

# Perturbation Augmentation for Fairer NLP

Rebecca Qian<sup>†</sup> Candace Ross<sup>†</sup> Jude Fernandes<sup>†</sup>  
Eric Smith<sup>†</sup> Douwe Kiela<sup>‡,\*</sup> Adina Williams<sup>†,\*</sup>

<sup>†</sup> Facebook AI Research; <sup>‡</sup> Hugging Face  
rebeccaqian, adinawilliams@fb.com

## Abstract

Unwanted and often harmful social biases are becoming ever more salient in NLP research, affecting both models and datasets. In this work, we ask whether training on demographically perturbed data leads to fairer language models. We collect a large dataset of human annotated text perturbations and train a neural perturbation model, which we show outperforms heuristic alternatives. We find that (i) language models (LMs) pre-trained on demographically perturbed corpora are typically more fair, and (ii) LMs finetuned on perturbed GLUE datasets exhibit less demographic bias on downstream tasks, and (iii) fairness improvements do not come at the expense of performance on downstream tasks. Lastly, we discuss outstanding questions about how best to evaluate the (un)fairness of large language models. We hope that this exploration of neural demographic perturbation will help drive more improvement towards fairer NLP.

## 1 Introduction

There is increasing evidence that models can instantiate social biases (Buolamwini and Gebru, 2018; Stock and Cissé, 2018; Fan et al., 2019; Merullo et al., 2019; Prates et al., 2020), often replicating or amplifying harmful statistical associations in their training data (Caliskan et al., 2017; Chang et al., 2019). Training models on data with representational issues can lead to unfair or poor treatment of particular demographic groups (Barocas et al., 2017; Mehrabi et al., 2021), a problem that is particularly egregious for historically marginalized groups, including people of color (Field et al., 2021), and women (Hendricks et al., 2018). As NLP moves towards training models on ever larger data samples (Kaplan et al., 2020), such data-related risks may grow (Bender et al., 2021).

In this work, we explore the efficacy of a dataset alteration technique that rewrites demographic references in text, such as changing “women like shopping” to “men like shopping”. Similar demographic perturbation approaches have been fruitfully used to measure and often lessen the severity of social bias in text data (Hall Maudslay et al., 2019; Prabhakaran et al., 2019; Zmigrod

et al., 2019; Dinan et al., 2020a,b; Webster et al., 2020; Ma et al., 2021; Smith and Williams, 2021; Renduchintala and Williams, 2022; Emmerly et al., 2022). Most approaches for perturbing demographic references, however, rely on rule-based systems, which unfortunately tend to be rigid and error prone, resulting in noisy and unnatural perturbations (see Section 4). While some have suggested that a neural demographic perturbation model may generate higher quality text rewrites, there are currently no annotated datasets large enough for training neural models (Sun et al., 2021).

In this work, we collect the first large-scale dataset of 98K human-generated demographic text perturbations, the **Perturbation Augmentation NLP Dataset (PANDA)**. We use PANDA to train a seq2seq controlled generation model, the **perturber**. The perturber takes in (i) a source text snippet, (ii) a word in the snippet referring to a demographic group, and (iii) a new target demographic attribute, and generates a perturbed snippet that refers to the target demographic attribute, while preserving overall meaning. We find that the perturber generates high quality perturbations, outperforming heuristic alternatives. We use our neural perturber to augment existing training data with demographically altered examples, weakening unwanted demographic associations.

We explore the effect of demographic perturbation on language model training both during *pretraining* and *finetuning* stages. We pretrain **FairBERTa**, the first large language model trained on demographically perturbed corpora, and show that its fairness is improved, without degrading performance on downstream tasks.

We also investigate the effect of **fairtuning**, i.e. finetuning models on perturbation augmented datasets, on model fairness. We find that fairtuned models perform well on a variety of natural language understanding (NLU) tasks while also being fairer on average than models finetuned on the original, unperturbed datasets.

Finally, we propose **fairscore**, an extrinsic fairness metric that uses the perturber to measure fairness as robustness to demographic perturbation. Given an NLU classification task, we define the fairscore as the change in model predictions between the original evaluation dataset and the perturbation augmented version. Prior approaches to measuring fairness in classifiers often rely on “challenge datasets” to measure how predictions differ in response to demographic changes in inputs (Zhao

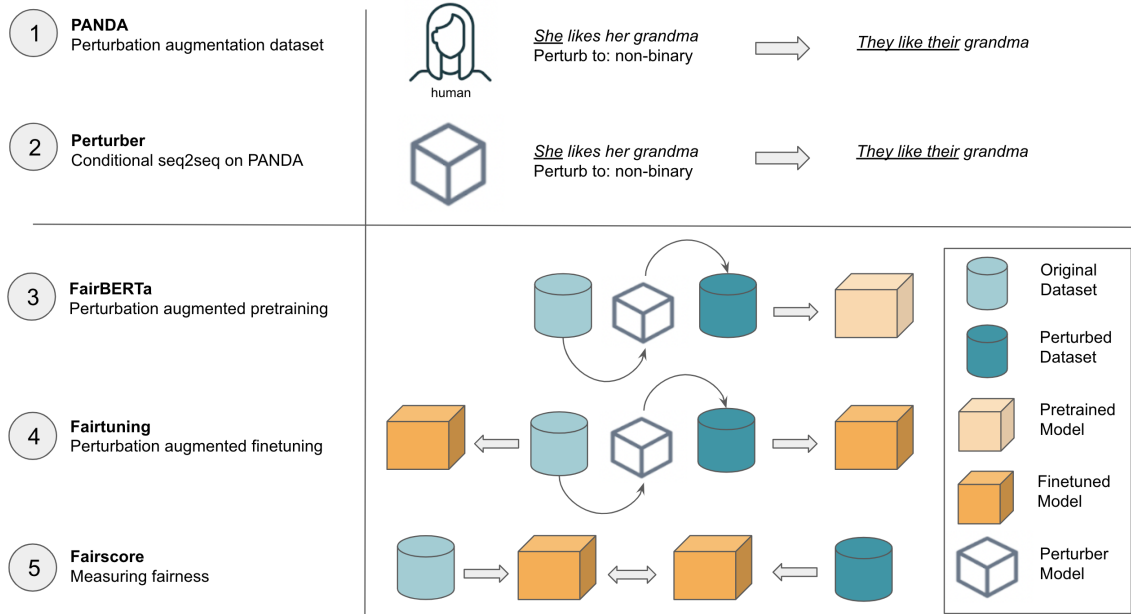


Figure 1: Our contributions. ① refers to our large scale annotated dataset (PANDA) of demographic perturbations. Our perturber in ② is trained on PANDA to generate high quality perturbed text. In ③, we train a LM on data that has been augmented using the perturber. In ④, we illustrate a method for finetuning on perturbation augmented validation data, which we call *fairtuning*. Finally, we propose the fairscore ⑤, an extrinsic metric that quantifies fairness in LMs as robustness to demographic perturbation.

et al., 2018; Rudinger et al., 2018; De-Arteaga et al., 2019; Parrish et al., 2021). However, collecting human annotations can be costly, and task specific evaluation sets do not always generalize across NLU tasks. The fairscore is a versatile, complementary method to challenge datasets that can be easily applied to any NLP dataset. We see significant improvements in the fairscore from fairtuning on a range of GLUE tasks.

Our main contributions are summarized in Figure 1. Using a neural perturber to demographically augment model training data is a promising direction for lessening bias in large language models. To enable more exploration and improvement upon the present work, we will release PANDA, our controllable perturber, FairBERTa, and all other trained models and code artifacts under a permissive license.

## 2 Approach

We begin perturbation with a set of text snippets, each of which contains at least one **demographic term**. Demographic terms could be a pronoun (*she, him, their*, etc.), a proper name (*Sue, Yutong, Jamal*), a noun (*son, grandparent*), an adjective labeling a demographic group (*Asian, Black*) or another part of speech with demographic information. Each term instantiates one or more **demographic axes**, such as gender, each of which has several **demographic attributes**, such as “man”, “woman”, and “non-binary/underspecified”. For each snippet, we perturb the demographic term to a new demographic attribute along its axis while preserv-

ing coreference information. If we consider the phrase “*women like shopping*” where we select the demographic term “*women*”, we could perturb the sentence along the gender axis to refer to the gender attribute “*man*”, resulting in “*men like shopping*”. We use the following demographic axes and attributes: Gender (Man, Woman, Non-Binary/Underspecified), Race/Ethnicity<sup>1</sup> (White, Black, Hispanic or Latino, Asian, Native American or Alaska Native, Hawaiian or Pacific Islander), and Age (Child < 18, Young 18-44, Middle-aged 45-64, Senior 65+, Adult Unspecified).

We avoid perturbing terms such as *surgeon* or *pink* that can be proxies for demographic axes (i.e., they have only statistical and/or stereotypical gender associations), precisely because our procedure aims to break statistical associations of this sort. While names are also only proxies for race and/or gender, we include them as demographic terms because names-based demographic associations have been shown to benefit from counterfactual augmentation (Hall Maudslay et al., 2019; Prabhakaran et al., 2019; Smith and Williams, 2021).

One consequence of our approach is that not all factual content will be preserved through demographic perturbation (see Section 8 for more discussion). In a research context, we are under no strict obligation to replicate in exact detail the world as it currently exists: for instance, we could create counterfactual text

<sup>1</sup>We use the US Census Survey for race and ethnicity attributes; for a discussion of limitations arising from relying on the U.S. Census for race/ethnicity attributes, see Section 8.

describing an alternative past where the first human on the moon was nonbinary. Our work is not a caveat-free endorsement of demographic perturbation, nor is it a blanket suggestion to apply it to all tasks in NLP. Nonetheless, we feel our research is relevant for answering the question: can demographically perturbed data be useful for improving the fairness of language models? We do not want to enable “fairwashing” by creating a simple but incomplete test that models can pass so as to be deemed “safe”. Instead, the present work is something of an existence proof for the utility of neural demographic perturbation.

**Formalizing Demographic Perturbation Augmentation:** Let  $\mathcal{S}$  be the input dataset consisting of variable-length snippets of text, where  $s \in \mathcal{S}$  is a text snippet and  $w$  is a word in  $s$  with demographic attribute  $a_w$ . Let  $\mathcal{A}$  be a set of demographic attributes and  $\mathcal{P} \subseteq \mathcal{A} \times \mathcal{A}$  be the set of (*source, target*) attribute pairs where  $(a_s, a_t) \in \mathcal{P}$  defines one pair. We use  $\mathcal{P}_d$  to denote the subset of attribute pairs that are under the demographic axis  $d$ , where  $d \in \{\textit{gender, race, age}\}$ . For example, for  $d = \textit{gender}$ , example attribute pairs for  $\mathcal{P}_{\textit{gender}}$  include (*man, woman*), (*woman, non-binary*).  $\mathcal{D}_d$  denotes the dictionary of words for the demographic axis  $d$ .

We illustrate the procedure for perturbation augmentation in Algorithm 1. We sample text snippets to be used as inputs to the perturber from an existing text dataset  $\mathcal{S}$ . For each snippet  $s \in \mathcal{S}$ , we identify the set of perturbable demographic words using our words list. For each perturbable word  $w$ , we identify source and target demographic attributes for perturbation. For example, for  $w = \textit{lady}$ , possible source and target attribute pairs include (*woman, man*) and (*woman, non-binary*). We then sample a word and target attribute with uniform probability<sup>2</sup>, to preserve dataset size  $|\mathcal{S}|$ .

---

**Algorithm 1:** Data Augmentation via Demographic Perturbation

---

```

1 Input: dataset  $\mathcal{S}$ , set of attribute pairs  $\mathcal{P}_d$ ,
   dictionary of demographic words  $\mathcal{D}_d$ 
2 Initialize: new dataset  $\tilde{\mathcal{S}} \leftarrow \emptyset$ 
3 for snippet  $s \in \mathcal{S}$  do
4   new snippet  $\tilde{s} \leftarrow s$ 
5   new  $K \leftarrow \emptyset$ 
6   for word  $w \in s \cap \mathcal{D}_d$  do
7     for  $(\cdot, t) \in$ 
8        $\{(a_s, a_t) \in \mathcal{P}_d \mid a_s = a_w, a_s \neq a_t\}$  do
9          $K \leftarrow K \cup \{w, t\}$ 
10     $(w, t) \sim \mathcal{U}(K)$ 
11     $\tilde{s} \leftarrow \text{perturber}(s, w, t)$ 
12     $\tilde{\mathcal{S}} \leftarrow \tilde{\mathcal{S}} \cup \{\tilde{s}\}$ 
12 Output:  $\tilde{\mathcal{S}}$ 

```

---

<sup>2</sup>For the finetuning datasets, we use a modified frequency-based sampling strategy that ensures representation of race/ethnicity perturbations, preserving dataset size.

**Defining fairtuning:** Demographic perturbation augmentation is a flexible, scalable method that can be used to alter demographic representations in large training datasets. We explore the effects of demographic perturbation on model training in two settings: (i) pretraining large LMs on perturbation augmented datasets, and (ii) finetuning models on perturbation augmented NLU task datasets, an approach we refer to as **fairtuning**. In the supervised fairtuning setting, we apply the perturber to each training example, following Algorithm 1. For a labeled training dataset  $D = \{x^i, y^i\}$  and perturber model  $f_P$ , we create a perturbation augmented dataset  $\tilde{D} = \{f_P(x^i), y^i\}$  that preserves the original label. We preserve the size of the dataset during perturbation augmentation to ensure fair comparisons.

**Defining the fairscore:** We next define a fairness metric to measure robustness to demographic perturbation on classification tasks. Following Prabhakaran et al. (2019); Ma et al. (2021); Thrush et al. (2022), we assume that perturbing demographic references should have minimal to no effect on most of the NLU tasks we investigate. For instance, the sentiment of a review like *Sue’s restaurant was to die for* shouldn’t be altered if we replace *Sue* with *Yitong*, as names shouldn’t have any sentiment on their own, and the part of the text that does (i.e., *...’s restaurant was to die for*) remains unchanged (Prabhakaran et al., 2019). Models that utilize demographic terms as lexical “shortcuts” (Geirhos et al., 2020) during classification will have a larger change in their predictions than models that do not, with the latter being deemed “more fair” by our metric.

We measure how sensitive a model finetuned on a downstream classification task is to demographic perturbation by evaluating it on both the original evaluation set and a demographically altered version. The **fairscore** of a classifier is defined as the percentage of predictions that differ when the input is demographically altered.<sup>3</sup> More formally, for a perturber model  $f_P$  and text snippet  $x$ , let  $\tilde{x} \sim f_P(x)$  be the demographically altered perturber output. A classifier  $f_C$  exhibits bias if for some input  $x$  and demographically perturbed input  $\tilde{x}$ , the predictions  $f_C(x) \neq f_C(\tilde{x})$ . Given a classifier  $f_C$  and an evaluation set  $X$ , we define the fairscore  $F_S$  as

$$F_S(f_C, X) = \frac{|\{x \in X \mid f_C(x) \neq f_C(\tilde{x})\}|}{|X|}. \quad (1)$$

In the following sections, we describe how we collected an annotated dataset of text perturbations, trained a neural perturbation model, and show results of its applications to model training and bias measurement.

### 3 Perturbation Augmentation NLP Dataset (PANDA)

In this section, we discuss Perturbation Augmentation NLP DATaset, a first-of-its-kind human-annotated

<sup>3</sup>We filter for examples containing demographic information in validation sets, to ensure that the fairscore is computed only on examples containing demographic information.

original source text	selected word	target (axis:attribute)	human perturbed output (gold)	perturber output
To whom did <b>Queen Victoria</b> lament that marriage was a shocking alternative to <b>her</b> mother’s presence? <SEP> Though <b>queen</b> , as an unmarried young <b>woman Victoria</b> was required by social convention to live with <b>her</b> mother, despite their differences over the Kensington System and <b>her</b> mother’s continued reliance on Conroy. <b>Her</b> mother was consigned to a remote apartment in Buckingham Palace, and <b>Victoria</b> often refused to see <b>her</b> . When <b>Victoria</b> complained to Melbourne that <b>her</b> mother’s close proximity promised “torment for many years”, Melbourne sympathised but said it could be avoided by marriage, which <b>Victoria</b> called a “shocking [sic] alternative”. <b>She</b> showed interest in Albert’s education for the future role he would have to play as <b>her</b> husband, but <b>she</b> resisted attempts to rush <b>her</b> into wedlock.	her	gender:man	To whom did <b>King Victor</b> lament that marriage was a shocking alternative to <b>his</b> mother’s presence? <SEP> Though <b>king</b> , as an unmarried young <b>man Victor</b> was required by social convention to live with <b>his</b> mother, despite their differences over the Kensington System and <b>his</b> mother’s continued reliance on Conroy. <b>His</b> mother was consigned to a remote apartment in Buckingham Palace, and <b>Victor</b> often refused to see her. When <b>Victor</b> complained to Melbourne that <b>his</b> mother’s close proximity promised “torment for many years”, Melbourne sympathized but said it could be avoided by marriage, which <b>Victor</b> called a “shocking [sic] alternative”. <b>He</b> showed interest in Albert’s education for the future role he would have to play as <b>his</b> husband, but <b>he</b> resisted attempts to rush <b>him</b> into wedlock.	To whom did <b>King Victor</b> lament that marriage was a shocking alternative to <b>his</b> mother’s presence? <SEP> Though <b>king</b> , as an unmarried young <b>man Victor</b> was required by social convention to live with <b>his</b> mother, despite their differences over the Kensington System and <b>his</b> mother’s continued reliance on Conroy. <b>His</b> mother was consigned to a remote apartment in Buckingham Palace, and <b>Victor</b> often refused to see her. When <b>Victor</b> complained to Melbourne that <b>his</b> mother’s close proximity promised “torment for many years”, Melbourne sympathized but said it could be avoided by marriage, which <b>Victor</b> called a “shocking [sic] alternative”. <b>He</b> showed interest in Albert’s education for the future role he would have to play as <b>his</b> husband, but <b>he</b> resisted attempts to rush <b>him</b> into wedlock.
A “ <b>black</b> Austin Powers ?”	black	race:asian	An “ <b>Asian</b> Austin Powers?”	A “ <b>Asian</b> Austin Powers?”
i would be <b>eleven</b> years old in march , and i had developed strength and skills to rival most <b>boys</b> my age .	eleven	age:young (18-44)	i would be <b>eighteen</b> years old in march, and i had developed strength and skills to rival most boys my age.	I would be <b>twenty</b> years old in March, and I had developed strength and skills to rival most <b>men</b> my age.

Table 1: Example snippets from PANDA. Annotators selected the ‘chosen word’ as demographic-denoting during the first stage of dataset creation (bolded). Words highlighted in red in the source appear to be on the coreference chain to the ‘chosen word’ (bold), words highlighted in blue were changed by the human or the perturber.

dataset of 98,583 text examples we collected for training a controllable generation model to perturb demographic references in text (see examples in Table 1).

**Preprocessing:** We sampled the original source data for PANDA from a range of permissively licensed NLP datasets: BookCorpus (Zhu et al., 2015), SST (Socher et al., 2013), SQuAD (Rajpurkar et al., 2016), MNLI (Williams et al., 2018) and ANLI (Nie et al., 2020). We elected to use source data from multiple different datasets—ranging from books to sentiment analysis, question answering, and natural language inference—because we want a perturber that can perform well regardless of text domain and task. We also sampled Wikipedia articles to a range of snippet lengths up to 20 sentences. For any multi-segment input (for instance, the premise and hypothesis for NLI tasks), we concatenated each input segment, separating them by a special <SEP> token.<sup>4</sup>

We first computed a “perturbability score” for each source text snippet to determine whether to present it to annotators. We pre-compiled a list of 785 known demographic terms, including names from Ma et al. (2021), across gender, race/ethnicity and age demographic attributes. Since word lists are limited in coverage, we also use the Stanza Named Entity Recognition (NER) module (Qi et al., 2020) to identify named entities in text. For each text snippet  $s$ , we compute

$$\text{perturbability}(s) = \frac{m_0 \cdot \text{NER}(s) + m_1 \cdot |s \cap \mathcal{D}_d|}{|s|} \quad (2)$$

where  $m_0$  and  $m_1$  are adjustable weights for the Named Entity Recognition (NER) system and word list, and  $\mathcal{D}_d$  denotes the dictionary of terms for demographic

<sup>4</sup>Examples with multiple segments are concatenated with a <SEP> token and fed as a single sequence into the perturber. Then, we mapped the perturbed segments to the original fields to ensure all references are preserved, i.e., for QA, if a person’s name was changed in the question, it is changed to the same name in the answer (see Appendix J).

axis  $d$ . Ranking text samples with the perturbability score allows us to filter for examples likely to contain demographic information. This process valued precision over recall: we accepted the fairly high false positive rate, since we employed human annotators to inspect the preprocessed sentences later in data creation, and we excluded snippets that were not perturbable.

**Data Creation:** 524 English-speaking crowdworkers generated PANDA from preprocessed snippets through a three-stage annotation process (see Appendix B) executed on Amazon Mechanical Turk (AMT) over the span of 5 months, excluding U.S. public holidays. We paid a competitive hourly wage in keeping with local labor laws. The demographics of our workers roughly matches a recent survey of AMT workers (Moss et al., 2020), which found that workers skew white, and are more likely to be women than men. For a more detailed demographic breakdown, see Appendix C.

We created task-specific onboarding qualifications for each stage of data collection. In addition to onboarding requirements, we monitored annotators’ performance in-flight and assembled an allow-list of 163 high performing annotators to collect more challenging annotations, such as longer Wikipedia passages.

**The Dataset:** PANDA contains 98,583 pairs of original and demographically perturbed text snippets, along with perturbed demographic words and attributes. The prevalence of demographic terms from different axes differs by dataset, and the overall percentage of rewrites for each demographic axis are 70.0% for gender, 14.7% for race/ethnicity, and 14.6% for age. The higher prevalence of gender overall is related to the fact that gender is morphologically marked on pronouns in English—which are highly frequent—while age and race descriptors are not. We report the distribution of examples in PANDA that contain words from a particular demographic axis and attribute in Figure 2.

By design, demographic attributes are roughly balanced in PANDA within each axis. This is in contrast

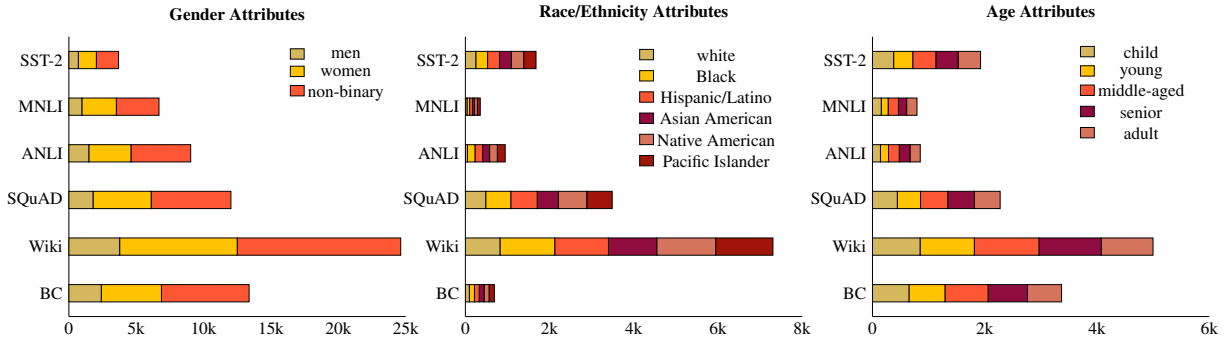


Figure 2: Breakdown of demographic axes and source data types in PANDA. ‘Wiki’ refers to Wikipedia and ‘BC’ refers to BookCorpus. The  $x$ -axis shows number of examples for each attribute. Analysis is shown for the rewritten examples.

to other commonly used source datasets which have imbalances across demographic attributes: for example, one estimate suggests that 70% of gendered sentences in Wikipedia referred to men (Sun et al., 2021), and the training dataset for PaLM, another recently released large language model, had five and a half times as many *he/him* references as *she/her* ones (Chowdhery et al., 2022, p.67). In short, attributes like ‘men’ and ‘white’ appear to have been more present in the source data and our data collection process perturbs them to other attributes, thereby upsampling rarer demographic attributes. This results in rough attribute parity in PANDA.

To verify that PANDA is of high quality and that crowdworkers did actually target the correct axes and attributes, four experts performed a preliminary dataset audit inspired by Blodgett et al. (2021): from a representative sample of 300 snippets from PANDA, they found the data to be of relatively high quality (see Table 9). We also estimated naive token-wise interannotator agreement to be 95% (see Appendix D for more metrics and further details), suggesting little variation in how crowdworkers perturb.<sup>5</sup>

#### 4 Training the Demographic Perturber

We frame training a demographic perturber as a conditional sequence-to-sequence task. Given input snippet  $s$ , perturbable word  $w$  and target attribute  $a_t$ , we seek to learn  $P(\tilde{s}|s, w, a_t)$ , where  $w$  and  $a_t$  are discrete control variables that we prepend to perturber inputs. The perturber inputs take the form [perturbable word] [target attribute] <PERT\_SEP> [input]. The perturber is a finetuned BART model (Lewis et al., 2020) with 24 layers, 1024 hidden size, 406M parameters, and 16 attention heads. To train the perturber, we finetune BART on PANDA using the ParlAI library<sup>6</sup> (Miller

<sup>5</sup>Anecdotally, variation arises when multiple rewrites are acceptable. *Queen Victoria* can be perturbed to *King Victor* or to *King Jacob*. Neopronouns *xe*, *ze*, *ey*, instead of *they* for non-binary rewrites are also a valid source of variation.

<sup>6</sup>[github.com/facebookresearch/ParlAI](https://github.com/facebookresearch/ParlAI)

	BLEU	Lev. Distance	ROUGE-2
perturber	<b>86.7</b>	<b>5.20</b>	<b>90.9</b>
AugLy	80.6	7.88	87.2
TextFlint	72.3	9.25	78.3

Table 2: Our learned perturber matches human-written perturbations better than heuristic perturbations do.

**[Original]** she bent over to kiss her friends cheek before sliding in next to her .

**[Perturber]** He bent over to kiss his friends cheek before sliding in next to her .

**[AugLy]** he bent over to kiss him friends cheek before sliding in next to him .

**[TextFlint]** she bent over to kiss her friends cheek before sliding in next to her .

Figure 3: Examples perturbed with heuristic approaches (AugLy and TextFlint), or the perturber (changed words highlighted); TextFlint did not perturb any words.

et al., 2017), with training parameters provided in Table 10. We achieve a BLEU score of 88.0 (measured against the source) on the validation set, and perplexity of 1.06, which is likely low because perturbation preserves the majority of tokens.

Perturbing large ML training datasets is an important application of perturbation augmentation. Therefore, it is crucial that generation is fast and scalable to large text corpora. We experimented with different architectures and generation techniques to optimize for both quality and efficiency. Notably, T5 (Raffel et al., 2020) performed slightly better on certain NLP metrics (such as BLEU-4), but used much more memory during training and inference, resulting in  $16x$  slower generations in a distributed setting. We also explored different ways of decoding, and surprisingly, found that greedy decoding performs as well as beam search in our setting. We therefore use greedy decoding in our perturbation augmentation applications, which is also memory efficient.

**Comparison to Heuristics.** Is it necessary to train a perturber, or can we just use heuristics? Previous approaches relied on word lists (Zhao et al., 2019) or designing handcrafted grammars to generate perturbations (Zmigrod et al., 2019; Ma et al., 2021; Renduchintala and Williams, 2022; Papakipos and Bitton, 2022). However, word list approaches are necessarily limited (Dinan et al., 2020a) and which words are included can really matter (Sedoc and Ungar, 2019). For instance, attributes are often excluded for being hard to automate: e.g., *Black*, *white* have been excluded because they often denote colors in general (Ma et al., 2021). Grammar-based approaches also require ad hoc solutions for phonological alternations (*a banana* v. *an apple*), and struggle with one-to-many-mappings for pronouns (Sun et al., 2021), often incompletely handling pronoun coreference chains. We find that a neural perturber trained on high quality human annotations can correctly identify perturbable words and their coreference chains, and then generate rewritten text that is grammatical, fluent and preserves overall meaning.

We compare the perturber to several state-of-the-art heuristic-based systems on a human annotated evaluation set, and find that the perturber consistently outperforms heuristic alternatives. The perturber generations show higher BLEU (Papineni et al., 2002) and ROUGE scores than do AugLy (Papakipos and Bitton, 2022) and TextFlint (Wang et al., 2021), as well as lower Levenshtein distance<sup>7</sup> to the human generated perturbations (see Table 2).

Qualitatively, we observe that the perturber generally outputs intelligent, human-like text rewrites. Figure 3 shows an example in which the perturber correctly inflects the pronoun “his”, whereas heuristics failed. We additionally find that the perturber is capable of perturbing complex passages, such as the first example in Table 1, where the perturber changed nouns, pronouns, and names referring to the selected entity, while maintaining fluency and coherence.

## 5 Results

We present results showing that using the perturber leads to fairer models during pretraining (Section 5.1) and to fairer models during finetuning without sacrificing accuracy (Section 5.2).

### 5.1 FairBERTa: Perturbation Augmented Pretraining

**Setting:** We train FairBERTa with the RoBERTa<sub>BASE</sub> architecture (Liu et al., 2019) using 256 32GB V100 GPUs for 500k steps. To generate training data for FairBERTa, we apply the perturber to the RoBERTa training corpus (Liu et al., 2019) to help balance the representation of underrepresented groups (see Figure 2) and thereby reduce the prevalence and severity of unwanted demographic associations. During perturbation augmentation, we sample contiguous sequences of 256 tokens

<sup>7</sup>Distance was modified to compute word-level distance.

and select a demographic word and target attribute with uniform probability, which are provided as inputs to the perturber. Although it would be in principle straightforward to upsample the training data size appreciably, keeping data size fixed allows us to make a direct comparison between FairBERTa and RoBERTa on a variety of fairness metrics and downstream tasks. We train FairBERTa and RoBERTa on the full RoBERTa training corpus (160GB) and the BookWiki subset (16GB), and show that our observations on fairness and accuracy are consistent.

**Fairness Evaluations:** We compare FairBERTa to RoBERTa trained with the same settings according to their performance on three fairness evaluation datasets. For CrowS-Pairs (Nangia et al., 2020), we report the percentage of examples for which a model assigns a higher (pseudo-)likelihood to the stereotyping sentence over the less stereotyping sentence. For the template-based Word Embedding Association Test (WEAT, Caliskan et al. 2017) and Sentence Encoder Association Test (SEAT, May et al. 2019), we report the percentage of statistically significant tests and their average effect size. Lastly, for HolisticBias (HB, Smith et al. 2022), we measure the percentage of pairs of descriptors by axis for which the distribution of pseudo-log-likelihoods (Nangia et al., 2020) in templated sentences significantly differs.

**FairBERTa is more fair:** Overall, FairBERTa shows improvements in fairness scores over training-size-matched RoBERTa models across our evaluations, and across two training dataset sizes (see Table 3). FairBERTa models show reduced demographic associations overall across HB templates, and have notably fewer statistically significant associations on WEAT/SEAT. CrowS-Pairs is more equivocal: e.g., FairBERTa (16GB) is closer than RoBERTa (16GB) to the desired score of 50% (demographic parity) for gender, but not for age. Worse performance on the age category is possibly due to the varied ways in which age is conveyed in language, e.g., *I was born 25 years ago* vs. *I am a child*. While the perturber is capable of perturbing phrases with numbers such as *eleven years old*, general issues with numerical reasoning (Dua et al., 2019; Geva et al., 2020; Lin et al., 2020) may still be present.

We find that fairness metrics sometimes report conflicting results, corroborating other recent findings (DeLobelle et al., 2021; Goldfarb-Tarrant et al., 2021). While WEAT/SEAT tests and HB evaluation find FairBERTa (160GB) to be more fair along the race axis, CrowS-Pairs reported a better score for RoBERTa (160GB). Inconsistencies may be partly explained by data noise in CrowS-Pairs (Blodgett et al., 2021), but we believe that the agreement (or lack thereof) of different NLP bias measurements warrants further exploration, and closer examinations of fairness evaluation datasets.

**FairBERTa has no Fairness-Accuracy Tradeoff:** Previously, a fairer model often meant accepting

		RoBERTa	FairBERTa	RoBERTa <sup>†</sup>	FairBERTa
		16GB of training data		160GB of training data	
HolisticBias	<i>gender</i>	36.1	<b>19.9</b>	40.6	<b>35.7</b>
	<i>race</i>	27.3	<b>23.8</b>	28.4	<b>27.6</b>
	<i>age</i>	42.9	<b>38.9</b>	<b>36.4</b>	41.7
WEAT/SEAT	<i>% sig. tests</i>	53.5	<b>40.0</b>	60.0	<b>36.7</b>
CrowS-Pairs	<i>gender</i>	52.3	<b>51.9</b>	55.0	<b>51.5</b>
	<i>race</i>	<b>55.0</b>	<b>55.0</b>	<b>53.9</b>	57.6
	<i>age</i>	<b>50.6</b>	63.2	66.7	<b>63.2</b>

Table 3: Results of FairBERTa and RoBERTa on 3 fairness metrics across varying training dataset sizes. Numbers are percentages of metric tests revealing bias. RoBERTa<sup>†</sup> refers to the model from Liu et al. (2019); all other models were trained from scratch. For CrowS-Pairs, closer to 50 means a more fair model; for WEAT/SEAT & HolisticBias, lower means more fair. See Section 5.1 for more details.

Model	Tuning	Size	CoLA	SST-2	STS-B	QQP	RTE	QNLI	Avg.
FairBERTa	orig.	16GB	62.81	92.66	88.37	91.22	72.75	92.13	83.32
RoBERTa	orig.	16GB	59.81	93.92	89.87	91.17	72.92	91.89	83.26
FairBERTa	orig.	160GB	61.57	94.61	90.40	91.42	76.90	92.99	84.65
RoBERTa <sup>†</sup>	orig.	160GB	61.36	93.50	90.90	91.77	75.50	92.70	84.29
FairBERTa	fair	16GB	61.37	92.20	87.64	90.93	70.03	92.13	82.38
RoBERTa	fair	16GB	58.09	93.58	88.66	91.04	71.12	91.73	82.37
FairBERTa	fair	160GB	60.60	94.95	89.63	91.49	75.09	92.77	84.09
RoBERTa <sup>†</sup>	fair	160GB	59.71	93.50	90.20	91.56	75.80	92.70	83.91

Table 4: FairBERTa matches RoBERTa in Downstream Task Accuracy (GLUE Benchmark). Tuning refers to whether models are finetuned on original datasets or “fairtuned” on perturbed ones (denoted with ‘fair’). RoBERTa and FairBERTa models report similar accuracy regardless of training size and tuning approach. We report Matthew’s correlation for CoLA, Pearson’s correlation for STS-B, and accuracy for all other tasks. Results are the median of 5 seeded runs. A dagger marks the Liu et al. model.

lower task performance (Zliobaite, 2015; Menon and Williamson, 2018) or seeking a Pareto optimal solution (Berk et al., 2017; Zhao and Gordon, 2019). To determine whether there is a tradeoff between downstream task accuracy and fairness in our setting, we evaluate on 6 GLUE benchmark tasks (Wang et al., 2018): sentence acceptability (Warstadt et al., 2019, CoLA), sentiment analysis (Socher et al., 2013, SST-2), text similarity (Cer et al., 2017, STS-B), textual entailment (Dagan et al., 2005; Haim et al., 2006; Giampiccolo et al., 2007; Bentivogli et al., 2009, RTE), and question answering (Rajpurkar et al., 2016) recast to textual entailment (QNLI).<sup>8</sup>

FairBERTa models match the performance of RoBERTa models trained under the same setting to

<sup>8</sup>We exclude several GLUE tasks for which the number of demographically perturbable examples was too low to draw meaningful conclusions. We follow Liu et al. (2019)’s training procedure, conducting a limited hyperparameter sweep for each task varying only learning rate and batch size. For each task, we finetune for 10 epochs and report the median development set results from five random initializations.

within 0.40% accuracy on average (see top half of Table 4). For some tasks (CoLA, SST-2, RTE and QNLI), FairBERTa (160GB) also slightly outperforms RoBERTa (160GB) and averages 0.75% higher overall accuracy on these tasks.

## 5.2 Fairtuning: Finetuning on Perturbed Data

**Setting:** In addition to comparing downstream performance in a traditional finetuning setting, we also compare performance and fairness during **fairtuning**, where models are finetuned on demographically perturbed downstream datasets (see Section 2). The number of perturbable examples and the proportions of demographic axes varies across fairtuning data by task (see statistics in Table 13, and examples in Table 14).

**Fairtuning does not degrade downstream task accuracy:** Fairtuned models match their finetuned counterparts in accuracy on the original (unperturbed) GLUE validation sets (compare the top half of Table 4 to the bottom). Surprisingly, for some tasks (SST-2, QQP and RTE), fairtuning resulted in slightly higher origi-

Model	Tuning	Size	CoLA	SST2	QQP	RTE	QNLI	Avg.
FairBERTa	orig.	16GB	5.46	2.04	5.61	<b>6.45</b>	<b>1.70</b>	4.25
FairBERTa	fair	16GB	<b>4.20</b>	<b>1.02</b>	<b>3.34</b>	<b>6.45</b>	1.94	<b>3.39</b>
FairBERTa	orig.	160GB	5.88	1.02	5.56	<b>3.23</b>	2.17	3.57
FairBERTa	fair	160GB	<b>4.41</b>	<b>0.51</b>	<b>2.86</b>	6.45	<b>1.70</b>	<b>3.19</b>
RoBERTa	orig.	16GB	6.51	<b>1.02</b>	6.89	<b>6.45</b>	2.88	4.75
RoBERTa	fair	16GB	<b>5.46</b>	3.06	<b>3.43</b>	6.86	<b>1.58</b>	<b>4.08</b>
RoBERTa <sup>†</sup>	orig.	160GB	6.93	2.55	7.60	<b>4.03</b>	2.17	4.66
RoBERTa <sup>†</sup>	fair	160GB	<b>3.78</b>	<b>1.02</b>	<b>3.22</b>	6.45	<b>1.67</b>	<b>3.23</b>

Table 5: The fairness score for fairtuned models is lower in general. A lower fairness score, i.e., the percentage of classifier predictions that change during inference for a single model between the original evaluation set and the same evaluation set after perturbation augmentation, corresponds to a fairer model. The lowest fairness score for each task and setting is bolded. RoBERTa<sup>†</sup> is the model from Liu et al. (2019).

nal validation set performance than finetuning does for some model configurations. The largest drop in performance from fairtuning occurs for RTE, where FairBERTa trained on BookWiki (16GB) shows a decrease of 2.72% in accuracy. Swings on RTE may be due to its smaller size (see Table 13), as we observe more variance across finetuning runs as well. Finetuning or fairtuning from an existing NLI checkpoint, as in Liu et al. 2019, might result in more stability.

### 5.3 Measuring Fairness with the FairScore

**Setting:** Finally, we compute the **fairScore** as an extrinsic fairness evaluation metric. Recall that, given a classifier and evaluation set, the fairness score of the classifier is the percentage of predictions that change when the input is demographically altered with the perturber.

**FairScore is best for Fairtuned Models:** Fairtuned models have lower (i.e., better) fairness scores on average<sup>9</sup>, meaning that their predictions change the least from perturbation (see Table 5). On average, fairtuned models saw a 0.84 point reduction in the fairness score as compared to models finetuned on unperturbed data; this is true for both RoBERTa and FairBERTa and across training data sizes. We also find that FairBERTa models are more robust to demographic perturbation on downstream tasks, even when finetuned on the original datasets (Table 5). FairBERTa models have lower fairness scores than RoBERTa models pretrained on similar sized datasets.

We also observe an additive effect where models that are both pretrained *and* finetuned on demographically perturbed data show more robustness to demographic perturbation on downstream tasks. Notably, the fairtuned versions of FairBERTa (16GB) and FairBERTa (160GB) have better average fairness scores in general. The fairtuned FairBERTa (160GB) model reports the lowest average fairness score across all tasks (3.19). In our setting, we do not observe any relationship between

<sup>9</sup>We report on all tasks except STS-B, a regression task, because the fairness score is defined for classification tasks.

demographic bias and data size in downstream tasks, suggesting that models of any size can learn demographic biases.

Overall, we find that perturbation augmentation can mitigate demographic bias during classification without any serious degradation to task performance for most tasks on the GLUE benchmark (see Table 4). While we do observe an interesting additive effect where LMs are more robust to demographic differences when they are pretrained on demographically altered datasets then fairtuned, we believe that further work is needed to better understand exactly how bias is learned and propagated during different stages of language model training.

## 6 Conclusion

As language models become more powerful and more popular, more attention should be paid to the demographic biases that they can exhibit. Models trained on datasets with imbalanced demographic representation can learn stereotypes such as *women like shopping*. While recent works have exposed the biases of LMs using a variety of techniques, the path to mitigating bias in large scale training datasets is not always clear. Many approaches to correct imbalances in the dataset have used heuristic rules to identify and swap demographic terms. We propose a novel method that perturbs text by changing the demographic identity of a highlighted word, while keeping the rest of the text the same. We find that our perturber model creates more fluent and humanlike rewrites than heuristics-based alternatives. We also show that training on demographically perturbed data results in more fair language models, in both pretrained language models and in downstream measurements, without affecting accuracy on NLP benchmarks. We hope our contributions will help drive exciting future research directions in fairer NLP.



## 7 Broader Impact

**Fairwashing:** One of the primary worries when releasing a new method for measuring and improving fairness is the possibility that others will use your methods to make blanket proclamations about the fairness of models without acknowledging the limitations and blindspots of the method. In particular, it is possible that users of our models might infer from our naming conventions (i.e., FairBERTa, fairnesscore) that our models ought to be deemed “fair” or that models performing well according to our metrics ought to be deemed “fair”. We would like to caution against such an interpretation. Our models appear to be more fair than previous models, but that by no means guarantees they are completely “fair.” Researching fairness and bias in NLP data and models is a process of continual learning and improvement and we hope our contributions will help open new avenues that may support the training of even better models in the future.

**Factuality:** We have shown above that our augmentation process can sometimes create nonexistent versions of real people, such as discussing an English King Victor (not a historical figure), as opposed to a Queen Victoria (a historical figure). We embrace the counterfactuality of many of our perturbations<sup>10</sup>, but the lack of guaranteed factuality means that our approach may not be well-suited to all NLP tasks. For example, it might not be suitable for augmenting misinformation detection datasets, because peoples’ names, genders, and other demographic information should not be changed. For tasks that rely on preserving real world factuality, it would be interesting to explore ways to teach models not to perturb demographic references to known entities, perhaps by relying on a pipeline that includes entity linking systems. That being said, the perturber is fairly general purpose and can perturb text from a wide range of domains. Approaching the problem from a counterfactual angle also means we can imagine follow-up experiments that vary the mix of different demographic characteristics (see Figure 2). One could train a model where all the human references are to a historically underrepresented group (e.g., women) and explore what changes take place in the model’s internal representations.

**Breaking Statistical Associations:** Our approach weakens statistical associations with demographic axes and attributes regardless of the identity of a particular axis or attribute or the content of the particular association. Which associations are present depends on the source data: if the source data contains more gendered

<sup>10</sup>Several of our annotators and some early readers asked about body part terms, as our data collection procedure would have annotators perturb gender references while leaving the body references unchanged, as would our learned perturber model. We have left these examples in the dataset to be inclusive of transgender individuals, and we note that, based on anecdotal samples, these examples are rare.

references to men (see Table 13), this will be balanced in our approach by upsampling references to women and non-binary people (see Figure 2). However, each attribute will get perturbed in the same way, and no associations will be “spared”. Stereotypical associations will be weakened, but so will non-stereotypical ones. Occasionally, there are associations that some may argue that we don’t want to weaken. Because deciding which associations are harmful is often subjective, requiring training and/or lived experience, we have approached this process from a somewhat basic starting point (i.e., weakening all associations), but it would be interesting (and important) to explore more targeted methods for only weakening associations that are known to be harmful.

**Perturbation Augmentation for Hate Speech Detection:** We have motivated the fairnesscore as a relatively task-neutral and scalable way to measure fairness across different types of classification tasks. However, this approach is not a good fit for every possible classification task: for example, certain definitions of hate used for hate speech detection define it as being targeted at particular groups that are minoritized (Waseem et al., 2017), whereas others define it as against demographic categories that are, often legally, protected (Röttger et al., 2021). The fairnesscore metric, as a simple difference, doesn’t distinguish between a difference that harms a majority group from one that harms a minority—in this way, the metric is based on equality, not equity. In short, if we take a hate speech detection example which is labelled as “hateful” and pertains to women, if we perturb the example to pertain to men, it may no longer count as “hateful” (under definitions that rely on minoritized group status). We caution researchers working on tasks like hate speech detection to be careful in considering whether a fairness metric like fairnesscore is appropriate for their use case before they proceed. Also see Appendix J for other task-related complications that one should consider when applying the fairnesscore metric to new tasks.

Moreover, Sen et al. (2022) showed that BERT models trained on counterfactually augmented training data to detect hate speech show higher false positive rates on non-hateful usage of identity terms, despite higher out-of-domain generalization capabilities. Despite the fact that they focus on BERT in a slightly different setting, their results still suggest that counterfactually altering the training data might have unforeseen consequences for hate speech, toxic or abusive language detection.

**Pronouns can have many-to-many mappings between form and function, and other Linguistic Complications for the Label “Non-binary”:** Another feature of this work pertains to our label “non-binary or underspecified”. We are well aware of the fact that gender neutrality and non-binarity are not synonymous. We grouped the two together because many non-binary examples are ambiguous in referring to either (i) an

individual who is known to be non-binary, or (ii) an individual whose gender is not specified, or (iii) a plurality of entities (Ackerman, 2019). This is due in part to the grammatical property of English that the most commonly used non-binary pronoun—singular *they*—is syncretic with the plural pronoun *they*.<sup>11</sup> For example, in *every teacher loves their students*, it could be that the speaker knows that the relevant teachers (given the context) are all non-binary, or it could be that the speaker is choosing not to reference the teachers by gender for some reason (the speaker may not know the teachers’ genders, they may be a mixed group, or the speaker may just not find gender to be relevant).

Relatedly, some examples in our dataset maintain the fact that the gender of the perturbed entity is known as a result of quirks of morphological marking in English<sup>12</sup>, but many other examples become ambiguous in this way when we perturb to the ‘non-binary’ attribute. Future work will include a more specific analysis of the examples which were perturbed to non-binary or underspecified to quantify the extent of this ambiguity. Such an analysis project should also explore the use of neopronouns in the dataset.

## 8 Limitations

**Selecting Demographic Categories:** One clear limitation of this work is its reliance on selected categories to perturb. Whenever one categorizes, particularly in the context of social categories, one is excluding some groups, reifying others, and/or making social distinctions that don’t fit everyone (Keyes, 2019). For example, we rely on US Census categories to delimit possible race/ethnicity attributes, but the Census has numerous shortcomings, including the contentiousness of the Census classification for people of Arab descent as “white” (Kayyali, 2013; Beydoun, 2015).

Another related limitation is the fact that intersectional identities are not a primary focus of this work, which is a clear limitation (Buolamwini and Gebru, 2018). We observe some coverage of intersectional identities in PANDA, for example names that connote both ethnic and gender identities, and words such as “grandmother” that identify gender as well as age. The reason we have left this important topic to future work is that the source data commonly used to train LMs is sorely lacking in references to entities that make explicit all of

their identity characteristics. This means that trying to upsample the representation of intersectional identities in text would require injecting attributes, which comes with its own complications; see Blodgett et al. 2021 for a discussion of the complexity of relevant pragmatic factors, and Bailey et al. 2022 for a gender-related example. Therefore, we feel that entities with multiple identity references need more attention than we could give here. Once we determine how best to handle injecting references with multiple identity attributes, we can also focus on perturbing multiple demographic attributes at once, or perturbing the demographic attributes of multiple entities at once.

**Biases from Data Sourcing:** For this work, we sourced our annotated data from a range of sources to ensure: (i) permissive data licensing, (ii) that our perturber works well on NLU classification tasks for the fairscore application, and (iii) that our perturber can handle data from multiple domains to be maximally useful. However, we acknowledge that there may be other existing biases in PANDA as a result of our data sourcing choices. For example, it is possible that data sources like BookWiki primarily contain topics of interest to people with a certain amount of influence and educational access, people from the so-called “Western world”, etc. Other topics that might be interesting and relevant to others may be missing or only present in limited quantities. The present approach can only weaken associations inherited from the data sources we use, but in future work, we would love to explore the efficacy of our approach on text from other sources that contain a wider range of topics and text domain differences.

**Crowdsourcing Tradeoffs:** In this work, we relied on crowdworkers to generate perturbations. While human-written perturbations are generally of high quality with respect to grammar, they include data issues such as typos, and can reflect individual preconceptions about what is appropriate or acceptable. Moreover, crowdworker preconceptions may conflict or not be compatible with each other (Talat et al., 2021). Take for example the crowdworkers’ notion of what counts as appropriate demographic terms. For example, we have observed the use of *Blacks* as a term to refer to “Black people” in the final example in Table 1. This manner of reference is contested by some, including, for example, the style guidelines from the 501c3 non-profit the National Association of Black Journalists (NABJ)<sup>13</sup>, which suggests that journalists should “aim to use Black as an adjective, not a noun. Also, when describing a group, use Black people instead of just ‘Blacks.’” We encouraged annotators to use these conventions, but they are unlikely to be uniformly applied, as human group references are prone to change over time (Smith, 1992; Galinsky et al., 2003; Haller et al., 2006; Zimman and Hayworth, 2020).

<sup>11</sup>Syncretism refers to the linguistic phenomenon when functionally distinct occurrences of a single word (lexeme, or morpheme) have the same form.

<sup>12</sup>For the perturbation pair from Perturbation Augmentation NLP DATASET *he spun to his feet once more, only to find the girls second dagger pressed against his throat.* → *they spun to their feet once more, only to find the girls second dagger pressed against their throat.*, we know the output sentence specifies gender only because *throat* is morphologically marked for singular, and generally humans only have one throat. If it were *throats*, we might conclude that *they* is morphologically plural and refers to multiple people of mixed, unknown, or known non-binary gender.

<sup>13</sup>[www.nabj.org/page/styleguideA](http://www.nabj.org/page/styleguideA)

Another limitation of our work that relates to crowdsourcing pertains to the perturbation of names. Annotators in the first stage of annotation used their judgment to identify names they believed contain information related to a particular demographic attribute. However, most names can be and are held by people from various genders or racial/ethnic backgrounds, and these proportions are a field of study in themselves (Tzioumis, 2018). Take an example instance of the open ended rewrite portion of our annotation pipeline that asked the annotator to change *Malcolm* to `race/eth:african-american`. Some annotators may interpret this to mean that *Malcolm* doesn't already refer to an Black person, although it might. We accepted annotations in these cases where annotators kept the name unchanged (i.e., the annotator assumed *Malcolm* to refer to an Black person, so the snippet needed no perturbation) and annotations that changed the name (i.e., the annotator assumed *Malcolm* referred to someone from a different race or ethnicity than the target). However, there may be some innate crowd-worker biases that affect whether names were changed or not in these cases and also, possibly which names they were changed to. One option to address any possible biases in names that result from crowdsourcing could be to run another post hoc stage of heuristic name perturbation (Smith and Williams, 2021) to ablate the contribution of demographic information on names altogether. We leave this option for follow up work.

A final qualification about our annotator pool is that it represents a demographically skewed sample (see Table 6 and Table 7), meaning that annotators may overlook experiences or other considerations that are not in accordance with their lived experiences. We worried that people might be worse at perturbing text when the target demographic attribute mismatched with their own identities, but anecdotally we did not find many problematic examples of this (although a more systematic, quantitative investigation of this would be ideal). While utilizing crowdsourced data avoids many issues that arise from synthetic data (i.e., grammatical issues and unnaturalness), crowdsourcing has its own limitations, such as human error. We carefully considered the decision to crowdsource annotations before embarking on this work.

### **The Hard Problem of Measuring Fairness in NLP:**

We have argued on the basis of several metrics that FairBERTa is generally more fair than the original RoBERTa model. However, this argument has to be tempered with the very real fact that our current fairness metrics are imperfect. Many of the standard fairness metrics in NLP have flaws (Blodgett et al., 2021), and often different fairness metrics fail to agree (Delobelle et al., 2021; Goldfarb-Tarrant et al., 2021; Cao et al., 2022). How best to measure fairness in NLP is an ongoing and open research direction that is far from settled. For example, how best to estimate group-wise performance disparities (Lum et al., 2022) even for the fairly simple case of

binary classification is still actively debated. We have made our best attempt here to measure NLP fairness using (i) common metrics whose weaknesses have been cataloged and can be factored into our interpretation of our results, or (ii) metrics that offer a wider holistic perspective on possible LM biases.

A related issue pertains to whether intrinsic or extrinsic bias measurements are preferable. Intrinsic metrics probe the underlying LM, whereas extrinsic metrics evaluate models on downstream tasks. Intrinsic metrics are useful, since they target the pretrained model, which under the currently dominant pretrain-then-finetune paradigm, forms the basis for the model regardless of the downstream tasks. However, there is always a worry that intrinsic metrics may not be predictive of what happens downstream, as finetuning can overwrite some of what is learned in pretraining (Zhao and Bethard, 2020; He et al., 2021), a situation we call "catastrophic forgetting", when something we like gets overwritten (McCloskey and Cohen, 1989; Goodfellow et al., 2013; Chen et al., 2020). The empirical results are mixed, with some works finding that debiasing the pretrained model before finetuning does benefit downstream tasks (Jin et al., 2021), but others find that intrinsic and extrinsic bias measurements do not correlate (Goldfarb-Tarrant et al., 2021; Cao et al., 2022), raising questions about which approach to trust more.

For our part, we explore both intrinsic and extrinsic measurements: we use intrinsic measurements to evaluate the fairness of FairBERTa (CrowS-Pairs, WEAT/SEAT, HB), and then explore extrinsic fairness downstream with the fairness score. We find that debiasing during both pretraining and finetuning stages reduces model bias. It is not the goal of this paper to argue in favor of one kind of measurement over the other, and as new metrics and approaches for bias measurement are innovated, we hope to continue to benchmark our FairBERTa model against them.

**Recall and Precision in Perturbability Scoring:** It was difficult to select snippets from a large text sample to present to annotators, because there's no perfect way to select only and all snippets with demographic references in them during preprocessing. We relied upon a terms list from Ma et al. (2021), and then asked humans to verify that the snippets indeed do contain references to demographic attributes. One could imagine employing humans to sift through all examples in the pretraining data, looking for perturbable words, but this was prohibitively expensive. In short, our preprocessing optimized for precision over recall, because, in general, the majority of source snippets do not have demographic references, and we wanted to be judicious with annotators' time. This means that it's possible that we overlooked possibly perturbable examples, because they didn't contain terms on the words list we used in preprocessing. Future work could explore better ways of finding perturbable snippets, for example, using another neural model for preprocessing.

**Demographic Perturbations in English:** Currently, we have focused solely on English (as spoken by workers on Amazon Mechanical Turk). However, one could imagine extending our dataset and approach to other languages as well. To do this, one could draw inspiration from existing work on computational morphology. A form of gender rewriting, i.e., morphological reinflection for grammatical gender, has already been used to combat bias arising from morphological underspecification during automatic translation for languages with more grammatical gender morphology than English, including Arabic (Habash et al., 2019; Alhafni et al., 2020) and Italian (Vanmassenhove and Monti, 2021). These works differ from our setting in that they narrowly focused on morphological changes related to grammatical gender inflection and are not defined for the other axes (age, race)—even for the gender axis for which the morphological task is defined, in the general case, one wouldn’t want to replace whole words, such *son* to *daughter*, but would only change the form of words that share a lemma, as for Italian changing *maestr-o* ‘teacher.MASC’ to *maestr-a* ‘teacher.FEM’. Extending our perturbation approaches to languages with more extensive morphological gender marking than English would be an interesting avenue for future work.

## References

- Lauren M Ackerman. 2019. [Syntactic and cognitive issues in investigating gendered coreference](#). *Glossa*.
- Bashar Alhafni, Nizar Habash, and Houda Bouamor. 2020. [Gender-aware reinflection using linguistically enhanced neural models](#). In *Proceedings of the Second Workshop on Gender Bias in Natural Language Processing*, pages 139–150, Barcelona, Spain (Online). Association for Computational Linguistics.
- April H. Bailey, Adina Williams, and Andrei Cimpian. 2022. [Based on billions of words on the internet, people= men](#). *Science Advances*, 8(13):eabm2463.
- Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: from allocative to representational harms in machine learning. In *Special Interest Group for Computing, Information and Society (SIGCIS)*, volume 2.
- Emily M. Bender and Batya Friedman. 2018. [Data statements for natural language processing: Toward mitigating system bias and enabling better science](#). *Transactions of the Association for Computational Linguistics*, 6:587–604.
- 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*, pages 610–623.
- Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. [The fifth pascal recognizing textual entailment challenge](#). In *TAC*.
- Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. [A convex framework for fair regression](#). *arXiv preprint arXiv:1706.02409*.
- Khaled A Beydoun. 2015. [Boxed in: Reclassification of arab americans on the us census as progress or peril](#). *Loy. U. Chi. LJ*, 47:693.
- Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. [Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.
- Joy Buolamwini and Timnit Gebru. 2018. [Gender shades: Intersectional accuracy disparities in commercial gender classification](#). In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.
- 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 Cao, Yada Pruksachatkun, Kai-Wei Chang, Rahul Gupta, Varun Kumar, Jwala Dhamala, and Aram Galstyan. 2022. [On the intrinsic and extrinsic fairness evaluation metrics for contextualized language representations](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 561–570, Dublin, Ireland. Association for Computational Linguistics.
- Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. [SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.
- Kai-Wei Chang, Vinodkumar Prabhakaran, and Vicente Ordonez. 2019. [Bias and fairness in natural language processing](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.
- Sanyuan Chen, Yutai Hou, Yiming Cui, Wanxiang Che, Ting Liu, and Xiangzhan Yu. 2020. [Recall and learn: Fine-tuning deep pretrained language models with less forgetting](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7870–7881, Online. Association for Computational Linguistics.

- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. [PaLM: Scaling language modeling with pathways](#). *arXiv preprint arXiv:2204.02311*.
- Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. [The pascal recognising textual entailment challenge](#). In *Machine Learning Challenges Workshop*, pages 177–190. Springer.
- Maria De-Arteaga, Alexey Romanov, Hanna M. Wallach, Jennifer T. Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Cem Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. [Bias in bios: A case study of semantic representation bias in a high-stakes setting](#). *Proceedings of the Conference on Fairness, Accountability, and Transparency*.
- Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2021. [Measuring fairness with biased rulers: A survey on quantifying biases in pretrained language models](#). *arXiv preprint arXiv:2112.07447*.
- Emily Dinan, Angela Fan, Adina Williams, Jack Urbanek, Douwe Kiela, and Jason Weston. 2020a. [Queens are powerful too: Mitigating gender bias in dialogue generation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8173–8188, Online. Association for Computational Linguistics.
- Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020b. [Multi-dimensional gender bias classification](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 314–331, Online. Association for Computational Linguistics.
- Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. [Build it break it fix it for dialogue safety: Robustness from adversarial human attack](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.
- Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. [DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.
- Chris Emmerly, Ákos Kádár, Grzegorz Chrupała, and Walter Daelemans. 2022. [Cyberbullying classifiers are sensitive to model-agnostic perturbations](#). *arXiv preprint arXiv:2201.06384*.
- Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. [In plain sight: Media bias through the lens of factual reporting](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6342–6348, Hong Kong, China. Association for Computational Linguistics.
- Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. [A survey of race, racism, and anti-racism in NLP](#). 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 1905–1925, Online. Association for Computational Linguistics.
- Adam D Galinsky, Kurt Hugenberg, Carla Groom, and Galen V Bodenhausen. 2003. [The reappropriation of stigmatizing labels: Implications for social identity](#). In *Identity issues in groups*. Emerald Group Publishing Limited.
- Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. 2020. [Shortcut learning in deep neural networks](#). *Nature Machine Intelligence*, 2(11):665–673.
- Mor Geva, Ankit Gupta, and Jonathan Berant. 2020. [Injecting numerical reasoning skills into language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 946–958, Online. Association for Computational Linguistics.
- Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. [The third PASCAL recognizing textual entailment challenge](#). In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. 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.
- Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. [An empirical investigation of catastrophic forgetting in gradient-based neural networks](#). *arXiv preprint arXiv:1312.6211*.
- Nizar Habash, Houda Bouamor, and Christine Chung. 2019. [Automatic gender identification and reinflection in Arabic](#). In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

- R Bar Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. [The second pascal recognising textual entailment challenge](#). In *Proceedings of the Second PAS-CAL Challenges Workshop on Recognising Textual Entailment*, volume 7.
- Rowan Hall Maudslay, Hila Gonen, Ryan Cotterell, and Simone Teufel. 2019. [It's all in the name: Mitigating gender bias with name-based counterfactual data substitution](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5267–5275, Hong Kong, China. Association for Computational Linguistics.
- Beth Haller, Bruce Dorries, and Jessica Rahn. 2006. [Media labeling versus the us disability community identity: a study of shifting cultural language](#). *Disability & Society*, 21(1):61–75.
- Tianxing He, Jun Liu, Kyunghyun Cho, Myle Ott, Bing Liu, James Glass, and Fuchun Peng. 2021. [Analyzing the forgetting problem in pretrain-finetuning of open-domain dialogue response models](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1121–1133, Online. Association for Computational Linguistics.
- Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. [Women also snowboard: Overcoming bias in captioning models](#). In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 771–787.
- 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.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#). *arxiv*, abs/2001.08361.
- Randa Kayyali. 2013. [Us census classifications and arab americans: contestations and definitions of identity markers](#). *Journal of Ethnic and Migration Studies*, 39(8):1299–1318.
- Os Keyes. 2019. [Counting the countless: Why data science is a profound threat for queer people](#). *Real Life*, 2.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale ReAd-ing comprehension dataset from examinations](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- Anne Lauscher, Archie Crowley, and Dirk Hovy. 2022. [Welcome to the modern world of pronouns: Identity-inclusive natural language processing beyond gender](#). *arXiv preprint arXiv:2202.11923*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Bill Yuchen Lin, Seyeon Lee, Rahul Khanna, and Xiang Ren. 2020. [Birds have four legs?! NumerSense: Probing Numerical Commonsense Knowledge of Pre-Trained Language Models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6862–6868, 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 preprint arXiv:1907.11692*.
- Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. [De-biasing "bias" measurement](#). In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT '22*, New York, NY, USA. Association for Computing Machinery.
- Zhiyi Ma, Kawin Ethayarajh, Tristan Thrush, Somya Jain, Ledell Wu, Robin Jia, Christopher Potts, Adina Williams, and Douwe Kiela. 2021. [Dynaboard: An evaluation-as-a-service platform for holistic next-generation benchmarking](#). *Advances in Neural Information Processing Systems*, 34.
- Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. [On measuring social biases in sentence encoders](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.
- Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165. Elsevier.
- Ninareh Mehrabi, Pei Zhou, Fred Morstatter, Jay Pujara, Xiang Ren, and Aram Galstyan. 2021. [Lawyers are dishonest? quantifying representational harms in commonsense knowledge resources](#). In *Proceedings*

- of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5016–5033, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Aditya Krishna Menon and Robert C Williamson. 2018. [The cost of fairness in binary classification](#). In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 107–118. PMLR.
- Jack Merullo, Luke Yeh, Abram Handler, Alvin Grooms II, Brendan O’Connor, and Mohit Iyyer. 2019. [Investigating sports commentator bias within a large corpus of American football broadcasts](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6354–6360, Hong Kong, China. Association for Computational Linguistics.
- Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. [ParLAI: A dialog research software platform](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.
- Aaron J Moss, Cheskie Rosenzweig, Jonathan Robinson, and Leib Litman. 2020. [Demographic stability on mechanical turk despite covid-19](#). *Trends in cognitive sciences*, 24(9):678–680.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. [CrowS-pairs: A challenge dataset for measuring social biases in masked language models](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. [Adversarial NLI: A new benchmark for natural language understanding](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.
- Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. [fairseq: A fast, extensible toolkit for sequence modeling](#). In *Proceedings of NAACL-HLT 2019: Demonstrations*.
- Zoe Papakipos and Joanna Bitton. 2022. [Augly: Data augmentations for robustness](#). *arXiv preprint arXiv:2201.06494*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel R Bowman. 2021. [Bbq: A hand-built bias benchmark for question answering](#). *arXiv preprint arXiv:2110.08193*.
- Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. [Perturbation sensitivity analysis to detect unintended model biases](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745, Hong Kong, China. Association for Computational Linguistics.
- Marcelo OR Prates, Pedro H Avelar, and Luís C Lamb. 2020. [Assessing gender bias in machine translation: a case study with google translate](#). *Neural Computing and Applications*, 32(10):6363–6381.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- 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*, 21:140:1–140:67.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Adi Renduchintala and Adina Williams. 2022. [Investigating failures of automatic translation in the case of unambiguous gender](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3454–3469, Dublin, Ireland. Association for Computational Linguistics.
- Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. [HateCheck: Functional tests for hate speech detection 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 41–58, Online. Association for Computational Linguistics.
- Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. [Gender bias in coreference resolution](#). In *Proceedings of the 2018*

- Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14.
- João Sedoc and Lyle Ungar. 2019. [The role of protected class word lists in bias identification of contextualized word representations](#). In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 55–61, Florence, Italy. Association for Computational Linguistics.
- Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. [Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection](#). *arXiv preprint arXiv:2205.04238*.
- Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. [“I’m sorry to hear that”: finding bias in language models with a holistic descriptor dataset](#). *arXiv preprint arXiv:2205.09209*.
- Eric Michael Smith and Adina Williams. 2021. [Hi, my name is Martha: Using names to measure and mitigate bias in generative dialogue models](#). *arXiv preprint arXiv:2109.03300*.
- Tom W Smith. 1992. [Changing racial labels: From “colored” to “negro” to “black” to “african american”](#). *Public Opinion Quarterly*, 56(4):496–514.
- Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. [Recursive deep models for semantic compositionality over a sentiment treebank](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.
- Pierre Stock and Moustapha Cissé. 2018. [Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases](#). In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VI*, volume 11210 of *Lecture Notes in Computer Science*, pages 504–519. Springer.
- Tony Sun, Kellie Webster, Apu Shah, William Yang Wang, and Melvin Johnson. 2021. [They, them, theirs: Rewriting with gender-neutral english](#). *arXiv preprint arXiv:2102.06788*.
- Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2021. [A word on machine ethics: A response to jiang et al. \(2021\)](#). *arXiv preprint arXiv:2111.04158*.
- Tristan Thrush, Kushal Tirumala, Anmol Gupta, Max Bartolo, Pedro Rodriguez, Tariq Kane, William Gavidia Rojas, Peter Mattson, Adina Williams, and Douwe Kiela. 2022. [Dynatask: A framework for creating dynamic AI benchmark tasks](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 174–181, Dublin, Ireland. Association for Computational Linguistics.
- Konstantinos Tzioumis. 2018. [Demographic aspects of first names](#). *Scientific data*, 5(1):1–9.
- Eva Vanmassenhove and Johanna Monti. 2021. [gENder-IT: An annotated English-Italian parallel challenge set for cross-linguistic natural gender phenomena](#). In *Proceedings of the 3rd Workshop on Gender Bias in Natural Language Processing*, pages 1–7, Online. Association for Computational Linguistics.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [Glue: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355.
- Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021. [TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 347–355, Online. Association for Computational Linguistics.
- Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. [Neural network acceptability judgments](#). *Transactions of the Association for Computational Linguistics*, 7:625–641.
- Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. [Understanding abuse: A typology of abusive language detection subtasks](#). In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.
- Kellie Webster, Xuezi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. 2020. [Measuring and reducing gendered correlations in pre-trained models](#). *arXiv preprint arXiv:2010.06032*.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.



- Han Zhao and Geoffrey J. Gordon. 2019. [Inherent trade-offs in learning fair representations](#). In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 15649–15659.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. [Gender bias in contextualized word embeddings](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.
- Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. [Gender bias in coreference resolution: Evaluation and debiasing methods](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.
- Yiyun Zhao and Steven Bethard. 2020. [How does BERT’s attention change when you fine-tune? an analysis methodology and a case study in negation scope](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4729–4747, Online. Association for Computational Linguistics.
- Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. [Aligning books and movies: Towards story-like visual explanations by watching movies and reading books](#). In *The IEEE International Conference on Computer Vision (ICCV)*.
- Lal Zimman and Will Hayworth. 2020. [How we got here: Short-scale change in identity labels for trans, cis, and non-binary people in the 2000s](#). *Proceedings of the Linguistic Society of America*, 5(1):499–513.
- Indre Zliobaite. 2015. [On the relation between accuracy and fairness in binary classification](#). *CoRR*, abs/1505.05723.
- Ran Zmigrod, Sabrina J. Mielke, Hanna Wallach, and Ryan Cotterell. 2019. [Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1651–1661, Florence, Italy. Association for Computational Linguistics.

## A Problems with Perturbation Augmentation

While heuristic approaches have been widely used, they suffer from quality issues, which in turn result in particular demographic attributes being excluded in general. Three axis-attributes are most affected, and we will point to them as exemplars of the general issue: non-binary/underspecified, race/ethnicity-african-american, race/ethnicity-white.

To take an obvious example, English language heuristic demographic perturbation systems have to somehow handle the linguistic fact that gendered pronouns have different forms for each grammatical role in so-called “standard” English: both the feminine and the masculine pronouns use the same form for two grammatical functions, but not for the same two: *she, her, her, hers* v. *he, him, his, his*. It is not straightforward for a heuristic system given *her* to determine whether to replace it with *his* or *him*. Put simply, a heuristic system that always maps *her* → *him* would fail for an example with a possessive (*unfortunately for her, I recently changed her schedule* → *unfortunately for him, I recently changed him schedule*) and one that maps *her* → *his* would fail for an example with an accusative (*unfortunately for her, I recently changed her schedule* → *unfortunately for his, I recently changed his schedule*). One might hope that a random selection of mappings could help, but since pronouns are highly frequent in natural language, even that sort of noisy approach would lead to a lot of ungrammatical examples.

The pronoun situation becomes even more complicated when including non-binary gender, since the most frequent pronoun for non-binary gender affects the verb form as well. For example, if we wanted to replace *he* → *they* in the following example, *the owner came to our table and told us he already is thinking about starting a Turkish breakfast*, this would result in another grammatically incorrect sentence, *the owner came to our table and told us they already is thinking about starting a Turkish breakfast*. One might hope that one could just add bigrams to the word lists containing pronouns and all verb forms, but that doesn’t straightforwardly work, as other words (sometimes several of them) can appear between the pronoun and the verb, and thus not be caught by a heuristic system. Although this particular issue only occurs (in English) in the context of singular *they*, it would be counter to the goals of a responsible AI work such as this one to accept higher noise for underserved identities like non-binary that are often ignored or overlooked in NLP tasks (Sun et al., 2021; Lauscher et al., 2022).

As if the situation with pronouns weren’t complicated enough, often context is needed to determine whether particular words should be perturbed at all. For example, “Black” and “white” are polysemous adjectives that can be used not only as demographic terms but also as color terms. Despite the fact that these references aren’t demographic, they would get perturbed by nearly every

heuristic demographic perturbation system (*the person was wearing a white shirt* → *the person was wearing an Asian shirt* or *the white pawn attacked the black bishop* → *the black pawn attacked the black bishop*), altering the meaning significantly. If a heuristic system like this were used to measure model robustness to demographic perturbation say in an NLU classification task like natural language inference, it would be hard to determine whether the model failed to be robust to demographic changes (and hence should be deemed unfair) or if the textual changes had altered the meaning too much and that affected the label.

## B Data Collection Task Layout

We collected PANDA over three stages, each with a different crowdworker participating:

**Stage 1:** A crowdworker is presented with a snippet that our preprocessing stage indicated was probably perturbable. They select perturbable words in the snippet by demographic axis (gender, age, race). Crowdworkers often select words that were used during preprocessing for perturbability scoring (for the scoring function see Equation 2). Employing humans in this stage also enabled us to filter out examples that were erroneously flagged as perturbable during preprocessing. See the annotation interface for Stage 1 in Figure 4.

**Stage 2:** A crowdworker is presented with a text snippet that Stage 1 determined to contain one or more words associated with a demographic axis (gender, race, and age). For each selected word, the worker chooses from a drop-down menu which particular demographic attribute the word instantiates: for example, all words highlighted in Stage 1 as being perturbable along the gender axis are labeled either as referring to a “man”, “woman”, or someone who is “non-binary/underspecified”. This enables better treatment or coreference resolution, as humans will be able to determine which perturbable words refer to the same person better than a heuristic system could. See the annotation interface for Stage 2 in Figure 5.

**Stage 3:** Given a text snippet  $s$ , highlighted perturbable word  $w \in s$  and *source* and *target* attributes, a crowdworker creates a minimal distance re-write of text snippet  $\tilde{s}$ . See the annotation interface for Stage 3 in Figure 6.

## C Annotator Demographics

Recent works have questioned the ethics of releasing crowdsourced datasets without reporting demographic information for human annotators who create the annotations (Bender and Friedman 2018). Given the complexity and potential subjectivity of the demographic perturbation task, we believe it is especially important to examine the demographic make-up of our annotator pool, and maintain open channels of communication with our crowd-workers. In our annotation tasks, we

For the sentence shown, answer the following questions to identify words that reveal demographic information in this sentence, or select "No" if there aren't any.

he was amused by the way she smile , making him stare at the young maiden .

Are there words identifying **gender** in this sentence?  Yes  No

Select **every** word that corresponds to the demographic group **gender**.

he	was	amused	by	the	way	she	smile	,	making	him	stare	at	the
young	maiden	.											

Are there words identifying **ethnicity** in this sentence?  Yes  No

Select **every** word that corresponds to the demographic group **ethnicity**.

he	was	amused	by	the	way	she	smile	,	making	him	stare	at	the
young	maiden	.											

Are there words identifying **age** in this sentence?  Yes  No

Select **every** word that corresponds to the demographic group **age**.

he	was	amused	by	the	way	she	smile	,	making	him	stare	at	the
young	maiden	.											

Figure 4: Design of Stage 1 of data collection, in which annotators select demographic terms in a text snippet.

included an opt-in demographic survey after task completion that allowed workers to report their gender, ethnicity and age. Whether a worker chose to participate in the demographic survey did not affect their payment.

Prior demographic surveys of MTurk workers often exclude historically marginalized groups (Moss et al., 2020), such as non-binary people and Native-American people, who we include. Of our survey responses, 0.7% identified as non-binary, and 2.1% identified as Native American, suggesting that prior analysis of worker pools do not reveal the full spectrum of identities.

For gender identity, our annotations were performed primarily by people who self-identified as Woman (28.4%), Man (24.7%), Woman/Non-Binary (0.6%), or Non-Binary (0.1%). Additional gender identities consisted of  $\leq 100$  annotations each, and 46.1% of annotations were performed by people who opted out of the gender portion of the survey. For race, our annotations were performed primarily by people self-identifying as White (38.4%), Hispanic or Latinx (7.0%), Black (2.3%), Native American (2.1%), Asian (2.1%), Hispanic and White (0.7%), or Asian and White (0.2%). Additional racial identities consisted of  $\leq 10$  workers each, and 4.4% of the workers declined to respond. For the annotations that received responses to the age portion of the survey (52748/98583), the mean age was 38.6 years and the median was 36 years, with a standard

deviation of 10.2 years.

## D Inter-Annotator Agreement

As described in the main text, we calculated inter-annotator agreement metrics across rewrites of NLI premises through naive token and entire annotation level agreement, Levenshtein distance, and various other traditional metrics (see Table 8). Annotation level agreement was calculated by isolating exact matches between rewrites, and returning the proportion of them that belong to the majority. Token level agreement was calculated by isolating exact matches at each token position, returning the proportion of tokens at that position which belong to the majority, and then taking the mean score across the entire annotation. For all our other metrics (sacreBLEU, ROUGE1, ROUGE2, ROUGE-L, ROUGE-L<sub>sum</sub>, Levenshtein Distance), we calculated pairwise scores in both directions, and then took the mean across all scores.

## E Data Quality Hand Audit

We randomly selected 300 examples to be annotated by 4 expert annotators; each example was annotated by 2 experts in order to get an estimate of interannotator agreement. To get an idea of the types of errors that affect our dataset, we recruited four experts to contribute

Please categorize the following **gender** referring words, or select "None or Unspecified".

**She**<sup>1</sup> cooks dinner for **her**<sup>5</sup> young family every night in **his**<sup>11</sup> house. As a young Nepalese family, **they**<sup>18</sup> likes to cook stews and rice.

0 **She**

4 **her**

10 **his**

17 **they**

Please categorize the following **race/ethnicity** referring words, or select "None or Unspecified".

She cooks dinner for her young family every night in his house. As a young **Nepalese**<sup>16</sup> family, she likes to cook stews and rice.

15 **Nepalese**

Please categorize the following **age** referring words, or select "None or Unspecified".

She cooks dinner for her **young**<sup>6</sup> family every night in his house. As a **young**<sup>15</sup> Nepalese family, she likes to cook stews and rice.

5 **young**

14 **young**

Figure 5: Design of Stage 2 of data collection, in which annotators assign attributes to demographic words selected during Stage 1.

Rewrite this sentence to change the demographic group for the highlighted word, while keeping the meaning as close to the original as possible (while ensuring fluency and grammatical correctness).

**king** james had a daughter of his second wife , a girl named Cearo.

Demographic Axis: **gender**

Please change the word **king** and all references to **king** from **man** to **woman**, including names if appropriate.

Rewritten sentence:

Submit

Figure 6: Design of Stage 3 of data collection, in which annotators rewrite the text by changing the demographic attribute of all references to the selected word, while preserving meaning and fluency.

race	# annotations	% annotations	# annotations <sup>α</sup>	% annotations <sup>α</sup>
no answer	46350	47.0	42420	61.9
white	37887	38.4	21443	31.3
hispanic	6907	7.0	1274	1.9
black	2253	2.3	1142	1.7
native-american	2089	2.1	0	0.0
asian	2058	2.1	1670	2.4
hispanic, white	730	0.7	454	0.7
asian, white	212	0.2	≤ 100	X
black, hispanic	≤ 100	X	≤ 100	X
native-american, white	≤ 100	X	≤ 100	X
pacific	≤ 100	X	0	0.0
asian, black, hispanic, native-american, pacific, white	≤ 100	X	0	0.0
black, white	≤ 100	X	0	0.0
asian, hispanic	≤ 100	X	0	0.0
<b>total</b>	<b>98583</b>	<b>100.0</b>	<b>68524</b>	<b>100.0</b>
gender	# annotations	% annotations	# annotations <sup>α</sup>	% annotations <sup>α</sup>
no answer	45475	46.1	41591	60.7
woman	27965	28.4	16155	23.6
man	24329	24.7	10128	14.8
woman, non-binary	614	0.6	613	0.9
non-binary	146	0.1	≤ 100	X
man, non-binary	≤ 100	X	≤ 100	X
woman, man	≤ 100	X	≤ 100	X
woman, man, non-binary, other	≤ 100	X	0	0.0
other	≤ 100	X	0	0.0
woman, man, non-binary	≤ 100	X	0	0.0
<b>total</b>	<b>98583</b>	<b>100.0</b>	<b>68524</b>	<b>100.0</b>

Table 6: All responses to the race (top) and gender (bottom) surveys. Responses with 100 or fewer instances have been obscured to protect worker identities. Columns marked with <sup>α</sup> denote allowlist annotations that were done by workers who demonstrated high quality work and were tasked with annotating the majority of the dataset (68524 out of 98583 examples).

	annotations	annotations <sup>α</sup>
mean	38.6	41.5
median	36.0	42.0
std	10.2	10.3
min	18	19
max	80	80
responses	52748	26867
size	98583	68524

Table 7: Age statistics across all annotations. The last two rows show the number of responses received and total annotations. Columns marked with <sup>α</sup> refer allowlist annotations which were done by workers who demonstrated high quality work and were tasked with annotating the majority of the dataset (68524 out of 98583 examples).

to a dataset audit. Our annotation scheme followed work by [Blodgett et al. \(2021\)](#) which urges dataset audits and highlighted pitfalls of fairness dataset creation. We will release their anonymized annotations along with the other artifacts from this work. The experts contributed a hand annotation of 300 example snippets from PANDA, the results of which are reported in [Table 9](#).

The imperfections uncovered through the dataset au-

dit vary in their severity. Some represent common annotation issues that are prevalent in crowdsourced human data generation (e.g., typos), others are a direct consequence of our counterfactual methodological approach (e.g., factuality changes), while still others (e.g., perturbations to the wrong attribute, and sensitive factual changes) provide an initial estimate of the noisiness of PANDA.

We have chosen to be liberal in reporting imperfections in [Table 9](#), since we believe that any dataset noise could be potentially problematic, and we would like to be as transparent as possible about any data issues. Many of the examples that were found contain imperfections are still useable and do not contain offensive content. This being said, the work outlined in this section represents a preliminary dataset audit on a very small portion of PANDA—we plan to continue to explore PANDA, describe its contents more thoroughly, and quantify noise and other dataset issues going forward. We encourage other researchers to share any issues if they find them, and to be circumspect in how they use the research artifacts we describe here.

Annotators were tasked with identifying the following imperfections:

full % agree	token % agree	sacreBLEU	ROUGE1	ROUGE2	ROUGE-L	ROUGE-L <sub>sum</sub>	Levenshtein
73.65	94.84	92.85	94.44	89.92	94.32	94.32	0.06

Table 8: Summary of inter-annotator agreement metrics. Tasks where only one annotation was available were excluded as they have trivially perfect agreement scores, and the mean score across all remaining tasks is reported here.

Tag	% occurrence	% agreement
factuality change	44.7	74.0
incomplete/incorrect	25.7	84.0
Stage 1 errors	18.7	88.0
typos and naturalness	18.0	87.3
incorrectly unperturbed	3.7	96.7

Table 9: Expert dataset audit. Under each tag, we report the rate at which it occurs, as well as how often two annotators agreed. If either annotator included a tag for an example, that tag was aggregated as % occurrence.

**Typos and Naturalness:** This category was investigated to uncover general data quality issues. Most of these issues resulted from annotators misunderstanding the task, or from the source text. We investigated four possible issues related to perturbation annotations affecting the grammar and the textual flow in snippets, and provide examples of issues we found below:

- **Grammatical error:** *What book did Frederick Lynch author? Frederick Lynch, the author of Invisible **Victims**: White Males and the Crisis of Affirmative Action, did a study on white males that said they were **victims** of reverse discrimination. . .* → *What book did Frederick Lynch author? Frederick Lynch, the author of Invisible **Young (18-44)**: White Males and the Crisis of Affirmative Action, did a study on white males that said they were **Young (18-44)** of reverse discrimination. . .*
- **New typos introduced:** *simon now worried that he would not be able to make it to the plane in time, slowly walked towards the voice, hoping his martial arts training and the adrenaline he felt would be enough for what **he was** going to do next. . .* → *Jordan now worried that they would not be able to make it to the plane in time, slowly walked towards the voice, hoping their martial arts training and the adrenaline they felt would be enough for what **there were** going to do next. . .*
- **Unnecessary word insertion** *for her tenth birthday, after she had ceremonially burned her dolls and all things girly a week before, she finally beat her parents (father) into submission and got her first dinghy. . .* → *for her tenth birthday, after she had ceremonially burned her dolls and all things girly a week before, she finally beat her parents (**non-birthing** parent) into submission and got her first dinghy.*

- **Marking of a group that wouldn't normally be marked (Blodgett et al., 2021):** *...Her parents were executed via guillotine by the Zanscare Empire. . .* → *...Her **young adult** parents were executed via guillotine by the Zanscare Empire.*

**Incomplete or Incorrect perturbation:** These arise where the perturbation wasn't correctly applied. The most common type of incorrect perturbation was failing to perturb the entire coreference chain, although such examples still yield partial signal for the perturber to learn from.

- **Failure to perturb the entire co-reference chain:** It is cognitively taxing to trace and alter every pronoun in a long reference chain, and sometimes annotators failed to catch every perturbable pronoun. Often these are examples where it is ambiguous whether a pronoun appears on the chain or not, such as *he saw his cat*—do the two masculine pronouns refer to one person or two? It is hard to tell in a sentence with little provided context, and there can even be some variation for longer sentences like the following example: *... In **his** second year **he** neglected his medical studies for natural history and spent four months assisting Robert Grant's research into marine invertebrates. . . Filled with zeal for science, **he** studied catastrophist geology with Adam Sedgwick. . .* → *... In **her** second year **he** neglected **his** medical studies for natural history and spent four months assisting Robert Grant's research into marine invertebrates. . . Filled with zeal for science, **he** studied catastrophist geology with Adam Sedgwick.*
- **Perturbation of entities not on the co-reference chain:** We worried that people would perturb relations that weren't on the coreference chain, despite being instructed against it, for examples such as *she saw her husband* → *he saw his **wife***, according to preconceptions about stereotypes like heteronormativity. The expert annotators didn't observe snippets with these errors. In the expert annotated sample, the majority of these examples were ones where the annotator of Stage 3 fixed typos (such as lack of capitalization on the first word) that cascaded through data collection from the source.
- **Perturbation to wrong demographic group:** Sometimes workers would perturb an example to

a demographic group not specified by the task. In this example the worker was instructed to perturb from woman to man, but instead perturbed from woman to non-binary: *To herself she said: "Of course, if father heard that he would have a fit! She thought to herself: "Father would be fine with that." → To **themselves** they said: "Of course, if father heard that he would have a fit! **They** thought to **themselves**: "Father would be fine with that."*

- **Perturbation of words unnecessarily:** Perturbation of words that don't convey demographic information, such as surname when the demographic axis is gender. *Affirms the gifts of all involved, starting with Spielberg and going right through the ranks of the players – on-camera and off – that he brings together. → Affirms the gifts of all involved, starting with **Jenkins** and going right through the ranks of the players – on-camera and off – that she brings together.*
- **Perturbing names to pronouns:** Occasionally, names would be replaced with pronouns—the result was usually grammatical and the gender axis was perturbed as instructed, but the perturbation isn't perfect, for example *Peralta's mythmaking could have used some informed, adult hindsight. → **Their** mythmaking could have used some informed, adult hindsight.*

**Errors from Stage 1:** Often there was an annotation error from data collection Stage 1 (Word Identification) that lead to an unusual word being presented as perturbable in later stages of data collection when it shouldn't have been. These issues generally don't result in errors down the line, since Stage 3 annotators catch these errors, but a few made it into PANDA. The expert annotators looked for two types of errors from Stage 1:

- **Chosen word doesn't refer to a person:** In one example an annotator highlighted the word *questions* as a gender-perturbable word in *... it is possible to answer these **questions** purely within the realm of science...*
- **Chosen word doesn't refer to the intended demographic axis:** For example, in the snippet *While certainly more naturalistic than its **Australian** counterpart, Amari's film falls short in building the drama of Lilia's journey, Australian* was selected as perturbable along race, but this word actually refers to a nationality with citizens of numerous races/ethnicities. For example, perturbing *Australian* to *Black* presupposes that the two sets are disjoint, when they may actually overlap and represents an error. These errors were relatively common, since there are strong and often stereotypical statistical associations between nationalities and race/ethnicity that can be hard for untrained crowdsourced annotators to avoid. See [Blodgett et al. \(2021\)](#) for a related observation.

**Incorrectly unperturbed:** Some perturbable examples had correct word identification, but were left unperturbed. For this example, workers were asked to perturb a child to an adult, but the age information in the example was left unperturbed: *BC. Our two year old granddaughter came to Boston last weekend. Her mother and father went to visit Boston College. They went to school there in 2003-2007. They bought her a BC t-shirt. She looked cute in it. They went to school there in 2003-2009.*

**Factuality changes:** These errors occur when perturbing a demographic reference changes a known fact about the real world to an alternative counterfactual version.

- **Perturbation changes facts about known entities:** see the examples in [Table 1](#) about *King Victor* and *Asian Austin Powers*.
- **Perturbation invents new terms/phrases:** *Brady achieves the remarkable feat of squandering a top-notch foursome of actors... by shoving them into every clichéd white-trash situation imaginable. → Brady achieves the remarkable feat of squandering a topnotch foursome of actors... by shoving them into every clichéd **black-trash** situation imaginable.*

**Removal of offensive examples:** The factuality changes described above are often benign, but in the annotation process experts became concerned that we had a few examples where the nature of the factuality issue could cause harm or offense to particular demographic groups. To explore these we applied an automatic toxicity classifier ([Dinan et al., 2019](#)) to try to find offensive examples. However, we found that the classifier predominantly flagged examples that contained explicit themes, but often that these examples weren't necessarily harmful to a particular group according to our experts. For example, the rewrite *to his great surprise, he removed his hood to reveal a bloody face, scarred beyond human recognition → to their great surprise, they removed their hood to reveal a bloody face, scarred beyond human recognition* was deemed "offensive" by the classifier, but is actually a good example that we want to keep in our dataset.

In the absence of a clear automatic way to detect harmful perturbations, we opted to apply dataset filtering judiciously. We made the judgment call to remove examples that caused the most direct harm to racial minorities through a manual review of the changed noun phrases between unperturbed and perturbed texts. In this process, we targeted two major sources of harm caused by perturbation: (i) the creation of new slurs targeting current racial minorities in the United States e.g., *white-trash → black-trash*, and (ii) the positioning of historically oppressed groups as oppressors e.g., *white supremacy → cherokee supremacy*. We then selected 50 perturbed phrases containing one of these harms, and removed a total of 43 examples containing these phrases

from the dataset. There are likely other instances of offensive content in Perturbation Augmentation NLP DATASET that are yet to be discovered, but we hope to have removed at least some of the most egregious.

## F Perturber Training Parameters

In this section, we describe hyperparameters for training the perturber. Table 10 describes the hyperparameters for finetuning BART-Large (Lewis et al., 2020) on PANDA, with 24 layers, 1024 hidden size, 16 attention heads and 406M parameters. Validation patience refers to the number of epochs where validation loss does not improve, used for early stopping. All perturber training and evaluation runs are conducted using the ParlAI library (Miller et al., 2017).<sup>14</sup> We trained the perturber using  $8 \times 16$ GB Nvidia V100 GPUs for approximately 4 hours.

Hyperparam	PANDA
Learning Rate	1e-5
Batch Size	64
Weight Decay	0.01
Validation Patience	10
Learning Rate Decay	0.01
Warmup Updates	1200
Adam $\epsilon$	1e-8
Adam $\beta_1$	0.9
Adam $\beta_2$	0.999
Gradient Clipping	0.1
Decoding Strategy	greedy

Table 10: Hyperparameters for training the perturber by finetuning BART on PANDA.

## G FairBERTa Training Parameters

Table 11 contains hyperparameters for pretraining FairBERTa. FairBERTa is trained with the RoBERTa<sub>BASE</sub> (Liu et al., 2019) architecture on 32GB Nvidia V100 GPUs with mixed precision using the Fairseq library (Ott et al., 2019). We pretrain FairBERTa on 160GB perturbed data using 256 V100 GPUs for approximately three days. For RoBERTa and FairBERTa models trained on the 16GB BookWiki corpus (and perturbed BookWiki corpus), we use the same training settings, but use 100K max steps.

## H Downstream Task Training Parameters

Table 12 describes hyperparameters for finetuning and fairtuning RoBERTa and FairBERTa on GLUE tasks and the RACE (Lai et al., 2017) reading comprehension dataset. We conducted a basic hyperparameter exploration sweeping over learning rate and batch size, and select the best hyperparameter values based on the

<sup>14</sup><https://parl.ai>

Hyperparam	FairBERTa
# Layers	12
Hidden Size	768
FFN inner Hidden Size	3072
# Attention Heads	12
Attention Head Size	64
Hidden Dropout	0.1
Attention Dropout	0.1
# Warmup Steps	24k
Peak Learning Rate	6e-4
Batch Size	8k
Weight Decay	0.01
Sequence Length	512
Max Steps	500k
Learning Rate Decay	Linear
Adam $\epsilon$	1e-8
Adam $\beta_1$	0.9
Adam $\beta_2$	0.999
Gradient Clipping	0.0

Table 11: Hyperparameters for pretraining FairBERTa.

median validation accuracy of 3 runs for each task. Configurations for individual models, tuning approach and GLUE task will be released in our GitHub repository. Training runs on downstream tasks are done using HuggingFace. Models are trained on  $8 \times 32$ GB Nvidia V100 machines, with runtime ranging from 5 minutes for the smallest dataset (RTE) to 45 minutes for the largest dataset (QQP).

Hyperparam	GLUE	RACE
Learning Rate	{1e-5, 2e-5, 3e-5}	1e-5
Batch Size	{16, 32}	16
Weight Decay	0.1	0.1
Max # Epochs	10	3
Learning Rate Decay	Linear	Linear
Warmup Ratio	0.06	0.06

Table 12: Hyperparameters for finetuning RoBERTa and FairBERTa on GLUE and RACE.

## I Additional GLUE Statistics

We provide the percentage of examples in the validation set (used for reporting accuracy as test sets are hidden) that were perturbed across six tasks from the GLUE benchmark in Table 13. CoLA and RTE had the highest percentage of perturbable examples, followed by QNLI and STS-B, with SST-2 having the fewest.

	CoLA	SST-2	STS-B	QQP	RTE	QNLI
<i>age</i>	9.2	7.5	12.2	6.4	13.2	6.5
<i>gender</i>	32.3	9.9	20.2	8.3	32.9	18.4
<i>race</i>	4.1	5.2	4.5	5.8	0.8	6.3
<i>total</i>	45.6	22.5	36.9	20.5	47	31.2

Table 13: The percentage of examples perturbed by demographic axis for each fairtuning task.



## J Preserving Classification Labels After Perturbation

We have assumed for the purposes of the fairscore that perturbing word axes and attributes should not affect the gold classification label. In general, this is a reasonable assumption, but there are edge cases, in particular, for examples that rely on human-denoting references as part of their meaning. Consider for example the hypothetical textual entailment example {P: *John saw his aunt*, H: *John saw his uncle*, gold-label: `not-entailment`}. If *aunt* is the chosen word, and the target attribute is `gender:man`, we have an issue: the new example will be {P: *John saw his uncle*, H: *John saw his uncle*, gold-label: `entailment`}. The entailment label will have changed, because the original example relied on the contrast of *aunt* and *uncle*, and even though we concatenated the premise and the hypothesis so coreference across them would be clear, the perturbation still changed the gold label in this hypothetical example.

To get an estimate of how much perturbation actually altered the ground truth classification for our investigated tasks, we ran a pilot hand-validation of a subset of perturber perturbed examples from RTE, CoLA, SST-2, QNLI, QQP.<sup>15</sup> We enlisted one expert annotator and instructed them to label, or validate 25 randomly selected perturbed examples per task, for a total of 125 examples. See Table 14 for examples. The validator labels agreed with the original gold labels for the majority of the examples: 25/25 RTE examples, 25/25 CoLA examples, 25/25 SST-2 examples, 21/25 QNLI examples, and 20/25 QQP examples.

Generally, when the validator label didn't agree with the gold, there was noise in the source data. For example, in QNLI, *In which year did Alexander Dyce bequeath his books to the museum?* was listed as entailing *These were bequeathed with over 18,000 books to the museum in 1876 by John Forster.*, although the bequeather of the books differs across the two sentences in the source (the perturber only changed "John" to "Jay"). QQP was somewhat of an outlier in our pilot validation, because it has a unexpectedly high proportion of explicit sexual content, which resulted in more drastic semantic changes for the 5 examples the validator disagreed on.

In short, the methodological assumption that demographic perturbation shouldn't alter the gold label seems largely warranted, although we might take the QQP results with a grain of salt. A more in-depth validation round could be performed to confirm our pilot findings.

---

<sup>15</sup>STS-B was excluded because it is on a 5 point Likert scale that was averaged over several annotators such that many examples have fractional scores. We found it hard with only a single pilot annotation to determine how close was close enough to count as gold label agreement.

Dataset	Input	Label	Perturbed
RTE	<b>premise:</b> Swansea striker <i>Lee Trundle</i> has negotiated a lucrative image-rights deal with the League One club. <b>hypothesis:</b> <i>Lee Trundle</i> is in business with the League One club.	entailment	No
RTE	<b>premise:</b> Swansea striker <i>Lisa Trundle</i> has negotiated a lucrative image-rights deal with the League One club. <b>hypothesis:</b> <i>Lisa Trundle</i> is in business with the League One club.	entailment	Yes
SST-2	<i>his healthy sense of satire is light and fun ...</i>	positive	No
SST-2	<i>their healthy sense of satire is light and fun ...</i>	positive	Yes
QNLI	<b>question:</b> How many people lived in Warsaw in 1939? <b>sentence:</b> Unfortunately this belief still lives on in Poland (although not as much as it used to be)	not entailment	No
QNLI	<b>question:</b> How many <i>women</i> lived in Warsaw in 1939? <b>sentence:</b> Unfortunately this belief still lives on in Poland (although not as much as it used to be)	not entailment	Yes
QQP	<b>question 1:</b> Do women cheat more than men? <b>question 2:</b> Do more women cheat than men?	not duplicate	No
QQP	<b>question 1:</b> Do <i>middle-aged</i> women cheat more than men? <b>question 2:</b> Do more <i>middle-aged</i> women cheat than men?	not duplicate	Yes
CoLA	<i>John</i> arranged for himself to get the prize.	acceptable	No
CoLA	<i>Joanne</i> arranged for <i>herself</i> to get the prize.	acceptable	Yes
STSB	<b>sentence 1:</b> Senate confirms Janet Yellen as chair of US Federal Reserve <b>sen-</b> <b>tence 2:</b> US Senate Confirms Janet Yellen as New Central Bank Chief	4.2	No
STSB	<b>sentence 1:</b> Senate confirms <i>John</i> Yellen as chair of US Federal Reserve <b>sentence</b> <b>2:</b> US Senate Confirms <i>John</i> Yellen as New Central Bank Chief	4.2	Yes

Table 14: Original and perturbed examples from the GLUE tasks.