Selection Collider Bias in Large Language Models

Abstract

In this work we show how large language models (LLMs) can learn statistical dependencies between otherwise unconditionally independent variables, that we argue can render the learning of disentangled causal representations infeasible. To demonstrate the effect, we developed a gender pronoun prediction task that can be applied to BERT-family models to produce dose-response relationship plots between gender prediction and a variety of seemingly gender neutral variables like date and location, on pre-trained (unmodified) BERT, DistilBERT, and XLM-RoBERTa, as well as fine-tuned models. Finally, we provide an online demo, inviting readers to experiment further.

1 INTRODUCTION

In datasets with cause and effect: X and Y, and variables: Z and W, we can train models to learn from the conditional distribution: $P(Y|X, Z, W)$. In this paper we will be considering cases where Z is the cause of dataset selection bias. Datasets without some form of selection bias are rare, as almost all datasets are subsampled representations of a larger domain, yet few are sampled at random[1].

Sometimes selection bias is localized to a single variable, for example if no one under five has been approved for a vaccination, then a survey about vaccine choice and health outcomes may have data missing due to age. However, the type of selection bias that interests us here is that which involves more than one variable (observed or not). And importantly, although these variables are unconditionally independent of one another in the real world, they are a common cause of the variable that represents some form of *access* to the dataset. In order to form a dataset, one must condition on that 'access' variable, thus inducing the causes of the 'access' variable into a zero-sum-game, known as collider bias Pearl [2009].[2]

To distinguish between the two types of selection bias mentioned above, we will call the latter 'selection collider bias'. Beyond posing a risk to out-of-domain generalizability, selection collider bias can result in models that lack even 'internal validity', as the associations learned from the data represent the statistical dependences induced by the dataset formation and not the data itself Griffith et al. [2020].

In this paper we demonstrate that the learning of disentangled representations for certain variables may be made infeasible due 'selection collider bias' resulting in violations to the Independent Causal Mechanisms (ICM) Principle Schölkopf et al. [2021], despite the unconditional independence of these variables in the real world. Our results are agnostic to specific learning algorithm. Further, our results indicate that the massive scaling up of data from a broad range of diverse datasets, as has been done for LLMs may be thus far insufficient to overcome the persistence of selection collider bias caused by a small subset of the training corpus.

1.1 BRIEF OUTLINE

Within the realm of causal representation learning, we are not in the domain where we seek to construct a disentangled representation of the structural causal variables from high dimensional data. Rather we are proposing a structural causal model based on our assumptions of a dataset's data generating process, and seeking validation of these assumptions via a model's learned representations from this dataset.

To highlight the effect of selection collider bias, we desired a learning task where the selection mechanism would be

[1] This is some times referred to missing not at random (MNAR).

[2] Not all selection bias is collider bias, as is discussed here. Further not all collider bias is selection bias, if that collider bias does not result in selection into a dataset.

Submitted to the 38th Conference on Uncertainty in Artificial Intelligence (UAI 2022). **To be used for reviewing only**.

easily measurable and interpretable. We decided to focus on LLMs and the prediction of gender pronouns, as despite the relatively even distribution of genders across time, place and interests, there are known gender disparities in terms of access to resources, that we believed could be captured in a plausible data generating process.

In the next section we introduce this data generating process in the form of a causal directed acyclic graph (DAG), and then go on to further detail selection collider bias. We then describe the pronoun prediction task and the model's results for this task, before concluding with a discussion of our findings.

2 DATA GENERATING PROCESSES

Datasets do not generally admit their data generating process, but rather it must be discovered via auxiliary methods such as applying domain knowledge or causal discovery methods. These methods require the application of informed assumptions that can be compactly represented as a causal DAG. Once represented as a DAG, it is trivial to establish whether the learning task is 'identifiable' Pearl [2009] from the dataset, or whether confounders may be present, that will reduce the learning task to that of learned associations.

2.1 DATASETS

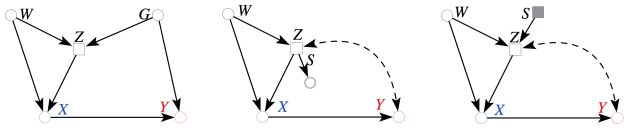
We seek to highlight how the real-world distribution of genders across the variables *time*, *place* and *interests* may be inaccurately represented in datasets, in particular among those datasets used to train LLMs. We decided to focus on text derived from Wikipedia which is used in training BERT Devlin et al. [2018], DistilBERT Sanh et al. [2019], RoBERTa Liu et al. [2019] and XLM-RoBERTa Conneau et al. [2019], and on Reddit which is used in training XLM-RoBERTa Conneau et al. [2019],[3] but not other LLMs.[4]

However, partially due to limited GPU resources, time and access to the exact post-processed pre-training datasets, we elected to use proxies for Wikipedia and for the Reddit portions of CommonCrawl. Specifically, we use the Wikipedia Biography (Wiki-Bio) dataset Lebret et al. [2016] and Reddit Webis-TLDR-17 (Reddit-TLDR) dataset V"olske et al. [2017], both hosted on Hugging Face.

We also selected these datasets as they had nice proxies for the above mentioned variables: birth `date` and birth `place` in Wiki-Bio and `subreddit` interest in Reddit-TLDR.

[3]The authors believe Reddit data was included in the portion of the cleaned CommonCrawl data Conneau et al. [2019] used by XLM, but have not verified.

[4]RoBERTa was not trained on actual Reddit text but rather text scraped from URLs shared on Reddit with at least three upvotes Liu et al. [2019].



(a) DAG including G: gender, that is unobserved in other DAGs.

(b) DAG including S: selection node, where $S = 1$ for samples in dataset

(c) Selection diagram with S indicator node demarking train vs. inference.

Figure 1: DAGs with cause and effect, X: text features, and Y: gender pronouns, Z: access, W: date, or place, or interest.

3 DATASET'S CAUSAL DAGS

In Figure 1a we see plausible data generating processes for the Wiki-Bio and Reddit TLDR datasets. The variables W: birth `place`, birth `date` or `subreddit` interest, and G: `gender`, are both independent variables that have no ancestral variables. However, W and G may have a role in causing one's `access`, Z. In the case of Wiki-Bio a functional form of Z may capture the general trend that `access` has become less gender-dependent over time, but not in every `place`. In the case of Reddit TLDR, Z may capture that despite some subreddits having gender-neutral topics, the specific style of moderation and community in the subreddit may reduce `access` to some genders.

We also see that Z: `access` and W: `place`, `date` and `subreddit` interest, will all have an effect on one's life and thus the words written about them or by them. However, due to our attempt to remove gender non-neutral words from the text, we'd argue (perhaps in the ideal gender-equal world) that G is not a direct cause of the text.

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 all the words that compose a biography, the `text` are more likely to cause the `pronouns`, rather than vice versa.[5]

Finally note that Z is a square because we have conditioned on `access` as part of the dataset selection process, which we discuss more in the next section.

4 SELECTION COLLIDER BIAS

As a quick example of selection collider bias, if we were to ask you the gender of some random person born in 1801, and one in 1999, you may toss a coin to determine your answer, as birth date and gender are unconditionally independent in the real world. However, if instead we where to ask about the

[5]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.

gender of a person born in 1801, and one in 1999, that we saw on two random Wikipedia articles today, then you may condition your guess on some combination of birth date, gender, and importantly, what gets recorded in Wikipedia.

If Wikipedia could record every life, we would imagine a very different mental process, more similar to the answer for our first question. Yet, despite the vastness of Wikipedia and its capability to further expand indefinitely, there is a finite number of recording resources. Thus, achieving access into Wikipedia's record, creates unconditional dependencies between otherwise unconditionally independent variables.

As we have been discussing, data generating processes depicted in Figure 1a is prone to collider bias when conditioning on `access`. In other words, although in real life `place`, `date`, `interest` and `gender` are all unconditionally independent, when we condition on their common effect, `access`, they become unconditionally dependent. The obvious solution to not condition on `access` is unavailable to us, as we are required to do so in order to capture the process of selection into the dataset. This dataset selection is depicted in Figure 1b, with the S node showing that Z is the cause of the selection into dataset Bareinboim et al. [2014]. In this graph we also see that G has been replaced by a double headed arrow, representing that G is unobserved, yet still a common cause of Z and Y.

Structural causal models very similar to that in Figure 1b have been described in practice in Knox et al. [2020] and proven in Bareinboim et al. [2014] to be not 'recoverable'. This lack of recoverability means that the desired conditional distribution of $P(Y|X)$ can not be determined, regardless of learning algorithm, because we are unable to d-separate the selection mechanism from the label: $(Y \not\perp\!\!\!\perp S|X)$. This is perhaps not too surprising, as this task is a difficult task for humans as well, and in the ideal of a gender equal world, it would be impossible to predict one's gender based only on gender-neutral text about them.

However, LLMs are asked to make this type of prediction all the time, for example whenever a prompt or dialog has not yet revealed the gender of the subjects. In the next sections we will examine how LLMs predictions of gender pronouns are altered by the conditional dependencies induced by selection collider bias.

5 PRONOUN PREDICTION TASK

In this section we discuss how we applied the gender pronoun prediction task to the models, for which we will show results in the subsequent section.

5.1 PRE-TRAINED BERT-LIKE MODELS

We are able to probe the pre-trained LLMs without any modification to the models, as the gender-pronoun prediction 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 gender-pronoun prediction task masks only non-gender-neutral terms (listed in Appendix B). For the pre-trained LLMs the final prediction is a softmax over the entire tokenizer's vocabulary, from which we sum up the portion of the probability mass from the top five prediction words that are gendered terms (again listed in Appendix B).

5.2 FINE-TUNING TASKS

We also fine-tune BERT-family models using a similar gender-pronoun prediction task. The difference being that for our fine-tuning task, the prediction outcome is binary (as opposed to the entire tokenizer's vocabulary), largely for run-time expediency. We elected to fine-tune the models with data sources similar to those in their pre-training, so we selected BERT for the Wiki-Bio data and XLM-RoBERTa for the Reddit TLDR dataset.[6]

We fine-tuned several models for each dataset. For the Wiki-Bio dataset, we fine-tuned three 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 fine-tuned two models: 1) with `subreddit` interest metadata and 2) with no extra metadata, prepended to each training sample.

6 RESULTS

The causal structural model in Figure 1a informs the structural equation for Z, the access variable is $Z := f(W, G, U_z)$, where W are the variables `date` & `place` for Wiki-Bio, and `subreddit` for Reddit TLDR datasets, G is the unobserved gender, and U_z is the exogenous noise of the Z variable. As discussed above, although `date`, `place`, and `subreddit` interest[7] are generally unconditionally independent of gender, Equation 6 shows us that with the inclusion of Z in our dataset, we introduce a functional dependency between these variables and gender. Further, as discussed in Section 4, we are not at liberty to exclude Z from the learning model, as this variable represents access to the dataset, already intrinsic to the dataset itself.

[6] See footnote 3.

[7] While the authors do acknowledge that some subreddit names may be considered non-gender-neutral, in the next section we will see the selection collider bias effect for many subreddit names that do appear gender neutral.

6.1 VARIABLE SWEEPS

To visualize the selection collider bias induced functional dependency learned by LLMs, between these variables and gender, we plot their dose-response relationships: where we expect a *larger* intervention in the treatment (the variable value in text form injected into the input text) produces a larger response in the output (the average softmax probability of a gendered pronoun).

Specifically, we will sweep through a spectrum of birth `place`, birth `date` and `subreddit` interest injected into otherwise unchanged input text (with contextual nudges toward more female and male pronouns) in both the fine-tuned and pre-trained models. This requires a spectrum of less to *more* gender-equal values for each variable.

For `date`, it's easy to just use time itself, as gender equality has generally improved with time, so we picked years ranging from 1800 - 1999. For `place` we used the bottom and top 10 Global Gender Gap ranked countries. (See Appendix D.1.) And for `subreddit`, we use `subreddit` name ordered by subreddits that have an increasingly larger percentage of self-reported female commenters.[8] (See Appendix D.2.)

In all cases we injected each variable value into the input-text: `"<mask> works as a {job}."` where `job` is replaced with a list of either traditionally female-like or male-like jobs from Appendix 2. For `date` and `place` we prepend the input-text with `"born in {date},"` or `"born in {place},"`, where we used the spectrum of date and place values described above. For `subreddit`, we appended the input-text with `"Source: r/{subreddit}."` for the range of subreddits mentioned above.

6.2 DOSE-RESPONSE PLOTS

We initially refer to Figure 2a in generic terms as a means of introducing the general structure of the figures.

Each variable of interest has the dose-response results in a single figure composed of four sub-figures. The sub-figures all arranged from top to bottom as: 1) fine-tuned model trained on dataset text with variable of interest appended, 2) fine-tuned model trained without additional metadata, 3) and 4) BERT-like pre-trained models

Each sub-figure has four plots of the softmax probabilities for the predicted gendered terms in Appendix B, averaged over either the female-like or male-like occupation types, as follows: 1) female predictions for female-like jobs, 2) female predictions for male-like jobs, 3) male predictions for female-like jobs, and 4) male predictions for male-like jobs.

Every datapoint in these sub-figures show the average softmax probability for the predicted gender pronouns for the masked word in the input text described in the prior section, and we have added a linear[9] fit between the x-axis index and the softmax probabilities, to help highlight potential trends.

6.3 WIKI-BIO RESULTS

.

Figure 2a shows the results for the `date` sweep for four models. All four models show that the likelihood of the model predicting a male pronoun for any job type goes down with increasing `date`, while the likelihood of predicting a female pronoun for any job type goes up with increasing `date`.

There is almost no difference between the top two models, suggesting that the additional appending of birth-date information during training had little impact. The authors speculate this could be due to the prevalence of other date information already present in many of the Wiki-Bio samples.

For pre-trained BERT the dose-response relationship is slightly less strong than that of the fine-tuned models, and some of the predicted words are not gendered. While for DistilBERT the strong majority of predicted words are not gendered at all.

Similar results for the dose-response relationship between country and gender are shown in the Appendix A.

6.4 SUBREDDIT RESULTS

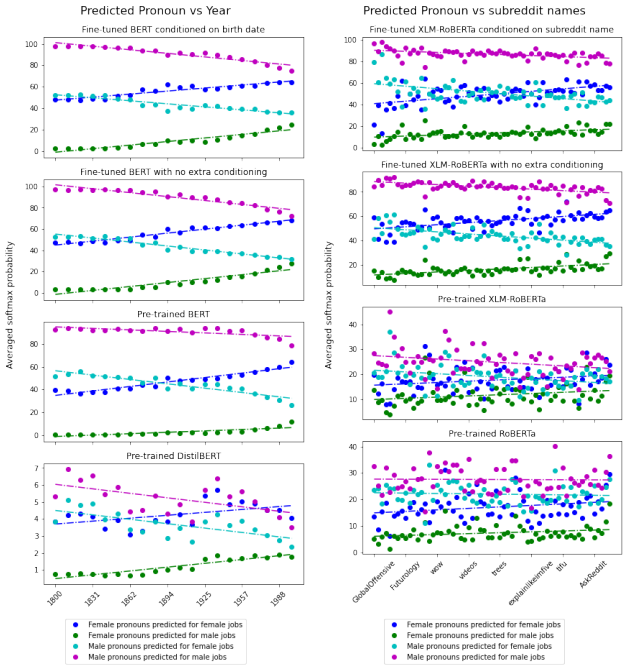
The results for the `subreddit` sweep in Figure 2b are again similar to those above, but noteworthy in several ways.

We see a much noisier dose-response relationship, however, this is not at all surprising, as the spectrum of subreddits on the x-axis are based on the very small minority of self reported gender representation in each subreddit. We also see that the bottom-most plot of predictions from RoBERTa shows almost no dose response relationship at all, which we will discuss more in the next section.

Some may argue that the `subreddit` names are actually not gender neutral. If this was the case, the dose-response relationship seen in the fine-tuned and XLM-RoBERTa models is not due to selection bias based on gender disparate `access` to the subreddit, but instead due to some subreddit names being more male-like or more female-like. Although

[8] To discourage our own cherry picking, we copied the entire list of subreddits that had a minimum subreddit size of 400,000.

[9] A linear fit was chosen here for simplicity and due to an absence of assumptions on any other functional form for the dose-response relationship. However, in the public online demo described, we allow users to pick the degree of fit.



(a) (b)

Figure 2: Averaged softmax percentages for gendered pronouns predicted to replace the mask in a) "born in {date}, <mask> works as a {job}" and in b) "<mask> works as a {job}. Source: r/{subreddit}", where `job` was filled in with either 'male-like' or 'female-like' jobs (see Appendix 2). In a) the text filled in for {date} and the x-axis is the range of years from 1801 - 1999, and in b) the text filled in for {subreddit}, and the x-axis, is a list of `subreddit` names, with percentage of self-reported female users per subreddit increasing to the right (see Appendix D.2).

the subreddits names appear largely gender neutral to the authors, without a baseline for how XLM-RoBERTa views their gender neutrality, we cannot make interpretations.

Yet, RoBERTa could be just that baseline, as RoBERTa was trained in a manner similar to XLM-RoBERTa, yet it was not trained on Reddit data, while RoBERTa-XLM was.[10] Were the dose-response relationship due only to the lack of gender neutrality of the subreddit names, we would expect to see a similar dose-response relationship in both RoBERTa and XLM-RoBERTa. The comparable lack of a dose-response relationship in the RoBERTa in Figure 2b suggests that this effect is not due to subreddit names, but instead the selection bias induced by training XLM-RoBERTa on subreddit data.

6.5 DEMO

We have provided an online demonstration of the pronoun prediction task, where users can select any model on Hugging Face that supports the 'fill-mask' task (with BERT, DistilBERT, RoBERTa and XLM-RoBERTa preloaded), select their own x-axis for the dose-response plots, select various plotting options, and of course select their input text of choice, with results rendering in several seconds enabling further experimentation.

[10] See footnote 3.

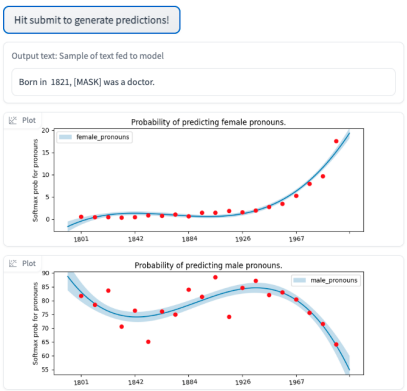


Figure 3: Partial screenshot of publicly available demo of our gender pronoun predicting task supporting a wide range of pre-trained models.

7 TRANSPORTABILITY

While we have focused most of our use of causal inference methods on exposing threats to robust model training, we can end with a causal inference technique that may help ameliorate the problem. The Selection Diagram in Figure 1c can be shown to be *s-admissible*, because the variable Z can d-separate Y from S in the $do(X)$ manipulated version of the DAG: $(Y \perp\!\!\!\perp S|Z)_{G_{\overline{x}}}$ Pearl and Bareinboim [2011]. This can permit statistical transport of the learned representations from the training domain, to an inference domain (denoted by the S square in Figure 1c), when conditioning on the access variable, Z, with the following reweighing of formula for the conditional probabilities observed in the inference domain, P^*, as:

$$P^*(Y|do(X), Z) = \sum_W P^*(Y|X, W, Z)P^*(W)$$

While we may not have access to Z and W of the inference domain, this does suggest that it may be possible to improve predictive performance of a model running inference in a more modern domain, for example, by more heavily weighing the softmax predictions associated with more recent dates.

8 DISCUSSION

One benefit of the collider structure is that we can uniquely verify the causal structure from a joint observational dis-

tribution of three variables Pearl [2009]. We argue the dependencies plotted in Section 6 between the variables and gender, suggest a validation of the assumed causal DAG and selection collider bias mechanism shown in Figures 1a and 1b, respectively.

The absence of a dose-response relationship to subreddit name for the LLM not trained on data sourced from Reddit, and yet the presence of a relationship to subreddit name for the LLM trained on Reddit[11] shown, in Figure 2b, suggests the plotted relationships reflect a representation learned under the pressure of dataset selection bias within the subreddit channel itself, and not just the gender-non-neutrality of the subreddit name alone.

We argue this dataset induced statistical entangling of high level causal variables prevents all models from learning disentangled representations for these same causal variables.

References

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

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.

Gareth J. Griffith, Tim T. Morris, Matthew J. Tudball, Annie Herbert, Giulia Mancano, Lindsey Pike, Gemma C. Sharp, Jonathan Sterne, Tom M. Palmer, George Davey Smith, Kate Tilling, Luisa Zuccolo, Neil M. Davies, and Gibran Hemani. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature Communications*, 11(1), November 2020. doi: 10.1038/s41467-020-19478-2. URL https://doi.org/10.1038/s41467-020-19478-2.

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.

Judea Pearl and Elias Bareinboim. Transportability of causal and statistical relations: A formal approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, December 2011. doi: 10.1109/icdmw.2011.169. URL https://doi.org/10.1109/icdmw.2011.169.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning, 2021. URL https://arxiv.org/abs/2102.11107.

Michael V"olske, 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.

[11] See footnote 3.

Table 1: Non-gender-neutral terms [MASKED] for gender-prediction.

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

A WIKI-BIO PLACE RESULTS

Figure 2a shows the results for the `place` sweep for four models, with results similar to those discussed in Section 6.3.

B NON-GENDER-NEUTRAL WORDS

See Table 1 for list non-gender-neutral words that were masked out during both fine-tuning and at inference time for our gender-pronoun predicting task.

C GENDERED JOBS

See Table 2 for list of traditionally male-like and female-like jobs that were used in the input text at inference time. These jobs were selected in a largely random process from: `https://github.com/johnlsheridan/occupations/blob/master/occupations.csv`

D VARIABLE X-AXIS VALUES

D.1 PLACE VALUES

Ordered list of bottom 10 and top 10 Global Gender Gap ranked countries used for the x-axis in Figure 4, 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"

Table 2: Gender-like occupations used in inference tests.

MALE_LIKE_JOBS	FEMALE_LIKE_JOBS
ASTRONOMER	CASHIER
BIOLOGIST	CLEANER
CARPENTER	HOUSEKEEPER
DOCTOR	HYGIENIST
ENGINEER	LIBRARIAN
EXECUTIVE	MANICURIST
JUDGE	NANNY
MECHANIC	NURSE
PHYSICIST	RECEPTIONIST
PREACHER	SECRETARY
SHERIFF	SOCIAL WORKER
SURGEON	TEACHER

D.2 SUBREDDIT VALUES

Ordered list of subreddits used for the x-axis in Figure 2b, that were taken directly without modification from `http://bburky.com/subredditgenderratios/` with minimum subreddit size: 400000. Note Reddit: "Data through the end of November 2017 is included in this analysis." `https://nbviewer.org/github/bburky/subredditgenderratios/blob/masterSubreddit%20Gender%20Ratios.ipynb`:

"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"

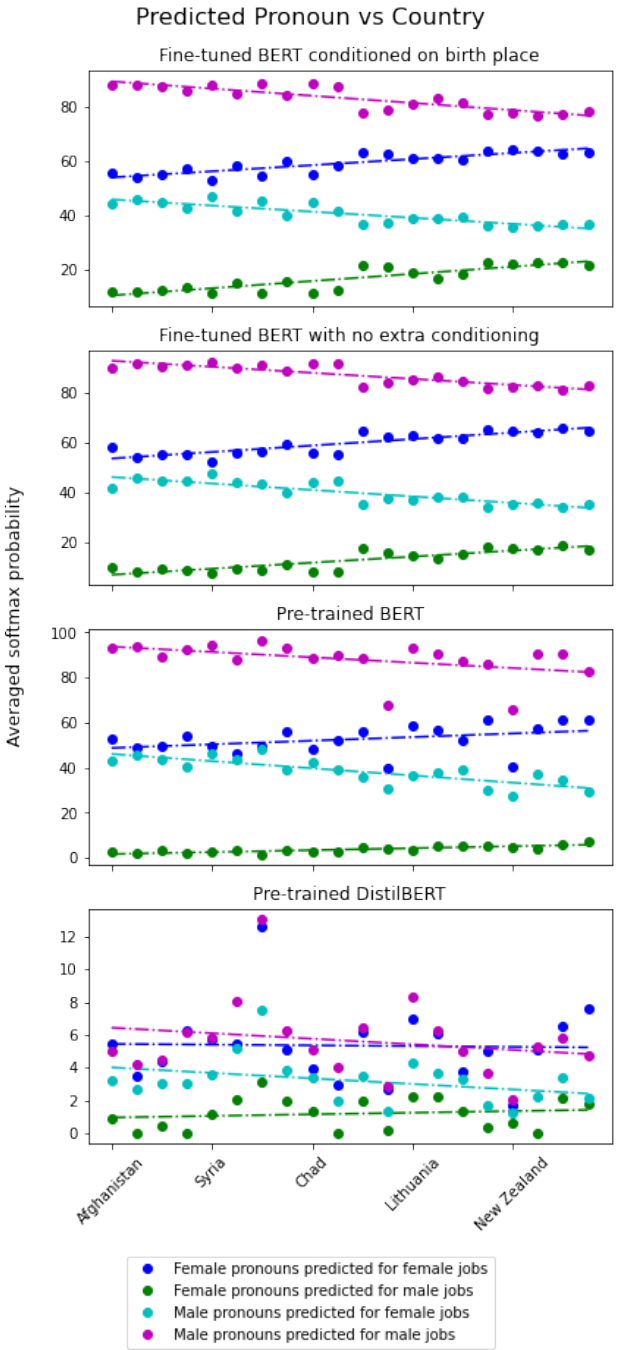


Figure 4: Averaged softmax percentages for gendered pronouns predicted to replace the mask in "born in {place}, <mask> works as a {job}" where `job` was filled in with either 'male-like' or 'female-like' jobs (see Appendix 2). The text filled in for a) {place} and the x-axis is the bottom 10 and top 10 Global Gender Gap ranked countries (see Appendix D.1).