# Selection Bias Induced Spurious Correlations in Large Language Models

**Anonymous Authors**[1]

## Abstract

In this work we explore the role of dataset selection bias in inducing and amplifying spurious correlations in large language models (LLMs). To highlight known discrepancies in gender representation between what exists in society and what is recorded in datasets, we developed a gender pronoun prediction task. We demonstrate and explain a dose-response relationship in the magnitude of the correlation between gender pronoun prediction and a variety of seemingly gender neutral variables like date and location on pre-trained (unmodified) BERT, DistilBERT, and XLM-RoBERTa models. We also fine-tune several models with the gender pronoun prediction task to further highlight the spurious correlation mechanism, and make an argument about its generalizability to far more datasets. Finally, we provide an online demo, inviting readers to experiment with their own interventions.

## 1. Introduction

Although genders are relatively evenly distributed across time, place and interests, there are also known gender disparities in terms of access to resources. We propose that this access disparity can result in dataset selection bias, causing models to learn a surprising range of spurious associations. These spurious associations are often considered undesirable, as they do not match our intuition about the real-world domain from which we derive samples for inference-time prediction. This discrepancy between the training domain and the inference domain is a common problem to many machine learning tasks and can be depicted as follows.

In models trained on datasets with cause and effect: $X$ and $Y$, covariates: $Z$, and selection bias: $S$, we first train the model on the conditional distribution sampled from the selection-biased dataset as:

$$P(Y|X,Z,S) \tag{1}$$

And then run inference on the interventional distribution, in which users choose which samples to select for the model as:

$$P(Y|X,Z,do(S)) \tag{2}$$

In Equation (2) we have deployed the *do*-operator described by Pearl in (2009), to capture that *do*-ing text selection at inference-time serves as an intervention upon the original dataset selection mechanism.

### 1.1. Brief Outline

For any given dataset, in to order determine which variables serve as $X$, $Y$, $Z$, and $S$ from above, we must first consider the dataset's data generating process, which we describe in the next section. We then go on to describe the learning task we used to fine-tune our models to perform token-level binary classification of masked out non-gender-neutral words. We then show in the results section how these fine-tuned models, as well as their pre-trained counterparts, alter their pronoun prediction with increasing magnitude, along a spectrum of covariate values in otherwise neutral text. Finally we discuss how these findings may extend to a broader class of datasets and trained models.

## 2. Data Generating Processes

Datasets do not generally emit 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 directed acyclic graph (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 *covariates* of *time*, *place* and *interests*

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the SCIS workshop.

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)[1], but not other LLMs[2].

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 *covariates*: birth `date` and birth `place` in Wiki-Bio and `subreddit` interest in Reddit-TLDR.

## 3. Dataset's Causal DAGs

In Figure 1 we see plausible data generating processes for the Wiki-Bio dataset on the left and the Reddit TLDR dataset on the right.

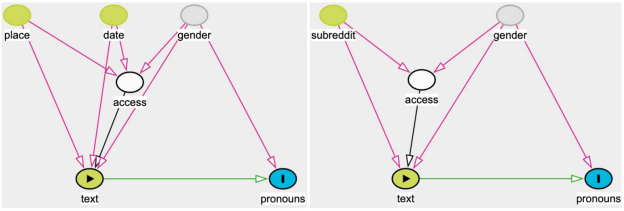


*Figure 1.* Plausible causal DAG representing the relevant portion of the data generating process for Wiki-Bio (left) and Reddit TLDR (right) datasets.

Starting with Wiki-Bio, we can see that birth `place`, birth `date` and `gender` are all independent variables that have no ancestral variables. However, `place`, `date` and `gender` may all have a role in causing one's `access` to resources, with the general trend that `access` has become less gender-dependent over time, but not in every `place`, with recent events in Afghanistan providing a stark counterexample to this trend. Further, `access` directly effects one's chances of achieving the level of fame or infamy necessary to be included in Wikipedia.

We also see that `access`, `place`, `date` and `gender` will all have an effect on the life one lives, and thus the `text` that would be written about one's life.

At the bottom we see our dataset's features: `text`, and labels: `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.[3]

Additionally, `gender` is a confounder of both `text` and `pronouns`, leading to a confounding bias that matches our intuition about the world, and will lead to associations that few would consider to be spurious in nature. Note also that `gender` is grayed out in Figure 1 as we are intentionally attempting to obscure it from the text.[4] Finally note that `access` is circled in black, because we have conditioned on `access` as part of the selection process, which we discuss more in the next section.

Much described above for Wiki-Bio also holds for the Reddit TLDR dataset, on the right of Figure 1. The main distinction is that `place` and `date` have been replaced with `subreddit`. Here we have assumed that despite some subreddits having gender-neutral topics, the specific style of moderation and community in the subreddit may reduce `access` to some genders, again leading to selection bias (at the subreddit granularity).

## 4. Selection Bias

The data generating processes depicted in Figure 1 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, forced into a zero-sum game where some must lose for others to win. 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. Thus, the act of selecting samples for a specific dataset is inducing the same collider bias that we would see if conditioned on `access` in our model.

In Figure 1 the selection covariate, `access`, is a direct cause of the `text`, and is statistically associated with the `pronouns` (via unobserved `gender`). Similar causal selection relationships defined theoretically in (Bareinboim et al., 2014) and described in practice in (Knox et al., 2020),

[1]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.

[2]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).

[3]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.

[4]Although this seems a contrived process, this scenario of unobservable `gender` is common to many text generation tasks.

are shown to be not 'recoverable'[5] due to the lack of conditional independence of the selection criteria from treatment and outcome. This lack of conditional independence is mathematically represented below, where {} represents the empty set, or the absence of any covariate that we may condition upon:[6] $\texttt{access} \not\perp\!\!\!\perp \texttt{pronouns}, \texttt{text} | \{\}$.

It is noteworthy that if we were able to intervene on access, then we'd break the flow of statistical association to the treatment: $\texttt{access} \perp\!\!\!\perp \texttt{pronouns}$

# 5. Models

## 5.1. Pre-trained BERT-like models

We are able to test 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 Table 1). 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 Table 1).

## 5.2. Finetuning tasks

We also fine-tune BERT-like 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 runtime 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[7].

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 models were trained to learn the conditional distribution of Equation (1) where $X$= text, $Y$=pronouns, $S$=access and $Z$ is one of the above mentioned metadata covariants.

At inference time, we are probing the interventional distribution in Equation (2). We apply the same variables for $X$, $Y$, $Z$, and $S$ as above, but now we can intervene upon who gets access into the model, regardless of their covariate values. Whereas in training we may only see male (and the very occasional extraordinary female) doctors in 1850, at inference we can expose the model to any desired distribution of doctor genders across time.[8]

## 6.1. Covariate Sweeps

One intuitive way to see the impact that changing one variable may have upon another is to look for a dose-response relationship, in which a *larger* intervention in the treatment (the covariate value in text form injected in the otherwise unchanged text sample) 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, while intervening on access, in both the fine-tuned and pre-trained models. This requires a spectrum of less to *more* gender-equal values for each covariate.

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 D.1.) And for subreddit, we use subreddit name ordered by subreddits that have an increasingly larger percentage of self-reported female commenters.[9] (See D.2.)

In all cases we injected each covariate 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 Table 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 2 in generic terms as a means of introducing the general structure of the figures.

Each covariate of interest has the dose-response results in a

[5] (Bareinboim et al., 2014) uses selection nodes, with distinct requirements necessary for their generalized treatment of the problem which our specific case does not require.

[6] Gender is unavailable due to being unobserved.

[7] See footnote 1.

[8] We argue this *do*-operation of Equation (2) is a very practical action that often takes place when out-of-domain samples are fed to a model at inference-time.

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

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 covariate 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 Table 1, 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.

### 6.3. Wiki-Bio Date Results

Figure 2 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.

### 6.4. Wiki-Bio Place Results

The results for the `place` sweep in Figure 3 are comparable to those discussed above, although the dose-relationship is noisier. We speculate this is due to the subjective nature by which the countries are ordered along the x-axis and due to the regional shifting nature of gender equality.

### 6.5. Subreddit Results

The results for the `subreddit` sweep in Figure 4 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.

### 6.6. Demo

In the Appendix we describe a demo hosted on a public website where users experiment with their own input text as shown in Figure 5.

## 7. Discussion

While the spurious association in LLMs of gender pronouns vs jobs types is well documented, we have now also shown spurious association between: gender pronouns vs dates, vs countries, and vs subreddit names. We have shown a fine-tuning task that can amplify the strength of these associations.

These new spurious associations may be more surprising at first because, unlike jobs which are known sources of gender disparity, `date`, `place`, and `subreddit`[10] names are largely considered gender neutral. However, as we have argued in this paper, the selection of these covariates into datasets is a zero-sum-game, with even our high quality datasets forced to trade off one for another, thus inducing selection bias into the learned associations of the model.

## References

Bareinboim, E., Tian, J., and Pearl, J. 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.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. Unsupervised cross-lingual repre-

[10]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 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 (see footnote 1). 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 4 suggests that this effect is not due to subreddit names, but instead the selection bias induced by training XLM-RoBERTa on subreddit data.

sentation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

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

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

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

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

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108, 2019. URL http://arxiv.org/abs/1910.01108.

V"olske, M., Potthast, M., Syed, S., and Stein, B. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 59–63, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4508. URL https://www.aclweb.org/anthology/W17-4508.

*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 |

*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 |

## A. Demo

See Figure 5 for a partial screen shot of a publicly available demo of our gender pronoun predicting task on both the fine-tuned and pre-trained models.

## 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. Covariate 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 3, Bottom 10 and top 10 Global Gender Gap ranked countries 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"

## D.2. Subreddit Values

Ordered list of subreddits used for the x-axis in Figure 4, 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"

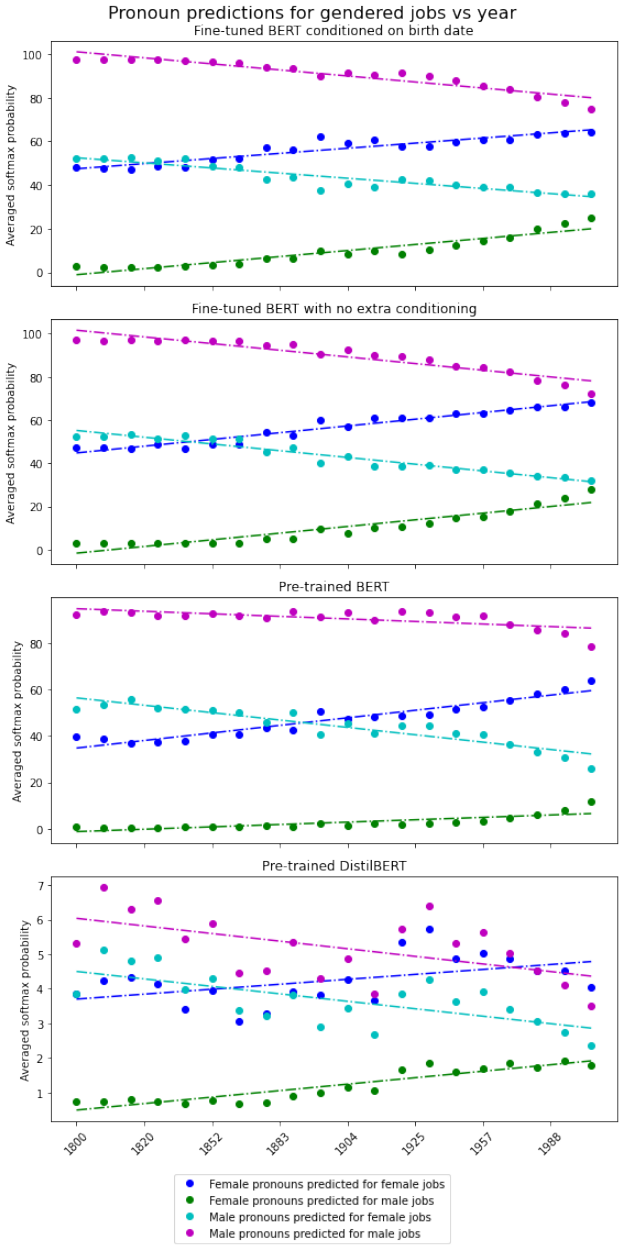


*Figure 2.* Averaged softmax percentages for gendered pronouns predicted to replace the mask in: "born in {date}, <mask> works as a {job}.", where `job` was filled in with either 'male-like' or 'female-like' jobs (see Table 2). The text filled in for {date}, and the x-axis, is a range of `date` values from 1800 to 1999.

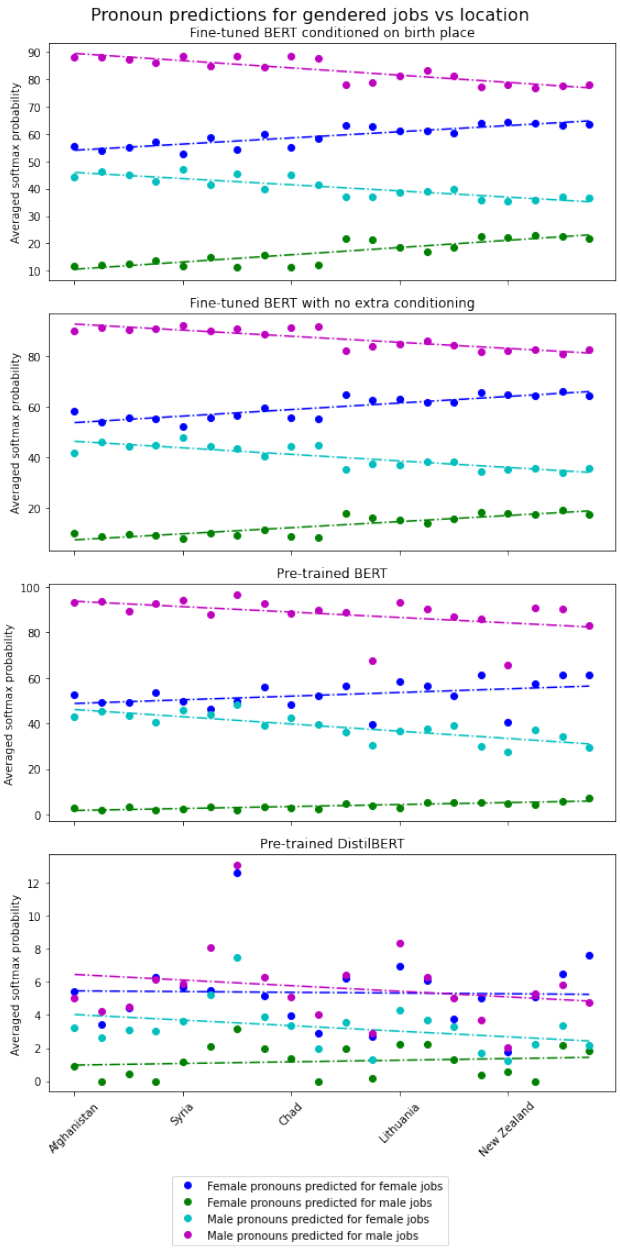


*Figure 3.* Averaged softmax percentages for gendered pronouns predicted to replace the mask in: "born in {place}, <mask> works as a {job}.", where `job` was filled in with either 'male-like' or 'female-like' jobs (see Table 2). The text filled in for {place}, and the x-axis, comes from a list of countries ranked by gender equality scores increasing to the right (see D.1).

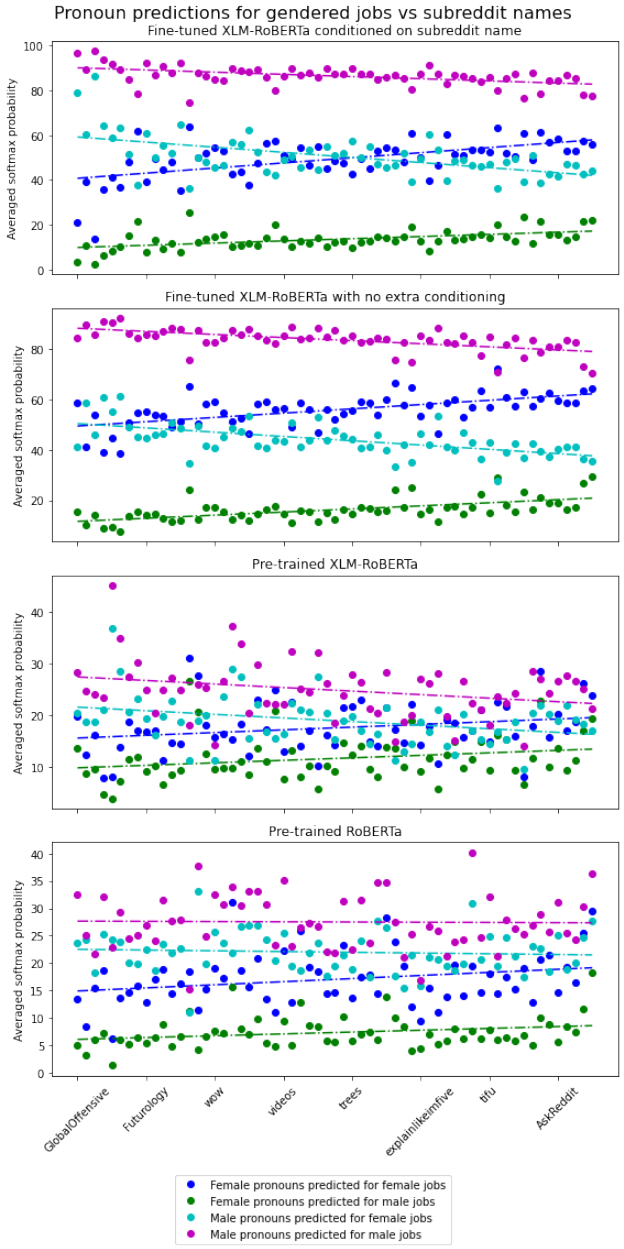


*Figure 4.* Averaged softmax percentages for gendered pronouns predicted to replace the mask in: "<mask> works as a {job}. Source: r/{subreddit}", where `job` was filled in with either 'male-like' or 'female-like' jobs (see Table 2). The text filled in for {subreddit}, and the x-axis, is a list of `subreddit` names, with percentage of self-reported female users per subreddit increasing to the right (see D.2).

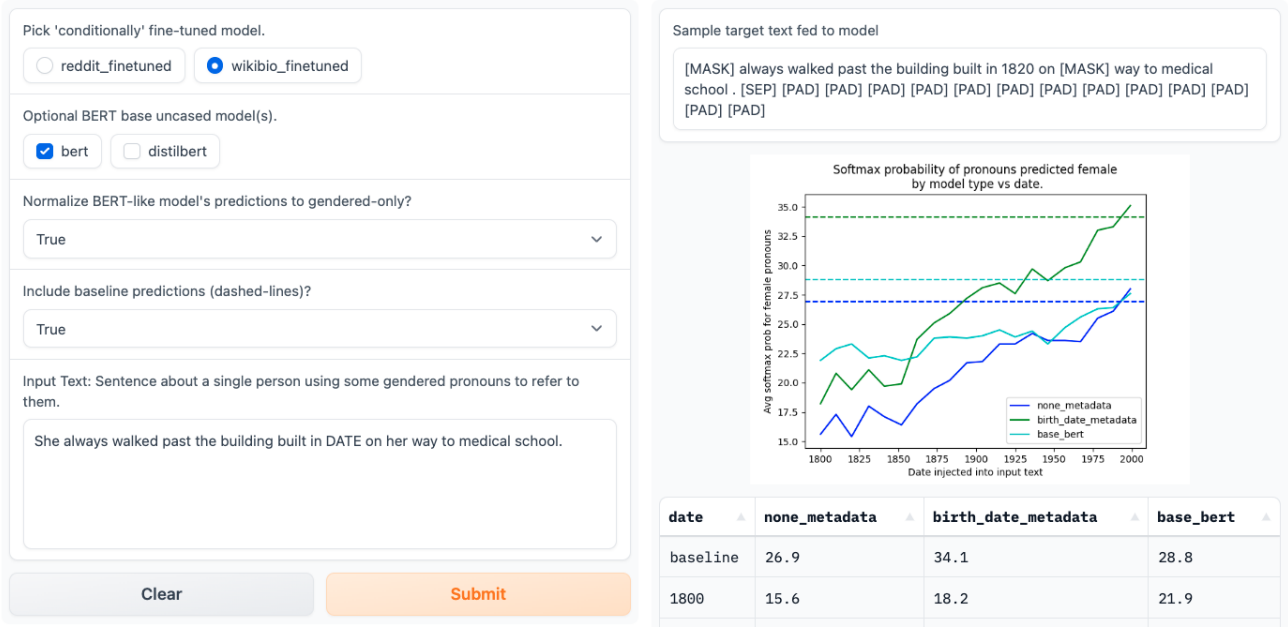


*Figure 5.* Publicly available demo of our gender pronoun predicting task on both the fine-tuned and pre-trained models. Here we see the presence of spurious correlation between predicted gender pronouns and the date in which a building was constructed.