Upstream Mitigation Is *Not* All You Need: Testing the Bias Transfer Hypothesis in Pre-Trained Language Models

Ryan Steed
Carnegie Mellon University
ryansteed@cmu.edu

Swetasudha Panda, Ari Kobren, Michael Wick
Oracle Labs
{swetasudha.panda,ari.kobren,michael.wick}@oracle.com

A few large, homogenous, pre-trained models undergird many machine learning systems — and often, these models contain harmful stereotypes learned from the internet. We investigate the *bias transfer hypothesis*: the theory that social biases (such as stereotypes) internalized by large language models during pre-training transfer into harmful task-specific behavior after fine-tuning. For two classification tasks, we find that reducing intrinsic bias with controlled interventions *before* fine-tuning does little to mitigate the classifier's discriminatory behavior *after* fine-tuning. Regression analysis suggests that downstream disparities are better explained by biases in the fine-tuning dataset. Still, pre-training plays a role: simple alterations to co-occurrence rates in the fine-tuning dataset are ineffective when the model has been pre-trained. Our results encourage practitioners to focus more on dataset quality and context-specific harms.

1 Introduction

Large language models (LLMs) and other massively pre-trained "foundation" models are powerful tools for task-specific machine learning (Bommasani et al., 2021). Models pre-trained by well-resourced organizations can easily adapt to a wide variety of downstream tasks in a process called *fine-tuning*. But massive pre-training datasets and increasingly homogeneous model design come with well-known, immediate social risks beyond the financial and environmental costs (Strubell et al., 2019; Bender et al., 2021).

Transformer-based LLMs like BERT and GPT-3 contain quantifiable *intrinsic* social biases encoded in their embedding spaces (Goldfarb-Tarrant et al., 2021). These intrinsic biases are typically associated with representational harms, including stereotyping and denigration (Barocas et al., 2017; Blodgett et al., 2020; Bender et al., 2021). Separately, many studies document the *extrinsic* harms of the downstream (fine-tuned & task-specific) applications of fine-tuned LLMs, including discriminatory medical diagnoses (Zhang et al., 2020), over-reliance on binary gender for coreference resolution (Cao and Daumé, 2021), the re-inforcement of traditional gender roles in part-of-speech tagging (Garimella et al., 2019), toxic text generation (Gehman et al., 2020), and censorship of inclusive language and AAVE (Blodgett and O'Connor, 2017; Blodgett et al., 2018; Park et al., 2018; Sap et al., 2019).

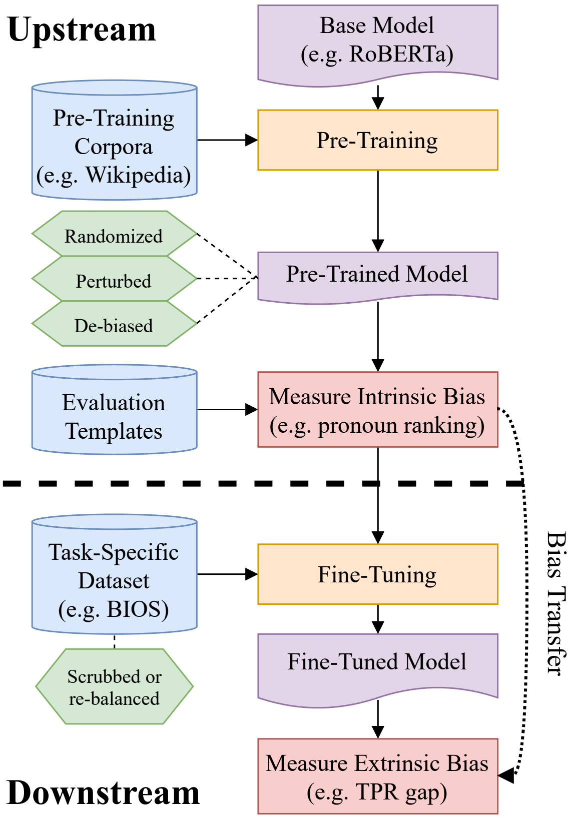

Upstream
Base Model (e.g. RoBERTa)
Pre-Training Corpora (e.g. Wikipedia)
Pre-Training
Randomized
Perturbed
De-biased
Pre-Trained Model
Evaluation Templates
Measure Intrinsic Bias (e.g. pronoun ranking)
Bias Transfer
Task-Specific Dataset (e.g. BIOS)
Fine-Tuning
Scrubbed or re-balanced
Fine-Tuned Model
Downstream
Measure Extrinsic Bias (e.g. TPR gap)

Figure 1: Full pre-training to fine-tuning pipeline, with experimental interventions (green hexagons).

Despite these risks, no research has investigated the extent to which downstream systems inherit social biases from pre-trained models.[1] Many stud-

[1] We use the term "bias" to refer to statistical associations

ies warn that increasing intrinsic bias upstream may lead to an increased risk of downstream harms (Bolukbasi et al., 2016; Caliskan et al., 2017). This hypothesis, which we call the **Bias Transfer Hypothesis**, holds that stereotypes and other biased associations in a pre-trained model are transferred to post-fine-tuning downstream tasks, where they can cause further, task-specific harms. A weaker version of this hypothesis holds that downstream harms are at least mostly determined by the pre-trained model (Bommasani et al., 2021).

In the pre-training paradigm, the extent to which the bias transfer hypothesis holds will determine the most effective strategies for responsible design. In the cases we study, reducing upstream bias does little to change downstream behavior. Still, there is hope: instead, developers can carefully curate the fine-tuning dataset, checking for harms in context.

We test the bias transfer hypothesis on two classification tasks with previously demonstrated performance disparities: occupation classification of online biographies (De-Arteaga et al., 2019) and toxicity classification of Wikipedia Talks comments (Dixon et al., 2018). We investigate whether reducing or exacerbating intrinsic biases encoded by RoBERTa (Liu et al., 2019) decreases or increases the severity of downstream, extrinsic harms (Figure 1). We find that the bias transfer hypothesis describes only part of the interplay between pre-training biases and harms after fine-tuning:

- Systematically manipulating upstream bias has little impact on downstream disparity, especially for the most-harmed groups.

- With a regression analysis, we find that most variation in downstream bias can be explained by bias in the fine-tuning dataset (proxied by co-occurrence rates).

- Altering associations in the fine-tuning dataset can sometimes change downstream behavior, but only when the model is not pre-trained.

Without absolving LLMs or their owners of representational harms intrinsic to pre-trained models, our results encourage practitioners and application stakeholders to focus more on dataset quality and context-specific harm identification and reduction.

2 Related Work

Little prior work directly tests the bias transfer hypothesis. The closest example of this phenomena is the "blood diamond" effect (Birhane and Prabhu, 2021), in which stereotyping and denigration in the pre-training corpora pervade subsequently generated images and language even *before* the fine-tuning stage (Steed and Caliskan, 2021). Still, it is unclear to what extent undesirable values encoded in pre-training datasets or benchmarks—such as Wikipedia or ImageNet—induce task-specific harms *after* fine-tuning (Barocas et al., 2019).

Some work explores the consistency of intrinsic and extrinsic bias metrics: Goldfarb-Tarrant et al. (2021) find that intrinsic and extrinsic metrics are not reliably correlated for static embeddings like `word2vec`. We focus instead on state-of-the-art transformer-based LLMs—the subject of intense ethical debate (Bender et al., 2021; Bommasani et al., 2021)—which construct contextual, rather than static, embeddings. Contextual embeddings—token encodings that are conditional on other, nearby tokens—pose an ongoing challenge for intrinsic bias measurement (May et al., 2019; Kurita et al., 2019; Guo and Caliskan, 2021) and bias mitigation (Liang et al., 2020). We find that intrinsic and extrinsic metrics *are* correlated for the typical LLM—but that the correlation is mostly explained by biases in the fine-tuning dataset.

Other research tests the possibility that upstream mitigation could universally prevent downstream harm. Jin et al. (2021) show that an intermediate, bias-mitigating fine-tuning step can help reduce bias in later tasks. Likewise, Solaiman and Dennison (2021) propose fine-tuning on carefully curated "values-targeted" datasets to reduce toxic GPT-3 behavior. Our results tend to corroborate these methods: we find that the fine-tuning process can to some extent overwrite the biases present in the original pre-trained model. A recent *post-hoc* mitigation technique, on the other hand, debiases contextual embeddings before fine-tuning (Liang et al., 2020). Our results imply that while this type of debiasing may help with representational harms upstream, it is less successful for reducing harms downstream.

3 Methods

To empirically evaluate the bias transfer hypothesis, we examine the relationship between upstream

that result in representational or allocational harms to historically marginalized social groups (Blodgett et al., 2020).

bias and downstream bias for two tasks. We track how this relationship changes under various controlled interventions on the model weights or the fine-tuning dataset.

3.1 Model

For each task, we fine-tune RoBERTa[2] (Liu et al., 2019). We split the fine-tuning dataset into train (80%), evaluation (10%), and test (20%) partitions. To fine-tune, we attach a sequence classification head and train for 3 epochs.[3]

3.2 Occupation Classification

The goal of occupation classification is to predict someone's occupation from their online biography. We fine-tune with the BIOS dataset (De-Arteaga et al., 2019), which includes over 400,000 online biographies scraped from Common Crawl and annotated with (binary) gender. Since self-identified gender was not collected, we will refer instead to the pronouns used in each biography (each biography uses either he/him or she/her pronouns). Following De-Arteaga et al. (2019), we use the "scrubbed" version of the dataset—in which all the identifying pronouns have been removed—to measure just the effects of proxy words (e.g. "mother") and avoid overfitting on pronouns directly.

Downstream Bias.— Biographies with she/her pronouns are less frequently classified as belonging to certain traditionally male-dominated professions—such as "surgeon"—which could result in lower recruitment or callback rates for job candidates if the classifier is used by an employer. The empirical true positive rate (TPR) estimates the likelihood that the classifier correctly identifies a person's occupation from their biography.

We follow previous work (De-Arteaga et al., 2019) in measuring downstream bias via the empirical true positive rate (TPR) gap between biographies using each set of pronouns. First, define

$$\mathrm{TPR}_{y,g} = \mathbb{P}[\hat{Y} = y \mid G = g, Y = y],$$

where g is a set of pronouns and y is an occupation. Y and $\hat{Y}$ represent the true and predicted occupation, respectively. Then the TPR bias (TPB) is

$$\mathrm{TPB}_y = \frac{\mathrm{TPR}_{y,\text{she/her}}}{\mathrm{TPR}_{y,\text{he/him}}}. \tag{1}$$

For example, the classifier correctly predicts "surgeon" for he/him biographies much more often than for she/her biographies, so the TPR ratio for the "surgeon" occupation is low (see Appendix A).

Upstream Bias.— We adapt Kurita et al. (2019)'s pronoun ranking test to the 28 occupations in the BIOS dataset. Kurita et al. (2019) measure the encoded association of he/him and she/her pronouns by the difference in log probability scores between pronouns appearing in templates of the form `{pronoun} is a(n) {occupation}`. We augment this approach with 5 career-related templates proposed by Bartl et al. (2020) (see Appendix A). Formally, given a template sequence $\mathbf{x}_{y,g}$ filled in with occupation y and pronoun g, we compute $p_{y,g} = \mathbb{P}(\mathbf{x}_{y,g})$. As a baseline, we also mask the occupation and compute the prior probability $\pi_{y,g} = \mathbb{P}(\mathbf{x}^{\pi}_{y,g})$. The pronoun ranking bias (PRB) for this template is the difference in log probabilities:

$$\mathrm{PRB}_y = \log \frac{p_{y,\text{she/her}}}{\pi_{y,\text{she/her}}} - \log \frac{p_{y,\text{he/him}}}{\pi_{y,\text{he/him}}}. \tag{2}$$

3.3 Toxicity Classification

For toxicity classification, we use the WIKI dataset, which consists of just under 130,000 comments from the online forum Wikipedia Talks Pages (Dixon et al., 2018). The goal of the task is to predict whether each comment is toxic. Each comment has been labeled as `toxic` or `non-toxic` by a human annotator, where a toxic comment is a "rude, disrespectful, or unreasonable comment that is likely to make you leave the discussion" (Dixon et al., 2018). Following Dixon et al. (2018), we focus on 50 terms referring to people of certain genders, races, ethnicities, sexualities, and religions.

Downstream (Extrinsic) Bias.— Mentions of certain identity groups—such as "queer"—are more likely to be flagged for toxic content, which could result in certain communities being systematically censored or left unprotected if an online platform uses the classifier. The classifier's empirical false positive rate (FPR) estimates its likelihood to falsely flag a non-toxic comment as toxic. The FPR corresponds to the risk of censoring inclusive speech or de-platforming individuals who often mention marginalized groups.

Following Dixon et al. (2018), we express the classifier's bias against comments or commenters harmlessly mentioning an identity term as the FPR

[2] `roberta-base` from HuggingFace (Wolf et al., 2020).

[3] See Appendix D for more details. Epochs and other parameters were chosen to match prior work (Jin et al., 2021).

bias (FPB).

$$\mathrm{FPB}_i = \frac{\mathbb{P}[\hat{T} = 0 \mid I = i, T = 1]}{\mathbb{P}[\hat{T} = 0 \mid T = 1]}, \qquad (3)$$

where i is an identity term and $T = 1$ if the comment was deemed toxic by a human annotator.

Upstream Bias.— Following Hutchinson et al. (2020), we measure upstream bias via sentiment associations. We construct a set of templates of the form `{identity} {person} is [MASK]`, where identities are the identity terms from Dixon et al. (2018) (e.g. "gay" or "Muslim") and the person phrases include "a person," "my sibling," and other relations. We predict the top-20 likely tokens for the "[MASK]" position (e.g., "awesome" or "dangerous"). Using a pre-trained RoBERTA sentiment classifier trained on the TweetEval benchmark (Barbieri et al., 2020), we then measure the average negative sentiment score of the predicted tokens. The model's bias is the magnitude of negative association with each identity term.

RoBERTa sometimes suggests terms which refer back to the target identity group. To mitigate this effect, we drop any predicted tokens that match the 50 identity terms (e.g. "Latino") from Dixon et al. (2018), but we are likely missing other confounding adjectives (e.g. "Spanish"). We suspect this confounding is minimal: we achieve similar results with an alternative ranking-based bias metric (see Appendix C.2).

4 Experiments

We measure changes in upstream and downstream bias subject to the following interventions (Fig. 1):

- **No pre-training.** To control for the effects of pre-training, we test randomly initialized versions of both models that have not been pre-trained. We average over 10 trials.

- **Random perturbations.** We instantiate a pre-trained model and then add random noise e to every weight in the embedding matrix. We try both uniform noise $u \sim \mathrm{Unif}(-c, c)$ and Gaussian noise $z \sim \mathcal{N}(0, \sigma^2)$, varying c and σ^2. The final noise-added matrix is clipped so that its range does not exceed that of the original matrix.

- **Bias mitigation.** We apply the SENTDEBIAS algorithm to de-bias embeddings at the word-level (Liang et al., 2020). SENTDEBIAS estimates a bias subspace $\mathbf{V}$ with principal component analysis, then computes debiased word embeddings $\hat{\mathbf{h}} = \mathbf{h} - \gamma \sum_{j=1}^{k} \langle \mathbf{h}, \mathbf{v}_j \rangle \mathbf{v}_j$ by subtracting the projection of $\mathbf{h}$ onto $\mathbf{V}$. We add the multiplier γ to add or remove bias to various degrees—standard SENTDEBIAS uses $\gamma = 1.0$.

- **Re-balancing and scrubbing.** For BIOS, we re-balance the fine-tuning dataset by under-sampling biographies with the prevalent pronoun in each occupation. For WIKI, we randomly remove from the fine-tuning dataset α percent of comments mentioning each identity term.

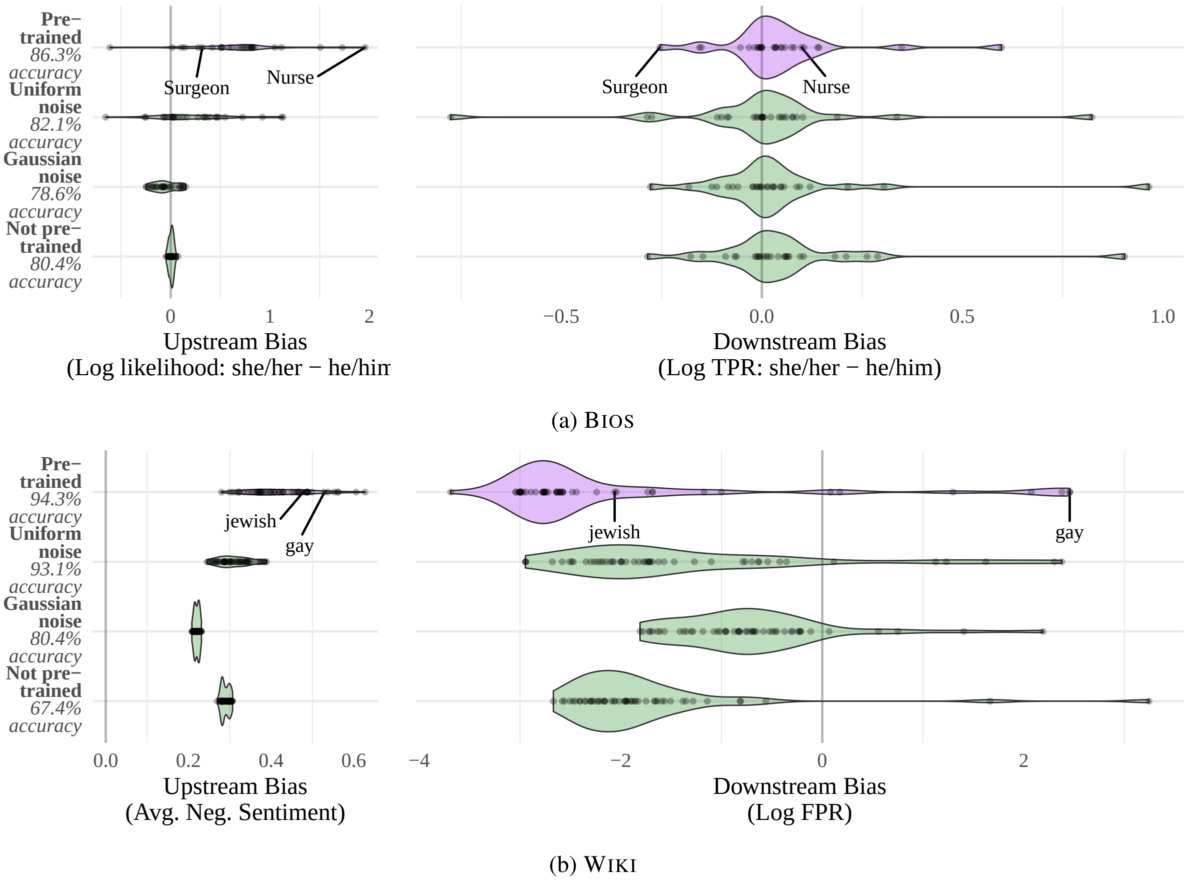
4.1 Upstream variations have little impact on downstream bias.

Our goal is to test the bias transfer hypothesis, which holds that upstream bias is transferred through fine-tuning to downstream models. By this view, we would expect changes to the pre-trained model to also change the distribution of downstream bias—but we find that for both tasks, downstream bias is largely invariant to upstream interventions. Figure 2 summarizes the similarity of biases before and after each randomized event. Though randomizing the model weights significantly reduces the mean and variance of upstream bias, the distribution of downstream bias changes very little.[4] For example, RoBERTa exhibits the same disparities in performance after fine-tuning regardless of whether the base model was pre-trained.

Likewise, although the SENTDEBIAS mitigation method reduces pronoun ranking (upstream) bias as intended, we detect roughly the same downstream biases no matter the level of mitigation applied (Figure 3). For example, in the BIOS task, surgeons with he/him pronouns are still 1.3 times more likely to have their biographies correctly classified than their she/her counterparts.

There is one notable exception to this trend: for the WIKI task, adding noise (uniform or Gaussian) to the pre-trained model's embedding matrix or not pre-training the model yields a modest reduction in median bias (Figure 2). As upstream bias shifts towards zero, downstream bias also moves marginally towards zero. Still, the largest biases (e.g., against the term "gay") do not decrease and may even increase after randomization.

[4] See Appendix B.2 for a full set of correlation tests.



(a) BIOS

(b) WIKI

Figure 2: Bias per occupation after randomized interventions, averaged over 10 trials. Despite drastic changes to the distribution of upstream bias (left), downstream bias remains roughly stable (right). For example, upstream bias *with* pre-training (purple) is not correlated with upstream bias *without* pre-training. But downstream bias *is* partially correlated with the control (Pearson's correlation coefficient $\rho_{\text{BIOS}} = 0.93$ and $\rho_{\text{WIKI}} = 0.64$, $p < 0.01$).

4.2 Most downstream bias is explained by the fine-tuning step.

Though the results in the preceding section suggest that there is no clear or consistent correspondence between *changes* in upstream bias and *changes* in downstream bias, there is still a noticeable correlation between baseline upstream and downstream bias (Pearson's $\rho = 0.43$, $p = 0.022$ for BIOS, $\rho = 0.59$, $p < 10^{-5}$ for WIKI—see Appendix A). There is an important third variable that helps explain this correlation: cultural artifacts ingrained in both the pre-training and fine-tuning datasets.[5] RoBERTa learns these artifacts through co-occurrences and other associations between words in both sets of corpora.

To test this explanation, we conduct a simple regression analysis across interventions (Figure 1) and evaluation templates. We estimate

$$\log \text{TPB}_{m,y} = \beta_0 + \beta_1 \text{PRB}_{m,y,t} + \beta_2 \pi_y + f_s + c_m. \tag{4}$$

for model treatment m, occupation y, and pronoun ranking template s. TPB is the TPR bias (downstream bias) from Eq. 1; PRB is the pronoun ranking bias (upstream bias) from Eq. 2; f_s and c_m are dummy variables (for ordinary least squares) or fixed effects to capture heterogeneous effects between templates and models (such as variations in overall embedding quality). We control for statistical "dataset bias" with π, the prevalence of "she/her" biographies within each occupation y in the fine-tuning data.

We find that the "dataset bias" in the fine-tuning stage explains most of the correlation between upstream and downstream bias. Under the strong bias transfer hypothesis, we would expect the coefficient on upstream bias β_1 to be statistically significant and greater in magnitude than the coefficient β_2 on our proxy for dataset bias. But for both tasks,

[5] For example, cultural biases about which pronouns belong in which occupations are likely to pervade both the pre-training dataset (e.g., Wikipedia) and the fine-tuning dataset (internet biographies).

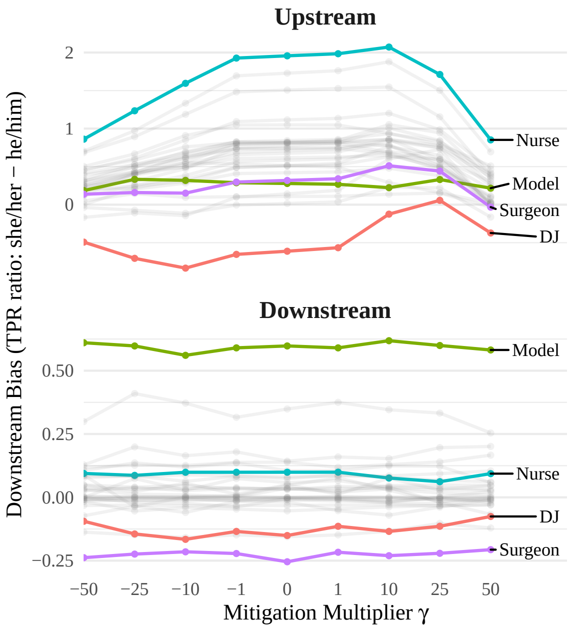


Figure 3: Bias per occupation after scaled SENTDEBIAS on the BIOS task. Mitigation significantly reduces pronoun ranking (upstream) bias compared to base RoBERTa (top); but even when upstream bias decreases, the TPR ratio (downstream bias) remains mostly constant (bottom). The distribution of downstream bias without any mitigation is almost perfectly correlated with the distribution at $\gamma = 50$ (Pearson's $\rho = 0.96$, $p < 0.01$).

the opposite is true: fine-tuning dataset bias has a larger effect than upstream bias. Figure 4 reports the coefficient estimates for these two variables. (See Appendix C.1 for all estimates, standard errors, assumptions and additional specifications.)

In the BIOS task, a large decrease in upstream bias corresponds to a small but statistically significant *increase* in downstream bias. On average, a reduction of 0.3 to the log likelihood gap—equivalent to the reduction in bias towards nurses after upstream mitigation—corresponds to a 0.5% increase in the TPR ratio. Almost all the downstream bias in the BIOS task is explained by dataset bias instead: a 10% increase in the prevalence of she/her pronouns within an occupation corresponds to a much larger 6.5% increase in the TPR ratio.

In the WIKI task, upstream bias has a more noticeable effect—but the effect of dataset bias is still much larger. The regression takes the same form as Eq. 4, where downstream bias is FPR bias (Eq. 3), upstream bias is negative sentiment, and π_i is the proportion of toxic mentions of identity i. We additionally control for the prevalence of each identity term and the average length of toxic mentions of each identity term—longer comments are less likely to result in erroneous screening (Appendix C.1).

As in the previous regression, dataset bias explains more of the variation in downstream bias than does upstream bias. On average, a *large* increase in average negative sentiment against a given identity term (e.g. 0.1, one standard deviation) corresponds to only a modest 3.7% increase in FPR. In comparison, only a 10% increase in the prevalence of toxic mentions of an identity corresponds to an even larger 6.3% increase in FPR.

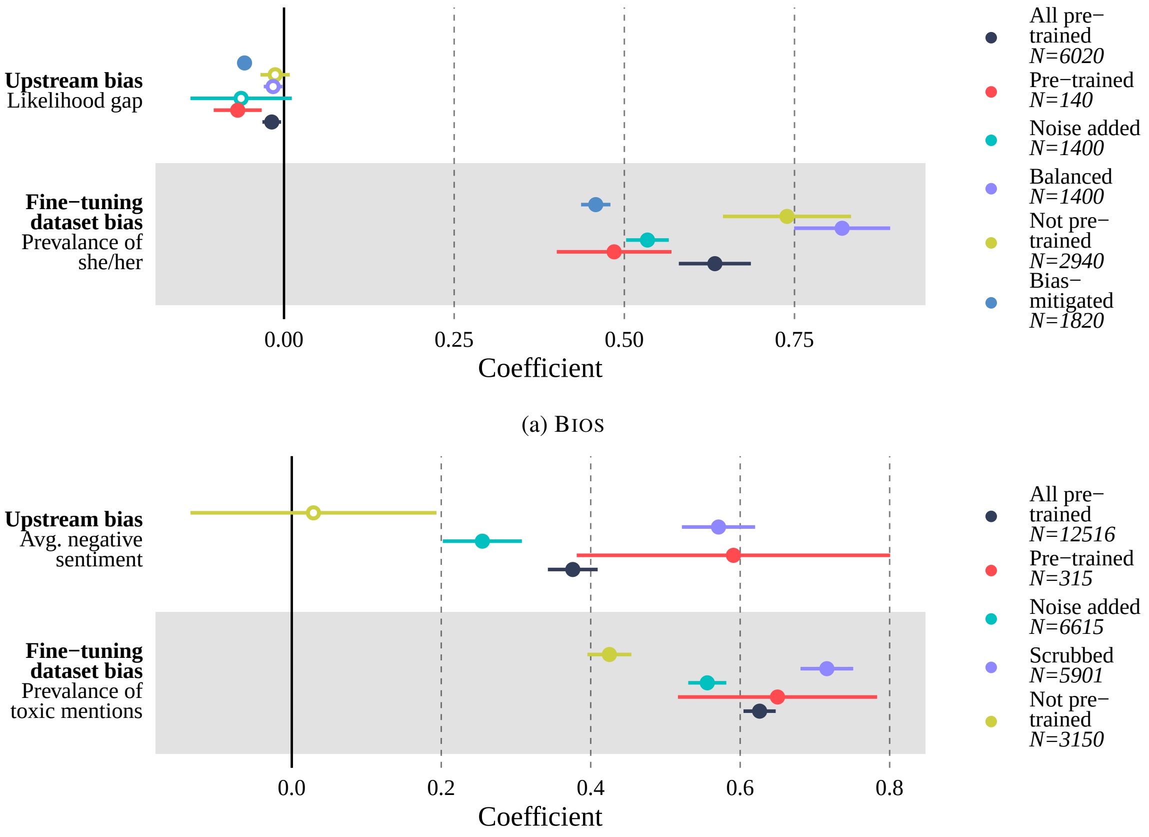
We also check that *intrinsic* downstream bias also changes due to fine-tuning. We measure intrinsic bias again after fine-tuning and regress on downstream intrinsic bias instead of downstream extrinsic bias (Eq. 4). The results are consistent: after controlling for the overall increase in log likelihood, the effect of upstream intrinsic bias on downstream intrinsic bias is explained almost entirely by fine-tuning dataset bias (Appendix C.1).

4.3 Re-sampling and re-scrubbing has little effect on downstream behavior.

Given the strong relation between our proxies for dataset bias and downstream bias, we test whether manipulating these proxies admits some control over downstream bias. For example, were the fine-tuning dataset to include exactly as many she/her nurse biographies as he/him, would the model still exhibit biased performance on that occupation?

Our findings suggest not. No matter the amount of re-sampling, downstream bias remains relatively stable for pre-trained RoBERTa. The distributions of downstream bias with and without re-balancing are almost perfectly correlated (Pearson's $\rho = 0.94$, $p < 0.01$—see Appendix B.1). Though co-occurrence statistics help to explain downstream bias, they are still only proxies for dataset bias. Directly altering these statistics via re-sampling the dataset does not alter the sentence-level context in which the words are used.

Based on this result, we also try completely removing mentions of identity terms. Scrubbing mentions of identity terms—in all comments or only in toxic comments—appears to reduce bias only when the model is not pre-trained and all mentions of the term are scrubbed (Figure 5). For a pre-trained model trained on scrubbed data, a 10% decrease in mentions of an identity term corresponds to a 7.2%



(a) BIOS

(b) WIKI

Figure 4: Effect of upstream bias vs. fine-tuning dataset bias on downstream bias, controlling for model & template fixed effects. Bars depict heteroskedasticity-consistent standard errors. Statistically insignificant ($p < 0.01$) coefficients are hollow. For both tasks, reduction in fine-tuning dataset bias corresponds to a greater alteration to downstream bias than an equivalent reduction (accounting for scale) in upstream bias.

decrease in FPR. We speculate that RoBERTa relies on its high quality feature embeddings to learn proxy biases about identity terms based on the way they are used in the pre-training corpora. For example, our model classifies a sequence containing only the term "gay" as toxic without any context. If a term like "gay" is often used pejoratively on the web, RoBERTa is likely to infer that sentences including "gay" are toxic even if the term never appears in the fine-tuning dataset.

But when the upstream model is *not* pre-trained, the fine-tuned model has no such prejudices. In this case, removing all mentions of identity results in a distribution of bias entirely uncorrelated with the control (Pearson's $\rho = 0.09$, $p > 0.1$). Notably, though, even a small number of mentions of an identity term like "gay" in the fine-tuning dataset are enough for a randomly initialized model to exhibit the same biases as the pre-trained model (Figure 5).

5 Limitations

Our approach comes with several limitations. First, our results may not generalize to all tasks—especially non-classification tasks—or all kinds of bias (e.g., bias against AAVE or non-English speakers). Also, while similar studies of bias have been successfully applied to vision transformers (Steed and Caliskan, 2021; Srinivasan and Uchino, 2021), our results may vary for substrates other than English language.

Second, Goldfarb-Tarrant et al. (2021) conclude that the lack of correlation between intrinsic bias indicators and downstream bias is because some embedding bias metrics are unsuitable for measuring model bias. To ensure our intrinsic and extrinsic metrics measure the same construct, we chose upstream indicators that correlate with real-world occupation statistics (Caliskan et al., 2017; Kurita et al., 2019). Pronoun ranking in particular may be more reliable for transformer models than

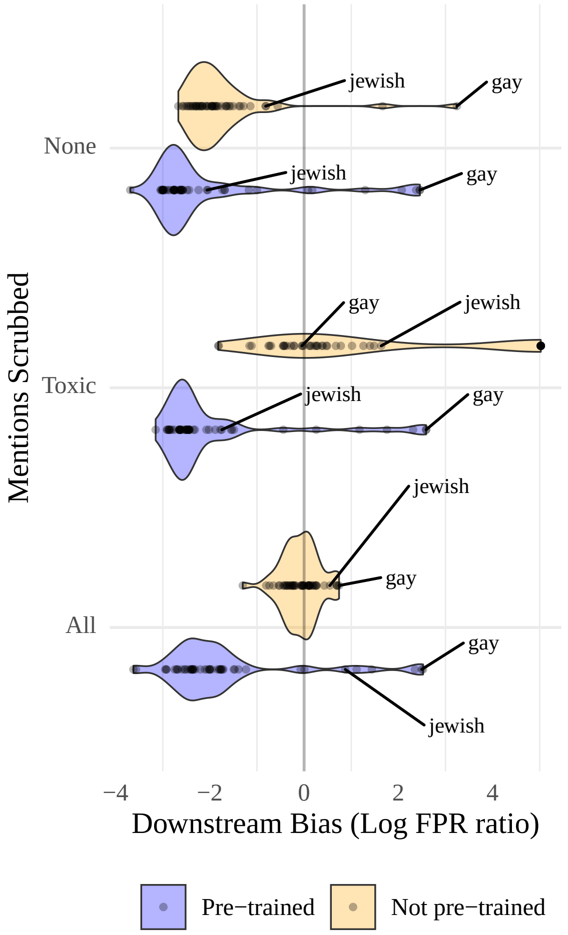


Figure 5: FPR gap (downstream bias) after scrubbing toxic mentions of identity terms from the WIKI fine-tuning dataset. A combination of scrubbing and not pre-training (orange) results in a zero-mean, noticeably re-ordered bias distribution. Scrubbing but still pre-training (purple) results in a bias distribution that is still correlated with the original bias distribution (Pearson's $\rho = 0.99, 0.93$ for toxic and all respectively, $p < 0.01$).

other metrics (Silva et al., 2021). Still, downstream, annotator prejudices and other label biases could skew our extrinsic bias metrics as well (Davani et al., 2021).

Third, there may be other explanations for the relationship between upstream and downstream bias: for example, decreasing the magnitude of upstream bias often requires a reduction in model accuracy, though we attempt to control for between-model variation with fixed effects and other controls. Alternate regression specifications included in Appendix C.1 show how our results change with the inclusion of controls.

6 Conclusion

Our results offer several points of guidance to organizations training and distributing LLMs and the practitioners applying them:

- Attenuating downstream bias via upstream interventions—including embedding-space bias mitigation—is mostly futile in the cases we study and may be fruitless in similar settings.
- For a typical pre-trained model trained for the tasks we study, the fine-tuning dataset plays a much larger role than upstream bias in determining downstream harms.
- Still, simply modulating co-occurence statistics (e.g., by scrubbing harmful mentions of certain identities) is not sufficient. Task framing, design, and data quality are also very important for preventing harm.
- If a model is pre-trained, it may be more resistant to scrubbing, re-balancing, and other simple modulations of the fine-tuning dataset.

But, our results also corroborate a nascent, somewhat optimistic view of pre-training bias. LLMs' intrinsic biases are harmful even before downstream applications, and correcting those biases is not guaranteed to prevent downstream harms. Increased emphasis on the role of fine-tuning dataset bias offers an opportunity for practitioners to shift to more careful, quality-focused and context-aware approach to NLP applications (Zhu et al., 2018; Scheuerman et al., 2021).

Ethical Considerations

This study navigates several difficult ethical issues in NLP ethics research. First, unlike prior work, we do not claim to measure gender biases—only biases related to someone's choice of personal pronouns. However, our dataset is limited to the English "he/him" and "she/her," so our results do not capture biases against other pronouns. Our study is also very Western-centric: we study only English models/datasets and test for biases considered normatively pressing in Western research. Second, our training data (including pre-training datasets), was almost entirely scraped from internet users without compensation or explicit consent. To avoid exploiting these users further, we only used already-scraped data and replicated already-existing classifiers, and we do not release these data or classifiers publicly. Finally, the models we trained exhibit toxic, offensive behavior. These models and datasets are intended only for studying bias and simulating harms and, as our results show, should not be deployed or applied to any other data except for this purpose.

Acknowledgements

Thanks to Maria De-Arteaga and Benedikt Boecking for assistance with BIOS data collection. Thanks also to the reviewers for their helpful comments and feedback.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org.

Marion Bartl, Malvina Nissim, and Albert Gatt. 2020. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 1–16, Barcelona, Spain (Online). Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Abeba Birhane and Vinay Uday Prabhu. 2021. Large image datasets: A pyrrhic win for computer vision? In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. ISSN: 2642-9381.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Su Lin Blodgett and Brendan O'Connor. 2017. Racial Disparity in Natural Language Processing: A Case Study of Social Media African-American English. *arXiv:1707.00061 [cs]*. ArXiv: 1707.00061.

Su Lin Blodgett, Johnny Wei, and Brendan O'Connor. 2018. Twitter Universal Dependency Parsing for African-American and Mainstream American English. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1415–1425, Melbourne, Australia. 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 D D Lee, M Sugiyama, U V Luxburg, I Guyon, and R Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 4349–4357. Curran Associates, Inc.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Kohd, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2021. On the Opportunities and Risks of Foundation Models. *arXiv:2108.07258 [cs]*. ArXiv: 2108.07258.

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

Yang Trista Cao and Hal Daumé, III. 2021. Toward Gender-Inclusive Coreference Resolution: An Analysis of Gender and Bias Throughout the Machine Learning Lifecycle*. *Computational Linguistics*, 47(3):615–661.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2021. Hate Speech Classifiers Learn Human-Like Social Stereotypes. *arXiv:2110.14839 [cs]*. ArXiv: 2110.14839.

Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin 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* '19, pages 120–128, New York, NY, USA. Association for Computing Machinery.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73, New Orleans LA USA. ACM.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's Syntactic Resilience and Men's Grammatical Luck: Gender-Bias in Part-of-Speech Tagging and Dependency Parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic Bias Metrics Do Not Correlate with Application Bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Wei Guo and Aylin Caliskan. 2021. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 122–133. Association for Computing Machinery, New York, NY, USA.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social Biases in NLP Models as Barriers for Persons with Disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, Online. Association for Computational Linguistics.

Xisen Jin, Francesco Barbieri, Brendan Kennedy, Aida Mostafazadeh Davani, Leonardo Neves, and Xiang Ren. 2021. On Transferability of Bias Mitigation Effects in Language Model Fine-Tuning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3770–3783, Online. 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 (ACL).

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

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. *arXiv:1907.11692 [cs]*. ArXiv: 1907.11692.

David P. MacKinnon, Jennifer L. Krull, and Chondra M. Lockwood. 2000. Equivalence of the Mediation, Confounding and Suppression Effect. *Prevention Science*, 1(4):173–181.

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*, pages 622–628, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Morgan Klaus Scheuerman, Alex Hanna, and Emily Denton. 2021. Do Datasets Have Politics? Disciplinary Values in Computer Vision Dataset Development. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):317:1–317:37.

Andrew Silva, Pradyumna Tambwekar, and Matthew Gombolay. 2021. Towards a Comprehensive Understanding and Accurate Evaluation of Societal Biases in Pre-Trained Transformers. In *Proceedings of the 2021 Conference of the North American Chapter of* *the Association for Computational Linguistics: Human Language Technologies*, pages 2383–2389, Online. Association for Computational Linguistics.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. In *Pre-Proceedings of Advances in Neural Information Processing Systems*, volume 34.

Ramya Srinivasan and Kanji Uchino. 2021. Biases in Generative Art: A Causal Look from the Lens of Art History. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 41–51, New York, NY, USA. Association for Computing Machinery.

Ryan Steed and Aylin Caliskan. 2021. Image Representations Learned With Unsupervised Pre-Training Contain Human-like Biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 701–713, New York, NY, USA. Association for Computing Machinery.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

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.

Haoran Zhang, Amy X. Lu, Mohamed Abdalla, Matthew McDermott, and Marzyeh Ghassemi. 2020. Hurtful words: quantifying biases in clinical contextual word embeddings. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, CHIL '20, pages 110–120, New York, NY, USA. Association for Computing Machinery.

Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):194:1–194:23.

A Descriptive Statistics

BIOS.— Biographies include the 28-most frequent occupations according to the BLS Standard Occupation Classification system.[6] See Figure 6 for a full list of occupations and the prevalence of each set of pronouns.

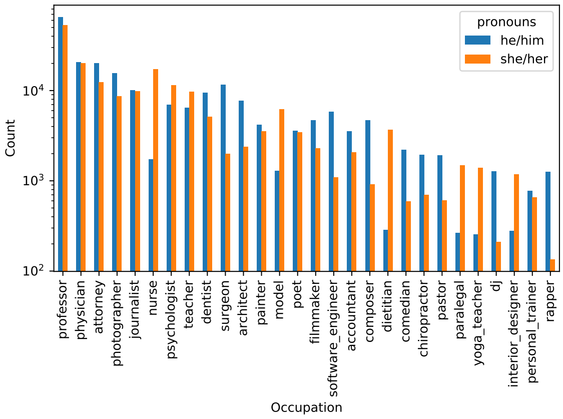


Figure 6: Frequency of occupations in BIOS dataset.

Upstream bias (measured with pronoun ranking) is depicted in Figure 7. Table 1 Gives the full list of templates used for testing. Traditionally female occupations (e.g., "nurse") are generally biased towards "she/her," with some exceptions (e.g., "software engineer"). Downstream bias is similarly distributed—Figure 9 depicts the relationship between upstream and downstream bias, which is generally linear (Pearson's $\rho = 0.43$, $p < 0.05$). There are noticeable outliers (e.g., "surgeon") for which real-world harms could be especially acute.

WIKI.— Identity terms occur relatively infrequently in the dataset and some terms appear in toxic comments more often than others. See Figure 8 for a full list of identity terms tested and their prevalence in toxic and non-toxic comments.

Upstream and downstream bias have an even stronger relationship for WIKI than for BIOS ($\rho = 0.59$, $p < 0.01$). The terms "homosexual" and "gay" have especially high biases both upstream and downstream, likely because of their use as common pejoratives on certain internet spaces represented in the pre-training and fine-tuning datasets.

B Additional Results

B.1 Partial Interventions

This section gives the full results for each partial intervention (e.g., scrubbing only half of the toxic mentions, rather than all of them). Figure 10 shows downstream bias at each level of pronoun re-balancing. To measure the effects of different levels of gender balance, we either randomly oversample the minority group to close α percent of the gap between the majority and minority pronouns, randomly under-sample the majority group to close β percent of the gap, or some combination of both.

Figure 11 shows the full set of partial scrubbing interventions tested - notably, there is a slight convergence to zero in downstream bias when the model is *not* pre-trained.

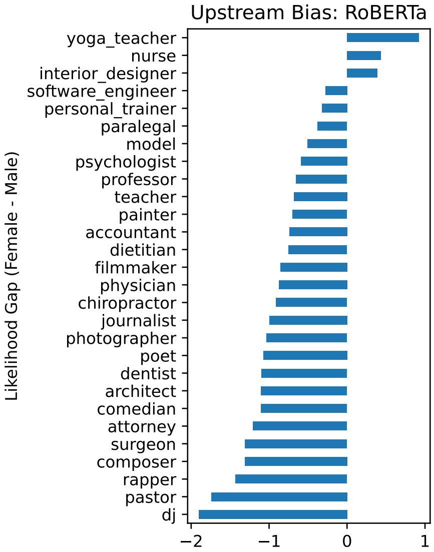


Figure 7: Distribution of pronoun ranking (upstream) bias across occupations in the BIOS task.

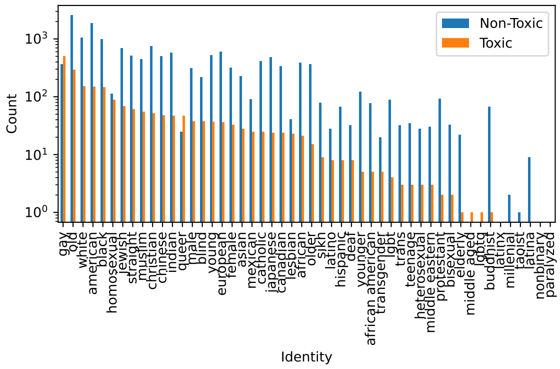


Figure 8: Frequency of comments mentioning each identity term in the WIKI dataset.

B.2 Correlation Tests

To quantify the effect of our experimental interventions on upstream and downstream bias, we compared the distribution of bias after each in-

[6] https://www.bls.gov/soc/

Bias	Template	Targets	Source
Occupation-Pronoun	[TARGET] is a(n) [OCCUPATION]	he, she	(Kurita et al., 2019)
Occupation-Pronoun	[TARGET] works as a(n) [OCCUPATION]	he, she	(Bartl et al., 2020)
Occupation-Pronoun	[TARGET] wants to become a(n) [OCCUPATION]	he, she	(Bartl et al., 2020)
Occupation-Pronoun	[TARGET] applied for the position of [OCCUPATION]	he, she	(Bartl et al., 2020)
Identity-Sentiment	[IDENTITY] [TARGET] are [MASK]	people	(Dixon et al., 2018)
Identity-Sentiment	[IDENTITY] [TARGET] is [MASK]	my sibling my friend my parent my partner my spouse	(Dixon et al., 2018)

Table 1: Templates used for bias measurement.

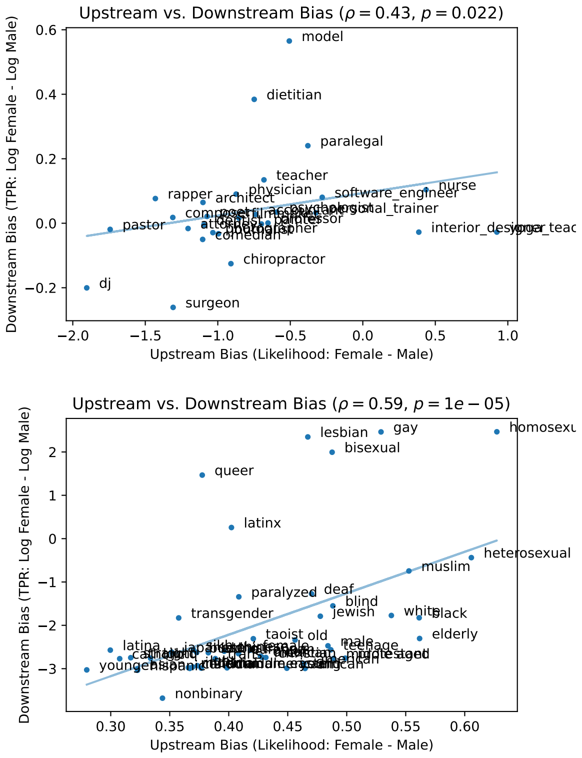


Figure 9: Correlation between upstream and downstream bias across occupations (BIOS) and identity terms (WIKI). ρ is the Pearson correlation coefficient.

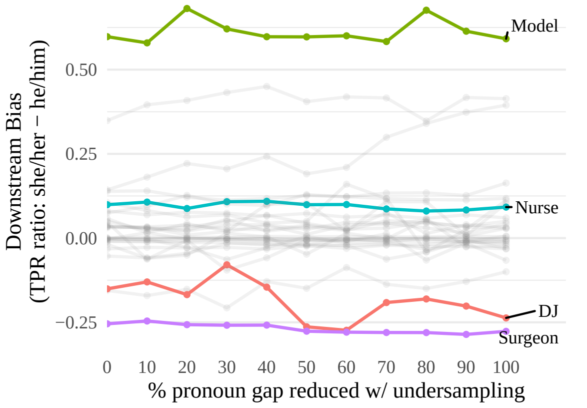


Figure 10: TPR gap (downstream bias) after balancing pronouns within each occupation of the BIOS finetuning dataset. As shown, balancing pronoun prevalence has little effect on downstream bias.

tervention to the distribution of bias after an unmodified pre-trained model. We tested for statistical correlation between these two distributions with both Pearson's correlation coefficient ρ and Kendall's correlation coefficient τ. For Pearson's, we assume that the two distributions are approximately normally distributed. This assumptions seems reasonable because our samples are not too small ($N = 28$ and $N = 50$ for BIOS and WIKI, respectively), but a Shapiro-Wilk test of normality shows that downstream bias for both tasks is likely non-normal ($W = 0.81$ and $W = 0.67$ for TPR ratio and FPR ratio respectively, $p < 0.01$). So, we also compute Kendall's correlation coefficient τ, which is a nonparametric test of ordinal association. The results are similar in magnitude and significance (Table 2).

C Full Regression Results

Tables 3 and 4 report the full set of coefficient estimates used to generate Figure 4 and the effects described in the paper. We use HC3

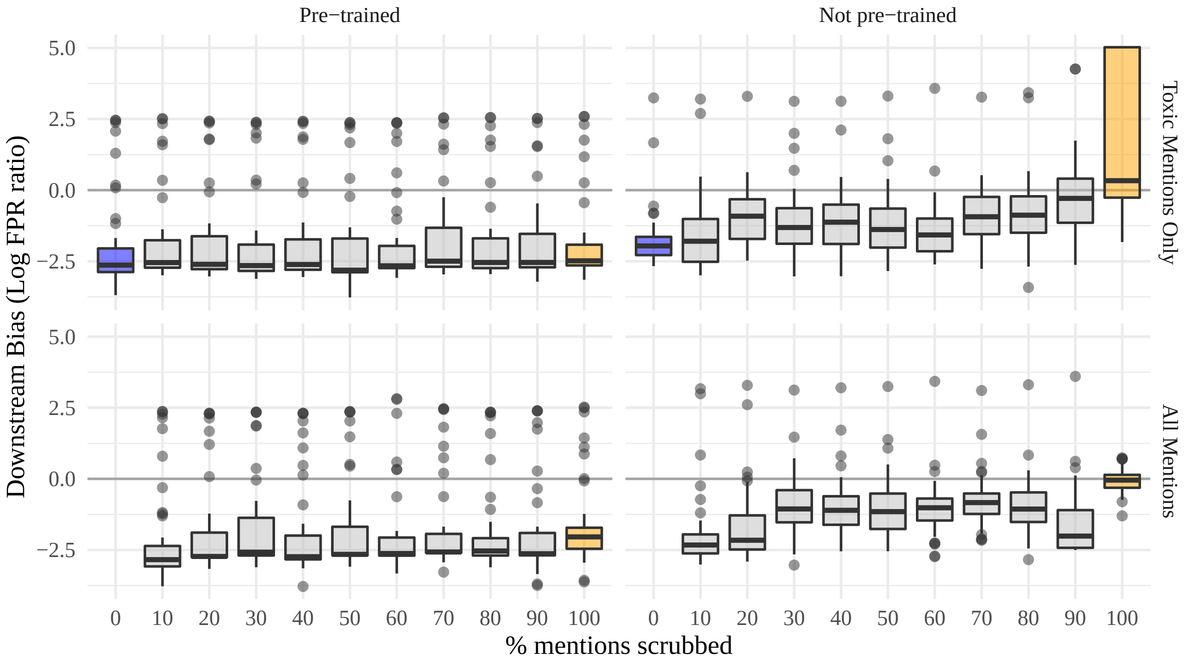


Figure 11: FPR gap (downstream bias) after scrubbing toxic mentions of identity terms from the WIKI fine-tuning dataset.

heteroskedasticity-consistent standard errors.[7] The Variance Inflation Factor (VIF) for the covariates are all less than 2.5 for an unmodified pre-trained model for both tasks (a sign that multicollinearity may not be too severe).

The fixed effects regression requires a few assumptions for unbiased, normally-converging estimates. First, we assume that the error is uncorrelated with every covariate (i.e., there are no omitted variables; we discuss this possibility in the limitations section). Second, we assume that the samples are independent and identically distributed (independence is assured by our experimental setup, which varies one factor at a time). Third, we assume that large outliers are unlikely (evident from the distribution plots presented).

C.1 Additional Specifications

We tested several regression specifications on just the unmodified, pre-trained model (Tables 5 and 6). For BIOS, note that the direct and indirect (after controlling for dataset bias) effects of upstream bias on downstream bias have opposite signs. The change is the effect of including dataset bias, a colinear confounder, in the regression. Confounders can be interpreted as "explaining" the relationship between the independent (upstream) and dependent (downstream) variables (MacKinnon et al., 2000).

To test whether the effect observed is mediated by a change in the model's weights, also include an estimate of the effect of upstream intrinsic bias (e.g., from pronoun ranking) on *downstream* intrinsic bias (intrinsic bias, measured after fine-tuning). We control for the overall increase in log likelihood by including in the regression the difference in log likelihood of the neutral pronouns "they/them" before and after fine-tuning. We find that a similar relationship holds between upstream bias and intrinsic bias downstream as holds between upstream bias and extrinsic bias downstream, suggesting that the model's internal representations change in concert with its downstream behavior.

C.2 Identity Ranking - Robustness Check

Because of the limitations of the sentiment-based approach, we check the robustness of our results with an identity ranking approach based on pronoun ranking. Included in the Dixon et al. (2018) study of toxicity classification bias is an extensive evaluation set composed of 89,000 templates such as "[IDENTITY] is [ATTRIBUTE]," where the attributes include both positive (for non-toxic examples) and extremely negative words (for toxic

[7] For a simple OLS specification on WIKI, the Breusch-Pagan test rejects the hypothesis that our errors are homoskedastic with $BP = 27.039$, $p < 10^{-3}$. For BIOS, the hypothesis is not rejected ($BP = 5.033$, $p = 0.41$).

Intervention	Upstream Bias Pearson's ρ	 Kendall's τ	Downstream Bias Pearson's ρ	 Kendall's τ
BIOS	*Pronoun ranking*		*TPR ratio*	
Not pre-trained	-0.08	0.04	0.93***	0.74***
Uniform noise	0.90***	0.62***	0.67***	0.60***
Gaussian noise	0.29	0.09	0.90***	0.56***
SENTDEBIAS ($\gamma = 50$)	0.87***	0.61***	0.96***	0.77***
Dataset re-balancing ($\beta = 1.0$)			0.94***	0.85***
WIKI	*Negative sentiment*		*FPR ratio*	
Not pre-trained	-0.21	-0.14	0.64***	0.39***
Uniform noise	0.56***	0.37***	0.91***	0.56***
Gaussian noise	0.36***	0.24**	0.78***	0.45***
Scrubbing toxic mentions			0.99***	0.85***
Scrubbing all mentions			0.93***	0.77***
Scrubbing toxic mentions, not pre-trained	0.30**	0.21**	-0.11	-0.07
Scrubbing all mentions, not pre-trained	0.30**	0.21**	0.09	0.10

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: Correlation between bias distributions before and after each intervention. Statistically insignificant correlation coefficients indicate bias has changed drastically (red). Notably, downstream bias is correlated with the control to some extent for every intervention except for scrubbing and not pre-training. Pearson's correlation coefficient ρ measure of correlation strength and direction; Kendall's τ is a measure of ordinal correlation. Randomized interventions (e.g., not pre-training, adding noise) tend to re-order the bias distribution more than others, indicated by a lower τ.

Table 3: Effect of upstream on downstream bias for pre-trained RoBERTa on the BIOS task. Panel linear models include model fixed effects.

	Dependent variable:					
	Log TPR ratio (downstream bias)					
	OLS			*panel linear*		
	Pre-trained	Mitigated	Noise added	Random	Balanced	All pre-trained
	(1)	(2)	(3)	(4)	(5)	(6)
Likelihood gap (upstream bias)	−0.068*** (0.018)	−0.058*** (0.005)	−0.063* (0.038)	−0.013 (0.011)	−0.016** (0.007)	−0.018*** (0.007)
Prevalance of she/her	0.485*** (0.043)	0.458*** (0.011)	0.534*** (0.016)	0.739*** (0.048)	0.820*** (0.036)	0.633*** (0.027)
Constant	−0.090*** (0.029)					
Template Dummies?	Yes	Yes	Yes	Yes	Yes	Yes
Observations	140	1,820	1,400	2,940	1,400	6,020
R^2	0.500	0.489	0.438	0.075	0.297	0.085
Adjusted R^2	0.477	0.484	0.432	0.067	0.289	0.078
Residual Std. Error	0.109 (df = 133)					
F Statistic	22.149*** (df = 6; 133)	287.282*** (df = 6; 1801)	179.585*** (df = 6; 1384)	39.276*** (df = 6; 2913)	97.257*** (df = 6; 1384)	92.307*** (df = 6; 5971)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4: Effect of upstream on downstream bias for pre-trained RoBERTa on the WIKI task. Panel linear models include model fixed effects.

	Dependent variable:				
	FPR				
	OLS	*panel linear*			
	Pre-trained	Noise added	Random	Scrubbed	All pre-trained
	(1)	(2)	(3)	(4)	(5)
Avg. negative sentiment (upstream bias)	0.591***	0.255***	0.029	0.571***	0.376***
	(0.107)	(0.027)	(0.084)	(0.025)	(0.017)
Prevalance of toxic mentions	0.650***	0.556***	0.425***	0.716***	0.626***
	(0.068)	(0.013)	(0.015)	(0.018)	(0.011)
Prevalance of identity term	−5.024	1.575**	6.526***	−7.274***	−1.231**
	(3.740)	(0.708)	(0.815)	(1.086)	(0.612)
Avg. length of toxic mentions	−0.373***				
	(0.077)				
Template Dummies?	Yes	Yes	Yes	Yes	Yes
Observations	315	6,615	3,150	5,901	12,516
R^2	0.296	0.225	0.210	0.283	0.241
Adjusted R^2	0.276	0.221	0.206	0.279	0.238
Residual Std. Error	0.221 (df = 305)				
F Statistic	14.283*** (df = 9; 305)	211.998*** (df = 9; 6585)	92.711*** (df = 9; 3131)	257.043*** (df = 9; 5873)	439.225*** (df = 9; 12467)
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 5: Effect of upstream on downstream bias for pre-trained RoBERTa on the BIOS task.

	Dependent variable:				
	TPR ratio (downstream bias)				Likelihood gap after fine-tuning (intermediate bias)
	OLS		*panel linear*		*panel linear*
	(1)	(2)	(3)	(4)	(5)
Likelihood gap (upstream bias)	0.043**	−0.068***	0.043**	−0.068***	0.046
	(0.021)	(0.018)	(0.021)	(0.018)	(0.575)
Prevalance of she/her		0.485***		0.485***	10.311***
		(0.043)		(0.043)	(1.424)
Difference in they/them log likelihood before and after pre-training					0.206**
					(0.086)
Constant	−0.013	−0.090***			
	(0.039)	(0.029)			
Template dummies?	Yes	Yes	No	No	No
Observations	140	140	140	140	140
R^2	0.031	0.500	0.031	0.500	0.366
Adjusted R^2	−0.005	0.477	−0.005	0.477	0.332
Residual Std. Error	0.151 (df = 134)	0.109 (df = 133)			
F Statistic	0.856 (df = 5; 134)	22.149*** (df = 6; 133)	4.279** (df = 1; 134)	66.447*** (df = 2; 133)	25.386*** (df = 3; 132)
Note:					*p<0.1; **p<0.05; ***p<0.01

Table 6: Effect of upstream on downstream bias for pre-trained RoBERTa on the WIKI task.

	Dependent variable:					
	FPR					Avg. negative sentiment after fine-tuning (intermediate bias)
	OLS			*panel linear*		*panel linear*
	(1)	(2)	(3)	(4)	(5)	(6)
Avg. negative sentiment (upstream bias)	0.569***	0.568***	0.586***	0.569***	0.591***	−0.004
	(0.110)	(0.106)	(0.108)	(0.110)	(0.107)	(0.014)
Prevalance of toxic mentions		0.657***	0.654***		0.650***	−0.020**
		(0.068)	(0.070)		(0.068)	(0.009)
Prevalance of identity term			−4.973		−5.024	0.656
			(3.749)		(3.740)	(0.481)
Avg. length of toxic mentions			0.00001			
			(0.00002)			
Constant	−0.281***	−0.371***	−0.375***			
	(0.079)	(0.077)	(0.078)			
Template Dummies?	Yes	Yes	Yes	No	No	No
Observations	350	315	315	350	315	315
R^2	0.073	0.292	0.297	0.073	0.296	0.024
Adjusted R^2	0.054	0.274	0.274	0.054	0.276	−0.004
Residual Std. Error	0.240 (df = 342)	0.221 (df = 306)	0.221 (df = 304)			
F Statistic	3.829*** (df = 7; 342)	15.801*** (df = 8; 306)	12.826*** (df = 10; 304)	26.802*** (df = 1; 342)	42.848*** (df = 3; 305)	2.551* (df = 3; 305)
Note:						*p<0.1; **p<0.05; ***p<0.01

examples). Templates of other forms are not included to reduce computation time.[8] For each of these templates, we mask the identity term and compute the log probability score as described in §3.2. The model's bias is described by the difference between the average log probability scores for the toxic templates and the non-toxic templates for each identity term.

For the regressions (Tables 7 and 8), the templates are not paired, so we average first across toxic and non-toxic templates, then calculate the ratio between the two. The relative size and statistical significance of the coefficients are the same as for the negative sentiment approach, suggesting the negative sentiment metric is robust for our purposes despite its limitations.

D Replication

We provide our full results (upstream and downstream bias for every intervention, for each task) and the scripts used to analyse them. We are not allowed to release the source code used to train our models and measure bias, but we include additional details on our implementation to help others understand and replicate our results.

- Our code for the pronoun ranking tests is adapted from Zhang et al. (2020)'s implementation available at https://github.com/MLforHealth/HurtfulWords.

- Our code for SENTDEBIAS is adapted from the original authors' (Liang et al., 2020), available at https://github.com/pliang279/sent_debias.

- Epochs and other parameters were chosen to match prior work on the same tasks (Jin et al., 2021). We train with 5 epochs, batch sizes 16 and 64 for training and evaluation respectively, and a learning rate of $5de-6$. Otherwise, we use the default hyperparameters for roberta-base (https://huggingface.co/roberta-base).

- Code for scraping the BIOS dataset is provided by the original authors at https://github.com/microsoft/biosbias. The WIKI dataset is available at https://github.com/conversationai/unintended-ml-bias-analysis.

- Fine-tuning a single model for either task takes from 4-6 hours on single NVIDIA Tesla V100 16GB GPU. Our results include approximately 60 model permutations for a total of 240-360 GPU hours. roberta-base has 125M parameters, but we did not pre-train any models from scratch.

[8] A full list of these templates can be found in (Dixon et al., 2018) or https://github.com/conversationai/unintended-ml-bias-analysis.

Table 7: Effect of upstream on downstream bias for pre-trained RoBERTa on the WIKI task.

	Dependent variable:		
	FPR		
	(1)	(2)	(3)
Avg. log likelihood ratio (upstream bias)	0.399**	0.292	0.286
	(0.192)	(0.176)	(0.187)
Prevalance of toxic mentions		0.624***	0.628***
		(0.180)	(0.186)
Prevalance of identity term			0.00001
			(0.0001)
Avg. length of toxic mentions	0.096***	0.008	0.004
	(0.034)	(0.040)	(0.058)
Observations	50	50	50
R^2	0.083	0.269	0.269
Adjusted R^2	0.064	0.238	0.222
Residual Std. Error	0.241 (df = 48)	0.217 (df = 47)	0.220 (df = 46)
F Statistic	4.345** (df = 1; 48)	8.649*** (df = 2; 47)	5.648*** (df = 3; 46)
Note:			*p<0.1; **p<0.05; ***p<0.01

Table 8: Effect of upstream on downstream bias for pre-trained RoBERTa on the WIKI task.

	Dependent variable:				
	FPR ratio (downstream bias)				
	OLS	*panel linear*			
	Pre-trained	Noise added	Random	Scrubbed	All pre-trained
	(1)	(2)	(3)	(4)	(5)
Avg. log likelihood ratio (upstream bias)	0.292	0.073**	0.165	0.301***	0.184***
	(0.176)	(0.032)	(0.167)	(0.039)	(0.022)
Prevalance of toxic mentions	0.624***	0.558***	0.421***	0.688***	0.536***
	(0.180)	(0.034)	(0.039)	(0.046)	(0.021)
Prevalance of identity term		2.611		−1.074	2.682**
		(1.777)		(2.700)	(1.177)
Constant	0.008				
	(0.040)				
Observations	50	1,050	500	1,000	3,050
R^2	0.269	0.216	0.196	0.247	0.200
Adjusted R^2	0.238	0.198	0.178	0.230	0.183
Residual Std. Error	0.217 (df = 47)				
F Statistic	8.649*** (df = 2; 47)	94.264*** (df = 3; 1026)	59.616*** (df = 2; 488)	106.960*** (df = 3; 977)	248.602*** (df = 3; 2986)
Note:					*p<0.1; **p<0.05; ***p<0.01