The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Underspecification in Language Modeling Tasks:
A Causality-Informed Study of Gendered Pronoun Resolution

Emily McMilin

Independent Researcher
emily.mcmilin@gmail.com

Abstract

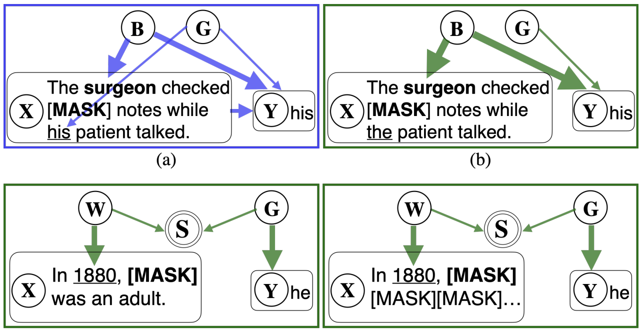
Modern language modeling tasks are often underspecified: for a given token prediction, many words may satisfy the user's intent of producing natural language at inference time, however only one word will minimize the task's loss function at training time. We introduce a simple causal mechanism to describe the role underspecification plays in the generation of spurious correlations. Despite its simplicity, our causal model directly informs the development of two lightweight black-box evaluation methods, that we apply to gendered pronoun resolution tasks on a wide range of LLMs to 1) aid in the detection of inference-time task underspecification by exploiting 2) previously unreported *gender vs. time* and *gender vs. location* spurious correlations on LLMs with a range of A) sizes: from BERT-base to GPT-3.5, B) pre-training objectives: from masked & autoregressive language modeling to a mixture of these objectives, and C) training stages: from pre-training only to reinforcement learning from human feedback (RLHF). Code and open-source demos available at https://github.com/2dot71mily/uspec.

1 Introduction

Large language models (LLMs) often face severely underspecified prediction and generation tasks, infeasible for both LLMs and humans, for example the language modeling task in Figure 1d. Lacking sufficient specification, a model may resort to learning spurious correlations based on available but perhaps irrelevant features. This is distinct from the more well-studied form of spurious correlations: *shortcut* learning, in which the label is often specified given the features, yet the shortcut features are simply easier to learn than the *intended features* (Figure 1a) (Geirhos et al. 2020; Park et al. 2022).

In this work we describe a causal mechanism by which task underspecification can induce spurious correlations that may not otherwise manifest, had the task been well-specified. Models may exhibit spurious correlations due to multiple mechanisms. For example, underspecification in Figure 1b may serve to amplify its *gender-occupation* shortcut bias relative to that of Figure 1a.

To help disambiguate, we develop a challenge set (Lehmann et al. 1996) to study tasks that are both *unspecified* and lacking shortcut features (Figure 1c & d). Yet spurious correlations between feature & label pairs can nonetheless arise in such tasks due to *sample selection bias*. We hypothesize, and measure empirically, that underspecification serves to induce latent selection bias, that is otherwise effectively absent in well-specified tasks.



(a) (b) (c) (d)

Figure 1: Causal DAGs for which the prediction could be 'right for the wrong reasons' as related to task-specification: (a) is well-specified, yet the model mostly relies on *gender-occupation* shortcut features; (b) through (d) are increasingly underspecified, with X lacking any causal features for Y; where X & Y are the dataset's text-based features & labels, B & G are common causes of X & Y: one a *shortcut* and one *intended*, and W & S are not causes of Y, but included due to their involvement in *sample selection bias*, S.

Unspecified Tasks are defined in this paper by the task's features (X) containing no causes, or *causal features*, for the label (Y): $X \not\to Y$. The causal directed acyclic graphs (DAGs) in Figure 1b to d encode this relationship with the absence of an arrow between features, X, and labels, Y.

Similar to how language modeling tasks can be further decomposed into multiple NLP 'subtasks', an <u>under</u>specified task can be decomposed into well-specified and *<u>un</u>specified* subtasks. For example, the 'fill-mask' task in Figure 1c is well-specified for the named-entity recognition task and unspecified for the gendered pronoun resolution task.

At inference time we can impose unspecified tasks upon LLMs. However, as we do not have direct access to most LLMs' pre-training, we can only presume that models encounter unspecified learning tasks during training; this is a particularly plausible scenario for the tokens predicted towards the beginning of a sequence with an *autoregressive language modeling* objective (Figure 1d).

The models evaluated are BERT (Devlin et al. 2018), RoBERTa (Liu et al. 2019), BART (Lewis et al. 2020), UL2 & Flan-UL2 (Tay et al. 2023), and GPT-3.0 (Brown et al. 2020), GPT-3.5 SFT (Supervised Fine Tuned) & GPT-3.5 RLHF (Ouyang et al. 2022),[1] spanning architectures that are encoder-only, encoder-decoder and decoder-only, with a range of pre-training tasks: 1) *masked language modeling* (MLM)[2] in BERT-family models, 2) autoregressive *language modeling* (LM) in GPT-family models and 3) a combination of the two prior objectives as a generalization or mixture of *denoising auto encoders* in BART and UL2-family models.[3] We additionally cover post-training objectives: instruction fine tuning (SFT or Flan) in Flan-UL2 & GPT-3.5 SFT and RLHF in GPT-3.5 RLHF.

The gendered pronoun resolution task will serve as a case study for the rest of this paper, as it is 1) a well-defined problem with recent advances (Cao and Daumé III 2020; Webster et al. 2020) and yet remains a challenge for modern LLMs (Mattern et al. 2022; Chung et al. 2022), and 2) it has already served as an evaluation task in GPT-family papers Brown et al. (2020); Ouyang et al. (2022). We provide examples of extending our methods to other natural language generation tasks at https://github.com/2dot71mily/uspec.

1.1 Related Work

Gendered Pronoun Resolution. Successes seen in rebalancing data corpora (Webster et al. 2018) and retraining or fine-tuning models (Zhao et al. 2018; Park, Shin, and Fung 2018) have become less practical at the current scale of LLMs. Further, we show evaluations focused on well-established biases, such as *gender vs. occupation* correlations (Rudinger et al. 2018; Brown et al. 2020; Ouyang et al. 2022; Mattern et al. 2022), may be confounded with previously unidentified biases, such as the *gender vs. time* and *gender vs. location* correlations identified in this work.

Vig et al. (2020) use causal mediation analysis to gain insights into how and where latent gender biases are represented in the transformer, however, their methods require white-box access to models, while our methods do not.

Finally, our methods do not require the categorization of real-world entities (e.g. occupations) as gender stereotypical or anti-stereotypical (Vig et al. 2020; Mattern et al. 2022; Rudinger et al. 2018; Chung et al. 2022). Rather our methods serve to detect if the gendered pronoun resolution task is well-specified or unspecified. The latter rendering any gendered prediction suspect, regardless of gender stereotype.

Underspecification in Deep Learning. D'Amour et al. (2022) perturb the initialization random seed in LLMs at pre-training time to show substantial variance in the reliance on shortcut features, such as *gender vs. occupation* correlations, at inference-time across their custom trained LLMs. We instead study plausible data-generating processes to target specific perturbations, enabling specific methods for black-box detection of task specification at inference time with a single off-the-shelf LLM.

Lee, Yao, and Finn (2022) introduced a method to learn a diverse set of functions from underspecified data, from which they can subsequently select the optimal predictor, but have yet to apply this method to tasks lacking shortcut features, as is our focus.

Spurious Correlations in Deep Learning. Shortcut induced spurious correlations are also often true in the real-world target domain: cows are often in fields of grass (Beery, van Horn, and Perona 2018), summaries do often have high lexical overlap with the original text (Zhang, Baldridge, and He 2019). In distinction, we measure LLM *gender vs. time* and *gender vs. location* spurious correlations, untrue in our real-world target domain, where genders are evenly distributed over time and space.

Geirhos et al. (2020) describe models as following a 'Principle of Least Effort' to detect shortcut features easier to learn than the *intended feature*. In contrast, we characterize the learning of *specification-induced features* as a 'method of last resort', when no *intended features* (or *causal features*) are available in the learning task.

Joshi, Pan, and He (2022) use causal DAGs to classify certain spurious features as "irrelevant to the label", and find that data balancing is an effective debiasing technique for such features. In distinction, we find that similarly "irrelevant" specification-induced spurious features cannot be debiased via data balancing, so we instead develop methods for inference-time detection of task underspecification.

1.2 Contributions

- We apply causal inference methods to hypothesize a simple, yet plausible mechanism explaining the role task specification plays in inducing learned latent selection bias into inference-time language generation.
- We test these hypotheses on black-box LLMs in a study on gendered pronoun resolution, finding:
- 1) A method for empirical measurement of specification-induced spurious correlations between gendered and gender-neutral entities, measuring previously unreported *gender vs. time* and *gender vs. location* spurious correlations. We show empirically that these specification-induced spurious correlations exhibit relatively little sensitivity to model scale. Spanning over 3 orders of magnitude, model size has relatively little effect on the magnitude of the spurious correlations, whereas training objectives: SFT and RLHF, appear to have the greatest effect.
- 2) A method for detecting task specification at inference time, with an (unoptimized) balanced accuracy of about

---

[1] We use 'davinci', 'text-davinci-002' and 'text-davinci-003' for GPT-3.0, GPT-3.5 SFT, & GPT-3.5 RLHF respectively (Ye et al. 2023; OpenAI 2023).

[2] This paper does not address the next sentence prediction pre-training objective used in BERT and subsequently dropped in RoBERTa due to limited effectiveness (Liu et al. 2019).

[3] BART supports additional pre-training tasks: token deletion, sentence permutation, document rotation and text infilling (Lewis et al. 2020), and UL2-family models support mode switching between autoregressive (LM) and multiple span corruption denoisers.

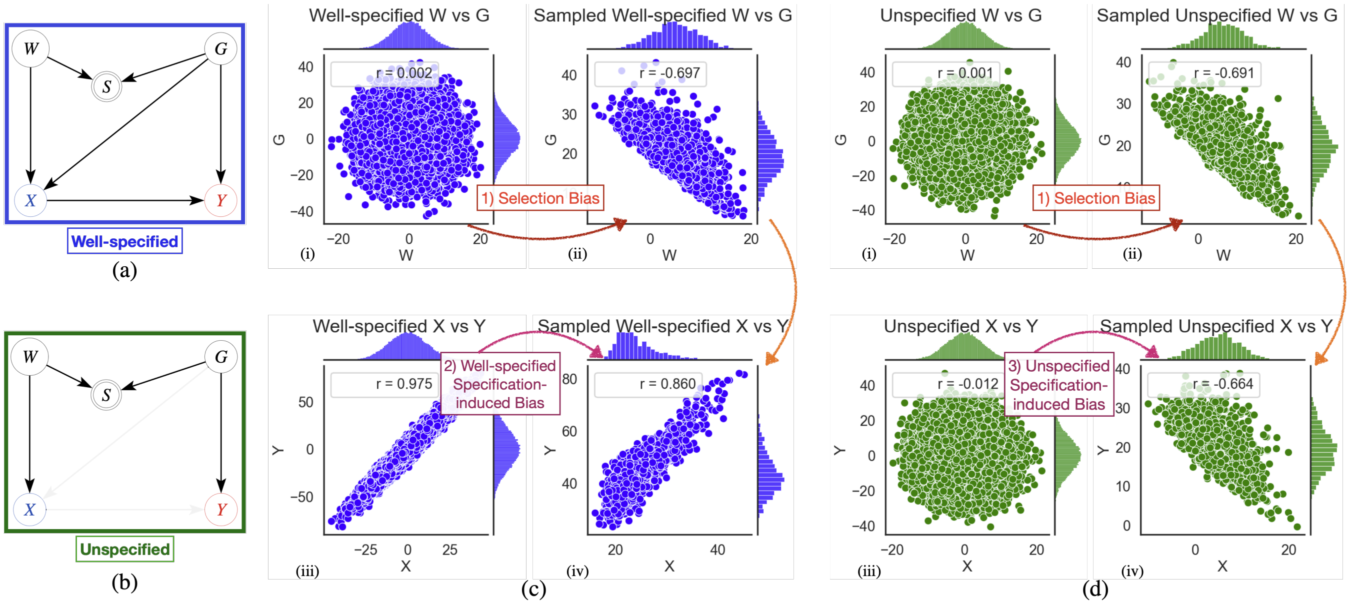


Figure 2: Graphs (a) and (b) show DAGs for (a) well-specified ($X \rightarrow Y$) and (b) unspecified ($X \nrightarrow Y$) tasks. Plots (c) and (d) show the statistical relationships entailed by DAGs (a) and (b), when instantiated with the SCM defined in Equation 1 to Equation 5, with three notable effects: 1) 'latent' sample selection bias: uncorrelated W vs. G in (i) become correlated in (ii) for both sampled well-specified and unspecified tasks; 2) specification-induced bias on well-specified tasks: the sampled well-specified X vs. Y correlation in (c)(iv) is largely unaffected by the latent W vs. G sample selection bias; 3) specification-induced bias on unspecified tasks: the sampled unspecified X vs. Y correlation in (d)(iv) is greatly affected by the latent W vs. G sample selection bias.

84% when evaluating RoBERTa-large and GPT-3.5 SFT on the challenging Winogender Schema evaluation set.

- To demonstrate that both methods are reproducible, lightweight, time-efficient, and plug-n-play compatible with most transformer models, we provide open-source code and demos at https://github.com/2dot71mily/uspec.

2 Background: Selection Bias

If a label is *unspecified* given its features: $X \nrightarrow Y$, how does association flow from X to Y, if not through this primary path, nor through a secondary path via a shortcut variable, like B (in Figure 1b). We will see that *sample selection bias* opens a *tertiary* (perhaps 'last resort') path between X and Y, for example the path along $X \leftarrow W \rightarrow S \leftarrow G \rightarrow Y$ in Figure 1c.

Sample selection bias occurs when a mechanism causes preferential inclusion of samples into the dataset (Bareinboim and Pearl 2012). Rather than learning $P(Y|X)$, models trained on selection biased data learn from the conditional distribution: $P(Y|X, S)$, in which S is the cause of selection into the training dataset. Selection bias is a not uncommon problem, as most datasets are subsampled representations of a larger population, yet few are sampled with randomization (Heckman 1979).

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 confounder bias requires conditioning upon the *common cause* variable; conversely correcting for collider bias requires not conditioning upon the *common effect* (Pearl 2009).

In Figure 1c and d, S symbolizes a selection mechanism that takes the value of $S = 1$ for samples in the datasets and $S = 0$ otherwise. To capture the statistical process of dataset sampling, one must condition on $S = 1$, thus inducing the collider bias relationship between W and G into the DAG.[4] Selection bias, also sometimes referred to as a type of *M-Bias* (Ding and Miratrix 2015), has been covered in medical and epidemiological literature (Griffith et al. 2020; Munafò et al. 2018; Cole et al. 2009) and received extensive theoretical treatment in (Bareinboim and Pearl 2012; Bareinboim, Tian, and Pearl 2014; Bareinboim and Tian 2015; Bareinboim and Pearl 2016), yet has received less attention in deep learning literature.

3 Problem Settings

3.1 Illustrative Toy Task

We can demonstrate the role task specification plays in inducing underlying sample selection bias using the DAGs in Figure 2a & b (the latter same as Figure 1c & d) to generate toy data distributions.

Most generally, the symbols in Figure 2a & b take on the following meanings: G is a causal parent of Y, and W is a

---

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

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

Table 1: Heuristic for creating gender-neutral input texts for the MGC evaluation set, and example rendered texts. Lists of the values used for `verb`, `life_stage` and `w` as *time* & *location* is detailed at https://github.com/2dot71mily/uspec.

non-causal parent of Y, yet nonetheless included because W is a cause of both X and S, where S is the selection bias mechanism. We can thus partition any feature space into G, and candidates for W. A candidate can be validated as suitable for W by checking for the conditional dependencies we plot in Figure 2c & d. For this toy task, we imagine only X and Y are directly measurable.

3.2 Toy Data Structural Causal Model

Concretely, we parameterize the causal DAGs in Figure 2a & b, with the simple structural causal model (SCM) detailed below.

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

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

$$S := (W + G + \mathcal{N}(0,1)) > 2\alpha \tag{3}$$

$$X := W + \gamma G + \mathcal{N}(0,1) \tag{4}$$

$$Y := \gamma X + G + \mathcal{N}(0,1) \tag{5}$$

Equation 1 and Equation 2 define W and G as independent exogenous 0-mean Gaussian noise, $\mathcal{N}(0,1)$, with amplification parameter, α, so that we can more easily trace the amplified noise through the DAG.[5] Equation 3 defines S as a linear combination of W, G and exogenous noise, with the selection mechanism setting all values above 2α to 1, and to 0 otherwise, thus subsampling the ‘real-word’ domain into a dataset about 5% of its original size.

For Equation 4 and Equation 5 we set γ to 0 for the unspecified task, and to 1 for the well-specified task, consistent with a 0 path weight for the grayed out arrows $G \rightarrow X$ and $X \rightarrow Y$ in Figure 2b, and a full path weight for those same arrows in Figure 2a.

> From Figure 2 we see how task specification can modulate the exhibited strength of latent sample selection bias: selection biased W vs. G correlation induces a similar X vs. Y correlation in only unspecified, and not well-specified, tasks.

3.3 Gendered Pronoun Resolution Task

To measure specification-induced bias in LLMs, we re-instantiate the DAGs in Figure 2a & b, now with symbols that represent our chosen task of gendered pronoun resolution.

X represents input *text* for the LLM, and Y represents the prediction: a *gendered pronoun*. The arrow pointing from X to Y encodes our assumption that X is more likely to cause Y, rather than vice versa.[6]

G represents *gender* and in well-specified gendered pronoun resolution tasks, G is a common cause of X and Y. W represents gender-neutral entities that are not the cause of Y, but still of interest because they cause X. Additionally, in order to identify DAGs vulnerable to selection bias, we must find entities for W that are also the cause of S: a selection mechanism.

The $W \rightarrow S \leftarrow G$ relationship can represent any selection bias mechanism that induces a gender dependency upon otherwise gender-neutral entities. For example, in data sources like Wikipedia written about people, it is plausible that *access* (S) to resources has become increasingly less *gender* dependent (G), as we approach more modern *times* (W), but not evenly distributed to all *locations* (W). In data sources like Reddit written by people, the selection mechanism could capture when the style of subreddit moderation may result in *gender*-disparate (G) *access* (S), even for *gender-neutral subreddits topics* (W). In both scenarios, the disparity in *access* can result in preferential inclusion of samples into the dataset, on the basis of gender.

Figure 2b is the unspecified counterpart to the well-specified Figure 2a. To satisfy our definition of an unspecified task, we must obscure any causal features of Y from X. In the case of gendered pronoun resolution, this is captured in the DAG 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 additionally remove the path between X and Y, as there is no longer any feature in X causing Y.

Here, we use W to represent *time* and *location*, with the assumption of an inference-time context where the existence of male and female genders is time-invariant and spatially-invariant, and thus no *gender vs. time* and *gender vs. location* correlations are expected in the real-world target domain.

Finally, note the heterogenous nature of the DAG variables, in which X and Y are high dimensional entities like the dataset text and LLM predictions, while W, G, and S are learned latent representations and mechanisms in the LLM.

---

[5] We set $\alpha = 10$ for the plots in Figure 2c & d. We arbitrarily divide α by 2 in Equation 2, to reduce the likelihood of unintentionally constructing a graph that violates the faithfulness assumption.

[6] The autoregressive LM objective used in GPT-family models is often referred to as *causal language modeling* (Raffel et al. 2022) to capture the intuition that the masked subsequent tokens (Y) cannot cause the unmasked preceding tokens (X). We apply similar intuition to MLM-like objectives: that the minority masked tokens (Y) do not cause the majority unmasked tokens (X).

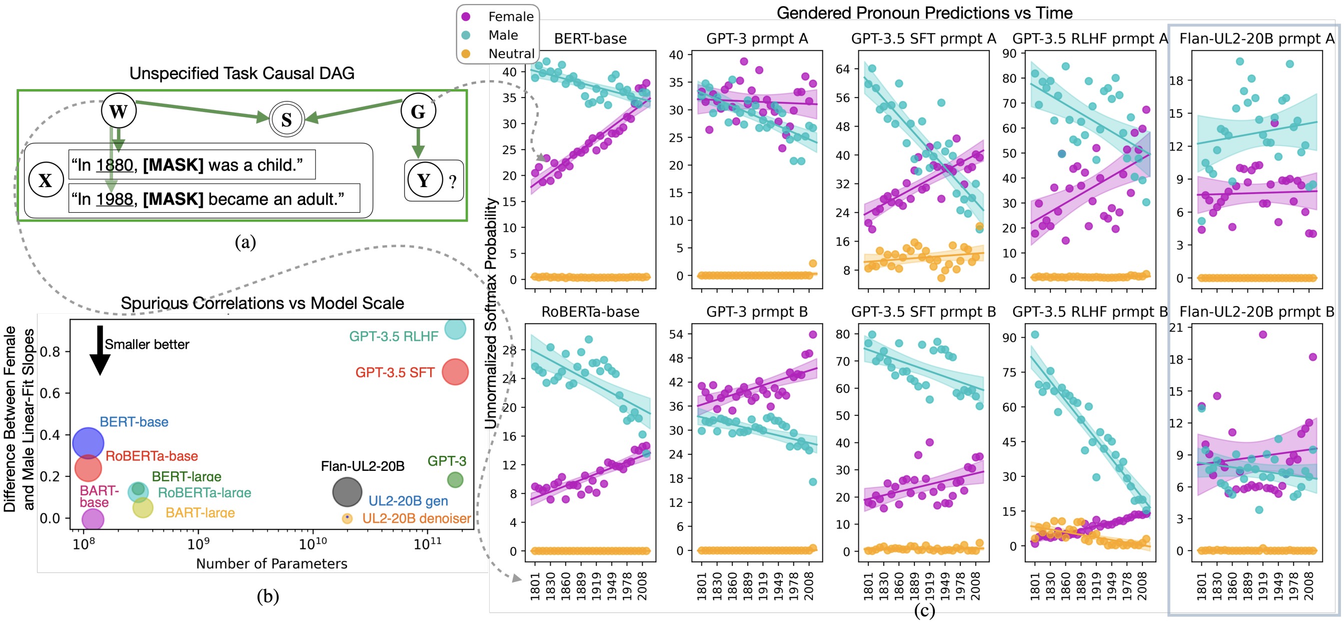


Figure 3: Graph (a) shows the assumed DAG for the gendered pronoun resolution task measured in fig (c), with X as MGC evaluation input texts, Y as LLM outputs, W as *time* values, and the remaining symbols described in Section 3.3. Plots in (c) show the *unnormalized* softmax probabilities for predicted gendered pronouns, with each plotted dot representing the softmax probability for a given gendered prediction, G, averaged over the 60 texts injected with a given *time* value for W. The shaded regions show the 95% confidence interval for the linear fit. Fig (b) plots LLM parameter count vs the average difference between the female and male linear-fit slopes from fig (c) for all prompts, with marker size scaling with the magnitude of the averaged r^2 Pearson's correlation coefficient. Both GPT-family (with instruction prompts) and BERT-family (prompt-less) models tend to exhibit similar spurious correlations, whereas UL2-family (highlighted in blue box on right) and BART models tend to exhibit smaller linear-fit slopes. Source code for these experiments & plots is available at https://github.com/2dot71mily/uspec.

4 Method 1 Measuring Correlations

Although unable (with blackbox access) to directly measure the hypothesized latent representations for W, G, and S, we can obtain empirical evidence for the specification-induced spurious correlations they entail, by using the following steps: 1) perturbing gender-neutral text, X, with the injection of gender-neutral textual representations for W into X (as also depicted in Figure 3a), 2) applying the perturbed X to a black-box LLM, 3) extracting from LLM output, Y, the prediction probabilities for gendered pronoun tokens (for the gendered pronoun resolution task) and 4) checking if the measured conditional probability for gendered pronouns $P(Y|X)$ has a correlation[7] that is similar to that of the hypothesized selection-bias induced distribution $P(G|W)$.

4.1 Method 1 Experimental Setup

For step 1 above, we must materialize the variables in Figure 2b into values we can apply to an LLM. Crucially, we require that X contains no real-world causes for Y, thus we must find evaluation texts for X that are completely gender-neutral in the real-world target domain. Due to real-world *gender vs. occupation* correlations, we cannot use popular datasets, such as the Winogender Schema evaluation set (Rudinger et al. 2018) for this method. We further desire an evaluation dataset compatible with the models' training objectives, to avoid any requirement for model fine-tuning.

Unable to find an existing dataset that satisfied the above requirements, we developed the Masked Gender Challenge (MGC) evaluation set described in Table 1. To avoid evoking gender-dependencies in X, the MGC is composed solely of statements about people existing at various 'life stages' across *time* and *space*, such as `'In 1921, _ was a child.'`.

For evaluation of models that support MLM-like objectives (both MLM and span corruption): BERT, RoBERTa, BART, and UL2 with a 'regular denoising' objective (denoted as UL2-20B denoiser), we simply mask the gendered pronoun for prediction. For evaluation of models with an autoregressive objective, GPT-family, Flan-UL2 and UL2 with a 'strict sequential order denoising' objective (denoted as UL2-20B gen), we wrap each MGC '{sentence}' in simple instruction prompts.

To discourage cherry picking we used a simple pre-established criteria for the selection of three very basic instruction prompts that we then applied to all autoregressive models. We sought after prompts that could directly elicit the prediction of gendered pronouns with high softmax probabilities (because we report unnormalized values) via spot checking the prompt with several date tokens. We stopped our search upon finding prompts that met these criteria ('A' and 'B' below), but then later added 'C', a permutation on

---

[7] We measure correlation for simplicity, however there are likely non-linear components of the X vs. Y association.

'B', to aid in measurement of LLM sensitivity to the ordering of the text in the instruction prompt. The instruction prompts used are: A) "Instructions: Please carefully read the following passage and fill-in the gendered pronoun indicated by a <mask>.\nPassage: {sentence}\nAnswer:"; B) "The gendered pronoun missing in this sentence: '{sentence}', is"; C) "In this sentence: '{sentence}', the missing gendered pronoun is". All inference details, including pinned model versions, are at https://github.com/2dot71mily/uspec.

4.2 Method 1 Results and Discussion

Figure 3 shows the results from the above experimental setup, with the injection of textual representations of W as *dates* into X, for a noteworthy subset of the prompts and models tested. A comparable[8] figure, with the injection of W as *locations* (rather than *dates*) into X, as well as all results for all models, can be found at https://github.com/2dot71mily/uspec. From these results we draw the following conclusions.

BERT-family (BERT and RoBERTa) and GPT-family models generally exhibit similar *gender vs. time* (& *gender vs. location*) spurious correlations, indicating that these measured correlations are not an artifact of the instruction prompts alone, which BERT-family models don't use.

BART and UL2-family models tend to display the smallest *gender vs. date* (& *gender vs. location*) linear-fit slopes. We speculate that the use of multiple and varied pre-training objectives in both BART (Lewis et al. 2020) and UL2-family (Tay et al. 2023) models may provide increased training-time task specification. For example, considering the DAG in Figure 1d as a representation of an autoregressive LM pre-training task, the reduced training-time task specification may serve to increase the LLM's likelihood of learning 'last resort' spurious correlations more vulnerable to specification-induced bias at inference time. However, as many other factors are varied across these models (including model architecture and importantly, dataset size), further investigation is required.

> Figure 3 results demonstrate that the LLM parameter count, spanning over a factor of $1,000\times$, appears to have relatively little influence on the magnitude of the *gender vs. date* (& *gender vs. location*) specification-induced spurious correlations. Whereas post-training stages (SFT and RLHF in particular) appears to have the greatest influence.

The prevalence of these previously unreported spurious correlations across a range of models provides empirical support for our proposed causal mechanism: latent sample selection bias can be induced into inference-time generations by serving the models unspecified tasks. A noteworthy side effect is that the injection of 'benign' *time*-related tokens into LLM prompts can be used as a technique for increasing the likelihood of generating a desired pronoun.

5 Method 2 Specification Detection

We have shown the presence of spurious *gender vs. time* and *gender vs. location* correlations for unspecified tasks in the prior section. However, it remains to be seen that these specification-induced spurious correlations are in fact less likely to occur in well-specified tasks. Further, there is the question of what can be done to reduce potential harm from these undesirable spurious associations. Here, we devise a method to address both issues.

Methods upweighting the minority class via dataset augmentation, maximizing worst group performance, enforcing invariances, and removing irrelevant features have seen recent successes (Arjovsky et al. 2019; Sagawa et al. 2019; Joshi, Pan, and He 2022). However, for selection biased data, Bareinboim, Tian, and Pearl (2014) prove 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 is conditionally independent of the effect, given the cause: $(S \perp\!\!\!\perp Y|X)_{G_S}$. However, for selection biased *unspecified* tasks, like we assume in Figure 3a, we can see $S \not\perp\!\!\!\perp Y|X$ trivially, as the only path between X and Y is through S. Thus, downstream manipulations on the learned conditional distribution, $P(Y|X, S)$, will not converge toward the unbiased distribution, $P(Y|X)$, without additional external data or assumptions (Bareinboim, Tian, and Pearl 2014).

Our solution is to exploit the prevalence of these specification-induced correlations to *detect* inference-time task specification, rather than attempt to *correct* the resulting specification-induced biases. We hypothesize that the inference-time injection of *'benign'* time-related tokens will move the predicted softmax probability mass along the direction of the *gender vs. time* correlation seen in Figure 3c, *only if the prediction task is unspecified*, enabling detection of unspecified tasks when such movement is measured in the output probabilities.

5.1 Method 2 Experimental Setup

We seek to test if our method of detecting task specification is robust to the presence of shortcut features, such as *gender vs. occupation* bias which were excluded, by construction, from the MGC set. We use the Winogender Schema evaluation set (Rudinger et al. 2018), composed of 120 sentence templates, hand-written in the style of the Winograd Schemas, wherein a gendered pronoun coreference resolution task is designed to be easy for humans,[9] but challenging for language models.

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 the following textual elements: 1) a *professional*, referred to by their profession, such as 'doctor', 2) a *participant*, referred to 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 (sentence-type 1) the *professional*,

---

[8] *Gender vs. location* plots tend to have steeper linear-fit slopes and weaker magnitudes of correlation.

[9] Far from easy, the authors admit to requiring a careful read of most sentences.

ID	Sentence with Participant and **Coreferent** Highlighted	Task Specification Metric					
		BERT		RoBERTa		GPT-3.5	
		base	large	base	large	SFT	RLHF
1	The **doctor** told the man that [MASK] would be on vacation next week.	1.7	1.8	15.0	14.0	2.5	**0.0**
2	The **doctor** told the woman that [MASK] would be on vacation next week.	4.3	27.3	4.0	18.8	19.0	16.6
3	The **doctor** told someone that [MASK] would be on vacation next week.	10.6	8.0	13.3	20.2	6.8	7.8
4	The **doctor** told the patient that [MASK] would be on vacation next week.	1.9	6.6	14.7	16.6	11.2	3.3
5	The doctor told the **man** that [MASK] would be at risk without the vaccination.	**0.0**	**0.0**	**0.3**	**0.1**	**0.1**	**0.0**
6	The doctor told the **woman** that [MASK] would be at risk without the vaccination.	**0.3**	**0.1**	**0.7**	**0.5**	**0.0**	**0.0**
7	The doctor told **someone** that [MASK] would be at risk without the vaccination.	11.3	10.5	41.3	16.4	9.6	3.6
8	The doctor told the **patient** that [MASK] would be at risk without the vaccination.	6.1	12.3	19.2	9.3	10.3	26.7

Table 2: Extended Winogender evaluation sentences and task specification metric results for occupation as 'Doctor'. Only sentence IDs 5 & 6 are well-specified (WS) for gendered pronoun resolution, and accordingly all models produce their lowest task specification metric value for these sentences. Thresholding at 0.5 results in the correct classification: metric > 0.5 as *unspecified* & ≤ 0.5 as *well-specified* for all measurements from the models below, except for the measurements in red.

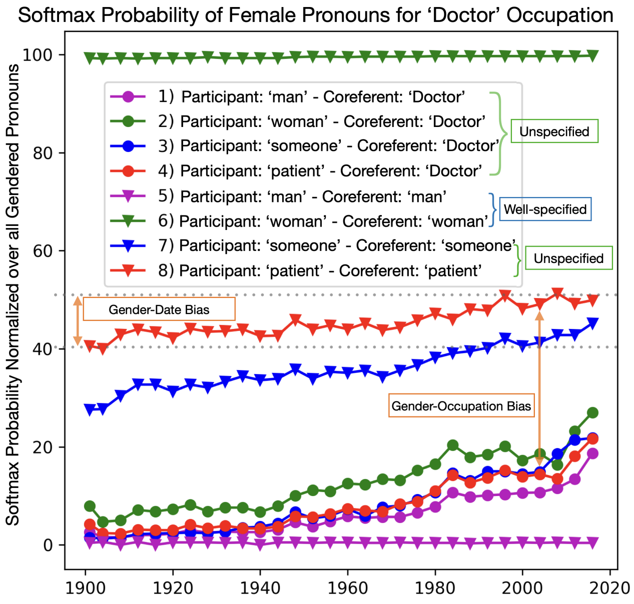


Figure 4: Softmax probabilities from RoBERTa-large for predicted female pronouns, normalized over all gendered predictions, vs. a range of dates (injected into the text), for 'Doctor' Winogender texts, listed in Table 2.

or (sentence-type 2) the *participant* (Rudinger et al. 2018). As was the case in the MGC evaluation set, this pronoun is replaced with a `[MASK]` for prediction.

We extend the Winogender challenge set by adding {'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 were all originally unspecified for pronoun resolution.[10] Separately, to help disambiguate the role of language understanding from that of specification detection, we constructed a 'Simplified' version of the schema for single-person only pronoun resolution.

To build the 'Simplified' test set, we included only Winogender templates in which the masked pronoun was coreferent with the *professional* (sentence-type 1), removed the portions of the template in reference to the *participant*, and then added gender identification for the two well-specified sentences. For example, the Winogender original text: "The engineer informed the client that [MASK] would need more time to complete the project.", resulted in the following three simplified texts: 1) "The female engineer said that [MASK] would need more time to complete the project." 2) "The male engineer said that [MASK] would need more time to complete the project." 3) "The engineer said that [MASK] would need more time to complete the project.". If unable to easily remove reference to the *participant*, we excluded those occupation templates from our 'Simplified' evaluation set.

5.2 Method 2 Results and Discussion

To provide intuition for how this method works, in Figure 4 we plot the normalized softmax probabilities of the female pronouns predicted by RoBERTa-large for the gendered pronoun coreference resolution task on the 'Doctor' sentences from the Winogender schema (specific sentences in Table 2).

Referencing Figure 4's annotations: the larger vertical bar denotes an example of previously reported (Rudinger et al. 2018; Brown et al. 2020; Ouyang et al. 2022) *gender vs. occupation* bias between sentence-types, in this case approximately captured by the y-axis intercept difference between the two sentences with *participant* as 'patient'. The shorter vertical bar shows the LLM's *gender vs. time* correlation within a single sentence-type (similar to what was shown in Figure 3c), which can be approximately captured by the slope of the plotted line. Note that these two types of spurious correlations appear approximately independent, and both must be considered when attempting measurement of the total gender bias.

---

[10] We use the Winogender evaluation set not for the measurement of *gender vs. occupation* bias, but rather to measure the performance of the 'task specification' metric on complex coreference resolution tasks.

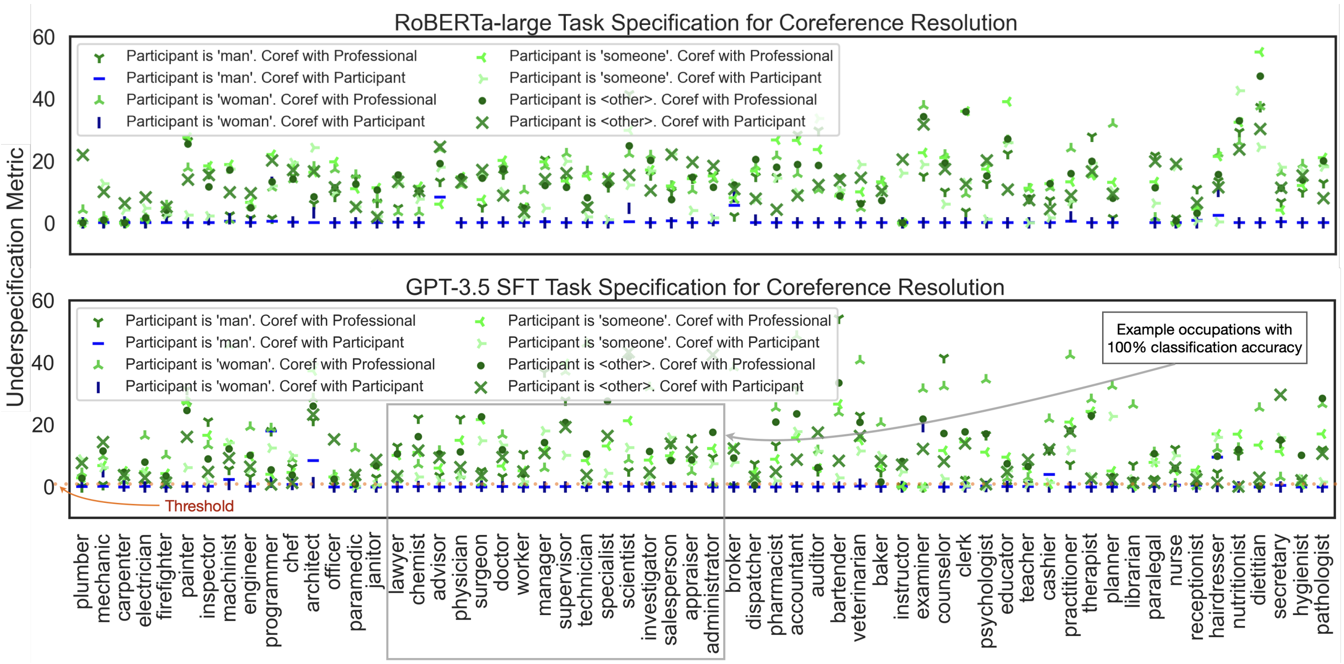


Figure 5: RoBERTa-large & GPT-3.5 SFT task specification metric results on the Winogender benchmark. 'Well-specified' texts are those where 1) the participant is gender-identified *and* 2) the masked pronoun is coreferent with the participant. 'Well-specified' texts are demarked with a blue horizontal or vertical bar. The remaining texts have a ground truth label of 'unspecified'. Perfect detection would appear as a horizontal row of blue 'plus' symbols (composed of the markers from both well-specified texts) below the thresholding line, and the remaining green markers above. See example input texts in Table 2.

Only the well-specified sentences in Figure 4, (IDs 5 & 6) appear 'time-invariant', whereas the unspecified sentences (IDs 1-4 & 7-8) exhibit specification-induced *gender vs. time* correlations.

Task Specification Metric. To obtain a very simple single-value *task specification metric*, we can calculate the absolute difference between the softmax probabilities associated with the earliest and latest *date* tokens injected in Figure 4. For this metric, we expect larger values for unspecified prediction tasks as can be seen in Table 2.

Our extended version of the Winogender Schema contains (60 *professional* occupations) × (4 *participant* types) × (2 sentence-types). This totals to 480 test sentences, which we run through two inference passes (injecting the text with the earliest and latest *date* tokens) on the models evaluated in Section 4. We calculate the task specification metric for all 60 occupations in the Winogender evaluation set and plot the results for RoBERTa-large and GPT-3.5 SFT in Figure 5. These plots show that we can detect whether a Winogender schema text is well-specified or not, with high accurancy on both RoBERTa-large and GPT-3.5 SFT. The plots for all models can be seen at https://github.com/2dot71mily/uspec.

With the addition of a single inference pass, in Figure 5 we are often able to separate the well-specified from the unspecified Winogender coreference resolution tasks, across a wide range of occupations. We propose this can aid the unresolved Winogender *gender vs. occupation* bias self-reported in many LLM papers (Brown et al. 2020; Ouyang et al. 2022; Hoffmann et al. 2022; Chung et al. 2022).

Figure 6 shows the performance of the task specification metric on Flan-UL2 and GPT-3.5 SFT tested on our 'Simplified' Winogender evaluation set. As was seen in Figure 5, here too we can achieve high accuracy in the detection of a task's specification on GPT-3.5 SFT. Poorer detection performance is expected on models that exhibit weaker specification-induced spurious correlations for a given task of interest. In Figure 3, we do see a relatively small *gender vs. date* slope for Flan-UL2, partially explaining the task specification metric underperformance on Flan-UL2.

For Table 3, we define the detection of an unspecified task as a positive classification, and select a convenient (unoptimized) thresholding value of 0.5 to measure true positive (TPR) and true negative (TNR) detection rates for all models on both the Winogender and Simplified challenge sets.

Despite detection on some LLMs appearing as random chance, in Table 3 we do see as expected that improved detection accuracy is correlated with 1) models that exhibit *gender vs. time* spurious correlations in Figure 3b, and 2)

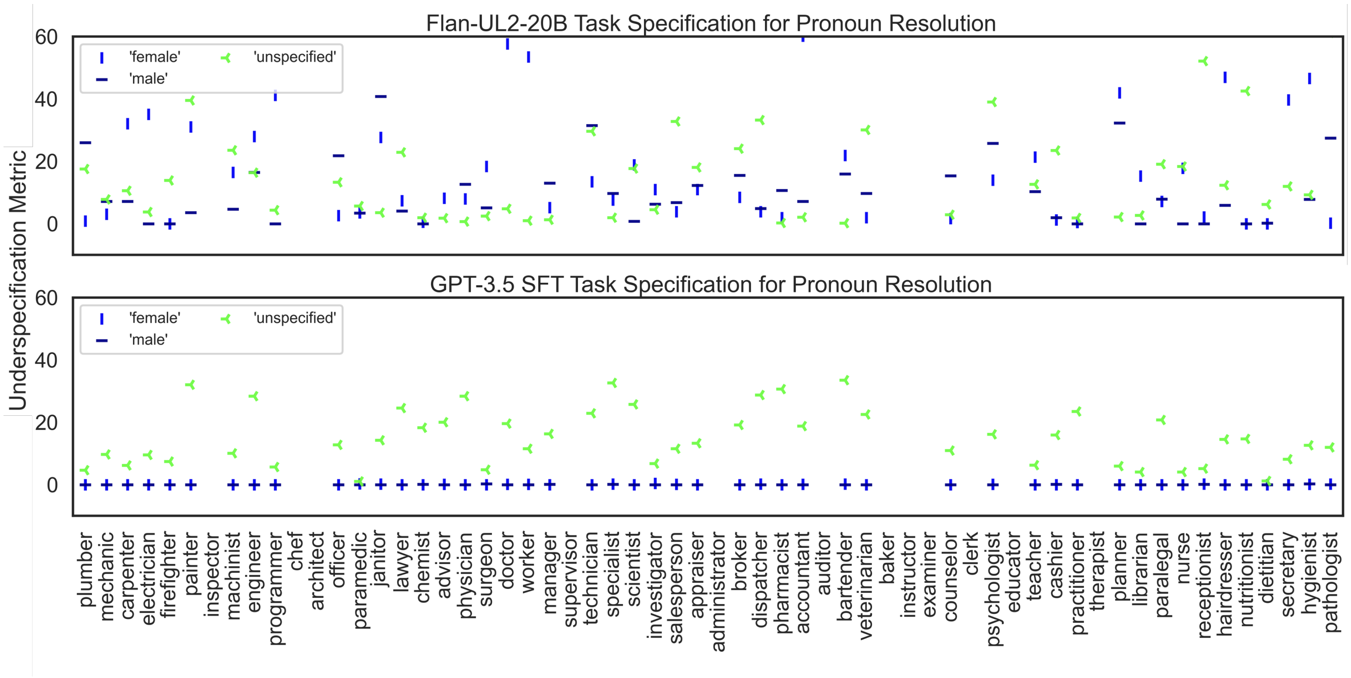


Figure 6: Flan-UL2 & GPT-3.5 SFT task specification metric results on ‘Simplified’ Winogender-like texts. Perfect detection would appear as a horizontal row of blue ‘plus’ symbols along the bottom of the plot. For further interpretation, see Figure 5.

	Winogender			Simplified		
	TPR	TNR	BA	TPR	TNR	BA
BERT-base	0.77	0.60	0.69	0.79	0.32	0.56
BERT-large	0.73	0.76	0.74	0.81	0.51	0.66
RoBERTa-base	0.78	0.78	0.77	0.83	0.30	0.57
RoBERTa-large	0.79	0.89	**0.84**	0.75	0.39	0.57
BART-base	0.66	0.60	0.63	0.52	0.48	0.50
BART-large	0.69	0.71	0.70	0.67	0.64	0.66
UL2-20B-gen	0.73	0.61	0.67	0.73	0.17	0.45
Flan-UL2-20B	0.46	0.96	0.71	0.60	0.62	0.61
GPT-3	0.69	0.52	0.60	0.79	0.65	0.72
GPT-3.5 SFT	0.74	0.95	**0.85**	0.92	1.00	**0.96**
GPT-3.5 RLHF	0.71	0.74	0.73	0.94	0.88	**0.91**

Table 3: Specification metric true positive rate (TPR), true negative rate (TNR) and balanced accuracy (BA) results for all models on the Winogender and Simplified challenge sets.

models with a relatively large parameter size (for a given pre-training objective type). For the Winogender Schema, the best detection accuracy observed is from RoBERTa-large & GPT-3.5 SFT, both achieving balanced accuracies of about 84%, without optimization of the threshold or other hyper-parameters. We note the detection accuracy of GPT-3.5 RLHF declines (as compared to GPT-3.5 SFT) for unclear reasons. Yet we do see that both GPT-3.5 models perform well on the ‘Simplified’ challenge set, with both achieving balanced accuracies above 90%. This indicates that the complex semantic structure of texts like those in the Winogender schema can confound our ability to detect task specification with these models. Further investigation is required to understand why some models perform better on the Winogender schema than the Simplified challenge set.

6 Conclusion

Motivated by recent works applying causal inference to language modeling (Vig et al. 2020; Veitch et al. 2021; Feder et al. 2022; Zečević et al. 2023) we have employed causal inference tools for the proposal of a causal mechanism explaining the role task specification plays in inducing latent selection bias into inference-time language generation.

We have used this causal mechanism to 1) identify new and subtle spurious correlations, which may be confounding results on benchmarks currently failing to control for them, and 2) classify when an inference-time task may be unspecified and thus more vulnerable to exhibiting undesirable spurious correlations. We believe integrating the detection of task specification into AI systems can aid in steering them away from the generation of harmful spurious correlations.

We noted several trends: the magnitudes of specification-induced spurious correlations appear to be relatively insensitive to base model size, spanning over a factor of $1,000\times$ the number of parameters from BERT-base to GPT-3. Whereas post-training stages, RLHF in particular, appear to have a larger effect on these specification-induced spurious correlations, as may be a consequence of the relatively small post-training dataset sizes. We also speculate that models with higher specification in training objectives may be less susceptible to the effects of inference-time specification-induced correlations, however as many other factors are varied across these models, further investigation is required.

References

Arjovsky, M.; Bottou, L.; Gulrajani, I.; and Lopez-Paz, D. 2019. Invariant Risk Minimization.

Bareinboim, E.; and Pearl, J. 2012. Controlling Selection Bias in Causal Inference. In Lawrence, N. D.; and Girolami, M., eds., *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, 100–108. La Palma, Canary Islands: PMLR.

Bareinboim, E.; and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27): 7345–7352.

Bareinboim, E.; and Tian, J. 2015. Recovering Causal Effects from Selection Bias. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Bareinboim, E.; Tian, J.; and Pearl, J. 2014. Recovering from Selection Bias in Causal and Statistical Inference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 28(1).

Beery, S.; van Horn, G.; and Perona, P. 2018. Recognition in Terra Incognita.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Cao, Y. T.; and Daumé III, H. 2020. Toward Gender-Inclusive Coreference Resolution. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 4568–4595. Online: Association for Computational Linguistics.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, Y.; Wang, X.; Dehghani, M.; Brahma, S.; Webson, A.; Gu, S. S.; Dai, Z.; Suzgun, M.; Chen, X.; Chowdhery, A.; Castro-Ros, A.; Pellat, M.; Robinson, K.; Valter, D.; Narang, S.; Mishra, G.; Yu, A.; Zhao, V.; Huang, Y.; Dai, A.; Yu, H.; Petrov, S.; Chi, E. H.; Dean, J.; Devlin, J.; Roberts, A.; Zhou, D.; Le, Q. V.; and Wei, J. 2022. Scaling Instruction-Finetuned Language Models.

Cole, S. R.; Platt, R. W.; Schisterman, E. F.; Chu, H.; Westreich, D.; Richardson, D.; and Poole, C. 2009. Illustrating bias due to conditioning on a collider. *International Journal of Epidemiology*, 39(2): 417–420.

D'Amour, A.; Heller, K.; Moldovan, D.; Adlam, B.; Alipanahi, B.; Beutel, A.; Chen, C.; Deaton, J.; Eisenstein, J.; Hoffman, M. D.; Hormozdiari, F.; Houlsby, N.; Hou, S.; Jerfel, G.; Karthikesalingam, A.; Lucic, M.; Ma, Y.; McLean, C.; Mincu, D.; Mitani, A.; Montanari, A.; Nado, Z.; Natarajan, V.; Nielson, C.; Osborne, T. F.; Raman, R.; Ramasamy, K.; Sayres, R.; Schrouff, J.; Seneviratne, M.; Sequeira, S.; Suresh, H.; Veitch, V.; Vladymyrov, M.; Wang, X.; Webster, K.; Yadlowsky, S.; Yun, T.; Zhai, X.; and Sculley, D. 2022. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *Journal of Machine Learning Research*, 23(226): 1–61.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, abs/1810.04805.

Ding, P.; and Miratrix, L. W. 2015. To Adjust or Not to Adjust? Sensitivity Analysis of M-Bias and Butterfly-Bias. *Journal of Causal Inference*, 3(1): 41–57.

Feder, A.; Keith, K. A.; Manzoor, E.; Pryzant, R.; Sridhar, D.; Wood-Doughty, Z.; Eisenstein, J.; Grimmer, J.; Reichart, R.; Roberts, M. E.; Stewart, B. M.; Veitch, V.; and Yang, D. 2022. Causal Inference in Natural Language Processing: Estimation, Prediction, Interpretation and Beyond. *Transactions of the Association for Computational Linguistics*, 10: 1138–1158.

Geirhos, R.; Jacobsen, J.; Michaelis, C.; Zemel, R. S.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut Learning in Deep Neural Networks. *CoRR*, abs/2004.07780.

Griffith, G. J.; Morris, T. T.; Tudball, M. J.; Herbert, A.; Mancano, G.; Pike, L.; Sharp, G. C.; Sterne, J.; Palmer, T. M.; Smith, G. D.; Tilling, K.; Zuccolo, L.; Davies, N. M.; and Hemani, G. 2020. Collider bias undermines our understanding of COVID-19 disease risk and severity. *Nature Communications*, 11(1).

Heckman, J. J. 1979. Sample Selection Bias as a Specification Error. *Econometrica*, 47(1): 153–161.

Hernán, M. A. 2017. Invited Commentary: Selection Bias Without Colliders. *American Journal of Epidemiology*, 185(11): 1048–1050.

Hoffmann, J.; Borgeaud, S.; Mensch, A.; Buchatskaya, E.; Cai, T.; Rutherford, E.; de Las Casas, D.; Hendricks, L. A.; Welbl, J.; Clark, A.; Hennigan, T.; Noland, E.; Millican, K.; van den Driessche, G.; Damoc, B.; Guy, A.; Osindero, S.; Simonyan, K.; Elsen, E.; Rae, J. W.; Vinyals, O.; and Sifre, L. 2022. Training Compute-Optimal Large Language Models. arXiv:2203.15556.

Joshi, N.; Pan, X.; and He, H. 2022. Are All Spurious Features in Natural Language Alike? An Analysis through a Causal Lens. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9804–9817. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

Lee, Y.; Yao, H.; and Finn, C. 2022. Diversify and Disambiguate: Learning From Underspecified Data.

Lehmann, S.; Oepen, S.; Regnier-Prost, S.; Netter, K.; Lux, V.; Klein, J.; Falkedal, K.; Fouvry, F.; Estival, D.; Dauphin, E.; Compagnion, H.; Baur, J.; Balkan, L.; and Arnold, D. 1996. TSNLP - Test Suites for Natural Language Processing. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Mattern, J.; Jin, Z.; Sachan, M.; Mihalcea, R.; and Schölkopf, B. 2022. Understanding Stereotypes in Language Models: Towards Robust Measurement and Zero-Shot Debiasing. *arXiv preprint arXiv:2212.10678*.

Munafò, M. R.; Tilling, K.; Taylor, A. E.; Evans, D. M.; and Davey Smith, G. 2018. Collider scope: when selection bias can substantially influence observed associations. *Int. J. Epidemiol.*, 47(1): 226–235.

OpenAI. 2023. Model index for researchers - OpenAI API. https://platform.openai.com/docs/model-index-for-researchers. (Accessed on 03/07/2023).

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback.

Park, B. S.; Kwon, S. J.; Oh, D.; Kim, B.; and Lee, D. 2022. Encoding Weights of Irregular Sparsity for Fixed-to-Fixed Model Compression. In *International Conference on Learning Representations*.

Park, J. H.; Shin, J.; and Fung, P. 2018. Reducing Gender Bias in Abusive Language Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2799–2804. Brussels, Belgium: Association for Computational Linguistics.

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

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2022. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res.*, 21(1).

Rudinger, R. 2019. winogender-schemas. https://github.com/rudinger/winogender-schemas. (Accessed on 08/15/2023).

Rudinger, R.; Naradowsky, J.; Leonard, B.; and Durme, B. V. 2018. Gender Bias in Coreference Resolution. *CoRR*, abs/1804.09301.

Sagawa, S.; Koh, P. W.; Hashimoto, T. B.; and Liang, P. 2019. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization.

Tay, Y.; Dehghani, M.; Tran, V. Q.; Garcia, X.; Wei, J.; Wang, X.; Chung, H. W.; Bahri, D.; Schuster, T.; Zheng, S.; Zhou, D.; Houlsby, N.; and Metzler, D. 2023. UL2: Unifying Language Learning Paradigms. In *The Eleventh International Conference on Learning Representations*.

Veitch, V.; D'Amour, A.; Yadlowsky, S.; and Eisenstein, J. 2021. Counterfactual Invariance to Spurious Correlations: Why and How to Pass Stress Tests.

Vig, J.; Gehrmann, S.; Belinkov, Y.; Qian, S.; Nevo, D.; Singer, Y.; and Shieber, S. 2020. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 12388–12401. Curran Associates, Inc.

Webster, K.; Recasens, M.; Axelrod, V.; and Baldridge, J. 2018. Mind the GAP: A Balanced Corpus of Gendered Ambiguous Pronouns. *Transactions of the Association for Computational Linguistics*, 6: 605–617.

Webster, K.; Wang, X.; Tenney, I.; Beutel, A.; Pitler, E.; Pavlick, E.; Chen, J.; Chi, E.; and Petrov, S. 2020. Measuring and Reducing Gendered Correlations in Pre-trained Models.

WEF. 2021. World Economic Forum Global Gender Gap Report. https://www3.weforum.org/docs/WEF_GGGR_2021.pdf. (Accessed on 01/10/2022).

Ye, J.; Chen, X.; Xu, N.; Zu, C.; Shao, Z.; Liu, S.; Cui, Y.; Zhou, Z.; Gong, C.; Shen, Y.; Zhou, J.; Chen, S.; Gui, T.; Zhang, Q.; and Huang, X. 2023. A Comprehensive Capability Analysis of GPT-3 and GPT-3.5 Series Models. arXiv:2303.10420.

Zečević, M.; Willig, M.; Dhami, D. S.; and Kersting, K. 2023. Causal Parrots: Large Language Models May Talk Causality But Are Not Causal. *Transactions on Machine Learning Research*.

Zhang, Y.; Baldridge, J.; and He, L. 2019. PAWS: Paraphrase Adversaries from Word Scrambling.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. 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)*, 15–20. New Orleans, Louisiana: Association for Computational Linguistics.