# Queens are Powerful too: Mitigating Gender Bias in Dialogue Generation

**Emily Dinan[*], Angela Fan[*†], Adina Williams, Jack Urbanek, Douwe Kiela, Jason Weston**
Facebook AI Research
†Laboratoire Lorrain d'Informatique et Applications (LORIA)

## Abstract

Social biases present in data are often directly reflected in the predictions of models trained on that data. We analyze gender bias in dialogue data, and examine how this bias is not only replicated, but is also amplified in subsequent generative chit-chat dialogue models. We measure gender bias in six existing dialogue datasets before selecting the most biased one, the multi-player text-based fantasy adventure dataset LIGHT (Urbanek et al., 2019), as a testbed for bias mitigation techniques. We consider three techniques to mitigate gender bias: counterfactual data augmentation, targeted data collection, and bias controlled training. We show that our proposed techniques mitigate gender bias by balancing the genderedness of generated dialogue utterances, and find that they are particularly effective in combination. We evaluate model performance with a variety of quantitative methods—including the quantity of gendered words, a dialogue safety classifier, and human assessments—all of which show that our models generate less gendered, but equally engaging chit-chat responses.

**Gendered word counts in dialogue datasets**

| Dataset | % gend. words | % male bias |
|---|---|---|
| LIGHT | 0.94 | **73.4** |
| Reddit | 1.32 | 69.76 |
| Wizard of Wikipedia | 0.076 | 65.9 |
| Daily Dialog | 1.02 | 59.04 |
| Empathetic Dialogues | 2.07 | 53.45 |
| ConvAI2 | 1.28 | 50.05 |

Table 1: **Counts of gendered words** in several dialogue datasets. We report the percent of gendered words (% gend. words) as well as the percentage of male-gendered words out of all gendered words (% male bias). Datasets are arranged in descending order with respect to *% male bias*. LIGHT has the most *% male bias*; thus we chose it as our main testbed.

## 1 Introduction

Machine learning algorithms learn to model patterns present in training datasets. In particular, they make predictions that directly reflect the harmful societal biases present in training datasets, such as racial bias in sports reports (Merullo et al., 2019) and political bias in news data (Fan et al., 2019). Such biases are rife in NLP, for example, in learned word embeddings (Bolukbasi et al., 2016; Brunet et al., 2018; Zhao et al., 2019), visual semantic role labeling (Zhao et al., 2017), natural language inference (He et al., 2019), abusive language classification (Park et al., 2018), and coreference resolution (Zhao et al., 2018a). Although research into bias in NLP writ large is maturing, bias in dialogue utterances has received somewhat less attention (Liu et al., 2019; Sheng et al., 2019; Henderson et al., 2018). As real-world use-cases for dialogue agents, such as interactive assistants, are rapidly developing, bias in dialogue models has the very real potential to invade downstream systems and exacerbate existing social biases. Thus, dialogue debiasing is becoming an increasingly important problem in NLP. In this work, we foreground dataset bias as a crucial cause of gender bias in dialogue models, and explore ways to address it.

Gender bias has been found in many machine learning datasets, in both images and text (Stock and Cissé, 2017; Zhao et al., 2017). Here, we analyze several existing dialogue datasets for gender bias (see Table 1, and §3 for more discussion) for the purpose of finding a good testbed for a deeper dive. Our analysis revealed that the dataset from the LIGHT text adventure world (Urbanek et al., 2019) was the most biased in our sample. LIGHT is also an interesting dataset for measur-

[*] Joint first authors.

*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 8173–8188,
November 16–20, 2020. ©2020 Association for Computational Linguistics

| **Persona Example (Original LIGHT Dataset)** | |
|---|---|
| *daughter:* | I spend most of my time doing household chores. I want to find meaning in life. I am energetic and happy. |
| *chief wife*: | I am the king's chief wife. Of all the women that he has married, or who are his concubines, I am the principal one. I represent the kingdom of my father, who is the king's biggest ally. My sons are the ones who will most likely become the king after the death of my husband. |
| *women:* | I live with my husband and 4 children in the village. I spend my days washing clothing and cleaning our home. My husband works for the royal army defending out town. |
| *farmer Bob's wife:* | I am farmer Bob's wife. I like to take care of all our animals. I help Farmer Bob everyday on the farm. |
| *mother:* | I am a mother of eight children. I live with my family in a cottage in the countryside. I spend every day tending to the needs of all of my little ones which can be overwhelming, but I always manage to maintain a pleasing disposition and a happy smile. |
| *wife:* | I am the wife of a farmer. While I may not be the most attractive woman ever, I am loyal and loving. My husband is a good man, but only seems to stay with me out of duty. |

Table 2: **Examples of gender biased personas** in LIGHT. In a review that we conducted in this work (section 3), none of these characters were flagged as sexist or offensive. For male examples, see Appendix Table 11.

| **Dialogue Example (Original LIGHT Dataset)** | |
|---|---|
| **wife:** I was married off by my family about five years ago. I spend my days cooking and cleaning so my husband will have something to eat when he returns from his work and can enjoy a clean home. I love my husband dearly because he works very hard to provide for us. | |
| *merchant:* | What a great day for more money. |
| *wife*: | Oh my. That is some thick dust! |
| *merchant:* | Indeed, it is very old. |
| *wife:* | This room is going to take a while to clean. You might want to come back later. |
| *merchant:* | It is fine I can set my booth up here. |
| *wife:* | With all the foot traffic? |
| *merchant:* | Yes it should be ok. |
| *wife:* | It doesn't appear that anyone ever comes up here! |
| *merchant:* | Well they will when they know I am here. |
| *wife:* | I have my doubts but I'll just go about my cleaning. |
| *merchant:* | Yea sounds like a good idea. |
| *wife:* | What is that supposed to mean? |
| *merchant:* | I am saying we should both do our jobs. |
| *wife:* | Don't take that tone with me! |

Table 3: A crowdsourced persona and dialogue from the original LIGHT dataset.

ing and mitigating gender bias for three additional reasons: first, it has multiple potential sources of bias—not just dialogues, but also characters and personas—second, it was crowdsourced, and thus susceptible to reflecting the gender biases known to be present in crowdworkers' annotations (Otterbacher et al., 2018; Barbosa and Chen, 2019), and third, LIGHT's medieval, fantasy setting might encourage crowdworkers to impart text with their gender biases.

After selecting LIGHT for particular scrutiny, we then explore three bias mitigation techniques, one of which is wholly novel, and another which is novel in its application to dialogue: (i) Counterfactual Data Augmentation (CDA) (Hall Maudslay et al., 2019; Zmigrod et al., 2019), (ii) a targeted data collection method, which we refer to as Positive-Bias Data collection, and (iii) Bias Controlled text generation. We show that these techniques are most effective in combination, resulting in dialogue models that produce engaging responses with measurably less gender bias and offensive content (see §5). Models and code are released at `https://parl.ai/projects/genderation_bias/`.

## 2 Related Work

Recently, the NLP community has focused on exploring gender bias in NLP systems (Sun et al., 2019), uncovering many gender disparities and harmful biases in algorithms and text (Cao and Daumé III 2020; Chang et al. 2019; Chang and McKeown 2019; Costa-jussà 2019; Du et al. 2019; Emami et al. 2019; Garimella et al. 2019; Gaut et al. 2020; Habash et al. 2019; Hashempour 2019; Hoyle et al. 2019; Lee et al. 2019a; Lepp 2019; Qian 2019; Qian et al. 2019; Sharifirad et al. 2019; Sharifirad and Matwin 2019; Stanovsky et al. 2019; O'Neil 2016; Blodgett et al. 2020; Nangia et al. 2020). Particular attention has been paid to uncovering, analyzing, and removing gender biases in word embeddings (Basta et al., 2019; Kaneko and Bollegala, 2019; Zhao et al., 2019, 2018b; Bolukbasi et al., 2016). This word embedding work has even extended to multilingual work on gender-marking (Gonen et al., 2019; Williams et al., 2019; Zhou et al., 2019; Williams et al., 2020). Despite these efforts, many methods for debiasing embeddings have only succeeded in *hiding* word embedding biases as opposed to *removing* them (Gonen and Goldberg, 2019)—making gender debiasing still an open area of research.

Despite the relatively ample literature on gender debiasing for word-level representations, very little work has focused on sentence representations (Liang et al., 2020; Liu et al., 2019; Sheng et al., 2019; Lee et al., 2019b). Until this point, most debiasing work on sentences mainly focus on measuring bias (Lee et al., 2019b; Sheng et al., 2019). Very few foreground the contribution of training data to gender bias in model outputs. For example, Kang et al. collect a corpus of text that is parallel across multiple stylistic categories, one of which is gender. Closer to our work, Liu et al. present a test dataset for dialogue and find that models can produce less diverse dialogues when prompted with sentences containing words describing individuals from underrepresented groups. Still, it differs from our work in that the data was created by combining templates and hand-created lists of word-pairs, rather than using real dialogue data. Liu et al. also proposes two methods for debiasing, one of which we also employ (i.e., CDA), and the other of which extends to sentences a word-embedding post-processing method (Bolukbasi et al., 2016) that has been shown to be ineffective at removing gender bias (Gonen and Goldberg 2019, but see Wang et al. 2020 for a more recent, perhaps more effective attempt). Finally—and as a direct extension of this work—Dinan et al. (2020) decomposes gender bias along three semantic-pragmatic dimensions, and show that train more fine-grained classifiers allow for more accurate classification of dataset gender biases. The novelty of the present contribution lies in how we measure bias, and in the joint application of our three gender debiasing methods.

## 3 Measuring Bias

Before one can mitigate bias, one must first measure it. As a first pass, we measured the counts of gendered words used (using a word list from Zhao et al. 2018b), and the percent of those which referred to male characters for six datasets (Table 1). We count the number of male and female gendered words in the training sets of several datasets (LIGHT, ConvAI2, Reddit, Wizard of Wikipedia, Daily Dialog, Empathetic Dialogues, and ConvAI2). We use this to calculate the percentage of gendered words out of all words, and the *% male bias*, that is the percentage of male gendered words among all gendered words in a dialogue. We find that LIGHT is the most gender imbalanced dataset among all datasets in this table, with a *% male bias* of 73%, although others, like Reddit, are close behind.

Since LIGHT was found to be the most gender biased, we qualitatively examine it more closely, and find many biased utterances present in the training data. For example, the *queen* persona adheres to negatively stereotyped gender roles when uttering the line *I spend my days doing embroidery and having a talk with the ladies*. Another character admires a *sultry wench with fire in her eyes*. We conclude from examples like this that presenting crowdworkers with gender biased personas often leads them to create even more gender biased dialogues (see Table 3): for example, a *wife* persona contains the text *I spend my days cooking and cleaning so my husband will have something to eat when he returns from his work...*, and, in dialogue with a *merchant*, discusses only her cleaning duties. The *merchant* even derisively refers to cleaning as the *wife's* job. This could be an effect of gender stereotype priming (Blair and Banaji, 1996; Steele and Ambady, 2006; Oswald, 2008; Derks et al., 2011; Verhaeghen et al., 2011).

Given this, we wonder how much biased character names and personas themselves lead to LIGHT dialogues being more biased than the others. Thus, we focus on persona-based dialogue text in particular for the remainder of the paper. Dialogue research has found that, while incorporating personas increases engagingness and improves consistency (Zhang et al., 2018; Shuster et al., 2018; Mazaré et al., 2018; Olabiyi et al., 2018; Li et al., 2016b), they can also crystallize gender bias (Clark et al., 2019; Henderson et al., 2018). Such bias propagates to subsequently generated conversations. Crowdworkers in particular might imbue their annotations with their particular gender biases at every stage of dataset creation. For example, LIGHT (Urbanek et al., 2019) was created by crowdworkers in stages: crowdworkers were first assigned a **character** (with previously crowdsourced names such as “farmer” or “witch”), as well as a previously crowdsourced **persona**, or short textual description of the char-

|  | # Characters |  |  |  | # Ref. |  |
|---|---|---|---|---|---|---|
|  | *F* | *M* | *N* | *All* | *F* | *M* |
| ***LIGHT*** |  |  |  |  |  |  |
| Orig. Data | 159 | 258 | 1460 | 1877 | 439 | 1238 |
| Swap Persona | 336 | 230 | 694 | 1260 | 1419 | 1030 |
| New Charac. | 151 | 120 | 1448 | 1719 | 357 | 275 |
| *Total* | 646 | 608 | 3602 | 4856 | 2215 | 2543 |
| ***ConvAI2*** |  |  |  |  |  |  |
| Orig. Data | 1109 | 1048 | 4214 | 6371 | 1283 | 1148 |

Table 4: **Analysis of gender in LIGHT and ConvAI2**: The LIGHT dataset is compared to similar novel datasets obtained after either gender-swapping character and personas or collecting wholly new ones. # Characters refers to the counts of gendered characters and # Ref. refers to counts of gendered references in personas. The original LIGHT dataset is skewed towards male characters, while ConvAI2 contains both male and female in a roughly equal proportions.

acter. Then, they were paired up, and tasked with generating a **dialogue** as those characters.

To determine with more granularity precisely how bias manifests in persona-based dialogue datasets, we investigate the text for (i) characters such as *fisherman* (Table 1), and (ii) personas such as *I love fishing* (Table 2). We ask: (i) do crowdworkers generate male and female characters at an equal rate, (ii) do they imbue characters' personas with sexism or undesirable gender biases?

**Bias in Number of Characters.** We first determine whether crowdworkers create an equal number of male and female characters. To quantify this, we asked annotators on Amazon Mechanical Turk to label the gender of each character name based on its persona description (choosing *neutral* if the gender was not explicit). This annotation is possible because many personas include text such as *I am a young woman.*[1] Since this measurement requires personas, we consider the two persona-based dialogue datasets in our sample: LIGHT and ConvAI2 (Zhang et al., 2018). LIGHT is highly gender imbalanced: there are over 1.6 times as many male characters as female ones[2]. LIGHT is also considerably less gender-balanced than ConvAI2, which has a nearly equal number of male and female gendered personas (see Table 4).

**Bias in Personas.** In addition to the stark underrepresentation of female characters, the medieval setting in LIGHT is likely to encourage crowdworkers to generate dialogues accentuating historical biases and inequalities of the time period (Bowman, 2010; Garcia, 2017). We investigate the number of references to men or women in the text of personas, as another source of bias. Take for example, a female persona that contains a gendered reference such as *I want to follow in my **father's** footsteps* rather than *in my **mother's**.* Although using gendered **relational nouns** (Barker, 1992; Williams, 2018), such as *father*, doesn't always signal sexism, if female characters are predominantly defined in reference to male characters, it becomes a problem. We count the appearance of gendered words in personas using the list compiled by Zhao et al. (2018b), and find that men are disproportionately referred to in the personas: there are nearly 3x as many mentions of men than women, which suggests that a large number of characters are defined by their relationships to men (see Table 2 for examples, and Table 4 for counts).

Gender bias and sexism are clearly present in many dialogue datasets (Henderson et al., 2018), but finding a clear way to define these terms (and others that categorize unsafe text), let alone measure their effects at scale, is very challenging. For example, the persona for the character *girl* contains the line *I regularly clean and cook dinner* (see Table 2 for more examples), which strikes us as stereotypical and sexist, but it might not be noticed by others. In this paper, we rely on each annotator's own, subjective, definition(s) of the term but aggregate multiple opinions. Three naïve annotators examined each persona for unsafe content. If annotators detected content was 'offensive' or 'maybe offensive', they were asked to select one of four categories—racist, sexist, classist, other—and to provide a reason for their response. Just over 2% of personas were flagged by at least one annotator, and these personas and their resulting dialogues were removed.

## 4 Mitigating Bias in Generative Dialogue

In this section, we present a *general* framework for mitigating bias in generative dialogue. More specifically, we explore data augmentation and other algorithmic methods to mitigate bias in generative Transformer models. We (i) extend counterfactual data augmentation to dialogue

[1] Note that our procedure doesn't preclude annotators from implicitly assuming genders for ungendered personas, such as "doctor", which may widen the gender gap.

[2] When we use "female" and "male"—rather than "woman" and "man"—we want our reference to include characters that are binarily gendered, but not necessarily human.

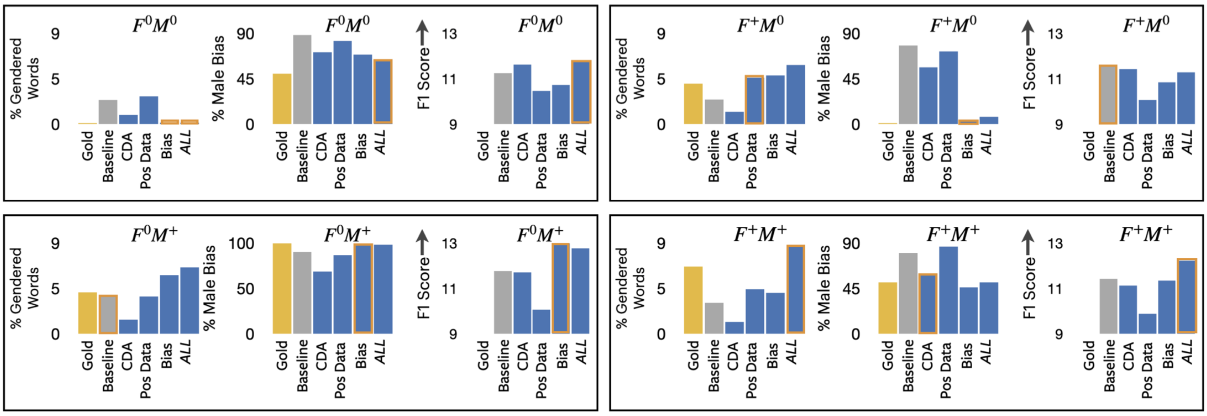

Figure 1: We compare the **performance of various bias mitigation methods**: Counterfactual Data Augmentation (CDA), Positive-Bias Data Collection (Pos. Data), Bias Control Model (Bias Ctrl), and combining these methods (*ALL*). We split test set across the four genderedness bins: $F^{0/+}M^{0/+}$. $X^0$ indicates there are no X-gendered words in the gold response, while $X^+$ indicates that there is at least one. We measure the percent of gendered words generated in the dialogue (% gend. words) and the percent of male bias (% male bias), i.e. the percent of male-gendered words out of all generated gendered words. While each of these methods yield some improvement, *combining them yields the best control over the genderedness of the utterances while improving the F1-score.* The orange outline represents the best performing model. For % Gendered words, lower is better. For % Male Bias, closer to 50 is better. For F1 Score, higher is better.

(Hall Maudslay et al., 2019; Zmigrod et al., 2019) following (Liu et al., 2019), (ii) perform positive data collection by augmenting the existing dataset via targeted data collection with crowdworkers, and lastly, (iii) apply controllable generation techniques to gender bias to control how many male and female gendered words models produce.

### 4.1 Counterfactual Data Augmentation

A straightforward solution for gender bias in embeddings is *Counterfactual Data Augmentation* (CDA) (Hall Maudslay et al., 2019; Zmigrod et al., 2019; Liu et al., 2019). CDA swaps, say, all instances of *grandmother* with *grandfather*, *she* with *he*, etc. We apply this word-based augmentation to dialogue by first copying every dialogue, then swapping all gendered words with their counterpart from the paired list in Zhao et al. (2018b). The augmentation is limited to words on the list, and the swapping is performed automatically. The model is then retrained on the augmented data. While CDA is somewhat effective strategy for mitigating bias in word embeddings, this method has several pitfalls: it may result in ungrammatical sentences, and it relies on existing (and perhaps incomplete) lists to determine and swap gender.

### 4.2 Positive-Bias Data Collection

To resolve the issues with CDA, we use humans to collect additional dialogue data via a two-pronged Positive-Bias Data Collection (Pos. Data) strategy. We first collect additional personas by having humans (i) manually swap the gender of the character name and all gendered references in the character's persona text (rather than relying on brittle word lists) and (ii) write additional, diversified personas. We then use these personas to seed the collection of additional, positively biased dialogue data, which we refer to as Pos. Data throughout.

**New Characters & Personas.** When a dataset contains more male characters and references to male characters than it contains female characters and references to female characters (see Table 4), we balance existing characters and personas with **gender swapping**. For every gendered character-persona pairing, annotators create a new opposite-gendered character-persona pairing for which animate nouns or pronouns are changed, but the rest of the persona remains unchanged. For example, for every persona describing a male character like a king, annotators will create a new one describing a female character like a queen. Annotators are instructed to swap the gender(s) of other animate references in the text (e.g., if an original persona describes a woman in relation to her father, the new male persona will describe a man in relation to his mother). This method ensures that the created sentences will be grammatical, unlike heuristic data augmentation.

However, simply balancing references to men and women is insufficient, as female characters might be specifically described in sexist ways (see §3). As detecting sexism is challenging (also see §3), we take our qualitative analysis to be sufficient motivation, and moved to further offset the bias by collecting a new set of *interesting* and *independent* female characters. We primed workers by showing examples of gender underspecified character names like *adventurer* with personas like *I am a woman passionate about exploring a world I have not yet seen. I embark on ambitious adventures*. We also provided crowdworkers with additional instruction to encourage them to create diverse characters: *We're looking for strong and diverse descriptions. Avoid descriptions that could be considered hateful, offensive, or stereotypical.* Even with explicit instruction, annotators created 3 times as many male characters as female characters, revealing the stubbornness of the inherent gender biases of the available crowdworker pool. We ultimately exclude all male-gendered personas created in this fashion from the new dataset, as including them would worsen the gender balance of the dataset. Our new dataset is approximately balanced then in the number of male or female characters and in the number of references to male or female characters (see Table 4). In total, we add 2,629 new characters and release the data for optional inclusion in the LIGHT dataset.

**New Dialogues.** After gender-balancing the personas, we moved on to using the gender-balanced personas to crowdsource additional, hopefully gender-balanced, dialogues. We selected more female-gendered characters for new dialogue collection, and explicitly instructed annotators to be mindful of gender bias. In particular, we encouraged them to assume *equality*—social, economic, political, or otherwise—between genders (Note: this is uniquely possible with a dataset like LIGHT, which is situated in a fully fictional world). We collected a total of 507 new dialogues containing 6,658 utterances (approximately 6% of the original dataset size). We refer to this additional dialogue data as Pos. Data.

### 4.3 Bias Controlled Training

Gender bias in dialogue can take the form of imbalanced use of gendered words. To create dialogue models that can generate an equal number of gendered words, we control model output with Bias Control (Bias Ctrl) via conditional training. Previous conditional training models learn to associate specific control tokens with some desired text properties (Kikuchi et al., 2016; Fan et al., 2018a; Oraby et al., 2018; See et al., 2019), but have not been applied to address bias issues.

|  | $F^0M^0$ | $F^0M^+$ | $F^+M^0$ | $F^+M^+$ |
|---|---|---|---|---|
| % of test set | 60.65 | 27.21 | 7.61 | 4.63 |

Table 5: **Percentage of dialogue examples in each of the four genderedness bins** —$F^{0/+}M^{0/+}$— for the LIGHT dialogue data test set.

We apply conditional training techniques to control gender bias in generative dialogue by learning to associate control tokens with properties of gender bias. Any general function that takes as input a dialogue utterance and outputs a continuous or discrete value that provides information about gender bias could be used as a control variable. In our case, prior to training, each dialogue response is binned into one of four bins—$F^{0/+}M^{0/+}$ —where $X^0$ indicates that there are zero X-gendered words in the response. $X^+$ indicates the presence of one or more X-gendered word. The percentage of test set examples that fall into each bin is in Table 5. Nouns and adjectives are binned into gendered bins via an aggregation of existing gendered word lists (Zhao et al., 2018b,a; Hoyle et al., 2019). Note that other functions could be used as well, such as a bias classifier (Dinan et al., 2020).

We append a special token to the input that indicates which bin the response falls into. During Bias Ctrl training, the model should learn to associate the special token with the genderedness of the dialogue response, such that at inference time, we could append different special tokens to control the genderedness of the model output. For example, a model trained with multiple gender control bins could be set to the gender neutral (in this case, $F^0M^0$) setting at inference time, to produce a response containing few (or no) gendered words.

### 4.4 Implementation Details

Following Urbanek et al. (2019), we fine-tune a large, pre-trained Transformer encoder-decoder on the dialogues in the LIGHT dataset for all generation experiments. Following Humeau et al. (2019), we pre-trained on Reddit conversations extracted and obtained by a third party, and made avail-

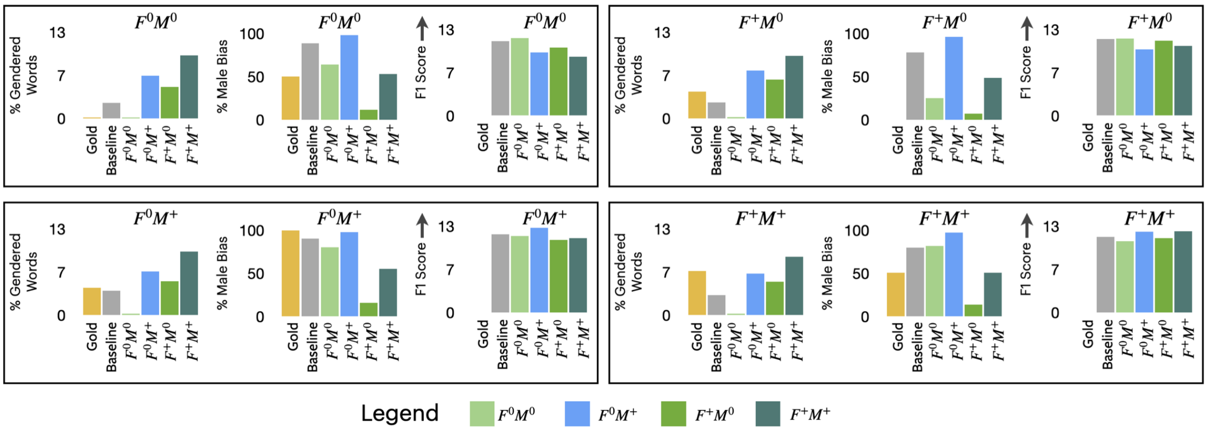

Figure 2: **Performance of the *ALL* debiasing model** controlled by indicating specific bins for all examples at test time. We report results for each possible conditioning bin choice. Across bins (at the top of graphs), the model maintains performance as measured by F1 whilst *radically changing the genderedness of the language generated.*

able on `pushshift.io`. During pre-training, models learned to generate a comment conditioned on the preceding conversation thread. All comments that contained URLs or were shorter than 5 characters long were removed, along with child comments, resulting in approximately 2.2 billion training examples. Similarly during fine-tuning, models were conditioned on the full preceding dialogue history. All models are 8-layer encoders, 8-layer decoders, with 512 dimensional embeddings and 16 attention heads based on the ParlAI transformer implementation (Miller et al., 2017). We decode with a beam search size of 5.

## 5 Results

We train five Transformer models: one baseline trained only on original LIGHT without any mitigation techniques, one Transformer for each of our three methods (see §4.1 for CDA, §4.2 for Positive-Bias Data Collection, and §4.3 for Bias Control), and a final one combining all three methods (*ALL*) that achieves the best results.

**Bias is Amplified in Generation.** Figure 1 compares the performance of the various techniques. We compare our methods to the gold labels from the test set and to the baseline. To do this, we divide the test set into four *genderedness bins* (as defined in §4.3)—$F^0M^0$, $F^0M^+$, $F^+M^0$, and $F^+M^+$—and calculate: (i) the F1 word overlap with the gold response, (ii) the percentage of gendered words generated (% gend. words), and (iii) the percentage of male-gendered words generated (relative to the sum total of gendered words generated by the model). We find that Transformer models not only reflect dataset biases, but also they *amplify* them. When the model produces gendered words, it generates male-gendered words the vast majority of the time. Even when the gold label only contains female-gendered words ($F^+M^0$), it still generates male-gendered words nearly 78% of the time.

**Comparing Debiasing Methods** As shown in Figure 1, each method improves on the metrics—*% gendered words*, *% male bias*, and F1—over the baseline Transformer, but we find that combining all methods (*ALL*) is most advantageous. While *ALL* has more data than CDA and Bias Ctrl, more data alone is not enough—the Positive-Bias Data Collection model does not achieve as strong results as *ALL* despite also having more data. Both the Bias Ctrl and *ALL* models benefit from knowing the data split ($F^0M^0$, for example), and both yield a gender ratio closest to ground truth.

**Bias Controlled Training Controls Gendered Words.** Our Bias Ctrl method can control the number of gendered words in generated dialogues (Figure 2). We examine the effect of Bias Ctrl by generating responses conditioning the *ALL* model on each bin. We observe that changing the bin radically changes the genderedness of generated text with only small differences in overall F1, which shows that the Bias Ctrl method is efficacious.

Examples of generated text from both the baseline and the *ALL* model are shown in Table 6. Further examples are provided in the Appendix in Table 12. The baseline model generates male-gendered words when the gold response contains no gendered words or only female-gendered

| Generation Examples | |
|---|---|
| Bin | F$^0$M$^0$ |
| Context | *[P1:]* I don't remember anything I was hoping you could find out.<br>*[P2:]* Your form is very hazy. Do you remember if you are a man or woman? |
| Baseline: | *i am not a man. i am a man of the forest.* |
| ALL: | *no, i don't remember.* |
| Gold: | *I don't know what's wrong with me!* |
| Bin | F$^+$M$^0$ |
| Context | *[P1:]* I do not believe my eyes, for an angel is upon me! Angel, please tell me your name.<br>*[P2:]* My name is Abigail! |
| Baseline: | *my name is abigail. i am the king of this kingdom.* |
| ALL: | *i am the queen's daughter!* |
| Gold: | *Abigail! Such a beautiful name. To what do I owe the pleasure of meeting you?* |

Table 6: **Example generations** from the baseline model and the proposed debiased models. Ground truth ('Gold') either contains no gendered words or only female-gendered words, but the baseline model still generates male-gendered words.

| | Gold Labels | Baseline | *ALL* |
|---|---|---|---|
| % Offensive | 13.0 | 14.25 | **10.37** |

Table 7: **Offensive language classification** of model responses on the LIGHT dialogue test set. The *ALL* model generates a lower percentage of offensive utterances.

words, even generating unlikely sequences such as *my name is abigail. i am the king of this kingdom.* For various methods, we compute the top 20 words generated on the test set (after removing stop words), shown in Appendix Table 8. We denote gendered nouns using an asterisk. Among the top 20 words generated by the baseline, there are only two gendered nouns—*knight* and *king*—both male-gendered. The *ALL* model generates similar words, but also features *queen* in its top 20, another indication that gender is more balanced.

## 5.1 Safety of Generated Text

To further evaluate our techniques, we investigate whether the *ALL* model generates fewer offensive utterances than (i) the baseline, and (ii) the human-generated gold labels. Our bias mitigation techniques have the ancillary benefit of producing models that generate proportionately fewer offensive utterances; see Table 7 for results.

We use a Transformer-based dialogue safety classifier to classify model-generated utterances as offensive or safe following Liu et al. (2019). The classifier was fine-tuned on an offensive language classification task (Dinan et al., 2019), and achieves state-of-the-art results. We apply this classifier to each utterance generated by the *ALL* model and baseline models on the test set, in addition to the gold (human generated) labels from the test set. The dialogue safety classifier rates our proposed *ALL* model as less offensive than both the baseline model and the ground truth (gold) labels, which argues in favor of the efficacy of our debiasing methods.

## 5.2 Human Evaluation: Bias and Quality

We compare the quality of our debiasing methods using human evaluation. One might hypothesize that some gender debiasing methods work by replacing contentful words (e.g., *witch*) with bleached or uninteresting ones (e.g., *person*, *thing*), effectively trading off gender bias with engagingness. Generative models in particular are well-known to produce generic text (Li et al., 2016a; Fan et al., 2018b), which is often less engaging. Overreliance on generic text might increase the chances of biases such as **androcentrism**, or the propensity of societies to consider men central but women peripheral (Bem, 1993; Bailey et al., 2020); in language, male-gendered words often act as a gender-neutral standard (Bailey et al., 2019), as in Neil Armstrong's 1969 quote "one small step for a *man*, one giant leap for *man*kind". We use the dialogue evaluation system Acute-Eval (Li et al., 2019) to ask evaluators to compare pairs of conversations from models and decide which model generates (i) more biased dialogues and (ii) more engaging dialogues. We collect 100 model conversations with crowdworkers per method. Then, we compare conversations between a human and the baseline model to conversations between a human and the *ALL* model with all generations set to the F$^0$M$^0$ gender-neutral control bin. We found that asking for predictions of speaker gender was more effective than asking about sexism directly.

As shown in Figure 3, predicting the gender accurately of *ALL* model generations is more challenging (significant at $p < 0.01$ with a t-test), but the responses are just as engaging according to human evaluators. We conclude our proposed methods are able to help mitigate gender bias without degrading dialogue quality.

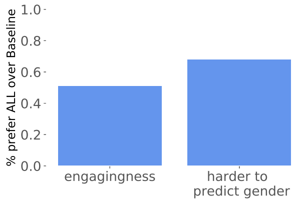

Figure 3: **Human Evaluation of *ALL* model** ($F^0M^0$) compared to baseline Transformer generative model. Evaluators choose which model output they prefer for dialogue engagingness and difficulty of predicting speaker gender. The ALL model produces less gendered text while engagingness is not affected.

## 6 Conclusion

We analyze gender bias in dialogue data and resulting model generations for models trained on dialogue data. We propose general purpose techniques for reducing gender bias in generated text. The methods described in this paper combine data augmentation, positive-bias data collection, and bias controlled training. We note that our results show that data collection techniques help mitigate issues, so when it is possible, bias should be considered at the earliest stages of a project. Newly collected or constructed datasets should consider how to carefully craft the collection to mitigate bias issues from the very start. When this is not possible, however, such as in the case of using real-world data or a dataset that already exists, the techniques presented in this paper are shown to be effective at reducing gender bias. They are especially effective when combined, producing less gendered, more balanced, safer utterances that maintain the engagingness of the dialogue.

## Acknowledgements

Thanks to Isabelle Kloumann, Ledell Wu, and Hila Gonen for comments and advice on this project.

## References

April H Bailey, Marianne LaFrance, and John F Dovidio. 2019. Is man the measure of all things? a social cognitive account of androcentrism. *Personality and Social Psychology Review*, 23(4):307–331.

April H Bailey, Marianne LaFrance, and John F Dovidio. 2020. Implicit androcentrism: Men are human, women are gendered. *Journal of Experimental Social Psychology*, 89:103980.

Natã M Barbosa and Monchu Chen. 2019. Rehumanized crowdsourcing: A labeling framework addressing bias and ethics in machine learning. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 543. ACM.

Chris Barker. 1992. *Possessive descriptions*. Ph.D. thesis, University of California, Santa Cruz.

Christine Basta, Marta R. Costa-jussà, and Noe Casas. 2019. Evaluating the underlying gender bias in contextualized word embeddings. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 33–39, Florence, Italy. Association for Computational Linguistics.

Sandra L Bem. 1993. *The lenses of gender: Transforming the debate on sexual inequality*. Yale University Press.

Irene V Blair and Mahzarin R Banaji. 1996. Automatic and controlled processes in stereotype priming. *Journal of personality and social psychology*, 70(6):1142.

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

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in neural information processing systems*, pages 4349–4357.

Sarah Lynne Bowman. 2010. *The functions of role-playing games: how participants create community, solve problems and explore identity*. McFarland and Co.

Marc-Etienne Brunet, Colleen Alkalay-Houlihan, Ashton Anderson, and Richard Zemel. 2018. Understanding the origins of bias in word embeddings. *arXiv preprint arXiv:1810.03611*.

Yang Trista Cao and Hal Daumé III. 2020. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online. Association for Computational Linguistics.

Kai-Wei Chang, Vinod Prabhakaran, and Vicente Ordonez. 2019. Bias and fairness in natural language processing. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): Tutorial Abstracts*, Hong Kong, China. Association for Computational Linguistics.

Serina Chang and Kathy McKeown. 2019. Automatically inferring gender associations from language. 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 5745–5751, Hong Kong, China. Association for Computational Linguistics.

Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. 2019. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. *arXiv preprint arXiv:1909.03683*.

Marta R Costa-jussà. 2019. An analysis of gender bias studies in natural language processing. *Nature Machine Intelligence*, pages 1–2.

Belle Derks, Colette Van Laar, Naomi Ellemers, and Kim De Groot. 2011. Gender-bias primes elicit queen-bee responses among senior policewomen. *Psychological science*, 22(10):1243–1249.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Emily Dinan, Samuel Humeau, Bharath Chintagunta, and Jason Weston. 2019. Build it break it fix it for dialogue safety: Robustness from adversarial human attack. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4537–4546, Hong Kong, China. Association for Computational Linguistics.

Yupei Du, Yuanbin Wu, and Man Lan. 2019. Exploring human gender stereotypes with word association test. 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 6133–6143, Hong Kong, China. Association for Computational Linguistics.

Ali Emami, Paul Trichelair, Adam Trischler, Kaheer Suleman, Hannes Schulz, and Jackie Chi Kit Cheung. 2019. The KnowRef coreference corpus: Removing gender and number cues for difficult pronominal anaphora resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3952–3961, Florence, Italy. Association for Computational Linguistics.

Angela Fan, David Grangier, and Michael Auli. 2018a. Controllable abstractive summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 45–54, Melbourne, Australia. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018b. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6342–6348, Hong Kong, China. Association for Computational Linguistics.

Antero Garcia. 2017. Privilege, power, and dungeons & dragons: How systems shape racial and gender identities in tabletop role-playing games. *Mind, Culture, and Activity*, 24(3):232–246.

Aparna Garimella, Carmen Banea, Dirk Hovy, and Rada Mihalcea. 2019. Women's syntactic resilience and men's grammatical luck: Gender-bias in part-of-speech tagging and dependency parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3493–3498, Florence, Italy. Association for Computational Linguistics.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online. Association for Computational Linguistics.

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.

Hila Gonen, Yova Kementchedjhieva, and Yoav Goldberg. 2019. How does grammatical gender affect noun representations in gender-marking languages? In *Proceedings of the 2019 Workshop on Widening NLP*, pages 64–67, Florence, Italy. Association for Computational Linguistics.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165, Florence, Italy. Association for Computational Linguistics.

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.

Reyhaneh Hashempour. 2019. A deep learning approach to language-independent gender prediction on Twitter. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 92–94, Florence, Italy. Association for Computational Linguistics.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 132–142, Hong Kong, China. Association for Computational Linguistics.

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical challenges in data-driven dialogue systems. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 123–129.

Alexander Miserlis Hoyle, Lawrence Wolf-Sonkin, Hanna Wallach, Isabelle Augenstein, and Ryan Cotterell. 2019. Unsupervised discovery of gendered language through latent-variable modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1706–1716, Florence, Italy. Association for Computational Linguistics.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Real-time inference in multi-sentence tasks with deep pretrained transformers. *arXiv preprint arXiv:1905.01969*.

Masahiro Kaneko and Danushka Bollegala. 2019. Gender-preserving debiasing for pre-trained word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1641–1650, Florence, Italy. Association for Computational Linguistics.

Dongyeop Kang, Varun Gangal, and Eduard Hovy. 2019. (male, bachelor) and (female, Ph.D) have different connotations: Parallelly annotated stylistic language dataset with multiple personas. 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 1696–1706, Hong Kong, China. Association for Computational Linguistics.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1328–1338, Austin, Texas. Association for Computational Linguistics.

Nayeon Lee, Yejin Bang, Jamin Shin, and Pascale Fung. 2019a. Understanding the shades of sexism in popular TV series. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 122–125, Florence, Italy. Association for Computational Linguistics.

Nayeon Lee, Andrea Madotto, and Pascale Fung. 2019b. Exploring social bias in chatbots using stereotype knowledge. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 177–180, Florence, Italy. Association for Computational Linguistics.

Haley Lepp. 2019. Pardon the interruption: Automatic analysis of gender and competitive turn-taking in united states supreme court hearings. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 143–145, Florence, Italy. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003, Berlin, Germany. Association for Computational Linguistics.

Margaret Li, Jason Weston, and Stephen Roller. 2019. Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. *arXiv preprint arXiv:1909.03087*.

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.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does gender matter? Towards fairness in dialogue systems. *CoRR*, abs/1910.10486.

Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training millions of personalized dialogue agents. *arXiv preprint arXiv:1809.01984*.

Jack Merullo, Luke Yeh, Abram Handler, Alvin Grissom II, Brendan O'Connor, and Mohit Iyyer. 2019. Investigating sports commentator bias within a large corpus of American football broadcasts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6354–6360, Hong Kong, China. Association for Computational Linguistics.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. *arXiv preprint arXiv:2010.00133*.

Oluwatobi O Olabiyi, Anish Khazane, and Erik T Mueller. 2018. A persona-based multi-turn conversation model in an adversarial learning framework. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 489–494. IEEE.

Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

Shereen Oraby, Lena Reed, Shubhangi Tandon, TS Sharath, Stephanie Lukin, and Marilyn Walker. 2018. Controlling personality-based stylistic variation with neural natural language generators. *arXiv preprint arXiv:1805.08352*.

Debra L Oswald. 2008. Gender stereotypes and women's reports of liking and ability in traditionally masculine and feminine occupations. *Psychology of Women Quarterly*, 32(2):196–203.

Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating user perception of gender bias in image search: the role of sexism. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 933–936. ACM.

Ji Ho Park, Jamin Shin, and Pascale Fung. 2018. Reducing gender bias in abusive language detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2799–2804, Brussels, Belgium. Association for Computational Linguistics.

Yusu Qian. 2019. Gender stereotypes differ between male and female writings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 48–53, Florence, Italy. Association for Computational Linguistics.

Yusu Qian, Urwa Muaz, Ben Zhang, and Jae Won Hyun. 2019. Reducing gender bias in word-level language models with a gender-equalizing loss function. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 223–228, Florence, Italy. Association for Computational Linguistics.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. 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 1702–1723, Minneapolis, Minnesota. Association for Computational Linguistics.

Sima Sharifirad, Alon Jacovi, Israel Bar Ilan Univesity, and Stan Matwin. 2019. Learning and understanding different categories of sexism using convolutional neural network's filters. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 21–23.

Sima Sharifirad and Stan Matwin. 2019. Using attention-based bidirectional LSTM to identify different categories of offensive language directed toward female celebrities. In *Proceedings of the 2019 Workshop on Widening NLP*, pages 46–48.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2019. The woman worked as a babysitter: On biases in language generation. 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 3398–3403, Hong Kong, China. Association for Computational Linguistics.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Engaging image chat: Modeling personality in grounded dialogue. *arXiv preprint arXiv:1811.00945*.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Jennifer R Steele and Nalini Ambady. 2006. "Math is hard!" the effect of gender priming on women's attitudes. *Journal of Experimental Social Psychology*, 42(4):428–436.

P. Stock and M. Cissé. 2017. Convnets and imagenet beyond accuracy: Explanations, bias detection, adversarial examples and model criticism. *arXiv:1711.11443v2*.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. 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 673–683, Hong Kong, China. Association for Computational Linguistics.

Paul Verhaeghen, Shelley N Aikman, and Ana E Van Gulick. 2011. Prime and prejudice: Co-occurrence in the culture as a source of automatic stereotype priming. *British Journal of Social Psychology*, 50(3):501–518.

Tianlu Wang, Xi Victoria Lin, Nazneen Fatema Rajani, Bryan McCann, Vicente Ordonez, and Caiming Xiong. 2020. Double-hard debias: Tailoring word embeddings for gender bias mitigation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5443–5453, Online. Association for Computational Linguistics.

Adina Williams. 2018. *Representing Relationality: MEG Studies on Argument Structure*. Ph.D. thesis, New York University.

Adina Williams, Damian Blasi, Lawrence Wolf-Sonkin, Hanna Wallach, and Ryan Cotterell. 2019. Quantifying the semantic core of gender systems. 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 5733–5738, Hong Kong, China. Association for Computational Linguistics.

Adina Williams, Ryan Cotterell, Lawrence Wolf-Sonkin, Damián Blasi, and Hanna Wallach. 2020. On the relationships between the grammatical genders of inanimate nouns and their co-occurring adjectives and verbs. *arXiv preprint arXiv:2005.01204*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender bias in contextualized word embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 629–634, Minneapolis, Minnesota. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. ”men also like shopping: Reducing gender bias amplification using corpus-level constraints”. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018a. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018b. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining gender bias in languages with grammatical gender. 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 5275–5283, Hong Kong, China. Association for Computational Linguistics.

Ran Zmigrod, Sebastian 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 Appendix

## A.1 Discussion of Generation Quality

**Generality of Gendered Words.** The gendered word lists used may not be comprehensive (Zhao et al., 2018a,b; Hoyle et al., 2019). For example, they do not include *hag* or *wench*, which are common in LIGHT. Further, a more continuous representation of gender should be used in the future.

**More Fine-Grained Control.** We present an effective method to control the quantity of gendered words generated by manipulating control bins. This technique is general and could be used to control other properties of generated utterances. For example, a sexism or bias classifier could be used instead of the gendered word list.

**Quality of Generated Dialogue.** Generative dialogue models are prone to overuse frequent words and produce generic utterances, the so-called *I don't know* problem (Li et al., 2016a). We also observe these effects which can affect bias.

| Model | Top 20 generated words |
|---|---|
| Baseline | sorry, hear, not, what, glad, doing, don, king*, thank, sure, will, your, can, much, do, know, but, knight*, blacksmith, going |
| *ALL* | sorry, hear, sure, not, what, help, doing, your, course, trying, glad, thank, queen*, don, good, king*, but, yes, know, sir* |
| *ALL* $F^0M^0$ | sorry, hear, sure, what, not, doing, glad, thank, your, yes, course, but, don, do, know, help, have, enjoying, fool, much |
| *ALL* $F^0M^+$ | sorry, hear, help, trying, sure, good, king*, sir*, not, your, day, course, father*, he*, don, thank, happy, guard*, glad, have |
| *ALL* $F^+M^0$ | sorry, hear, queen*, sure, miss*, not, your, thank, how, hello, today, guard*, she*, yes, course, kind, woman*, help, glad, what |
| *ALL* $F^+M^+$ | sorry, queen*, hear, guard*, help, trying, your, sure, good, course, day, knight*, not, protect, yes, friend, king*, woman*, she*, thank |

Table 8: **Genderedness bins control the genderedness of generated text**. The top 20 words (test set) with stop words removed. * indicates gendered nouns.

| *Data Split:* | $F^0M^0$ | | | $F^0M^+$ | | | $F^+M^0$ | | | $F^+M^+$ | | | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Model* | % gend. words | % male bias | F1 score | % gend. words | % male bias | F1 score | % gend. words | % male bias | F1 score | % gend. words | % male bias | F1 score | F1 score |
| Gold Lbl | 0 | 0 | - | 4.11 | 100 | - | 4.03 | 0 | - | 6.67 | 50.71 | - | - |
| Baseline | 2.37 | 88.39 | 11.24 | 3.66 | 90.26 | 11.77 | 2.44 | 77.99 | **11.54** | 3.05 | 80.05 | 11.43 | 11.42 |
| ConvAI2 FT | 0.79 | 71.09 | 7.78 | 1.1 | 78.31 | 7.94 | 1.35 | 51.6 | 8.75 | 1.97 | 67.23 | 8.99 | 7.95 |
| Reddit Base | 2.18 | 73.68 | 9.93 | 3.03 | 81.78 | 11.54 | 2.81 | 52.99 | 10.99 | 3.94 | 63.16 | 12.61 | 10.57 |
| CDA | 0.88 | 71.03 | 11.63 | 1.38 | 68.57 | 11.7 | 1.2 | 56.18 | 11.43 | 1.17 | 58.01 | 11.12 | 11.62 |
| Pos. Data | 2.76 | 82.44 | 10.46 | **3.68** | 86.43 | 10.07 | **4.59** | 72.1 | 10.07 | **4.43** | 86.5 | 9.88 | 10.44 |
| Bias Ctrl | **0.14** | 68.75 | 10.72 | 5.83 | **98.08** | 13.01 | 4.8 | **2.69** | 10.84 | 4.05 | 45.86 | 11.35 | 11.38 |
| *ALL* | **0.14** | **64.19** | **11.72** | 6.59 | 97.94 | **12.77** | 5.84 | 7.13 | 11.28 | 8.81 | **50.94** | **12.22** | **11.99** |

Table 9: We compare the **performance of various bias mitigation methods**—Counterfactual Data Augmentation (CDA), Positive-Bias Data Collection (Pos. Data), Bias Control Model (Bias Ctrl), and combining these methods (*ALL*)—on the test set, splitting the test set across the four genderedness bins: $F^{0/+}M^{0/+}$. $X^0$ indicates there are no X-gendered words in the gold response, while $X^+$ indicates that there is at least one. We measure the percent of gendered words in the generated utterances (% gend. words) and the percent of male bias (% male bias), i.e. the percent of male-gendered words among all gendered words generated. While each of these methods yield some improvement, *combining all of these methods in one yields the best control over the genderedness of the utterances while improving the F1-score.*

| *Data Split:* | $F^0M^0$ | | | $F^0M^+$ | | | $F^+M^0$ | | | $F^+M^+$ | | | All |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Model* | % gend. words | % male bias | F1 score | % gend. words | % male bias | F1 score | % gend. words | % male bias | F1 score | % gend. words | % male bias | F1 score | F1 score |
| Gold Lbl | 0 | 0 | - | 4.11 | 100 | - | 4.03 | 0 | - | 6.67 | 50.71 | - | - |
| Baseline | 2.37 | 88.39 | 11.24 | 3.66 | 90.26 | 11.77 | 2.44 | 77.99 | 11.54 | 3.05 | 80.05 | 11.43 | 11.42 |
| *ALL* $F^0M^0$ | 0.14 | 64.19 | 11.72 | 0.24 | 80.11 | 11.51 | 0.22 | 25.0 | 11.63 | 0.23 | 81.58 | 10.72 | 11.61 |
| *ALL* $F^0M^+$ | 6.47 | 97.97 | 9.58 | 6.59 | 97.94 | 12.77 | 7.22 | 96.33 | 10.0 | 6.27 | 97.52 | 12.21 | 10.6 |
| *ALL* $F^+M^0$ | 4.77 | 11.66 | 10.27 | 5.12 | 15.84 | 10.94 | 5.84 | 7.13 | 11.28 | 5.03 | 13.64 | 11.23 | 10.57 |
| *ALL* $F^+M^+$ | 9.53 | 53.34 | 8.89 | 9.6 | 55.35 | 11.19 | 9.42 | 48.65 | 10.5 | 8.81 | 50.94 | 12.22 | 9.79 |

Table 10: **Performance of the *ALL* debiasing model** controlled by indicating specific bins for all examples at test time. We report results for each possible conditioning bin choice. Across bins, the model maintains performance (F1) whilst *radically changing the genderedness of the language generated.*

| **Persona Example (Original LIGHT Dataset)** | |
|---|---|
| *son:* | I am spoiled and rich. I enjoy running in the castle. I like hide and seek. |
| *men:* | I am an average man in the village. I do what ever work that my King requires me to do. At night, I spend my time in the local pub with my fellow men. |
| *farmer Bob:* | I was born in a poor village. I eat what we grow. I love being close to the earth. |
| *father:* | I am a role model for my children. I provide for the family with meat and I keep a roof over their heads. I am stability to the family, and keep things together and provide safety to my children. |
| *husband:* | I try to be good to my wife. I want to provide for my family. I try to be strong. |

Table 11: **Examples of male gender biased personas** written for gendered characters in the LIGHT dataset.

| **Generation Examples** | |
|---|---|
| Bin | $F^0M^+$ |
| Context | *[P1:]* Yes, we must get started but while we paint, it would not do any harm to keep an eye out for passages and secrets.<br>*[P2:]* Great!<br>*[P1:]* I will start on this side if you like. I will start taping the area for us.<br>*[P2:]* Ok, I have this area covered. You do this with a level of enviable professionalism. |
| Baseline: | *thank you so much for your kind words.* |
| ALL: | *thank you, sir.* |
| Gold: | *I am just a worker with lots of patience. But come here and see, I believe I may have found something.* |
| Bin | $F^0M^+$ |
| Context | *[P1:]* Just calm down you mad fellow.<br>*[P2:]* I am going to slit your throat, nobody attacks me |
| Baseline: | *i will kill you first.* |
| ALL: | *i'm just trying to protect the king.* |
| Gold: | *You too violent. Lets call a truce* |
| Bin | $F^+M^0$ |
| Context | *[P1:]* you killed my mother and you are about to marry this woman to replace her<br>*[P2:]* This is your new mother and you must respect her! |
| Baseline: | *i am sorry father* |
| ALL: | *i am sorry mother* |
| Gold: | *Hey lady, how are you doing.* |
| Bin | $F^+M^0$ |
| Context | *[P1:]* I think that fear mixes with making any person brave. There are some that are braver<br>*[P2:]* I heard that ”fear is the mind killer”. Thank you for keeping us safe.<br>*[P1:]* If you do not have fear inside you, you will do something foolish trying to be brave.<br>*[P2:]* You are my hero! |
| Baseline: | *you are a brave man!* |
| ALL: | *you are a brave woman!* |
| Gold: | *You are sweet!* |

Table 12: **Additional example generations** from the baseline model and the proposed debiased models. We show examples for conditioning on both the $F^+M^0$ and $F^+M^0$ bins.