A. Gender Norms in Language

Different English words have gender-related connotations (Moon, 2014). Crawford et al. (2004) provide a corpus of 600 words and human-labeled gender scores, as scored on a 1-5 scale (1 is the most feminine, 5 is the most masculine) by samples of undergraduates at U.S. universities. They find that words referring to explicitly gendered roles such as *wife, husband, princess,* and *prince* are the most strongly gendered, while words such as *pretty, handsome* also skew strongly in the feminine/masculine directions respectively. Gender also affects the ways that people are perceived and described. A study on resumes describes "consistent differences in the way men and women with similar education, experience, and qualifications represent themselves" (Snyder, 2015). In letters of recommendation, women are described as more communal and less agentic than men, which harms the women who are perceived in this way (Madera et al., 2009). Wikipedia articles also contain significant gender differences, such as notable women's biographies focusing more on their romantic and family lives (Wagner et al., 2015).

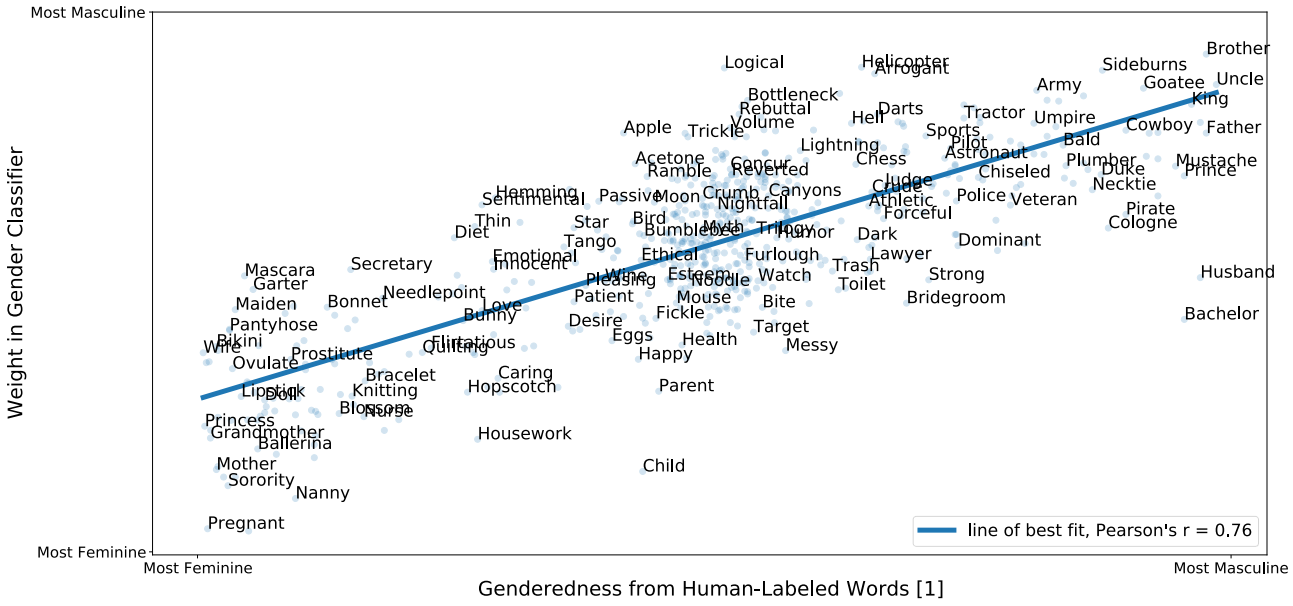

Figure 3. To validate that G effectively measures how the language of biographies adheres to gender norms, we use the corpus provided by Crawford et al. (2004). We compare the weights of the words (see Section D) in the classifier (y-axis) to the human-labeled gender scores (x-axis). We find a strong correlation (Pearson's r value 0.76) between the notion of masculine/feminine gender norms learned by G and the human-labeled gender scores. G has accuracy 0.68 for BOW and WE and 0.71 for BERT.

B. Group Fairness Metric: Gap$^{\text{RMS}}$

The fairness intervention approaches that we use focus on reducing the Gap$^{\text{RMS}}$, which is a measure of the difference in classifier performance, specifically true postiive rate (TPR), between "she" and "he" biographies. Let $a, \neg a$ be binary genders. We have

$$\text{TPR}_{c,a} = P[\hat{Y}_c = c | Y_c = c, A = a],$$

$$\text{GAP}_{c,a} = \text{TPR}_{c,a} - \text{TPR}_{c,\neg a}.$$

Then, Gap$^{\text{RMS}}$ is defined as:

$$\sqrt{\frac{1}{|C|}\sum_c \text{GAP}_{c,a}^2}.$$

Root mean square (RMS) penalizes larger values of the gender gap more heavily than a linear average across the occupations.

Table 2. Examples of Gendered Words Used in Occupation Classification

Occupation	Feminine	Masculine
Surgeon	girlfriend, hero	intern, women
Software Engineer	ball, chairman, wife	adventures, boyfriend, dress
Composer	basketball, epic, babies	entry, feminist, gorgeous, lesbian, lovely, poses, printmaker
Nurse	mary, mom, mother	boy, jesus, prison

C. Analysis on Nonbinary Dataset

We collected a dataset of biographies that use nonbinary pronouns by searching for biography-format strings (a name-like pattern followed by is a(n) ... *title*," where *title* is an occupation in the BLS Standard Occupation Classification system.[2]) with nonbinary pronouns such as "they", "xe", and "hir." Since "they" frequently refers to plural people, we manually filtered these results for the biographies that refer to individuals. PROFESSOR is the only occupation title with more than 20 such biographies; the other occupations have too few biographies to perform meaningful statistical analysis. We find that $r_{\text{PROFESSOR}}$ is negative across different approaches, following the same patterns as the values of $r_{\text{PROFESSOR}}$ obtained from the original dataset.

D. Comparing Weights of Words

To understand the mechanisms behind SNoB, we plot the importance of different words to the classifiers. We focus on the BOW and WE representations since the BERT representations are contextualized, so each word does not have a fixed importance to the model in an easily interpretable way.

For the BOW representation, each feature in the input vector v_b corresponds to a word in the vocabulary. Thus, the magnitude of the corresponding coefficient of the logistic regression classifier represents its importance to the occupation classification, while the sign (positive or negative) of the coefficient indicates whether it is correlated or anti-correlated with the positive class of the classifier.

For the WE representation, we compute the importance of each word by computing the cosine similarity, $\frac{e_w \cdot W_c}{|e_w||W_c|}$, between each word's fastText word embedding e_w and the coefficient weight vector W_c of the WE-representation classifier. Like in the BOW representation, the magnitude of this similarity represents its importance, while the sign indicates the direction of the association.

Figure 4 shows the weights in the WE, PO approach, which is the starting frame for the animations. See surgeon_words.gif and software_words.gif on Github (hyperlinked) for visualizations of how the weights of the words change between the WE, PO and WE, DE approaches.

Table 2 lists examples of words whose weights in the WE, PO approach and G are 1) magnitude at least 0.1 and 2) at least double those in the WE, DE approach. Such words are used more in WE, PO and are gendered, which suggests that they may contributing to the higher $\text{COV}(\mathbf{p}_c, \mathbf{r}_c)$ in WE, PO compared to WE, DE.

E. Robustness of Associations: Occupation-Irrelevant Gender Classifiers G^c_{irrev}

For some occupations, certain gendered words may be acceptable for the occupation classifier to use. For example, the word "computer" is learned to be masculine by the gender classifier (it is in the upper right quadrant of Figure 4). Since the word "computer" is related to the software engineering profession, it does not seem discriminatory that the occupation classifier relies on "computer" to determine whether an individual is a software engineer.

To address the concern that r_c and $\text{COV}(\mathbf{p}_c, \mathbf{r}_c)$ are quantifying simply the occurrence of occupation-related words in the biographies rather than SNoB, we trained gender classifiers G^c_{irrev}using only the words that are *task-irrelevant* for occupation c. We define task-irrelevance based on conditional independence of the word to the occupation, conditioned on gender: a word w is task-irrelevant for occupation c if

$$\text{FREQ}_w(s_c) \perp \text{FREQ}_w(\{s_c\}|c \in C), \text{FREQ}_w(h_c) \perp \text{FREQ}_w(\{h_c\}|c \in C),$$

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

where $\text{FREQ}_w(x)$ denotes the number of times w appears in x. We determine conditional independence using the chi-squared test with a 99% significance threshold. We find that this method is overly cautious in determining "task-irrelevant" words, i.e. it labels many words as task-relevant, even when they do not seem obviously related to the occupation (Figure 5). Since our purpose is to eliminate the use of task-relevant words in the gender classifier, it is acceptable to err on the side of having too few task-irrelevant words. This "conservative" estimate keeps around 40% of the words for each of the occupations. Computing the associations r_c and $\text{COV}(\mathbf{p}_c, \mathbf{r}_c)$ using G^c_{irrev}, we find that the same associations persist (Figure 6).

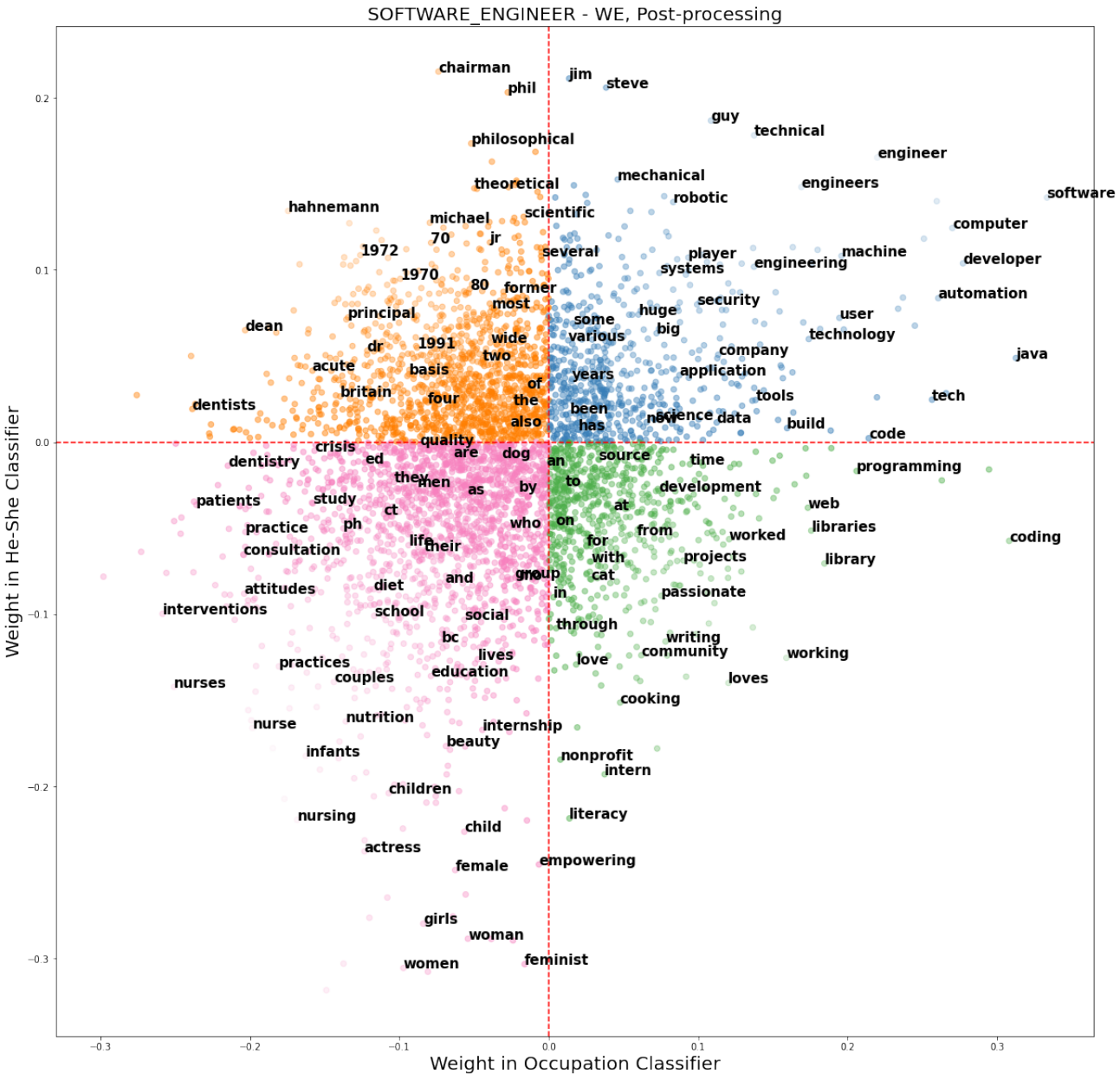

Figure 4. Visualizing words' significance in the "software engineer" occupation classifier versus the gender classifier. Each point represents a word; its x-position and y-position represents its weight in the occupation classification task and gender classifier respectively. Each point is colored based on its quadrant in the WE, PO approach. We observe that many points belong to different quadrants in the PO versus DE approaches. This is the starting frame for the animation software_words.gif (hyperlinked).

Figure 5. **Identifying Task-irrelevant Words.** Each point represents a word, plotted based on its weight in the occupation classifier (x-axis) and gender classifier (y-axis). Words are labeled as task-ir/relevant using the chi-squared test. G^c_{irrev} uses only the task-irrelevant words (blue) to learn a notion of gender norms. Our notion of task-irrelevance is overly cautious to avoid any possible proxies between gender and occupation; we see that many seemingly occupation-unrelated words are labeled as task-relevant (orange). Critically, words that are highly predictive of the occupation (large x-value) are determined as task-relevant and thus not used in G^c_{irrev}.

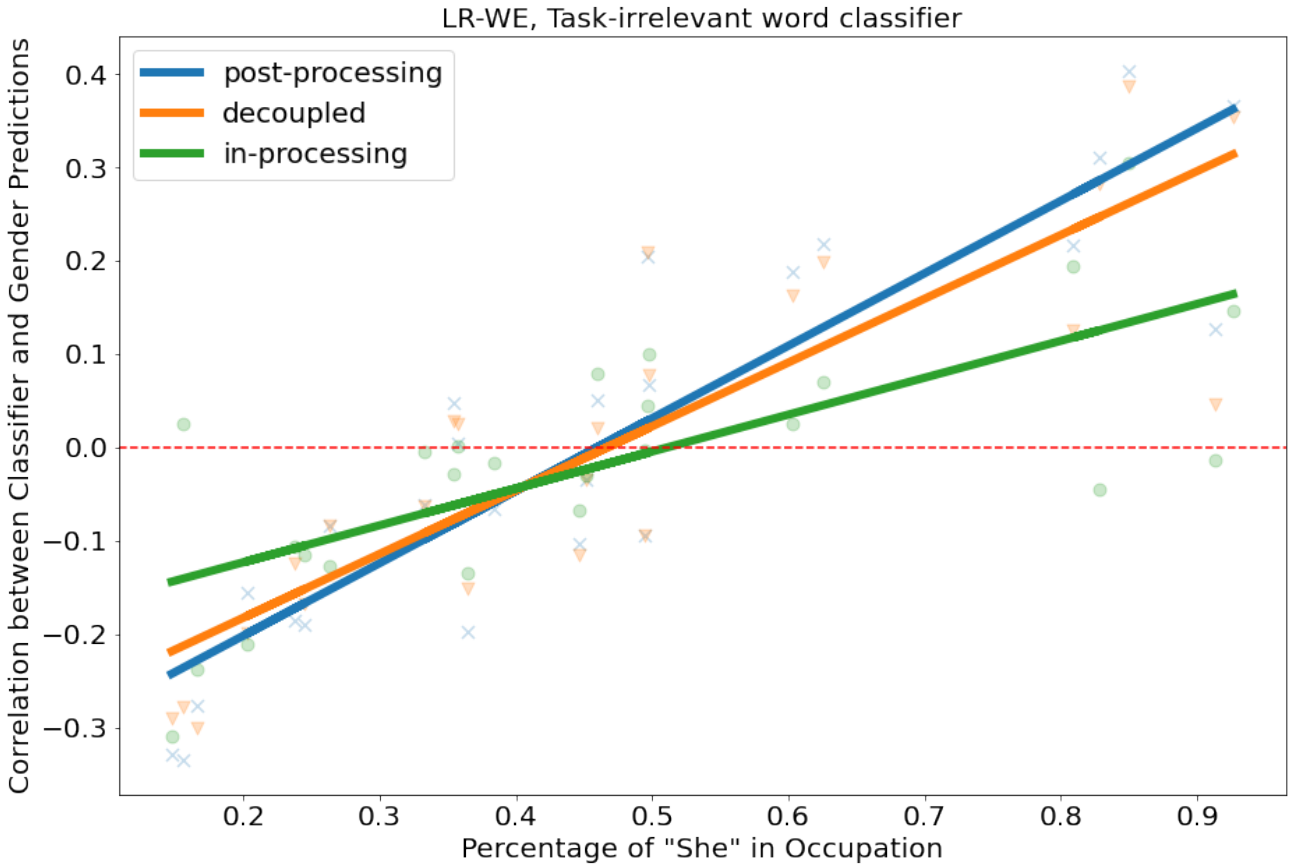

Figure 6. **SNoB using** G^c_{irrev}. This plot measures the same associations as Figure 2 in the main paper, using G^c_{irrev} instead of G to obtain gender scores. The trend is similar, which suggests that our method does not merely use occupation as a proxy for gender norms.