Selection Bias Induced Spurious Correlations in Large Language Models

Emily McMilin [1]

Abstract

In this work we show how large language models (LLMs) can learn statistical dependencies between otherwise unconditionally independent variables due to dataset selection bias. To demonstrate the effect, we developed a masked gender task that can be applied to BERT-family models to reveal spurious correlations between predicted gender pronouns and a variety of seemingly gender-neutral variables like date and location, on pre-trained (unmodified) BERT and RoBERTa large models. Finally, we provide an online demo, inviting readers to experiment further.

1. Introduction

For generalization to real-world target domains, a learned model's training data would ideally be a randomized sampled subset of the data in the target domain. However, achieving such randomization is often impractical, resulting in many datasets exhibiting sampling bias (Heckman, 1979). Models trained on datasets with sampling bias are vulnerable to learning spurious associations from this bias.

These spurious associations, often referred to as spurious correlations though the associations need not be linear, can reduce the predictive performance of the model in real-world domains. Specifically, although we desire the model to learn the conditional distribution: $P(Y|X)$, it has instead learned $P(Y|X, S{=}1)$, where $S{=}1$ represents selection into the dataset (Bareinboim and Pearl, 2012).

This paper focuses on a specific type of selection bias in which two variables: W and G, which are unconditionally independent in the real world ($G \perp\!\!\!\perp W$), become conditionally dependent within the dataset ($G \not\perp\!\!\!\perp W|S$), due to the selection process, S.

We hypothesize that a wide range of spurious correlations can be traced back to this type of bias. To expose subtle spurious associations that have not yet been reported, we desire an underspecified learning task, in which there are multiple equally plausible predictions. In natural language processing, one well-researched underspecified task is that of gender pronoun prediction (D'Amour et al., 2020), in which a gender is predicted from gender-neutral features.

Undesirable and spurious associations between gender: G, and variables: W', such as occupation (Webster et al., 2020) and college major (Rudinger et al., 2018) have been found in LLMs. Unfortunately, many of these spurious associations are in fact representative of undesirable gender inequities in our real world ($G \not\perp\!\!\!\perp W'$).

To show how selection bias can induce associations that do not match our real-world distributions of gender, we seek to demonstrate spurious associations between gender and real-world gender-neutral variables: W.

In this paper we introduce and use the 'masked gender task' to demonstrate previously unreported spurious correlations between gender pronouns and the following gender-neutral entities for W: time, location, and topics of interest, on unmodified pre-trained BERT large (Devlin et al., 2018) and RoBERTa large (Liu et al., 2019) models.

2. Data Generating Processes

A sample can only be selected into a dataset if it has *access* to the sampling process. We use the term *access* here, as it evokes processes of selection also experienced by human-beings, which we will further motivate when describing the data generating process.

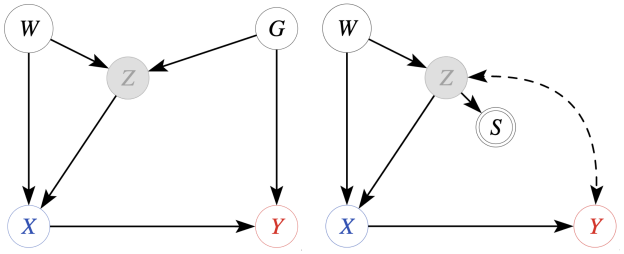
To understand the selection bias intrinsic to a dataset, one must consider the dataset's data generating process. Datasets do not generally admit their generating process, but rather it must be discovered via auxiliary methods such as applying domain knowledge or causal discovery methods. We use intuition and domain knowledge to describe what we hope are plausible data generating processes below.

2.1. Datasets

We ground this discussion with two example datasets, selected as they are representative of data sources used to pre-train LLMs.

[1]Independent Researcher. Correspondence to: Emily McMilin <emcmilin@cs.stanford.edu>.

Published at the ICML 2022 Workshop on Spurious Correlations, Invariance, and Stability. Baltimore, Maryland, USA.



(a) DAG including G: gender, that is later unobserved.

(b) DAG including node, S, representing the dataset selection.

Figure 1. Causal DAG representing the assumed data generating process for the Wiki-Bio and Reddit TLDR datasets.

Specifically, we use the Wikipedia Biography (Wiki-Bio) dataset (Lebret et al., 2016), composed of about 730,000 biographies from English Wikipedia, with which we finetuned models on the roughly 105,000 entries that contained birth *date* and birth *place* data, as well as at least one instance of an explicitly gendered word, from the list in Table 3[1].

And we use Reddit Webis-TLDR-17 (Reddit-TLDR) dataset (Völske et al., 2017), composed of content-summary pairs from Reddit from 2006-2016, with which we finetuned models on the roughly 320,000 entries that again contained at least one explicitly gendered word from Table 3. Both datasets are hosted on Hugging Face's Datasets Library.

2.2. Selection Variables

Recall we desire to demonstrate a learned statistical association between gender, G, and gender-neutral variables, W, that is driven by *access* (now referred as Z) to the dataset sampling process. For our datasets, suitable instantiations for W and Z as related to G are as follows.

For Wiki-Bio: *access* to resources has generally become less gender inequitable over time as *date* (W_0) increases, but not evenly in every *place* (W_1). Generally only those with *access* to resources will achieve the level of notoriety necessary for an entry in Wikipedia.

For Reddit-TLDR: despite many *subreddit* channels (W) having a focus on gender-neutral topics of interest, the style of moderation and community within a given subreddit may reduce gender-equal *access* to participation in that *subreddit*.

2.3. Causal DAGs

The above written descriptions of variables and their causal relationships can be compactly represented in a causal directed acyclic graph (DAG), where nodes are variables and arrows are the direction of causation, as can be seen in Figure 1(a) for our assumed data generating process for the Wiki-Bio and Reddit TLDR datasets.

[1]Further selection bias is introduced here during the filtering of ineligible entries, of unknown effect on finetuned models, which serve referential purposes for the pre-trained model's results.

From Figure 1(a), we see that W and G can cause one's *access*, Z, to dataset selection, as discussed above.

At the bottom we see our dataset's features: X for *text*, and labels: Y for *pronouns*. We argue that despite the complex causal interactions between words in a sentence, the *text* are more likely to cause the *pronouns*, rather than vice versa.[2]

Further we assume that Z and W have an effect on one's life and thus the *text* written about them or by them. And clearly G does cause the gender *pronouns*, Y. However, because the goal of the masked gender task is to mask out explicitly gendered words, we'd argue that G is not a direct cause of the *text*, X. Finally, note Z is grayed out, as it is not explicitly recorded in the dataset.

3. Selection Bias

To explain how selection bias can cause two unconditionally independent variables to become dependent, we revisit the causal DAG representing the data generating processes in Figure 1(a). Causal DAGs are associated with a set of structural equations that together compose a structural causal model (Pearl, 2009). For the *access* variable, the structural equation is $Z := f_z(W, G, U_z)$, where U_z is the exogenous noise of the Z variable, and again W are the variables *date* and *place* for Wiki-Bio, and *subreddit* for Reddit TLDR, and G is gender. Although W and G are independent in Figure 1(a), for all but trivial special cases, they will become statistically associated in the equation f_z for Z.

3.1. Conditioning on a Collider

Thus any model that conditions on the variable Z, will introduce this spurious association between W and G, known as collider bias (Pearl, 2009). Excluding Z from the predictive model seems trivial, especially because Z is not directly observed in the dataset. However, recall that Z is a cause of the selection process, depicted in the selection diagram, Figure 1(b), where the selection node, S, can only take values S=1 for a sample selected into the dataset and S=0 otherwise (Bareinboim and Pearl, 2012).

Recall from Section 1 that during dataset formation, we implicitly condition on S=1, as only selected samples appear in the dataset. Conditioning on a descendant of Z, induces the collider bias relationship, as if we had conditioned on Z directly. We are thus inducing the latent structural equation for f_z into the dataset, from which the statistical association between W and G can be learned by our models.

[2]For example, if the subject is a famous doctor and the object is her wealthy father, these context words will determine which person is being referred to, and thus which gendered-pronoun to use.

3.2. Selection Bias Recoverability

Figure 1(b) is a modified version of our original causal DAG, which satisfies the requirements of the masked gender task, specifically we have obscured G and replaced it with double headed arrows to represent an unobserved common cause of both Z and Y.

Although the act of obscuring gender for gender pronoun prediction may seem contrived, we argue that LLMs are often in similar circumstances, for example whenever a prompt, dialog, translation or classification task has not been provided gender features, yet predictions about gender are required.

Structural causal models very similar to that in Figure 1(b) have been described in practice in (Knox et al., 2020), and proven in (Bareinboim et al., 2014), to be not 'recoverable'. Formally, because we are unable to d-separate the selection mechanism from the label: $(Y \not\perp\!\!\!\perp S|X)$, the conditional distribution of $P(Y|X)$ cannot be determined without further assumptions or additional data about target populations (Bareinboim and Tian, 2015).

4. Masked Gender Task

This lack of recoverability for $P(Y|X)$ is consistent with our goals of underspecification for the masked gender task.

4.1. Inference Test Texts

Revisiting Figure 1(b), to maintain underspecification we now require gender-neutral text values for the input *text*, X, and the *date*, *place*, and *subreddit* variables, W. In addition to gender-neutral, for X we desire extremely simplistic texts, to avoid inducing unrelated spurious correlations. We were unable to find such a benchmark dataset, so we used the heuristic described in Table 1 to generate over 700 input texts for each of the three W categories.

4.2. W x-axis Values

Remaining undefined in the heuristic for input texts in Table 1, is the text values to use for W. These W values will also serve as our x-axis in the coming plots. We require values that are gender-neutral in the real world, yet are hypothesized, due to the selection bias process, to be a spectrum of gender-inequitable values in the dataset.

For W as *date*, it's easy to just use time itself, as over time women have become more likely to be recorded into historical documents reflected in Wikipedia, so we pick years ranging from 1801 - 2011. For W as *place*, we use the bottom and top 10 World Economic Forum Global Gender Gap ranked countries (see B.1). And for W as *subreddit*, we use *subreddit* name ordered by subreddits channels that have an increasingly larger percentage of self-reported female commenters, with a minimum size of 400,000 commenters overall (see B.2).

The original premise for the W variable was to use categories that are gender-neutral in the real world, but not necessarily so in the dataset. To achieve this, we considered filtering the *subreddit* list to only topics deemed gender-neutral, however this subjective process invited too much cherry picking on our behalf. Thus, for the *subreddit* (and *place*) values, we copied the referenced lists verbatim from their sources.

To help disambiguate the role of non-gender-neutral subreddit topic names, from the role of *access* based selection bias, in contributing to a correlation between W and G, we tested on one pre-trained model likely exposed to the selection bias effect during pre-training, and one that was likely not exposed. Specifically, RoBERTa was trained with the same data sources as BERT, plus additional data sources including OpenWebText[3] (Gokaslan and Cohen, 2019). Thus, we'd expect RoBERTa to exhibit a stronger *subreddit* to *gender* correlation, than that of BERT, due to the role of subreddit *access* selection bias during the pre-training of RoBERTa.

4.3. Pre-trained BERT-like models

For testing the masked gender task on pre-trained models, we selected BERT large and RoBERT large, as explained in Section 4.2, using the default weights hosted for the models on Hugging Face.

We are able to test the pre-trained LLMs without any modification to the models, as the masked gender task is simply a special case of the masked language modeling (MLM) task, with which all these models were pre-trained.

Rather than random masking, the masked gender task masks only explicitly gendered words (listed in Table 3). During LLM pre-training, the MLM prediction is a softmax over the entire tokenizer's vocabulary. The masked gender task sums the gendered portion (as listed in Table 3) of that probability mass from the top five predicted words.

4.4. Finetuned Models

We also finetune BERT-like models using a similar masked gender task. The difference being that for our finetuning task, the prediction outcome is binary (as opposed to the entire tokenizer's vocabulary), largely for run-time expediency. We elected to finetune the models with data sources similar to those in their pre-training, so we selected BERT

[3]Although OpenWebText does not explicitly include Reddit data, because it is composed of scraped web content from URLs shared on Reddit that received at least 3 upvotes (Liu et al., 2019), we conjecture subreddit topic names would appear in the context of Reddit in this dataset.

Table 1. Heuristic for creating gender-neutral input texts for the masked gender task for each W variable category, and example rendered text. For `verb` we used past, present and future tenses of the verb *to be*: ["was","is","will be"], and for `life_stage` we used proper and colloquial terms for a range of life stages: ["a child", "a kid", "an adolescent", "a teenager", "an adult", "all grown up"]. We didn't include any `life_stage` past adulthood, as there are not equal gender ratios of elderly men to women, in many locations. Finally for the text versions of `w`, we used a spectrum of values defined in Section 4.2.

W Category	Python f-string	Example text
Date & Place	`'f"[MASK] {verb} {life_stage} in {w}."'`	'[MASK] was a teenager, in 1953.'
	`'f"In {w}, [MASK] {verb} {life_stage}."'`	'In Mali, [MASK] will be an adult .'
Subreddit	`'f"[MASK] {verb} {life_stage}. {w}."'`	'[MASK] is a kid. gifs.'

for the Wiki-Bio dataset and RoBERTa for the Reddit TLDR dataset.

For the Wiki-Bio dataset, we finetune three BERT base models: 1) with birth *date* metadata, 2) birth *place* metadata, and 3) with no extra metadata, prepended to each training sample. In the case of the Reddit TLDR dataset we finetune two RoBERTa base models: 1) with *subreddit* metadata and 2) with no extra metadata, prepended to each training sample.

As these models are finetuned with a single dataset (Wiki-Bio or Reddit TLDR) for the single task of gendered-word prediction, we'd expect them to serve as an upper limit for the magnitude of spurious correlations a model could exhibit for the masked gender task. In particular, the models conditioned with the textual metadata values for W at train time are expected to learn the strongest relationship between W and G.

5. Results

In this section we share the results of the masked gender task, tested on pre-trained BERT and RoBERTa large, as well as our finetuned models which can serve as a rough upper limit for the magnitude of expected spurious correlations.

5.1. Wiki-Bio Date Results

Figure 2(a) shows the results for W as *date* vs gender pronoun predictions. The spurious correlations shown in these plots are consistent with our hypothesized outcome of the selection bias effect, since all of the models were trained on Wikipedia biographical data. Specifically, as *date* increases, women's *access* to resources increases. And thus their representation in Wikipedia increases, inducing the spurious correlation between *date* and *gender*. Due to the nature this collider bias, as female representation goes up, male representation tends to go down[4].

[4]However, in pre-trained models there exist many selection pressures, one of which appears to cause BERT and RoBERTa to be more likely to predict gender pronouns (as opposed to non-gendered pronouns) for both genders after 2000.

Table 2 shows the slope and Pearson's r correlation coefficient (following (Rudinger et al., 2018)) for all the plots in this section. These reported coefficients are limited in many ways, including the improbable assumption that the learned relationship between W and *gender* is linear. Further limiting is that we calculate these coefficients against the index of the x-axis, rather than the *date* value on the x-axis. We do this here for consistency with the coming *place* and *subreddit* plots, for which the most convenient quantitative value we can assign to an ordered list of countries or subreddits is their index in the list.

Thus, these coefficients only serve to compare one model's response to another's for a given W category. The most noteworthy comparison here is that the correlation coefficients of the pre-trained models are comparable to those of the finetuned models for female pronouns, all above a Pearson's r value of 0.75 (perhaps in part as a trade-off for the male pronoun coefficients).

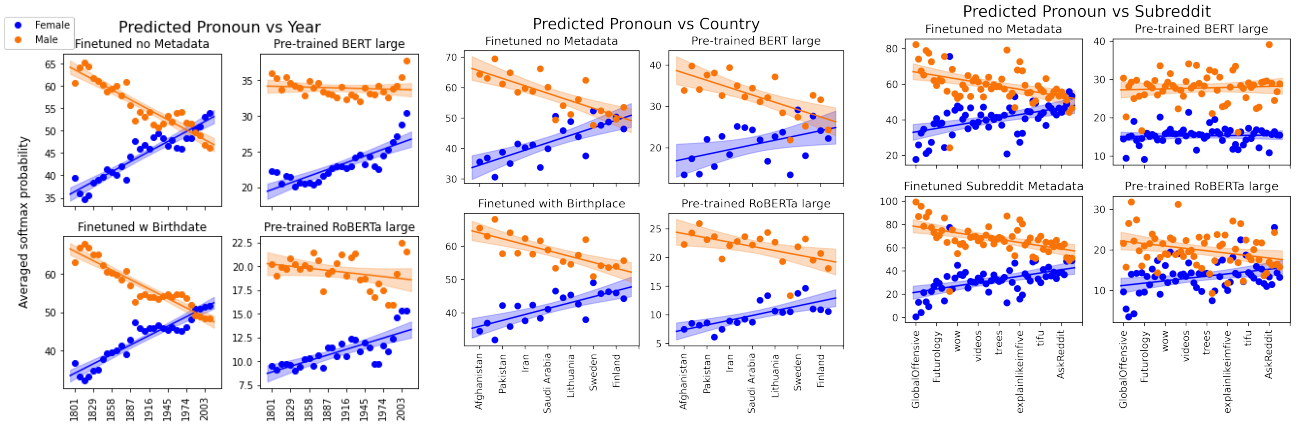
5.2. Wiki-Bio Place Results

The results for W as *place* in Figure 2(b) are similar to those discussed above, although the correlation coefficients appear slightly weaker. As mentioned, reliable comparisons of these coefficients across W variables are limited. We nonetheless conjecture that the slightly weaker correlations for *place* are perhaps in part due to the subjective nature by which the countries are ordered along the x-axis.

Despite this limitation, we still see comparable slope and correlation coefficients between the finetuned and pre-trained models for the spurious correlation between *place* and *gender* pronouns.

5.3. Subreddit Results

A challenge in interpreting the results for W as *subreddit* in Figure 2(c) is that our claim of W as gender-neutral in the 'real-word' is dubious for subreddit names, as compared to dates and country names. As discussed in Section 4.2, we elected against filtering the x-axis to only subreddit names that were more gender-neutral, as this was a subjective process that invited cherry picking.



(a) W as *date*, each dot is an average of 36 softmax probabilities for the mask in input text like "[MASK] will be an adult in 1945.".

(b) W as *place*, each dot is an average of 36 softmax probabilities for the mask in input text like "In Iran, [MASK] was a teenager.".

(c) W as *subreddit*, each dot is an average of 18 softmax probabilities for the mask in input text like "[MASK] is a kid. Futurology".

Figure 2. Averaged softmax percentages for predicted gender pronouns vs a range of W values as described in Section 4.2, for gender-neutral input texts described in Table 1. Shaded regions show the 95% confidence interval for a 1st degree linear fit between W and G. For all plots we expect a correlation between the x-axis and the predictions, except for Pre-trained Bert large's predictions vs subreddit.

To help disambiguate the role of selection bias vs the role of 'real-word' gendered terms for W, we tested one pre-trained model that was likely exposed to the selection bias effect during pre-training, and one that was likely not exposed. Specifically, RoBERTa should learn a more gendered representation for the otherwise gender-neutral subreddit channel names since the selection bias was likely present during its pre-training, whereas BERT should not.

The bold-faced correlation coefficients in Table 2 do appear to confirm this hypothesis, with BERT's slope and r coefficients roughly $\frac{1}{5}$ to $\frac{1}{10}$ that of RoBERTa, but the authors note more robust testing is desired to strengthen this argument.

6. Demonstration and Open-Source Code

The authors would greatly appreciate community feedback that supports or challenges our results. To enable greater access, we have developed a demo of the masked gender task, where users can choose their own input text, as well as the W variable, x-axis values, and the plotted degree of fit, to test spurious correlation to gender in almost any BERT-like model hosted on Hugging Face at https://huggingface.co/spaces/emilylearning/spurious_correlation_evaluation.

We additionally we will make all code available at https://github.com/2dot71mily/spurious_correlations_ICML_2022.

7. Conclusion

In this paper we have introduced and applied the masked gender task to reveal spurious correlations between gender pronouns and real-world gender-neutral entities like dates and countries, on BERT and RoBERTa large pre-trained models. We showed similar spurious correlations between gender pronouns and subreddit channel names, however as we lacked an objectively gender-neutral list of subreddit names, it is more difficult to disambiguate the role of selection bias vs that of non-gender-neutral topic names.

The measured correlations between date, country and subreddit-topic vs the probability of a man or woman existing (as a child, adolescent or adult), may be predictive for Wikipedia entries or Subreddit commenters, as $P(Y|X, W, S{=}1)$, but will not necessarily generalize to the probabilities of men and women existing in real-world inference domains of $P(Y|X, W)$.

Our results indicate that sentences previously considered as gender-neutral baselines for testing gender bias in LLMs (e.g. input text such as 'a woman is walking.' (D'Amour et al., 2020)), are also vulnerable to spurious correlations.

We explained the role of dataset selection bias in inducing the spurious association between otherwise unconditionally independent entities, such as gender and time, and suggested broad applicability beyond the particular relationships investigated here. As mentioned in Section 3.2, further assumptions or data can help to mitigate the effects of selection bias, which we hope to apply to LLMs in the future.

Acknowledgements

Thank you to the SCIS reviewers for their helpful comments, to Rosanne Liu and Jason Yosinski for their encouragement, to Hugging Face for their open source services, and to Judea Pearl, Elias Bareinboim, Brady Neal and Paul Hünermund for their fantastic online causal inference resources.

Table 2. Slopes and Pearson's *r* correlation coefficient for the plots in Figure 2 of spurious correlations between *gender* and several otherwise gender-neutral categories for the W variable: *date*, *place* and *subreddit* channel topics. Values in bold font are for the only experiment in which a model was tested against W variables for which the model under test has no selection bias pressures.

W Category	Model Type	Model Training Details	Female		Male	
			Slope	Pearson's r	Slope	Pearson's r
Date	finetuned	WikiBio no Metadata	0.558	0.929	-0.558	-0.929
		WikiBio w Birthdate	0.616	0.936	-0.616	-0.936
	pre-trained	BERT large	0.235	0.826	-0.016	-0.116
		RoBERTa large	0.149	0.759	-0.055	-0.290
Place	finetuned	WikiBio no Metadata	0.817	0.763	-0.817	-0.763
		WikiBio w Birthplace	0.591	0.752	-0.591	-0.752
	pre-trained	BERT large	0.381	0.476	-0.589	-0.701
		RoBERTa large	0.277	0.724	-0.247	-0.525
Subreddit	finetuned	Reddit TLDR no Metadata	0.243	0.452	-0.243	-0.452
		Reddit TLDR w Subreddit	0.345	0.494	-0.345	-0.494
	pre-trained	BERT large	**0.006**	**0.049**	**0.016**	**0.071**
		RoBERTa large	0.071	0.335	-0.074	-0.300

REFERENCES

Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 100–108, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL https://proceedings.mlr.press/v22/bareinboim12.html.

Elias Bareinboim and Jin Tian. Recovering causal effects from selection bias. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Mar. 2015. doi: 10.1609/aaai.v29i1.9679. URL https://ojs.aaai.org/index.php/AAAI/article/view/9679.

Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1), Jun. 2014. URL https://ojs.aaai.org/index.php/AAAI/article/view/9074.

Alexander D'Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification presents challenges for credibility in modern machine learning. *CoRR*, abs/2011.03395, 2020. URL https://arxiv.org/abs/2011.03395.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus, 2019. URL http://Skylion007.github.io/OpenWebTextCorpus.

James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979. ISSN 00129682, 14680262. URL http://www.jstor.org/stable/1912352.

Dean Knox, Will Lower, and Jonathan Mummolo. Administrative records mask racially biased policing. *American Political Science Review*, 114(3):619–637, 2020. doi: 10.1017/S0003055420000039.

Rémi Lebret, David Grangier, and Michael Auli. Generating text from structured data with application to the biography domain. *CoRR*, abs/1603.07771, 2016. URL http://arxiv.org/abs/1603.07771.

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

Judea Pearl. *Causality*. Cambridge University Press, Cambridge, UK, 2 edition, 2009. ISBN 978-0-521-89560-6. doi: 10.1017/CBO9780511803161.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. Gender bias in coreference resolution. *CoRR*, abs/1804.09301, 2018. URL http://arxiv.org/abs/1804.09301.

Michael Völske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4508. URL https://www.aclweb.org/anthology/W17-4508.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models, 2020. URL https://arxiv.org/abs/2010.06032.

Table 3. List of explicitly gendered words that are masked out for prediction as part of the masked gender task. These words were largely selected for convenience, as each is a single token in both the BERT and RoBERTa tokenizer vocabs, for ease of downstream token to word alignment. During finetuning, it is expected that this list will not fully mask gender in every sample, reducing the underspecification of the learning task and the potential learning of gender-neutral spurious associations to gender. At inference time, it is critical that all gendered words are masked, and because the inference input texts are constructed by a heuristic, this is trivial to achieve.

MALE-VARIANT	FEMALE-VARIANT
HE	SHE
HIM	HER
HIS	HER
HIMSELF	HERSELF
MALE	FEMALE
MAN	WOMAN
MEN	WOMEN
HUSBAND	WIFE
FATHER	MOTHER
BOYFRIEND	GIRLFRIEND
BROTHER	SISTER
ACTOR	ACTRESS

A. Explicitly Gendered Words

See Table 3 for list explicitly gendered words that were masked for prediction during both finetuning and at inference time for the masked gender task.

B. *W* variable x-axis values

B.1. Place Values

Ordered list of bottom 10 and top 10 World Economic Forum Global Gender Gap ranked countries used for the x-axis in Figure 2(b), that were taken directly without modification from https://www3.weforum.org/docs/WEF_GGGR_2021.pdf:

"Afghanistan", "Yemen", "Iraq", "Pakistan", "Syria", "Democratic Republic of Congo", "Iran", "Mali", "Chad", "Saudi Arabia", "Switzerland", "Ireland", "Lithuania", "Rwanda", "Namibia", "Sweden", "New Zealand", "Norway", "Finland", "Iceland"

B.2. Subreddit Values

Ordered list of subreddits used for the x-axis in Figure 2(c), that were taken directly without modification from http://bburky.com/subredditgenderratios/ with minimum subreddit size: 400,000.

"GlobalOffensive", "pcmasterrace", "nfl", "sports", "The_Donald", "leagueoflegends", "Overwatch", "gonewild", "Futurology", "space", "technology", "gaming", "Jokes", "dataisbeautiful", "woahdude", "askscience", "wow", "anime", "BlackPeopleTwitter", "politics", "pokemon", "worldnews", "reddit.com", "interestingasfuck", "videos", "nottheonion", "television", "science", "atheism", "movies", "gifs", "Music", "trees", "EarthPorn", "GetMotivated", "pokemongo", "news", "Fitness", "Showerthoughts", "OldSchoolCool", "explainlikeimfive", "todayilearned", "gameofthrones", "AdviceAnimals", "DIY", "WTF", "IAmA", "cringepics", "tifu", "mildlyinteresting", "funny", "pics", "LifeProTips", "creepy", "personalfinance", "food", "AskReddit", "books", "aww", "sex", "relationships"