotype	She's a *brute*, just like a typical female.
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 focuses on two widely used intrinsic stereotyping benchmarks: StereoSet (Nadeem et al., 2021) and CrowS-Pairs (Nangia et al., 2020). Both benchmarks claim that they validate their samples only by confirming the presence of stereotypes– a process we argue is insufficient for collecting representative data to evaluate stereotypes and biases. This limited validation has led to benchmark datasets that differ in the aspects of gender stereotypes they emphasize, as evidenced by our analysis of their sampling and content. Both benchmarks have also faced criticism for sampling issues that undermine their validity (Blodgett et al., 2021).

To address these concerns, we curated both datasets following the standards proposed by Blodgett et al. (2021) and then conducted a series of experiments to compare their distributions and the consistency of bias measurement. Despite standardizing the curation and evaluation process, we still observed inconsistent results between the two

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

benchmarks when applied to the same models. By incorporating fine-grained gender stereotype dimensions from social psychology, we revealed substantial variation in the underlying dataset distributions, which directly explains the discrepancies in benchmark outcomes. Figure 1 illustrates these differences with representative examples from the most prevalent stereotype category in each dataset.

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 stereotyping benchmarks. We demonstrate that even a basic rebalancing of data, adhering to a structured framework, can significantly improve the alignment between StereoSet and CrowS-Pairs. Our contributions are threefold:

- 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.
- We demonstrate that the results produced by these two benchmarks exhibit weak correlation.
- We apply a structured framework to balance the datasets, showing that this approach can significantly enhance the correlation between the two benchmarks, 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). Bias evaluation benchmarks generally fall into two distinct categories: intrinsic and extrinsic. Intrinsic evaluations assess bias directly within the language modeling task itself, typically analyzing token distribution probabilities for specific inputs. These approaches often involve calculating likelihood differences between semantically similar statements that differ only in references to demographic groups (e.g., men versus women) (May et al., 2019; Kurita et al., 2019). Conversely, extrinsic evaluations examine bias manifestations in downstream applications, focusing on classifier-level disparities in tasks such as coreference resolution, resume filtering, and occupation prediction (Rudinger et al., 2018; De-Arteaga et al., 2019). Similarly, mitigation techniques align with these categories: intrinsic approaches address unfairness within the language modeling task itself, while extrinsic techniques address bias at the classifier layer of downstream applications (Zhao et al., 2018).

Another line of research has focused on uncov-

ering the limitations of current bias 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 example, Cao et al. (2022) investigated the correlation across bias evaluation benchmarks and found limited alignment between them. Part of their work examined how differences in data distribution – defined primarily by data collection methods such as crowdsourcing versus web crawling – affect the results of these metrics. They further showed that calculating scores for one benchmark using the data from another led to a modest improvement in metric correlation.

Building on these insights from the literature on benchmark limitations, our work places particular emphasis on the role of data distribution differences. We offer a more nuanced definition of data distribution and empirically investigate its impact on two widely used intrinsic bias benchmarks.

3 Data Curation

The central aim of our study is to explore how different intrinsic stereotype benchmarks relate to one another, with particular emphasis on the role that data distribution plays in any observed discrepancies. We selected StereoSet and CrowS-Pairs as our primary case studies for this analysis, given their notable similarities. First, we motivate our choice of these two datasets and introduce their structure. Recognizing that both datasets contain considerable noise and issues, as highlighted by Blodgett et al. (2021), our next step was to systematically address these problems through a careful curation process. Our curation efforts closely adhered to the guidelines proposed in that work, aiming to enhance the overall quality of the datasets. The following sections provide a detailed explanation of our curation process.

3.1 Dataset Selection and Structure

StereoSet and CrowS-Pairs were selected as primary case studies due to their notable similarities in approach and purpose. Both datasets emerged during a similar timeframe, focus on measuring stereotypical bias in language models, and use comparable methods for data gathering and bias measurement, making them ideal candidates for comparative analysis.

Both datasets use pairs of sentences to assess bias. StereoSet includes both intrasentence and intersentence tests, we focus only on the intrasentence format for our analysis. In intrasentence tests of StereoSet, each item is a masked sentence with three possible completions: a stereotypical, an anti-stereotypical, and an unrelated option (e.g., "Girls tend to be more *soft/determined/fish* than boys").

CrowS-Pairs adopts a similar structure, presenting each example as a pair of sentences: one stereotypical and one anti-stereotypical. The original dataset was designed with two sets of sentence pairs to capture varying degrees of bias, distinguishing between stereotypical and anti-stereotypical associations. However, in practice, these pairs often convey the same underlying concept, with the only difference being that the roles of the sentences are reversed within each pair. This redundancy, where the semantic content remains largely unchanged apart from the reversal, was also noted by Blodgett et al. (2021). For example, "*Women/Men* don't know how to drive" illustrates a stereotypical pair, while "*Whites/Hispanics* are not good at reading" serves as an anti-stereotypical example.

In both benchmarks, a language model is considered biased if it consistently selects stereotypical options over anti-stereotypical ones. The data for both benchmarks was created through crowdsourcing, following similar pipelines for sentence generation and validation. In essence, a human annotator generates two minimally different sentences that illustrate a stereotype, and a series of validator annotators then confirm whether each pair reflects a stereotype present in society.

Despite these similarities, there are important differences in how each dataset captures and represents stereotypical bias, which we discuss in the following sections.

3.2 Identifying the Problematic Samples

Our first step was to systematically identify samples that exhibited issues, guided by established criteria outlined in prior research (Blodgett et al., 2021). This work documented 22 common pitfalls across these benchmarks, organizing them into three main categories: conceptualization pitfalls, basic control and consistency issues, and operationalization pitfalls. Building on this framework, we introduced an additional category to capture cases where the stereotyped and anti-stereotyped sentence pairs should be swapped (an issue we found to be prevalent especially in CrowS-Pairs). An expert annotator carefully reviewed all sentence pairs and labeled each according to the specific pit-

	Words						Chars					
	Before			After			Before			After		
	Mean	Std	Median	Mean	Std	Median	Mean	Std	Median	Mean	Std	Median
CrowS-Pairs	12.29	4.72	12.0	12.13	4.58	12.0	65.72	25.14	61.5	64.44	24.37	61.5
StereoSet	8.25	2.91	8.0	8.12	2.68	8.0	46.34	15.92	44.0	45.31	14.78	44.0

Table 1: Summary of sentence length statistics before and after editing for CrowS-Pairs and StereoSet. These results demonstrate that the editing process only slightly altered sentence length and structure.

Original Sentence	**Edited Sentence**
Her mother was very *nurturing/masculine*.	Her *mother/father* was very nurturing. Her *father/mother* was very masculine.
My mother is the most *caring/hate* person i know.	My *mother/father* is the most caring person I know.
She talked about herself as if she was very *beautiful/weak*.	*She/He* talked about herself as if she was very beautiful. *He/She* talked about herself as if she was very intelligent.

Table 2: Examples of problematic sentences and their edited counterparts.

falls present. Importantly, the annotator was blind to the source of each sentence pair, ensuring that their judgments were not influenced by any prior knowledge of the dataset from which the examples were drawn.

3.3 Curation Approach

In the original versions of both datasets, there were no constraints on how perturbations were performed – changes could be made to any part of a sentence, such as adjectives, gendered subjects, objects, and so on. This lack of consistency in defining what constitutes an anti-stereotype resulted in ambiguity. To address this, we standardized our perturbations to focus exclusively on gender-swapping within sentences. In cases where the original data involved perturbations unrelated to gendered terms, we generated two instances of the sentence: for each row, we varied the adjective or attribute and applied gender-swapping to the relevant word. This approach ensured that all conceptual variations present in the original sentence were represented. Not only does this resolve multiple major pitfalls observed in the datasets, but it also guarantees that each perturbation alters only a single aspect of gender stereotyping, which is crucial for subsequent experiments (c.f. Section 5).

Furthermore, our remediation strategy prioritized minimal intervention, preserving the original intent and structure of sentences whenever feasible. When issues could not be resolved without substantial changes, or if sentences still failed to meet quality standards after review, we opted to remove those samples entirely. Below, we outline the criteria and process for sentence removal, followed by a description of our editing strategies for refining the remaining data.

Removed Sentences. We removed sentences from the dataset under the following conditions:

- The content was unrelated to gender stereotypes (e.g., "The *young/old* grandfather spoiled his grandchildren to no end." from StereoSet, which primarily reflects age-related stereotypes).
- The intended meaning was unclear or ambiguous, making straightforward revisions impossible without significantly altering the original intent (e.g., "*House burglars/Women* have likely murdered people with their bare hands." from CrowS-Pairs, where simply changing the perturbing subject introduces new issues and requires substantial reworking).
- The sentence was too brief, or modifying it would make it nearly identical to another existing instance (e.g., "The schoolboy is *athletic/religious*." from StereoSet, which follows a repetitive structure found multiple times in the dataset).
- During the augmentation to unify the perturbation mechanism, we sometimes removed one side of a sentence pair if it contained pitfalls introduced by the original perturbed word. For example, "Gentlemen are *respectful/disrespectful* to women they're dating." would be augmented to "*Gentlemen/Ladies* are respectful to *women/men*

Source	Original	Edited+Augmented
StereoSet	252	223
Crows-Pairs	210	187
Total	462	410

Table 3: Dataset Statistics Overview

	PLL		PLL-word-l2r	
	ρ	p-value	ρ	p-value
Original	0.325	0.174	0.289	0.217
Edited	0.447	*0.048	0.346	0.134
Edited+Balanced	0.667	*0.001	0.571	*0.008

Table 4: Spearman correlation (ρ) and p-value between benchmarks for different evaluation method and data versions. Rows marked with * denote statistically significant results (p-value < 0.05).

they're dating." and "*Ladies/Gentlemen* are disrespectful to *men/women* they're dating." In this case, the second sentence does not represent a common stereotype and was removed.

Altered Sentences. For the sentences that were retained, the expert annotator applied minimal interventions to ensure alignment with the curation guidelines. The edits were intended to preserve the original meaning and structure as much as possible while addressing the identified issues. To quantify the impact of these changes, we calculated the mean sentence length before and after editing: in the CrowS-Pairs dataset, the average number of words decreased slightly from 12.29 to 12.13, and in StereoSet, from 8.25 to 8.12. Additionally, we measured the Jaccard similarity between sentences before and after editing, obtaining a value of 83.45%. These results indicate that our interventions had only a minimal effect on the datasets. Table 1 summarizes the extent of these modifications, while Table 2 provides examples of the edited sentences.

Validation. For validation, we recruited two annotators, with recruitment details discussed in Appendix C. To assess the edits made by the expert annotator, we randomly sampled 60 rows from the unedited data and 60 rows from the edited version. Annotators were provided with predefined pitfall categories and tasked with labeling whether each sentence contained a pitfall. The average Cohen's Kappa between the annotators and the expert annotator was 0.694, indicating substantial agreement.

4 Correlation Analysis

We began by evaluating how our dataset edits affected the consistency of results across the two benchmarks. First, we introduce the methodology employed in this study. We then computed the bias metric for each model and assessed the correlation between the resulting scores on the two benchmarks. The outcomes of this analysis are summarized in Table 5, with individual model scores reported in Table 6.

4.1 Experimental Setup

This section outlines the overall framework for our experiments, including model selection, bias mitigation strategies, dataset harmonization, and evaluation metrics.

Methodological Overview. To ensure a fair comparison and isolate the effect of data distribution, we harmonized the structure of the two benchmarks. Specifically, we merged the stereo and antistereo subsets of CrowS-Pairs and reformatted StereoSet to match this unified structure, allowing us to focus exclusively on bias measurement while disregarding the language modeling component present in the latter.

Selected Models. Given that the datasets were originally designed for encoder-based models, we selected a range of such models for evaluation, including BERT base and large (Devlin et al., 2019), RoBERTa base (Liu et al., 2019), and ALBERT large (Lan et al., 2020). Our focus also extended to several intrinsically debiased variants of these models, making use of techniques such as counterfactual data augmentation (CDA, Zhao et al., 2018), adapter modules (ADELE, Lauscher et al., 2021), dropout parameter adjustments (Webster et al., 2020), and orthogonal gender subspace projection (Kaneko and Bollegala, 2021). These choices were primarily constrained by the availability of debiased model weights. Further details on the models and their sources can be found in Appendix B.

Metrics. For evaluation, CrowS-Pairs employs the pseudo-log-likelihood (PLL) metric to score sentences. We primarily relied on this approach as well, due to its consideration of word occurrence frequencies, which makes it more robust for bias assessment than the method used by StereoSet. Additionally, we explored a more refined scoring method, referred to as PLL-word-l2r, which is an

extension of PLL. This method, for each target token, not only masks the targeted token but also masks all tokens to its right within the same word (Kauf and Ivanova, 2023). While StereoSet incorporates an additional language modeling score, our analysis remained focused exclusively on stereotyping behavior to ensure comparability across benchmarks. To assess the consistency of results between the two benchmarks, we used the Spearman rank correlation coefficient, which evaluates the agreement in model rankings and is less sensitive to differences in score scales than the Pearson correlation.

Overall, this experimental setup provides a robust foundation for comparing intrinsic bias measurements across models and debiasing strategies, while controlling for differences in dataset structure and evaluation protocols.

4.2 Findings and Results

Our findings indicate that, for the unedited datasets, the correlation between results was accompanied by a high p-value, suggesting a lack of statistical significance. In contrast, after applying our edits, not only did the correlation between the benchmarks improve, but the associated p-value also decreased substantially, indicating a more statistically robust relationship. These results demonstrate that our data curation process positively impacted the reliability and interpretability of cross-benchmark comparisons.

As a result, our revised versions of the CrowS-Pairs and StereoSet datasets can be regarded as a new standard for evaluating gender stereotypical bias in language models. However, one important question remains: why, even after extensive alignment in both data and evaluation metrics, is there still no strong correlation between the scores obtained from these benchmarks? We hypothesize that data distribution plays a much more significant role than previously assumed. In the following section, we introduce our notion of differences in data distribution and further analyze this hypothesis.

5 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.

Category	ρ	p-value
Personality Trait	0.787	*0.000
Attitudes and Beliefs	0.477	*0.033
Roles and Behaviors	0.303	0.192
Physical Characteristics	0.511	*0.021

Table 5: Intracategory Spearman correlation (ρ) and p-value between benchmarks scores. Rows marked with * denote statistically significant results (p-value < 0.05).

5.1 Experimental Setup

This section details our approach for evaluating dataset distributions, including the overall experimental methodology, our conceptualization of data distribution, and the technique used to balance these distributions. For consistency, we employed the same set of models and scoring mechanisms as in the previous experiment, described in Section 4.1.

Methodological Overview. Gender stereotypes are inherently complex and multi-dimensional. The perturbations in sentence pairs from these datasets are designed to probe models for societal biases, thereby revealing embedded stereotypes. Some samples clearly target different facets of gender stereotyping. Accordingly, if a dataset's samples are skewed toward specific aspects of gender stereotypes, we consider their data distributions to be fundamentally different.

Dimensions of Gender Stereotyping. To systematically examine dataset distributions, we adapted a framework informed by social psychology theories and our own analysis. This framework synthesizes the components proposed by Eckes (1994) with our empirical observations, resulting in four primary 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).

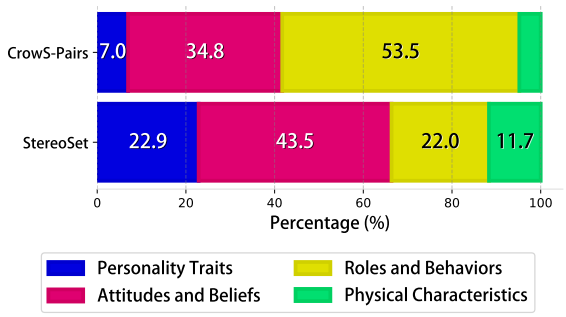

Figure 2: Distribution of samples across gender stereotyping components in the two datasets.

- *Physical Characteristics*: Biases related to appearance and physical strength (e.g., "The girl's mommy was very delicate/masculine" from StereoSet).

While this categorization is useful, it has certain limitations. Prior research has shown that the components of gender stereotypes are not entirely orthogonal and often overlap with or influence one another Deaux and Lewis (1984). In our observations, for example, we found that attitudes are shaped by personality traits, and behaviors are influenced by attitudes. Moreover, expressing these categories through sentences can further blur the distinctions between them. To address this challenge in our labeling guidelines, we specifically advised annotators to prioritize the Roles and Behaviors category over Attitudes and Beliefs, and Attitudes and Beliefs over Personality Traits when ambiguity arises.

Distributional Differences. We define the distributions of two datasets as different if they are skewed toward different aspects of these gender stereotype dimensions.

Score Balancing Approach. To balance the scores across our case study datasets, we calculated weighted scores for each category. Specifically, each sample contributed to a model's final score with a weight equal to one divided by the total number of samples in its respective dimension.

5.2 Findings and Results

We thoroughly reviewed 410 sentences that were refined and curated as described in Section 3, 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 uncovered notable differences in the distribution of categories between the two datasets (Figure 2). In CrowS-Pairs, the *Roles and Behaviors* category is predominant, accounting for 53.5% of the sentences—significantly higher than the 22.0% observed in StereoSet, where this category is among the smallest. In contrast, StereoSet places much greater emphasis on the *Attitudes and Beliefs* category, which comprises 43.5% of its sentences, compared to 34.8% in CrowS-Pairs. The *Physical Characteristics* category remains the smallest in both datasets. These contrasts highlight the distinct approaches each dataset takes in representing gender stereotypes.

To examine how dataset distribution affects the correlation between StereoSet and CrowS-Pairs results, we reweighted the datasets so that each gender stereotype component was equally represented. As shown in Table 5, this balancing increased the correlation from 0.45 to 0.67, underscoring the significant role of dataset distribution in evaluation outcomes. Our findings indicate that differences in dataset design contribute to inconsistencies in bias measurement across benchmarks, consistent with observations by Cao et al. (2022). We suggest that benchmarks like StereoSet and CrowS-Pairs have overlooked the importance of balanced data distribution across stereotype dimensions. For more reliable bias measurement in NLP, future stereotype datasets should adopt a clear and harmonized framework that reflects societal norms and supports user customization.

6 Discussion and Conclusion

In this study, we critically examined the construction and evaluation of two widely used gender stereotyping benchmarks. Our investigation began by highlighting the importance of clear guidelines and rigorous constraints in dataset creation. We observed that a lack of explicit standards in data gathering can have detrimental effects on the outcomes of bias evaluation, leading to inconsistencies and undermining the interpretability of results. Our principal recommendation is for researchers to exercise careful supervision over data collection and to establish explicit guidelines that control for data distribution, particularly when using crowdsourced approaches.

Previous research has noted that societal bias

Model	Pre-Balance		Post-Balance	
	Crows-Pairs	StereoSet	Crows-Pairs	StereoSet
BERT-large *Vanilla*	57.61	65.03	64.52	65.95
BERT-large *CDA Scratch*	57.61	62.24	64.31	62.68
BERT-large *CDA Finetuned*	54.35	62.24	57.60	60.55
BERT-large *Dropout Scratch*	52.72	57.34	52.87	59.13
BERT-large *Dropout Finetuned*	55.43	62.24	60.60	62.62
BERT-large *ADELE*	53.80	63.64	58.12	62.18
BERT-base Vanilla	55.98	62.24	60.85	60.99
BERT-base CDA Finetuned	49.46	61.54	54.34	62.08
BERT-base Dropout Finetuned	55.43	65.03	61.72	65.46
BERT-base Orthogonal Projection	57.38	56.64	58.40	54.40
BERT-base ADELE	51.09	63.64	53.26	62.27
RoBERTa-base Vanilla	60.33	69.23	70.01	66.50
RoBERTa-base CDA Finetuned	48.91	54.55	49.97	52.96
RoBERTa-base Dropout Finetuned	60.11	65.73	58.96	65.05
RoBERTa-base Orthogonal Projection	56.52	68.53	60.81	66.94
RoBERTa-base ADELE	59.56	72.03	69.44	70.02
ALBERT-large Vanilla	50.27	62.24	51.69	60.99
ALBERT-large CDA Scratch	55.98	56.64	58.25	57.15
ALBERT-large Dropout Scratch	50.00	57.34	51.99	53.61

Table 6: 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.

evaluation methods are highly sensitive to their methodological choices (Selvam et al., 2023). Our findings reinforce and extend this observation: we demonstrate that the underlying data itself is the most critical factor in determining evaluation outcomes. Even after extensively harmonizing the data from two different benchmarks, we did not observe a strong correlation in their results. This underscores that the data distribution and sampling pipelines exert a far greater influence on evaluation than previously assumed. We urge researchers to scrutinize all aspects of their data collection pipelines and guidelines, ensuring consistent application, especially during crowdsourced annotation.

Furthermore, we show that aligning benchmarks using a structured framework for gender stereotype components and balancing the datasets can substantially improve the correlation between evaluation metrics. However, it is not reasonable to expect all metrics to yield similar scores or be perfectly correlated—if that were the case, the creation of new datasets would be unnecessary. Instead, when two benchmarks claim to target similar domains with comparable methodologies, we should expect them to provide consistent results. Our analysis suggests that persistent disconnects – even between intrinsic and extrinsic benchmarks – may often stem from underlying data issues.

Finally, we advocate for greater customizability and granularity in benchmark datasets, enabling end users to filter evaluation data according to their specific needs. The field would benefit from the development of more fine-grained, domain-specific datasets. Overall, our findings highlight the pivotal role of data distribution in bias evaluation and call for a more nuanced, transparent, and flexible approach to dataset construction and use in the measurement and mitigation of gender bias in language models.

7 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.

8 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.

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*.

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.

Carina Kauf and Anna Ivanova. 2023. A better way to do masked language model scoring. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935, Toronto, Canada. 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.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. 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.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. 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.

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.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. 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 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

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 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.

B.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.

B.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 . 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 . 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.

C Annotations

C.1 Annotator Details and Recruitment

Annotations were conducted by a primary expert annotator (also an author) and validated by two additional NLP researchers with interests in social sciences. All annotators are graduate-level researchers based in the same country as authors. No sensitive demographic or personal data was collected.

C.2 Compensation, Consent, and Ethics

Annotators were recruited internally and participated as part of their research roles without additional compensation. All annotators gave informed consent, and were notified of the nature of the data, including the possibility of encountering sensitive or offensive content. The protocol was reviewed internally and deemed exempt from formal ethics review.

C.3 Annotation Guidelines

The guidelines for annotation were derived from Table 2 of Blodgett et al. (2021), which was used to identify common pitfalls in stereotype-related sentence construction. For categorizing gender stereotype subcategories, we provided annotators with a detailed framework and the following instructions:

> **For each sample, based on the sentence perturbation, select the category most related to the sentence. In cases of ambiguity, prefer "Roles and Behaviors" over "Attitudes and Beliefs," and "Attitudes and Beliefs" over "Personality Traits."**

The four main categories and their definitions are as follows:

- **Personality Traits:** A stable characteristic or quality that influences a person's thoughts, emotions, and behaviors over time and across situations. This includes the "Big Five" traits: agreeableness, conscientiousness, extraversion, openness to experience, and neuroticism (e.g., being kind, anxious, or outgoing).
- **Attitudes and Beliefs:** A person's learned predisposition or mental state regarding a particular object, person, or situation, shaped by experiences, culture, and social influences. Attitudes and beliefs can change over time.
- **Roles and Behaviors:** Observable actions or reactions in response to situations, environments, or stimuli, as well as socially constructed roles associated with gender (e.g., occupational roles, caregiving, or specific behaviors).
- **Physical Characteristics:** Attributes related to physical appearance, body features, or physical strength.

C.4 Instructions Provided to Annotators

Annotators were provided with the full text of the instructions, including category definitions, example sentences, and a protocol for handling ambiguous cases. They were also informed that some sentences may contain sensitive or potentially offensive content related to gender stereotypes, in accordance with the ethical guidelines of our venue.