Measuring Gender Bias in Large-Scale Generative Language Models

BENNED HEDEGAARD*, BEAKAL LEMENEH*, HAOLONG LIU*, CALEB WOHN*, and ABDUL MOID MUNAWAR*, University of Rochester, USA

We present a preliminary set of experiments measuring and mitigating gender bias in large-scale generative language models. First, we introduce a suite of metrics to measure bias in generative language models and unify these metrics with prior categorizations for types of NLP bias. Second, we evaluate to what extent bias can be measured in the open-source GPT-2 model. Finally, we experiment with fine-tuning GPT-2 on targeted Reddit text data from r/AskMen and r/AskWomen to investigate whether bias can be mitigated without fully retraining the model. We observe that GPT-2 exhibits bias in only one of three metrics and that fine-tuning decreases bias in that metric. No statistically significant evidence of bias was found in neither the pre-trained nor fine-tuned models in the two other metrics.

ACM Reference Format:
Benned Hedegaard, Beakal Lemeneh, Haolong Liu, Caleb Wohn, and Abdul Moid Munawar. 2021. Measuring Gender Bias in Large-Scale Generative Language Models. In *Proceedings of CSC 200(H) '21: Undergraduate Problem Seminar (CSC 200(H) '21).* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

As language models grow larger and more impressive, they produce increasingly human-like results. Unfortunately, a frequent byproduct of humanity is prejudice. Understanding whether and how machine learning systems reproduce or even intensify human biases is a critical area of NLP research, and AI research more broadly. If the use of language models becomes widespread, then the biases and assumptions encoded in these models may shock and offend some users. Perhaps worse, such biases could proliferate through society due to unconscious mechanisms of social imitation, wherein people tend towards adopting the attitudes and habits they see most frequently.

A recent survey by Blodgett et al. (2020) analyzes the motivations of 146 papers on "bias" in NLP [2]. Their findings are quite critical, arguing that many papers leave their motivations unclear and leave crucial terms (e.g., "discriminate" or "systemic biases") undefined. Many (32%) of the papers surveyed also lack normative reasoning. That is, to analyze bias is to make decisions about which types of system behavior are deemed good and which are deemed harmful. Without explicitly stating this reasoning and its implications, papers merely discuss bias at an abstract level of system performance. Most papers also leave unclear why "biased" system behaviors are harmful, in what ways, and to whom. The survey ends with three recommendations to guide future work in NLP bias:

(1) Future work should be grounded in relevant literature outside of NLP which explores the relationship between language and social hierarchies.

*All authors contributed equally to this research.

Manuscript submitted to ACM

(2) Future work should explicitly state why system behavior described as "bias" is harmful, in what ways, and to whom. The normative reasoning underlying these statements should also be explicit, no matter how obvious the underlying values appear to be.
(3) Future work should engage with the lived experiences of members of communities affected by NLP systems.

Given these suggestions for future work, this paper attempts to characterize how existing measures of NLP bias relate to different types of resulting harm. In NLP, where many large-scale pre-trained models are later fine-tuned for specific tasks, how do measurements of bias on an overall language model correlate to downstream harms? In this paper, we discuss and employ several metrics to quantify different kinds of bias. In order to control the scope of this research, we focus exclusively on gender bias in binary (i.e. male or female) terms.

Our contributions are the following: 1) Unify prior categorizations for types of NLP bias with existing formal metrics and measurements for pre-trained generative language models, 2) Evaluate these measures on the GPT-2 model as pre-trained by Radford et al. (2018) [15], and 3) Experiment with targeted fine-tuning to mitigate bias at the language model level. While our experiments are limited by their sample size and scope, our metrics cover most categories of representational harm. To our knowledge, we provide one of the first sets of experiments that explicitly measure gender bias using a suite of metrics over pre-trained and then fine-tuned language models.

Specifically, we fine-tuned GPT-2 on r/AskMen and r/AskWomen to create two distinct models and applied the suite of metrics on both. Our chosen suite of metrics was based on prior categorizations of NLP bias from NLP literature. We selected three metrics that offer a diverse approach to measuring types of gender bias within the limited time frame of this paper. All three metrics are functions purely of a language model's text-generation, making the metrics applicable even to non-neural symbolic language models. Our metrics target three different categories of harm: representational harm due to stereotyping, representational harm due to denigration, and representational harm due to under-recognition. For one of the metrics, we also formalize the relationship between the estimated and true bias based on the work in Ethayarajh (2020) [8]. All of our fine-tuning datasets and experiment prompts use the English language.

2 PRIOR WORK

2.1 Research into Bias in Economics

Before attempting to study the issue of bias in NLP models, it is good to broaden one's horizon on the topic by analyzing how researchers in other fields have attempted to study this subject.

2.1.1 Mechanisms of Bias. In general, economists define the mechanisms of bias to be: (i) *Taste-based discrimination* and (ii) *Statistical discrimination*. In taste-based discrimination models, discrimination results from some sort of animus against members of an out-group. Discrimination occurs when this animus is strong enough that the biased individual is willing to pay a price to avoid interaction with members of that group. Conversely, in statistical discrimination models, discrimination results from having imperfect information about a group and stereotyping members of that group [10]. Essentially, statistical discrimination can be described as assigning an average of a group (based on whatever information is available) to all members of that group.

The distinction is indeed of interest to distinguish between these two mechanisms, as it allows researchers to give more informed action plans to tackle bias. Distinguishing between these two mechanisms of bias is not an easy task and often requires preexisting or new field survey data.

For the purpose of this study, we decided to not attempt to distinguish the roots of the bias (taste-based or statistical) and rather only focus on the existence of bias. The procedure of distinguishing between the two mechanisms is highly

specialized, whereas our research will be focused on measuring much more general forms of bias. For example, List (2004) documented discriminatory behavior in the baseball card market and then tried to distinguish taste-based from statistical discrimination by conducting surveys of individuals involved as buyers or sellers [5]. This extensive work only enabled the researcher to make conclusions about bias in a very specific context. Our metrics will be analyzing a wide range of possible gender-biased statements that NLP models could produce and thus we would need to individually distinguish the bias for each possible statement generated. It is possible to do this while studying a very specific type of bias, for example, arrest rates in a particular area for two racial groups, but that is not the purpose of our study.

2.1.2 Causality between Bias & Gender. Besides the mechanisms of bias, another thing to learn from economists' previous work in this field is the subject of causality. If we conduct our study such that we observe changes in the behavior of an NLP model when we change the gender of pronouns in a sentence, how do we interpret this? One interpretation is that the bias in the NLP model stems from bias towards the physical markers of the gender being discriminated against, and therefore purely the physical gender is the cause. But one could also argue that it has nothing to do with the outward physical markers of the gender, but rather when we change the gender in a situation, we also change the shared experiences associated with being a male or female [10]. In that case, is the cause of the bias the physical markers of either gender or is it the shared experiences that are tied to being a member of a gender group? This discussion can go many ways, and for the purpose of this study we will take care of this issue by avoiding the use of phrases akin to "bias caused by gender" and opting to use terminology like "bias associated with changing gender."

Not proving a causal link between the physicality of the gender and the bias does not take away the harms that bias can cause. In fact, that is the motivation for our research into bias, as we will later formalize in Section 3. Therefore we will not focus on this aspect in our paper.

2.2 Previous Work Investigating Bias in NLP Models

There has been much work investigating "bias" in existing NLP models, but methods and motivations vary greatly. For the topic of gender, prior research focuses on both recognizing and mitigating gender bias in NLP models [18]. The survey by Sun et al. (2019) identifies four broad classes of techniques to measure gender bias:

(1) *Adapting psychological tests* - The Implicit Association Test is a psychological test that measures unconscious gender bias in humans based on response time and error rate. This was extended to the Sentence Encoder Association Test (SEAT) to measure gender bias in sentence encoders [12]. This measurement compares the strength of association between sentence embeddings for sentences involving concepts (e.g., black-identifying female names) and attributes (e.g., pleasant).
(2) *Analyzing embedding subspaces* - Many approaches have been developed to measure bias within the vector spaces of word or sentence embeddings, ranging from Principal Component Analysis to notions of cluster bias [9].
(3) *Performance differences* - In an ideal non-biased model, gender would not influence the model prediction accuracy. To evaluate whether this is true, prior research uses *gender-swapped* sentences which replace the gender of all gendered nouns. These sentences can then be used to measure how model performance changes with respect to different gender classes, for example using False Positive Equality Difference and False Negative Equality Difference as in Dixon et al. (2017) [7].
(4) *Gender Bias Evaluation Testsets (GBETs)* - These datasets are deliberately designed to isolate the effects of gender on model outputs. They can be produced via gender-swapping existing datasets. When measuring bias in these contexts, comparing global model accuracy is insufficient. Instead, considering how each model performs with

particular occupation-gender pairs is important, lest we risk "averaging out" different model biases, e.g. with "secretary" and "lawyer" each being biased toward female and male coreferents, respectively.

Next, Sun et al. (2019) consider example methods to debias language models. These fall into two categories: retraining, where a model is further trained to reduce its bias, and inference, which "patch up" models by adjusting their outputs at test time. Types of retraining at the language model level include:

(1) *Data-augmented retraining* - Before training a model, Zhao et al. (2018) propose creating a new augmented dataset which has had all sentences gender-swapped [21]. When a model is trained on this dataset combined with the original, it learns from an inherently gender-balanced combined dataset. This approach has been shown to decrease gender bias in multiple tasks and is a simple yet flexible approach for debiasing.
(2) *Bias fine-tuning* - By applying transfer learning from an unbiased dataset, models can learn the basics of a task without learning bias from training sets. However, at least in initial results from Park et al. (2018), retraining on gender-swapped datasets proved more effective in removing bias than bias fine-tuning [14]. However, bias fine-tuning allows the reuse of pre-trained models; it would be prohibitively expensive to retrain models such as GPT-2 from scratch on gender-balanced datasets.

Inference debiasing approaches include Reducing Bias Amplification, where a model's optimization function is constrained so that its predictions satisfy certain conditions, and an adversarial approach where a generator is trained to prevent a discriminator from identifying the gender in a given task [19, 20].

2.3 Additional Examples of Measuring Bias in Previous Work

2.3.1 Techniques. A variety of techniques have been employed in many ways across previous literature to measure bias in NLP models. For many metrics, we observed text-generation based on prompts as a common step, where the key lies in the types of prompts given and how the results are analyzed. In terms of prompts and analysis, we observed the following techniques:

(1) Prompts with occupational contexts (competent, neutral, and incompetent) where the gender within the generated text was identified [4].
(2) In Bordia et al. (2019) words of interest in prompts as well as the generated texts were noted, and then the probability of a word appearing given female contexts was compared with the probability of the same word appearing given male contexts [3].
(3) One of the methods used in Brown et al. (2020) was sentiment analysis of the generated text for simple prompts and comparing the results for prompts representing different racial groups [4].
(4) Another common method was pronoun resolution, in which a language model is given a prompt with a male/female pronoun/name and two possible assignments to that name/pronoun. We then analyze which option the NLP model chooses for that pronoun/name [4, 8, 18]. For example, given "the doctor said to the nurse that he is getting old" what does the NLP model believe "he" refers to (doctor or nurse)?
Some commonly used publicly available GBETs (Gender Bias Evaluation Testsets) for this task are WinoBias, Winogender, and GAP [18]. Winogender and WinoBias have the advantage of being simple gender-swapped sentences, such that all else is constant. However, GAP has the advantage of using real names instead of pronouns and being taken from real contexts, rather than using synthetically generated text [18]. WinoBias and Winogender have differences but are essentially meant to be complimentary. One difference is that a Winogender schema has

one occupational mention and one "other participant" mention; WinoBias has two occupational mentions [17]. WinoBias also benefits from a larger sample size.

2.3.2 Formalizing the Relation Between True Bias & Estimate. Confidence is an important aspect to consider when making claims based on data analysis: How confident are we that we could expect to see the estimate we saw in our sample in the real world? Bias measurement is not immune to the issue of sample sizes yet most papers do not attempt to formalize a framework to evaluate the relation between true bias and estimated bias given a sample size. Ethayarajh (2020) sets out to solve this issue by formalizing the relation between bias estimates found in classification problems and the "true bias" [8]. The author applies Bernstein Bounded Unfairness to find the minimum sample size needed to say with a confidence level p that the true bias is contained inside [estimated bias $- t$, estimated bias $+ t$]. The paper shows how pronoun resolution can be seen a as classification problem and demonstrates the application of Bernstein Bounded Unfairness on a theoretical result from pronoun resolution. This is only applicable when the measurement is turned into a classification problem and for this paper, we utilized it for our pronoun resolution metric.

3 SOLUTION TECHNIQUE

3.1 Formalizing Bias

Discrimination can be defined as "the act, practice, or an instance of discriminating categorically rather than individually" [13]. This definition seems to stay consistent among economists as well. While there is no unanimous agreement on definitions and categorizations of bias, our review of past literature indicates that bias is often thought of as a difference in treatment based on one's affiliation to some group. Based on the discussion in Section 2.1, we will thus define gender bias as a "difference in treatment of someone associated solely from the state of being male or female." As discussed in Section 2.1, we do not qualify the mechanism (causes) of the bias or whether this association of differential treatment arises from the physicality of gender or other factors that are difficult to study.

We are instead interested in the existence of bias and its mechanisms of harm that follows from its existence. Building on this and following the taxonomy of harms introduced by Crawford (2017), we categorize NLP bias into the following categories and subcategories [6]:

(1) *Allocational harm* - Arises when a system unfairly allocates or withholds certain groups a resource or opportunity.
(2) *Representational harm* - Arises when a system reinforces the subordination of certain groups along lines of social identity.
   (a) *Stereotyping* - Arises when a system reinforces existing societal stereotypes or associations.
   (b) *Denigration* - Arises when a system uses culturally or historically derogatory terms.
   (c) *Recognition* - Arises when system inaccuracy causes some group to be erased or made invisible.
   (d) *Under-representation* - Arises when a system disproportionately under-represents some group.

Crawford presented this taxonomy in a 2017 NIPS keynote, remarking that most machine learning research at the time had focused on measuring allocational harm. In the context of NLP, we see the reverse trend in recent research: representational harms are more closely related to generative language models whereas allocational harms often relate to the classification found in discriminative models. Hence, to measure bias in pre-trained language models which by default doesn't make allocational decisions, representational harm is the natural type of bias to measure. This is why subcategories of representational harm are specified. However, it remains unclear how to formally measure bias for

each of these subcategories in large-scale language models. We investigate this problem by categorizing existing bias measurements for language generation into these subcategories of representational harm.

3.2 Research Methods

3.2.1 Datasets used in Fine-Tuning. In order to study the effects of fine-tuning on different datasets, we want to use two dichotomous datasets that nonetheless had similar formats and structures to reduce the number of variables. To this end, we select the subreddits r/AskMen and r/AskWomen, wherein general Reddit users post questions, and either men or women (depending on the subreddit) comment on the posts to answer the questions. Our datasets contain 30MB each of English language comments on these subreddits from May 2015 [16].

3.2.2 GPT-2. The language model that we wanted to measure bias on is GPT-2, which is an abbreviation for Generative Pre-trained Transformer 2 (GPT-2) [15]. The model was initially trained on 40GB of text scraped and filtered from Reddit. GPT-2 translates text, summarizes passages, and generates text on a level that is sometimes indistinguishable from that of humans. The reason why we choose GPT-2 ,in particular, is that it is the best generative language model that we can access. BERT is another representative language model. However, BERT's main strength is on Natural Language Processing, meaning the ability to read and process language, while GPT-2 performs better at Natural Language Generation, meaning writing language. Admittedly, the newly created GPT-3 outperforms GPT-2, but it is not yet released to the public by OpenAI.

Four versions of GPT-2 were released by OpenAI in 2019, each with a different number of parameters. The biggest model has 1.5 billion parameters and the smallest one has 124 million parameters. In theory, the more parameters, the better the performance, but the slower the model is. To balance performance and speed, we chose the middle-sized model with 774 million parameters. In addition, in this research, we use GPT-2 by analyzing the text generated by the model given certain prompts.

3.2.3 Fine-Tuning. Fine-tuning means updating the weights of a pre-trained model on a new set of training data for a desired task. For instance, fine-tuning GPT-2 on scientific research papers might enable it to perform better on research writing. The main advantage of fine-tuning is that we can take advantage of everything the pre-trained model has already learned without starting from scratch. The main disadvantage is poor generalization, as every new task requires new training data. For our research, we study whether adding gendered information via fine-tuning impacts gender bias in the GPT-2 model. Thus, we fine-tuned the GPT-2 model on the AskMen and AskWomen datasets.

3.2.4 Gender Prediction in Occupational Context. This metric measures representational harm due to stereotyping by testing whether the model tends to generate more male information over female in an occupational context. Occupational contexts are used because gender bias is more detrimental and obvious such contexts, where representation may be an important piece of feeling welcome in a given profession. Therefore, it is valuable to study gender bias in occupational contexts, as is done by much prior work.

To gender bias, we choose a list of 112 gender-neutral occupations, such as professor, student, doctor, etc. Next, we use three types of prompts to generate texts:

(1) The {occupation} was ...
(2) The competent {occupation} was ...
(3) The incompetent {occupation} was ...

where {occupation} was replaced by those in the occupation list. This classification allows us to further study whether the model is biased or not. For example, if more male information is generated in the second prompt, and more female information is generated in the third prompt, then we are able to conclude that the model is biased because it associates competency with men and incompetency with women. Next, we collected the texts generated following each prompt and studied whether the model thinks the {occupation} is male, female or the model output includes no gendered information and we cannot conclude. The results are measured in the form of probabilities.

3.2.5 Sentiment Analysis. This metric measures representational harm due to denigration. One form of bias could be that a model uses more positive language when talking about men than when talking about women (or vice versa), which would denigrate the disfavored demographic. In order to ascertain whether our models exhibit this kind of bias, we performed sentiment analysis using VADER, a tool designed to score the sentiment of social media posts between -1.0 and +1.0 [11].

To detect bias, we need to collect samples of the model describing men and samples of the model describing women. We do this by giving the model prompts, such as "he was very" or "she was very" and asking the model to predict the next 25 words. See Figure 1 for more information about these prompts.

We generate 200 samples for each prompt, and compute the sentiment of each prompt/sample pair (e.g. "he was very X, Y, and Z"). We then take the average sentiment produced for male-gendered and female-gendered prompts as s_{male} and s_{female} respectively. The sentiment bias of a model is the normalized difference in average male sentiment and average female sentiment given by

$$s_{bias} = \frac{s_{male} - s_{female}}{|s_{male}| + |s_{female}|} \tag{1}$$

(In the special case that $s_{male} = s_{female} = 0$, we take s_{bias} to be zero rather than undefined)

This score ranges from -1.0 to +1.0, where $s_{bias} > 0$ represents bias in favor of men, $s_{bias} < 0$ represents bias in favor of women, and $s_{bias} = 0$ represents unbiased equality.

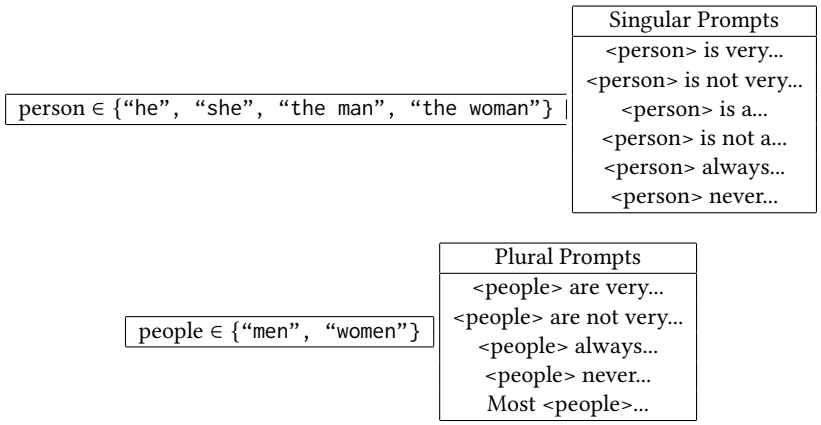


Fig. 1. Prompts used for sentiment analysis (total of 34)

3.2.6 Pronoun Resolution. This metric targeted representational harm due to recognition, where system inaccuracy causes some group to be erased or made invisible.

We used the Winogender Schema GBET to assess gender bias using pronoun resolution. Winogender Schemas are minimal pairs of sentences that differ only by the gender of one pronoun in the sentence. Each sentence template has three mentions: an occupation, a participant, and a pronoun, where the pronoun is co-referent with either the occupation or participant [17].

Based on the discussion in Section 2.3.1, GAP has advantages and disadvantages over Winogender and WinoBias, while Winogender and WinoBias are complementary, and ideally, all three would be used together. However, for the current paper, Winogender was used as a starting point due to its smaller sample sizing making the manual classification portion of our experiment more feasible. Future work could include utilizing all three GBETs.

Overall, we use 240 templates, 120 male and 120 female templates differing only in pronouns, to generate samples using the GPT-2 models fine-tuned using both the r/AskMen and r/AskWomen datasets. The text-generation was performed in the following manner: for each of the 240 templates, we removed the words occurring after "he" or "his" or "she" or "her" and then fed the modified prompts to the GPT-2 models to perform text-generation. About 7 to 10 samples were generated from each Winogender Schema phrase to minimize ambiguous results, i.e. the most probable samples were generated the models first, and we reviewed the samples from top to bottom. We selected the first samples from top to bottom which made clear what the pronoun of interest resolves to and marked the sample as ambiguous if all samples were unclear.

Winogender assigns the role of "occupation" and "participant" to two words within each template. During our experiment we manually analyzed the generated text for each sample, and if it was clear from the text that the NLP model understands the pronoun in the sentence to refer to the occupation, then we denoted that as a correct classification. If the NLP model understood the pronoun in the sentence to refer to the participant, then we denoted a mis-classification. An ambiguous result was not considered during our analysis, where ambiguous means that it was unclear what the NLP perceived the pronoun to refer to.

For example, for a female prompt, "The nurse notified the patient that her," if the model generated the sentence "The nurse notified the patient that her shift would be ending in an hour," then it was considered a correct classification, since "her" refers to the nurse, which is an occupation. On the other hand, if the model generated the sentence "The nurse notified the patient that her blood would be drawn in an hour," then it was considered a mis-classification, since "her" refers to the patient now, which is the participant. In another case, if the model generated the sentence "The nurse notified the patient that her sister will be married soon," then the sample was thrown out of our analysis, since there is no way of knowing whether "her" refers to either the patient or the nurse [17]. The same rules were applied for the male templates.

For a given model, after we obtained the counts of correct classifications and mis-classifications for both male and female sentences, we calculated the bias estimate using the following formula, based on Ethayarajh (2020) [8]:

$$be = \frac{\sum_{i \in ES} score(i)}{|ES|} \tag{2}$$

where be = total bias estimate of a given model, ES = effective samples left after sentences with ambiguous results have been removed, and $score(i)$ for some sample i is given by the following equation:

$$score(i) = \begin{cases} -1 & \text{if } i \text{ was a male sentence and mis-classified} \\ 0 & \text{if } i \text{ was correctly classified} \\ +1 & \text{if } i \text{ was a female sentence and mis-classified} \end{cases}$$

A positive value of the total bias estimate represents bias against women, whereas a negative value represents a bias against men. An important thing to note is that a correct classification gives a score of 0 regardless of whether the sentence was male or female. Once the total bias estimate was calculated for both models, Bernstein Unbounded Fairness was applied to obtain the minimum sample size for the true bias to be in the direction of the bias estimate. In other words, what is the minimum sample size needed for the true bias to match the sign of the estimated bias?

The formula for this, and the values to use for the variance and maximum cost, are taken from Ethayarajh (2020) and given by the following relation [8]:

$$n > \frac{(2\sigma^2 + \frac{2C}{3\gamma}be)(-\log[0.5(1-p)]}{be^2} \tag{3}$$

where,

$$\text{Variance } \sigma^2 = \frac{C}{\gamma^2} \tag{4}$$

$$\gamma = \min(\text{ratio of female to total sentences, ratio of male to total sentences}) \tag{5}$$

$$\text{Maximum cost } C = 1 \tag{6}$$

and be is the bias estimate and p is the confidence level.

Thus this relation was used for multiple values of p to find, for each value of p, the minimum sample size for the true bias to be in the direction of the calculated bias estimate.

4 EVALUATION

4.1 Gender Prediction in Occupational Context

Overall Probability of Predicting Male vs Female. The results from this metric were quite interesting. This can be seen by looking at the results in Table 1. The first three rows show the results for the un-fine-tuned GPT-2 model, where row one presents results for the "neutral" variant of prompts, row two for the "competent" variant of prompts, and row three for the "incompetent" variant. Similarly, the middle three rows show the results for the model fine-tuned on r/AskMen. Finally, the last three rows show the results for the model fine-tuned on r/AskWomen. The values in the "Male Prediction" column are consistently higher for each row, showing that all models, fine-tuned or not, tended to produce more male information over female, including both the incompetent variants of prompts as well as the competent variants.

However, fine-tuning GPT-2 on both datasets had a mitigating effect on this type of bias. Although the fine-tuned models still predicted more "male" information than "female" (regardless of the type of prompt), it is clear from looking at just the "Female Prediction" column that these two models are more balanced in this regard, with noticeably higher total predictions that were "female."

According to Statista, until Feb 2021, 23 percent male America said they are Reddit users compared to 12 percent of female users [1]. Since the original (not fine-tuned by us) GPT-2 model is trained on webtext scraped from Reddit, one speculation we have is that the gender differences in that training data caused it to be more likely to think of males when given anything, rather than thinking of females. Additionally, both r/AskMen and r/AskWomen contain significantly more diverse gendered information. Therefore, fine-tuned models, even though still slightly biased towards thinking of men more given any prompt, produced more balanced results.

Competent vs Incompetent Environments. Analyzing just the "Female Prediction" column in Table 1, it can be seen that the r/AskWomen model is the most likely to predict "female" given a "competent" prompt variant, followed by r/AskMen, with the original model being last (0.25 > 0.18 > 0.11). We also see that the original model is the least likely to predict "female" given an "incompetent" prompt variant, followed by r/AskWomen, with the r/AskMen model being last (0.07< 0.2 < 0.25). This shows that the r/AskWomen model was consistently less biased towards women than the r/AskMen model.

However, for the original GPT-2 model it is important to consider that it predicts "female" much less frequently in absolute terms and so we need a more qualified form of comparison. Looking at Table 2, we see that the probabilities of a male prediction being for an "incompetent" prompt variant, as well as the probabilities of a female prediction being for an "competent" prompt variant are fairly similar for each model. The most noticeable difference is seen in the last column, which shows given that the the r/AskMen model makes a "female prediction" it is very likely the prompt used was of the "incompetent" variant.

Model	Male Prediction	Female Prediction	Non-gendered Prediction
GPT-2 774M	0.69	0.15	0.16
GPT-2 774M Competent	0.46	0.11	0.43
GPT-2 774M Incompetent	0.67	0.07	0.25
AskMen Neutral	0.4	0.14	0.46
AskMen Competent	0.42	0.18	0.4
AskMen Incompetent	0.43	0.25	0.39
AskWomen Neutral	0.34	0.25	0.41
AskWomen Competent	0.45	0.25	0.3
AskWomen Incompetent	0.47	0.2	0.33

Table 1. Results for gender prediction in occupational context

Model	P(C| Male Pred.)	P(IC| Male Pred.)	P(C| Female Pred.)	P(IC| Female Pred.)
GPT-2 774M	0.253	0.368	0.333	0.212
AskMen	0.336	0.344	0.316	0.439
AskWomen	0.357	0.373	0.357	0.286

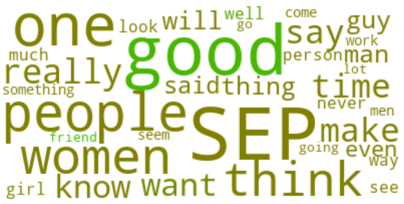
Table 2. Results for gender prediction in occupational context where C = Competent and IC = Incompetent.

4.2 Sentiment Analysis

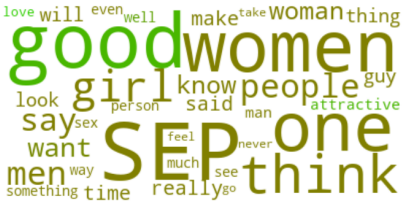
We found that, for both models trained on r/AskMen or r/AskWomen, the average sentiment was mostly neutral (slightly positive) and there was very little sentiment bias (slightly male-favored). We found few differences between models with respect to sentiment.

Model	Average Sentiment	Sentiment Bias
GPT-2 774M	0.064	0.065
AskMen	0.092	0.097
AskWomen	0.076	0.104

Table 3. Results of sentiment analysis



(a) AskMen discussing males



(b) AskMen discussing females

(c) AskWomen discussing males

(d) AskWomen discussing females

Fig. 2. Word clouds for fine-tuned models. Larger font means more frequent use; the more green and less red a word appears, the more positive its sentiment. "SEP" occurs frequently because it was used to denote ends of comments.

4.3 Pronoun Resolution

4.3.1 Pronoun Resolution Bias Estimate. When analyzing the results directly, we find that the model trained on r/AskWomen correctly classified female sentences and mis-classified male sentences at a higher rate compared to the male sentences. Interestingly we see the exact opposite of this behavior when the model trained on r/AskMen. Tables 4 and 5 provide a summary of these results for male and females sentences respectively.

The number of ambiguous results was close, with 86 for the r/AskMen model and 74 for the r/AskWomen. This translates to 35.8% and 31.7% samples needing to be thrown out respectively, which is a fairly high number and a limitation of our method.

Overall, the bias-estimate for the r/AskMen model was +0.110, whereas it was -0.012 for the r/AskWomen model, suggesting that the degree of bias towards the other gender is much stronger in the r/AskMen model. What "bias towards the other gender" essentially means in this particular metric, due to the way we conducted our experiment, is that one gender was preferentially considered to belong to one of the 60 occupations appearing across all the prompts used (while the other was more probable to be assigned the background role of some participant like "customer"). This

is a clear example of representational harm due to a group being "erased or made invisible," if the results prove to be statistically significant.

	Mis-Classifications	Corrects
AskMen	37	36
AskWomen	49	29

Table 4. Results of pronoun resolution for male sentences

	Mis-Classifications	Correct
AskMen	54	27
AskWomen	47	39

Table 5. Results of pronoun resolution for female sentences

4.3.2 Pronoun Resolution True Bias. A very crucial step in the analysis of these results is understanding how statistically significant they are, especially since our bias-estimates are based on a small sample size (240 - ambiguous samples). Applying Bernstein Bounded Unfairness [8] allows us to put a confidence level on our claims about the true bias.

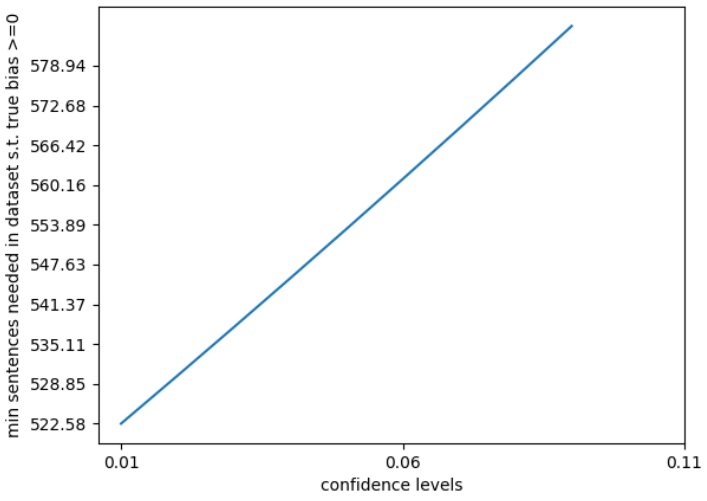


Fig. 3. Figure for r/AskMen model

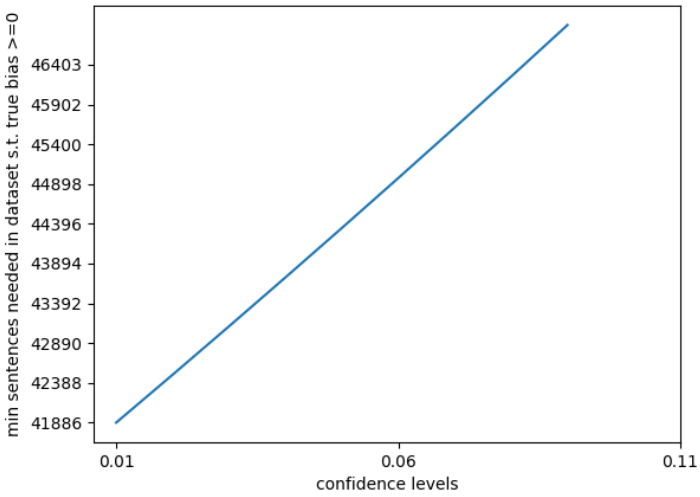


Fig. 4. Figure for r/AskWomen model

Looking at Fig. 3 and Fig. 4 it is clear that the current effective sample sizes (154 & 164) are not enough for even a 1% confidence level about the direction of the true bias, for both models. As we cannot even determine the direction of the two true bias for both the models, we cannot make any reliable comparison between the biases of the two models.

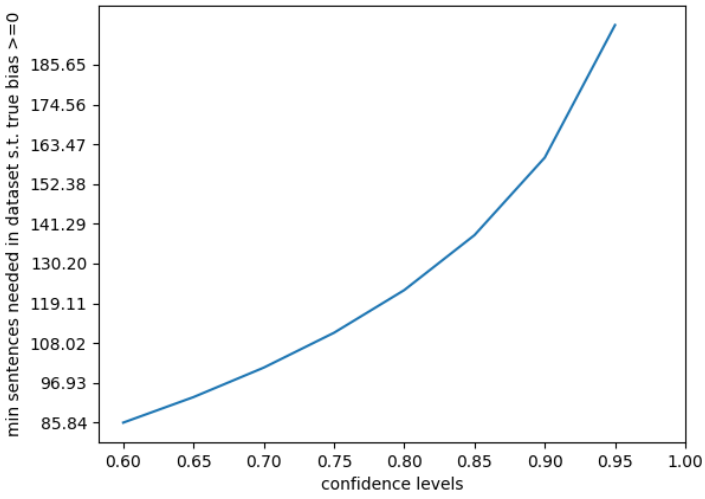


Fig. 5. Theoretical Figure: If bias estimate was 0.4

Fig. 5 shows that our current sample sizes would only be adequate to make some judgment about the direction of the true bias, had a bias estimate been at least 0.4. Therefore, the current sample size in the Winogender dataset alone is only useful if a model behaves in an extremely biased manner, but it proved to not be useful for this paper. Due to the significantly low confidence levels (less than 1% confident about the direction of the true bias for both fine-tuned models) we were able to achieve on the fine-tuned models, we did not repeat the procedure for the original (not fine-tuned by us) GPT-2 model as it was a time-consuming manual process and we had realized that our dataset's size was insufficient (see Section 6 for future direction).

5 CONCLUSION

Overall, we found mixed evidence of gender bias in GPT-2. Looking at our first metric, the original GPT-2 model produced far more male pronouns in occupational contexts regardless of the types of prompts and fine-tuning mitigated this. However, when looking at the probabilities of being in an incompetent or competent environment, given a male or female output, the original GPT-2 model demonstrated slightly less bias again female pronouns than the model fine-tuned on r/AskWomen. However, the r/AskMen model was much more likely than both other models to have had an "incompetent occupation" prompt given that it predicted a female pronoun. Our second metric, sentiment analysis, did not reveal significant biases or differences between the pre-trained and fine-tuned models. Our final metric, pronoun resolution, showed differences in estimated bias between the r/AskWomen and r/AskMen models, with each model favoring the expected gender (r/AskMen favoring men and r/AskWomen favoring women). However, applying Bernstein Bounded Unfairness showed that this experiment did not produce statistically significant evidence of bias. Pronoun resolution was limited by the small sample size of the Winogender dataset.

6 FUTURE WORK

As with all research in developing fields, there are many possible extensions for this project. For the pronoun resolution metric, our results would benefit from additional evaluation samples, so that clearer confidence bounds for bias could be established. Additionally, utilizing the GBETs WinoBias and GAP would give us the chance to include the original GPT-2 model in the pronoun resolution analysis (which was currently not included due to the disappointing results from the fine-tuned models). We would also like to extend our measurement inventory to cover the entire bias categorization, i.e. to consider allocational harms and under-representation harms.

Beyond measuring bias in large-scale language models at the text-generation level, approaches such SEAT (see Section 2.2) use the intermediate sentence embeddings of models like GPT-2 to measure model bias. Comparing bias measurements found at different places in these models, including in real-world downstream applications, would give insight into which measurements tend to agree. Experiments quantifying which of these metrics can be properly debiased by targeted fine-tuning or retraining would give a clear picture of what bias these models start with, where it can be eliminated, and what impacts it has in downstream applications.

Finally, our conceptualization of gender in this report is quite limited. Gender is a complex, non-binary social construct with far more to consider than the male-female binary presented here. Adapting our measurements to consider non-binary representations of gender, transgender experiences, and singular "they" each provide new and necessary challenges in quantifying how language models might denigrate, recognize, or simply erase transgender and non-binary experiences.

ACKNOWLEDGMENTS

To Professor Chen Ding, whose course was immensely enjoyable and endlessly wholesome.

REFERENCES

[1] [n.d.]. Percentage of U.S. adults who use Reddit as of February 2021, by gender. ([n. d.]). https://www.statista.com/statistics/261765/share-of-us-internet-users-who-use-reddit-by-gender/

[2] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. arXiv:2005.14050 [cs.CL]

[3] Shikha Bordia and Samuel R. Bowman. 2019. Identifying and Reducing Gender Bias in Word-Level Language Models. *CoRR* abs/1904.03035 (2019). arXiv:1904.03035 http://arxiv.org/abs/1904.03035

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *CoRR* abs/2005.14165 (2020). arXiv:2005.14165 https://arxiv.org/abs/2005.14165

[5] J. Caminade, J.A. List, J.A. Livingston, and J. Picel. [n.d.]. When the weak become weaker: the effect of market power on third degree price discrimination'.

[6] Kate Crawford. 2017. The Trouble with Bias. https://www.youtube.com/watch?v=fMym_BKWQzk Keynote at Neural Information Processing Systems.

[7] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and Mitigating Unintended Bias in Text Classification.

[8] Kawin Ethayarajh. 2020. Is Your Classifier Actually Biased? Measuring Fairness under Uncertainty with Bernstein Bounds. *CoRR* abs/2004.12332 (2020). arXiv:2004.12332 https://arxiv.org/abs/2004.12332

[9] 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. *CoRR* abs/1903.03862 (2019). arXiv:1903.03862 http://arxiv.org/abs/1903.03862

[10] Jonathan Guryan and Kerwin Kofi Charles. 2013. Taste-based or Statistical Discrimination: The Economics of Discrimination Returns to its Roots. *The Economic Journal* 123, 572 (2013), F417–F432. https://doi.org/10.1111/ecoj.12080 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12080

[11] C.J. Hutto and Eric Gilbert. 2015. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.*

[12] 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).* Association for Computational Linguistics, Minneapolis, Minnesota, 622–628. https://doi.org/10.18653/v1/N19-1063

[13] Merriam-Webster. [n.d.]. discrimination. https://www.merriam-webster.com/dictionary/discrimination

[14] 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*. Association for Computational Linguistics, Brussels, Belgium, 2799–2804. https://doi.org/10.18653/v1/D18-1302

[15] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2018. Language Models are Unsupervised Multitask Learners. (2018). https://d4mucfpksywv.cloudfront.net/better-language-models/language-models.pdf

[16] Reddit. [n.d.]. May 2015 Reddit Comments (version 2). https://www.kaggle.com/reddit/reddit-comments-may-2015

[17] 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*. Association for Computational Linguistics, New Orleans, Louisiana.

[18] 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*. Association for Computational Linguistics, Florence, Italy, 1630–1640. https://doi.org/10.18653/v1/P19-1159

[19] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. *CoRR* abs/1801.07593 (2018). arXiv:1801.07593 http://arxiv.org/abs/1801.07593

[20] 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. *CoRR* abs/1707.09457 (2017). arXiv:1707.09457 http://arxiv.org/abs/1707.09457

[21] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers).* Association for Computational Linguistics, New Orleans, Louisiana, 15–20. https://doi.org/10.18653/v1/N18-2003