SELECTION COLLIDER BIAS IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

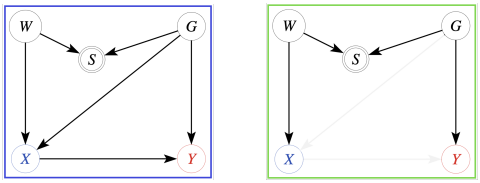
In this paper we motivate the causal mechanisms behind sample selection induced collider bias (selection collider bias) that can cause Large Language Models (LLMs) to learn unconditional dependence between entities that are unconditionally independent in the real world. We show that selection collider bias can become amplified in underspecified learning tasks, and although difficult to overcome, we describe a method to exploit the resulting spurious correlations for determination of when a model may be uncertain about its prediction. We demonstrate an uncertainty metric that matches human uncertainty in tasks with gender pronoun underspecification on an extended version of the Winogender Schemas evaluation set, and we provide online demos where users can evaluate spurious correlations and apply our uncertainty metric to their own texts and models. Finally, we generalize our approach to address a wider range of prediction tasks.

1 INTRODUCTION

This paper investigates models trained to estimate the conditional distribution: $P(Y|X, S)$, where S is the cause of sample selection bias in the training dataset. Selection bias is not an uncommon problem, as most datasets are subsampled representations of a larger population, yet few are sampled with randomization (Heckman, 1979).

Sample selection bias occurs when some mechanism, observed or not, causes samples to be included or excluded from the dataset. Employing the language of causal inference, selection bias is distinct from both confounder and collider bias. Confounder bias can occur when two variables have a common cause, whereas collider bias can occur when two variables have a common effect. Correcting for confounding bias requires that one condition upon the common cause variable; conversely correcting for collider bias requires that one does not condition upon the common effect (Pearl, 2009).

While sample selection bias can take many forms, the type of selection bias that interests us here is that which involves more than one variable (observed or not), whose common effect results in selection bias. Such relationships can be compactly represented as a causal model in the form of a directed acyclic graph (DAG), for example illustrated in Figure 1.



(a) Well-specified: G fully observed, in both dataset features and labels.

(b) Underspecified: Dataset features do not observe any causes of the labels.

Figure 1: Proposed data generating process for a range of NLP datasets for well-specified and underspecified prediction tasks. X and Y represent text: the dataset features and labels, while W, G, and S represent symbolic entities that may cause the text.

The absence of arrows connecting nodes in causal DAGs encode assumptions, for example that W and G in 1(a) are stochastically independent of one another. The direction of the arrowhead encodes our assumptions about the direction of causation. For example, two arrows departing from W and G toward S encode the assumption that S is a common effect of W and G.

In Figure 1, the twice-encircled node, S, symbolizes some mechanism that can cause samples to be selected into the dataset. To capture the statistical pro-

cess of sampling for dataset formation, one must condition on S, thus inducing the collider bias relationship between W and G in the causal model.

We will use the term *selection collider bias* to refer to circumstances such as this one, when the selection bias mechanism induces a collider bias relationship in the dataset, that would not have been there otherwise. Beyond posing a risk to out-of-domain generalizability (Arjovsky et al., 2019), selection collider bias can result in models that lack even 'internal validity', as the associations learned from the data represent the statistical dependencies induced by the dataset formation and not the data itself (Griffith et al., 2020).

2 UNDERSPECIFICATION

The link between underspecification in machine learning, and the resulting spurious correlations and risks to out-of-domain generalization is investigated thoroughly in (D'Amour et al., 2020). In this work, underspecification is defined in the context of having multiple predictors with equally low predictive risk (or some other risk minimization) scores. In the NLP setting, (Manning, 2011) provide an example of underspecification in a part-of-speech tagging task, in which a given sentence context makes it unclear whether the word, 'discontinued', should have the label 'adjective' or 'verbal participle'.

In this paper, we are interested in the underspecification that occurs when none of the features available to the model (at training or inference time) are causes of the label. With no causal features available, models must resort to learning any spurious associations that will reduce predictive risk, regardless of how tenuous the association may be. Thus, models trained with selection collider biased data are particularly vulnerable to learning statistical associations in underspecified learning tasks.

A single prediction task may be partitioned into well-specified and underspecified 'sub'-tasks. In the example from (Manning, 2011), we can assume a well-specified 'sub'-task reduced the set of all possible prediction classes to the set containing just 'adjective' and 'verbal participle', while the underspecified 'sub'-task must then determine which of the two remaining classes is correct. For much of this paper we will consider a similar scenario: where the well-specified portion of the masked language modeling (MLM) task may have correctly predicted that a gender pronoun should fill-in a masked-out word, however the 'sub'-task of determining which gender pronoun to predict is underspecified.

3 PLAUSIBLE CAUSAL MODELS

Figure 1 shows plausible data-generating processes for the relevant features and labels in the above-mentioned task of gender pronoun resolution.

In Figure 1(a), the symbol X represents *context* sentences in our dataset, and Y represents the label, specifically a *gender pronoun*. The arrow pointing from X to Y encodes our assumption that X, is more likely to cause Y, rather than vice versa. For example, in a sentence about a father and daughter going to the park, the sentence context determines which gender pronoun is most appropriate.

The remaining symbols in Figure 1(a) are not recorded in the dataset, but rather represent entities or mechanisms that may cause X or Y. The symbol G represents *gender* and causes both the *context* and *gender pronoun*. W represents various gender-neutral entities, such as *time*, *place*, or *gender-neutral hobbies*, and is included in the causal DAG because it may be a cause of both X and S. Finally, the selection node, S, takes the value of $S=1$ for samples in the data and $S=0$ otherwise.

In Figure 1, S is the common effect of W and G, and represents *access* to the dataset. The $W \rightarrow S \leftarrow G$ relationship can represent any selection bias mechanism that has a gender dependency upon otherwise gender-neutral entities. For example, in data sources like Wikipedia written about people, it is plausible that *access* has become increasingly less gender dependent as the *date* approaches more modern times, but not evenly in every *place*. In data sources like Reddit written by people, $W \rightarrow S \leftarrow G$ can plausibly capture the scenario that even in the case of subreddits about *gender-neutral hobbies*, the style of the moderation and community may result in gender-disparate *access* to a given subreddit.

Table 1: Heuristic for creating gender-neutral input texts for the MGT evaluation set for 2 W categories, and the example rendered text. For `verb` we used past, present and future tenses of the verb *to be*: ['was', 'became', 'is', 'will be', 'becomes'], 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'], excluding life stages past adulthood due to age gender-inequities. For the text versions of `w`, we used a spectrum of values defined in Appendix B.

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 .'

Now, in Figure 1(b), we can see the underspecified counterpart of the well-specified prediction task in Figure 1(a). As defined above, to achieve the underspecification of interest here, we must obscure any causal features from X. In the case of gender pronoun resolution, this is achieved by removing the path between *gender* and *context*. Further, because W is also gender-neutral, once we have removed any gender-identifying features from *context*, we should additionally remove the path between *context* and *gender-pronouns*, as there is no longer any feature of X causing Y.

With the causal path from $X \rightarrow Y$ removed, had there been no backdoor path from X to Y, then we would then expect X and Y to be unconditionally independent in the real world (RW). However, in the scenario depicted in Figure 1(b), selection collider bias has opened the backdoor path, $X \leftarrow W \rightarrow S \leftarrow G \rightarrow Y$, Thus, the learned model (LM) is encouraged to learn any spurious association along this path that reduces predictive risk, resulting in X and Y becoming unconditionally dependent. Formally, selection collider bias causes the transformation: $(Y \perp\!\!\!\perp X)_{RW} \overset{s}{\Rightarrow} (Y \not\perp\!\!\!\perp X)_{LM}$, due to even very weakly interacting spurious associations, such as those between *gender* and *time*. Whereas spurious associations learned in the well-specified model would be likely due to signals that are stronger than those along the causal path $X \rightarrow Y$ [1].

3.1 Masked Gender Task

Beyond just claiming plausibility, we can evaluate measured data from LLMs to support our arguments for the causal relationships in Figure 1, after making two additional assumptions: For a given causal graph, Gr, and a measured probability distribution Pr, we are assuming 1) the Markov assumption: $(X \perp\!\!\!\perp Y|Z)_{Gr} \Longrightarrow (X \perp\!\!\!\perp Y|Z)_{Pr}$ and 2) faithfulness: $(X \perp\!\!\!\perp Y|Z)_{Gr} \Longleftarrow (X \perp\!\!\!\perp Y|Z)_{Pr}$, where Z is the set $\{W, G, S\}$ (Peters et al., 2017).

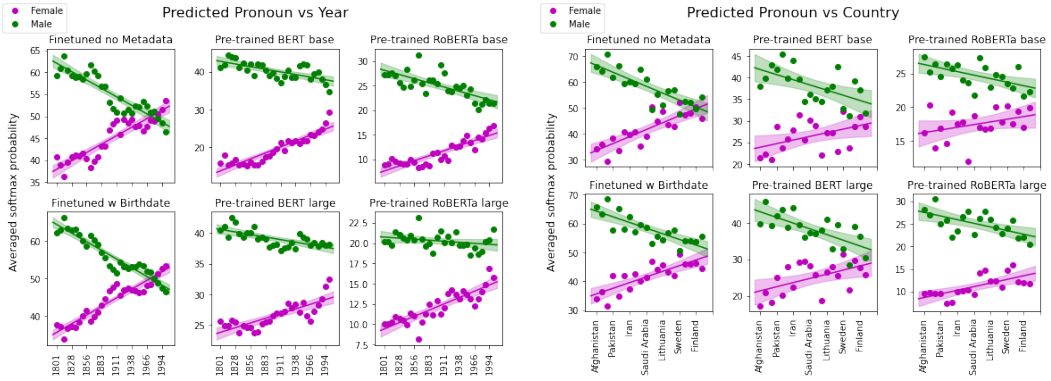
To obtain the measured probability distribution Pr, we developed the masked gender task (MGT), a special case of the Masked Language Modeling (MLM) objective, that uses a heuristic to build underspecified learning tasks by masking common gender-identifying words for prediction (see Appendix A), which we used to finetune two models. We use these finetuned models as comparison baselines against unmodified pre-trained BERT base and large (Devlin et al., 2018) and RoBERTa base and large (Liu et al., 2019) models, tested using our MGT evaluation set described in Table 1.

We argue it is not implausible that during MLM pre-training, gender-identifying words are masked for prediction in otherwise gender-neutral sequences. Further, at inference time, the prediction of gender-identifying words or labels from gender-neutral text is common to many downstream tasks such as dialog generation, machine translation from genderless to more gendered languages, or any task requiring gendered predictions from gender-underspecified features.

4 Selection Collider Bias

Before sharing our LLM measurements, we further motivate the mechanisms behind selection collider bias. Revisiting Figure 1(a), to estimate the causal effect of X on Y here, we could the use back-door adjustment (Pearl, 2009), with an admissible set $\{G\}$ to calculate: $P(Y|do(X)) = \sum_G P(Y|X, G)\, P(G)$. The observation of G in the dataset makes this a trivial problem to solve.

[1] One example of spurious associations learned despite a well-specified task is that of an image background (desert vs grass) providing a stronger signal than the animal foreground (camel vs cow) (Arjovsky et al., 2019).



(a) Spurious correlation between *gender* and *time* plotted as averaged softmax probabilities for predicted gender pronouns vs a range of dates.

(b) Spurious correlation between *gender* and *place* as list of countries, ordered by their Global Gender Gap rank (see Appendix B.1).

Figure 2: Softmax probabilities for LLM predictions displaying spurious correlation between *gender* and gender-neutral entities injected into gender-neutral input texts described in Table 1

Although the S node is not explicitly recorded in the dataset, even if S were available to us, we would not condition on it, as this would induce collider bias between G and W in the form of S's structural equation (Pearl, 2009): $S := f_s(W, G, U_s)$, where U_s is the exogenous noise of the S variable.

As mentioned above, to capture the statistical process of dataset formation, we implicitly condition on $S = 1$ for all the samples in the dataset. Now applying the above-mentioned Markov and faithfulness assumptions to the underspecified model in Figure 1(b), we can estimate the conditional probability of a gender-identifying word, Y, given gender-neutral text, X, in a LLM as shown in Equation (1) through Equation (4).

Equation (1) shows a mapping from the target unbiased quantity to the measured selection biased data, as defined in (Bareinboim and Pearl, 2012). Equation (2) assumes very high correlation between the textual form of gender in Y (as a 'gender-identifying word') with the symbolic variable for gender, G. Equation (3) replaces S with the variables in its structural equation, $S := f_s(W, G, U_s)$, which entails the conditional dependence $P(G|W) \neq P(G)$. Finally, Equation (4) assumes we have successfully achieved a gender-neutral text, X, and thus $P(G|X) = P(G)$.

$$
\begin{aligned}
P(Y|X) &= P(Y|X, S{=}1) && (1)\\
&\sim P(G|X, S{=}1) && (2)\\
&\sim P(G|X, W) && (3)\\
&\sim P(G|W) && (4)
\end{aligned}
$$

Equation (1) - Equation (4) show that conditional probability $P(Y|X)$ obtained by applying the underspecified task in Figure 1(b) to an LLM, should be distributed similarly to $P(G|W)$. This provides information about the latent representations for G and W, which are otherwise inaccessible. Additionally, these equations provide insights into the role of selection collider bias in the transformation from real-world (RW) unconditional independence to learned model (LM) unconditional dependence discussed in Section 3: $(Y \perp\!\!\!\perp X)_{RW} \approx (G \perp\!\!\!\perp W)_{RW} \overset{s}{\Rightarrow} (G \not\perp\!\!\!\perp W)_{LM} \approx (Y \not\perp\!\!\!\perp X)_{LM}$.

5 Pre-trained LLMs G, W Dependency

Figure 2(a) shows pre-trained BERT and RoBERTa base and large results, as well as results for models finetuned with the MGT objective[2]. Each plotted dot is the softmax probability (averaged over 60 gender-neutral texts) for the predicted gender pronouns vs *year*, where *year* in the x-axis matches that gender-neutral W value injected into the gender-neutral text[3].

[2]The finetuned models serve as a rough upper limit for the magnitude of expected spurious correlations.

[3]For example, the purple and green dots at the x-axis position of $W = 1938$ are the softmax predictions for the masked word in input texts like 'In 1938, [MASK] will became a teenager.', for female and male pronouns.

The shaded regions show the 95% confidence interval for the linear fit. Unlike the finetuned model's binary prediction, because the final layer in the pre-trained model is a softmax over the entire tokenizer's vocabulary, we sum the gender-identified portion (as listed in Table 3) of the probability mass from the top five predicted words[4].

We argue the spurious associations in Figure 2, between W along the x-axis and predicted gender, G, along the y-axis, supports our supports our hypothesis that selection collider bias has resulted in these LLMs learning $P(Y|X) \sim P(G|W)$, for entities G and W that would be otherwise unconditionally independent.

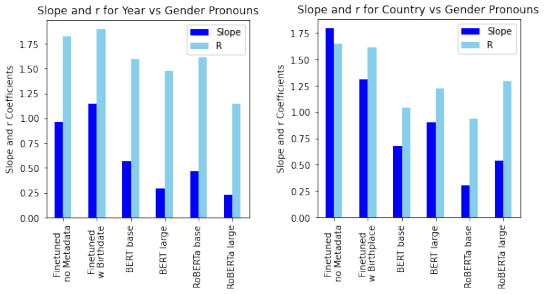


Figure 3: Difference between the female and male pronoun linear fit slope and Pearson's r coefficients from the plots in Figure 2.

Figure 3 shows the slope and Pearson's r correlation coefficient (following (Rudinger et al., 2018)) of the y-axis value against the *index* of the x-axis, for all the plots in Figure 2. As expected, we see the slope and correlation coefficients highest in the finetuned models. We do nonetheless see comparable coefficients for the spurious W, G dependency in the unmodified pre-trained models, supporting our hypothesis that these models are learning from underspecified tasks such as that in Figure 1(b).

6 Selection Bias Recovery

In (Bareinboim et al., 2014) it is proven that one can recover the unbiased conditional distribution $P(Y|X)$ from a causal DAG, G_S, with selection bias: $P(Y|X, S=1)$, if and only if the selection mechanism can become conditionally independent of the effect, given the cause: $(Y \perp\!\!\!\perp S|X)_{G_S}$. However, in Figure 1(b) we can see $(Y \not\perp\!\!\!\perp S|X)$ trivially, due to the lack of a direct path between X and Y. Thus, the conditional distribution learned by the underspecified model will not converge toward the unbiased distribution without additional data or assumptions (Bareinboim et al., 2014). However, in the next section we will exploit the outcomes of selection bias to our benefit.

7 Exploiting Spurious Correlations

As discussed in Section 3, underspecified models are more vulnerable to learning spurious associations when trained with selection collider biased data. In this section we turn the problem on its head and use spurious associations to determine if a prediction task may be underspecified, and therefore if the resulting prediction should be deemed uncertain, in a method we call W-injection.

We test this using the Winogender Schema evaluation set (Rudinger et al., 2018), composed of 120 sentence templates, hand-written in the style of the Winograd Schemas (Levesque et al., 2012). Originally the Winogender evaluation set was developed to demonstrate that many NLP pipelines produce spurious associations between gender and occupation, often well exceeding any occupation-based gender inequality in the real world.

7.1 Winogender Texts

The 'Sentence' column in Table 2 shows example texts from our extended version of the Winogender evaluation set, where the occupation is 'doctor'. Each sentence in the evaluation set contains: 1) a *professional*, referred to by their profession, such as 'doctor' 2) a context appropriate *participant*, referred by one of: {'man', 'woman', 'someone', *other*} where *other* is replaced by a context specific term like 'patient', and 3) a single pronoun that is either coreferent with (1) the *professional* or

[4]We pick the number $k = 5$ for the 'top_k' predicted words, because 5 is the default value for the 'top_k' argument in the Hugging Face 'fillmask()' function used for inference.

Table 2: Extended Winogender evaluation sentences and uncertainty results for occupation as 'Doctor'. Only sentence IDs 5 & 6 are not underspecified for gender pronoun prediction, and appropriately all models show the lowest relative uncertainty for these sentences, in particular the larger and longer trained models.

ID	Sentence with Participant and **Coreferent** Highlighted	Uncertainty			
		BERT		RoBERTa	
		base	large	base	large
1	In DATE: The **doctor** told the man that [MASK] would be on vacation next week.	1.7	1.8	15.0	14.0
2	In DATE: The **doctor** told the woman that [MASK] would be on vacation next week.	4.3	27.3	4.0	18.8
3	In DATE: The **doctor** told someone that [MASK] would be on vacation next week.	10.6	8.0	13.3	20.2
4	In DATE: The **doctor** told the patient that [MASK] would be on vacation next week.	1.9	6.6	14.7	16.6
5	In DATE: The doctor told the **man** that [MASK] would be at risk without the vaccination.	**0.0**	**0.0**	**0.3**	**0.1**
6	In DATE: The doctor told the **woman** that [MASK] would be at risk without the vaccination.	**0.3**	**0.1**	**0.7**	**0.5**
7	In DATE: The doctor told **someone** that [MASK] would be at risk without the vaccination.	11.3	10.5	41.3	16.4
8	In DATE: The doctor told the **patient** that [MASK] would be at risk without the vaccination.	6.1	12.3	19.2	9.3

(2) the *participant* in the sentence (Rudinger et al., 2018). For the masked gender task, this pronoun is replaced with a [MASK] for prediction.

Our extensions to the evaluation set are twofold: 1) we add {'man', 'woman'} to the list of words used to describe the *participant* in order to add well-specified tasks to the existing Winogender set, which are all underspecified, and 2) we perform *W*-injection by prepending each sentence with the phrase 'In DATE'[5], where 'DATE' is replaced by a range of years from 1901 to 2016[6], similar what was done in Figure 2(a).

In Sentence IDs 1 - 4 of Table 2, the masked pronoun is coreferent with the *professional*, who is always referred to as the 'doctor'. Whereas in Sentence IDs 5 - 8, the masked pronoun is coreferent with the *participant*, who is referred to as {'man', 'woman', 'someone', and 'patient'}, respectively. Of the eight sentences, six remain underspecified for the pronoun prediction task, with only IDs 5 & 6 becoming well-specified.

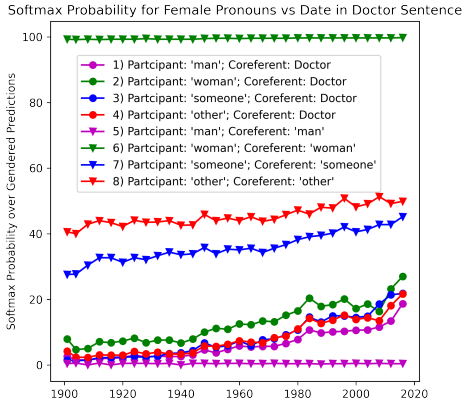


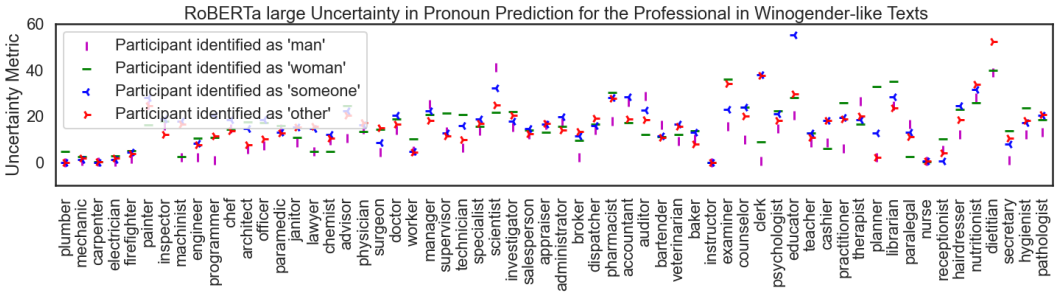
Figure 4: Averaged softmax percentages from RoBERTa large for predicted female gender pronouns (normalized over all gendered predictions) vs a range of dates (injected into the text), for the 'Doctor' Winogender texts listed in Table 2.

Figure 4 shows the predicted softmax probability for female pronouns for the masked words in the Table 2 sentences, normalized to the gendered predictions of the top five predicted words from pre-trained RoBERTa large. Similar to the findings in (Rudinger et al., 2018), the softmax probabilities for female pronouns are higher for masked pronouns coreferent with the patient as opposed to the doctor (for the underspecified sentences) indicating a specific gender bias for traditionally non-female occupations.

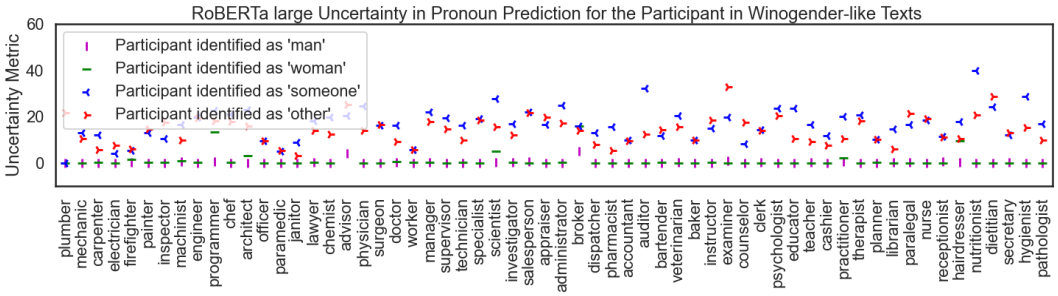
What is new here is that in Figure 4 we can see that the spurious associations due to the *W*-injection of an unrelated association (time vs gender) is additive with the existing spurious association between occupation and gender, but only takes on a non-zero value when the prediction task is underspecified.

[5]Similar results were obtained by using 'In PLACE' and replacing 'PLACE' with countries in Appendix B.1.

[6]We picked a slightly narrower and more modern date range as compared to that of Figure 2(a) for contextual consistency with some of the more modern occupations in the Winogender evaluation set.



(a) Masked pronoun is coreferent with the *professional* in the sentence, so all these sentences remain underspecified. Like human uncertainty, we see model uncertainty results above 0 for most occupations, regardless of the word injected into evaluation text for the *participant*, including co-occurring gender-identifying terms.



(b) Masked pronoun is coreferent with the *participant*, so the sentences containing 'man' and 'woman' become well-specified, while the rest remain underspecified. Like human uncertainty, we do see uncertainty results close to 0 for most occupations, when 'man' or 'woman' has been injected into the evaluation text for the *participant*, and generally above 0 otherwise

Figure 5: RoBERTa-large uncertainty results on all Winogender Schema occupations.

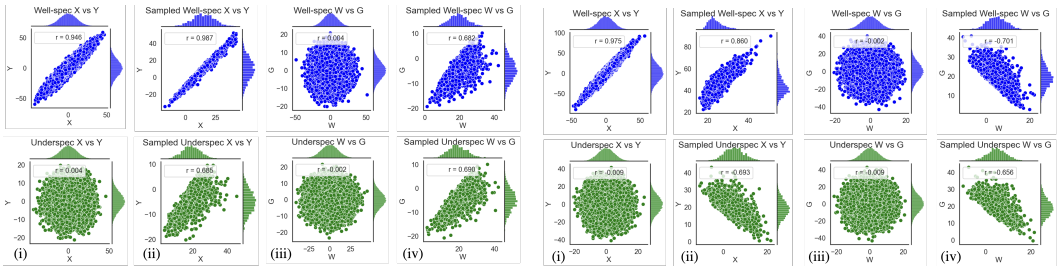
7.2 Uncertainty Metric

For an example single-value uncertainty metric, we can measure the absolute difference between the averaged softmax probabilities for the first and last several dates along the x-axis in Figure 4. For this uncertainty metric, we would expect larger values for underspecified prediction tasks, in which the spurious correlation between *gender* and *date* has a larger role in influencing the prediction. For the predictions in Figure 4, this metric is shown in the 'Uncertainty' columns in Table 2 for all four LLMs studied in this paper. Here we see uncertainty values closest to 0 for well-specified sentence IDs 5 & 6, consistent with human reasoning about the uncertainty of predicting gender pronouns for these sentences in Table 2[7].

Our extended version of the Winogender Schema contains 60 occupations for the *professional* $\times$ 4 words used to describe the *participant* $\times$ 30 values for DATE $\times$ 2 sentence templates (one in which the masked pronoun is coreferent with the *professional* and the other with the *participant*). This totals to 14,400 test sentences, which we provide as input text to the 4 pre-trained models thus far investigated in this paper: BERT base and large, and RoBERTa base and large.

We calculate the above-described uncertainty metric for all 60 occupations in the Winogender evaluation set and show the results from RoBERTa large 1) in Figure 5(a), with input sentences like IDs 1 - 4 where the masked pronoun is coreferent with the *professional*, and 2) in Figure 5(b), with input sentences like IDs 5 - 8 where the masked pronoun is coreferent with the *participant*. In these plots the x-axis is ordered from lower to higher female representation, according to Bureau of Labor Statistics 2015/16 statistics provided by (Rudinger et al., 2018), and the y-axis is the prediction uncertainty metric defined in the proceeding paragraphs.

[7] As can be seen further in Appendix C, this uncertainty metric appears to report results more consistent with human reasoning in RoBERTa large and generally as the model becomes increasingly over-parameterized.



(a) $(W - G + \mathcal{N}(0,1)) > 2\alpha$ sampled.

(b) $(W + G + \mathcal{N}(0,1)) > 2\alpha$ sampled.

Figure 6: Statistical relationships induced by the structural causal model defined by equations Equation (5) to Equation (9), with Equation (7) individually defined above, for the well-specified (top row in blue) and underspecified (bottom row in green) models in Figure 1.

In Figure 5, we again see the model reliably reporting high uncertainty for all six of the underspecified tasks and low uncertainty for the two well-specified tasks, for almost all Winogender evaluation sentences. In particular, the injection of gender-identifying text: 'man' and 'woman' into the sentence, reduces the model's uncertainty only when these gender-identifying terms are coreferent with the masked pronoun for prediction as in Figure 5(b), and not when gender-identifying terms are merely co-occurring as in Figure 5(a). We show similar results for BERT and RoBERTa base and BERT large in Appendix C, but note that the increased parameter count and hyper-parameter optimization in RoBERTa large appears to improve the uncertainty measurement.

8 Extending to More General Setting

We now explore a more general problem space where these symbols in Figure 1 take on the following meanings: G is the causal parents of Y, W is the non-causal parents of Y yet nonetheless included because W is a cause of both X and S, where S has the same meaning as before. We can thus partition any feature space into G, and candidates for W. A candidate can be validated as suitable W feature by checking for the conditional dependencies which we will be plotting below.

To make this hypothetical example slightly more concrete, we parameterize the structural causal models associated with Figure 1 in these slightly less generic terms shown on the right.

$$G := \alpha\mathcal{N}(0,1) \tag{5}$$

$$W := \frac{\alpha}{2}\mathcal{N}(0,1) \tag{6}$$

$$S := (W - G + \mathcal{N}(0,1)) > 2\alpha \tag{7}$$

$$X := W + \gamma G + \mathcal{N}(0,1) \tag{8}$$

$$Y := \gamma X + G + \mathcal{N}(0,1) \tag{9}$$

Equation (5) and Equation (6) define W and G as independent exogenous 0-mean Gaussian noise, for which we set $\alpha = 10$ so that we can more easily trace the amplified noise through the DAG[8]. Equation (7) defines S as an unweighted combination of W, G and exogenous noise, with the selection mechanism setting all values above 2α to 1, and to 0 otherwise. In Equation (8) and Equation (9) we set γ to 0 for the underspecified task and to 1 for the well-specified task, consistent with a 0 path weight for the grayed out arrows $G \to X$ and $X \to Y$ in Figure 1(b), and a full path weight for those same arrows in Figure 1(a).

Figure 6 plots the statistical relationships induced by the structural causal model defined by equations Equation (5) to Equation (9) for the well-specified (top row in blue) and underspecified (bottom row in green) causal models. Starting with Figure 6(a), columns (i) and (ii) show plots X vs Y for both the unsampled and the $S = 1$ sampled distributions, respectively. The plotted correlation between X and Y for the well-specified (top two) plots is not greatly affected by the sampling, with the Pearson's r coefficient going from 0.946 to 0.987. However, for underspecified plots (bottom two), sampling selection bias causes the Pearson's r coefficient to go from 0.004 to

[8]We set different noise weights to G and W by arbitrarily dividing α by 2 in Equation (6), to reduce the likelihood of unintentionally constructing an unfaithful graph.

0.685, consistent with the selection collider bias induced transformation discussed in Section 3: $(Y \perp\!\!\!\perp X)_{RW} \overset{S}{\Rightarrow} (Y \not\perp\!\!\!\perp X)_M$.

In Figure 6(a), columns (iii) and (iv) also show the unsampled and the $S = 1$ sampled distributions, but for W vs G. Here we also see that the S node subsampling causes W vs G to go from about 0.0 to about 0.7, but for both the underspecified model and well-specified models, rather than the underspecified model alone. This validates that selection collider bias induced $P(Y|X) \sim P(G|W)$ only holds underspecified models.

For Figure 6(b), we replace Equation (7) with $S := (W + G + \mathcal{N}(0,1)) > 2\alpha$, which differs in that we are combining W and G with an addition, rather than a subtraction. The results here are similar to those in Figure 6(a), with the only exception being that due to the updated structural equation for S, the direction of the correlation coefficient has flipped in both selection biased W vs G plots, and only the underspecified X vs Y plot. These results again validate $P(Y|X) \sim P(G|W)$ for only the underspecified model.

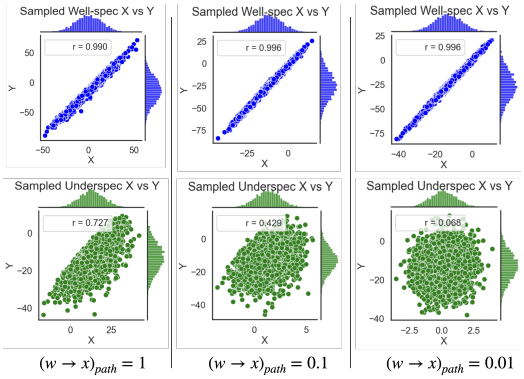


Figure 7: Decreasing the path coefficient, β, for $X \to Y$, as a toy representation of what we termed the W-Injection.

For the toy demonstration of W-injection in Figure 7, we replace Equation (8) with $X := \beta W + \gamma G + \mathcal{N}(0,1)$, where β takes on the values: 1) $\beta = 1$ (thus identical to Equation (8)) 2) $\beta = 0.1$ and 3) $\beta = 0.01$. Similar to the results in Section 7, probing $P(Y|X)$ for a range of W values can serve to classify as tasks as underspecified if $P(Y|X)$ is sensitive to W-injection, or well-specified if $P(Y|X)$ is unaffected.

9 Demonstration and Open-Source Code

We have developed demos where users can choose their own input text and select almost any BERT-like model hosted on Hugging Face to test for selection collider bias induced spurious correlations and model uncertainty, shown in Appendix D and Appendix E. We will make all code available on github.

10 Discussion

In this paper we have argued that underspecified prediction tasks leave models vulnerable to selection collider bias induced spurious associations, including weakly associating and previously unreported correlations between gender vs time and gender vs place, which we demonstrate on unmodified pre-trained BERT-like models.

We have introduced a technique for injecting spurious signals into inference tasks to determine if the task is well-specified or underspecified, and demonstrated this in the form of an uncertainty metric on an established evaluation set.

Finally, we have generalized our approach to address a wider range of prediction tasks and shown that our empirical results measured from LLMs can be demonstrated with toy data as well.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019. URL `https://arxiv.org/abs/1907.02893`.

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 Re-*

search, pages 100–108, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL https://proceedings.mlr.press/v22/bareinboim12.html.

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.

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.

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.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. The winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press, 2012. ISBN 9781577355601.

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.

Christopher D. Manning. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? In *Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing - Volume Part I*, CICLing'11, page 171–189, Berlin, Heidelberg, 2011. Springer-Verlag. ISBN 9783642193996.

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

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

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.

A Gender-identifying Words

See Table 3 for the list of gender-identifying words that were masked for prediction during both fine-tuning and at inference time for the Masked Gender Task, with the exclusion of 'man' & 'woman' that remained unmasked in the extended Winogender evaluation set.

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. 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

B W VARIABLE X-AXIS VALUES

For {w} we require a list of values that are gender-neutral in the real world, yet due to selection collider bias are hypothesized to be a spectrum of gender-inequitable values in the dataset. For W as *date*, we 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 - 2001. For W as *place*, we use the bottom and top 10 World Economic Forum Global Gender Gap ranked countries (see details in B.1).

B.1 PLACE VALUES

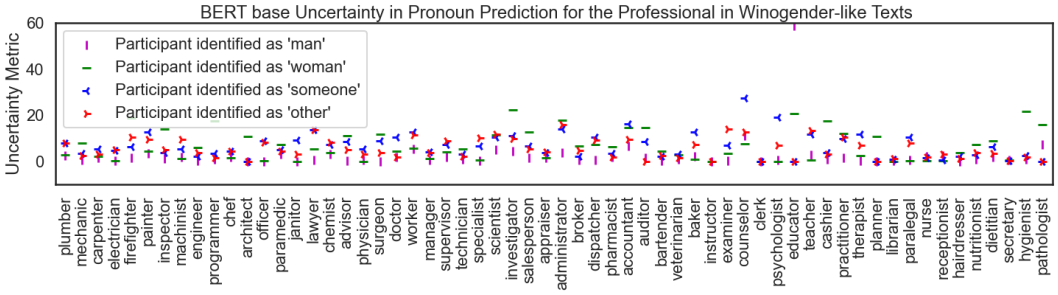
Ordered list of bottom 10 and top 10 World Economic Forum Global Gender Gap ranked countries used for the x-axis in Figure 2(b), that were taken directly without modification from `https://www3.weforum.org/docs/WEF_GGGR_2021.pdf`:

"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"

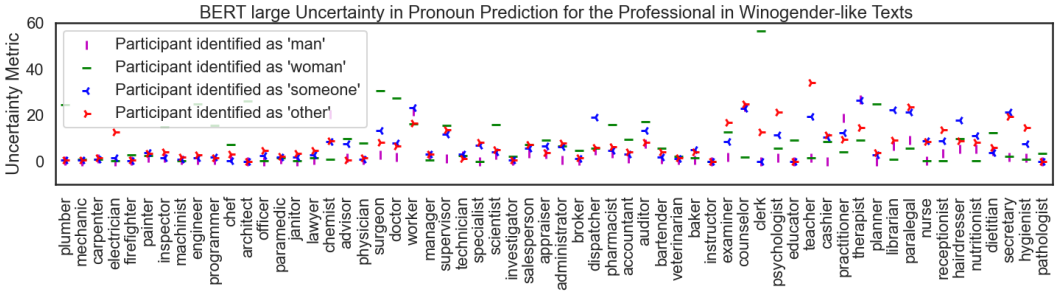
C EXTENDED WINOGENDER UNCERTAINTY RESULTS ON MORE LLMS

Figure 8 shows Uncertainty results for all Winogender occupations where the masked pronoun is coreferent with the *professional*. Because the injected text (one of: {'man', 'woman', 'someone', 'other'}) is referring to the *participant* and not the *professional*, all these sentences remain underspecified. The plots show all tested models tend to report uncertainty results above 0 for all occupations, regardless of the word injected into the evaluation text for the *participant*, thus the models do not become erroneously certain about gender when the words 'man' and 'woman' are injected into the text.

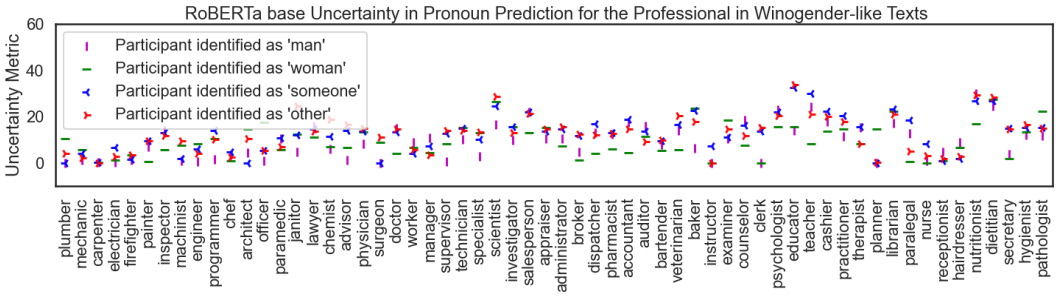
Figure 9 shows uncertainty results for all Winogender occupations where the masked pronoun is coreferent with the *participant*, unlike Figure 8 where the pronoun is coreferent with the *professional*. Because the injected text (again one of: {'man', 'woman', 'someone', 'other'}) is referring to the *participant*, the sentences containing 'man' and 'woman' become well-specified, while the rest remain underspecified. We see uncertainty results closer to 0 for most occupations when 'man' or 'woman' has been injected into the evaluation text for the *participant*, and generally above 0 otherwise, in particular for more highly over-parameterized models like BERT large and RoBERTA base & large in Figure 5(b).



(a) BERT base

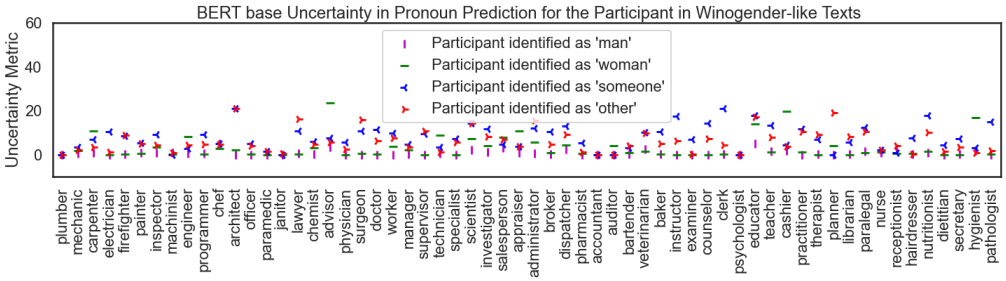


(b) BERT large

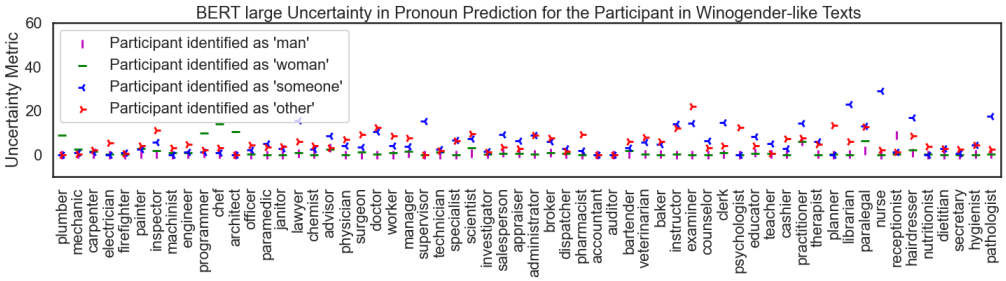


(c) RoBERTa base

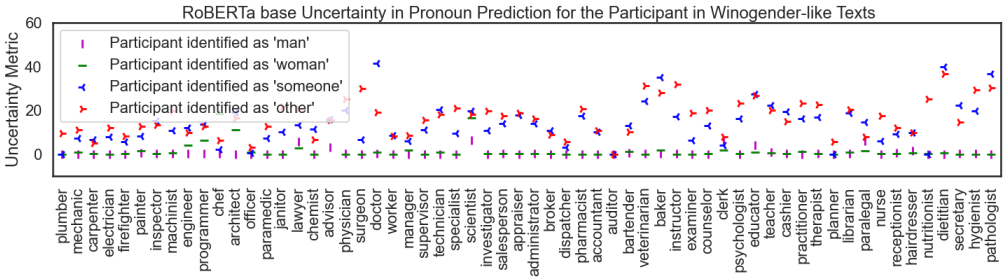
Figure 8: Uncertainty results for all Winogender occupations where the masked pronoun is coreferent with the gender-unidentified *professional*, thus all sentences remain underspecified. The plots show that generally, the models do not become erroneously certain about gender when the words 'man' and 'woman' are injected into the text.



(a) BERT base



(b) BERT large



(c) RoBERTa base

Figure 9: Uncertainty results for all Winogender occupations where the masked pronoun is coreferent with the *participant*, thus the sentences containing 'man' and 'woman' become well-specified, while the rest remain underspecified. Accordingly, the plots show that the uncertainty metric for the models is closer to 0 for the well-specified sentences containing 'man' and 'woman', and higher than 0 otherwise, particularly in the case of the more highly over-parameterized models like BERT large and RoBERTA base & large in Figure 5(b).

D SPURIOUS CORRELATIONS DEMO

See Figure 10 for our open source freely available demonstration where users can choose their own input text and select almost any BERT-like model hosted on Hugging Face to test for selection collider bias induced spurious correlations.

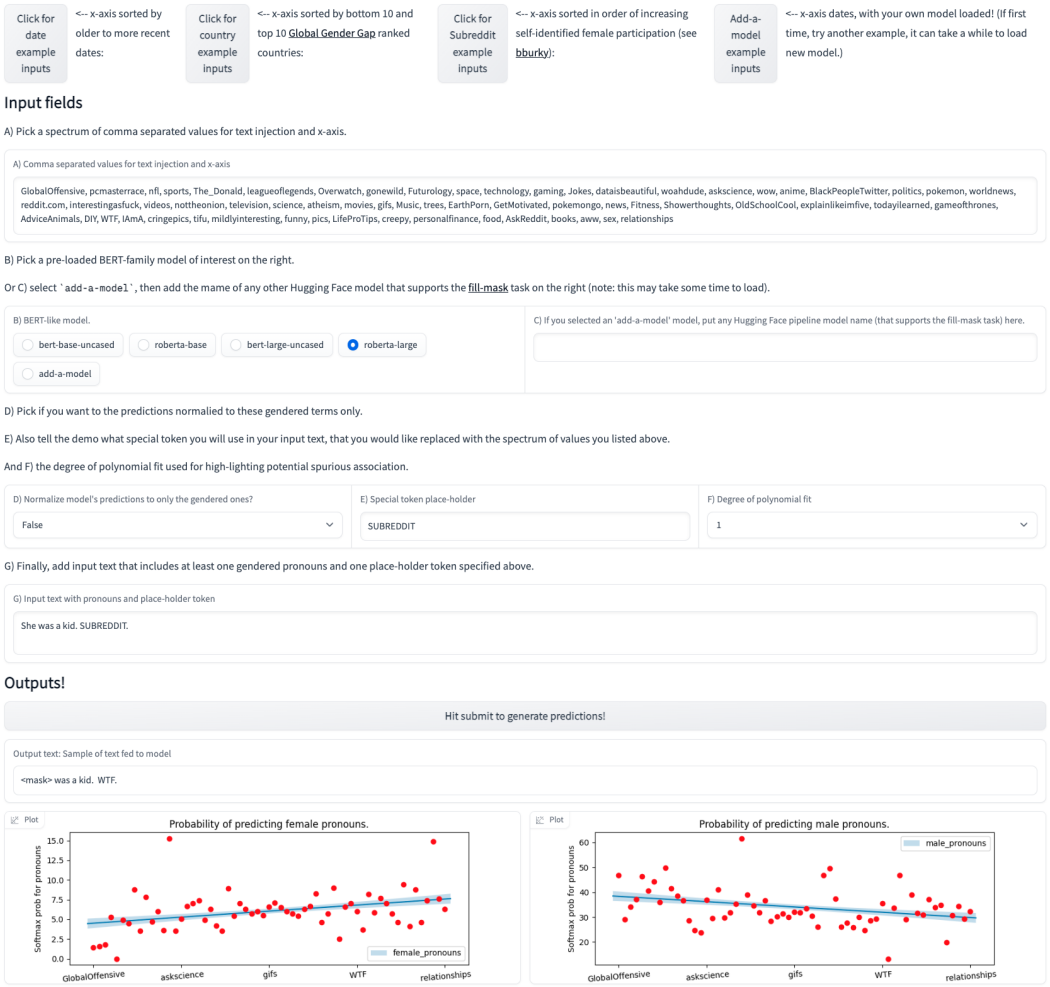


Figure 10: Demo where users can choose their own input text and select almost any BERT-like model hosted on Hugging Face to test for selection collider bias induced spurious correlations.

E MODEL UNCERTAINTY DEMO

See Figure 11 for our open source freely available demonstration where users can choose their own input text and select almost any BERT-like model hosted on Hugging Face to test for model uncertainty using selection collider bias induced spurious correlations.

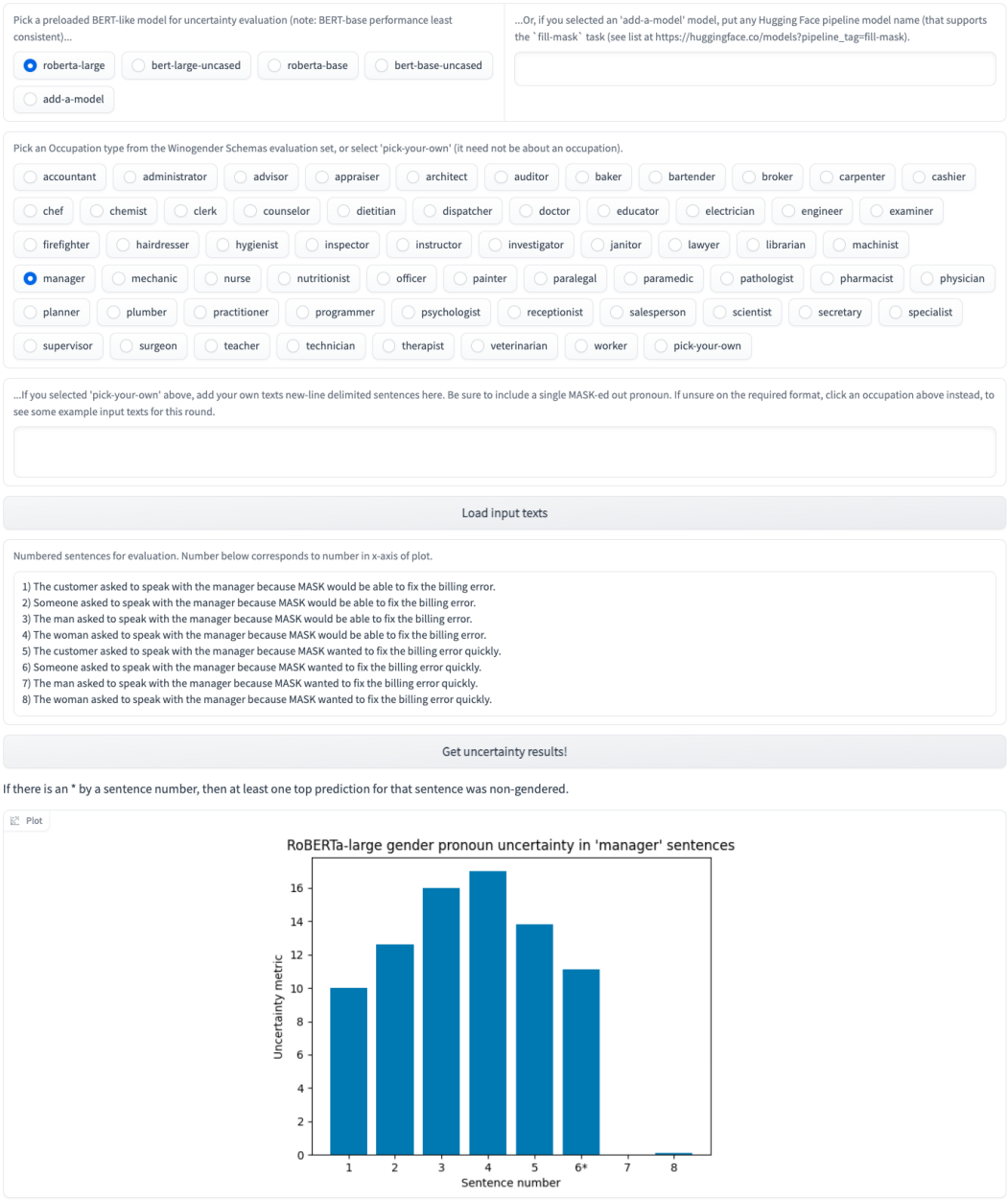


Figure 11: Demo where users can choose their own input text and select almost any BERT-like model hosted on Hugging Face to test for model uncertainty using selection collider bias induced spurious correlations.