Exploiting Selection Bias on Underspecified Tasks in Large Language Models

Emily McMilin
Independent Researcher
emcmilin@cs.stanford.edu

Background

Causal DAGs and Biases

- Sample selection bias: some mechanism, causes preferential inclusion of samples into the dataset [0]
- Confounder bias: when two variables have a common cause.
  - Correct by conditioning upon the common cause variable.
- Collider bias: when two variables have a common effect.
  - Correct by *not* conditioning upon the common effect variable.

(a) Well-specified: Entity G, a common cause of both dataset features, X, and labels, Y.

(b) Underspecified: G unobserved by features, and X contains no causes of Y.

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.

Underspecification & Spurious Associations

- We define underspecification: when none of the features are causes of the label.
  - Models must learn any spurious association that will reduce predictive risk.
  - Distinct from the learning of shortcut [1] spurious features:
  - Shortcut: feature & labels may have a common cause (e.g. 'background' grass for cow, sand for camel) of feature and label, but features still cause of label.
- How does spurious association flow from X to Y, if not through some confounding variable like 'background', nor though a direct causal path?
- Selection induced collider bias: Opens up a 3rd path between X and Y : the path along $X \leftarrow W \rightarrow S \leftarrow G \rightarrow Y$ [2].

Contributions

- Cast the problem of task underspecification in causal terms.
- Apply causal inference to hypothesize the effects of selection induced collider bias on underspecified tasks.
- Test hypotheses on unmodified and widely used pre-trained large language models (LLMs) via a case study of gender pronoun resolution, resulting in new findings:
  - A method for empirical measurement of spurious correlations between gender and gender-neutral entities: gender vs location and gender vs time.
  - A method for quantifying inference-time task underspecification with an accuracy of over 90% when testing RoBERTa-large with the Winogender Schema challenge test set.
- Demonstrate 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, with open-source and running demos:
  - https://huggingface.co/spaces/emilylearning/spurious_correlation_evaluation
  - https://huggingface.co/spaces/emilylearning/llm_uncertainty
- Generalize approach to address a wider range of prediction tasks, provide results on a generic DGP that are consistent with our empirically measured results on LLMs.

Problem Setting: Gender Pronoun Resolution

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

a) Underspecified ($X \nrightarrow Y$): Masked pronoun coreferent with *psychologist*.

Figure 2: DGP with augmented Winogender Schema [3] tests sentences

Plausible Data-generating Processes (DGP)

High dimensional variables, in the dataset:

- X: the *text* input sentences.
- Y: the label: a *gender pronoun*.
  - The arrow pointing from X to Y encodes our assumption that X is more likely to cause Y

Low dimensional entities, not recorded in the dataset, but may be causes of X or Y:

- G: *gender*
  - In well-specified gender resolution tasks: G causes both X and Y.
- W: *gender-neutral entities* that are not the cause of Y
  - Of interest because they cause X, and have been identified as also the cause of S.
- S: selection into the dataset.
  - $S = 1$ for samples in dataset, $S = 0$ otherwise.

[0] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. PMLR. Apr 2012.
[1] Martin Arjovsky, Le´on Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization, 2019.
[2] Gareth J Griffith. et. al. Collider bias undermines our understanding of COVID-19 disease risk and severity. Nature Communications, 11(1), November 2020.
[3] Rachel Rudinger, et. al. Gender bias in coreference resolution. CoRR, abs/1804.09301, 2018.

Method: Spurious Associations due to Selection Induced Collider Bias

W Category	Python f-string templates	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.'

Table 1: Heuristic for creating gender-neutral input texts for our challenge set, and 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']. For text versions of {w}, we used a range of 30 *dates* and 20 *places* (ordered by WEC Global Gender Gap rank). Totaled as (*2* Python f-string templates) × (*5* tenses of verb to-be) × (*6* life stages) × (*30* values for W as date + 20 values for W as place) = 3000 gender-neutral test sentences.

(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

Figure 3: Softmax probabilities (each dot an averag over 60 input texts) and linear fit coefficients for LLM predictions displaying spurious correlation between gender, G, and gender-neutral entities, W, injected into the gender-neutral input texts described in Table 1.

Method: Exploiting for Uncertainty Metric

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

Table 2: Extended Winogender evaluation sentences and uncertainty results for occupation as 'Doctor'. Only sentence IDs 5 & 6 are well-specified for gender pronoun prediction, and appropriately all models show the lowest uncertainty for these sentences. Cutoff threshold of 1.0 for our uncertainty metric would correctly classify all.

Figure 4: 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. Note: *no* gender-time spurious associations in the well-specified sentence IDs 5 & 6, and spurious associations due to the W-injection of an unrelated spurious association (time vs gender) appears *additive* with the existing spurious association between occupation vs gender.

Figure 5: RoBERTa-large uncertainty metric on all Winogender Schema occupations. Top: Masked pronoun is coreferent with the *professional* (similar to IDs 1-4 in Table 2), so all these sentences are underspecified. Bottom: 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. Setting a threshold at 1.0 for the metric, produces greater than 90% outcomes for both the true-positive classification of the underspecified tasks and true-negative classification of well-specified tasks.

Extending to More General Setting

Generic SCM reproducing selection induced collider bias above and W-injection to the right.

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

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

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

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

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

Demos & code