SELECTION INDUCED COLLIDER BIAS IN LLMS: A GENDER PRONOUN UNCERTAINTY CASE STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we cast the problem of task underspecification in causal terms, and develop a method for empirical measurement of spurious associations between gender and gender-neutral entities for unmodified LLMs, detecting previously unreported spurious correlations. We then describe a method to exploit the resulting spurious associations for determination of when a model may be uncertain about its prediction, which can correctly classify task underspecification on over 90% of the Winogender Schemas evaluation set. Finally, we generalize our approach to address a wider range of prediction tasks.

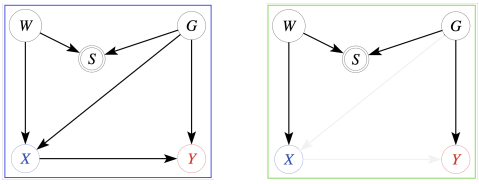
1 INTRODUCTION AND RELATED WORK

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).

1.1 CAUSAL DAGS AND BIASES

Sample selection bias occurs when some mechanism, observed or not, causes preferential inclusion of samples into the dataset (Bareinboim and Pearl, 2012). 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).

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 the selection bias. Such assumed relationships can be compactly and transparently represented as a causal data-generating process (DGP) in the form of a directed acyclic graph (DAG), for example illustrated in Figure 1. 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.



(a) Well-specified: Symbolic entity G, a common cause of both dataset features and labels.

(b) Underspecified: G unobserved by features, thus features contain no causes of the labels.

Figure 1: Data generating process for high dimensional data, such as in NLP, where X and Y represent high dimensional text features: the dataset features and labels, while W, G, and S represent low dimensional symbolic entities that may cause the text.

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

We will use the term *selection induced 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[1]. Selection induced collider bias has been covered in medical and epidemical literature (Griffith et al., 2020) (Munafò et al., 2018) (Cole et al., 2009) and received extensive theoretical treatment from Pearl and Bareinboim in (Bareinboim and Pearl, 2012), (Bareinboim et al., 2014), (Bareinboim and Tian, 2015) and Bareinboim and Pearl (2016), yet has received very little attention in deep learning literature.

1.2 Underspecification and Spurious Associations

We define a learning task as *underspecified* when none of the features available to the model (at training or inference time) are causes of the label. Figure 1(b) encodes this relationship with the absence of an arrow between features, X, and labels, Y. 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. We refer to these as *otherwise non-interacting* spurious associations.

We would like to draw a distinction between the types of spurious association induced by underspecification, and the spurious associations most often addressed in today's literature. For example, the task of predicting cow vs camel (perhaps based on *spurious* grassy vs sandy background pixel features) is not considered an underspecified task, due to the availability of the *causal* cow vs camel pixel features in the foreground. From a causal perspective, the symbolic background entity is a common cause of both the pixel features and the labels, inducing confounder bias and thus the learning of spurious associations along a *secondary* spurious path (Arjovsky et al., 2019), in addition to the *primary* direct causal path from feature to label.

A natural question to ask is, how does spurious association flow from X to Y, if not through some confounding variable like background, nor though a direct causal path. As demonstrated in (D'Amour et al., 2020), such weakly-interacting prediction tasks display significant variance due to even changes in the random seed initialization. In this work, by focusing on variables engaged in a relationship of selection induced collider bias, we are able to open up a *tertiary* path between X and Y: the path along $X \leftarrow W \rightarrow S \leftarrow G \rightarrow Y$ in Figure 1(b). We argue this causal perspective facilitates the identification of otherwise non-interacting (and previously unreported) spurious associations, and importantly enables the injection of these 'benign' spurious tokens into text at inference time, to achieve an uncertainty measurement.

2 Contributions

In this paper we make the following contributions:

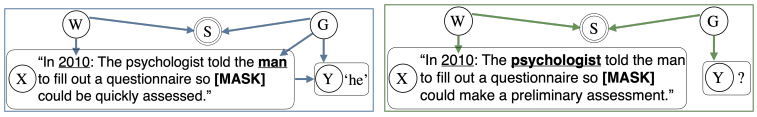
- We cast the problem of task underspecification in causal terms and apply causal inference methods to hypothesize the effects of selection induced collider bias on underspecified tasks.
- We test these hypotheses on unmodified and widely used pre-trained LLMs via a case study of gender pronoun resolution, resulting in two new findings:
  - A method for empirical measurement of spurious correlations between gender and gender-neutral entities for unmodified LLMs which permits measurement of previously unreported spurious correlations between gender vs location and time.
  - A method for quantifying inference-time task uncertainty that can correctly report when a task is underspecified or not on over 90% of the Winogender Schema challenge test set.
- To demonstrate that both above methods are reproducible, lightweight (dozens of lines of code), time-efficient (takes seconds), and plug-n-play compatible with almost any BERT-like LLM, we provide open source and running demos:
  - Spurious Correlations: `https://huggingface.co/spaces/paper5186/spurious`.
  - Uncertainty: `https://huggingface.co/spaces/paper5186/uncertainty`.
- We generalize our approach to address a wider range of prediction tasks and provide results on a generic data-generating process that are consistent with our empirical measured results on LLMs.

[1] Although conflated, collider bias can occur independent of selection bias and vice versa (Hernán, 2017)

3 A GENDER PRONOUN RESOLUTION TASK DATA-GENERATING PROCESS

An example of underspecification can be found in the gender pronoun resolution task in Figure 2(b). Note, a single prediction task may be partitioned into well-specified and underspecified 'sub'tasks. For example, Figure 2(b) may be well-specified for part-of-speech tagging. Gender pronoun resolution will serve as a case study for the rest of this paper, for largely two reasons: 1) it is a well-studied problem with many recent advances (Cao and Daumé III, 2020), (Lu et al., 2018), (Webster et al., 2020), (Zhao et al., 2018) and yet remains a challenging problem for modern LLMs (Deng et al., 2022) 2) it provides many human-relatable underspecified task scenarios.

To apply causal inference methods to gender pronoun resolution, we must make some initial assumptions about a plausible DGP for the relevant features and labels, that we can later support with empirical measurement. These assumptions have been transparently represented in Figure 1 and Figure 2. Note the heterogenous nature of the DAG variables, in which X and Y represent high dimensional features like text in the dataset, while W, G, and S represent low dimensional symbolic entities that may cause the text.



(a) Well-specified ($X \rightarrow Y$): Masked pronoun coreferent with the *man*.

(b) Underspecified ($X \not\rightarrow Y$): Masked pronoun coreferent with *psychologist*.

Figure 2: Gender pronoun resolution task with modified Winogender Schema (Rudinger et al., 2018) test sentences.

Specifically in Figure 2, the symbol X represents the *text* sentences in our dataset, and Y represents the label: 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. The remaining symbols in Figure 2 are not recorded in the dataset. The symbol G represents *gender* and in well-specified gender resolution tasks, G causes both X and Y, in Figure 2(a) only. W represents gender-neutral entities that are not the cause of Y, but still of interest because they cause X. Additionally, in order to measure the effects of selection induced collider bias, we must find entities for W that are also the cause of S: selection into 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. In both cases, the disparity in *access* can result in preferential inclusion of samples into the dataset, on the biases of gender. In this paper, we use W to represent *date* and *place*, for our LLM measurements.

Figure 1(b) and Figure 2(b) are the underspecified counterparts to the well-specified prediction tasks in Figure 1(a) and Figure 2(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 G and X. Further, because W is also gender-neutral, once we have removed any gender-identifying features from X, we should additionally remove the path between X and Y, as there is no longer any feature in X causing Y.

With the causal path from $X \rightarrow Y$ removed, 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) and Figure 2(b), selection induced 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 induced collider bias causes the transformation: $(Y \perp\!\!\!\perp X)_{RW} \overset{S}{\Rightarrow} (Y \not\perp\!\!\!\perp X)_{LM}$, due to otherwise non-interacting spurious associations, such as those between *gender* and *time*. Whereas spurious associations learned in the well-specified model would be more likely due to

signals that are stronger than those along the causal path $X \rightarrow Y$, as was discussed previously in the camel vs cow example.

4 SELECTION INDUCED COLLIDER BIAS

In addition to transparently stating our assumptions, the causal DAGs in Figure 1 and Figure 2 also entail specific conditional probabilities that should be empirically measurable, under the condition of two additional assumptions: For a given causal graph, Gr, and a measured probability distribution Pr, we are assuming 1) the Markov assumption: $(X \perp\!\!\!\perp Y|Z)_{Gr} \implies (X \perp\!\!\!\perp Y|Z)_{Pr}$ and 2) faithfulness: $(X \perp\!\!\!\perp Y|Z)_{Gr} \impliedby (X \perp\!\!\!\perp Y|Z)_{Pr}$ (Peters et al., 2017), where Z is the set $\{W, G, S\}$.

To capture the statistical process of dataset formation, we implicitly condition on $S = 1$ for all the samples in the dataset. Conditioning on S induces 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), entailing a conditional dependency between otherwise non-interacting entities: G, W.

However, at inference-time we only have access to X, Y. Here we will show that for underspecified tasks, we would expect $P(Y|X)$ to be distributed similarly as $P(G|W)$. Revisiting the underspecified DAGs in Figure 1(b) and Figure 2(b) and applying the Markov and faithfulness assumptions we can estimate the conditional probability of a gender pronoun, Y, given gender-neutral text, X.

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 pronoun*), 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 $P(G|W) \neq P(G)$, and thus we add W behind G's conditioning bar. Finally, Equation (4) assumes we have an underspecified (gender-neutral) text, X, so $P(G|X) = P(G)$, and thus we remove X from behind the conditioning bar.

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

Equation (1) - Equation (4) show that $P(Y|X) \sim P(G|W)$ in underspecified tasks, providing information about the otherwise inaccessible latent representations for G and W and specifing a measurable relationship between X and Y, that we can validate in the next section.

4.1 MASKED GENDER TASK (MGT) CHALLENGE SET

To validate spurious associations between *otherwise non-interacting* entities, such as *time* and *gender*, we desire a challenge test set of gender-neutral text. However, the text from the example in Figure 2(b) does not satisfy this requirement due to associations between gender and occupation. We also desire a test set compatible with the Masked Language Modeling (MLM) objective used in the pretraining of LLMs[2], for the greatest applicability of the results. Table 1 shows the heuristic we used to create a challenge set composed of (2 Python f-string templates) $\times$ (5 tenses of `verb` to-be) $\times$ (6 `life_stages`) $\times$ (30 values for W as *date*) $\times$ (20 values for W as *place*) $= 36000$ gender-neutral test sentences for X.

4.2 PRE-TRAINED LLMS GENDER-DATE AND GENDER-PLACE DEPENDENCIES

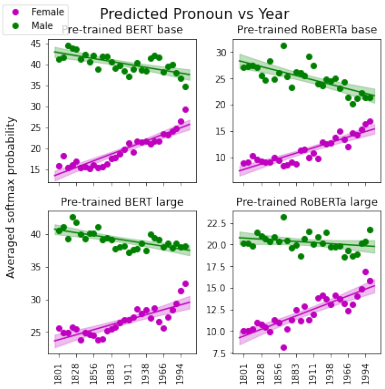
Figure 3 shows pre-trained BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), base and large results on our MGT challenge set. Each plotted dot is the softmax probability (each averaged over 60 gender-neutral texts) for predicted gender pronouns vs *year*, or vs *country*, where the *year* or *country* along the x-axis matches the gender-neutral W value injected into the gender-neutral text[3].

[2]Tests compatible with the MLM objective can avoid fine-tuning weights during measurement procedures.

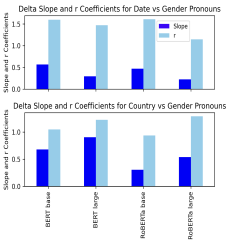
[3]For example, the purple and green dots at the x-axis position of $W = 1938$ are the softmax predictions for the masked word in input texts like 'In 1938, [MASK] will became a teenager.', for female and male pronouns.

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 range of 30 *dates* and 20 *places*, defined in Appendix C.

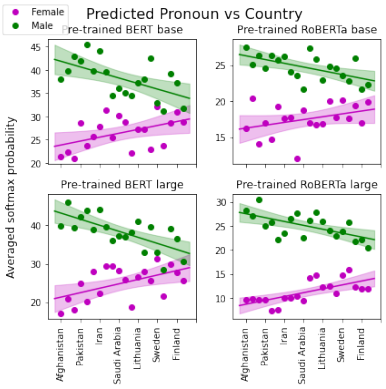
W Category	Python f-string templates	Example text
Date & Place	`f"[MASK] {verb} {life_stage} in {w}."` `f"In {w}, [MASK] {verb} {life_stage}."`	'[MASK] was a teenager, in 1953.' 'In Mali, [MASK] will be an adult .'



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



(b) Difference between the female and male pronoun linear fit slope and Pearson's r coefficients for Figure 3(a) (top) and Figure 3(c) (bottom).



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

Figure 3: Softmax probabilities and linear fit line coefficients for LLM predictions displaying spurious correlation between *gender*, G, and gender-neutral entities, W, injected into gender-neutral input texts described in Table 1.

As the final layer in the pre-trained LLM is a softmax over the entire tokenizer's vocabulary, we sum (without normalization) the gender-identified portion (as listed in Table 3) of the probability mass from the top five predicted words[4].

The shaded regions in Figure 3(a) and Figure 3(c) show the 95% confidence interval for the linear fit, and Figure 3(b) 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 these plots. It is noteworthy that the magnitude of the spurious correlations appear similar across model size, but from this data alone, it is unclear the role of model overparameterization in the learning of otherwise non-interacting spurious associations.

Overall, in Figure 3 we can see that the injection of different W values into X results in a changed prediction for Y consistent with the $P(G|W)$ conditional dependency entailed by collider bias at S. These spurious associations of *gender* vs *time* and *place* have not been reported before this work. Further, with the derivation that $P(Y|X) \sim P(G|W)$ in underspecified tasks, additional such spurious associations should be measurable via only repeated inference runs.

[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.

5 EXPLOITING FOR UNCERTAINTY MEASUREMENT

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 on an underspecified task with selection induced collider bias will not converge toward the unbiased distribution without additional data or assumptions (Bareinboim et al., 2014). However, in this section we will exploit the outcomes of selection bias to our benefit, in a method we call *W*-injection and describe below.

5.1 WINOGENDER TEXTS

We test *W*-injection 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), and already seen in Figure 2.

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 (2) the *participant* in the sentence (Rudinger et al., 2018). As was the case in the MGT challenge set, 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 3(a).

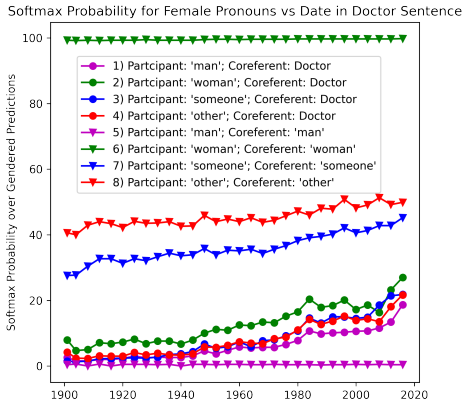


Figure 4: Averaged softmax percentages from RoBERTa large for predicted female gender pronouns (normalized over all gendered predictions) vs a range of dates (injected into the text), for the 'Doctor' Winogender texts listed in Table 2.

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.

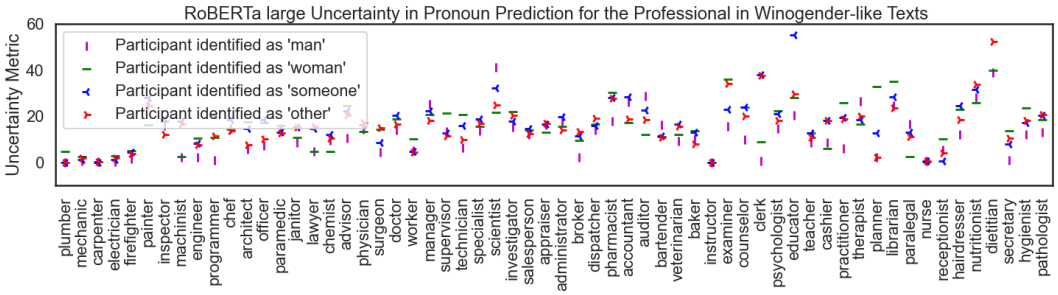
Figure 4 shows the predicted softmax probability for female pronouns for the masked words in the Table 2 sentences, normalized to the gendered predictions of the top five predicted words from pre-trained RoBERTa large. Similar to the findings in (Rudinger et al., 2018), at any given W value, the softmax probabilities for female pronouns are higher for masked pronouns coreferent with the patient as opposed to the doctor, indicating the now well known learned spurious association between gender and occupation.

[5]Similar results can be obtained with the *W*-injection of countries, as was done in Figure 3(c).

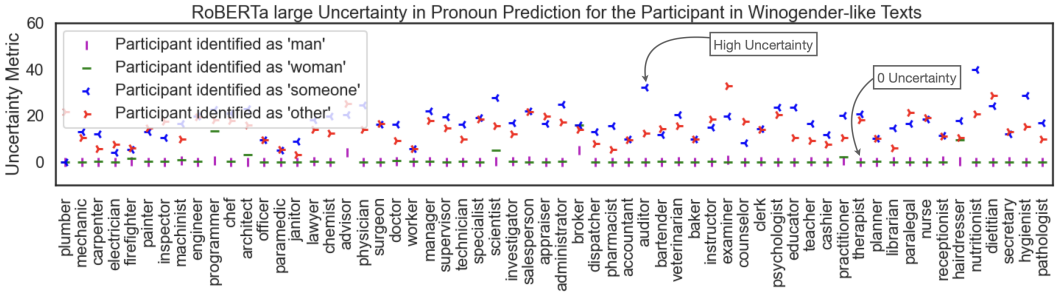
[6]We picked a slightly narrower and more modern date range as compared to that of Figure 3(a) for contextual consistency with some of the more modern occupations in the Winogender evaluation set.

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 uncertainty for these sentences. For example, a cutoff threshold of 1.0 for our uncertainty metric would correctly classify each sentence as well or underspecified.

ID	Sentence with Participant and **Coreferent** Highlighted	Uncertainty BERT base	BERT large	RoBERTa base	RoBERTa 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



(a) Masked pronoun is coreferent with the *professional* (similar to IDs 1-4 in Table 2), so all these sentences remain underspecified. We see the accurate reporting of model uncertainty results above 0 for most occupations, including tricky sentences containing a co-occurring gender-identifying term for the *participant*.



(b) Masked pronoun is coreferent with the *participant* (similar to IDs 5-8 in Table 2), thus sentences containing 'man' and 'woman' are well-specified, and the rest are underspecified. We see the accurate reporting of uncertainty results close to 0 for the well-specified 'man' and 'woman' sentences, and generally above 0 for the underspecified 'someone' and 'other' sentences.

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

Our contribution here is that in Figure 4 we can see that the spurious associations due to the *W*-injection of an unrelated spurious association (time vs gender) appears *additive* with the existing spurious association between occupation and gender, and *both spurious relationships only occur* when the prediction task is underspecified.

5.2 Uncertainty Metric

As a simple and naive 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 *W*-injection has a larger influence on 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.

Our extended version of the Winogender Schema contains (60 *professional* occupations) $\times$ (4 *participant* text tokens) $\times$ (30 values for DATE) $\times$ (2 sentence templates[7]). 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 uncertainty metric defined in the proceeding paragraphs.

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. Focusing exclusively on the most challenging Winogender sentence-types, in which a gender-identifying text: 'man' or 'woman' is used for the *participant* in the sentence, and using an unoptimized thresholding value of 1.0, we still find the uncertainty metric correctly classifies the well-specified sentences for 91.7%, and underspecified sentences for 90.0%, of the Winogender occupations. We show similar results for BERT and RoBERTa base and BERT large in Appendix D, but note the uncertainty metric appears more accurate for more overparameterized models.

6 Extending to More General Setting

We now explore a more general problem space where the symbols in Figure 1 take on the following meanings: G is the causal parent of Y, W is the non-causal parent 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 a 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 DAGs in Figure 1, with the structural causal model (SCM) shown on the right.

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

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

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

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

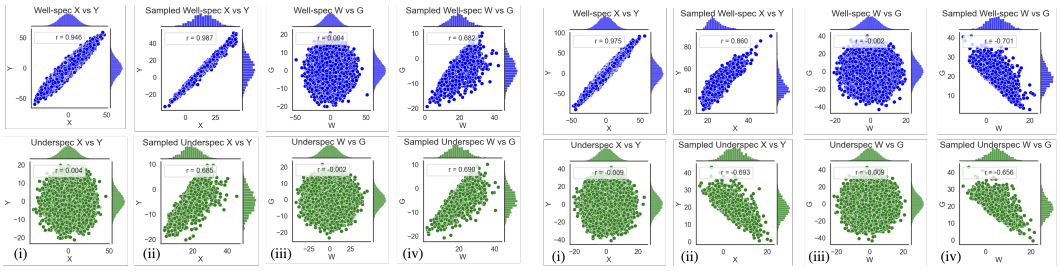
$$Y := \gamma X + G + \mathcal{N}(0,1) \tag{9}$$

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, reducing the dataset to about 5% of its original size. 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 entailed by the SCM above. Columns (i) plots X vs Y for the unsampled, and columns (ii) plots X vs Y for the $S = 1$ sampled distributions. In both Figure 6(a) and Figure 6(b), we can see that selection induced collider bias has little effect on the distributions (i) vs (ii) for the well-specified SCM, but causes the underspecified SCM's Pearson's r coefficient to go from 0 to 0.7. The latter is consistent with Section 4.2 measurements of gender-date (and gender-place) correlations on underspecified tasks in LLMs trained on sampled datasets,

[7] One template with the masked pronoun coreferent with the *professional* and the other with the *participant*.

[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.



(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 SCM defined in Equation (5) to Equation (9), with Equation (7) separately defined above, with the well-specified (top row in blue) and underspecified (bottom row in green) DGP in Figure 1.

while we know that gender and date (as well as gender and place) are uncorrelated in the unsampled population.

Columns (iii) and (iv) plots the unsampled and the $S = 1$ sampled distributions for W vs G. Comparing Figure 6(a) and Figure 6(b), we can see the direction of the correlation coefficient has been flipped for both the well and underspecified SCMs. However, the direction of the correlation coefficient flips in the X vs Y distributions of column (ii) only for the underspecified SCM. This is consistent with our derivation of $P(Y|X) \sim P(G|W)$ for underspecified tasks in Section 4.

In Figure 7, we replace Equation (8) with $X := \beta W + \gamma G + \mathcal{N}(0,1)$, where β takes increasingly larger values: 1) $\beta = 0.01$ 2) $\beta = 0.1$ and 3) $\beta = 1$ (thus identical to Equation (8)), as a toy demonstration of W-injection. Similar to the uncertainty results in Section 5, we see underspecified tasks are more sensitive to W-injection, while well-specified are not.

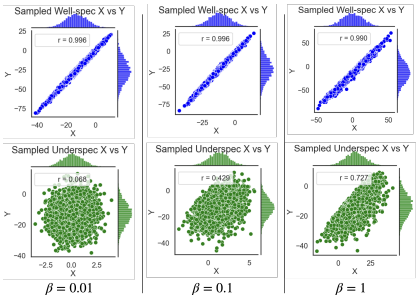


Figure 7: W-injection implemented via increasingly larger values of β as the $W \to X$ path weight.

7 Discussion and Reproducibility

As underscored in (D'Amour et al., 2020), the prevalence of underspecified tasks in machine learning requires the development of nuanced stress tests. In this work we strived to make our methods accessible and extensible to other tasks in machine learning. Please see more details in Appendix A.

We have argued that underspecified prediction tasks leave models vulnerable to selection induced collider bias and the resulting spurious associations, including non-interacting and previously unreported associations between gender vs time and gender vs place, which we demonstrate on unmodified pre-trained BERT-like models. We have introduced a lightweight inference-time technique for injecting spurious signals into prediction 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. While not a universal solution, we hope our work can lead to the development of targeted heuristics using our method to determine when the model is uncertain about the prediction task, and an alternate value should be returned (for example 'they' in the case of gender pronoun resolution).

For further research, we believe our work complements that of (Vig et al., 2020), which uses causal mediation analysis to intervene on LLMs at the individual attention head and neuron level to provide insights into the model's internal causal mechanisms mediating gender bias. Despite the more limited nature of our non-invasive investigation, our surprising empirical findings about the additivity of unrelated spurious associations is consistent with their surprising finding that gender bias appears decomposable between the elements of the direct and indirect effect within the model. We hope to be able to complement more of their internal model findings via external methods.

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

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016. doi: 10.1073/pnas.1510507113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1510507113.

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

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

Yang Trista Cao and Hal Daumé III. Toward gender-inclusive coreference resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4568–4595, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.418. URL https://aclanthology.org/2020.acl-main.418.

Stephen R Cole, Robert W Platt, Enrique F Schisterman, Haitao Chu, Daniel Westreich, David Richardson, and Charles Poole. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, 39(2):417–420, 11 2009. ISSN 0300-5771. doi: 10.1093/ije/dyp334. URL https://doi.org/10.1093/ije/dyp334.

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.

Yuntian Deng, Volodymyr Kuleshov, and Alexander M. Rush. Model criticism for long-form text generation, 2022. URL https://arxiv.org/abs/2210.08444.

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.

Miguel A. Hernán. Invited Commentary: Selection Bias Without Colliders. *American Journal of Epidemiology*, 185(11):1048–1050, 05 2017. ISSN 0002-9262. doi: 10.1093/aje/kwx077. URL https://doi.org/10.1093/aje/kwx077.

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.

Kaiji Lu, Piotr Mardziel, Fangjing Wu, Preetam Amancharla, and Anupam Datta. Gender bias in neural natural language processing, 2018. URL https://arxiv.org/abs/1807.11714.

Marcus R Munafò, Kate Tilling, Amy E Taylor, David M Evans, and George Davey Smith. Collider scope: when selection bias can substantially influence observed associations. *Int. J. Epidemiol.*, 47(1):226–235, February 2018.

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.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural nlp: The case of gender bias, 2020. URL https://arxiv.org/abs/2004.12265.

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

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-2003. URL https://aclanthology.org/N18-2003.

A Reproducibility

Please see the following sources to reproduce the methods and measurements in this paper:

- Spurious Correlations Open Source Hugging Face Space (screen shot at Figure 10): https://huggingface.co/spaces/paper5186/spurious.
- Uncertainty Measurement Open Source Hugging Face Space (screen shot at Figure 11) : https://huggingface.co/spaces/paper5186/uncertainty.
- More General Setting Toy SCM : https://tinyurl.com/2ub4xyjs.
- Github repo to replicate all the plots will be coming very soon.

B 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

C 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 induced 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 C.1).

C.1 PLACE VALUES

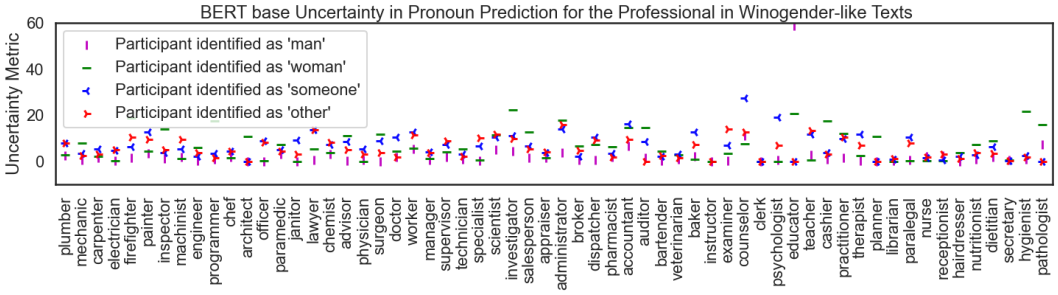
Ordered list of bottom 10 and top 10 World Economic Forum Global Gender Gap ranked countries used for the x-axis in Figure 3(c), that were taken directly without modification from `https://www3.weforum.org/docs/WEF_GGGR_2021.pdf`:

'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'

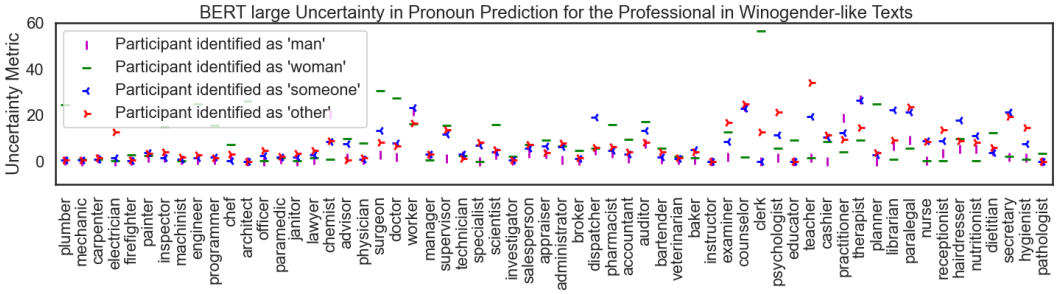
D EXTENDED WINOGENDER UNCERTAINTY RESULTS ON MORE LLMS

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.

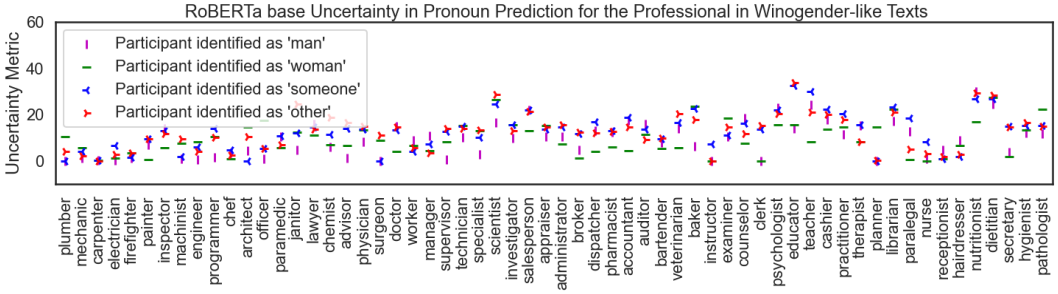
Figure 9 shows uncertainty results for all Winogender occupations where the masked pronoun is coreferent with the *participant*, unlike Figure 8 where the pronoun is coreferent with the *professional*. Because the injected text (again one of: {'man', 'woman', 'someone', 'other'}) is referring to the *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

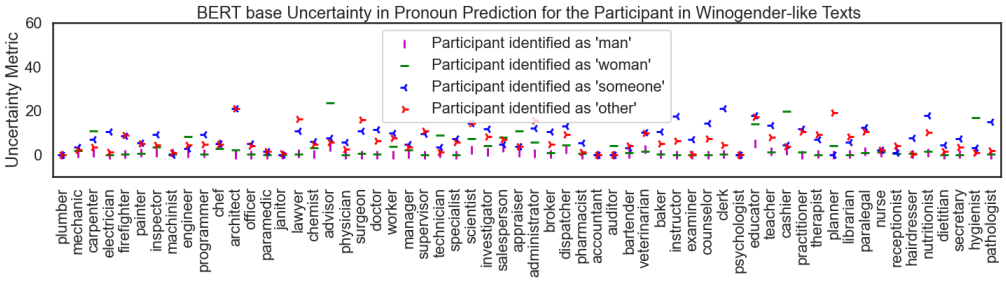


(b) BERT large

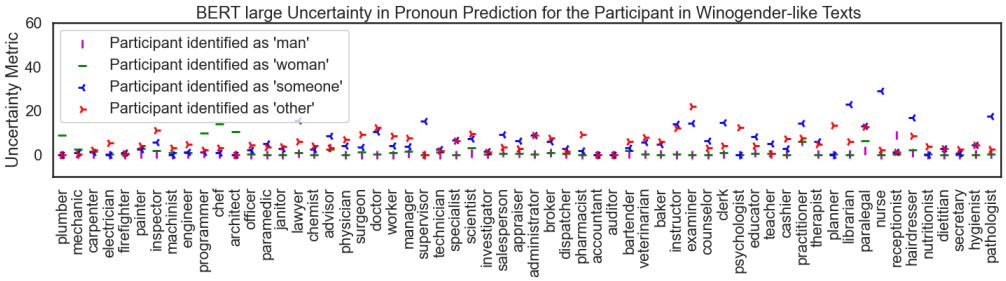


(c) RoBERTa base

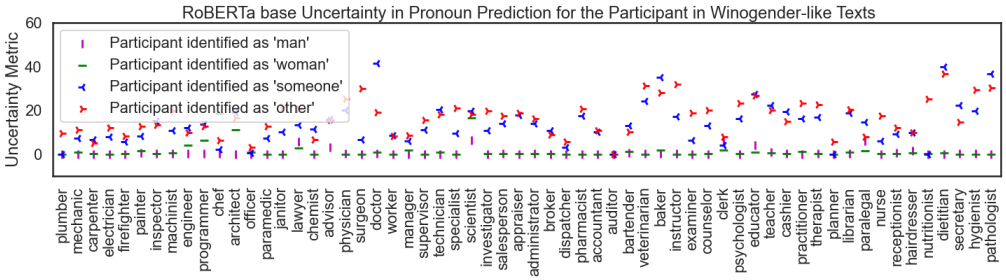
Figure 8: Uncertainty results for all Winogender occupations where the masked pronoun is coreferent with the gender-unidentified *professional*, thus all sentences remain underspecified. The plots show that generally, the models do not become erroneously certain about gender when the words 'man' and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

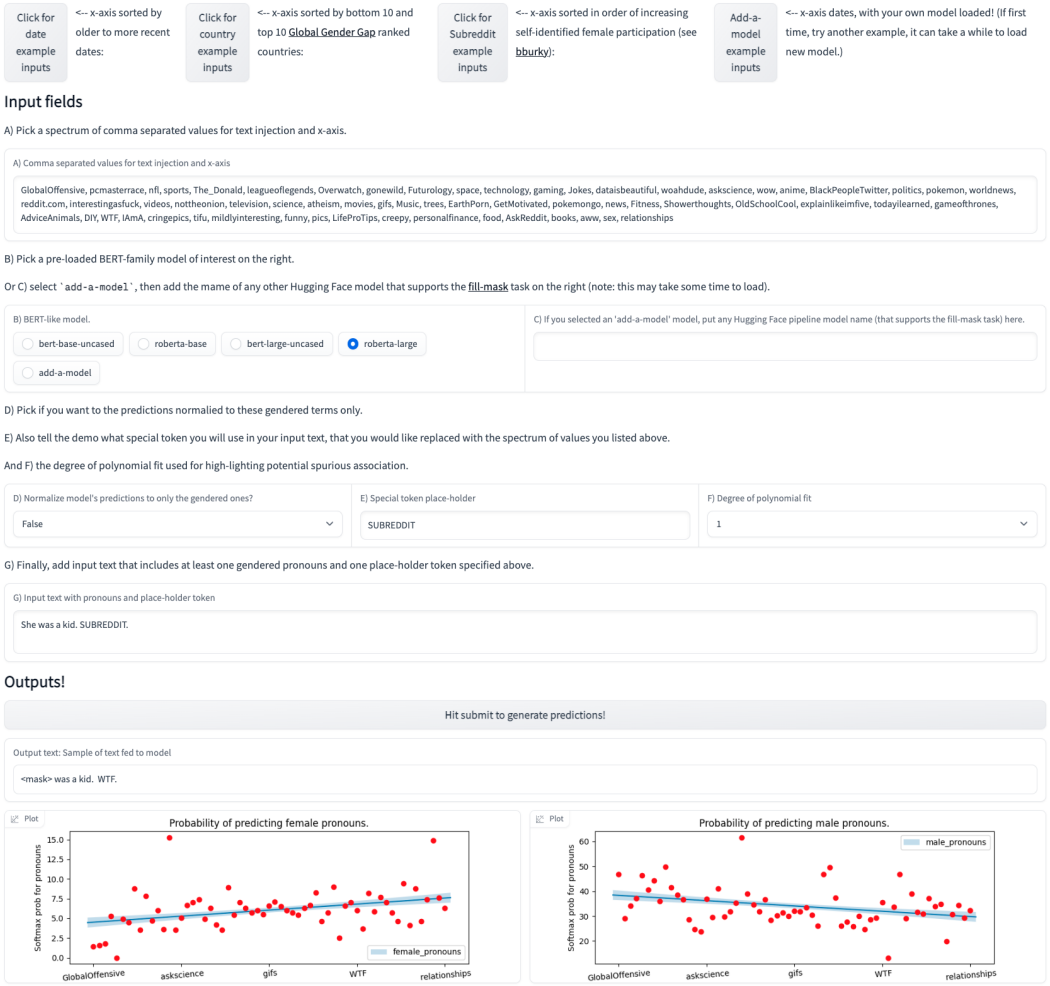


Figure 10: Demo where users can choose their own input text and select almost any BERT-like model hosted on Hugging Face to test for selection induced collider bias induced spurious correlations.

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 induced collider bias induced spurious correlations.

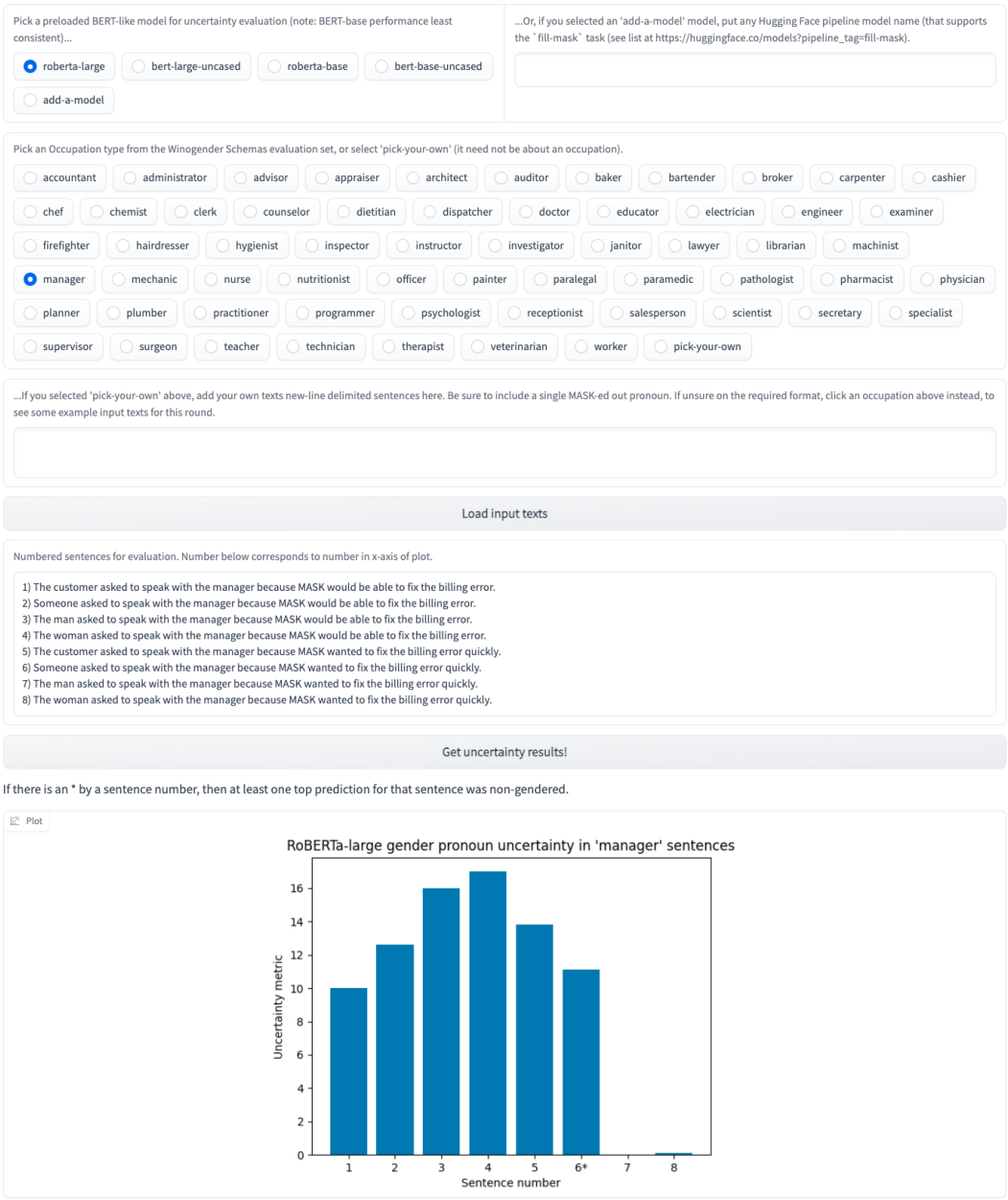


Figure 11: Demo where users can choose their own input text and select almost any BERT-like model hosted on Hugging Face to test for model uncertainty using selection induced collider bias induced spurious correlations.