Exploiting Selection Bias on Underspecified Tasks in Large Language Models

Emily McMilin
Independent Researcher
Palo Alto, CA 94306
emcmilin@cs.stanford.edu

Abstract

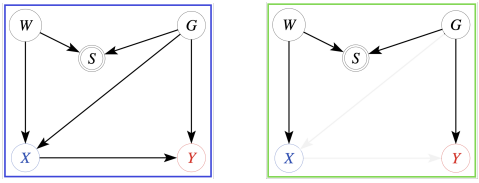
In this paper we motivate the causal mechanisms behind sample selection induced collider bias (selection collider bias) that can cause Large Language Models (LLMs) to learn unconditional dependence between entities that are unconditionally independent in the real world. We show that selection collider 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.

NeurIPS 2022Workshop on Causal Machine Learning for Real-World Impact (CML4Impact 2022).

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 process 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]. In [McMilin, 2022], we have shown how *selection collider bias* can cause spurious associations, including weakly associating and previously unreported correlations between gender vs time and gender vs place, which we demonstrated on unmodified pre-trained BERT [Devlin et al., 2018] and BERT-like models.

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 paper we will consider the 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 Selection Collider Bias

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.

To capture the statistical process of dataset formation, we implicitly condition on $S = 1$ for all the samples in the dataset. Now applying the 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).

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

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

Table 1: 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

gender-neutral text, X, and thus $P(G|X) = P(G)$. 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 measured on BERT-family LLMs in [McMilin, 2022]: $(Y \perp\!\!\!\perp X)_{RW} \approx (G \perp\!\!\!\perp W)_{RW} \overset{S}{\Rightarrow} (G \not\perp\!\!\!\perp W)_{LM} \approx (Y \not\perp\!\!\!\perp X)_{LM}$.

4 Exploiting Selection Collider Bias

Here we 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.

The 'Sentence' column in Table 1 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]. For the masked gender task, this pronoun is replaced with a [MASK] for prediction.

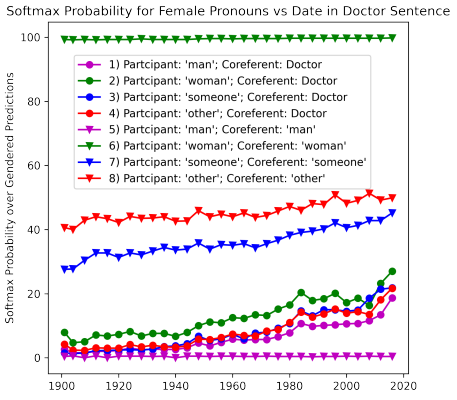
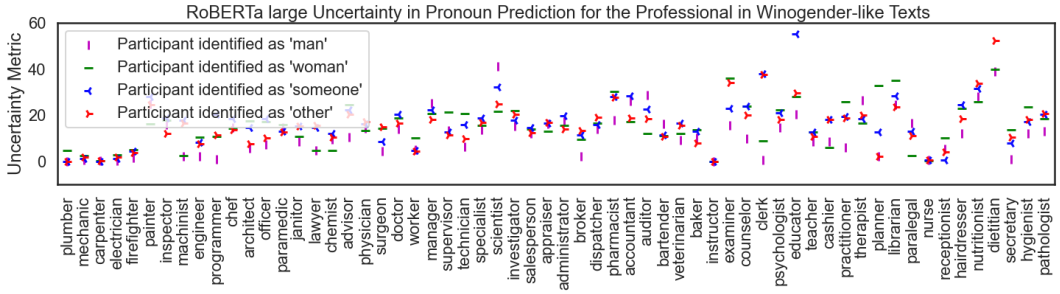
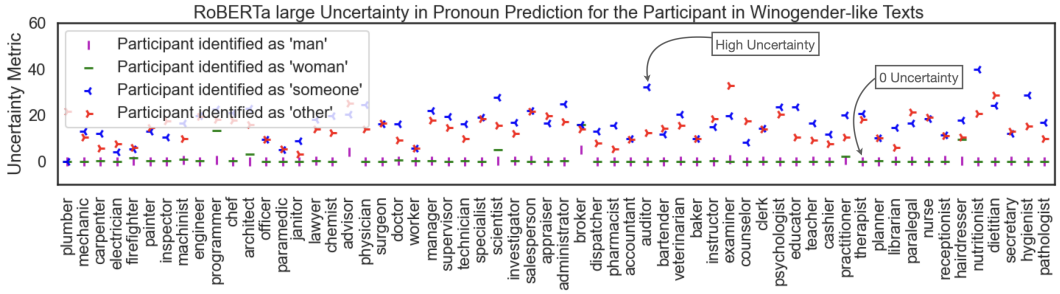


Figure 2: Averaged softmax percentages from RoBERTa large for predicted female gender pronouns (normalized over all gendered predictions) vs a range of dates (injected into the text), for the 'Doctor' Winogender texts listed in Table 1.

Our extensions to the evaluation set are two-fold: 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



(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 3: RoBERTa-large uncertainty results on all Winogender Schema occupations.

phrase 'In DATE', where 'DATE' is replaced by a range of years from 1901 to 2016[1], similar to what was done in [McMilin, 2022].

In Sentence IDs 1 - 4 of Table 1, 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 2 shows the predicted softmax probability for female pronouns for the masked words in the Table 1 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 2 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.

4.1 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 2. For this uncertainty metric, we would expect larger values for underspecified prediction tasks, in which

[1]We picked a slightly narrower and more modern date range as compared to that in [McMilin, 2022] for contextual consistency with some of the more modern occupations in the Winogender evaluation set.

the spurious correlation between *gender* and *date* has a larger role in influencing the prediction. For the predictions in Figure 2, this metric is shown in the 'Uncertainty' columns in Table 1 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 1[2].

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 3(a), with input sentences like IDs 1 - 4 where the masked pronoun is coreferent with the *professional*, and 2) in Figure 3(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.

In Figure 3, 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. We show similar results for BERT and RoBERTa base and BERT large in Appendix A, but note that the increased parameter count and hyper-parameter optimization in RoBERTa large appears to improve the uncertainty measurement.

5 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.

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[3]. 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).

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

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

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

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

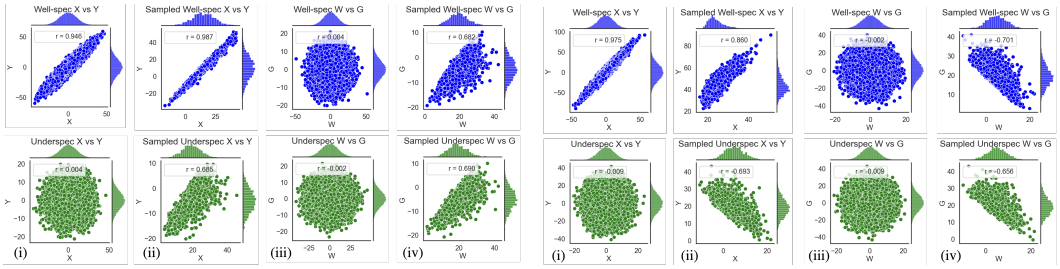
$$Y := \gamma X + G + \mathcal{N}(0,1) \tag{9}$$

Figure 4 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 4(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 0.685, consistent with the selection collider bias induced transformation: $(Y \perp\!\!\!\perp X)_{RW} \overset{S}{\Rightarrow} (Y \not\perp\!\!\!\perp X)_{M}$.

In Figure 4(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

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

[3]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 4: Statistical relationships induced by the structural causal model defined by Equation (5) to Equation (9), with Equation (7) separately defined above, for the well-specified (top row in blue) and underspecified (bottom row in green) models in Figure 1.

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 for underspecified models. For Figure 4(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 4(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.

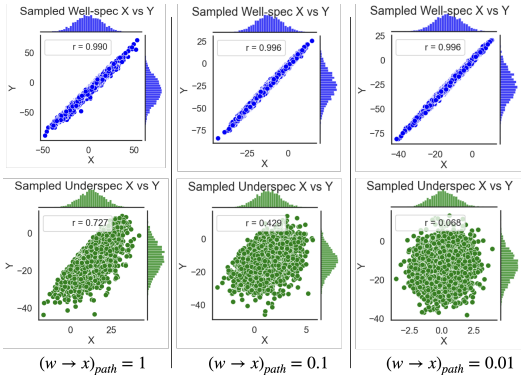


Figure 5: Decreasing the path coefficient, β, for $X \rightarrow Y$, as a toy representation of what we termed the W-Injection.

For the toy demonstration of W-injection in Figure 5, 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 4, probing $P(Y|X)$ for a range of W values can serve to classify tasks as underspecified if $P(Y|X)$ is sensitive to W-injection, or well-specified if $P(Y|X)$ is unaffected.

6 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 B and Appendix C and available at `https://huggingface.co/spaces/emilylearning/spurious_correlation_evaluation` and `https://huggingface.co/spaces/emilylearning/llm_uncertainty`, respectively. We will make all code available at `https://github.com/2dot71mily/exploiting_selection_bias`.

7 Discussion

In this paper we have argued that underspecified prediction tasks leave models vulnerable to selection collider bias induced spurious associations, and 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.

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.

8 Acknowledgments

Thank you to the CML4Impact reviewers for their helpful comments, to Rosanne Liu and Jason Yosinski for their encouragement, to Jen Iofinova and Sara Hooker with Cohere For AI for their support, to Hugging Face for their open source services, and to Judea Pearl, Elias Bareinboim, Brady Neal and Paul Hünermund for their fantastic online causal inference resources.

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.

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.

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.

Emily McMilin. Selection bias induced spurious correlations in large language models, 2022. URL https://arxiv.org/abs/2207.08982.

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

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

A Extended Winogender Uncertainty Results on More LLMs

Figure 6 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.

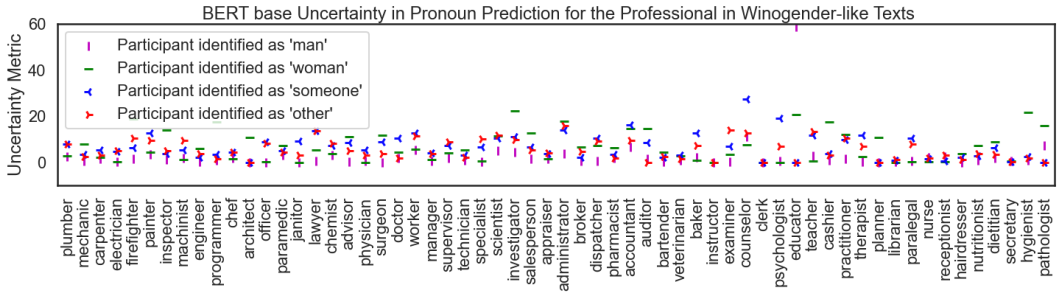
Figure 7 shows uncertainty results for all Winogender occupations where the masked pronoun is coreferent with the *participant*, unlike Figure 6 where the pronoun is coreferent with the *professional*. Because the injected text (again one of: {'man', 'woman', 'someone', 'other'}) is referring to the *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 3(b).

B Spurious Correlations Demo

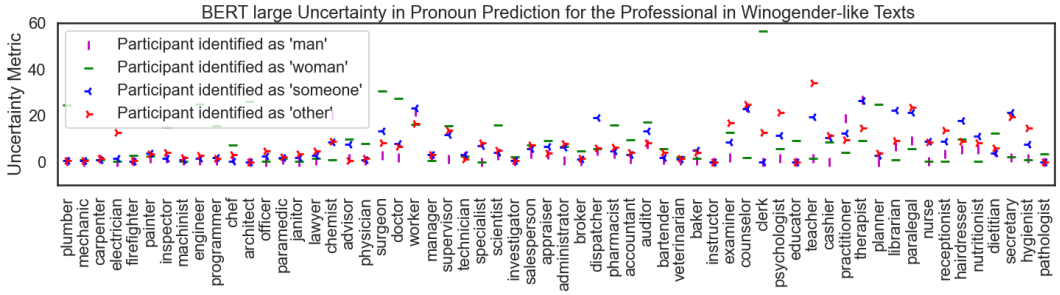
See Figure 8 for our open source freely available demonstration where users can choose their own input text and select almost any BERT-like model hosted on Hugging Face to test for selection collider bias induced spurious correlations.

C Model Uncertainty Demo

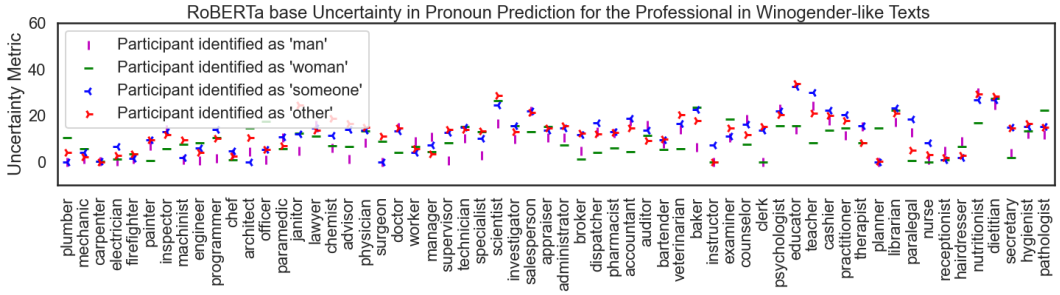
See Figure 9 for our open source freely available demonstration where users can choose their own input text and select almost any BERT-like model hosted on Hugging Face to test for model uncertainty using selection collider bias induced spurious correlations.



(a) BERT base

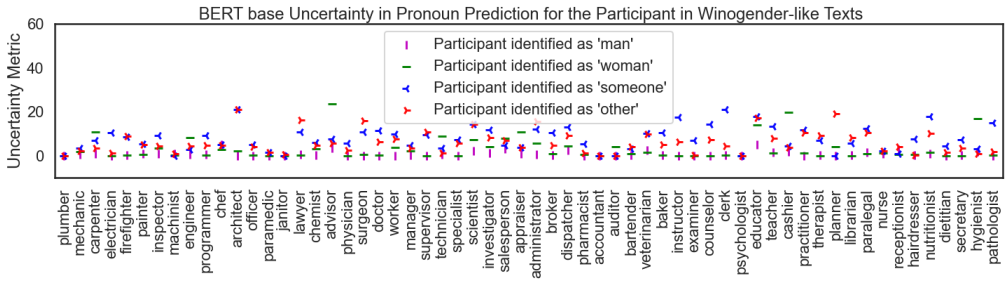


(b) BERT large

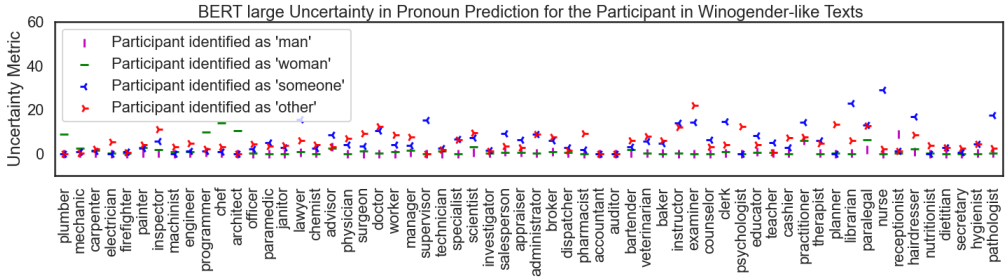


(c) RoBERTa base

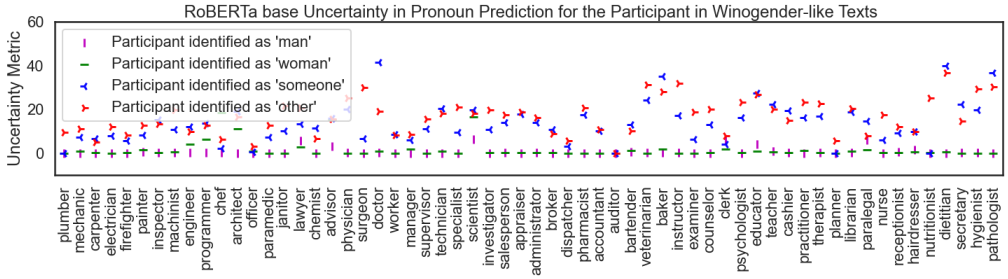
Figure 6: Uncertainty results for all Winogender occupations where the masked pronoun is coreferent with the gender-unidentified *professional*, thus all sentences remain underspecified. The plots show that generally, the models do not become erroneously certain about gender when the words 'man' and 'woman' are injected into the text.



(a) BERT base



(b) BERT large



(c) RoBERTa base

Figure 7: 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 3(b).

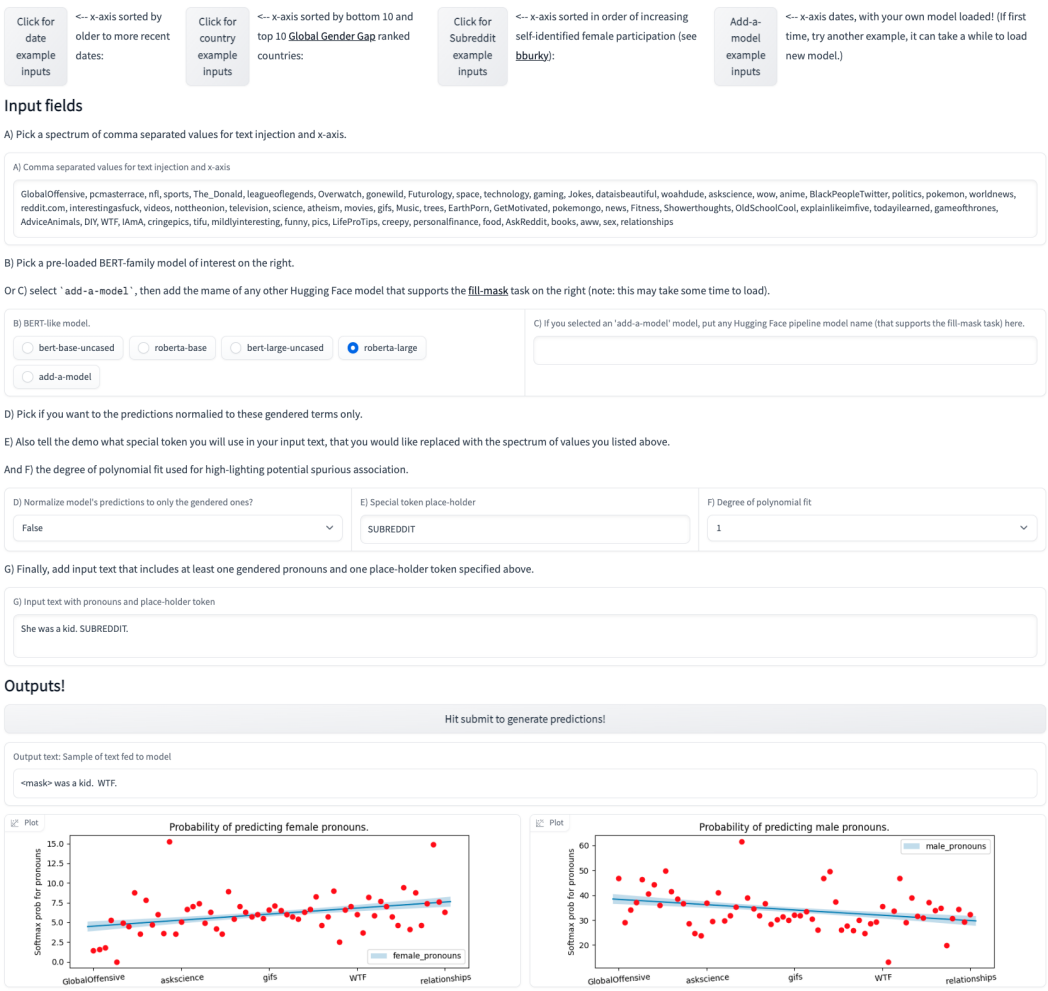


Figure 8: Demo where users can choose their own input text and select almost any BERT-like model hosted on Hugging Face to test for selection collider bias induced spurious correlations.

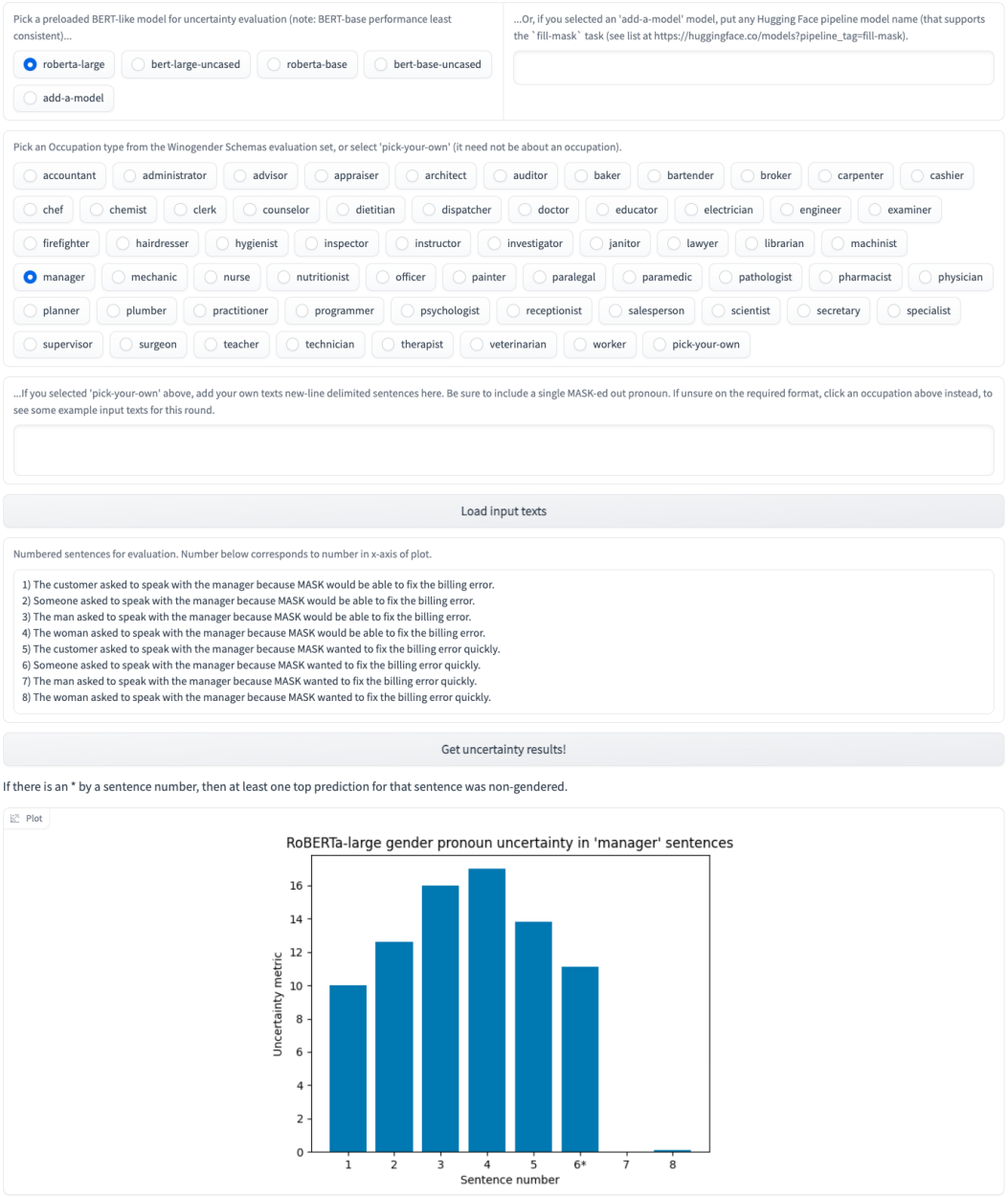


Figure 9: Demo where users can choose their own input text and select almost any BERT-like model hosted on Hugging Face to test for model uncertainty using selection collider bias induced spurious correlations.