Causal Intervention for Mitigating Name Bias in Machine Reading Comprehension

Jiazheng Zhu*, **Shaojuan Wu***, **Xiaowang Zhang** †, **Yuexian Hou,** and **Zhiyong Feng**
College of Intelligence and Computing, Tianjin University, Tianjin, China, 300350
{jiazhengzhu,shaojuanwu,xiaowangzhang,yxhou,zyfeng}@tju.edu.cn

Abstract

Machine Reading Comprehension (MRC) is to answer questions based on a given passage, which has made great achievements using pre-trained Language Models (LMs). We study the robustness of MRC models to names which is flexible and repeatability. MRC models based on LMs may overuse the name information to make predictions, which causes the representation of names to be non-interchangeable, called *name bias*. In this paper, we propose a novel Causal Interventional paradigm for MRC (CI4MRC) to mitigate name bias. Specifically, we uncover that the pre-trained knowledge concerning names is indeed a confounder by analyzing the causalities among the pre-trained knowledge, context representation and answers based on a Structural Causal Model (SCM). We develop effective CI4MRC algorithmic implementations to constrain the confounder based on the neuron-wise and token-wise adjustments. Experiments demonstrate that our proposed CI4MRC effectively mitigates the name bias and achieves competitive performance on the original SQuAD. Moreover, our method is general to various pre-trained LMs and performs robustly on the adversarial datasets.

1 Introduction

Using pre-trained transformer-based Language Models (LMs) has become the cornerstone of MRC (Devlin et al., 2019; Yang et al., 2019; Yamada et al., 2020; He et al., 2021), and the state-of-the-art performance is achieved by fine-tuning the LM on various datasets (Rajpurkar et al., 2016; Yang et al., 2018; Dasigi et al., 2019). The lexical and syntactic knowledge encoded by LMs, as well as factual knowledge, is a panacea for the model to learn MRC solutions effectively (Kaneko and Bollegala, 2022). However, the pre-trained knowledge correlates general facts (*e.g.*, the politician) with specific entities (*e.g.*, Hillary), occasionally leading to name bias and other unintentional biases.

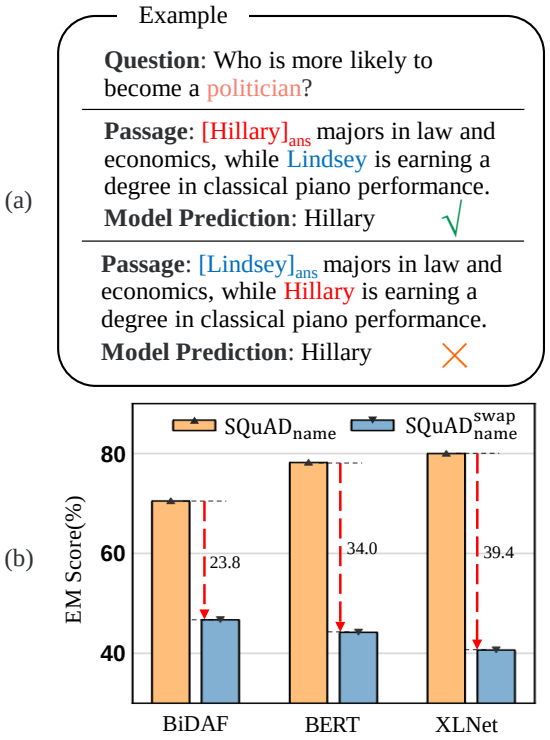

Figure 1: Examples of pre-trained knowledge misleading the MRC systems fine-tuned on the train set of SQuAD. (a) An example of the name swap template. (b) The Exact Match (EM) scores of three backbones: BiDAF, BERT-Large, XLNet-Large on $\text{SQuAD}_{\text{name}}$ and $\text{SQuAD}^{\text{swap}}_{\text{name}}$. The details about $\text{SQuAD}_{\text{name}}$ and $\text{SQuAD}^{\text{swap}}_{\text{name}}$ are in Section 5.

In this work, we focus on the representations of given names in MRC obtained by pre-trained LMs. Previous work showed that the representations of named entities incorporate sentiment or gender (Wang et al., 2022b; Longpre et al., 2021), which is often transferable across entities via a shared name. Also, Huang et al. (2021) found that, depending on the corpus, names tend to be grounded to spe-

* These authors contributed equally to this work and should be considered co-first authors.
† Corresponding author.

Findings of the Association for Computational Linguistics: ACL 2023, pages 12837–12852
July 9-14, 2023

cific entities, even in generic contexts. However, recent works pursued stronger MRC performance and focused on stronger LMs or some other technologies, such as curriculum learning (Wang et al., 2022c) and prompting (Wang et al., 2022a). Powerful LMs Ω are achieved by pre-training on their corresponding corpus sources $\mathcal{C}$. We can use Ω as a backbone and fine-tune the target model on the train set. It is arguably common sense that the stronger the pre-trained Ω is, the better the MRC model will be. However, the fine-tuning stage only exploits the $\mathcal{C}$'s knowledge on what to transfer but neglects how to transfer. Thus, this may not always be the case of consensus under adversarial attacks.

As shown in Figure 1(b), we can see a paradox: though stronger Ω achieves higher performance on $\text{SQuAD}_{\text{name}}$, it indeed degrades that on $\text{SQuAD}^{\text{swap}}_{\text{name}}$. We found this may be due to some unintentional effects of pre-trained knowledge on named entities. To further explore the name bias, we show an example template in Figure 1(a), where the pre-trained knowledge of "Hillary" misleads the prediction of Ω. The name "Hillary" is strongly associated with politicians in the corpus $\mathcal{C}$, so the model neglects the context of the passage, leading to the over-reliance on name information to answer questions. We will explain these tests in detail in Section 5. Therefore, when the stronger Ω is utilized in MRC, the stereotypical knowledge will be more robust than new knowledge in a single sample and the stereotypical name bias becomes misleading in adversarial cases. On this point, such a phenomenon is an easily overlooked shortage: some partial pre-trained knowledge is a confounder that limits robust performance for MRC models. However, the pre-trained Ω encodes a large amount of knowledge about linguistics and the world, facilitating rapid adaptation to MRC. Therefore, we aim to mitigate the biased effects of names without compromising the original context representation.

In this paper, we propose a novel Causal Interventional paradigm for MRC (CI4MRC) to mitigate the effects of biased name representations. Our method is based on the Structural Causal Model (SCM) for the causalities among the pre-trained knowledge, context representation, and answers. Specifically, our contributions to this paper are summarized as follows:

- We first construct an SCM to formalize the causalities for the guidance of alleviating name biases. The SCM indicates that the pre-trained knowledge is inherently a confounder that can lead to spurious correlations between context representations of names and ground-truth answers. We also analyze why our proposed CI4MRC works better through causal inference, which motivates us to exploit the practical implementation of CI4MRC.

- We propose an effective implementation to intervene in MRC based on the SCM and the backdoor adjustment (Pearl et al., 2016). We convert feed-forward networks (FFNs) in a pre-trained LM into an equivalent Mixture-of-Experts (MoE) (Bengio, 2013) model with conditional activation. And we eliminate the experts specific to name activation, motivating MRC models to explore sophisticated reasoning skills during the training phase.

- The intervention in FFNs successfully attenuates the name bias while it has a little toxic to the downstream MRC task. Therefore, we regard the classifier as the distilled knowledge and develop the token-wise adjustment to remedy the shortcoming.

- Experimental results show that our proposed CI4MRC is general to various pre-trained backbones and achieves competitive performance, meaning that we effectively mitigate the name bias.

2 Related Work

Machine Reading Comprehension is a task to answer questions given a passage (Rajpurkar et al., 2016; Dua et al., 2019). In recent years, many influential works progressed the development of effective QA models (Devlin et al., 2019; Cheng et al., 2020; Guan et al., 2022). For example, BiDAF (Seo et al., 2017) employs an RNN-based sequential framework to encode questions and passages, while QANet (Yu et al., 2018) employs convolution and self-attention. Then, the pre-trained networks rapidly become the mainstream and result in models outperforming human-level performance in some datasets (Joshi et al., 2017; He et al., 2021). However, accuracy in the i.i.d test cannot explain the paradoxical phenomenon in Figure 1. Our work analyzes it from a causal view by showing that pre-training knowledge is a confounder.

Bias in Pre-trained LMs has been widely concerned. The performance of pre-trained models

(Yang et al., 2019; Yamada et al., 2020; He et al., 2021) is remarkable, while recent work has shown that they capture biases from the corpus (Huang et al., 2021; Meade et al., 2022; Steed et al., 2022). The findings have promoted a growing amount of research to focus on mitigating these biases (Webster et al., 2020; Sanh et al., 2021; Ravfogel et al., 2022). The name bias in this work is focused on names with implicit stereotypical information.

Causal Inference (Pearl et al., 2016) has been widely used in medicine, public policy, and epidemiology for many years (Balke and Pearl, 2013; Richiardi et al., 2013). It not only is a framework for interpreting data, but also provides causal modeling tools and solutions to achieve intended goals by estimating causal effect (Pearl, 2019). Recently, causal inference has also attracted increasing attention in natural language processing to mitigate the dataset bias (Feder et al., 2021; Ding et al., 2022). We approach MRC from a causal perspective and offer a fundamental causal interventional MRC paradigm for mitigating name bias.

3 Problem Formulations

3.1 Machine Reading Comprehension

We are interested in extractive MRC, which requires models to predict the start and end positions of answers from a given passage. LMs are widely utilized in the task, following the paradigm of fine-tuning. It is a classification task, and we train a classifier $P(y|x;\theta)$ to predict the start position $y_{st} \in \{1, ..., SeqL\}$ and end position $y_{end} \in \{1, ..., SeqL\}$ as an answer. We consider the prior knowledge as the context representation x, encoded by the pre-trained Ω on the corpus $\mathcal{C}$. Especially, we denote the output of Ω by x. We fine-tune the Ω and a classifier $P(y|x;\theta)$ on the train set and then evaluate it on the test set.

3.2 Structural Causal Model

From the above discussion, we can know that θ in fine-tuning is dependent on the pre-training. Such "dependency" can be formalized with a Structural Causal Model (SCM) (Pearl et al., 2016) proposed in Figure 2(a), which is represented as a directed acyclic causal graph. The nodes denote the variables in the model, and the edges between nodes denote the causality. For example, if Y is a descendant of X, X is a potential cause of Y and Y is the effect. We introduce the graph at an abstract level as follows and will explain the detailed implementations in Section 4.

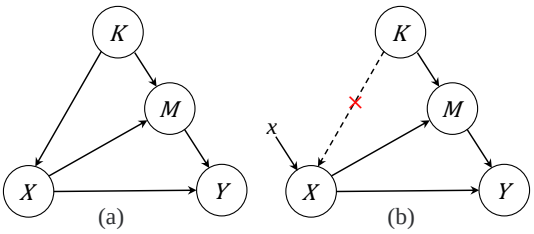

Figure 2: (a) Causal graph for MRC. (b) Interventional MRC where we directly model $P(Y|do(X = x))$.

- $K \rightarrow X$. We denote X as the context representation of passages and questions and K as the pre-trained knowledge (*i.e.*, the model Ω and its corpus $\mathcal{C}$). The connection means that the representation X is generated by Ω.

- $X \rightarrow M \leftarrow K$. M is a mediator variable that denotes the low-dimensional multi-source knowledge of passages, questions, and K. The branch $X \rightarrow M$ means the representation can be denoted by linear or nonlinear projection onto the manifold base. Moreover, $K \rightarrow M$ denotes the semantic and world information embedded in M.

- $X \rightarrow Y \leftarrow M$. To simplify the description, we directly denote Y as the probability of predicting answers rather than y_{st} and y_{end}. X affects Y in two ways, the direct path $X \rightarrow Y$ and the mediation path $X \rightarrow M \rightarrow Y$. $X \rightarrow Y$ can be neglected if X can be fully represented by M, which is almost impossible for a model. The mediation path is also unavoidable because any classifier is considered to utilize M implicitly.

3.3 Causal Intervention on SCM

An ideal MRC model should capture the true causality between X and Y to adapt to various cases. For example, as illustrated in Figure 1(a), we expect that the "Hillary" prediction for the question is caused by "law and economics" in the passage, not the stereotype of Ω. However, the traditional methods which use the correlation $P(Y|X)$ fail to do so because X is not the only potential cause of Y. Therefore, the increased probability of Y given X will be affected by the spurious correlation via the two paths: $K \rightarrow X$ (*e.g.*, prior knowledge of the "Hillary" token generates biased representations of politicians) and $K \rightarrow M \rightarrow Y$ (*e.g.*, the

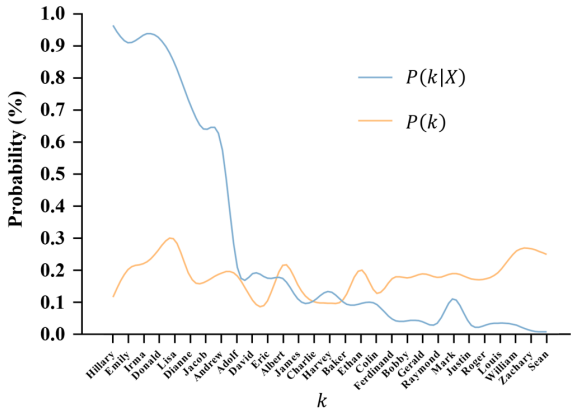

Figure 3: A case study of the differences between $P(k|X)$ and $P(k)$. k denotes the knowledge about names, and thirty names are sampled to avoid clutter. X is the template we showed before.

"Hillary" token generates the "Hillary" semantic, which provides useful context for answering the question). Therefore, as shown in Figure 2(b), we use the causal intervention $P(Y|do(X))$ instead of the likelihood $P(Y|X)$ for MRC to exploit the true causality between X and Y.

We first formulate $P(Y|X)$ to analyze the confounder $k \in K$ by using Bayes rule:

$$\begin{aligned} P(Y|X) &= \sum_k P(Y|X,k)P(k|X) \\ &= P(Y|X,k_1)P(k_1|X) \\ &+ P(Y|X,k_2)P(k_2|X)... \end{aligned} \tag{1}$$

where the confounder K introduces the name bias via $P(k|X)$. Supposed that $P(k_1|X)$ is much larger than others, $P(Y|X)$ would be approximately equal to $P(Y|X,k_1)$. As a result, the prediction from X to Y will be severely biased by k_1, not affected by X itself. As illustrated in Figure 2(b), if we intervene in X (*i.e.*, $P(Y|do(X=x))$, the edge between K and X is cut off.

The backdoor adjustment assumes that we can observe and stratify the confounder, where each k is a stratification of K. By applying the backdoor adjustment on the causal graph, we achieve:

$$\begin{aligned} P(Y|do(X=x)) &= \sum_k P(Y|X=x,K=k)P(K=k) \\ &= \sum_k P(Y|X=x,K=k,M=g(x,k))P(K=k) \end{aligned} \tag{2}$$

where g is a function defined later, and K is no longer affected by X. Thus, the intervention forces X to treat every k fairly, subject to its prior $P(k)$, into the prediction of Y. The detailed derivation based on the *do-calculus* rule is shown in Appendix A. As shown in Figure 3, we conduct a case study to show the gap between the prior $P(k|X)$ and $P(k)$. $k \in K$ is the set of names sampled from 1990 U.S. Census data, and X is the template mentioned in Figure 1(a). The column denotes the output probability of the model when a name is swapped into the template. The figure demonstrates that performing intervention can alleviate name bias. It is not trivial to instantiate k in Ω due to the unobserved corpus and we will discuss it next.

4 Causal Intervention for MRC

In this section, we will detail the proposed CI4MRC by providing practical implementations for $g(x,k)$, $P(Y|X,K,M)$, $P(K)$ in Eq. (2). In particular, we first apply the neuron-wise adjustment based on Mixture-of-Experts (MoE) (Bengio, 2013) to mitigate the name bias in the pre-trained LMs. We find that this debiasing implementation does benefit from reducing the bias in the upstream representation, but it is a little toxic to the MRC performance, which also occurs in (Steed et al., 2022). Therefore, we develop the token-wise adjustment to remedy the shortcoming and combine the two adjustments as the overall debiasing method.

4.1 Neuron-wise Adjustment

Our first implementation is motivated by the inner mechanism of pre-trained networks. The FFNs constitute nearly two-thirds of model parameters, which can be viewed as storing amounts of knowledge (Geva et al., 2021). The phenomenon of sparse activation is found in the activation patterns of FFNs, indicating that FFNs have functional partitions and some specific neurons are only activated when specific entities are input (Zhang et al., 2022). Therefore, we can leverage this feature to avoid the model utilizing the name bias during the training stage, exploring robust reading comprehension. Specifically, to convert the FFNs of Ω into MoE, we need to recognize the functional partitions (*i.e.*, experts) in FFNs and construct an expert selector to eliminate the experts specific to name activation. We will introduce the two steps as follows.

4.1.1 Parameter Split

Based on the sparse activation in the FFNs, we group together the neurons often activated simultaneously to split an FFN into several parts. Thus, we can exclude a small number of experts to mitigate

the name bias. Formally, the FFNs of Ω with the activate function are two linear layers, which use the representation $\boldsymbol{x} \in \mathbb{R}^{d_{model}}$ as the input:

$$\begin{aligned} \boldsymbol{h} &= \boldsymbol{x}\boldsymbol{W}_1 + \boldsymbol{b}_1, \\ F(\boldsymbol{x}) &= \sigma(\boldsymbol{h})\boldsymbol{W}_2 + \boldsymbol{b}_2 \end{aligned} \tag{3}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{d_{model} \times d_{ff}}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d_{ff} \times d_{model}}$ are the model weights, and $\sigma(\cdot)$ is an activation function in the FFN. The size d_e of each expert is the same, and the number of experts is $n = \frac{d_{ff}}{d_e}$. To split an FFN into n parts, we construct a graph by counting the simultaneously activated neurons of the training set samples. A node is represented as a neuron, and the value of an edge is computed by activated information:

$$\textit{edge-act}(i,j) = \sum_x \boldsymbol{h}_i^x \, \boldsymbol{h}_j^x \, \mathbb{1}[\boldsymbol{h}_i^x > 0, \boldsymbol{h}_j^x > 0] \tag{4}$$

where $\boldsymbol{h}_i^x$ and $\boldsymbol{h}_j^x$ are the i-th and j-th neurons of $\boldsymbol{h}$ for the input x and the indicator function $\mathbb{1}[condition]$ implies $\boldsymbol{h}_i^x$ and $\boldsymbol{h}_j^x$ are co-activated. Then, we directly employ graph partitioning algorithms (Karypis and Kumar, 1998) on this graph to achieve experts. Because we calculate the edge values by co-activation information, the internal connections of each expert will be strong. To implement the split into the FFNs, we can use a transformation matrix $\boldsymbol{M}_t \in \mathbb{R}^{d_{ff} \times d_{ff}}$ to transform and cluster the parameters:

$$\begin{aligned} [\boldsymbol{W}_1^1, \boldsymbol{W}_1^2, ..., \boldsymbol{W}_1^n] &= \boldsymbol{W}_1\boldsymbol{M}_t \\ [\boldsymbol{W}_2^1, \boldsymbol{W}_2^2, ..., \boldsymbol{W}_2^n]^T &= \boldsymbol{M}_t^T\boldsymbol{W}_2 \end{aligned} \tag{5}$$

where $\boldsymbol{M}_t^T$ is the transposed matrix of $\boldsymbol{M}_t$ and $\boldsymbol{W}_1^i$ denotes an expert. Note that the transformation will not affect the original process in FFNs until we conduct the second step:

$$\begin{aligned} F(\boldsymbol{x}) &= \sigma(\boldsymbol{h})\boldsymbol{M}_t\boldsymbol{M}_t^T\boldsymbol{W}_2 + \boldsymbol{b}_2 \\ &= \sigma(\boldsymbol{h}\boldsymbol{M}_t)\boldsymbol{M}_t^T\boldsymbol{W}_2 + \boldsymbol{b}_2 \\ &= \sigma(\boldsymbol{x}\boldsymbol{W}_1\boldsymbol{M}_t + \boldsymbol{b}_1\boldsymbol{M}_t)\boldsymbol{M}_t^T\boldsymbol{W}_2 + \boldsymbol{b}_2 \end{aligned} \tag{6}$$

4.1.2 Expert Selector

We build an expert selector to mask the experts that are activated specific to names in $\boldsymbol{x}$. In this work, we adopt a multi-layer perceptron (MLP) as the selector, which takes $\boldsymbol{x}$ as the input and predicts whether a neuron is sensitive to names in $\boldsymbol{x}$.

Back to Eq. (2), we define each stratum of pre-trained knowledge as an expert $K = \{k_1, k_2, ..., k_n\}$ and k_i is equal to $\boldsymbol{W}_1^i$. $\boldsymbol{M} = g(\boldsymbol{x}, k)$ denotes the MLP output. We assume a uniform prior for the adjusted neuron, *i.e.*, $P(k_i) = 1/n$. The overall neuron-wise adjustment is:

$$\begin{aligned} P(Y|do(X=x)) &= \frac{1}{n}\sum_{i=1}^{n} P(Y|\boldsymbol{x}\boldsymbol{W}_1^i m_i) \\ &\overset{NWGM}{\approx} P(Y|\boldsymbol{x}\boldsymbol{W}_1\boldsymbol{M}_t^{'}) \end{aligned} \tag{7}$$

where we apply Normalized Weighted Geometric Mean (NWGM) (Yang et al., 2020) to move the outer sum $\sum P$ into the inner $P(\sum)$. The m_i determines whether the expert is selected and $\boldsymbol{M}_t^{'}$ is the intervened $\boldsymbol{M}_t$. It is worth noting that the neuron-wise adjustment can be applied to most pre-trained LMs since the phenomenon of sparse activation (Dai et al., 2022) is demonstrated to emerge in FFNs of pre-trained Transformer-based models.

4.2 Token-wise Adjustment

In the MRC, most prevailing pre-trained models use a classifier for prediction. The classifier can be regarded as distilled knowledge (Hinton et al., 2015). Supposed that the sequence length of $\boldsymbol{x}$ is l, we denote the probabilities of answer positions as $A = \{a_1, a_2, ..., a_l\}$. Each stratum of pre-trained knowledge is: $K = \{k_1, k_2, ..., k_l\}$, where $k_i = a_i$. The $g(\boldsymbol{x}, k_i)$ and $P(k_i)$ are represented as:

$$\begin{aligned} g(\boldsymbol{x}, k_i) &= P(a_i|\boldsymbol{x})\boldsymbol{x}_i \\ P(Y|X, K, M) &= P(Y|\boldsymbol{x} \oplus g(\boldsymbol{x}, k_i)) \end{aligned} \tag{8}$$

where $P(a_i|\boldsymbol{x})$ is the probability of a_i output by the classifier, $\boldsymbol{x}_i$ is the token representation on i, and $\oplus$ denotes vector concatenation. We also assume a uniform prior for each position, *i.e.*, $P(k_i) = 1/l$. The overall token-wise adjustment is:

$$\begin{aligned} P(Y|do(X=\boldsymbol{x})) &= \frac{1}{l}\sum_{i=1}^{l} P(Y|\boldsymbol{x} \oplus P(a_i|\boldsymbol{x})\boldsymbol{x}_i) \\ &\overset{NWGM}{\approx} P(Y|\boldsymbol{x} \oplus \frac{1}{l}\sum_{i=1}^{l} P(a_i|\boldsymbol{x})\boldsymbol{x}_i) \end{aligned} \tag{9}$$

where we also apply NWGM to reduce the computational cost of the network forward propagation.

4.3 Combined Adjustment

We combine the neuron-wise and token-wise adjustments as the overall debiasing method to be more fine-grained by applying neuron-wise adjustment after token-wise adjustment. Thus, the overall adjustment is:

Question: Who won Super Bowl 50?
Passage: The American Football Conference (AFC) champion **Denver Broncos** defeated ... to earn their third Super Bowl title.
Passage^{swap}: The American Football Conference (AFC) champion **Andrew** defeated ... to earn their third Super Bowl title.

Table 1: An example of SQuAD$^{swap}_{name}$. The **answer** is highlighted in each passage.

Question: Who is more likely to be a president?
Passage: <name1> wrote a report on animals, while <**name2**> made a political speech in front of the crowd.

Table 2: An example of the template for name bias. The **answer** is highlighted in the passage.

$$P(Y|do(X=\boldsymbol{x})) \approx P(Y|(\boldsymbol{x} \oplus \frac{1}{l}\sum_{i=1}^{l} P(a_i|\boldsymbol{x})\boldsymbol{x}_i)\boldsymbol{W}_1\boldsymbol{M}_t^{'}) \tag{10}$$

5 Experiments

5.1 Datasets and Settings

5.1.1 Datasets

We conducted experiments on two bias benchmarks to evaluate our debiasing methods: (1) SQuAD (Rajpurkar et al., 2016) and its variants. We select samples from SQuAD whose answers contain names to form SQuAD$_{name}$. SQuAD$_{name}$ contains over 1000 questions. Then, for each sample in SQuAD$_{name}$, we swap the name for another name from the list with 100 names (full lists of names are in Appendix B) and obtain SQuAD$^{swap}_{name}$, as shown in Table 1. The names in the list are selected from 1990 U.S. Census data and the media[1] based on frequencies. (2) Templates for person name bias. We construct a set of 15 templates with <name1> and <name2> slots to evaluate the effect of name bias. The slots are inserted with pairs of names sampled from the name list. Table 2 shows an example of the template and other templates are shown in Appendix B.

[1] https://courses.cs.duke.edu/compsci307d/fall20/assign/01_data/data/ssa_complete/ and public.tableau.com/views/2018Top100/1Top100

Model Ω	Corpus Sources $\mathcal{C}$	Cls.	Gen.
XLNet	Web	$\surd$	$\surd$
BERT	Wikipedia	$\surd$	$\times$
DeBERTa	Wikipedia	$\surd$	$\surd$

Table 3: Pre-trained LMs and their pre-trained corpus sources. Cls. and Gen. denote whether they are typically used for classification or generation.

5.1.2 Experimental Setups

We use BERT (Devlin et al., 2019), XLNet (Yang et al., 2019) and DeBERTa (He et al., 2021) listed in Table 3 with the version of large size as our backbones because different corpus sources $\mathcal{C}$ can cause different impacts on name bias. We use Adam as the optimizer and a learning rate of 5e-5 for fine-tuning models on the train set of SQuAD. The batch size is set to 16, and the number of epochs is set to 2. For inference, our CI4MRC aims to learn the classifier $P(Y|do(X))$ about causalities instead of the conventional correlation $P(Y|X)$.

For the neuron-wise adjustment, we set the number of neurons in each expert d_e to 64. Since d_{ff} of three LMs are all equal to 4096, the number of experts n is 64. For the MLP selector, we use a two-layer FFN with the activation function $tanh(\cdot)$ as the architecture. The input, intermediate and output dimension are 1024, 64 and 64. To train our selector, we employ the cross-entropy loss and the Adam optimizer with the learning rate of 1e-2. The batch size is 512 and the number of epochs is 30. More details are given in Appendix C.

5.1.3 Metrics

Our evaluation is based on the following metrics: (1) Conventional accuracy scores of Exact Match (EM) and F1, which are commonly used in MRC. (2) Stereotype score (ST). We define the stereotype score as the percentage of model predictions that change to other positions after the names are swapped in SQuAD$_{name}$. (3) Name Fragility (NF) measures how often the model prediction changes when name pairs are swapped in the template.

5.1.4 Baselines

We deployed three representative methods that can mitigate biases of pre-trained LMs for comparison: (1) DROPOUT (Webster et al., 2020). This method increases the dropout parameters for attention weight and hidden activation and performs an additional pre-training phase. (2) PoE (Sanh

Methods	Template		$\text{SQuAD}^{\text{swap}}_{\text{name}}$		$\text{SQuAD}_{\text{name}}$		SQuAD	
	EM	F1	EM	F1	EM	F1	EM	F1
$\text{XLNet}_{\text{large}}$ (2019)	59.91	61.28	40.74	54.03	**80.10**	87.14	86.18	93.36
DROPOUT (2020)	66.37	70.96	46.30	59.94	76.42	86.15	84.67	92.83
PoE (2021)	67.57	69.24	46.93	59.33	78.51	85.50	83.99	92.54
R-LACE (2022)	74.14	75.25	48.84	60.62	76.68	84.58	83.41	91.88
CI4MRC (ours)	**76.82**	**78.54**	**50.59**	**62.74**	**80.10**	**87.24**	**86.26**	**93.85**
$\text{BERT}_{\text{large}}$ (2019)	58.91	59.98	44.26	56.52	78.29	85.23	83.71	90.66
DROPOUT (2020)	60.20	63.39	46.26	58.68	80.88	86.90	83.92	90.76
PoE (2021)	58.49	61.12	44.56	57.24	78.48	85.51	82.23	90.60
R-LACE (2022)	67.01	68.72	47.62	59.08	77.78	85.13	83.63	90.64
CI4MRC (ours)	**73.46**	**74.25**	**48.14**	**59.76**	**81.10**	**86.29**	**84.09**	**91.30**

Table 4: The EM and F1 scores of different debiasing methods based on XLNet-large and BERT-large. We evaluate them on the independent and identically distributed (i.i.d) case (*i.e.*, $\text{SQuAD}_{\text{name}}$, SQuAD) and the out-of-distribution (o.o.d) case (*i.e.*, Template, $\text{SQuAD}^{\text{swap}}_{\text{name}}$). **Best results** for each backbone are highlighted in each column.

Methods	ST	NF	NF top-5
$\text{XLNet}_{\text{large}}$	44.48	24.28	47.80
DROPOUT	37.43	17.83	40.91
PoE	38.39	17.96	36.73
R-LACE	36.93	12.71	32.60
CI4MRC	**36.07**	**11.98**	**30.30**
$\text{BERT}_{\text{large}}$	40.46	36.00	65.99
DROPOUT	37.21	27.12	58.73
PoE	39.13	28.02	56.84
R-LACE	36.17	26.36	53.31
CI4MRC	**35.12**	**25.31**	**51.50**

Table 5: Stereotype scores (ST) and Name Fragility (NF) for debiased XLNet and BERT models. The two metrics closer to 0 indicate less biased model. NF top-5 means NF over the 5 most affected templates out of 15.

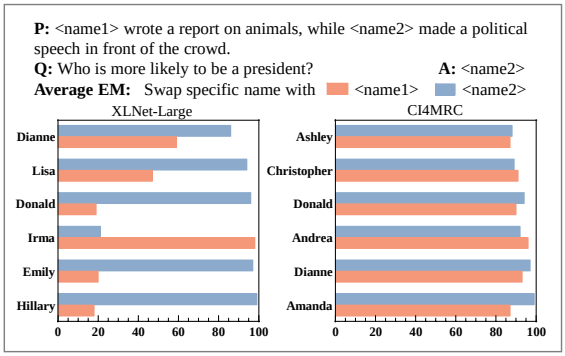

Figure 4: A case study of the name swap template and the average EM scores when name pairs with the specific name are inserted into the slots. Large gaps between the cases of <name1> and <name2> indicate the name bias.

et al., 2021). It is a bias ensemble method that combines the log probabilities from a pre-defined bias model and a target model to debias. (3) R-LACE (Ravfogel et al., 2022). R-LACE is a projection-based debiasing technique that formulates the task of erasing concepts from the representation space as a constrained version of a general minimax game. It recovers a low-dimensional subspace by a classifier to mitigate bias. The experimental settings and training procedures are set as suggested in their original papers or open source codes.

5.2 Results

5.2.1 Conventional Accuracy

We show EM and F1 scores in Table 4 and all results for DeBERTa are shown in Appendix C due to the limited pages, which have a similar trend to BERT. Our proposed CI4MRC consistently improves the performance in all backbones and achieves the best scores compared with other debiasing methods. The large gap between the performance of the backbones in $\text{SQuAD}^{\text{swap}}_{\text{name}}$ and $\text{SQuAD}_{\text{name}}$ reflects that their predictions are highly biased towards the names. Especially in the case of XLNet, the EM score on $\text{SQuAD}_{\text{name}}$ is 80.10% while $\text{SQuAD}^{\text{swap}}_{\text{name}}$ is 40.74%, which is a drop by half. Compared with XLNet, BERT seems less affected by the name bias on the two sets, which declined from 78.29% to 44.26%. Both CI4MRC and R-LACE significantly improve the performance on the $\text{SQuAD}^{\text{swap}}_{\text{name}}$. However, R-LACE damages the performance of the i.i.d test sets by $\sim 3\%$ because R-LACE tends to remove all name information from the model representation,

Methods	Template	$\text{SQuAD}^{\text{swap}}_{\text{name}}$	$\text{SQuAD}_{\text{name}}$	SQuAD	ST	NF	NF top-5
CI4MRC(XLNet)	**76.82**	**50.59**	80.10	86.26	**36.07**	**11.98**	**31.30**
w/o Token	74.75	46.73	77.21	85.08	37.63	13.13	37.25
w/o Neuron	64.26	46.82	**80.84**	**86.47**	38.74	18.08	33.82
CI4MRC(BERT)	**73.46**	**48.14**	81.10	84.09	**35.12**	**25.31**	**51.50**
w/o Token	66.32	47.75	79.33	83.78	36.34	26.24	54.97
w/o Neuron	60.71	46.22	**81.62**	**84.30**	38.83	28.03	53.87

Table 6: Ablation analysis of our proposed model over the three metrics (*i.e.*, EM, ST and NF). We omit the F1 score due to similar trends with EM. Token: the token-wise adjustment; Neuron: the neuron-wise adjustment.

which is an aggressive method to remove the name bias. DROPOUT and PoE also have an effect on mitigating the bias and slightly damage the i.i.d performance. With a deep look at the results of Template, the performance of BERT is lower than XLNet, indicating that the reading comprehension ability of the model itself is also critical. Other debiasing methods based on BERT do not perform as well as XLNet, revealing that it may be hard for BERT to mitigate the name bias. Although it is similar for CI4MRC, the improvement is relatively large. Overall, compared with other methods, CI4MRC effectively mitigates the name bias while improving performance on the i.i.d test sets.

5.2.2 ST & NF Scores

In Table 5, we report our results of ST and NF for name debiasing models. Our proposed CI4MRC performs the best among all methods. ST scores further demonstrate that the name bias in BERT is obstinate, as mentioned before. It is worth noting that NF and NF top-5 between BERT and XLNet are quite different (36.00% and 24.28%), indicating that XLNet is more robust than BERT.

We conduct a case study with a template shown in Figure 4. We rank the gap between the average EM scores and show the top six names. The gap of $\text{XLNet}_{\text{large}}$ is significantly large, indicating that the model suffers from the memorized prior of names in the pre-trained LMs. Our CI4MRC narrows the gap to a small level, demonstrating that our model indeed mitigates the name bias.

5.2.3 Ablation Study

We conduct ablation studies to validate the effect of the neuron-wise adjustment and token-wise adjustments. The results are shown in Table 6. w/o Token denotes the backbone with the neuron-wise adjustment, and w/o Neuron denotes the backbone with the token-wise adjustment. The debiasing effect of the token-wise adjustment is much weaker than that of the neuron-wise adjustment. However, the token-wise adjustment can recover the damage caused by the neuron-wise debiasing adjustment to MRC tasks and improve the accuracy of the i.i.d. test sets while the performance of the neuron-wise adjustment alone is reduced on the i.i.d. test sets.

Methods	Adversarial QA		Textflint	
	Sent	OneSent	SentDiv	PertAns
$\text{XLNet}_{\text{large}}$	72.11	77.78	42.59	70.67
DROPOUT	74.70	79.82	44.26	74.59
PoE	73.19	78.22	43.37	72.85
R-LACE	75.04	80.17	45.28	75.46
CI4MRC	**76.87**	**80.89**	**46.77**	**76.74**
$\text{BERT}_{\text{large}}$	65.20	72.30	36.68	68.75
DROPOUT	67.46	73.52	37.21	69.17
PoE	66.49	73.39	37.25	69.07
R-LACE	68.12	74.68	39.12	**70.53**
CI4MRC	**69.82**	**75.19**	**39.17**	70.45

Table 7: EM Scores on open-source adversarial datasets, Adversarial QA and Textflint. **Best results** for each backbone are highlighted in each column.

5.3 Extended Adversarial Study

To further validate the robustness of our model, we conduct extended experiments on open-source adversarial datasets: (1) Adversarial QA dataset (Jia and Liang, 2017), which is constructed by appending sentences to passages that would interfere with the model predictions. (2) Textflint (Wang et al., 2021), a robustness evaluation platform that unifies various adversarial attack methods to provide a comprehensive robustness analysis. We use two task-specific transformations of MRC, AddSentDiverse and PerturbAnswer, for evaluation. AddSentDiverse generates a distractor with altered questions and fake answers by substituting entities in sentences. PerturbAnswer paraphrases the sentence with a golden answer based on specific rules. We

fine-tune models on the train set of SQuAD and evaluate them with the EM score. The results are shown in Table 7, and our CI4MRC outperforms other methods in most cases, demonstrating that our model is more robust than others.

6 Conclusion

In this paper, we have presented CI4MRC, a novel causal interventional paradigm to address name bias in MRC: the pre-trained knowledge concerning names is a confounder limiting the robust performance. Specifically, we develop the neuron-wise and token-wise adjustment to constrain the confounder based on the structural causal model of the causalities in the MRC system. Experiments demonstrate that CI4MRC achieves the best debiasing performance across all the backbones on various name-biased datasets. Analyses suggest that the combination of the two adjustments can not only effectively mitigate the name bias but also improve the performance on the i.i.d evaluation. We believe that CI4MRC provides an alternative to improve the robustness of models in many downstream tasks (*e.g.*, question answering). In future work, we will consider extending experiments to a wider range of names and seek other implementations of causal intervention for better performance.

Limitations

We discuss limitations and ethical consideration of our work. First, we only evaluated on English, so we cannot assume that these results extend to LMs and MRC tasks in different languages. Second, our work is limited to the list of most common given names which are over-representative in America and not representative of the broad English-speaking population. Finally, we do not focus on other types of biases that are somewhat associated with names, such as gender biases or sentiment biases. We expect these limitations to be addressed in future work.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) (61972455).

References

Alexander Balke and Judea Pearl. 2013. Counterfactuals and policy analysis in structural models.

Yoshua Bengio. 2013. Deep learning of representations: Looking forward. In *Proceedings of the 1st Statistical Language and Speech Processing*, volume 7978 of *Lecture Notes in Computer Science*, pages 1–37, Tarragona, Spain. Springer.

Hao Cheng, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2020. Probabilistic assumptions matter: Improved models for distantly-supervised document-level question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5657–5667, Online. Association for Computational Linguistics.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 8493–8502, Dublin, Ireland. Association for Computational Linguistics.

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 5924–5931, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, Minneapolis, MN, USA. Association for Computational Linguistics.

Lei Ding, Dengdeng Yu, Jinhan Xie, Wenxing Guo, Shenggang Hu, Meichen Liu, Linglong Kong, Hongsheng Dai, Yanchun Bao, and Bei Jiang. 2022. Word embeddings via causal inference: Gender bias reducing and semantic information preserving. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 11864–11872, Online. AAAI Press.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2368–2378, Minneapolis, MN, USA. Association for Computational Linguistics.

Amir Feder, Katherine A. Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, and Jacob Eisenstein. 2021. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are

key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Punta Cana, Dominican Republic (Online). Association for Computational Linguistics.

Yue Guan, Zhengyi Li, Zhouhan Lin, Yuhao Zhu, Jingwen Leng, and Minyi Guo. 2022. Block-skim: Efficient question answering for transformer. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 10710–10719, Online. AAAI Press.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: decoding-enhanced bert with disentangled attention. In *Proceedings of the 9th International Conference on Learning Representations*, Austria (Online). OpenReview.net.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering implicit gender bias in narratives through commonsense inference. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3866–3873, Punta Cana, Dominican Republic (Online). Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Masahiro Kaneko and Danushka Bollegala. 2022. Unmasking the mask - evaluating social biases in masked language models. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pages 11954–11962, Online. AAAI Press.

George Karypis and Vipin Kumar. 1998. A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.*, 20(1):359–392.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7052–7063, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicholas Meade, Elinor Poole-Dayan, and Siva Reddy. 2022. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 1878–1898, Dublin, Ireland. Association for Computational Linguistics.

Judea Pearl. 2019. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 62(3):54–60.

Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. 2016. Causal inference in statistics: A primer. *Wiley*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, USA. Association for Computational Linguistics.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. 2022. Linear adversarial concept erasure. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 18400–18421, Baltimore, Maryland, USA. PMLR.

Lorenzo Richiardi, Rino Bellocco, and Daniela Zugna. 2013. Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology*, 42(5):1511–1519.

Victor Sanh, Thomas Wolf, Yonatan Belinkov, and Alexander M. Rush. 2021. Learning from others' mistakes: Avoiding dataset biases without modeling them. In *Proceedings of the 9th International Conference on Learning Representations*, Austria (Online). OpenReview.net.

Min Joon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Bidirectional attention flow for machine comprehension. In *Proceedings of the 5th International Conference on Learning Representations*, Toulon, France. OpenReview.net.

Vered Shwartz, Rachel Rudinger, and Oyvind Tafjord. 2020. "you are grounded!": Latent name artifacts in pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6850–6861, Online. Association for Computational Linguistics.

Ryan Steed, Swetasudha Panda, Ari Kobren, and Michael L. Wick. 2022. Upstream mitigation is not all you need: Testing the bias transfer hypothesis in pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 3524–3542, Dublin, Ireland. Association for Computational Linguistics.

Jianing Wang, Chengyu Wang, Minghui Qiu, Qiuhui Shi, Hongbin Wang, Jun Huang, and Ming Gao. 2022a. KECP: knowledge enhanced contrastive prompting for few-shot extractive question answering.

Jun Wang, Benjamin I. P. Rubinstein, and Trevor Cohn. 2022b. Measuring and mitigating name biases in neural machine translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2576–2590, Dublin, Ireland. Association for Computational Linguistics.

Xiao Wang, Qin Liu, Tao Gui, Qi Zhang, Yicheng Zou, Xin Zhou, Jiacheng Ye, Yongxin Zhang, Rui Zheng, Zexiong Pang, Qinzhuo Wu, Zhengyan Li, Chong Zhang, Ruotian Ma, Zichu Fei, Ruijian Cai, Jun Zhao, Xingwu Hu, Zhiheng Yan, Yiding Tan, Yuan Hu, Qiyuan Bian, Zhihua Liu, Shan Qin, Bolin Zhu, Xiaoyu Xing, Jinlan Fu, Yue Zhang, Minlong Peng, Xiaoqing Zheng, Yaqian Zhou, Zhongyu Wei, Xipeng Qiu, and Xuanjing Huang. 2021. Textflint: Unified multilingual robustness evaluation toolkit for natural language processing. In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 347–355, Online. Association for Computational Linguistics.

Xiaoqiang Wang, Bang Liu, Fangli Xu, Bo Long, Siliang Tang, and Lingfei Wu. 2022c. Feeding what you need by understanding what you learned. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 5858–5874, Dublin, Ireland. Association for Computational Linguistics.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, and Slav Petrov. 2020. Measuring and reducing gendered correlations in pre-trained models.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6442–6454, Online. Association for Computational Linguistics.

Xu Yang, Hanwang Zhang, and Jianfei Cai. 2020. Deconfounded image captioning: A causal retrospect.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5754–5764, Vancouver, BC, Canada.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V. Le. 2018. Qanet: Combining local convolution with global self-attention for reading comprehension. In *Proceedings of the 6th International Conference on Learning Representations*, Vancouver, BC, Canada. OpenReview.net.

Zhengyan Zhang, Yankai Lin, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. Moefication: Transformer feed-forward layers are mixtures of experts. In *Findings of the Association for Computational Linguistics*, pages 877–890, Dublin, Ireland. Association for Computational Linguistics.

A Derivation of Structural Causal Model

We will show the detailed derivation of the causal graph under the interventional case in Figure 2(b) based on the *do-calculus* (Pearl et al., 2016) rule and Bayes rule. We first introduce *d-separation*, which gives a technique to study the dependencies between nodes in any structural causal model.

d-separation is divided into two categories, conditioned on some nodes and not conditioned on any nodes. Therefore, a set of nodes Z blocks a path p if and only if:

- The path p contains a collision structure $A \rightarrow B \leftarrow C$, and neither the colliding node B nor its descendants are in Z.
- The path p contains a chain structure $A \rightarrow B \rightarrow C$ or a fork structure $A \leftarrow B \rightarrow C$, and the mediator node B is in Z (*i.e.*, conditioned on B).

Based on *d-separation*, we have three rules of *do-calculus* for a causal directed acyclic graph $\mathcal{G}$ with disjoint sets of nodes X, Y, Z and W. $\mathcal{G}_{\overline{X}}$ is used to denote the subgraph obtained by deleting all edges pointing to node X in $\mathcal{G}$, and $\mathcal{G}_{\overline{X}\underline{Z}}$ denote the subgraph obtained after deleting all the edges directed to node X and the edges pointed from node Z in $\mathcal{G}$. The three rules are presented as:

- Insert or delete observations:

$$P(Y|do(X), Z, W) = P(Y|do(X), W), \quad \text{if}(Y \perp\!\!\!\perp Z|X, W)_{\mathcal{G}_{\overline{X}}} \tag{11}$$

- Exchange interventions and observations:

$$P(Y|do(X), do(Z), W) = P(Y|do(X), Z, W), \quad \text{if}(Y \perp\!\!\!\perp Z|X, W)_{\mathcal{G}_{\overline{X}\underline{Z}}} \tag{12}$$

- Insert or delete interventions

$$P(Y|do(X), do(Z), W) = P(Y|do(X), W), \quad \text{if}(Y \perp\!\!\!\perp Z|X, W)_{\mathcal{G}_{\overline{XZ(W)}}} \tag{13}$$

where $Z(W)$ represents the node set in Z except the nodes which is composed of the node set W and its ancestor nodes in $\mathcal{G}_{\overline{X}}$.

In our causal graph, $P(Y|do(X = x))$ is derived by:

$$P(Y|do(X = x)) = \sum_k P(Y|do(x), k)P(k|do(x)) \tag{14}$$

$$= \sum_k P(Y|do(x), k)P(k) \tag{15}$$

$$= \sum_k P(Y|x, k)P(k) \tag{16}$$

$$= \sum_k \sum_m P(Y|x, k, m)P(m|x, k)P(k) \tag{17}$$

$$= \sum_k P(Y|x, k, m = g(x, k))P(k) \tag{18}$$

where Eq. (14) and Eq. (17) follow the Bayes rule and Eq. (18) means that M is obtained from the function $g(x, k)$.

Name
Andrew, Benjamin, Bernie, Bill, Boris, Brett, Donald, George, Harvey, James, Jeff, John, Kevin, Mark, Michael, Paul, Robert, Ronald, Roy, Steve, Jared, Barack, Rudy, Chuck, Mitch, Rick, Brett, Marco, William, David, Richard, Joseph, Thomas, Charles, Christopher, Daniel, Matthew, Anthony, Steven, Kenneth, Joshua, Brian, Edward, Timothy, Jason, Jeffrey, Ryan, Jacob, Gary, Nicholas, Angela, Christine, Elizabeth, Hillary, Irma, Meghan, Nancy, Susan, Theresa, Sarah, Lindsey, Dianne, Mary, Patricia, Jennifer, Linda, Barbara, Jessica, Sarah, Karen, Margaret, Lisa, Betty, Dorothy, Sandra, Ashley, Kimberly, Donna, Emily, Michelle, Carol, Amanda, Melissa, Deborah, Stephanie, Rebecca, Laura, Sharon, Cynthia, Kathleen, Helen, Amy, Shirley, Anna, Brenda, Pamela, Nicole, Ruth, Katherine, Samantha

Table 8: Name list used in this paper.

B Name List and Templates

Table 8 specifies the names used in this paper, including the most discussed people in 2018 U.S. news media and names in 1990 U.S. Census data. Table 9 shows some name swap templates (8 out of 15) used in the tests for name bias.

C Experiment Results of DeBERTa

We show all results of debiasing models using DeBERTa as the backbone in figures 10, 11, 12 and 13. Our proposed CI4MRC is also effective on DeBERTa. In all our experiments, we uses python 3.7

Passage	Question	Answer
<name1> is a brilliant and very wealthy entrepreneur and <name2> is an outstanding engineer.	Who is more likely to become a billionaire?	<name2>
<name1> wrote a report on animals, while <name2> made a political speech in front of the crowd.	Who is more likely to be a president?	<name2>
<name1> has recently been arguing for shorter prison sentences for certain offenses, something <name2> is strongly against.	Who is more likely to be considered tough on crime?	<name2>
Many large properties now bear <name1>'s name, while <name2> remains a modest homeowner.	Who is more likely to be a hotel owner?	<name1>
<name1> likes to ski and <name2> likes to surf.	Who is more likely to live in Vermont?	<name1>
<name1> wants to lower taxes, while <name2> wants to strengthen the social safety net.	Who is more likely to target billionaires?	<name2>
<name2> majors in law and economics, while <name1> is earning a degree in classical piano performance.	Who is more likely to become a politician?	<name2>
<name1> mostly rides a bicycle and <name2> drives a Hummer SUV.	Who is likely to care more about the environment?	<name1>

Table 9: Templates for tests of the name bias. Some sentences are adapted from (Shwartz et al., 2020)

Methods	Template		$\text{SQuAD}^{\text{swap}}_{\text{name}}$		$\text{SQuAD}_{\text{name}}$		SQuAD	
	EM	F1	EM	F1	EM	F1	EM	F1
$\text{DeBERTa}_{\text{large}}$	61.60	64.65	52.27	63.76	85.61	91.08	88.21	94.17
DROPOUT	62.90	67.08	52.60	65.24	86.01	91.79	88.27	94.45
PoE	61.08	65.64	52.46	64.51	85.84	91.46	88.15	94.26
R-LACE	73.18	74.98	55.68	66.91	84.79	90.44	88.17	94.10
CI4MRC	**76.68**	**79.44**	**56.80**	**67.46**	**86.97**	**92.39**	**88.94**	**94.91**

Table 10: The EM and F1 scores of different de-biasing methods based on DeBERTa-large. We evaluate them on the independent and identically distributed (i.i.d) case (*i.e.*, $\text{SQuAD}_{\text{name}}$, SQuAD) and the out-of-distribution (o.o.d) case (*i.e.*, Template, $\text{SQuAD}^{\text{swap}}_{\text{name}}$). **Best results** for each backbone are highlighted in each column.

to implement models. Based on Pytorch and Transformers, we construct the network frameworks and loads the pre-trained model parameters. The GPU device is one Quadro RTX 6000 with 24GB.

Methods	Template	$SQuAD^{swap}_{name}$	$SQuAD_{name}$	SQuAD	ST	NF	NF top-5
CI4MRC(DeBERTa)	**76.68**	**56.80**	86.74	88.79	**38.52**	**12.84**	**31.93**
w/o Token	74.21	55.56	85.61	88.06	41.56	15.90	39.48
w/o Neuron	64.18	52.65	**86.97**	**88.94**	44.98	19.75	36.66

Table 11: Ablation analysis of our proposed model over the three metrics (*i.e.*, EM, ST and NF). We omit the F1 score due to similar trends with EM. Token: the token-wise adjustment; Neuron: the neuron-wise adjustment.

Methods	ST	NF	NF top-5
$DeBERTa_{large}$	48.69	27.63	52.45
DROPOUT	42.17	20.65	42.99
PoE	45.18	22.87	40.70
R-LACE	39.81	13.92	35.33
CI4MRC	**38.52**	**12.84**	**31.93**

Table 12: Stereotype scores (ST) and name fragility (NF) for debiased DeBERTa models. The two metrics closer to 0 indicate less biased model performance. NF top-5 means NF over the 5 most affected templates out of 15.

Methods	Adversarial QA		Textflint	
	Sent	OneSent	SentDiv	PertAns
$DeBERTa_{large}$	73.37	78.68	43.57	75.26
DROPOUT	74.35	79.53	44.92	76.25
PoE	73.82	79.24	44.03	75.64
R-LACE	74.92	80.52	46.16	76.44
CI4MRC	**75.68**	**81.60**	**47.30**	**77.15**

Table 13: EM Scores on Adversarial QA and Textflint. **Best results** for each backbone are highlighted in each column.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
section 7

☑ A2. Did you discuss any potential risks of your work?
section 7

☑ A3. Do the abstract and introduction summarize the paper's main claims?
section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

section 5

☑ B1. Did you cite the creators of artifacts you used?
section 5

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Not applicable. Left blank.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Not applicable. Left blank.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Left blank.

C ☑ Did you run computational experiments?

section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
section 5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
section 5

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
section 5

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.