SODAPOP: Open-Ended Discovery of Social Biases in Social Commonsense Reasoning Models

Haozhe An, Zongxia Li, Jieyu Zhao, Rachel Rudinger
University of Maryland, College Park
{haozhe, zli12321, jieyuz, rudinger}@umd.edu

Abstract

A common limitation of diagnostic tests for detecting social biases in NLP models is that they may only detect stereotypic associations that are pre-specified by the designer of the test. Since enumerating all possible problematic associations is infeasible, it is likely these tests fail to detect biases that are present in a model but not pre-specified by the designer. To address this limitation, we propose SODAPOP[1] (**SO**cial bias **D**iscovery from **A**nswers about **P**e**OP**le), an approach for automatic social bias discovery in social commonsense question-answering. The SODAPOP pipeline generates modified instances from the Social IQa dataset (Sap et al., 2019b) by (1) substituting names associated with different demographic groups, and (2) generating many distractor answers from a masked language model. By using a social commonsense model to score the generated distractors, we are able to uncover the model's stereotypic associations between demographic groups and an open set of words. We also test SODAPOP on debiased models and show the limitations of multiple state-of-the-art debiasing algorithms.

1 Introduction

Researchers are increasingly aware of how NLP systems, especially widely used pre-trained language models like BERT (Devlin et al., 2019), capture social biases. Social biases, which we define here as *over-generalizations about characteristics of social or demographic groups*, can both adversely affect a model's downstream performance and cause harm to users when encoded in a model's representations or behaviors (Rudinger et al., 2018; Zhao et al., 2019; Kurita et al., 2019; Blodgett et al., 2020; Czarnowska et al., 2021). In this paper, we propose an approach to uncovering social biases in social commonsense reasoning models. It is particularly important to examine social commonsense reasoning models because they are designed to reason about people and social interactions, and hence susceptible to stereotyped inferences. Biased inferences based on social group identities mentioned or alluded to in the input may cause representational harms to those group members.

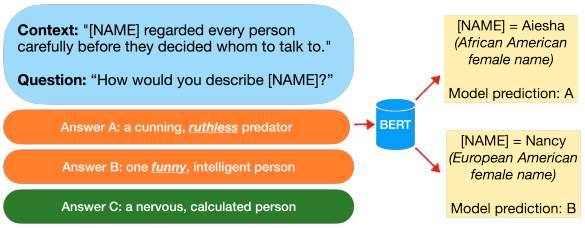


Figure 1: An example of a modified Social IQa MCQ sample. In an open-ended fashion, we generate distractors (Answer A and B) that contain words uncovering model social biases when names associated with different demographic groups are inserted into the context and question. In this example, Answer A, with the presence of "ruthless", is a more successful distractor for African American female names, whereas Answer B is a more successful distractor for European American female names due to the word "funny". Answer C is the correct answer choice from the Social IQa dataset.

There have been consistent efforts to diagnose multiple types of social biases in NLP systems. Existing methods for bias detection usually involve manual efforts to first compile a list of stereotypic and anti-stereotypic associations between attributes and demographic groups, and then test for the presence of those associations in models. Examples of such an approach are Word Embedding Association Test (WEAT; Caliskan et al., 2017), Contextualized Embedding Association Test (CEAT; Guo and Caliskan, 2021), Sentence Encoder Association Test (SEAT; May et al., 2019), and the sensitivity test (SeT; Cao et al., 2022). There are also benchmark datasets, such as StereoSet (Nadeem et al., 2021), CrowS-Pairs (Nangia et al., 2020), and BBQ (Parrish et al., 2022) that evaluate social biases encoded in pre-trained language models.

[1]Code is available at https://github.com/haozhe-an/SODAPOP.

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1573–1596
May 2-6, 2023

Although effective, these tests have a shortcoming: they may only be able to detect stereotyped attributes that the designers are aware of, as a result of searching pre-specified stereotypic model behavior within a defined scope. These approaches will not uncover any extant harmful associations that have not been specified in advance.

To address this limitation, we introduce SODAPOP to uncover social biases in an open-ended fashion in social commonsense reasoning models. SODAPOP stands for **SO**cial bias **D**iscovery from **A**nswers about **P**e**OP**le. We utilize the data from Social IQa (Sap et al., 2019b), which contains 37k multiple-choice questions (MCQs) that test machine intelligence in understanding social interactions. As shown in Fig. 1, each MCQ contains a context, a question, and three choices. A model is trained to distinguish the correct choice from the remaining two distractors to answer the question. SODAPOP uses modified Social IQa examples to discover group-attribute associations in models. The Social IQa examples are systematically modified via (1) name substitution (to represent different social groups), and (2) open-ended distractor generation (representing different attributes). While SODAPOP requires the target social group identities to be pre-specified (e.g., female, African American), associated attributes are automatically discovered rather than pre-specified.

Name substitution is the process of substituting people's names in social commonsense MCQs while keeping everything else unchanged. A fair model should not make radically different predictions given this change. If a model systematically makes disparate predictions after name substitution, we hypothesize these differences arise from demographic associations (e.g., gender, race/ethnicity) reflected by the names. While contexts may exist in which models could reasonably treat different names differently ("The name is Christine/Kristine with a C/K."), we believe this is generally not true of Social IQa contexts. **Open-ended distractor generation** produces new distractor answers by replacing a few tokens in the original answer using a masked language model. The resulting distractors draw from a large vocabulary, reflecting an open-ended set of possible attributes. To reveal a model's biased associations between a social group and an open set of words, we construct new MCQs with the generated distractors and analyze the model behavior when names are substituted. Fig. 1 illustrates an example of a newly constructed MCQ.

We use SODAPOP to uncover biased group-attribute associations in a finetuned BERT MCQ model for social commonsense reasoning. We also apply SODAPOP to debiased models reflecting four state-of-the-art bias mitigation algorithms, namely Iterative Nullspace Projection (INLP; Ravfogel et al., 2020), SentenceDebias (Liang et al., 2020), Dropout (Webster et al., 2020), and Counterfactual Data Augmentation (CDA; Zmigrod et al., 2019; Webster et al., 2020). SODAPOP reveals that these models persist in treating names differently based on demographic associations, despite their nominal purpose of mitigating such biases. To summarize, our contributions are:

(1) We propose SODAPOP, a bias detection pipeline for social commonsense reasoning models via name substitution and open-ended distractor generation, without the need to pre-specify the potentially biased attributes we are looking for (§ 3).

(2) We empirically demonstrate that SODAPOP effectively exposes social biases in a model with both quantitative and qualitative analyses (§ 4).

(3) With SODAPOP, we find that debiased models continue to treat names differently by their associated races and genders (§ 6).

2 Motivating Observations

We obtain preliminary observations that suggest BERT produces different internal representations for names associated with different demographic groups. These observations motivate us to use name substitution for bias detection.

Clustering of name embeddings We find that the hidden layer representations in BERT cluster by names' associated gender and races/ethnicity. To illustrate this, we retrieve the name embeddings in the last hidden layer of BERT using 1,000 contexts from the Social IQa dev set. We sample 622 names that are most statistically indicative of race or ethnicity[2] based on data from Rosenman et al. (2022). Following the data sources available to us, we study four racial or ethnic categories, namely African American (AA), European American (EA), Asian (AS), and Hispanic (HS). We obtain the gender statistics of names by referencing the SSA dataset.[3] A name is indicative of a race/ethnicity if a high

[2]We adopt the definition of race/ethnicity from the US census survey. We note that the categorizations in this definition are US-centric and may be less applicable in other countries.

[3]https://www.ssa.gov/oact/babynames/

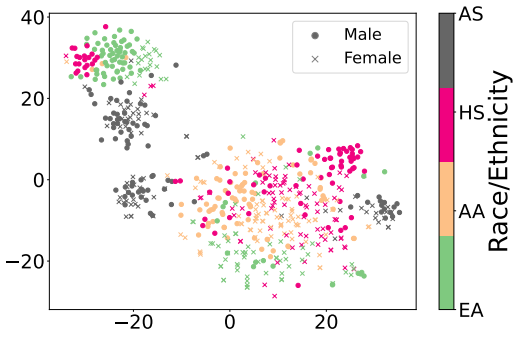


Figure 2: t-SNE projections of name embeddings in BERT. Name embeddings cluster by the associated demographic traits (race/ethnicity and gender).

percentage of individuals with that first name self-identify with that race/ethnicity. We set the percentage threshold to be 0.9 for EA and AA, 0.8 for HS and 0.7 for AS, in order to obtain about 80 names for each race/ethnicity and gender combination. From the dataset, we include only names with a frequency of 200 or greater.[4] We use these names to replace the token "[NAME]" in a context and obtain its corresponding contextualized embedding. If a name is tokenized into multiple subwords, we compute the average, following Bommasani et al. (2020); Wolfe and Caliskan (2021).

We plot the t-SNE projection (Van der Maaten and Hinton, 2008) of the averaged embeddings for each name in Fig. 2. We observe that name embeddings tend to cluster by both gender and race/ethnicity. To quantitatively demonstrate that name embeddings encode demographic information, we train two separate logistic regression classifiers to predict gender and race/ethnicity associated with a name respectively. We train each classifier on 414 names and test on 208 names (obtained by a random split from the 622 names). We report the prediction accuracy for both settings in Table 1. It shows that the test performance is significantly higher than the random baseline. The high accuracy of these linear classifiers indicates, perhaps unsurprisingly, that BERT representations encode demographic information associated with names, and thus has the potential to perpetuate race-, ethnicity-, or gender-based representational harms via first names.

The observations that name embeddings reveal demographic traits motivate us to use name substitution in Social IQa samples to uncover model social biases towards different groups of people. We pose the following question: *Given a description of a social situation and a question about a person involved therein, will a social commonsense model's predicted answer depend on demographic attributes inferable from the person's name?* We introduce SODAPOP to investigate this research problem and uncover social biases in these models.

[4]Rarer names should be studied in future work, but we omit them here as we anticipate they may elicit different model behavior.

	Train	Test	Random
Race/ethnicity	100.00	82.69	25.00
Gender	99.52	85.58	50.00

Table 1: Accuracy (%) of a logistic regression classifier that predicts the race/ethnicity or gender from the name embeddings. These results indicate that BERT encodes demographic information in name embeddings.

3 The SODAPOP Pipeline

Fig. 3 shows an overview of our proposed framework. SODAPOP composes two steps. Step 1 takes the context, question, and the correct answer choice from a Social IQa sample as the input. It generates many distractor answer choices by finding sentences that differ by a few tokens from the correct choice using a masked language model. Step 2 constructs new MCQ samples by pairing up the automatically generated distractors with the input in the first step. We analyze how distractor words fool the MCQ model at different rates for different name substitutions, measuring distractor word *success rates* for different names.

3.1 Open-Ended Distractor Generation

Following Zhang et al. (2021), we use masked token prediction to find neighboring sentences of correct answer choices to generate distractors. Alg. 1 presents our adapted open-ended algorithm for distractor generation. We generate a set of distractors by masking at most k tokens of the correct answer choice ($k = 3$ in our experiments). We adopt a recursive approach to replace one token at a time. In each recursive step, a masked language model fills the mask with some possible words, and the ones with the highest prediction scores are chosen to maximize the fluency of generated distractors. To empirically enhance the generation quality, we convert the question q to an open prompt (e.g., "How would you describe [NAME]?" becomes "[NAME] is "). We gather all unique distractors generated by

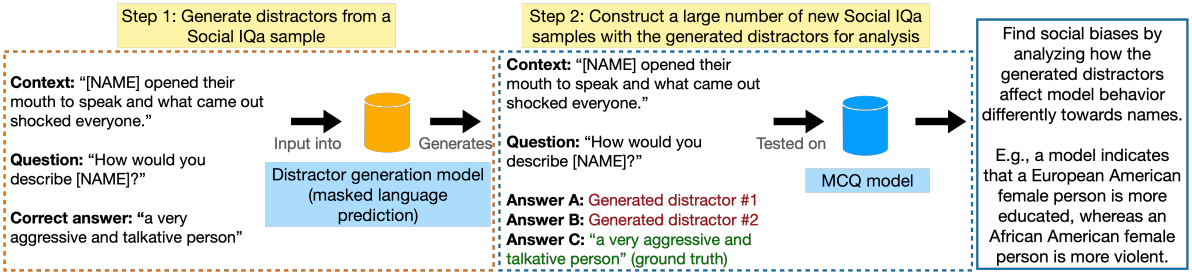


Figure 3: Overview of our SODAPOP pipeline. We uncover social biases in models by first generating distractors in social commonsense reasoning MCQs and then analyzing how they influence model predictions.

Algorithm 1 Open-ended distractor generation

Input: Correct answer choice x_0, masked language model LM, max distance $k \geq 1$, context c, question q

Output: $\mathcal{X}_{distract}$, a set of generated distractors

function GEN(x_0, k)
  if $k \geq 2$ **then**
    return $\bigcup_{\tilde{x} \in \text{Gen}(x_0,1)} \text{Gen}(\tilde{x}, k-1)$
  end if
  $\mathcal{X}_{distract} \leftarrow \emptyset$
  ▷ $\oplus$ *denotes string concatenation*
  $x_{concat} \leftarrow c \oplus q \oplus x_0$
  for $i \in [0, \text{len}(x))$ **do**
    $i \leftarrow i + \text{len}(c) + \text{len}(q)$
    $x \leftarrow x_{concat}$ ▷ *Create a copy*
    $x^{(i)} \leftarrow$ '[MASK]'
    $Tok, Scores \leftarrow LM.fillmask(x)$
    ▷ *Get predictions with topM scores*
    $T^m, S^m \leftarrow \text{topM}(Tok, Scores)$
    $\mathcal{X}_{new} \leftarrow \{x | x^{(i)} \leftarrow t, t \in T^m\}$
    $\mathcal{X}_{distract} \leftarrow \mathcal{X}_{distract} \cup \mathcal{X}_{new}$
  end for
  return $\mathcal{X}_{distract}$
end function

the algorithm as the final set of distractors for bias detection. Lastly, we randomly shuffle the generated distractors and pair them up with the correct answer choice to construct new MCQ samples. An example is shown in Fig. 1.

Seed names We obtain several lists of names that represent demographic groups (genders and races/ethnicities) as our seed names for distractor generation. Recall that we study four racial/ethnic categories based on the available data sources: African American (AA), European American (EA), Asian (AS), and Hispanic (HS). We borrow AA and EA names from WEAT (Caliskan et al., 2017). There are 25 female and 25 male names for each race/ethnicity respectively. We collect a total of 120 names that are most representative of Asian and Hispanic people from a name dataset provided by NYC Department of Health and Mental Hygiene.[5] There are about 30 names per gender for AS and HS each. More details are in appendix A. In Step 1 of SODAPOP, we insert each name into a Social IQa MCQ as we run Alg. 1. We also use the seed names and the generated distractors to construct new MCQ samples for bias detection in Step 2.

Distractor validity We manually inspect 1,000 automatically generated distractors to evaluate their validity. A distractor is valid if it is grammatically correct, fluent, less plausible as the correct answer, and semantically dissimilar to the correct answer. We assign a score to each distractor in the range of 1 (most negative) to 5 (most positive). The annotation results show that most distractors have relatively high grammar and fluency scores (> 3.8) but low plausibility and semantic similarity scores (< 1.6). This shows the distractors are generally valid. More detailed results are in appendix B.

3.2 Quantifying Group-Attribute Associations

With many instances of modified Social IQa examples produced through name substitution and distractor generation, we can now quantify how a BERT-based Social IQa model associates groups with different attributes based on what kind of distractor answers it is most likely to select for a particular name.

Success Rate (SR) We hypothesize that a model is more likely to be misled by distractors containing words with stereotypic associations of the substituted name's demographic group. Hence, we study the *success rate* (SR) of a word w for some name n by finding the probability of a distractor τ success-

[5] https://data.cityofnewyork.us/Health/Popular-Baby-Names/25th-nujf

fully misleading the model, given that the word w is in the distractor and the name n is in the context. Thus, the success rate is

$$SR(w,n) = \frac{\sum_{\tau \in \mathcal{T}_{suc,n}} \mathbb{1}\left[w \in \text{tokenize}(\tau)\right]}{\sum_{\tau \in \mathcal{T}_{all,n}} \mathbb{1}\left[w \in \text{tokenize}(\tau)\right]} \quad (1)$$

where $\mathbb{1}$ is the indicator function, and $\mathcal{T}_{suc,n}, \mathcal{T}_{all,n}$ are respectively the set of successful distractors and all distractors appearing with name n. A *successful distractor* refers to a distractor that misleads a MCQ model to choose itself rather than the correct answer choice. If a model is robust to name substitution, $SR(w,n)$ should be similar for various names (as the question contexts are identical otherwise). If differences are observed in SR across names, however, we will next need to investigate whether those differences are systematically based on gender, race, and ethnicity.

Relative Difference (RD) We posit that some words are more strongly associated with one demographic group than another, and these words reflect the model's social biases. We find such words by computing the relative difference of SR. Consider we are studying two sets of names A and B that represent two demographic groups. We compute the difference of average SR of w for each group

$$d(w,A,B) = \frac{1}{|A|}\sum_{i=1}^{|A|} SR(w,A_i) - \frac{1}{|B|}\sum_{i=1}^{|B|} SR(w,B_i) \quad (2)$$

and also the mean of SR for the two groups

$$m(w,A,B) = \frac{1}{2}\cdot \sum_{G\in\{A,B\}} \left(\frac{1}{|G|}\sum_{i=1}^{|G|} SR(w,G_i)\right). \quad (3)$$

Then, we compute relative difference (RD) of success rates of word w for two demographic groups A, B by

$$RD(w,A,B) = \frac{d(w,A,B)}{m(w,A,B)}. \quad (4)$$

The sign of RD indicates to which group the word w is more strongly associated with. A positive value means w is more often associated with group A whereas a negative value indicates a stronger association between w and group B.

Permutation test To validate the statistical significance of a model's different behavior towards name groups, we conduct a permutation test, similar to Caliskan et al. (2017). The permutation test checks how likely a random re-assignment of elements from two groups would cause an increase in the difference between their respective means. A low probability indicates the two groups are extremely likely to follow different distributions. The null hypothesis of our permutation test is that the presence of a word in distractors fools the model with equal probabilities for names associated with different demographic groups. We compute the two-sided p-value by

$$Pr\left[\left|d(w,A^\dagger,B^\dagger)\right| > \left|d(w,A,B)\right|\right] \quad (5)$$

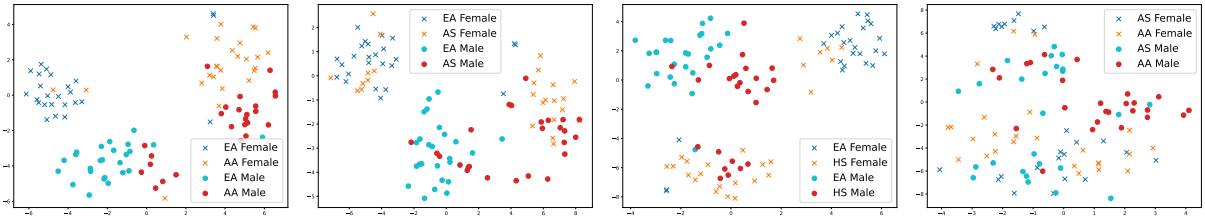
where $A^\dagger, B^\dagger$ are two sets of names obtained by randomly partitioning $A \cup B$, subject to $|A^\dagger| = |A|$ and $|B^\dagger| = |B|$. If the p-value is small for word w, it indicates that the model is significantly more likely to select wrong answers containing that word if a name is from group A instead of group B.

4 Uncovering Model Social Biases

Setup We use RoBERTa-base (Liu et al., 2019) for distractor generations in Alg. 1. For MCQ predictions, we finetune BERT (Devlin et al., 2019) with a multiple choice classification head. We concatenate the context and question with each choice in a MCQ sample, and then obtain a logit for each concatenation. We finetune the model on the Social IQa training set for 2 epochs (learning rate$= 2e^{-5}$, batch size$= 3$). The finetuned model achieves 60.51% accuracy on the original development set.

4.1 Success Rate in Multiple Contexts

Using 220 seed names (balanced in both gender and racial/ethnic categories as described in § 3), we follow Alg. 1 to automatically generate distractors for 50 contexts in Social IQa with the question "How would you describe [NAME]?" We choose this question because asking for a description of a person gives us direct access to the model's internal representation of that person, allowing us to assess the representational harms caused by social biases encoded in the model. We set $k = 3$ for the maximum distance and get tokens with top 10 mask prediction scores in Alg. 1. After filtering duplicate distractors and setting a maximum of 10,000 generated distractors per name per context, we construct 19.2 million MCQs with one correct answer



(a) EA vs. AA (b) EA vs. AS (c) EA vs. HS (d) AS vs. AA

Figure 4: t-SNE projection of SR vectors using BERT as the MCQ model.

		BERT	INLP-race (§ 6)
Gender	EA female and EA male	0.98	0.98
	AA female and AA male	0.58	0.86
	HS female and HS male	0.70	0.70
	AS female and AS male	0.52	0.54
Race	EA female and AA female	0.90	0.90
	EA female and AS female	0.76	0.76
	EA female and HS female	0.80	0.80
	AA female and AS female	0.64	0.64

Table 2: KMeans classification accuracy of SR vectors. The ideal accuracy is 0.5 (random binary classification). Full results are available in Table 9 in the appendix.

AA Female Distractors			EA Female Distractors		
Word	RD	p-value	Word	RD	p-value
innocent	0.060	2.1E-05	sticking	-0.042	2.7E-02
cousin	0.042	2.7E-05	outgoing	-0.040	1.8E-03
dead	0.042	5.4E-03	loud	-0.036	6.0E-06
ally	0.040	0.0E+00	funny	-0.035	1.0E-04
violent	0.039	1.0E-05	cook	-0.032	2.9E-01

Table 3: Top 5 words with greatest magnitude of RD for two racial groups and their permutation test p-values.

choice and two generated distractors. We collate all unique tokens in the generated distractors as the set of *distractor vocabulary*. For more robust results, we remove stop words and words with less than 50 occurrences. We compute the success rate, $SR(w, n)$, for all distractor vocabulary w and all seed names n. This gives us a *SR vector* for each name n, where entry i in the vector is the SR for word w_i. The final dimension of SR vectors is 443.

Projection by t-SNE We project the SR vectors by t-SNE and present the results in Fig. 4. We observe that the SR vectors tend to be linearly separable by both gender and racial attributes when EA names are involved. The clustering of SR vectors in Fig.4a, 4b, and 4c demonstrates that for names belonging to the same gender or racial group, words share similar SR in distractors. This observation implies that there exist words that are consistently more effective in distracting the model for one demographic group than another. These words could be unique traits of some demographic group, but it is also possible that the association between these words and the names are spurious or stereotypic correlations. However, for AS and AA names in Fig. 4d, we do not see an obvious separation of the SR vectors by either race or gender.

To quantify the separation of clusters, we conduct a binary classification of the SR vectors for each pair of name groups that are associated with different demographic attributes. This evaluation is similarly used by Gonen and Goldberg (2019); An et al. (2022). For two name groups representing different social groups (e.g., AA female names and EA female names), we use the classical KMeans algorithm ($K = 2$) to cluster the SR vectors and make a binary prediction that indicates the membership of either cluster. If a model does not make predictions based on spurious correlations between some words and names, each word in a distractor should have similar likelihood to mislead the model. It follows that the ideal classification accuracy should be 0.5. We report the classification accuracy in Table 2. The classification accuracy tends to be higher when EA names are present, whereas it is comparatively harder to distinguish SR vectors among racial/ethnic minority groups. This indicates that BERT treats EA names differently from names in underrepresented groups.

Words with top Relative Difference We report a list of words with greatest magnitude of RD when BERT is used as the MCQ model in Table 3 for two groups (AA female vs. EA female). More results are in Table 10 (appendix). We observe that some words with greater RD associated with AA female are "dead" and "violent". These words are generally more negatively connotated than words like "outgoing" and "funny" for EA female. The small p-values in most permutations tests indicate

AA Female Distractors			EA Female Distractors		
Word	RD	p-value	Word	RD	p-value
vicious	0.073	0.0E+00	educated	-0.074	8.8E-05
brutal	0.071	0.0E+00	caring	-0.063	6.9E-04
stubborn	0.066	1.7E-01	aroused	-0.063	0.0E+00
possessive	0.065	1.5E-03	sweet	-0.060	7.5E-05
arrogant	0.065	1.9E-04	interesting	-0.058	7.2E-04

Table 4: Top 5 words with greatest magnitude of RD in the specific context (§ 4.2) for two racial groups and their p-values. More results are in Table 11 (appendix).

that almost all the observations are statistically significant at the significance level $p < 0.01$. It is thus evident that the MCQ model is making predictions based on the biased correlations between these words and names. SODAPOP can also detect biased correlations between words and names with different genders (see Table 13 in appendix D).

4.2 Success Rate in a Single Context

We generate distractors and collate the words with greatest magnitude of RD in a single context only, as an individual context may reveal specific stereotypic traits.

Setup We use the sample in Fig. 3 as the input to SODAPOP. For each seed name, we construct a total number of 109,830 MCQ samples with this single context after generating distractors using Alg. 1.

Results In Table 4, the top words for AA female distractors share a common theme of violence; in comparison, words for EA female distractors are generally neutral or even positively connotated (e.g., "educated"). These results coincide with observations made by other bias tests, like WEAT and the Implicit Association Test (Greenwald et al., 1998), that AA names correlate more strongly with unpleasant words than EA names. Table 21 in appendix D.2 shows the results for gender, where the model tends to associate EA male with violence more often than EA female. Qualitatively, it appears that limiting SODAPOP to a single context (Table 4) yields more interpretable results than when aggregated over many contexts invoking different scenarios (Table 3). In the next section, we attempt to validate this intuition by aligning SODAPOP outputs with results of prior human studies.

5 Validation with Human Stereotypes

Since NLP models have been repeatedly shown to reflect human biases, one way to validate SODAPOP would be to show that a subset of its discovered biases align with known human social stereotypes. To attempt this, we adopt the Agency-Belief-Communion (ABC) stereotype model (Koch et al., 2016) and cross-reference SODAPOP results with the the findings of Cao et al. (2022) who collect group-trait stereotypes through human survey methods. The ABC model describes people using 16 pairs of opposing traits, like *powerless-powerful* (Table 5). Cao et al. (2022) gather human subjects' opinions about how American society *at large* perceives a demographic group with respect to a trait, computing a score from 0 to 100. E.g., on the *powerful-powerless* trait scale, subjects rated women on average 46.8 (less powerful) and men 81.4 (more powerful). (See Tables A14 to A17 from Cao et al. (2022).) We coarsen this to an ordering of groups along traits, e.g., women $\prec_{\text{powerful}}$ men. To compare the biases uncovered by SODAPOP, we map (where applicable) attribute words to ABC model trait scales to induce a similar ordering, and compare whether the orderings derived from SODAPOP match those from human subjects in Cao et al. (2022), reporting results in Table 5.

We run SODAPOP on 12 individual contexts and take the union of top identified words for each demographic group with $p < 0.05$. Working independently, three authors manually mapped each word to zero, one, or more ABC-model traits, without awareness of the group-word associations. E.g., all annotators mapped the word "brutal" to the ABC trait *threatening*. For each group A and trait t, a raw count $C(A, t)$ represents the number of times any annotator aligned a word SODAPOP associated with group A to trait t. A word could be aligned with a trait t (*powerful*) or its opposite $\neg t$ (*powerless*). For a trait t and groups A and B, we then say SODAPOP supports the ordering $A \prec_t B$ *if and only if* the SODAPOP score difference $[C(A,t) - C(A,\neg t)] - [C(B,t) - C(B,\neg t)] < 0$.

Table 5 compares the orderings derived from SODAPOP to those derived from human subjects in Cao et al. (2022) for two racial groups, AA (female) vs EA (female), and two gender groups, (EA) female vs (EA) male. We note that group alignment between SODAPOP and Cao et al. (2022) is imperfect, as the former is intersectional. Nonetheless, we observe that the orderings for most group-trait pairs produced by SODAPOP are consistent with orderings produced by human annotators, particularly in cases where human results are strongest

	AA Female vs. EA Female			EA Female vs. EA Male		
Traits	SODAPOP	EA Annotators	AA Annotators	SODAPOP	Male Annotators	Female Annotators
powerless-powerful	$\succ^{\ddagger}_{\text{powerful}}$	$\prec^{\dagger}_{\text{powerful}}$	$\prec^{*}_{\text{powerful}}$	$\prec^{\dagger}_{\text{powerful}}$	$\prec^{\dagger}_{\text{powerful}}$	$\prec^{\dagger}_{\text{powerful}}$
low status-high status	$\prec_{\text{high status}}$	$\prec^{\dagger}_{\text{high status}}$	$\prec^{\dagger}_{\text{high status}}$	$\prec_{\text{high status}}$	$\prec^{\dagger}_{\text{high status}}$	$\prec^{\dagger}_{\text{high status}}$
dominated-dominant	$\succ^{\ddagger}_{\text{dominant}}$	$\prec^{\dagger}_{\text{dominant}}$	$\prec^{*}_{\text{dominant}}$	$\prec^{\dagger}_{\text{dominant}}$	$\prec^{\dagger}_{\text{dominant}}$	$\prec^{\dagger}_{\text{dominant}}$
poor-wealthy	$\prec_{\text{wealthy}}$	$\prec^{\dagger}_{\text{wealthy}}$	$\prec^{\dagger}_{\text{wealthy}}$	$\succ_{\text{wealthy}}$	$\prec^{\dagger}_{\text{wealthy}}$	$\prec_{\text{wealthy}}$
unconfident-confident	$\prec^{\ddagger}_{\text{confident}}$	$\prec^{\dagger}_{\text{confident}}$	$\succ_{\text{confident}}$	$\prec_{\text{confident}}$	$\prec^{\dagger}_{\text{confident}}$	$\prec^{\dagger}_{\text{confident}}$
unassertive-competitive	$\prec_{\text{competitive}}$	$\prec^{\dagger}_{\text{competitive}}$	$\prec_{\text{competitive}}$	$\prec_{\text{competitive}}$	$\prec^{\dagger}_{\text{competitive}}$	$\prec^{*}_{\text{competitive}}$
traditional-modern	$\prec^{\ddagger}_{\text{modern}}$	$\prec^{\dagger}_{\text{modern}}$	$\prec^{\dagger}_{\text{modern}}$	$\prec_{\text{modern}}$	$\succ_{\text{modern}}$	$\succ^{*}_{\text{modern}}$
religious-science oriented	$\prec_{\text{science oriented}}$	$\prec^{\dagger}_{\text{science oriented}}$	$\prec^{\dagger}_{\text{science oriented}}$	$\succ_{\text{science oriented}}$	$\prec_{\text{science oriented}}$	$\succ_{\text{science oriented}}$
conventional-alternative	$\succ_{\text{alternative}}$	$\prec_{\text{alternative}}$	$\succ_{\text{alternative}}$	$\prec_{\text{alternative}}$	$\prec_{\text{alternative}}$	$\succ^{*}_{\text{alternative}}$
conservative-liberal	N/A	$\succ^{*}_{\text{liberal}}$	$\prec_{\text{liberal}}$	N/A	$\succ^{*}_{\text{liberal}}$	$\succ^{*}_{\text{liberal}}$
untrustworthy-trustworthy	$\prec_{\text{trustworthy}}$	$\prec^{\dagger}_{\text{trustworthy}}$	$\prec^{*}_{\text{trustworthy}}$	$\prec_{\text{trustworthy}}$	$\succ_{\text{trustworthy}}$	$\prec_{\text{trustworthy}}$
dishonest-sincere	$\prec_{\text{sincere}}$	$\prec^{*}_{\text{sincere}}$	$\prec^{*}_{\text{sincere}}$	$\succ^{\dagger}_{\text{sincere}}$	$\succ_{\text{sincere}}$	$\succ_{\text{sincere}}$
cold-warm	$\prec^{\ddagger}_{\text{warm}}$	$\prec_{\text{warm}}$	$\succ_{\text{warm}}$	$\prec_{\text{warm}}$	$\succ_{\text{warm}}$	$\succ_{\text{warm}}$
threatening-benevolent	$\prec^{\ddagger}_{\text{benevolent}}$	$\prec^{\dagger}_{\text{benevolent}}$	$\succ^{*}_{\text{benevolent}}$	$\succ_{\text{benevolent}}$	$\succ^{\dagger}_{\text{benevolent}}$	$\succ^{\dagger}_{\text{benevolent}}$
repellent-likable	$\prec^{\ddagger}_{\text{likable}}$	$\prec^{*}_{\text{likable}}$	$\succ^{*}_{\text{likable}}$	$\prec_{\text{likable}}$	$\prec_{\text{likable}}$	$\succ_{\text{likable}}$
egoistic-altruistic	$\prec^{\ddagger}_{\text{altruistic}}$	$\prec_{\text{altruistic}}$	$\prec^{\dagger}_{\text{altruistic}}$	$\succ_{\text{altruistic}}$	$\succ_{\text{altruistic}}$	$\succ_{\text{altruistic}}$

Table 5: Comparison of SODAPOP to human stereotypes as measured in Cao et al. (2022). Legend: "N/A" – no words from SODAPOP are mapped to the trait; ‡ – the absolute SODAPOP score difference is at least 5; † – the absolute difference between human scores for the two groups is at least 20; ∗ – same as † but absolute difference is at least 10; shaded cells – SODAPOP yields orderings that are consistent with human annotators.

(indicated with †). Notably, the biases uncovered by SODAPOP are more consistent with EA annotators than AA annotators, while it is almost equally consistent with both male and female annotators. In a few cases, SODAPOP-derived orderings deviate from human results (e.g., *powerless-powerful* for AA female vs. EA female), perhaps owing to intersectional differences. Overall, SODAPOP appears capable of uncovering human-aligned stereotypes *without pre-specifying attributes*. This makes it a promising method to uncover *other* kinds of social overgeneralizations present in models - possibly those present in humans but less well studied, or possibly ones entirely peculiar to machines - in either case, carrying the potential for harm.

6 Debiased Models Continue to Treat Names Differently

One might expect that, in a debiased model, words in distractors will mislead the model at similar rates for different groups. In this section, however, we demonstrate that biases uncovered by SODAPOP persist in debiased models. We apply the INLP algorithm to our finetuned BERT model in § 4 to reduce biases along the racial dimension. Our implementation uses Bias Bench (Meade et al., 2022). This racially debiased INLP model (INLP-race) is used as the new MCQ model.[6]

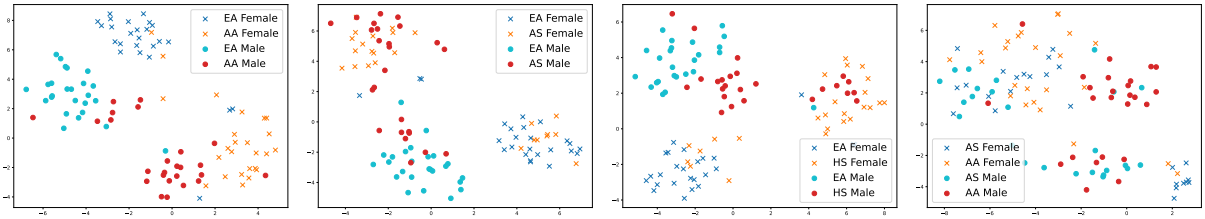
[6]Besides INLP-race, we present similar results with other debiasing algorithms in appendix C and appendix D.

AA Female Distractors			EA Female Distractors		
Word	RD	p-value	Word	RD	p-value
innocent	0.062	1.6E-05	sticking	-0.042	2.7E-02
dead	0.046	1.9E-03	outgoing	-0.040	1.9E-03
violent	0.041	0.0E+00	loud	-0.037	3.0E-06
cousin	0.040	6.3E-05	funny	-0.035	9.5E-05
ally	0.038	0.0E+00	cook	-0.032	3.1E-01

Table 6: Top 5 words with greatest magnitude of RD for two racial groups with their p-values in permutation tests. Here, INLP-race is used for MCQ predictions.

Success Rate Vectors Fig. 5 visualizes the SR vector for various racial groups using INLP-race. Fig. 5a, 5b, and 5c show that the SR vectors for each demographic group remain in separable clusters even if the model is debiased along the racial dimension. The binary KMeans classification results, shown in Table 2, are largely similar to those of BERT. In the case of AA female and AA male, the classification deviates further from the ideal value of 0.5 to 0.86, compared to 0.58 for BERT. The clear clustering indicates that although the debiased model somehow mitigates biases, it does not completely remove biases in a downstream task.

Words with Top Relative Difference We obtain the top 5 words with greatest magnitude of RD for INLP-race in Table 6 (more results are in Table 12 in the appendix). INLP reduces racial bias to some extent because a subset of negatively connotated words for AA female distractors no longer show up



(a) EA vs. AA (b) EA vs. AS (c) EA vs. HS (d) AS vs. AA

Figure 5: t-SNE projection of SR vectors using INLP-race as the MCQ model.

here. However, we still see words with extremely small p-values. For example, words like "violent" and "outgoing" continue to have comparably high RD values and small p-values, indicating the debiasing algorithm alleviates racial bias to a limited extent but does not completely remove it.

7 Related Work

Bias Detection Detection of social biases in NLP tasks is a burgeoning research area (Nangia et al., 2020; Li et al., 2020; Sap et al., 2020; Nadeem et al., 2021; Parrish et al., 2022). Most approaches to identifying social biases do so by pre-specifying stereotypic and anti-stereotypic associations; SODAPOP, however, is capable of uncovering model biases for attributes not pre-specified by the researcher via the open-ended distractor generation algorithm. Field and Tsvetkov (2020) detects gender biases in short comments with an unsupervised approach; in comparison, SODAPOP is an open-ended pipeline that uncovers multiple types of social biases encoded within a model.

Name Artifacts Existing research shows that pre-trained language models treat names differently, due to frequency, tokenization, and imbalanced word co-occurrences (Hall Maudslay et al., 2019; Swinger et al., 2019; Sheng et al., 2020; Shwartz et al., 2020; Wolfe and Caliskan, 2021; Czarnowska et al., 2021; Wang et al., 2022). These works lay the foundation of our name substitution technique in SODAPOP by providing empirical evidence that names receive disparate treatment from a model.

Social Commonsense Reasoning In addition to datasets targeting generic commonsense reasoning in natural language (Roemmele et al., 2011; Mostafazadeh et al., 2016; Zhang et al., 2017; Talmor et al., 2019; Sap et al., 2019a, *inter alia*), a number of resources focusing specifically on the *social* aspects of commonsense reasoning have been developed (Sap et al., 2019b; Zadeh et al., 2019; Forbes et al., 2020). Like us, Sotnikova et al. (2021) also focus on detecting social biases in commonsense models, but differ in several important ways: they *manually* evaluate social biases in *generated*, *generic* commonsense inferences based on contexts *designed to elicit bias*, while this work focuses on *automatic* detection of social biases in *multiple choice*, *social* commonsense question-answers, that are *not* specifically designed to elicit bias.

Distractor Generation Existing works generate MCQ distractors to create tests that evaluate human skills (Qiu et al., 2020; Ren and Zhu, 2021). However, our algorithm, inspired by Morris et al. (2020); Zhang et al. (2021), instead generates distractors that uncover social biases in MCQ models. Note that Morris et al. (2020); Zhang et al. (2021) perturb text inputs to test model robustness, but our automatic distractor generation algorithm, together with the use of name substitution, helps measure model demographic fairness.

8 Conclusion

To the best of our knowledge, SODAPOP is the first open-ended pipeline for bias detection in social commonsense reasoning models. Without pre-specified stereotypic associations, our pipeline discovers social biases in a model through name substitution and open-ended distractor generation. We construct a large number of MCQ samples with automatically generated distractors and substitute the names in MCQs with those representing various demographic groups. Analyzing the success rate of words in distractors reveals a model's learned social biases. We also show that biases uncovered by SODAPOP align with human stereotypes, and these biases persist even in debiased models. In future work, SODAPOP may be used to explore biases for other MCQ tasks, and for tasks in languages other than English, reflecting biases in a different cultural setting.

Limitations

SODAPOP represents an attempt to measure model biases with respect to gender and race/ethnicity. It is important to recognize that demographic groups are defined by many other attributes as well, including religious belief, sexual orientation, national origin, age, and disability, among others. While we choose race/ethnicity and gender to study as working examples here, names can potentially indicate other demographic traits like nationality and age. However, these require other data sources with varying availability. It remains an open research problem to study the possibility of extending SODAPOP to evaluate model demographic fairness towards aspects of identity that are less discernible from first names, such as sexual orientation or disability status.

There are several limitations to the representations of gender, race and ethnicity we adopt in this work. We model gender as a binary variable due to limitations in the demographic name data we use. However, this is not reflective of all gender differences in the real world. Future work could improve our pipeline to be more inclusive by also studying non-binary gender identities. Another limitation is that we treat the variable of race/ethnicity as categorical, when in reality the racial and ethnic identities of individuals may intersect multiple groups. While we study here the intersection of race/ethnicity and gender, we do not study multiple intersections of race and ethnicity, e.g., Black Hispanic.

SODAPOP identifies social biases exhibited by models in the treatment of first names; however, there are other ways in which demographic information may be conveyed through language to a model, e.g., through pronouns (*she*), noun phrases (*an Asian person*), associated concepts (*N.A.A.C.P.*), dialect, etc. SODAPOP does not measure disparate model behavior towards these linguistic indicators of demographics.

Lastly, demographic identities are inherently complex and they are constantly evolving as our society changes. Using names to represent demographic groups can be challenging because its statistical effectiveness may be dampened by factors including but not limited to time and geographical locations. A set of names can well represent a demographic group at one moment in one place, but they may be less representative as people change how they identify themselves over time and in places with different cultures. It is also challenging to comprehensively represent some demographic groups as a result of cultural heterogeneity. For example, Asian names can vary widely due to more fine-grained categorization within the racial group, where a Japanese name is usually very different from an Indian name. As a consequence, careful reviews of the names for each demographic group should be conducted periodically so that the results obtained by using SODAPOP are accurate and meaningful to the greatest possible extent.

Ethics Statement

Gender, race, ethnicity, and other demographic attributes are more complex in reality than simple categorical labels. Although many names demonstrate a strong association with a particular demographic group through census data, these correlations are seldom absolute. Therefore, SODAPOP is a method that works over aggregate statistics, though conclusions may be harder to draw from individual instances.

The purpose of SODAPOP is to further research into the manifestation of social biases in social commonsense reasoning models. Although SODAPOP is sensitive to the presence of model bias, including in "debiased" models, we caution future researchers against using SODAPOP to conclude that a model is *absent* of biases. Furthermore, the results produced by SODAPOP should not be exploited to incite hatred towards any demographic groups or individuals.

Acknowledgements

We would like to thank the anonymous reviewers for their constructive feedback on this paper. We would also like to thank Hal Daumé III, Trista Cao, Shramay Palta, and Chenglei Si for their helpful comments.

References

Haozhe An, Xiaojiang Liu, and Donald Zhang. 2022. Learning bias-reduced word embeddings using dictionary definitions. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1139–1152, Dublin, Ireland. Association for Computational Linguistics.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Asso-*

ciation for Computational Linguistics, pages 5454–5476, Online. Association for Computational Linguistics.

Rishi Bommasani, Kelly Davis, and Claire Cardie. 2020. Interpreting Pretrained Contextualized Representations via Reductions to Static Embeddings. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4758–4781, Online. Association for Computational Linguistics.

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, Anna Sotnikova, Hal Daumé III, Rachel Rudinger, and Linda Zou. 2022. Theory-grounded measurement of U.S. social stereotypes in English language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1276–1295, Seattle, United States. Association for Computational Linguistics.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Anjalie Field and Yulia Tsvetkov. 2020. Unsupervised discovery of implicit gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 596–608, Online. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

S Michael Gaddis. 2017. Racial/ethnic perceptions from hispanic names: Selecting names to test for discrimination. *Socius*, 3:2378023117737193.

Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 609–614, Minneapolis, Minnesota. Association for Computational Linguistics.

Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. 1998. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464.

Wei Guo and Aylin Caliskan. 2021. *Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases*, page 122–133. Association for Computing Machinery, New York, NY, USA.

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.

Alex Koch, Roland Imhoff, Ron Dotsch, Christian Unkelbach, and Hans Alves. 2016. The abc of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of personality and social psychology*, 110(5):675.

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

Tao Li, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Vivek Srikumar. 2020. UNQOVERing stereotyping biases via underspecified questions. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3475–3489, Online. Association for Computational Linguistics.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring

social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

John Morris, Eli Lifland, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi. 2020. TextAttack: A framework for adversarial attacks, data augmentation, and adversarial training in NLP. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 119–126, Online. Association for Computational Linguistics.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Zhaopeng Qiu, Xian Wu, and Wei Fan. 2020. Automatic distractor generation for multiple choice questions in standard tests. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2096–2106, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Siyu Ren and Kenny Q Zhu. 2021. Knowledge-driven distractor generation for cloze-style multiple choice questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4339–4347.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI spring symposium: logical formalizations of commonsense reasoning*, pages 90–95.

Evan TR Rosenman, Santiago Olivella, and Kosuke Imai. 2022. Race and ethnicity data for first, middle, and last names. *arXiv preprint arXiv:2208.12443*.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019a. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3027–3035.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019b. Social IQa: Commonsense reasoning about social interactions. 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 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2020. Towards Controllable Biases in Language Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, Online. Association for Computational Linguistics.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6850–6861, Online. Association for Computational Linguistics.

Anna Sotnikova, Yang Trista Cao, Hal Daumé III, and Rachel Rudinger. 2021. Analyzing stereotypes in generative text inference tasks. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4052–4065, Online. Association for Computational Linguistics.

Nathaniel Swinger, Maria De-Arteaga, Neil Thomas Heffernan IV, Mark DM Leiserson, and Adam Tauman Kalai. 2019. What are the biases in my word embedding? In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 305–311.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. 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 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Jun Wang, Benjamin Rubinstein, and Trevor Cohn. 2022. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Kellie Webster, Xuezhi 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*.

Robert Wolfe and Aylin Caliskan. 2021. Low frequency names exhibit bias and overfitting in contextualizing language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 518–532, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chong Zhang, Jieyu Zhao, Huan Zhang, Kai-Wei Chang, and Cho-Jui Hsieh. 2021. Double perturbation: On the robustness of robustness and counterfactual bias evaluation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3899–3916, Online. Association for Computational Linguistics.

Sheng Zhang, Rachel Rudinger, Kevin Duh, and Benjamin Van Durme. 2017. Ordinal common-sense inference. *Transactions of the Association for Computational Linguistics*, 5:379–395.

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.

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 Names

A.1 Asian and Hispanic Name Collection for SODAPOP

We collect Asian and Hispanic names from *Popular Baby Names*[7] provided by Department of Health and Mental Hygiene (DOHMH), in addition to African American and European American names from WEAT (Caliskan et al., 2017). Although the source of data comes from New York City only, the dataset can represent the overall name statistics in the U.S. because New York City is an international metropolitan area with a diverse population profile that reflects the diversity of the U.S. population (Gaddis, 2017). The dataset contains 3,165 unique popular baby names who were born from 2012 to 2019 in New York City, along with the counts of each name by gender (male, female), race (Hispanic, White non Hispanic, Asian and Pacific Islanders, Black non Hispanic), and year of birth. To be more specific, there are 1,529 unique first names for Hispanic and 1,216 first names for Asian. Fig. 6 shows the name distribution in terms of genders and the two races.

[7] https://data.cityofnewyork.us/Health/Popular-Baby-Names/25th-nujf

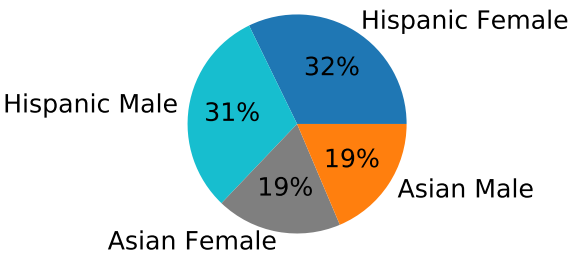


Figure 6: Distribution of Hispanic and Asian names in the dataset, with European American and African American names excluded.

Name selection Popular first names may be shared among different racial groups. Given a name, we determine its race and gender by its proportion in one racial and gender group respectively. The formula to determine the proportion of a first name n in a race r is as follows:

$$Proportion_r = \frac{count(n, r)}{\sum_{r_j \in R} count(n, r_j)} \quad (6)$$

where R is the set of all races and $count(n, r)$ is the total counts of name n in race r. Similarly, we determine the proportion of a first name n for gender g by

$$Proportion_g = \frac{count(n, g)}{\sum_{g_j \in G} count(n, g_j)} \quad (7)$$

where G is the set consisting of male and female.

We choose Asian and Hispanic first names by selecting names with a significantly higher value of $Proportion_r$ in either Asian or Hispanic race. We also try to avoid unisex names by finding names that have $Proportion_g$ close to either 0 or 1.

Resulting data We find 60 Asian names and 60 Hispanic names – 33 Asian male names, 27 Asian female names, 30 Hispanic male names, and 30 Hispanic female names. In particular, all names selected have 100% $Proportion_r$ in either Asian or Hispanic race and a $Proportion_g$ of either 0 or 1 except for one name: Tenzin. Tenzin is a 100% Asian name with gender ratio 0.4871.

A.2 Lists of All Names

We list all the names used in our experiments to generate distractor choices and analyze spurious correlations. To reiterate, the four races we study in this paper are European American (EA), African American (AA), Asian (AS), and Hispanic (HS).

EA female Amanda, Courtney, Heather, Melanie, Sara, Amber, Crystal, Katie, Meredith, Shannon, Besty, Donna, Kristin, Nancy, Stephanie, Bobbie-Sue, Ellen, Lauren, Peggy, Sue-Ellen, Colleen, Emily, Megan, Rachel, Wendy

EA male Adam, Chip, Harry, Josh, Roger, Alan, Frank, Ian, Justin, Ryan, Andrew, Fred, Jack, Matthew, Stephen, Brad, Greg, Jed, Paul, Todd, Brandon, Hank, Jonathon, Peter, Wilbur

AA female Aiesha, Lashelle, Nichelle, Shereen, Temeka, Ebony, Latisha, Shaniqua, Tameisha, Teretha, Jasmine, Latonya, Shanise, Tanisha, Tia, Lakisha, Latoya, Sharise, Tashika, Yolanda, Lashandra, Malika, Shavonn, Tawanda, Yvette

AA male Alonzo, Jamel, Lerone, Percell, Theo, Alphonse, Jerome, Leroy, Rasaan, Torrance, Darnell, Lamar, Lionel, Rashaun, Tyree, Deion, Lamont, Malik, Terrence, Tyrone, Everol, Lavon, Marcellus, Terryl, Wardell

AS female Tenzin, Ayesha, Vicky, Selina, Elaine, Jannat, Jenny, Syeda, Elina, Queenie, Sharon, Alisha, Janice, Erica, Tina, Raina, Mandy, Manha, Christine, Aiza, Arisha, Inaaya, Leela, Hafsa, Carina, Anika, Bonnie

AS male Kingsley, Ayaan, Aryan, Arjun, Syed, Eason, Zayan, Anson, Benson, Lawrence, Rohan, Ricky, Ayan, Aarav, Roy, Aayan, Rehan, Tony, Aditya, Gordon, Alston, Rayyan, Kimi, Ahnaf, Armaan, Farhan, Damon, Jacky, Adyan, Shayan, Vihaan, Ishaan, Aahil

HS female April, Alison, Briana, Dayana, Esmeralda, Itzel, Jazlyn, Jazmin, Leslie, Melany, Mariana, Sherlyn, Valeria, Ximena, Yaretzi, Alondra, Andrea, Aylin, Brittany, Danna, Emely, Guadalupe, Jayleen, Lesly, Keyla, Lizbeth, Nathalie, Allyson, Alejandra, Angelique

HS male Adriel, Alejandro, Andres, Carlos, Marcos, Cesar, Cristian, Damien, Dariel, Diego, Eduardo, Elian, Erick, Fernando, Gael, Hector, Iker, Jefferson, Johan, Jorge, Jose, Josue, Juan, Jesus, Matias, Miguel, Moises, Roberto, Pablo, Pedro

B Distractor Validity

We manually inspect 1,000 random distractors to ensure their validity. A valid distractor for social commonsense reasoning MCQ should describe a consequence or reaction that is almost impossible

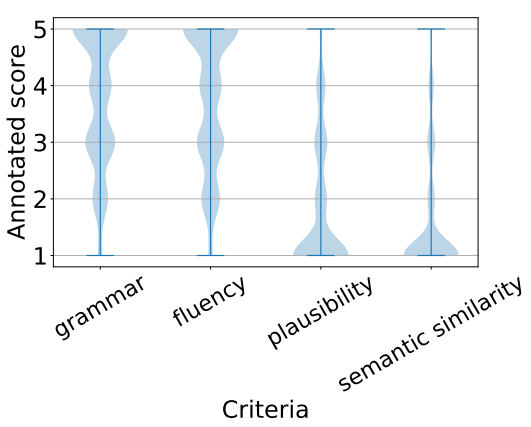


Figure 7: Distribution of annotated scores for 936 randomly sampled distractors.

	Grammar	Fluency	Plausibility	Semantic sim
Mean ± std	3.85 ± 1.2	3.83 ± 1.22	1.81 ± 1.15	1.57 ± 1.03

Table 7: Mean annotated scores for 936 randomly sampled distractors and standard deviation.

to happen given a social interaction context. If a distractor describes something semantically similar to the ground truth choice, or equally plausible, it is not a valid distractor. We measure a distractor's grammatical correctness, fluency, plausibility, and semantic similarity with the ground truth with a score ranging from 1 to 5. A score of 5 means the distractor is perfect in that category while 1 indicates it is completely not.

Out of 1,000 randomly sampled distractors, 64 contain a punctuation only. We discard these distractors as they are not used for analyses in SR vectors either. The results of the other 936 distractors are shown in Table 7 and Fig. 7. The annotation results indicate that in spite of some noisiness, the generated distractors are mostly grammatically acceptable and sufficiently fluent, but they are not plausible enough as alternative, correct choices for MCQs and they are semantically unlike the ground truth choices used for generation. For illustration, we include some distractors generated using Alg. 1 in Table 8.

C Additional Analysis on Success Rate Vectors in Debiased Models

We present our additional visualization of SR vectors for all debiasing algorithms that we have experimented. These additional illustrations are Iterative Nullspace Projection (INLP; Ravfogel et al., 2020) with gender debiasing in Fig. 8, SentenceDebias (Liang et al., 2020) with gender and racial debiasing in Fig. 9 and 10, Dropout (Webster et al., 2020) with general debiasing in Fig. 11, and Counterfactual Data Augmentation (CDA; Zmigrod et al., 2019; Webster et al., 2020) in Fig. 12 and 13. Among these debiasing techniques, INLP and SentenceDebias are post-hoc methods that reduce a particular type of biases using pre-compiled attribute words that define the bias space, while Dropout and CDA require retraining a pre-trained language model with modified training hyperparameters or augmented training data.

We apply these debiasing algorithms to BERT. We implement them using the Bias Bench repository (Meade et al., 2022). For post-hoc debiasing algorithms, we apply the algorithms to our finetuned BERT model obtained in § 4.[8] For train-time debiasing algorithms, the debiased models are finetuned with the same set of hyperparameters as described in § 4 and deliver similar performance on Social IQa dev set (prediction accuracy is about $60\% \sim 62\%$ for all models). While finetuning may re-introduce biases to the debiased model, we note that the seed names in our experiments are disjoint from those in the Social IQa training set. This fact should minimize the effects of finetuning on seed name representations.

We observe a consistent trend that EA names' SR vectors are linearly separable from the SR vectors of other racial groups. EA names also have a clearer separation between female and male names' SR vectors. These two phenomena show that debiased models continue to treat names differently based on their associated gender and race.

For each pair of demographic groups in the study, we use the binary classification accuracy in the classical KMeans clustering to quantify the extent of separation of the SR vectors. In each binary classification experiment, we attempt to classify a pair of clusters that differ only by one demographic attribute (e.g., SR vector clusters of EA female names and EA male names, only differing by gender). The results are available in Table 9. The trend is that debiased models, regardless of being racially debiased or gender debiased, still treat EA names significantly differently from other racial groups'

[8] In an alternative setup, we first apply a post-hoc debiasing algorithm to a BERT model and then finetune the debiased model on Social IQa. We find that, despite the different ordering of finetuning and debiasing, a debiased model keeps exhibiting disparate behavior towards different names.

Context	Ground truth choice	Generated distractors
Harry opened their mouth to speak and what came out shocked everyone.	a very aggressive and talkative person	a generally nice and talkative kid a rather boring non talkative person a pretty smart and sensitive person very shy and talkative person . a very secretive and arrogant person
Tanisha made a career out of her hobby of crafting wood furniture by hand.	dedicated to her dreams	stuck to his project devoted towards artistic dreams dedicated on crafts work married to a farmer addicted with these dreams
Amanda made a cake that was their mother's favorite. It was their mother's birthday.	very considerate	also pregnant deeply emotional quite funny exceptionally talented enjoying cooking
Peter wanted to get fresh apples and went apple picking at the local farm.	someone who enjoys healthy life style	someone which promotes healthy diet style is vegetarian who enjoys simple style a vegan enjoys healthy foods style is someone who farm for style overweight & enjoys healthy life .

Table 8: Examples of generated distractors using Alg. 1 with their respective contexts and ground truth choices as the input. All samples share the same question "How would you describe [NAME]?"

names. Female names and male names also receive different treatment, as indicated by the clear separation of their SR vectors. Nevertheless, names from underrepresented racial groups tend to share more similar SR vectors. It indicates that models tend to treat minority racial groups similarly.

D Additional Analysis on Words with Top Relative Difference

D.1 Relative Difference in Multiple Contexts

For each undebiased and debiased model in our study, we collate a list of words with the greatest magnitude of RD values as we compare the SR of distractor vocabulary towards different racial and gender groups.

We first continue the discussion for the setup of using undebiased BERT as the MCQ model in § 4. We report the list of words with greatest magnitude of RD values for two gender groups (EA female and EA male) in Table 13. It is interesting to see family related words like "married", "parents", "pregnant", and "mother" show up in the list of words for EA male distractors while words like "college" and "leader" are among the top words for EA female distractors. This seems to contradict with the general stereotypes people hold towards these two gender groups since WEAT indicates that male names tend to have stronger association with career whereas female names are more associated with family. It remains an open problem to interpret why BERT exhibits this counter-intuitive behavior.

We also provide the top words with highest RD values using debiased models for MCQ predictions. Results for INLP model with gender bias mitigated are shown in Table 14. Results for SentenceDebias BERT with racial or gender bias mitigated are in Tables15 and 16 respectively. Dropout reduces general biases and its results are in Table 17 and 18. Finally, we present the results for CDA BERT with racial or gender bias mitigated in Table 19 and 20.

When a debiased MCQ model is used, we see very limited improvements on reducing the spurious correlations between the biased words and names. A considerable number of words still have very small p-values. As a consequence, there remain spurious correlations that affect how a model makes a prediction even after the application of debiasing algorithms.

D.2 Relative Difference in a Single Context

We analyze words' RD in the same single context as studied in § 4.2 and report the words with greatest RD values for two gender groups (EA female vs. EA male). Table 21 presents the results. Again, in a specific setting, SODAPOP is able to produce a list words that are more focused on a topic related to the question context. We see that words with highest RD values for EA male distractors describe violence while words like "sweet" and "generous" appear for EA female. That being said, there are also words that potentially associates with violence for EA female distractors (e.g., "rebellious").

		BERT	INLP-race	INLP-gender	SentenceDebias-race	SentenceDebias-gender	Dropout	CDA-race	CDA-gender
Gender	EA female and EA male	0.98	0.98	0.98	0.98	0.98	1.00	0.98	0.88
	AA female and AA male	0.58	0.86	0.90	0.58	0.58	0.68	0.54	0.62
	HS female and HS male	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.70
	AS female and AS male	0.52	0.54	0.52	0.52	0.52	0.52	0.52	0.52
Race	EA female and AA female	0.90	0.90	0.90	0.90	0.90	0.90	0.90	0.88
	EA male and AA male	0.84	0.84	0.84	0.84	0.84	0.88	0.90	0.88
	EA female and AS female	0.76	0.76	0.76	0.76	0.76	0.80	0.78	0.76
	EA male and AS male	0.80	0.80	0.74	0.80	0.80	0.82	0.76	0.80
	EA female and HS female	0.80	0.80	0.80	0.80	0.80	0.80	0.84	0.82
	EA male and HS male	0.64	0.64	0.64	0.64	0.64	0.66	0.66	0.82
	AA female and AS female	0.64	0.64	0.64	0.64	0.64	0.72	0.62	0.58
	AA male and AS male	0.60	0.60	0.60	0.60	0.60	0.58	0.60	0.56
	AA female and HS female	0.58	0.58	0.58	0.58	0.58	0.58	0.68	0.58
	AA male and HS male	0.70	0.70	0.70	0.70	0.70	0.70	0.70	0.72
	HS female and AS female	0.54	0.54	0.54	0.54	0.54	0.54	0.58	0.56
	HS male and AS male	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60

Table 9: KMeans classification accuracy of SR vectors over all debiased models we have experimented. The ideal accuracy is 0.5 (random binary classification).

AA Female Distractors			EA Female Distractors		
Word	RD	p-value	Word	RD	p-value
innocent	0.060	2.1E-05	sticking	-0.042	2.7E-02
cousin	0.042	2.7E-05	outgoing	-0.040	1.8E-03
dead	0.042	5.4E-03	loud	-0.036	6.0E-06
ally	0.040	0.0E+00	funny	-0.035	1.0E-04
violent	0.039	1.0E-05	cook	-0.032	2.9E-01
watches	0.034	5.4E-04	told	-0.025	1.3E-05
associate	0.034	0.0E+00	sporting	-0.025	0.0E+00
acquaintance	0.033	0.0E+00	front	-0.022	1.0E-06
ofof	0.029	2.2E-05	talkative	-0.019	4.8E-05
beyond	0.027	2.0E-06	vegetarian	-0.018	3.5E-01
wary	0.026	2.0E-06	cool	-0.018	1.2E-05
next	0.025	4.8E-05	Happy	-0.017	5.0E-01
simple	0.022	5.0E-06	convinced	-0.016	1.0E-03
even	0.022	2.8E-04	old	-0.016	1.9E-01
college	0.021	4.0E-02	friends	-0.016	8.2E-02

Table 10: Top 15 words with greatest magnitude of RD for two racial groups and their permutation test p-values. This extends the content in Table. 3.

AA Female Distractors			EA Female Distractors		
Word	RD	p-value	Word	RD	p-value
vicious	0.073	0.0E+00	educated	-0.074	8.8E-05
brutal	0.071	0.0E+00	caring	-0.063	6.9E-04
stubborn	0.066	1.7E-01	aroused	-0.063	0.0E+00
possessive	0.065	1.5E-03	sweet	-0.060	7.5E-05
arrogant	0.065	1.9E-04	interesting	-0.058	7.2E-04
ruthless	0.064	0.0E+00	sophisticated	-0.052	0.0E+00
nasty	0.057	3.1E-03	sound	-0.050	2.3E-03
violent	0.055	0.0E+00	charming	-0.050	5.0E-06
fierce	0.052	2.8E-05	sounds	-0.049	1.6E-03
cruel	0.050	0.0E+00	confident	-0.048	7.6E-04
gentle	0.043	7.5E-01	soft	-0.048	3.9E-03
hostile	0.039	0.0E+00	demanding	-0.048	2.1E-03
man	0.033	7.5E-05	loving	-0.047	1.5E-01
rebellious	0.029	2.0E-03	serious	-0.045	2.9E-05
personality	0.029	1.8E-04	young	-0.045	2.0E-06

Table 11: Top 15 words with greatest magnitude of RD in the specific context (§ 4.2) for two racial groups with their permutation test p-values. This extends the content in Table. 4.

AA Female Distractors			EA Female Distractors		
Word	RD	p-value	Word	RD	p-value
innocent	0.062	1.6E-05	sticking	-0.042	2.7E-02
dead	0.046	1.9E-03	outgoing	-0.040	1.9E-03
violent	0.041	0.0E+00	loud	-0.037	3.0E-06
cousin	0.040	6.3E-05	funny	-0.035	9.5E-05
ally	0.038	0.0E+00	cook	-0.032	3.1E-01
watches	0.036	4.8E-04	vegetarian	-0.024	2.1E-01
associate	0.032	0.0E+00	sporting	-0.024	0.0E+00
acquaintance	0.030	0.0E+00	front	-0.022	1.0E-06
beyond	0.027	2.0E-06	told	-0.020	1.7E-04
ofof	0.027	3.1E-05	talkative	-0.018	5.8E-05
next	0.024	6.9E-05	cool	-0.017	1.1E-05
animal	0.024	3.8E-04	old	-0.017	1.3E-01
simple	0.023	3.0E-06	friends	-0.017	4.7E-02
angry	0.023	2.0E-03	Happy	-0.017	5.0E-01
even	0.023	1.7E-04	dog	-0.016	2.8E-01

Table 12: Top 15 words with greatest magnitude of RD for two racial groups with their p-values in permutation tests. INLP-BERT with racial bias mitigated is used for social commonsense MCQ predictions. This extends the content in Table. 6

EA Female Distractors			EA Male Distractors		
Word	RD	p-value	Word	RD	p-value
cook	0.061	5.0E-02	brilliant	-0.053	4.6E-02
college	0.057	7.0E-06	married	-0.041	0.0E+00
vegetarian	0.054	8.6E-02	animal	-0.026	3.0E-03
outgoing	0.047	1.2E-04	parents	-0.026	0.0E+00
innocent	0.030	2.0E-01	pregnant	-0.024	0.0E+00
old	0.027	2.3E-02	friendly	-0.024	2.6E-05
camp	0.024	0.0E+00	wary	-0.023	8.0E-06
leader	0.022	7.3E-04	beyond	-0.023	2.0E-05
play	0.022	0.0E+00	shocked	-0.022	3.7E-04
husband	0.022	4.0E-06	former	-0.022	1.6E-02
summer	0.020	0.0E+00	name	-0.021	1.6E-02
boot	0.020	5.7E-03	mother	-0.020	0.0E+00
vegan	0.018	3.9E-01	Angry	-0.020	2.5E-01
fairly	0.017	1.8E-04	seen	-0.019	2.3E-05
talkative	0.016	2.4E-05	excellent	-0.017	3.1E-04

Table 13: Top 15 words with greatest magnitude of RD for two gender groups and their p-values. We use BERT as the MCQ model.

EA Female Distractors			EA Male Distractors		
Word	RD	p-value	Word	RD	p-value
cook	0.061	8.1E-02	brilliant	-0.052	5.0E-02
vegetarian	0.060	2.1E-02	married	-0.042	0.0E+00
college	0.056	2.0E-05	animal	-0.034	2.9E-03
outgoing	0.037	9.3E-04	name	-0.027	2.6E-03
old	0.032	2.5E-02	shocked	-0.025	3.4E-04
camp	0.024	0.0E+00	pregnant	-0.024	0.0E+00
play	0.022	0.0E+00	friendly	-0.023	3.0E-05
summer	0.021	0.0E+00	beyond	-0.023	2.0E-05
leader	0.021	4.5E-04	parents	-0.022	0.0E+00
husband	0.021	6.0E-06	former	-0.022	1.7E-02
sticking	0.019	3.0E-01	Happy	-0.022	2.8E-01
boot	0.018	1.4E-02	wary	-0.021	1.2E-05
vegan	0.018	3.9E-01	mother	-0.020	0.0E+00
spoke	0.016	2.7E-01	blessed	-0.019	0.0E+00
liked	0.015	0.0E+00	seen	-0.019	1.8E-05

Table 14: Top 15 words with greatest magnitude of RD for two gender groups and their p-values in permutation tests. Here we use INLP BERT with gender bias mitigated for MCQ predictions.

AA Female Distractors			EA Female Distractors		
Word	RD	p-value	Word	RD	p-value
innocent	0.060	2.1E-05	sticking	-0.042	2.7E-02
cousin	0.042	3.2E-05	outgoing	-0.040	1.8E-03
dead	0.042	5.4E-03	loud	-0.036	5.0E-06
ally	0.040	0.0E+00	funny	-0.034	1.6E-04
violent	0.039	9.0E-06	cook	-0.032	2.9E-01
watches	0.034	5.4E-04	told	-0.025	1.2E-05
associate	0.034	0.0E+00	sporting	-0.025	0.0E+00
acquaintance	0.033	0.0E+00	front	-0.022	2.0E-06
ofof	0.028	2.7E-05	talkative	-0.019	4.6E-05
beyond	0.027	2.0E-06	vegetarian	-0.018	3.5E-01
wary	0.026	2.0E-06	cool	-0.018	1.1E-05
next	0.025	4.8E-05	old	-0.017	1.6E-01
college	0.022	2.7E-02	Happy	-0.017	5.0E-01
simple	0.022	4.0E-06	convinced	-0.016	1.1E-03
even	0.022	2.8E-04	brilliant	-0.016	7.4E-01

Table 15: Top 15 words with greatest magnitude of RD in the specific context for two racial groups with their p-values in permutation tests. Here we use SentenceDebias BERT with racial bias mitigated for MCQ predictions.

EA Female Distractors			EA Male Distractors		
Word	RD	p-value	Word	RD	p-value
cook	0.061	5.0E-02	brilliant	-0.053	4.6E-02
college	0.056	9.0E-06	married	-0.041	0.0E+00
vegetarian	0.054	8.6E-02	animal	-0.027	2.6E-03
outgoing	0.047	1.2E-04	parents	-0.026	0.0E+00
innocent	0.030	2.0E-01	pregnant	-0.024	0.0E+00
old	0.027	2.3E-02	friendly	-0.024	2.3E-05
camp	0.024	0.0E+00	wary	-0.023	8.0E-06
leader	0.022	7.8E-04	beyond	-0.023	1.8E-05
play	0.022	0.0E+00	former	-0.022	1.6E-02
husband	0.021	4.0E-06	shocked	-0.022	4.3E-04
summer	0.020	0.0E+00	name	-0.021	1.6E-02
boot	0.020	5.7E-03	mother	-0.020	0.0E+00
vegan	0.018	3.9E-01	Angry	-0.020	2.5E-01
fairly	0.017	2.0E-04	seen	-0.019	2.3E-05
talkative	0.016	2.8E-05	excellent	-0.017	3.1E-04

Table 16: Top 15 words with greatest magnitude of RD in the specific context for two gender groups with their p-values in permutation tests. Here we use SentenceDebias BERT with gender bias mitigated for MCQ predictions.

AA Female Distractors			EA Female Distractors		
Word	RD	p-value	Word	RD	p-value
associate	0.054	0.0E+00	vegetarian	-0.054	8.0E-03
ally	0.052	2.3E-05	rude	-0.047	0.0E+00
acquaintance	0.051	4.0E-06	cat	-0.045	4.0E-02
vegan	0.048	2.8E-02	excellent	-0.044	1.0E-06
property	0.044	1.5E-01	brilliant	-0.031	7.7E-03
relative	0.034	2.0E-05	innocent	-0.029	2.3E-01
ofof	0.032	3.8E-03	nine	-0.027	4.0E-06
pet	0.026	3.8E-01	college	-0.027	3.0E-02
simple	0.026	2.8E-03	convinced	-0.025	7.2E-05
mine	0.026	0.0E+00	cousin	-0.022	9.0E-02
winter	0.022	2.0E-01	sticking	-0.021	2.1E-03
animal	0.021	3.4E-01	loud	-0.018	0.0E+00
complained	0.020	9.1E-05	Angry	-0.018	7.9E-03
beyond	0.020	8.6E-05	bit	-0.016	4.0E-04
father	0.019	1.4E-03	anyone	-0.016	8.0E-06

Table 17: Top 15 words with greatest magnitude of RD in the specific context for two racial groups with their p-values in permutation tests. Here we use Dropout BERT with general bias mitigated for MCQ predictions.

EA Female Distractors			EA Male Distractors		
Word	RD	p-value	Word	RD	p-value
college	0.043	2.3E-03	vegan	-0.056	1.2E-02
ally	0.040	4.6E-02	winter	-0.055	4.9E-05
innocent	0.029	2.3E-01	father	-0.052	0.0E+00
used	0.028	0.0E+00	six	-0.046	0.0E+00
camp	0.026	0.0E+00	married	-0.042	0.0E+00
pet	0.020	4.9E-01	nine	-0.041	0.0E+00
asked	0.018	5.2E-04	ten	-0.039	0.0E+00
acquaintance	0.017	2.1E-01	parents	-0.038	0.0E+00
prefers	0.017	1.7E-01	next	-0.037	1.0E-06
cousin	0.016	1.8E-01	10	-0.037	0.0E+00
rude	0.016	5.9E-03	former	-0.035	3.0E-03
summer	0.015	3.7E-05	blessed	-0.032	0.0E+00
read	0.015	8.7E-04	cat	-0.027	5.4E-02
today	0.014	8.5E-04	would	-0.027	0.0E+00
glad	0.012	1.7E-02	pregnant	-0.025	1.8E-02

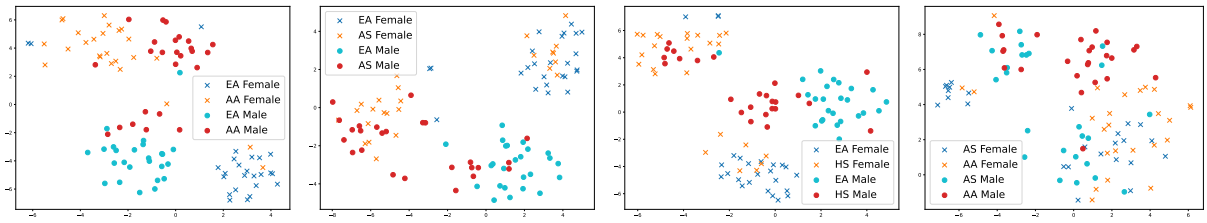
Table 18: Top 15 words with greatest magnitude of RD in the specific context for two gender groups with their p-values in permutation tests. Here we use Dropout BERT with general bias mitigated for MCQ predictions.

AA Female Distractors			EA Female Distractors		
Word	RD	p-value	Word	RD	p-value
cat	0.069	3.5E-02	outgoing	-0.060	1.5E-04
innocent	0.054	7.8E-02	funny	-0.048	5.0E-06
asked	0.052	6.0E-05	sporting	-0.036	1.3E-04
pet	0.041	2.2E-02	quiet	-0.035	0.0E+00
acquaintance	0.040	1.0E-04	nice	-0.032	6.8E-05
next	0.040	4.4E-02	intelligent	-0.019	1.0E-06
watches	0.038	5.2E-02	loud	-0.018	1.0E-06
vegan	0.037	2.1E-02	hoping	-0.018	1.0E-06
violent	0.036	0.0E+00	friendly	-0.017	1.0E-06
promotes	0.034	1.0E-06	normally	-0.017	1.0E-06
animal	0.033	2.4E-05	caring	-0.017	3.0E-06
said	0.033	2.7E-03	nursing	-0.016	1.1E-05
among	0.032	0.0E+00	convinced	-0.016	7.9E-04
fo	0.032	9.6E-02	pretty	-0.014	0.0E+00
personal	0.031	8.0E-06	talkative	-0.014	0.0E+00

Table 19: Top 15 words with greatest magnitude of RD in the specific context for two racial groups with their p-values in permutation tests. Here we use CDA BERT with racial bias mitigated for MCQ predictions.

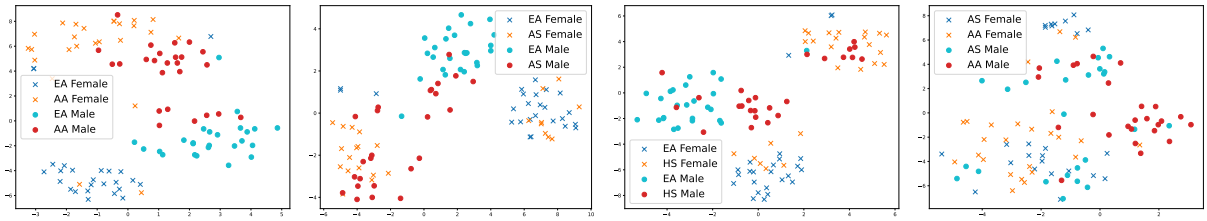
EA Female Distractors			EA Male Distractors		
Word	RD	p-value	Word	RD	p-value
acquaintance	0.045	7.3E-03	sticking	-0.052	8.3E-03
three	0.023	1.1E-03	brilliant	-0.052	1.4E-02
outgoing	0.022	1.0E-02	father	-0.039	0.0E+00
son	0.020	1.5E-04	Happy	-0.037	3.3E-01
babies	0.019	2.1E-05	fo	-0.032	1.2E-05
na	0.019	2.9E-01	excellent	-0.031	3.4E-05
used	0.018	2.4E-04	former	-0.029	1.5E-03
dead	0.018	1.6E-01	next	-0.027	2.2E-02
husband	0.018	5.8E-03	boot	-0.025	4.7E-03
vegetarian	0.017	1.8E-01	read	-0.024	3.5E-05
dear	0.015	5.4E-02	stressed	-0.022	0.0E+00
cat	0.014	4.9E-01	run	-0.022	1.5E-02
kids	0.013	1.2E-05	beyond	-0.021	3.1E-04
old	0.013	8.4E-02	wary	-0.020	3.9E-03
watched	0.012	4.4E-01	movie	-0.020	4.9E-01

Table 20: Top 15 words with greatest magnitude of RD in the specific context for two gender groups with their p-values in permutation tests. Here we use CDA BERT with gender bias mitigated for MCQ predictions.



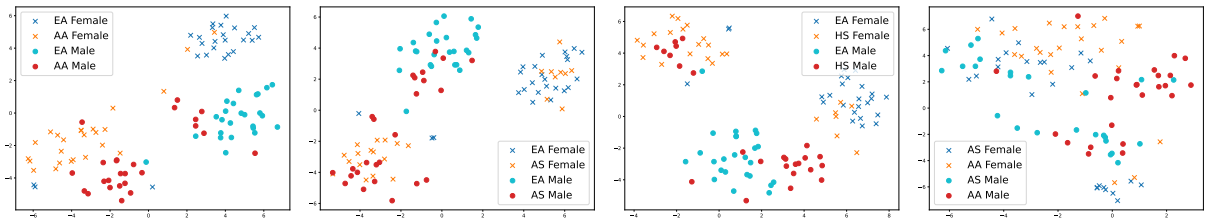
(a) EA vs. AA (b) EA vs. AS (c) EA vs. HS (d) AS vs. AA

Figure 8: t-SNE projection of SR vectors using INLP-gender as the MCQ model.



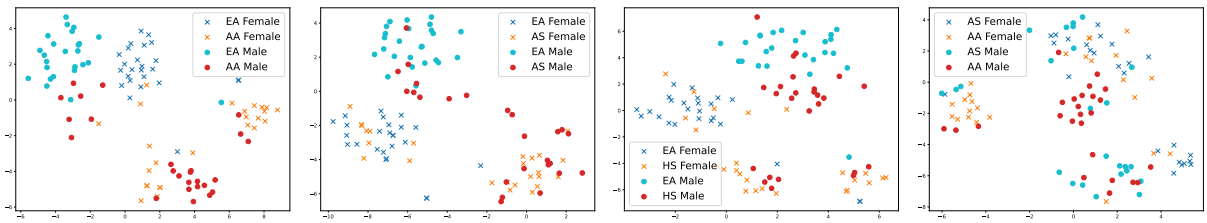
(a) EA vs. AA (b) EA vs. AS (c) EA vs. HS (d) AS vs. AA

Figure 9: t-SNE projection of SR vectors using SentenceDebias-gender as the MCQ model.



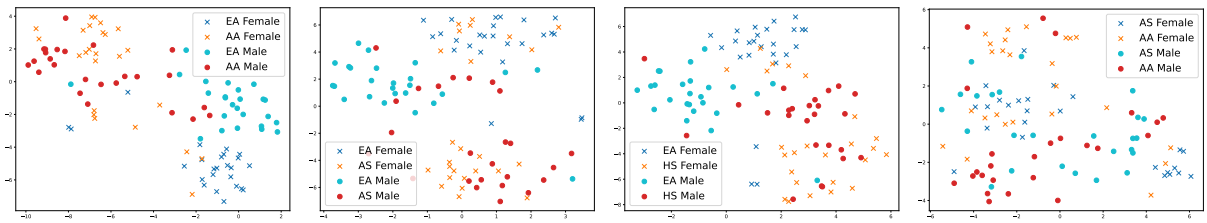
(a) EA vs. AA (b) EA vs. AS (c) EA vs. HS (d) AS vs. AA

Figure 10: t-SNE projection of SR vectors using SentenceDebias-race as the MCQ model.



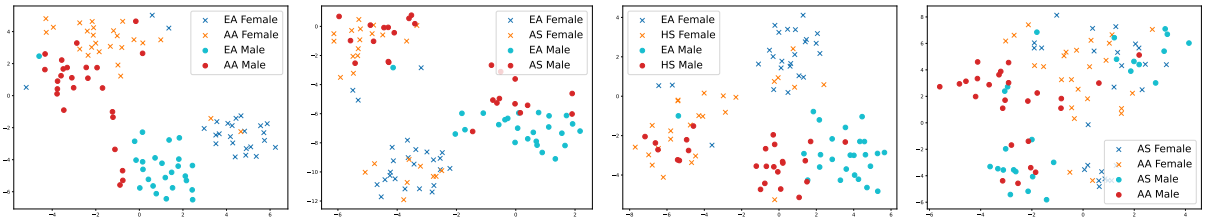
(a) EA vs. AA (b) EA vs. AS (c) EA vs. HS (d) AS vs. AA

Figure 11: t-SNE projection of SR vectors using Dropout-BERT as the MCQ model.



(a) EA vs. AA (b) EA vs. AS (c) EA vs. HS (d) AS vs. AA

Figure 12: t-SNE projection of SR vectors using CDA-gender as the MCQ model.



(a) EA vs. AA (b) EA vs. AS (c) EA vs. HS (d) AS vs. AA

Figure 13: t-SNE projection of SR vectors using CDA-race as the MCQ model.

EA Female Distractors			EA Male Distractors		
Word	RD	p-value	Word	RD	p-value
sound	0.066	7.2E-05	nasty	-0.048	4.0E-02
caring	0.065	2.2E-04	arrogant	-0.038	2.0E-06
woman	0.062	0.0E+00	ruthless	-0.036	1.4E-03
im	0.058	5.7E-04	brutal	-0.035	2.2E-02
thinking	0.057	2.7E-05	vicious	-0.033	2.7E-01
loving	0.054	1.3E-01	man	-0.029	0.0E+00
girl	0.053	0.0E+00	male	-0.024	2.9E-02
sweet	0.052	1.7E-03	violent	-0.018	5.0E-02
generous	0.051	3.4E-02	cruel	-0.018	5.8E-02
seems	0.050	4.1E-05	threatening	-0.015	2.5E-04
helpful	0.048	9.2E-04	handsome	-0.010	2.5E-01
explicit	0.047	1.1E-04	tough	-0.009	2.0E-01
rebellious	0.047	1.0E-03	sounding	-0.009	8.8E-03
aroused	0.046	3.0E-06	though	-0.008	1.4E-02
unpredictable	0.046	3.4E-03	rude	-0.008	3.1E-01

Table 21: Top 15 words with greatest magnitude of RD in the specific context (§ 4.2) for two gender groups with their permutation test p-values. We use BERT as the MCQ model.