When Debiasing Backfires: Counterintuitive Side Effects of Preprocessing-Based Stereotype Mitigation

Yahan Zheng
Dartmouth College
yahan.zheng.gr@dartmouth.edu

John J. Guerrerio
Dartmouth College
john.j.guerrerio.26@dartmouth.edu

Soroush Vosoughi
Dartmouth College
soroush.vosoughi@dartmouth.edu

Weicheng Ma
Oakland University
weichengma@oakland.edu

Abstract

Preprocessing-based methods for stereotype mitigation, such as pre-/post-training on debiased corpora, are widely used in NLP. While these approaches reduce measurable stereotypes for targeted groups, we find they often induce unintended shifts—*side effects*, where stereotyping or counter-stereotyping can *increase* relative to neutral baselines for other demographics, including across unrelated demographic categories. We demonstrate these side effects across two model families (encoder-only and decoder-only), multiple preprocessing strategies (removing stereotypical sentences, removing group mentions, and swapping group references), and both pre- and post-training at different data scales on Wikipedia. Standard benchmarks frequently miss these shifts. Using attention-rollout analysis, we observe that such side effects are not accompanied by large changes in attention flow, complicating mechanistic explanations. We discuss implications for evaluation, provide actionable diagnostics, and argue for side-effect-aware, transparent mitigation practices.

1 Introduction

Pre-trained language models (PLMs) encode and propagate social stereotypes (Bolukbasi et al., 2016; Caliskan et al., 2017; Zhao et al., 2019), raising concerns about safe and responsible deployment. Among mitigation strategies, *preprocessing*-based methods seek to modify training data, for instance, by removing or altering stereotypical content, because they are simple to apply and impose no inference-time cost (Gallegos et al., 2024). Intuitively, such interventions should hypothetically prevent models from learning unwanted associations by isolating them from stereotypical patterns. For consistency, we use the term *stereotype* to refer to any unwanted social bias encoded by language models from their training data, and *mitigation* or *debiasing* to refer to interventions aimed at reducing these encoded stereotypes.

Despite the intuitive appeal and partial effectiveness of preprocessing-based methods, to date, no PLM has achieved complete freedom from stereotypes, raising questions about their true efficacy. More importantly, it is unclear whether data-level mitigation *eliminates* harmful associations or merely *redistributes* them.

We revisit this question empirically by curating debiased Wikipedia corpora targeting six demographic groups spanning three categories (gender, race, religion). We use these corpora for both pre- and post-training of two PLMs (encoder-only TinyBERT (Jiao et al., 2020) and decoder-only GPT-2 (Radford et al., 2019)) and then measure changes in stereotype expression toward *all* groups. We report three findings:

1. **Unintended shifts within and across categories.** While stereotyping toward the group *targeted* for debiasing decreases, we frequently observe increased stereotyping or counter-stereotyping for *other* groups (including across bias categories). We define this phenomenon as a *side effect*: a case where a mitigation method decreases stereotype measures for the target group but simultaneously induces undesired changes for one or more non-target groups. Notably, these shifts are often asymmetric and hard to anticipate. We quantify trends on StereoSet and CrowS-Pairs (Nadeem et al., 2021; Nangia et al., 2020). These patterns cannot be explained solely by changes in the distribution of stereotypical/anti-stereotypical training examples.
2. **Robustness across settings.** Side effects appear under three preprocessing strategies (removing stereotypical sentences, removing group mentions, swapping references), during both pre- and post-training, *and across training data scales*, for encoder-only (TinyBERT) and

Findings of the Association for Computational Linguistics: ACL 2026, pages 10001–10021

decoder-only (GPT-2) models. A larger decoder-only model exhibits the same qualitative phenomenon on a single post-training slice.

3. **Mechanism remains elusive.** Attention-rollout analysis (Abnar and Zuidema, 2020) shows small attribution shifts even when stereotype scores move, suggesting that attention routing alone does not explain the effect and motivating distributional and causal follow-ups.

Overall, our results highlight a reliability gap for preprocessing-based mitigation: interventions that help the target group can unpredictably harm others. These unintended consequences persist across diverse experimental conditions, including multiple preprocessing strategies, reduced training data scales, and different PLM architectures. In addition, standard evaluation benchmarks frequently fail to surface these side effects, and they are largely undetectable by inspecting models' internal attention patterns. Taken together, this *calls into question the reliability and safety of data-level debiasing when used in isolation*. We recommend reporting side-effect–aware diagnostics (see Sections 4–5) and emphasize the urgent need for *more robust, interpretable, and controllable* mitigation methods capable of reducing social biases in language models. Our code for this paper has been publicly released at `https://github.com/InDaCS-Lab/Stereotype-Mitigation-Side-Effects`.

2 Background

Stereotype encoding and mitigation have been studied extensively in computational linguistics. Early work such as Bolukbasi et al. (2016) demonstrated gender bias in Word2Vec embeddings trained on the Google News corpus. Shortly thereafter, Caliskan et al. (2017) showed that GloVe embeddings likewise inherit human-like biases. Similar studies demonstrated these biases extended to contextualized models: Zhao et al. (2019) quantified and mitigated gender bias in ELMo's contextual word vectors. As larger, less interpretable Transformer-based models like BERT, GPT-2, and RoBERTa became dominant, the community shifted toward using specialized stereotype evaluation datasets, such as StereoSet (Nadeem et al., 2021), WinoGender (Zhao et al., 2018), and CrowS-Pairs (Nangia et al., 2020), to benchmark bias in commonly-used language models. For instance, Nadeem et al. (2021) empirically demonstrated BERT and GPT-2 have stereotypical tendencies, a finding validated by Nangia et al. (2020) for BERT. With the rise of Large Language Models (LLMs), prompt-based benchmarks such as Esiobu et al. (2023), Dhamala et al. (2021), and Akyürek et al. (2022) have been employed to examine biases in modern models.

The phenomenon of encoded stereotypes has prompted the creation of targeted mitigation strategies. As defined by Gallegos et al. (2024), these mitigation approaches can be divided into 4 broad categories.

Pre-Processing Mitigation involves changes to the training data to prevent the model from learning stereotypes. One common form of preprocessing-based stereotype mitigation is data augmentation. Lu et al. (2020) first formalized this approach to mitigate gender bias by creating pairs of semantically invariant sentences with flipped gendered words (e.g., "he" to "she"). Ghanbarzadeh et al. (2023) extended this approach by masking gendered words and using a language model to predict a replacement. Another form of preprocessing-based stereotype mitigation approach is dataset filtering, which focuses on identifying examples to either emphasize or exclude. Garimella et al. (2022) and Borchers et al. (2022) identify underrepresented or low-bias examples to focus on during post-training. Raffel et al. (2020) uses a word list to filter out biased examples, an approach refined by Ngo et al. (2021) and Sattigeri et al. (2022) with more advanced filtering techniques. Panda et al. (2022) identifies demographic identifying words and removes them prior to post-training. Notably, data augmentation and data filtering can be combined; Zayed et al. (2023) generates counterfactual examples for data that contribute the most to fairness and filters out other stereotypical examples.

In-Training Mitigation incorporates changes to the training procedure or additional post-training steps. For instance, Guo et al. (2022), Yang et al. (2023), and Gaci et al. (2022) introduce new loss functions to mitigate the biases the model potentially learns. Gira et al. (2022) fine-tune a very small subset of model parameters on the WinoBias and CrowS-Pairs datasets. Ouyang et al. (2022) employs a reinforcement learning-based post-training approach with human feedback to better align LLMs with human values.

Intra-Processing Mitigation entails modifications to the model's inference behavior. Works such as Ma et al. (2023) and Zhou et al. (2024) identify biased model components (e.g., attention

heads) and disable them at inference time. Tong et al. (2024) uses smaller stereotypical and anti-stereotypical expert models to re-balance next token probabilities toward anti-stereotypical tokens and away from stereotypical ones. Similarly, Liu et al. (2023) learns small, tunable bias vectors at inference time to shift the model's logits away from toxic tokens. Finally, Saunders et al. (2022) employs beam search to find more diverse model outputs at inference time.

Post-Processing Mitigation encompasses modifications to the model's output text generation. Tokpo and Calders (2022) frames debiasing as a style transfer problem, and uses LIME to identify biased keywords to be replaced via style transfer to a neutral domain. Dhingra et al. (2023) employ Shapley values to identify biased words in model output and re-prompt the LLM to rephrase the given sentence without those words.

Preprocessing-based methods offer several important advantages over other stereotype mitigation approaches. As described in Gallegos et al. (2024), these approaches only modify the input of the model during training. This allows debiasing of models without introducing additional constraints to the training process or requiring additional compute at inference time. However, they also have significant limitations. Many data augmentation techniques swap terms using word lists, which can be incomplete and change the semantic meaning of a sentence. Gallegos et al. (2024) argues this limitation is especially salient for words describing social groups. Assuming the interchangeability of social groups ignores the fact that stereotypes are nuanced and specific to each group. Similarly, removing or replacing identity words does not eliminate the harm within a stereotypical statement, but only redirects it toward a potentially irrelevant group.

Similar limitations apply to data filtering techniques. Incomplete and misrepresentative word lists can lead to the removal of minority voices while leaving behind harmful documents. Such techniques can also introduce distributional imbalances into the training data, exacerbating bias.

Our work presents a comprehensive investigation into the limitations of these methods, while also characterizing the side effects they introduce.

3 Experimental Settings

We investigate the encoded stereotypes of TinyBERT (4-layer, 14M parameters) and GPT-2 (124M parameters) to examine potential side effects of preprocessing-based stereotype mitigation methods. TinyBERT is an encoder-only transformer trained with a masked language modeling (MLM) objective, whereas GPT-2 is a decoder-only transformer trained with a causal language modeling (CLM) objective. Using two architecturally distinct yet compact PLMs provides diversity in the models we study (strengthening the generalizability of our findings) while keeping multiple rounds of pre-/post-training and evaluation computationally feasible.

All pre-/post-training uses the June 1, 2023 English Wikipedia snapshot, tokenized and filtered to prose articles. For TinyBERT, this domain matches the original training data used for the released model. For evaluation, we employ the intra-sentence portion of StereoSet and CrowS-Pairs. In addition to reporting overall scores, we report group-specific stereotype metrics for gender- (female and male), race- (Black and Caucasian), and religion- (Muslim and Christian) based demographic groups. This allows us to capture more fine-grained effects of stereotype mitigation.

We consider three commonly used data-level interventions. Note, we conduct each intervention solely for *one* of the six demographic groups or three demographic categories we consider (referred to as the *target group* or *target category* respectively). After completing an intervention, we measure stereotyping toward all demographic groups.

1. **Debias-A-Group (DG):** remove sentences flagged as *stereotypical* toward the target group using a pretrained sentence-level stereotype detector.[1] On CrowS-Pairs, this detector attains F1=0.98 (precision=0.98; recall=0.98).
2. **Remove-A-Group (RG):** remove *all* sentences mentioning the target group, regardless of sentiment or potential stereotypes.
3. **Swap-References (SR):** replace all identity mentions within the target category. For gender stereotypes, we apply direct antonym swaps (e.g., "female" $\leftrightarrow$ "male"). For racial and religious stereotypes, where clear antonyms do not exist, we define group mappings and generate replacements using a constrained LLM prompt

[1] https://huggingface.co/wu981526092/Sentence-Level-Stereotype-Detector

that enforces grammaticality, number, and syntactic role preservation. Three NLP experts annotated these samples, confirming that 96 out of 100 instances were of satisfactory quality (receiving approval from at least two annotators).

We run the full grid of the three strategies $\times$ two training stages (pre-/post-) $\times$ two data scales (100% and 5%), yielding 12 settings. This setup enables us to compare the effects of data scale and training stage on the expression of stereotypes.

Models are implemented with HuggingFace.[2] All runs use a single RTX 2080 Ti GPU, and the random seed is set to 42.

Evaluation protocol. We report overall StereoSet stereotype scores (SS; 50 is neutral; farther from 50 indicates stronger stereotyping/anti-stereotyping, which is undesirable), language modeling score (LMS), and iCAT (which balances SS and LMS), alongside demographic group-level SS. For CrowS-Pairs we report per-demographic category scores following the original work. Because absolute SS is known to be sensitive to instance composition, we emphasize *directional changes* relative to the corresponding base model in matched conditions. Where applicable, we additionally analyze attribution via attention-rollout (Abnar and Zuidema, 2020).

4 Data Debiasing Reduces Stereotypes

We first evaluate the extent to which preprocessing-based methods reduce stereotype scores for the targeted groups. We evaluate the effects of three preprocessing strategies (DG, RG, SR) under both pre-training and post-training of TinyBERT and GPT-2 on the full Wikipedia snapshot and report StereoSet's stereotype score (SS), language modeling score (LMS), and iCAT together with per-group SS, following Nadeem et al. (2021). We also report CrowS-Pairs results for completeness (Appendix A.2).

Pre-training. As shown in Fig. 1, SS for the *target* group typically decreases relative to the baseline models, moving toward 50 for both TinyBERT and GPT-2, with some limited exceptions (e.g., GPT-2 under SR-religion shows increased SS for Muslims). Similar trends are observed in the CrowS-Pairs benchmark (Appendix A.2). Aggregating across groups, overall SS often decreases while LMS remains comparable, yielding iCAT that is similar to or higher than the base models.

[2] https://huggingface.co

Post-training. Post-training generally yields larger reductions in target-group SS than pre-training (Fig. 2), again with broadly stable LMS, yielding higher iCAT scores overall. Detailed StereoSet and CrowS-Pairs results are provided in Appendices A.1.2 and A.2.2.

Data scale. Using 5% of Wikipedia, both pre-training and post-training still consistently reduce SS for targeted groups. These improvements occur alongside stable or slightly enhanced LMS and iCAT scores. These results are shown in Appendix A.

Taken together, these results show that preprocessing-based debiasing can reduce stereotype scores for targeted groups in a data-efficient manner. However, as we show in the next section, these improvements often coincide with unintended side-effects in non-target groups.

5 Data Debiasing Incurs Side Effects

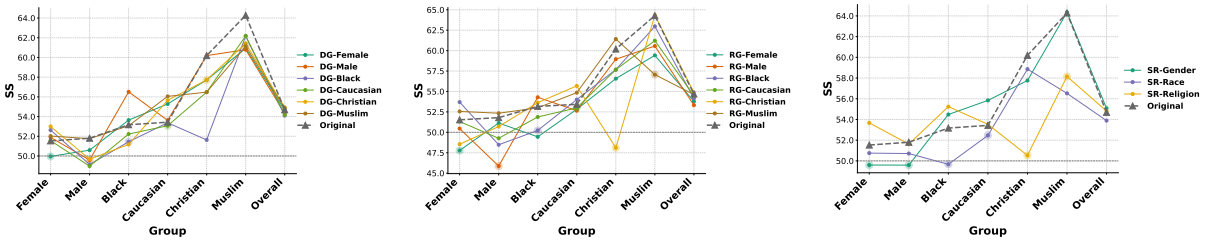
Alongside the targeted improvements documented in Section 4, we observe *side effects*: stereotype scores (SS) for *non-target* groups sometimes move *away from 50*. These shifts vary by preprocessing strategy and training stage (pre- vs. post-training), and they can cross stereotype categories. We therefore emphasize directional changes relative to the corresponding base model and highlight representative patterns below.

5.1 Stereotype Shifts Within Categories

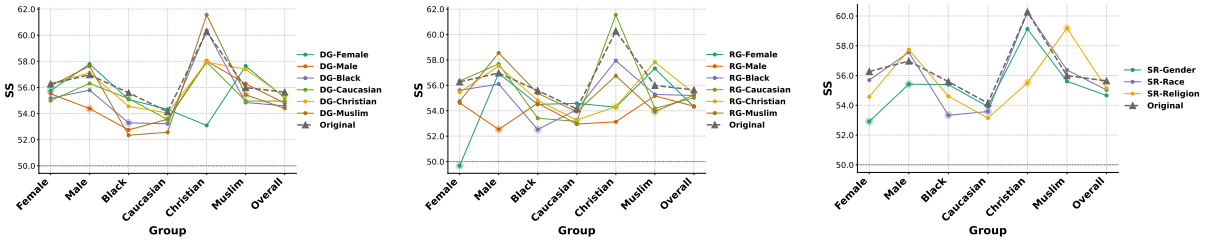
For each category (gender, race, religion), we evaluate two demographic groups to enable within-category comparisons. While debiasing a given group often moves its SS toward 50, the *other* group in the same category does not behave uniformly.

For example, pre-training TinyBERT on the full Wikipedia corpus debiased for females (DG-female) reduces the model's stereotypes toward males (green line in Figure 1a), whereas debiasing for males (DG-male) increases stereotypes toward females (red line in Figure 1a). Post-training shows analogous behavior: post-training on DG-male increases stereotypes toward females (red line in Figure 2a), and post-training on RG-female increases stereotypes toward males (green line in Figure 2b).

Interestingly, these spillovers depend on the debiasing method, which complicates a simple distributional explanation. TinyBERT pre- and post-trained on DG-male both show increased stereotypes to-

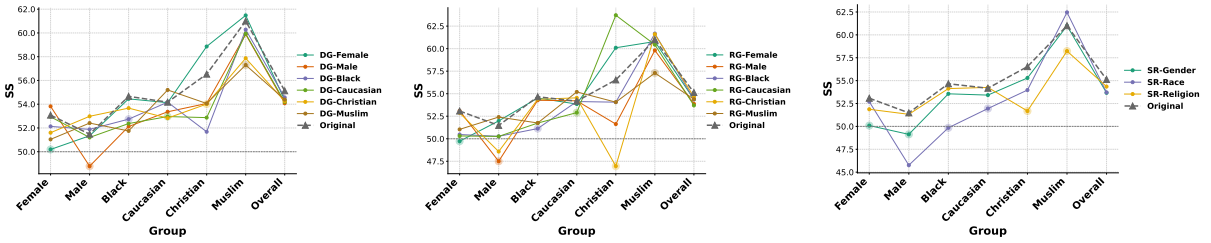


(a) TinyBERT-DG (b) TinyBERT-RG (c) TinyBERT-SR

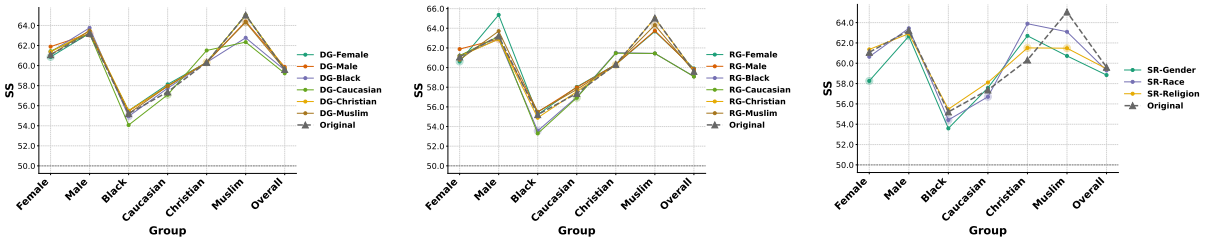


(d) GPT-2-DG (e) GPT-2-RG (f) GPT-2-SR

Figure 1: StereoSet stereotype scores (SS; 50 is neutral, closer to 50 indicates *less* stereotyping/anti-stereotyping) for models pre-trained on debiased Wikipedia under three preprocessing strategies. Dashed lines mark the original model; Larger semi-transparent markers indicate the target group used for data cleaning. Full numeric results appear in Appendix Tables A1-A3 and A7-A9.



(a) TinyBERT-DG (b) TinyBERT-RG (c) TinyBERT-SR



(d) GPT-2-DG (e) GPT-2-RG (f) GPT-2-SR

Figure 2: StereoSet stereotype (SS) for models *post-trained* on debiased Wikipedia. Full numeric results appear in Appendix Tables A10-A12 and A16-A18.

ward females (Figures 1a and 2a), whereas models trained on RG-male exhibit inconsistent patterns across pre- versus post-training (Figures 1b and 2b).

Unintended effects also extend beyond directly opposing groups. For instance, pre-training on RG-Muslim increases stereotypes toward Christians (Figure 1b), even though the two groups are not strict opposites in the StereoSet framework.

Taken together, preprocessing-based debiasing induces non-uniform and hard-to-predict shifts. The observed changes are not readily explained by altered data distributions and appear non-systematic across groups and training regimes. This raises a practical concern: *removing stereotypes documented in existing benchmarks may not mitigate bias, but instead redirect it toward other groups, including groups not represented in those benchmarks.*

5.2 Stereotype Variability Across Categories

Cross-category side effects are often more pronounced than within-category shifts and can reverse across training stages. For instance, the TinyBERT model pre-trained on the preprocessed

Wikipedia corpus under RG-female setting shows a consistent reduction in stereotype levels across all 6 groups evaluated in our study (Figure 1b), whereas post-training on the same RG–female corpus increases Christian SS (Figure 2b).

These cross-category side effects are not symmetric either. Comparing 2 TinyBERT models pre-trained on the full Wikipedia corpus cleaned under RG-female and RG-Christian settings, we observe that the former displays elevated stereotype levels toward Christians, whereas the latter actually reduces stereotypes toward females (Figure 1b). We observed similar trends when manually analyzing the models' word preferences before and after stereotype-mitigation training. Figure 4 provides a concrete example.

As shown in Figure 4, pre-training TinyBERT on DG-cleaned data for the Caucasian group shifts the model's behavior from exhibiting mild stereotypical tendencies to predominantly generating anti-stereotypical completions. For the sentence "That man went to the store to buy [BLANK]," the model prefers "food" over "alcohol," countering the stereotype that "men are heavy drinkers." However, this effect is not symmetric. When TinyBERT is instead pre-trained on DG-male data, it becomes highly stereotypical when presented with the sentence "Todd was white and [BLANK]," assigning a much higher probability to "racist" than to neutral alternatives such as "inclusive."

These effects are accompanied by substantial shifts in token-level probabilities, as shown in Figure 4. The lack of a consistent relationship between data interventions and behavioral outcomes, together with the observed asymmetry, highlights the unpredictability of stereotype-mitigation side effects. This unpredictability complicates the assessment of mitigation effectiveness and poses significant challenges for the safe and reliable deployment of PLMs.

Even more notably, the stereotype that "men are heavy drinkers." does not appear in StereoSet, the benchmark used to guide the debiasing of the Wikipedia corpus. This raises additional concerns about preprocessing-based debiasing approaches, as they may inadvertently introduce or amplify stereotypes that are not covered by existing benchmarks. As a result, it becomes difficult or impossible to fully assess the potential harms introduced by such debiasing, given that the evaluation framework lacks coverage of these emergent or untracked biases.

These side effects happen to models trained using other preprocessing settings as well. For example, the model pre-trained on DG-male shows lower-than-usual stereotype levels toward Christians while pre-training the model on DG-Christian leads to much higher stereotype levels toward males (Figure 1a). This observation counters the potential explanation that side effects are results of the stereotype/anti-stereotype content distributions being affected by text removal, since if stereotypical content toward males overlaps with stereotypical content toward Christians, DG for Christians should also reduce the model's stereotype levels on the male group. Hence, these results cannot be fully explained by changes in stereotype/anti-stereotype content distributions alone, suggesting that additional mechanisms may contribute to the observed side effects. Importantly, such side effects can undermine the effectiveness and reliability of preprocessing-based stereotype mitigation approaches by introducing unintended shifts in non-target groups.

The stereotype evaluation results of models pre- or post-trained on the SR-cleaned corpora further consolidate the randomness of such side effects. Theoretically, switching the references to a pair of groups would by no means affect the ratio of stereotypical contents toward other groups in a corpus. Yet, we noted that the TinyBERT model pre-trained on the SR-cleaned corpus for racial stereotypes (affecting black people and Caucasian people) shows a higher stereotype level toward Christians and a higher anti-stereotype level toward males (Figure 1c). Leveraging the same data, the model post-trained on the SR-cleaned data for racial stereotypes shows harsher stereotypes toward Muslims while a lower stereotype level toward Christians (Figure 2c). Due to the existence of random side effects with high frequency in our experiments, it raises concerns about reliability the effectiveness and robustness of preprocessing-based stereotype mitigation approaches.

5.3 Model Agnosticity of Side Effects

The experimental results on GPT-2 demonstrate that the side effects of preprocessing-based stereotype mitigation extend beyond small encoder-only models trained with the MLM objective. As shown in Figure 1d, pre-training GPT-2 on DG-female data amplifies stereotypes toward males, whereas DG-male consistently reduces SS for both gender groups. Unexpectedly, both DG-male and DG-

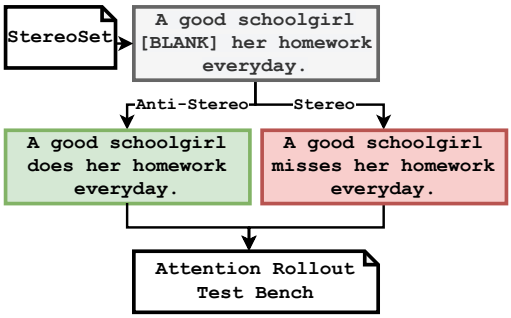


Figure 3: Curation process for the attention-rollout test bench derived from StereoSet instances.

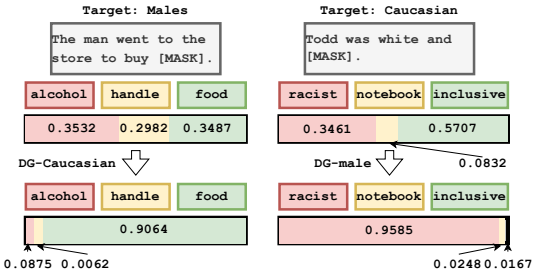


Figure 4: Example of cross-group stereotype shifts induced by preprocessing-based mitigation. Pretraining TinyBERT on DG-Caucasian data leads to anti-stereotypical behavior toward males (e.g., preferring "food" over "alcohol"), while DG-male induces stronger stereotypical behavior toward Caucasians (e.g., preferring "racist" over "inclusive"). These asymmetric shifts, reflected in token-level probabilities, highlight the unpredictability of side effects across groups.

female increase stereotypes toward Muslims. Removing female mentions entirely (RG-female) minimizes negative impacts on male-oriented stereotypes but continues to strengthen Muslim-oriented stereotypes (Figure 1e). Surprisingly, switching male and female references (SR-Gender) mitigates stereotypes across all 6 groups, resulting in overall lower stereotype levels on StereoSet (Figure 1f). Similar unpredictable and non-systematic stereotype shifts appear in other models, indicating that stereotype expression is a confounding issue across different types of LLMs. Although smaller in magnitude, these side effects persist in GPT-2 post-training experiments (Figures 2d–2f).

These findings confirm that the unintended consequences of preprocessing-based stereotype mitigation are not confined to small encoder models. Rather, they may undermine the reliability of debiasing across a wide range of PLMs, highlighting the need for more robust evaluation frameworks and mitigation strategies that explicitly account for cross-group side effects.

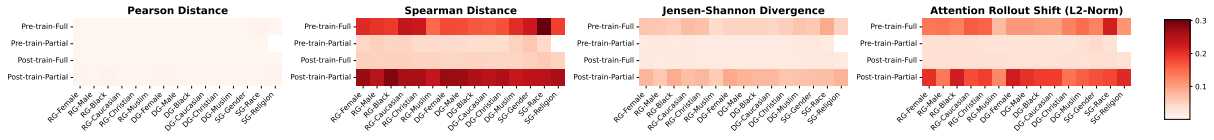
6 Side Effects Beyond Semantics

Removing content from LLM training corpora can alter models' semantic representations due to attention allocation changes. To assess whether semantic shifts might still contribute, we compute attention rollout (Abnar and Zuidema, 2020) for each model on StereoSet inputs and compare rollout values between stereotype-mitigated models and their corresponding base models.

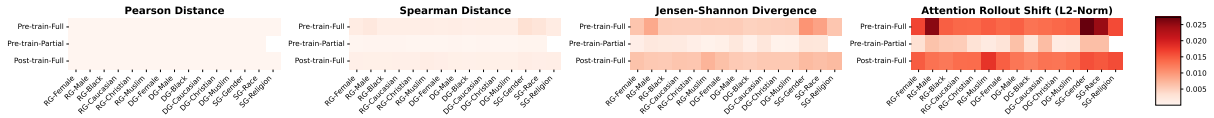
Attention rollout estimates how information from original input tokens contributes to representations at a chosen layer. In contrast to raw, per-layer attention, which shows only same-layer interactions among already-mixed embeddings, rollout aggregates heads, accounts for residual connections, and multiplies attention matrices across layers to propagate the influence. The resulting joint map provides a row-normalized distribution approximating relative influence along all attention paths. In our setting, rollout identifies the input tokens that most strongly affect internal representations and enables direct, cross-model comparisons of attribution patterns to reveal how debiasing affects attention.

We use StereoSet data to ensure that all test text directly concerns the minority groups under study, avoiding dilution from unrelated content. For data preparation, each StereoSet instance is converted into 2 sentences (one stereotypical and one anti-stereotypical), as illustrated in Figure 3. For completeness, we compare each stereotype-mitigated model with its base counterpart using Pearson distance (PD), Spearman distance (SD), Jensen–Shannon divergence (JSD), and the L2-norm of attention shift (AS) on rollout distributions. Among these metrics, PD and SD capture distributional correlation differences, AS quantifies absolute changes in attention allocation, and JSD measures divergence between normalized attention distributions (Lu et al., 2021; Hu et al., 2023). Results are shown as heatmaps in Figure 5.

Overall, the attention rollout patterns remain highly consistent between stereotype-mitigated models and their base counterparts, with very low maximum distances across all metrics (PD: 0.0061; SD: 0.3029; JSD: 0.0925; AS: 0.2235). This consistency holds at both the distribution level (PD, SD, JSD) and the magnitude level (AS), and it is robust across model sizes, from smaller architectures such



(a) TinyBERT



(b) GPT-2

Figure 5: Distances/divergences between attention-rollout distributions of stereotype-mitigated models and their base counterparts, plus L2 norms of rollout shifts (darker indicates larger differences). All differences shown are statistically significant ($p < 0.001$).

as TinyBERT to larger ones like GPT-2.

Finer-grained, group-specific analyses reach the same conclusion, even for models that show pronounced side effects on stereotype metrics. For example, the TinyBERT model pre-trained under the DG-male setting increases stereotype scores for women and, more strongly, for Black people (Figure 1a), yet rollout shifts remain small (PD: 0.001279, SD: 0.162548, JSD: 0.050337, AS: 0.103921 for women; PD: 0.001300, SD: 0.157040, JSD: 0.049205, AS: 0.101581 for Black people). Similarly, GPT-2 pre-trained under the DG-female setting lowers stereotype scores for Christians and increases them for Muslims, while rollout shifts remain comparably small (PD: 0.000020, SD: 0.001176, JSD: 0.004793, AS: 0.013177 for Christian; PD: 0.000020, SD: 0.001015, JSD: 0.004882, AS: 0.012909 for Muslim).

Taken together, these results suggest that the observed side effects of stereotype mitigation are not well explained by changes in how models semantically route information, as captured by attention rollout. While we note that our representation-level analysis only provides surface-level insights into each model, it serves as a useful diagnostic probe to demonstrate the unclear origin of the side effects we observe.

Their underlying mechanism remains uncertain, motivating methods that probe alternative causal pathways and a reconsideration of stereotype benchmarking practices to ensure faithful, side-effect-aware safety evaluation of LLMs. We note that identifying the precise layers, cues, or circuits responsible for stereotype redistribution is beyond the scope of this work.

7 Discussion

We additionally test the robustness of our findings across models of different training data sizes (Section 7.1), and for massive LLMs (Section 7.2).

7.1 Data Size Agnosticity of Side Effects

According to our experimental results, the DG, RG, and SR approaches using a tiny subset of the Wikipedia corpus (5%) still lead to less biased models for the target groups (Appendix A). This further highlights the sensitivity of models to preprocessing-based stereotype mitigation approaches.

Alongside the benefits, the side effects of stereotype mitigation still exist and are unpredictable. For example, stereotypes on the Caucasian group become more pronounced when the TinyBERT model is post-trained on the SR-cleaned data for gender stereotypes (Table A6). Gender stereotypes toward both groups get worse in the model post-trained on the SR-cleaned data for racial stereotypes (Table A15) while they are reduced in the pre-trained model on the same data (Table A6).

From a finer-grained lens, a higher level of stereotypes is observed toward male people when all the content related to black people is removed from the pre-training data of the TinyBERT model (Table A5), despite the existence of many shared stereotypes between the 2 groups in stereotype benchmarks, e.g., being aggressive and athletic. This does not occur when the same data is used to post-train the original TinyBERT model (Table A14), further suggesting that side effects are not consistent across training stages. We similarly observe unexpected side effects from models trained with DG-based stereotype mitigation. For example, pre-training the TinyBERT model using the DG-cleaned data for the female group leads to lower

Category	LLaMa2	LLaMa2-RG-F
Female SS	70.72	66.30
Male SS	65.48	65.01
Black SS	64.14	**65.43**
Caucasian SS	67.70	65.12
Christian SS	69.75	**71.03**
Muslim SS	59.50	55.50
Overall SS	64.63	62.97
Overall LMS	90.74	89.71
Overall iCAT	64.19	66.44

Table 1: StereoSet results for LLaMA2 and its variant post-trained on the RG-cleaned Wikipedia corpus (female group). Underlining marks SS moving *toward* 50 (less stereotyping/anti-stereotyping); bold marks SS moving *away* from 50. Targeted-group improvements coincide with unintended shifts for non-targeted groups.

stereotype levels for all the groups except for males, and pre-training on the DG-cleaned data for black people raises the stereotype level of the model on males and Christians (Table A4). Post-training the TinyBERT model using the DG-cleaned data for the female group, surprisingly, causes severer stereotypes toward Christians and Muslims, 2 uncorrelated groups with the female group (Table A13).

These observations reinforce our claim that the side effects of preprocessing-based stereotype mitigation can emerge across settings and are not consistently predictable from the intervention alone.

7.2 Side Effects Persist in Massive Models

Since TinyBERT and GPT-2 are relatively small models, we ask whether preprocessing-based debiasing methods remain effective in larger language models, or whether the observed side effects persist at scale. To investigate this, we conducted an additional experiment with LLaMa-2-7B (LLaMa2), a substantially larger transformer-based model than GPT-2 (117M parameters) and TinyBERT (14.5M parameters). Given the substantial computational cost, we limited training to the RG-female setting (LLaMa2-RG-F) and evaluated the model on StereoSet (Table 1).

The results indicate that stereotypes remain prevalent in massive models. In particular, female-oriented stereotypes were strongest in the base model (SS=70.72), motivating our focus on the female group. Post-training on RG-female data reduced the stereotype score toward females to 66.30 and produced slight reductions for the male, Caucasian, and Muslim groups. However, the same training unexpectedly increased stereotypes toward the black and Christian groups, though these groups are not directly related to the intervention.

These findings demonstrate that the side effects of stereotype mitigation persist even in massive LLMs. This underscores the need for continued, systematic efforts to evaluate, interpret, and refine stereotype mitigation strategies to ensure that scaling up models does not simply shift or amplify biases in unpredictable ways.

8 Conclusion & Future Work

Stereotype mitigation is central to the safe deployment of PLMs. Preprocessing-based approaches are appealing because they remove stereotypical content before it can be learned, yet our experiments show that removing material targeting a single group from pre- or post-training data can induce unintended cross-group spillovers: stereotypes toward other, sometimes unrelated, groups may increase. These effects vary across categories, corpora, and debiasing recipes and are difficult to predict, which undermines the reliability of preprocessing as a primary mitigation strategy. Although preprocessing can improve targeted metrics, unmonitored side effects can erode reliability; data debiasing should therefore be treated as an intervention whose side effects are monitored alongside benefits. Future work includes (i) moving beyond attention-based attribution toward causal and distributional analyses of how interventions reshape co-occurrences and inductive biases, and (ii) expanding evaluation beyond StereoSet/CrowS-Pairs to additional identity axes and languages.

Ethics Statement

We examine the side effects of data-level debiasing without releasing any post-trained models. We report only aggregate statistics to reduce risk. We curated corpora from the June 1, 2023, English Wikipedia snapshot and evaluated them on publicly released benchmarks (StereoSet; CrowS-Pairs). We used constrained LLM prompts for swapping references and removed generations containing slurs or identity-targeted insults. Because debiasing can redistribute harm, we recommend documenting both targeted and non-targeted groups in model cards and auditing side effects before deployment.

Regarding the use of generative AI (GenAI) tools, we acknowledge the use of these tools for assistance in improving the clarity and readability of the manuscript text. The use of these tools was limited to language refinement; no GenAI models were used for ideation, experimental design, or any other substantive aspects of the research.

Limitations

We focus on the English Wikipedia and six identity groups. The effects may differ for other identities, domains, and languages. Most experiments use compact models (TinyBERT, GPT-2) for control and tractability, and we include only a single post-training slice for a larger decoder-only model. We leave broader coverage to future work. We fix the random seed to 42 for reproducibility and do not report multi-seed uncertainty for SS/CrowS-Pairs, so claims are calibrated to directional trends rather than significance tests. Finally, our DG detector, although high-recall, may include false positives, and SR swaps may introduce subtle distributional artifacts, despite manual spot checks.

We acknowledge these as important directions for future work and encourage follow-up studies to build on our findings with broader model coverage and more diverse representation.

Acknowledgments

We are grateful to the Templeton Foundation and Georgia State University for supporting this work.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197.

Afra Feyza Akyürek, Sejin Paik, Muhammed Kocyigit, Seda Akbiyik, Serife Leman Runyun, and Derry Wijaya. 2022. On measuring social biases in prompt-based multi-task learning. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 551–564, Seattle, United States. Association for Computational Linguistics.

Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam T. Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, volume 29.

Conrad Borchers, Dalia Sara Gala, Benjamin Gilburt, Eduard Oravkin, Wilfried Bounsi, Yuki M. Asano, and Hannah Rose Kirk. 2022. Looking for a handsome carpenter! debiasing gpt-3 job advertisements. *arXiv preprint arXiv:2205.11374*.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

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 *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 862–872.

Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. 2023. Queer people are people first: Deconstructing sexual identity stereotypes in large language models. *arXiv preprint arXiv:2307.00101*.

David Esiobu, Xiaoqing Tan, Saghar Hosseini, Megan Ung, Yuchen Zhang, Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, and Eric Michael Smith. 2023. Robbie: Robust bias evaluation of large generative language models. *arXiv preprint arXiv:2311.18140*.

Yacine Gaci, Boualem Benatallah, Fabio Casati, and Khalid Benabdeslem. 2022. Debiasing pretrained text encoders by paying attention to paying attention. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9582–9602. 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.

Aparna Garimella, Rada Mihalcea, and Akhash Amarnath. 2022. Demographic-aware language model fine-tuning as a bias mitigation technique. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 311–319.

Somayeh Ghanbarzadeh, Yan Huang, Hamid Palangi, Radames Cruz Moreno, and Hamed Khanpour. 2023. Gender-tuning: Empowering fine-tuning for debiasing pre-trained language models. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5448–5458, Toronto, Canada. Association for Computational Linguistics.

Michael Gira, Ruisu Zhang, and Kangwook Lee. 2022. Debiasing pre-trained language models via efficient fine-tuning. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 59–69, Dublin, Ireland. Association for Computational Linguistics.

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

Lijie Hu, Yixin Liu, Ninghao Liu, Mengdi Huai, Lichao Sun, and Di Wang. 2023. Seat: stable and explainable attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12907–12915.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for natural language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4163–4174, Online. Association for Computational Linguistics.

Xin Liu, Muhammad Khalifa, and Lu Wang. 2023. Bolt: Fast energy-based controlled text generation with tunable biases. *arXiv preprint arXiv:2305.12018*.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. 2020. Gender bias in neural natural language processing. In *Logic, language, and security: essays dedicated to Andre Scedrov on the occasion of his 65th birthday*, pages 189–202.

Yu Lu, Jiali Zeng, Jiajun Zhang, Shuangzhi Wu, and Mu Li. 2021. Attention calibration for transformer in neural machine translation. 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 1288–1298.

Weicheng Ma, Henry Scheible, Brian Wang, Goutham Veeramachaneni, Pratim Chowdhary, Alan Sun, Andrew Koulogeorge, Lili Wang, Diyi Yang, and Soroush Vosoughi. 2023. Deciphering stereotypes in pre-trained language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11328–11345.

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 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.

Helen Ngo, Cooper Raterink, João GM Araújo, Ivan Zhang, Carol Chen, Adrien Morisot, and Nicholas Frosst. 2021. Mitigating harm in language models with conditional-likelihood filtration. *arXiv preprint arXiv:2108.07790*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, and 1 others. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744.

Swetasudha Panda, Ari Kobren, Michael Wick, and Qinlan Shen. 2022. Don't just clean it, proxy clean it: Mitigating bias by proxy in pre-trained models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5073–5085, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Prasanna Sattigeri, Soumya Ghosh, Inkit Padhi, Pierre Dognin, and Kush R. Varshney. 2022. Fair infinitesimal jackknife: Mitigating the influence of biased training data points without refitting. In *Advances in Neural Information Processing Systems*, volume 35, pages 35894–35906.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2022. First the worst: Finding better gender translations during beam search. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3814–3823, Dublin, Ireland. Association for Computational Linguistics.

Ewoenam Kwaku Tokpo and Toon Calders. 2022. Text style transfer for bias mitigation using masked language modeling. *arXiv preprint arXiv:2201.08643*.

Schrasing Tong, Eliott Zemour, Rawisara Lohanimit, and Lalana Kagal. 2024. Towards resource efficient and interpretable bias mitigation in natural language generation. In *NeurIPS Safe Generative AI Workshop*.

Ke Yang, Charles Yu, Yi R. Fung, Manling Li, and Heng Ji. 2023. Adept: A debiasing prompt framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 10780–10788.

Abdelrahman Zayed, Prasanna Parthasarathi, Gonçalo Mordido, Hamid Palangi, Samira Shabanian, and Sarath Chandar. 2023. Deep learning on a healthy data diet: Finding important examples for fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14593–14601.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*.

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.

Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. 2024. Unibias: Unveiling and mitigating llm bias through internal attention and ffn manipulation. *arXiv preprint arXiv:2405.20612*.

A Stereotype Evaluation Results

This section details the stereotype evaluation results of all the models in our experiments.

A.1 StereoSet

A.1.1 Pre-trained Models

Tables A1–A3 present the StereoSet evaluation results for TinyBERT models pre-trained on the full Wikipedia corpus debiased under the DG, RG, and SR settings, respectively. Results for TinyBERT models pre-trained on a subset of the corpus (5%) are shown in Tables A4–A6.

Similarly, the StereoSet evaluation results for GPT-2 models pre-trained on the full preprocessed corpus are reported in Tables A7–A9. Since GPT-2 training failed to converge on the 5% subset of Wikipedia, we omit results for that setting.

A.1.2 Post-trained Models

Tables A10–A12 report the StereoSet evaluation results for TinyBERT models post-trained on the fully preprocessed Wikipedia corpus, while Tables A13–A15 show results for models post-trained on the 5% subset. For GPT-2, StereoSet evaluation results on the fully preprocessed corpus are presented in Tables A16–A18, and results on the 5% subset are provided in Tables A19–A21.

A.2 CrowS-Pairs

A.2.1 Pre-trained Models

We show the CrowS-Pairs evaluation results of TinyBERT models pre-trained on the full preprocessed Wikipedia corpus (Tables A22 - A24) and on 5% preprocessed data (Tables A25 - A27) Likewise, Tables A28–A30 present the CrowS-Pairs evaluation results for GPT-2 models pre-trained on the fully preprocessed Wikipedia corpus. Since GPT-2 pre-training did not converge on the 5% subset, we omit CrowS-Pairs evaluation for those models.

A.2.2 Post-trained Models

Tables A31–A33 present the CrowS-Pairs stereotype evaluation results for TinyBERT models post-trained on the full preprocessed Wikipedia corpus, while Tables A34–A36 show the corresponding results for TinyBERT models trained on a 5% subset of the same corpus. For GPT-2 models, the evaluation results following post-training on the full preprocessed Wikipedia corpus are reported in Tables A37–A39, and those for models post-trained on the subsampled data are displayed in Tables A40 - A42.

Stereotype Category	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender-female SS	49.96	**51.89**	**52.63**	**51.56**	**53.00**	**52.02**	51.53
gender-male SS	50.59	49.59	49.14	48.97	49.67	51.76	51.80
race-black SS	**53.64**	**56.51**	51.44	52.23	51.17	53.07	53.17
race-caucasian SS	**55.28**	**53.62**	53.37	53.07	**55.65**	**56.05**	53.43
religion-christian SS	57.72	60.19	51.63	56.48	57.72	56.48	60.19
religion-muslim SS	60.92	60.79	62.16	62.19	61.43	61.14	64.28
Overall SS	54.63	54.32	54.10	54.17	54.94	54.83	54.69
Overall LMS	77.39	78.83	78.68	78.91	78.41	77.83	78.69
Overall iCAT Score	70.22	72.02	72.24	72.33	70.66	70.30	71.31

Table A1: StereoSet evaluation results of TinyBERT models pre-trained on DG-cleaned data for each group. SS, LMS, and iCAT score indicate stereotype score, language modeling score, and idealized context association test score, respectively. Group-specific SS in the debiased models that are closer to or farther away from 50 than in the original model are underlined and bolded, respectively. Original refers to the original release of TinyBERT model without stereotype mitigation.

Stereotype Category	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender-female SS	**47.78**	50.46	**53.70**	51.32	**48.55**	**52.56**	51.53
gender-male SS	51.14	**45.87**	48.48	49.27	50.71	**52.34**	51.80
race-black SS	49.45	**54.29**	50.23	51.88	**53.63**	53.00	53.17
race-caucasian SS	52.82	52.65	**54.01**	52.99	**55.68**	**54.85**	53.43
religion-christian SS	56.57	58.95	57.67	57.67	48.10	**61.42**	60.19
religion-muslim SS	59.42	60.56	62.97	61.22	**64.34**	57.06	64.28
Overall SS	53.78	53.32	54.16	54.47	54.88	54.62	54.69
Overall LMS	76.67	77.04	78.15	77.39	77.36	78.24	78.69
Overall iCAT Score	70.88	71.93	71.64	70.48	69.81	71.01	71.31

Table A2: StereoSet evaluation results of TinyBERT models pre-trained on RG-cleaned data for each group.

Category	Gender	Race	Religion	Original
Female SS	49.60	50.75	**53.67**	51.53
Male SS	49.59	50.71	**51.62**	51.80
Black SS	**54.48**	49.67	**55.24**	53.17
Caucasian SS	**55.84**	**52.46**	**53.53**	53.43
Christian SS	**57.76**	**58.86**	50.53	60.19
Muslim SS	**64.35**	56.52	58.15	64.28
Overall SS	55.08	53.89	54.81	54.69
Overall LMS	76.60	77.59	79.01	78.69
Overall iCAT	68.82	71.55	71.41	71.31

Table A3: StereoSet evaluation results of TinyBERT models pre-trained on SR-cleaned data for each stereotype category.

Stereotype Category	Female	Male	Black	Caucasian	Muslim	Christian	Original
gender-female SS	49.25	48.70	51.14	48.38	50.24	48.94	51.73
gender-male SS	50.15	51.76	49.43	51.88	51.83	**52.16**	48.10
race-black SS	50.59	50.93	50.15	50.83	49.98	46.11	55.31
race-caucasian SS	50.62	52.03	52.77	50.11	51.90	50.36	53.84
religion-muslim SS	55.87	56.77	56.11	51.46	52.65	53.31	59.78
religion-christian SS	55.38	54.14	**57.67**	**57.72**	55.25	50.37	56.57
Overall SS	52.36	51.94	52.23	51.76	52.19	51.65	52.89
Overall LMS	64.60	64.51	66.24	64.81	65.17	62.01	63.14
Overall iCAT Score	61.55	61.99	63.23	62.53	62.33	59.96	59.50

Table A4: StereoSet evaluation results of TinyBERT models pre-trained on 5% DG-cleaned data for each group.

Stereotype Category	Female	Male	Black	Caucasian	Muslim	Christian	Original
gender-female SS	**41.41**	49.77	51.00	**48.11**	48.50	49.74	51.73
gender-male SS	**53.97**	**47.25**	49.59	50.48	50.72	50.89	48.10
race-black SS	46.63	49.51	47.92	46.21	50.78	50.43	55.31
race-caucasian SS	51.51	50.19	53.00	49.37	51.63	52.37	53.84
religion-muslim SS	55.43	57.75	54.46	59.33	46.83	58.23	59.78
religion-christian SS	52.91	52.91	**58.86**	49.12	**58.99**	45.77	56.57
Overall SS	51.24	51.20	52.40	51.03	51.57	53.04	52.89
Overall LMS	63.12	63.91	66.31	66.42	65.15	63.35	63.14
Overall iCAT Score	61.55	62.38	63.12	65.05	63.10	59.50	59.50

Table A5: StereoSet evaluation results of TinyBERT models pre-trained on 5% RG-cleaned data for each group.

Category	Gender	Race	Religion	Original
Female SS	**46.55**	48.82	50.24	51.73
Male SS	50.54	**47.53**	48.75	48.10
Black SS	48.77	46.80	48.86	55.31
Caucasian SS	52.06	48.67	50.75	53.84
Christian SS	56.53	**58.95**	48.10	56.57
Muslim SS	56.03	55.46	56.41	59.78
Overall SS	51.39	50.96	51.82	52.89
Overall LMS	64.17	65.60	63.58	63.14
Overall iCAT	62.39	64.34	61.27	59.50

Table A6: StereoSet evaluation results of TinyBERT models pre-trained on 5% SR-cleaned data for each stereotype category.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender-female SS	55.75	55.52	55.18	54.99	**56.29**	56.11	56.27
gender-male SS	**57.79**	54.38	55.79	56.31	**57.12**	**57.64**	56.97
race-black SS	55.08	52.75	53.29	55.14	54.55	52.34	55.59
race-caucasian SS	**54.32**	53.57	53.22	53.55	54.04	52.57	54.14
religion-christian SS	53.09	58.02	**60.36**	57.94	57.94	**61.55**	60.27
religion-muslim SS	**57.64**	**56.28**	54.85	54.96	**57.39**	55.44	55.99
Overall SS	55.21	54.80	54.59	54.95	55.29	54.40	55.64
Overall LMS	82.25	82.74	83.52	82.88	83.13	83.07	83.32
Overall iCAT Score	73.69	74.79	75.85	74.68	74.34	75.75	73.93

Table A7: StereoSet evaluation results of GTP-2 models pre-trained on full DG-cleaned Wikipedia data for each group.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender-female SS	49.67	54.62	55.61	**56.34**	55.46	54.74	56.27
gender-male SS	56.95	52.53	56.10	**57.68**	**57.52**	**58.55**	56.97
race-black SS	54.50	54.60	52.52	53.41	54.82	55.37	55.59
race-caucasian SS	**54.58**	52.95	54.09	53.16	53.27	53.93	54.14
religion-christian SS	54.28	53.13	57.94	**61.55**	54.32	56.75	60.27
religion-muslim SS	**57.33**	55.17	55.28	54.18	**57.83**	53.93	55.99
Overall SS	54.31	54.36	55.20	55.02	55.38	55.24	55.64
Overall LMS	80.81	80.25	83.32	82.94	82.91	82.91	83.32
Overall iCAT Score	73.84	73.25	74.66	74.61	73.99	74.23	73.93

Table A8: StereoSet evaluation results of GPT-2 models pre-trained on full RG-cleaned Wikipedia data for each group.

Category Type	Gender	Race	Religion	Original
gender-female SS	52.91	55.69	54.57	56.27
gender-male SS	55.42	**57.56**	**57.74**	56.97
race-black SS	55.39	53.33	54.61	55.59
race-caucasian SS	53.91	53.58	53.14	54.14
religion-christian SS	59.13	60.27	55.51	60.27
religion-muslim SS	55.61	**56.37**	**59.18**	55.99
Overall SS	54.66	55.03	55.14	55.64
Overall LMS	81.47	82.97	83.55	83.32
Overall iCAT Score	73.87	74.63	74.95	73.93

Table A9: StereoSet evaluation results of GPT-2 models pre-trained on full SR-cleaned Wikipedia data.

Stereotype Category	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender-female SS	50.19	**53.82**	52.13	52.94	51.60	51.04	53.08
gender-male SS	51.34	48.78	**51.88**	51.20	**52.98**	**52.41**	51.48
race-black SS	54.45	52.12	52.74	52.37	53.67	51.75	54.66
race-caucasian SS	54.13	53.36	**54.17**	52.96	52.83	**55.20**	54.17
religion-christian SS	**58.86**	54.06	51.68	52.87	54.01	54.06	56.53
religion-muslim SS	**61.48**	59.89	60.28	59.94	57.87	57.30	61.00
Overall SS	54.34	54.08	54.57	54.19	54.25	54.39	55.13
Overall LMS	78.68	79.25	79.77	80.29	80.03	79.35	79.98
Overall iCAT Score	71.84	72.78	72.49	73.56	73.23	72.37	71.77

Table A10: StereoSet evaluation results of TinyBERT models post-trained on DG-cleaned data for each group.

Stereotype Category	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender-female SS	49.73	**53.10**	50.47	50.31	52.87	51.04	53.08
gender-male SS	**51.98**	**47.50**	50.27	50.26	48.58	**52.41**	51.48
race-black SS	54.52	54.27	51.13	51.74	54.38	51.75	54.66
race-caucasian SS	53.85	**54.21**	54.13	52.90	**54.54**	**55.20**	54.17
religion-christian SS	**60.10**	51.63	54.06	**63.71**	46.96	54.06	56.53
religion-muslim SS	60.79	59.81	**61.67**	60.44	**61.60**	57.30	61.00
Overall SS	54.38	53.88	54.38	53.70	54.65	54.39	55.13
Overall LMS	78.73	79.14	80.24	79.73	79.66	79.35	79.98
Overall iCAT Score	71.84	73.01	73.22	73.83	72.26	72.37	71.77

Table A11: StereoSet evaluation results of TinyBERT models post-trained on RG-cleaned data for each group.

Category	Gender	Race	Religion	Original
Female SS	50.09	52.70	51.87	53.08
Male SS	49.13	**45.76**	51.32	51.48
Black SS	53.56	49.85	54.13	54.66
Caucasian SS	53.42	51.95	**54.26**	54.17
Christian SS	55.29	53.97	51.68	56.53
Muslim SS	60.99	**62.46**	58.23	61.00
Overall SS	53.67	53.71	54.35	55.13
Overall LMS	79.60	79.99	80.11	79.98
Overall iCAT	73.75	74.06	73.14	71.77

Table A12: StereoSet evaluation results of TinyBERT models post-trained on SR-cleaned data.

Stereotype Category	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender-female SS	52.79	52.70	**53.52**	**53.53**	**53.58**	**53.00**	52.93
gender-male SS	**52.07**	49.48	**52.42**	48.37	51.22	**54.58**	51.73
race-black SS	51.95	51.10	52.85	50.56	52.74	51.45	56.12
race-caucasian SS	**56.16**	**55.82**	**56.18**	54.74	54.61	**55.74**	54.95
religion-christian SS	57.67	**58.86**	**60.05**	**58.91**	56.48	56.48	58.82
religion-muslim SS	60.02	57.06	58.66	**60.96**	57.15	60.07	60.09
Overall SS	54.93	54.71	54.82	54.48	54.52	55.05	54.90
Overall LMS	79.60	79.72	79.70	79.85	79.35	79.36	80.22
Overall iCAT Score	71.75	72.21	72.02	72.69	72.17	71.34	72.36

Table A13: StereoSet evaluation results of TinyBERT models post-trained on 5% DG-cleaned data for each group.

Stereotype Category	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender-female SS	52.67	**53.10**	**53.94**	**53.03**	52.20	**54.42**	52.93
gender-male SS	49.52	**48.15**	**51.93**	**53.67**	**54.29**	51.49	51.73
race-black SS	51.94	54.54	52.49	53.08	54.23	51.44	56.12
race-caucasian SS	**55.35**	**55.91**	**55.33**	54.57	**55.84**	**55.12**	54.95
religion-christian SS	**61.24**	56.48	**61.33**	55.29	51.76	**58.91**	58.82
religion-muslim SS	57.37	59.47	56.54	58.14	**62.17**	49.78	60.09
Overall SS	53.92	54.62	55.34	54.87	55.55	54.72	54.90
Overall LMS	78.46	78.94	79.71	79.01	79.35	79.61	80.22
Overall iCAT Score	72.31	71.64	71.19	71.31	70.54	72.09	72.36

Table A14: StereoSet evaluation results for TinyBERT models post-trained on 5% RG-cleaned data for each group.

Category	Gender	Race	Religion	Original
Female SS	52.38	**54.04**	52.34	52.93
Male SS	51.17	**53.50**	**53.86**	51.73
Black SS	53.13	48.72	53.78	56.12
Caucasian SS	**56.24**	54.40	**56.43**	54.95
Christian SS	57.67	**60.05**	55.46	58.82
Muslim SS	59.64	57.59	57.86	60.09
Overall SS	54.77	54.58	55.65	54.90
Overall LMS	78.81	79.27	79.49	80.22
Overall iCAT	71.29	71.99	70.50	72.36

Table A15: StereoSet evaluation results of TinyBERT models post-trained on 5% SR-cleaned data for each stereotype category.

Stereotype Category	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender-female SS	60.86	**61.89**	**61.36**	**61.44**	**61.36**	61.08	61.08
gender-male SS	**63.23**	63.18	**63.76**	63.17	**63.44**	**63.49**	63.21
race-black SS	**55.50**	**55.47**	54.94	54.07	**55.50**	**55.22**	55.21
race-caucasian SS	**58.13**	**57.97**	**57.74**	57.09	**57.94**	**57.95**	57.37
religion-christian SS	60.32	60.32	60.32	**61.51**	60.32	60.32	60.32
religion-muslim SS	65.04	64.39	62.77	62.34	65.04	64.32	65.04
Overall SS	59.60	59.86	59.56	59.23	59.72	59.62	59.61
Overall LMS	90.31	90.37	90.39	90.52	90.44	90.34	90.39
Overall iCAT Score	72.98	72.54	73.11	73.82	72.86	72.96	73.02

Table A16: StereoSet evaluation results of GTP-2 models post-trained on full DG-cleaned Wikipedia corpus for each group.

Stereotype Category	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender-female SS	60.61	**61.88**	**61.21**	**61.21**	**61.15**	60.79	61.08
gender-male SS	**65.36**	62.91	62.90	63.17	62.89	**63.71**	63.21
race-black SS	55.21	**55.47**	53.56	53.29	54.89	**55.50**	55.21
race-caucasian SS	**58.03**	**57.98**	57.00	56.92	**57.55**	**57.77**	57.37
religion-christian SS	60.32	60.32	**61.51**	**61.46**	60.27	60.32	60.32
religion-muslim SS	63.68	63.75	61.44	61.44	65.04	64.32	65.04
Overall SS	59.91	59.85	59.10	59.08	59.64	59.60	59.61
Overall LMS	90.32	90.34	90.55	90.56	90.55	90.41	90.39
Overall iCAT Score	72.41	72.54	74.08	74.10	73.10	73.04	73.02

Table A17: StereoSet evaluation results of GPT-2 models post-trained on full RG-cleaned Wikipedia corpus for each group.

Stereotype Category	Gender	Race	Religion	Original
gender-female SS	58.26	60.64	**61.36**	61.08
gender-male SS	62.66	**63.44**	62.91	63.21
race-black SS	53.59	54.42	**55.48**	55.21
race-caucasian SS	**57.58**	56.70	**58.11**	57.37
religion-christian SS	**62.70**	**63.89**	**61.51**	60.32
religion-muslim SS	60.72	63.10	61.48	65.04
Overall SS	58.84	59.29	59.46	59.61
Overall LMS	89.97	90.45	90.30	90.39
Overall iCAT Score	74.07	73.65	73.22	73.02

Table A18: StereoSet evaluation results of GPT-2 models post-trained on full SR-cleaned Wikipedia corpus.

Stereotype Category	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender-female SS	61.56	**62.37**	62.05	**62.56**	61.31	**62.77**	62.09
gender-male SS	**62.78**	61.99	60.17	**64.81**	61.93	**64.02**	62.41
race-black SS	**54.97**	**56.74**	54.16	53.35	**56.26**	**55.61**	54.94
race-caucasian SS	56.34	56.45	56.91	56.52	**57.58**	**57.40**	57.11
religion-christian SS	**65.17**	61.51	**66.31**	**65.12**	63.89	63.93	65.08
religion-muslim SS	61.83	62.46	**63.85**	**66.12**	60.99	61.92	62.68
Overall SS	59.49	59.84	59.40	59.73	60.01	60.13	59.58
Overall LMS	89.91	89.78	89.92	89.91	89.70	89.87	89.69
Overall iCAT Score	72.84	72.11	73.01	72.42	71.73	71.65	72.50

Table A19: StereoSet evaluation results of GPT-2 models post-trained on 5% DG-cleaned Wikipedia data for each group.

Stereotype Category	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender-female SS	61.82	**62.82**	61.58	**62.80**	**62.34**	**62.83**	62.09
gender-male SS	**64.07**	61.87	**63.60**	**63.25**	**63.26**	**63.10**	62.41
race-black SS	**56.74**	**56.96**	54.71	**55.23**	**55.85**	**55.76**	54.94
race-caucasian SS	56.74	**58.00**	56.33	56.54	**57.62**	**58.15**	57.11
religion-christian SS	63.89	62.70	63.93	62.65	63.93	62.74	65.08
religion-muslim SS	**63.75**	62.66	**66.19**	**66.09**	62.39	61.34	62.68
Overall SS	60.18	60.01	59.81	60.15	59.53	60.33	59.58
Overall LMS	89.41	89.72	90.10	89.86	89.75	89.43	89.69
Overall iCAT Score	71.20	71.76	72.42	71.61	72.64	70.96	72.50

Table A20: StereoSet evaluation results of GPT-2 models post-trained on 5% RG-cleaned Wikipedia data for each group.

Category	Gender	Race	Religion	Origin
Female SS	60.01	**62.87**	62.03	62.09
Male SS	60.09	**64.32**	**62.62**	62.41
Black SS	51.90	54.14	54.06	54.94
Caucasian SS	56.10	56.33	56.06	57.11
Christian SS	**66.31**	61.51	62.74	65.08
Muslim SS	60.54	**63.33**	62.09	62.68
Overall SS	58.61	59.42	59.51	59.58
Overall LMS	88.17	90.09	89.87	89.69
Overall iCAT Score	72.99	73.11	72.77	72.50

Table A21: StereoSet evaluation results of GPT-2 models post-trained on 5% SR-cleaned Wikipedia data for each stereotype category.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	51.53	53.82	53.44	51.53	55.34	52.29	51.15
race	61.55	61.75	59.03	58.83	60.78	61.75	65.24
religion	60.95	60.00	61.90	62.86	62.86	62.86	46.67
overall score	58.01	58.52	58.12	57.74	59.66	58.97	54.35

Table A22: CrowS-Pairs evaluation results of TinyBERT models pre-trained on the full DG-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	54.58	51.53	54.58	48.85	50.76	54.58	51.15
race	51.07	60.19	57.09	64.27	62.91	59.03	65.24
religion	69.52	60.95	74.29	66.67	63.81	47.62	46.67
overall score	58.39	57.56	61.99	59.93	59.16	53.74	54.35

Table A23: CrowS-Pairs evaluation results of TinyBERT models pre-trained on the full RG-cleaned Wikipedia data.

Category Type	Gender	Race	Religion	Original
gender	44.27	52.29	51.53	51.15
race	57.09	49.13	51.46	65.24
religion	63.81	55.24	62.86	46.67
overall score	55.06	52.22	55.28	54.35

Table A24: CrowS-Pairs evaluation results of TinyBERT models pre-trained on the full SR-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	54.58	51.15	44.27	49.62	58.02	51.53	51.15
race	58.06	62.33	67.77	61.75	62.33	58.64	65.24
religion	60.95	53.33	60.00	60.00	57.14	56.19	46.67
overall score	57.86	55.60	57.35	57.12	59.16	55.45	54.35

Table A25: CrowS-Pairs evaluation results of TinyBERT models pre-trained on the 5% DG-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	53.05	51.91	53.05	47.71	53.05	52.67	51.15
race	60.78	59.42	64.85	64.47	64.66	62.52	65.24
religion	55.24	50.48	71.43	54.29	45.71	38.10	46.67
overall score	56.36	53.94	63.11	55.49	54.47	51.10	54.35

Table A26: CrowS-Pairs evaluation results of TinyBERT models pre-trained on the 5% RG-cleaned Wikipedia data.

Category Type	Gender	Race	Religion	Original
gender	46.18	51.15	51.91	51.15
race	61.94	48.35	57.86	65.24
religion	51.43	56.19	69.52	46.67
overall score	53.18	51.90	59.76	54.35

Table A27: CrowS-Pairs evaluation results of TinyBERT models pre-trained on the 5% SR-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	48.09	53.05	53.05	51.53	52.29	52.29	51.15
race	57.36	59.88	56.78	58.33	56.2	56.98	58.53
religion	44.76	43.81	45.71	41.9	45.71	38.1	43.81
overall score	53.65	55.11	54.44	54.58	54.05	53.65	54.31

Table A28: CrowS-Pairs evaluation results of GPT-2 models pre-trained on the full DG-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	45.04	46.56	51.53	52.29	52.29	51.15	51.15
race	57.75	61.24	56.2	57.75	59.5	56.78	58.53
religion	43.81	39.05	44.76	43.81	68.57	36.19	43.81
overall score	52.79	52.85	53.65	54.64	56.96	53.38	54.31

Table A29: CrowS-Pairs evaluation results of GPT-2 models pre-trained on the full RG-cleaned Wikipedia data.

Category Type	Gender	Race	Religion	Original
gender	40.46	53.05	51.53	51.15
race	57.17	60.66	58.91	58.53
religion	47.62	47.62	53.33	43.81
overall score	51.53	55.77	55.44	54.31

Table A30: CrowS-Pairs evaluation results of GPT-2 models pre-trained on the full SR-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	51.15	54.96	50.38	45.04	53.82	51.53	51.15
race	52.43	51.46	52.23	58.83	54.76	57.67	65.24
religion	64.76	64.76	64.76	63.81	60.95	57.14	46.67
overall score	56.11	57.06	55.79	55.89	56.51	55.45	54.35

Table A31: CrowS-Pairs evaluation results of TinyBERT models post-trained on the full DG-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	53.82	53.44	50.38	45.42	52.29	54.58	51.15
race	55.15	61.55	56.50	63.69	59.22	61.36	65.24
religion	64.76	62.86	76.19	60.95	59.05	50.48	46.67
overall score	57.91	59.28	61.02	56.69	56.85	55.47	54.35

Table A32: CrowS-Pairs evaluation results of TinyBERT models post-trained on the full RG-cleaned Wikipedia data.

Category Type	Gender	Race	Religion	Original
gender	43.89	50.00	52.67	51.15
race	50.29	51.46	55.15	65.24
religion	59.05	64.76	63.81	46.67
overall score	51.08	55.41	57.21	54.35

Table A33: CrowS-Pairs evaluation results of TinyBERT models post-trained on the full SR-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	55.34	55.34	53.82	53.82	53.44	53.44	51.15
race	56.31	50.29	58.64	53.40	54.76	55.15	65.24
religion	63.81	65.71	68.57	63.81	62.86	59.05	46.67
overall score	58.49	57.11	60.34	57.01	57.02	55.88	54.35

Table A34: CrowS-Pairs evaluation results of TinyBERT models post-trained on the 5% DG-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	58.02	54.58	56.49	46.95	51.91	55.34	51.15
race	58.83	58.64	54.76	56.50	57.67	55.53	65.24
religion	64.76	68.57	74.29	69.52	65.71	51.43	46.67
overall score	60.54	60.60	61.85	57.66	58.43	54.10	54.35

Table A35: CrowS-Pairs evaluation results of TinyBERT models post-trained on the 5% RG-cleaned Wikipedia data.

Category Type	Gender	Race	Religion	Original
gender	42.75	55.73	56.11	51.15
race	58.64	45.44	54.95	65.24
religion	62.86	60.95	66.67	46.67
overall score	54.75	54.04	59.24	54.35

Table A36: CrowS-Pairs evaluation results of TinyBERT models post-trained on the 5% SR-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	58.78	53.44	58.78	59.92	58.78	59.54	59.16
race	58.14	57.75	58.14	58.53	58.14	57.95	58.14
religion	57.14	57.14	57.14	56.19	57.14	57.14	57.14
overall score	58.69	57.29	58.75	58.82	58.69	58.69	58.69

Table A37: CrowS-Pairs evaluation results of GPT-2 models post-trained on the full DG-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	57.63	53.44	59.92	59.92	59.54	59.16	59.16
race	57.95	58.14	58.53	58.33	59.88	58.14	58.14
religion	57.14	57.14	56.19	56.19	56.19	56.19	57.14
overall score	58.09	57.49	58.82	58.69	59.22	58.75	58.69

Table A38: CrowS-Pairs evaluation results of GPT-2 models post-trained on the full RG-cleaned Wikipedia data.

Category Type	Gender	Race	Religion	Original
gender	55.73	57.63	59.16	59.16
race	59.88	59.3	58.14	58.14
religion	55.24	57.14	57.14	57.14
overall score	58.16	58.82	58.75	58.69

Table A39: CrowS-Pairs evaluation results of GPT-2 models post-trained on the full SR-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	54.20	57.63	56.87	54.58	58.02	53.82	55.34
race	59.50	67.05	61.05	63.57	63.37	62.60	62.02
religion	54.29	55.24	56.19	55.24	59.05	53.33	56.19
overall score	58.09	61.87	59.02	59.95	60.48	59.28	59.28

Table A40: CrowS-Pairs evaluation results of GPT-2 models post-trained on the 5% DG-cleaned Wikipedia data.

Category Type	Female	Male	Black	Caucasian	Christian	Muslim	Original
gender	45.80	57.63	52.67	53.05	56.49	53.05	55.34
race	60.27	65.89	60.66	62.21	62.79	61.05	62.02
religion	56.19	56.19	57.14	55.24	59.05	54.29	56.19
overall score	57.56	61.41	58.16	59.02	60.08	58.62	59.28

Table A41: CrowS-Pairs evaluation results of GPT-2 models post-trained on the 5% RG-cleaned Wikipedia data.

Category Type	Gender	Race	Religion	Original
gender	51.15	51.91	53.44	55.34
race-color	62.60	63.57	61.24	62.02
religion	53.33	56.19	55.24	56.19
overall score	58.22	59.08	58.36	59.28

Table A42: CrowS-Pairs evaluation results of GPT-2 models post-trained on the 5% SR-cleaned Wikipedia data.