Causal-Guided Active Learning for Debiasing Large Language Models

摘要

虽然生成式大语言模型（LLMs）在诸多任务中展现出了优异的性能，但相关研究发现，现有的 LLMs 可能会从数据集中吸收偏见信息，并在生成文本时加以运用，这不仅会削弱其泛化能力，也可能产生潜在的危害。鉴于数据集偏见的复杂多样以及模型过度优化的现状，传统的基于先验知识的去偏方法和基于微调的去偏方法可能不适用于现有的 LLMs。为此，论文提出了一种融合主动学习与因果机制的因果指导主动学习（CAL）框架。该框架借助 LLMs 自身的功能，能够自动筛选出富含偏见信息的关键样本，并主动诱导出偏见模式，进而有效降低偏见的负面影响，并通过高效的上下文学习方法，阻止 LLMs 在生成过程中利用数据集中的偏见。实验结果充分证明，CAL 框架能够精准识别典型的偏见案例并诱导出多种偏见模式，成功消除 LLMs 在文本生成环节的偏见问题。

关键词：大语言模型；因果推断；主动学习

1 引言

文本是人类思想的记录与反映，而性别、种族等固有偏见一直深植于人类认知之中，相应地，这些偏见也在各类语料库及特定任务的数据集中有所体现。LLMs 在生成预训练过程中会广泛利用各种语料库与数据集，这一过程犹如双刃剑，其在汲取丰富知识的同时，也难免将其中蕴含的偏见一并纳入，例如位置偏见和刻板印象偏见等。位置偏见的产生根植于人类潜意识，人们往往下意识地认为第一个选项更为出色，从而使第一个选项在语料库中的出现频率相对较高，LLMs 在对语料库分布进行建模的过程中，便捕捉到了这一偏见关联。这些偏见的存在，会致使 LLMs 的泛化能力受限，并可能引发潜在的有害影响，究其原因，在于 LLMs 于预训练阶段被动地学习语料库中上下文间的关联，而语料库本身恰是人类固有偏好和偏见的映射。

当要求 LLM 对某个选项的优劣进行评估时，它可能会受到位置偏见的影响，倾向于选择排在首位的选项，哪怕该选项的优劣其实与其所处位置并无直接联系。因此，在第二选项于某些情况下更具优势时，LLM 的性能可能会出现明显下滑。此外，语料库中存在的刻板印象偏见等各类偏见，可能会致使 LLMs 生成诸如“女性在 STEM 领域缺乏能力”等有害内容，进而进一步加剧有害的社会刻板印象。

为了解决这些问题，鉴于 LLMs 在模式识别和归纳方面的强大能力，该论文探索性地将主动学习与因果机制相融合，提出了一种因果指导的主动学习（CAL）框架。该框架借助 LLMs

自身的能力，实现对偏见样本的自动自主识别以及偏见模式的诱导。主动学习的核心在于挑选出最具信息量的实例，进而向外部信息源查询以对这些数据点进行标注。在去偏的场景下，CAL 通过锁定 LLMs 未能精准建模上下文间因果不变语义关系的实例来精准识别偏见实例，随后依据数据集偏见对 LLMs 生成内容影响程度的大小，筛选出最具信息价值的偏见实例。因果不变性可用于厘清语义信息与数据集偏见之间的纠缠关系，因为后续文本的内容是由前文的语义所决定的（即体现“因果”关系），而这种关系在各类语料库中普遍存在（即“不变”特征）；反观后续文本与数据集偏见之间的关联，虽有联系，却在不同数据集上呈现出变化多端的特点。

2 研究方法

2.1 因果视角下的文本语料库数据集偏见

如图 1 所示，给定一段文本 X，语料库 D 中的后续文本 Y 会受到两个因素的影响：(1) X 与 Y 之间的语义关系，(2) D 中数据集存在的偏见。例如，给定 X = **医生雇佣了秘书**，因为，由于性别偏见，语料库中的后续文本 Y 更有可能是使用**他**来指代秘书，而不是**她**。这种**偏见关系特征化**了数据集偏见所带来的上下文之间的不当关联。这里将语义关系记作 $f_S(\cdot)$，将偏见关系记作 $g_B(\cdot)$。因此，给定 X，语料库 D 中 Y 在 X 的条件分布可以形式化为 $P(Y|X) = P(f_S(X), g_B(X)|X)$。

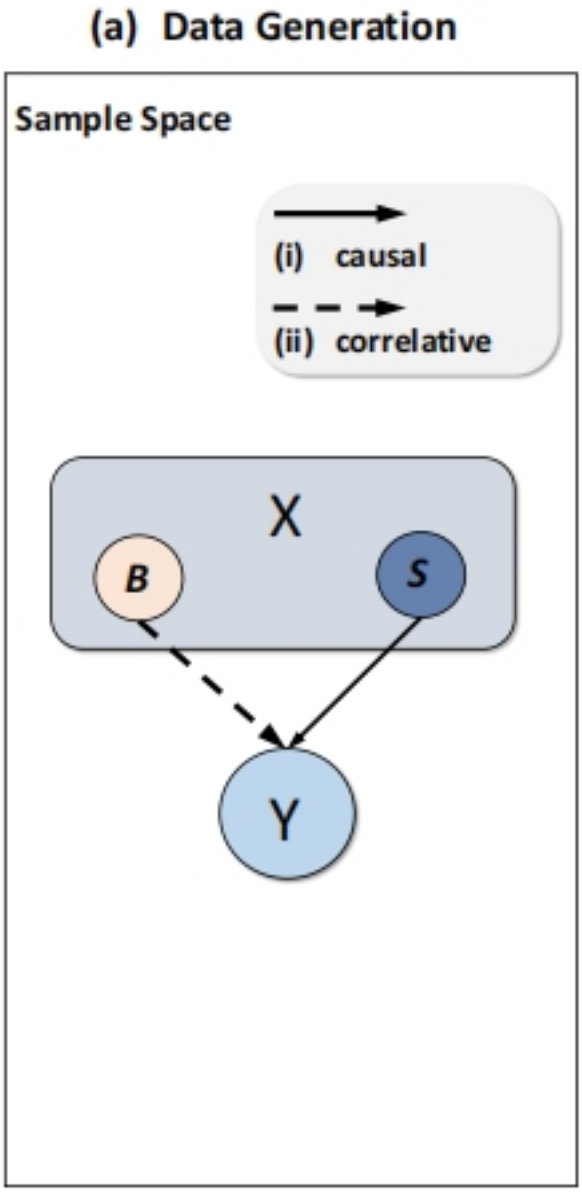

图 1. 从因果角度看数据集偏差

语义关系与偏见关系的关键区别在于，语义关系具有因果不变性，而偏见关系则没有。具体来说，对于所有数据集上的所有实例，给定前面的文本 X，后续文本 Y 将由语义关系决定，而偏见关系仅描述 X 与 Y 之间某种表面的统计关联。考虑一个例子，LLM 作为评审来评估两个 AI 助手的回答，如图 1 所示：答案 (Y) 由提示 X 和答案 Y 之间的语义关系决定。而在语料库中，某些偏见（如答案的位置与答案之间的相关性）可以被预测。然而，Y 实际上并不受偏见决定，这种关联在其他实例中可能无法被预测。因此，由于 Y 由 X 决定，它们的语义关系是一种“因果”关系，这在所有实例中是恒定的，而偏见关系仅仅是相关的，也就是相关性不等于因果性。

2.2 基于因果不变性偏见实例识别

2.2.1 反例对与偏见模式的定义

在预训练和任务特定的监督微调过程中，给定语料库 D 中的 X，Y 的分布可以形式化为 $P(Y|X) = P(f_S(X), g_B(X)|X)$，生成 LLMs 不可避免地被训练来建模 $f_S(X)$ 和 $g_B(X)$。因此，给定前面的文本 X_i，LLMs 不仅会关注 X_i 的语义，还会关注偏见模式，如否定词、性别指示符、选择位置等，以生成 Y。因此，在推理过程中，模型生成的 $\hat{Y}$ 不可避免地受到数据集偏见的影响。为简洁起见，将 X_i 中的语义信息记作 S_i，将偏见模式记作 B_i。

在偏见实例识别的过程中，有两个关键问题：(1) 找出哪个实例包含偏见；（2）找出最具信息性的偏见实例。为了解决上述两个问题，因果指导的主动学习框架将基于因果不变性标准的指导下识别偏见实例，并通过识别数据集偏见对 LLMs 生成影响最大的实例来找到最具信息性的偏见实例。

如图 2 所示，因果指导的主动学习（CAL）包含两个主要组件：(i) 基于因果不变性的偏见实例识别；(ii) 典型偏见实例选择和偏见模式诱导。根据识别出的偏见模式，我们提出了一种基于上下文学习的去偏方法用于对大型语言模型进行正则化。

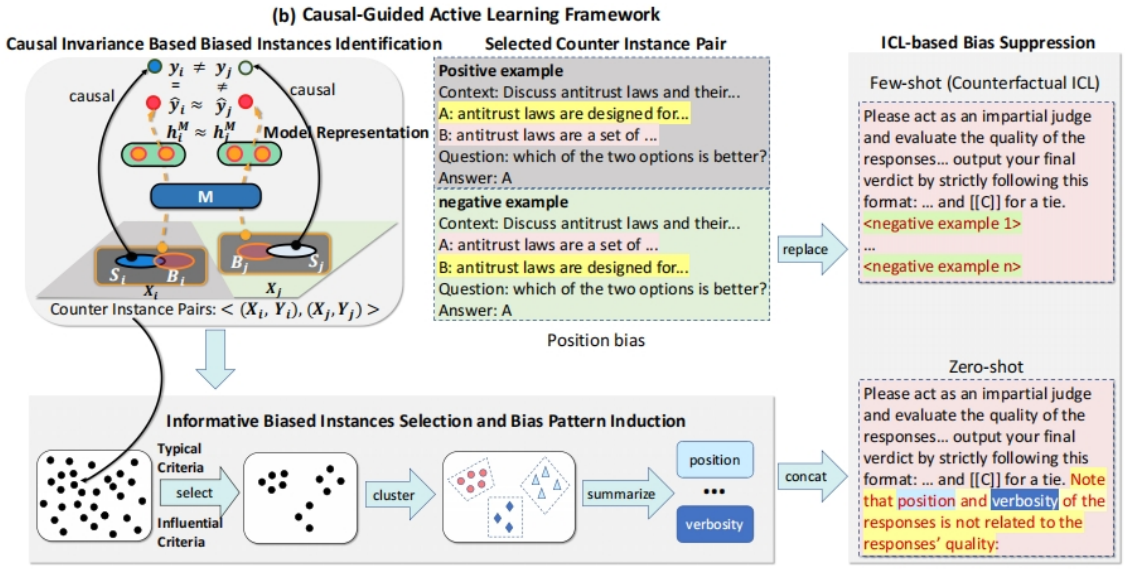

图 2. 因果引导主动学习框架的示意图

首先利用因果方差视角中语义信息与偏见信息之间的差异，识别出一组反映 LLMs 内在偏见的偏见实例。与语义信息相比，偏见信息的本质特征在于 B 与后续文本之间没有不变的因果关系，这使得偏见信息与语义信息得以解耦。注意，生成 LLMs 会捕获偏见信息以获取输入文本的表示（例如，隐藏状态）。因此，如果能够找到模型获取的表示不是不变预测的实例，那么这些实例的表示将包含偏见信息，这表明这些实例很可能包含偏见，也就能被识别为偏见实例。

由于输入前文 X 包含语义 S 和数据集偏见 B，假设对于大型数据集中任意实例 (X_i,Y_i)，可能存在其他实例 (X_j,Y_j)，与 (X_i,Y_i) 具有以下关系：$(B_i, S_i) \subset X_i$，$(B_i, S_i) \subset X_i$，$B_i = B_j$，$S_i \neq S_j$。换句话说，这一对实例几乎共享相同类型的数据集偏见，而输入文本中所隐含的语义信息却不同。这样实例对的存在使得利用因果不变性识别偏见实例成为可能。

在此假设下，考虑一对实例 $(X_i,Y_i),(X_j,Y_j)$，如果模型 M 主要捕获了语义信息 S_i 和 S_j，并且 $H_i^{\mathcal{M}}$ 接近于 $H_j^{\mathcal{M}}$，那么 S_i 与 S_j 相似，因此 $\mathrm{Sim}(Y_i,Y_j) \to 1$。换句话说，LLM 捕获了用于生成的恒定预测信息。

反过来，如果找到一对实例 $(X_i,Y_i),(X_j,Y_j)$，在这一对实例上 $H_i^{\mathcal{M}}$ 接近于 $H_j^{\mathcal{M}}$，而 $\mathrm{Sim}(Y_i,Y_j)$ 较低，则可以将 $(X_i,Y_i),(X_j,Y_j)$ 视为模型 M 违反因果不变性的实例，并且可以利用这一实例对来表征 LLMs 所捕获的偏见。为明确起见，定义这样的实例对 $(X_i,Y_i),(X_j,Y_j)$ 为反例对：

反例对： $\forall(X_i,Y_i),(X_j,Y_j) \in D, i \neq j$，如果

$$S(H_i^{\mathcal{M}}, H_j^{\mathcal{M}}) > \tau, s.t. Sim(Y_i, Y_j) < \alpha,$$

那么 $(X_i,Y_i),(X_j,Y_j)$ 能被视为一组反例对。其中 D 是数据集，S(•) 是测量 $H_i^{\mathcal{M}}$ 与 $H_j^{\mathcal{M}}$ 之间相似性的评分函数，τ 是控制 $H_i^{\mathcal{M}}$ 和 $H_j^{\mathcal{M}}$ 可以视为足够接近的阈值，而 α 是另一个阈值，确保 Y_i 和 Y_j 可以被视为充分不同。

上述反例对的定义可以用于检测数据集 D 中的所有反例对。在这些反例对上，不变性被违反，从而后续文本是基于偏见信息生成的。因此，$H_i^{\mathcal{M}}$ 和 $H_j^{\mathcal{M}}$ 包含偏见信息 $B_i = B_j$。然而，上述理论是建立在 LLMs 已捕获预测信息（包括偏见和语义信息）的假设之上的。实际上，当 X_i 非常困难或模糊时，不能排除 LLM 没有捕获任何预测信息的可能性。为了排除这些实例，我们引入了一个额外的过滤过程，使用预测标准，该标准要求模型 M 至少能对实例 i 或 j 进行合理生成，因为如果在 i 和 j 上模型生成不合理，那么模型在 X_i 或 X_j 上没有捕获任何预测信息的可能性很大：

$$Sim(\hat{Y}_i, Y_i) > \beta \vee Sim(\hat{Y}_j, Y_j) > \beta,$$

其中 $\hat{Y}i$ 和 $\hat{Y}j$ 是生成的后续文本，β 是一个阈值，确保 $\hat{Y}i$ 和 Y_i 可以被视为足够相似，从而可以视为正确答案（对于 $\hat{Y}j$ 也相同）。

2.2.2 典型偏见实例识别

首先，对于任何输入文本 X_i，如果 Y_i 被正确生成的概率相当低，说明偏见信息显著阻碍了 LLM 的表现。因此，这样的例子将包含较高程度的偏见，可以被视为信息性偏见实例。其次，对于反例对 $(X_i,Y_i),(X_j,Y_j)$，如果 LLMs 的对应生成结果 $\hat{Y}i$ 和 $\hat{Y}j$ 相当不同，这表示数据集偏见的影响是多样化的，因此基于这些样本总结一个统一的偏见模式将是具有挑战性的。相反，如果 $\hat{Y}i$ 和 $\hat{Y}j$ 相似，则更容易得出偏见造成的影响，因为数据集偏见的影响是典型的。基于这两个特征，引入以下两个标准来选择信息性偏见实例：

$$\textbf{影响标准：}\ \hat{p}_{j,l_j} < \tau_p, s.t. Sim(\hat{Y}_j, Y_j) < \alpha,$$

$$\textbf{典型标准：}\ Sim(\hat{Y}_i, \hat{Y}_j) > \beta,$$

其中 l_j 是黄金后续文本，$\hat{p}i$、lj 是黄金后续文本的预测概率，$\tau_p \in [0,1]$ 是控制模型 M 生成黄金后续文本概率的阈值。

2.2.3 偏见模式诱导

基于识别出的信息性偏见实例，进一步诱导出某些可解释的模式，以表征语料库中几种主要类型的数据集偏见。为此，首先将反例对分组为几个集群，然后为每个集群诱导偏见模式。

反例对的集群是基于反例对的偏见表示向量派生的，偏见表示向量是反例对的偏见部分的表示向量。我们通过提取两个示例（即 $H_i^{\mathcal{M}}$ 和 $H_j^{\mathcal{M}}$）表示中的相似部分来获取反例对 $(X_i,Y_i),(X_j,Y_j)$ 的偏见表示向量。这是因为，如反例对的定义所述，$H_i^{\mathcal{M}}$ 和 $H_j^{\mathcal{M}}$ 的相似部分携带偏见信息。

在获取每个反例对中的偏见表示向量后，应用主成分分析将偏见表示向量的维度降至两维。随着数据维度的增加，数据点之间的距离变得越来越相似，因此传统的距离度量（如欧几里得距离）将变得不太有效，从而影响聚类算法的性能。然后，使用基于密度的聚类方法 DBSCAN，根据降维后的偏见表示向量进行聚类。最后，获得每个集群内的反例对，并将其提供给大语言模型（原文使用 GPT-4）以总结偏见模式。例如，从图 2 中选择的反例对，可以总结为位置偏见。

2.3 基于上下文学习的偏见抑制

为防止 LLMs 在生成过程中利用数据集偏见，同时避免基于微调方法的缺点，可一种具有成本效益和高效的基于上下文学习（ICL）的方法。具体而言：

在零样本场景中，如图 2 所示，我们使用自动诱导的偏见模式明确告诉 LLM 在推理过程中不应使用何种信息，通过在原始提示的结尾附加文本“【偏见 xxx】与【任务目标】无关”。

在少样本场景中，使用一种反事实 ICL 方法，向 LLMs 提供自动派生的反事实示例以纠正 LLM 对偏见的认知。具体而言，如果能够找到“反事实示例”，使用偏见信息进行推理将导致错误的生成。那么，通过在提示中向 LLMs 提供此类示例，LLMs 将被隐式告知偏见信息与后续文本无关，从而被正则化为不使用偏见信息进行推理。为了寻找这样的“反事实示例”，根据影响标准，对于任意反例对 $(X_i,Y_i),(X_j,Y_j)$，在实例 i 或 j 上，LLM 将产生不当生成。为简便起见，将此实例记为 i，实例 i 可以被视为去偏的反事实示例。直观地说，在实例 i 中，数据集偏见导致了不当生成，这与语料库中的大多数情况相反，因此我们将实例 i 称为反事实示例。

3 复现细节

3.1 实验环境配置

在复现过程中，选择使用使用 llama2-13B-chat 和 vicuna-13B-v1.5 大模型进行实验。为了不失去一般性，选择具有明确回答集合的数据集进行实验，这样，就可以使用字符串的精确匹配来实现上文中的 Sim(•) 函数，如果匹配，则函数值为 1；否则为 0。因此，α 和 β 可以是介于 0 和 1 之间的任何值。此外，通过采用 LLM 层顶部最后一个 token 的嵌入向量来获取输入文本的表示，余弦函数被用作评分函数 S(•) 来测量这些隐藏状态之间的相似性。

为了获得反例对的偏见表示向量，需要提取反例对两个示例对应的隐藏状态中的相似部分。这是因为，隐藏状态中的相似部分携带了偏见信息。为此，通过逐元素的方式获取两个隐藏状态的相似组分。具体而言，使用以下函数：

$$f(H_{ik}, H_{jk}) = \begin{cases} (H_{ik} + H_{jk})/2 & if \frac{|H_{ik} - H_{jk}|}{H_{ik} + H_{jk}} < \mu \\ 0 & otherwise \end{cases}$$

其中 H_{ik} 和 H_{jk} 是两个隐藏状态 $H_i^{\mathcal{M}}$ 和 $H_j^{\mathcal{M}}$ 的第 k 个元素。如果 H_{ik} 和 H_{jk} 足够相似，则它们的差异应该相对较小。可以使用 $|H_{ik} - H_{jk}|/|H_{ik} + H_{jk}|$ 来测量这种差异，然后使用阈值 μ 来确定 $|H_{ik} - H_{jk}|/|H_{ik} + H_{jk}|$ 是否足够小，换句话说，H_{ik} 和 H_{jk} 是否足够相似。如果它们足够相似，我们使用 H_{ik} 和 H_{jk} 的平均值来表示反例对的偏见表示向量的第 k 个元素。如果不相似，则用 0 表示偏见表示向量的第 k 个元素。在实践中，通过控制在使用 llama2-13B-chat 时某一位置上的两个元素被视为足够相似的比例来选择 μ。在 MNLI 数据集中，将该比例的严格阈值设置为 0.15，以确保反例对的偏见表示向量具有更纯粹的偏见信息。

由于原论文中已经证明，使用数据集的 20% 子集和使用整个数据集对进行偏见识别和去偏影响不大，因此为提高效率，复现实验中也采用 20% 的数据集进行实验。

	MNLI	HANS
ZS	65.9	52.9
ZS-CAL	67.4	55.5
ZS-CAL(20%)	67.1	55.4
FS	66.1	53.1
FS-CAL	64.1	59.3
FS-CAL(20%)	64.0	59.7

图 3. 原文使用 20% 子集进行偏差模式归纳实验结果

3.2 生成式 LLM 特定偏见识别与去偏

将 Chatbot 和 MT-Bench 数据集作为基准。在这两个数据集上，LLM 需要从两个候选响应中选择一个更好的响应。实验中先在 Chatbot 数据集上诱导偏见模式，然后测试基于 Chatbot 的偏见模式是否可以用于在 Chatbot 和 MT-Bench 数据集上去偏 LLMs。

在 Chatbot 数据集上，

在 Chatbot 数据集，基于 CAL 框架，共找到 13896 反例对，经过 Qwen2.5-7B-Instruct 模型总结，可归纳为 7 种类别的偏见，如图 4 和图 5 所示。

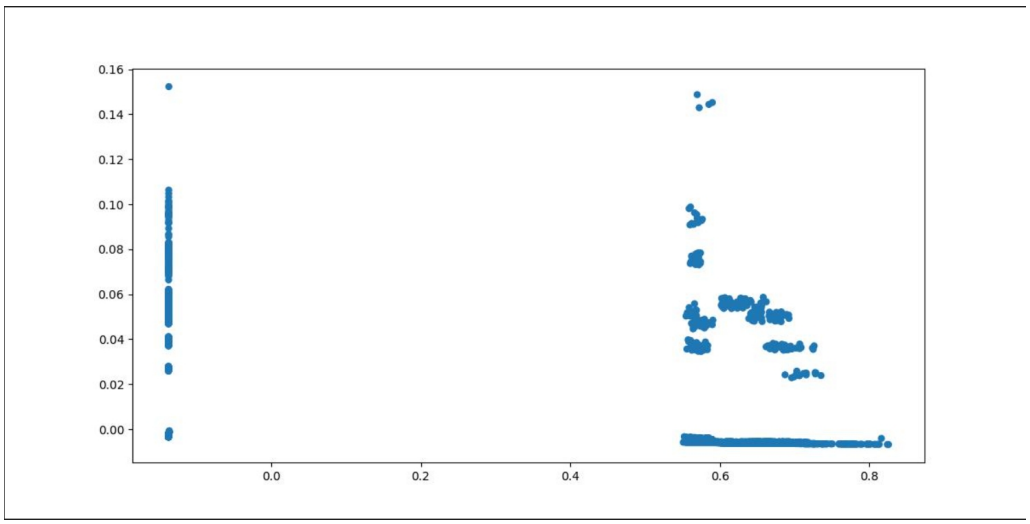

图 4. 初始 PCA 降维结果

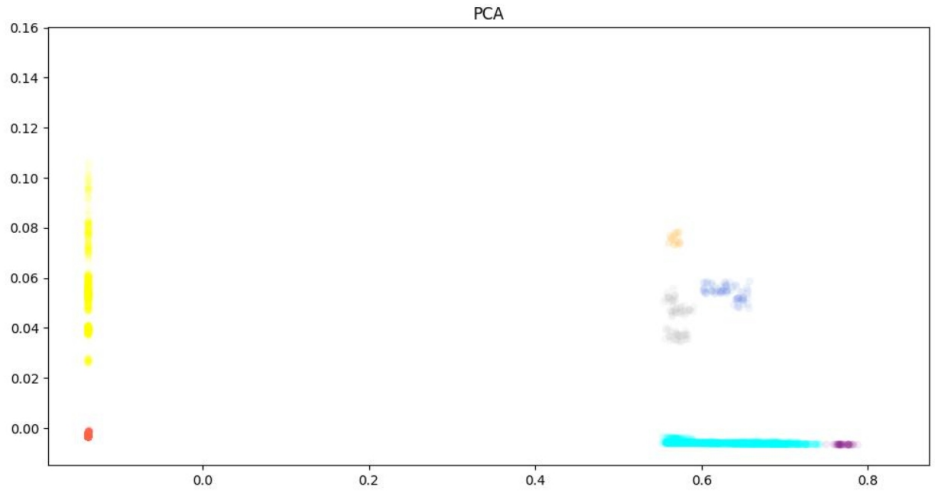

图 5. 加入 DBSCAN 聚类后的 PCA 结果

3.3 任务特定偏见识别与去偏

选择自然语言推理数据集 MNLI 和相应手动去偏的数据集 HANS 作为基准。因此，仅利用偏见信息的模型在 HANS 上通常表现接近随机基线。实验中先从 MNLI 数据集中诱导偏见模式，然后测试 CAL 是否能够利用诱导的偏见模式在 MNLI 和 HANS 数据集上去偏 LLMs。

在 MNLI 数据集上，基于 CAL 框架，共找到 29051 反例对，45 对负例，经过 Qwen2.5-7B-Instruct 模型总结，可归纳为 6 种类别的偏见，如图 6 和图 7 所示。

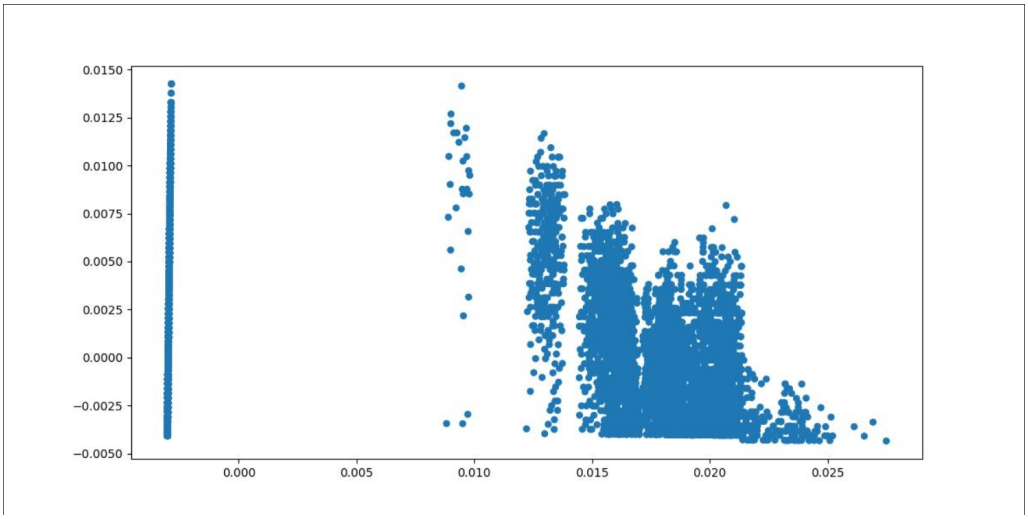

图 6. 初始 PCA 降维结果

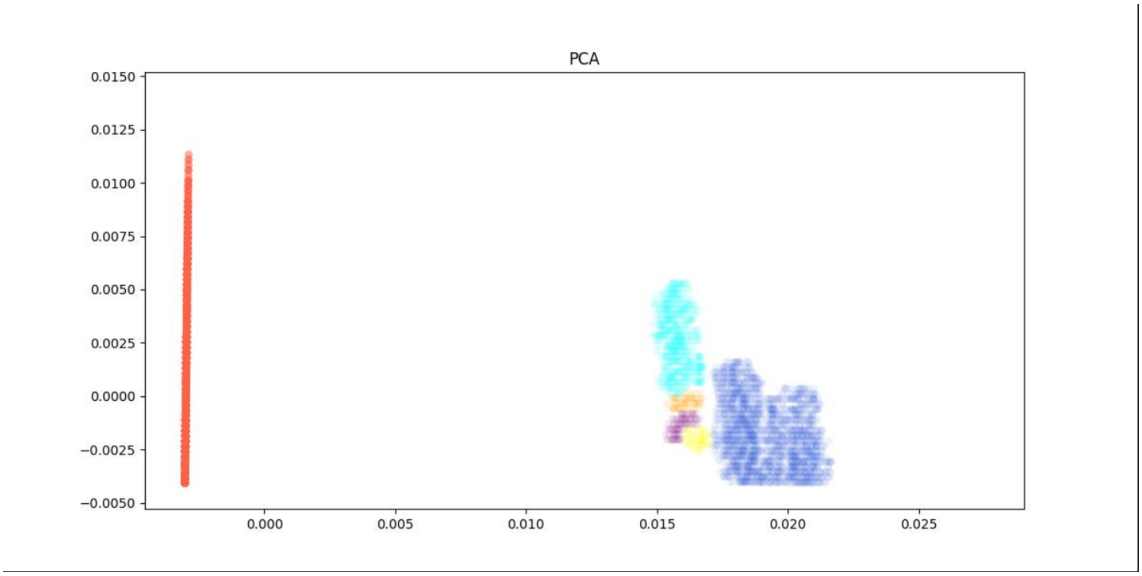

图 7. 加入 DBSCAN 聚类后的 PCA 结果

3.4 响应无害性评估

BBQ 和 UNQOVER 数据集旨在评估 LLMs 的响应中是否存在刻板印象偏见（如性别偏见和种族偏见）。两个数据集分别包含 9 种和 4 种类型的刻板印象偏见。在这两个数据集上，如果模型获得更高的准确率，则可以认为其包含刻板印象的可能性较低。

对于 BBQ 和 UNQOVER 数据集上的零样本 CAL 方法，我们尝试使用去偏提示“[偏见 xxx] 与 [任务的目标] 无关”。实验结果见表 6。尽管与零样本基线方法相比，性能有所提升，但 UNQOVER 数据集的表现远低于零样本已知偏见方法。我们怀疑这个问题是由于去偏提示模板造成的。因此，我们在零样本 CAL 方法中将去偏提示模板替换为零样本已知偏见方法中使用的模板“我们应该平等对待来自不同诱导刻板偏见模式和其他偏见模式的人。当我们没有足够的信息时，我们应该选择未知选项，而不是基于我们的刻板印象或其他偏见模式进行假设。”我们发现这导致性能显著提升。因此，为了在零样本 CAL 方法与零样本已知偏见方法之间进行公平比较，我们在主要实验中使用了这个模板。

在 BBQ 数据集上，基于 CAL 框架，共找到 22004 反例对，29 对负例，经过 Qwen2.5-7B-Instruct 模型总结，可归纳为 3 种类别的偏见，如图 8 和图 9 所示。

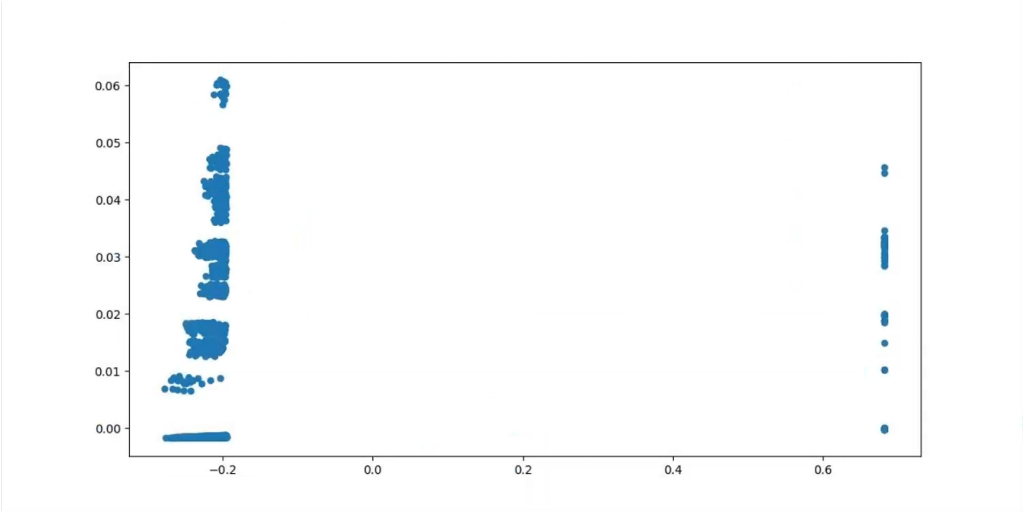

图 8. 初始 PCA 降维结果

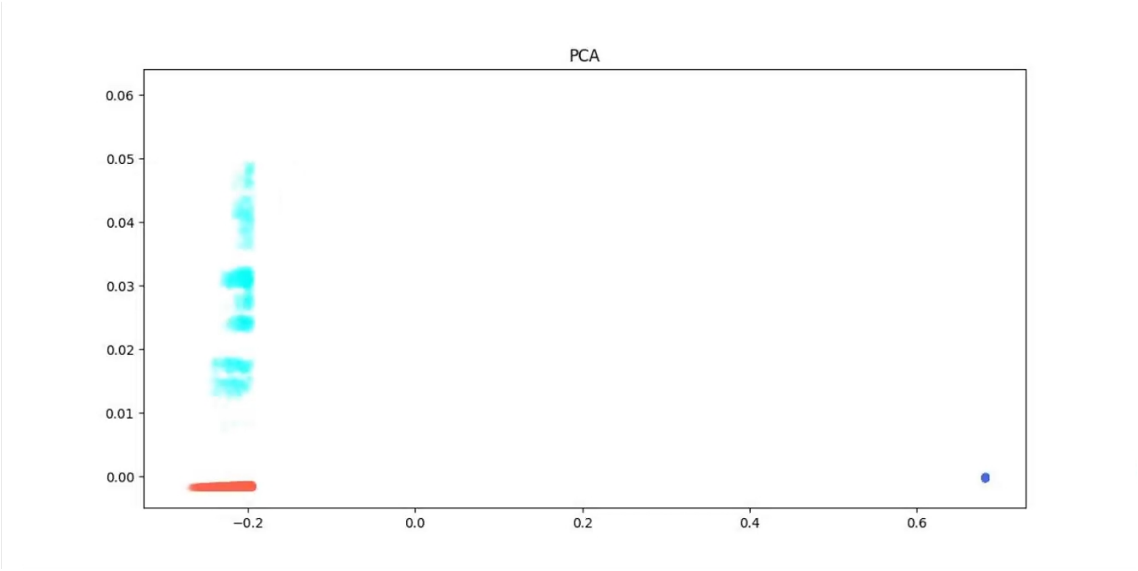

图 9. 加入 DBSCAN 聚类后的 PCA 结果

4 实验结果分析

复现实验结果如下：

基于两个 LLMs 在六个数据集上的实验结果，可以发现：

(1) 与原生零样本相比，基于先验知识的零样本去偏方法在所有数据集上普遍表现出改进的性能。这表明，通过上下文学习（ICL），LLMs 不仅可以有效去偏，还可以避免与微调方法相关的分布内性能下降，从而表明基于 ICL 的去偏方法的优越性。

(2) 与零样本基线和少样本基线相比，总体而言，少样本 CAL 在两类基准上实现了一致的性能提升。这表明 CAL 可以提高 LLMs 的泛化能力和无害性，并且通过利用语义

表 1. 大语言模型在零样本和少样本场景中 CAL 与基线的比较。ZS、ZS-known、FS 分别表示零样本、已知偏差的零样本和少样本。

	Chatbot	**MT**	**MNLI**	**HANS**	**BBQ**	**UQ**
LLAMA2						
ZS	38.9	34.5	65.9	52.9	47.6	23.4
ZS-known	42.7	41.2	67.2	55.0	51.1	59.4
FS	40.4	46.9	66.1	53.1	49.3	23.1
ZS-CAL	40.3	43.3	67.3	55.4	51.4	60.1
FS-CAL	41.8	49.8	64.1	59.2	53.4	32.4
Vicuna						
ZS	35.2	43.8	66.7	38.3	47.9	33.3
ZS-known	38.2	50.0	69.6	55.0	49.9	35.2
FS	37.3	46.9	71.0	62.5	59.7	48.5
ZS-CAL	39.8	50.1	69.7	57.0	48.4	35.3
FS-CAL	39.8	49.3	69.4	63.7	65.4	58.4

信息之间的基本差异，CAL 能够识别一组偏见实例，而基于反事实的 ICL 提示可以有效利用偏见反事实示例去偏 LLMs。

(3) 与原生零样本基线相比，零样本 CAL 在所有数据集上都能持续提高模型性能，甚至在部分基准上超过了少样本方法的表现。零样本 CAL 的有效性表明，CAL 诱导的偏见模式是典型的，确实存在于数据集中。这是因为，通过利用因果不变性结合影响与典型标准，可以选择出一组典型偏见实例，从而有效诱导出偏见模式。

(4) 与基于先验知识的零样本去偏方法相比，零样本 CAL 在两类基准上显示出可比或更好的性能。一方面，数据集偏见的分布复杂性给精确和全面检测潜在偏见带来了挑战。另一方面，零样本 CAL 与基于先验知识的零样本去偏方法之间的可比性能显示了我们方法的有效性，以及在现实场景中的应用潜力，因为在各种现实语料库中调查所有偏见将是不切实际的。

5 总结与展望

5.1 总结

复现的论文提出了一种因果指导的主动学习框架。依赖于数据集偏见和语义在因果不变性中的差异，该框架可以自动识别包含偏见的反例对。在复现实验中，利用影响标准和典型标准选择更具信息性的反例对以诱导偏见模式，并使用一种节省成本和有效的基于 ICL 的去偏方法，防止 LLM 在生成时利用偏见。实验结果表明，该方法能够有效自动识别各种偏见模式，并对 LLMs 进行去偏以增强其泛化能力和无害性。

5.2 不足与展望

在原文的实验中，作者发现在 HANS 数据集上，零样本 CAL 方法和少样本 CAL 方法对每种偏见的去偏效果存在较大差异。例如，在词汇重叠，子序列重叠和句子成分偏见三种偏见中，零样本 CAL 方法对句子成分偏见类别的效果不佳，但对词汇重叠和子序列偏见类别有效，而少样本 CAL 方法对三类偏见类别中均有效。

造成零样本 CAL 方法在某些偏见中效果较差的原因可能与提示词有关。在零样本 CAL 去偏场景中，提示词为 “【偏见 xxx】与【任务目标】无关”，实验中可以观察到，当提示词中包含两个以上偏见模式的去偏提示可能会导致性能下降，而单独使用一种去偏提示会导致性能提升。

参考文献

[1] Zhouhao Sun, Li Du, Xiao Ding, Yixuan Ma, Yang Zhao, Kaitao Qiu, Ting Liu, and Bing Qin. Causal-guided active learning for debiasing large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14455–14469, Bangkok, Thailand, August 2024. Association for Computational Linguistics.