

基于可控性解释的混合数据增强框架*

孙泽辰, 肖义胜, 李俊涛, 张 民, 周国栋

(苏州大学 计算机科学与技术学院, 江苏 苏州 215008)

通信作者: 李俊涛, E-mail: ljt@suda.edu.cn



摘 要: 先前的预训练语言模型已在众多自然语言理解任务中展现了其卓越的性能. 然而, 它们常表现出捷径学习的问题, 即学习了非鲁棒性特征与标签之间的虚假关联, 导致模型在不同于训练分布的测试场景中的泛化能力不佳. 近期, 生成式预训练大模型在理解任务中的出色表现引起了广泛的关注, 但它们是否受到捷径学习的影响尚未被充分研究. 以 LLaMA 系列模型与 FLAN-T5 模型为代表, 探究生成式预训练大模型在多个自然语言理解任务中的捷径学习现象. 研究结果表明, 近期流行的生成式大模型仍然存在捷径学习的问题. 进而, 提出针对生成式预训练大模型的捷径学习问题的缓解策略——基于可控性解释的混合数据增强框架. 该框架以数据为中心, 基于模型生成的可控性解释数据与部分原始提示性数据构造小规模混合数据集, 开展模型微调. 在 3 个具有代表性的自然语言理解任务中的大量实验结果表明, 使用该框架所构造的数据集训练模型能够有效缓解模型的捷径学习问题, 提升模型在分布外测试场景中的鲁棒性与泛化能力, 同时不牺牲甚至提升模型在分布内测试场景中的性能. 代码已公开发布在 <https://github.com/Mint9996/HEDA>.

关键词: 捷径学习; 生成式预训练语言模型; 自然语言理解

中图法分类号: TP18

中文引用格式: 孙泽辰, 肖义胜, 李俊涛, 张民, 周国栋. 基于可控性解释的混合数据增强框架. 软件学报. <http://www.jos.org.cn/1000-9825/7215.htm>

英文引用格式: Sun ZC, Xiao YS, Li JT, Zhou GD. Hybrid Data Augmentation Framework Based on Controllable Explanation. Ruan Jian Xue Bao/Journal of Software (in Chinese). <http://www.jos.org.cn/1000-9825/7215.htm>

Hybrid Data Augmentation Framework Based on Controllable Explanation

SUN Ze-Chen, XIAO Yi-Sheng, LI Jun-Tao, ZHANG Min, ZHOU Guo-Dong

(School of Computer Science and Technology, Soochow University, Suzhou 215008, China)

Abstract: Previous pre-trained language models (PLMs) have demonstrated excellent performance in numerous tasks of natural language understanding (NLU). However, they generally suffer shortcut learning, which means learning the spurious correlations between non-robust features and labels, resulting in poor generalization in out-of-distribution (OOD) test scenarios. Recently, the outstanding performance of generative large language models (LLMs) in understanding tasks has attracted widespread attention, but the extent to which it is affected by shortcut learning has not been fully studied. In this paper, the shortcut learning effect of generative LLMs in three NLU tasks is investigated for the first time using the LLaMA series models and FLAN-T5 models as representatives. The results show that the shortcut learning problem still exists in generative LLMs. Therefore, a hybrid data augmentation framework is proposed based on controllable explanations as a mitigation strategy for the shortcut learning problem in generative LLMs. The framework is data-centric, constructing a small-scale mix dataset composed of model-generated controllable explain data and partial original prompting data for model fine-tuning. The experimental results in three representative NLU tasks show that the framework can effectively mitigate shortcut learning, and significantly improve the robustness and generalization of the model in OOD test scenarios while avoiding sacrifice of or even improving the model performance in in-distribution test scenarios. The solution code is available at <https://github.com/Mint9996/HEDA>.

* 基金项目: 国家自然科学基金 (62206194); 江苏省自然科学基金 (BK20220488)

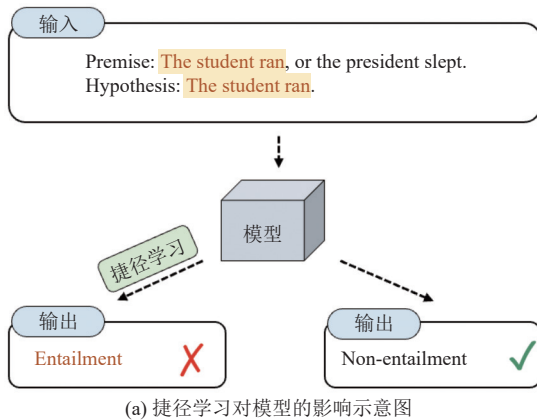
孙泽辰和肖义胜为共同第一作者.

收稿时间: 2023-10-18; 修改时间: 2024-02-03, 2024-03-27; 采用时间: 2024-04-15; jos 在线出版时间: 2024-06-20

Key words: shortcut learning; generative pre-trained language model; natural language understanding

预训练语言模型,例如 BERT^[1]、RoBERTa^[2]等,在经过任务特定数据集的微调之后,往往能够在多种自然语言理解任务中展现出优秀的性能.比如, BERT 模型在使用 MNLI 数据集^[3]进行微调之后,在自然语言推理任务上的准确率可以达到 84.6%^[4].然而,当测试集与训练集的数据分布不一致时,模型的性能将显著下降^[5].研究者们认为造成该现象的一个主要原因是模型的捷径学习 (shortcut learning)^[6],即模型依赖于数据集中捷径进行预测.捷径具体表现为非鲁棒性特征与目标标签之间的虚假关联,往往极易被模型捕捉.如图 1(a) 所示,以自然语言推理任务为例,在训练阶段,模型学习了前提句 (premise) 和假设句 (hypothesis) 之间的单词重叠与标签“蕴含 (entailment)”之间虚假关联,即学习到了捷径;因此,在推理阶段,模型倾向于通过捷径将存在单词重叠的样本标记为“蕴含”关系,而非通过语义进行预测.当测试数据与训练数据的分布不同时,捷径学习可能导致错误的预测,降低了模型的预测准确率.为了更好地揭示模型的捷径学习现象,研究者们构造了用于检测捷径学习的测试场景(如 HANS 数据集^[7]和 MNLI-Hard 数据集^[8]),并且通过对比模型在分布内测试场景与这些具有挑战性的分布外测试场景中的性能差异,反应模型的捷径学习程度^[6].图 1(b) 的实验数据直观地反映了经过 MNLI 数据集训练后的 BERT 模型的捷径学习问题:在分布内测试场景中,该模型的准确率高达 84.3%,而在 HANS 数据集集中的准确率仅为 61.1%,相差 23.2%.这个显著的性能差异表明, BERT 模型受捷径学习问题的影响较为严重.总之,捷径学习问题反映了模型在训练中倾向于学习虚假关联,并且利用捷径进行推理,该问题严重损害了模型在分布外数据中的鲁棒性与泛化性^[9],需要引起研究者的关注和重视.

虚假关联:在训练集中,标签“entailment”与句子中的词汇重叠之间存在联系.



Model	MNLI	HANS	MNLI-Hard
BERT-base	84.3	61.1	75.9
LLaMA (7B)	89.9	72.3	85.8

注: HANS 与 MNLI-Hard 均为常用于探究捷径学习的测试集. HANS 主要用于检测词汇重叠捷径 (lexical-overlap bias); MNLI-Hard 主要用于检测仅假设捷径 (hypothesis-only bias).

(b) 捷径学习对模型的性能影响

图 1 捷径学习的表现与影响

近年来,已有一些关于模型捷径学习问题的探索和研究.众多研究显示,使用参数规模更大的模型、更先进的预训练策略或者更为平衡的训练数据,均有助于缓解模型的捷径学习问题^[10,11].然而,我们注意到上述研究主要围绕 BERT 类别式预训练模型展开,在模型规模以及结构的变种上存在明显的局限性.同时,随着计算资源的扩展和预训练数据量的增加,生成式预训练大模型,如 LLaMA^[12]、LLaMA 2^[13]、FLAN-T5^[14]等,已在多个自然语言理解任务中显示出卓越的潜力,性能超越了以往的最优模型.尽管这些生成式大语言模型拥有强大的泛化能力,但是它们训练过程中是否仍存在捷径学习的问题尚不明确.因此,为了解答上述疑惑,我们基于近期具有代表性且使用广泛的生成式预训练大模型 LLaMA 系列模型与 FLAN-T5 模型,探索并分析了该类模型在多个自然语言理解任务中的捷径学习现象,并与传统的 BERT 类别式模型进行比较.

我们在图 1(b) 展示了 LLaMA 模型的捷径学习问题的部分实验结果.实验结果表明,经过 MNLI 数据集微调后, LLaMA 模型在分布内外测试场景中的准确率相比于 BERT 模型均有所提升.然而, LLaMA 模型在 MNLI 分布内测试集与 HANS 测试集中的准确率差值仍高达 17.6%,表明该模型仍存在明显的捷径学习问题.基于上述探

索, 我们提出了基于可控性解释的混合数据增强框架, 以缓解生成式大模型的捷径学习问题, 提升模型的鲁棒性与泛化性。具体来说, 我们首先采样一定数量的训练数据赋予提示 (prompt), 以便生成式语言模型能够更好地完成自然语言理解任务, 该部分数据被称为提示性数据; 其次, 我们将提示性数据分为两部分, 一部分输入 ChatGPT 模型 (GPT-3.5-turbo)^[15] 生成可控性解释, 并使用生成的解释替换样本的原始目标标签, 我们称替换后的样本为可控性解释数据; 最后, 我们将另一部分提示性数据与可控性解释数据以相等比例混合, 形成基于可控性解释的混合数据集用于模型训练。在多个自然语言理解任务中的实验结果表明, 使用我们的框架构造的数据集微调生成式大语言模型, 有效缓解了模型的捷径学习问题, 提升了模型在分布外测试场景中的鲁棒性。

本文的贡献可以总结如下。

- 探索并分析了近期流行的生成式大语言模型在多个自然语言理解任务中的捷径学习问题。实验结果表明, 模型的捷径学习问题仍然存在, 影响了模型在分布外测试场景中的鲁棒性。

- 提出了基于可控性解释的混合数据增强框架以缓解生成式大模型的捷径学习问题。该方法不需要依赖关于捷径的先验知识, 且适用于不同的任务与模型。在 3 个流行的自然语言理解任务上的广泛实验表明, 该框架能够有效缓解模型的捷径学习问题, 显著提升了模型在多个分布外的测试场景中的鲁棒性, 并且不会牺牲甚至提升模型在分布内测试场景中的性能。

- 进行了大量的分析实验, 探究训练数据的规模、提示性数据与可控性解释数据的混合比例、解释性数据的获取方案等因素对模型捷径学习的影响。我们发现, 小规模可控性解释数据的使用能够有效缓解模型的捷径学习问题。并且, 对于生成式大模型来说, 过多的训练数据可能在较为简单的任务中加剧模型的捷径学习问题。我们建议利用高质量、多样性强的小规模数据集来缓解模型的捷径学习, 提升模型鲁棒性的同时节约计算资源。

1 相关工作

模型的捷径学习问题已经在众多自然语言理解任务中暴露出来。许多研究表明, BERT 类预训练模型在自然语言推理任务中的预测依赖于数据集中的虚假关联^[8,16]。同样, 对于阅读理解任务, 模型依赖于问题与原文中的词汇匹配进行预测, 而忽略了对其语义特征的理解^[17]。捷径学习问题严重影响了模型在现实场景下的表现和性能^[9], 损害了模型的鲁棒性与泛化性, 因此越来越多的研究致力于缓解模型的捷径学习问题。

以往关于缓解模型捷径学习问题的方法主要分为以下两个思路: 以模型为中心和以数据为中心。以模型为中心的缓解方法侧重于在模型的训练阶段, 通过改变模型架构或者权重, 以减少模型对捷径特征的依赖。Liu 等人^[18]和 Han 等人^[19]利用人工标注与捷径先验知识, 强迫模型专注于鲁棒性特征; Clark 等人^[20]、He 等人^[21]、Sanh 等人^[22]提出在 PoE (product-of-experts) 框架的基础上, 结合纯偏差模型与主模型, 抑制模型学习捷径。张大操等人^[23]提出因果去偏推理方法并设计对比学习模块, 实现了针对情景感知的捷径学习问题缓解方法。此外, 在模型训练中对核心特征赋予更高的权重也有助于缓解捷径学习问题^[24,25]。而以数据为中心的研究思路则专注于从本质上减少数据集中的虚假特征所占比例。Wu 等人^[26]提出了一种基于 z 统计 (z-statistics)^[27]的过滤机制, 用于筛选出数据集中带有虚假特征的数据样本。多种数据增强方法同样也被应用于缓解捷径学习问题, 如 Kaushik 等人^[28]使用反事实数据增强方法, 使用反事实数据进一步训练模型以增强其鲁棒性, Si 等人^[29]提出对抗性和混合数据增强方法用于提升模型的对抗攻击能力。相比于以模型为导向的缓解方法, 基于数据的方法通常不会增加额外的训练模块或者训练目标, 适用场景更为广泛^[30,31]。

然而局限的是, 面向自然语言理解任务, 以往关于捷径学习问题的探索与缓解大多基于判别式预训练语言模型 (如 BERT、RoBERTa 等)。目前, 生成式语言模型通过基于提示的微调方法已被证明能够出色地完成自然语言理解任务^[32,33]。同时, 随着计算资源的增长和预训练架构的优化, 拥有更大模型参数的生成式大语言模型已经在多种理解任务展现强大的能力, 甚至超过了以往的最优结果^[12]。那么, 使用近期流行的生成式大语言模型进行微调, 是否能够缓解模型在自然语言理解任务中的捷径学习问题值得探索。因此, 我们基于当前广泛使用且具有代表性的 LLaMA 系列模型与 FLAN-T5 模型展开实验, 首次探索了该类模型在多个自然语言理解任务中的捷径学习问

题. 然而, 多个场景下的实验表明, 捷径学习问题仍然明显存在于微调后的生成式大模型, 影响模型的鲁棒性与泛化性, 需要得到缓解.

基于上述探索, 我们以数据为中心提出了面向生成式预训练大模型捷径学习问题的缓解方案: 基于可控性解释的混合数据增强框架. 数据的解释 (explanation) 在自然语言的多个场景中发挥着重要的作用, Wei 等人^[34]、Lampinen 等人^[35]探索了解释性数据对于提高模型推理能力的作用. 在使用解释性数据对于模型鲁棒性的影响方面, Stacey 等人^[36]提出利用解释性数据监督注意力权重的方法, 并证明该方法能够同时提升模型在分布内外测试集中的性能. Chen 等人^[37]与 Schuster 等人^[38]证实了使用解释性数据能够提升模型对于对抗攻击的鲁棒性. 与上述工作不同的是, 我们专注于使用解释性数据缓解生成式大模型的捷径学习问题. 并且, 借助当前生成式大模型强大的生成能力, 我们通过可控性提示工程使 ChatGPT 模型自动生成样本的解释性数据, 而不依赖人工标注或者抽取式方法. 最后, 我们提出了数据混合的策略, 将可控性解释数据与提示性数据进行混合, 以构造多样性更强的混合数据集用于微调模型. 我们的方法有效地缓解了生成式大语言模型的捷径学习问题, 提升了模型在分布外测试场景中的鲁棒性, 同时未牺牲甚至提升模型在分布内测试场景中的性能.

2 基于可控性解释的混合数据增强框架

本节详细介绍了我们提出的基于可控性解释的混合数据增强框架. 内容分两部分: 方法的实现流程以及与其他方法的对比分析.

2.1 方法实现流程

本节详细阐述了基于可控性解释的混合数据增强框架的具体实现流程. 如图 2 所示, 我们展现了该框架的整体工作流程. 接下来我们将从混合数据集的构建过程及模型的微调过程两个方面进行详细的介绍.



图 2 框架总体操作流程

2.1.1 数据集的构造

图 3 中展示了我们提出的框架中混合数据集的具体构造过程, 主要包括以下步骤.

- 采样数据: 对于当前的大规模语言模型来说, 如 LLaMA 系列模型, 基于小规模数据集的微调足以激发其对特定任务的理解和推理能力. 因此, 结合对算力开销的考虑, 我们仅从原始的大规模训练集中选取小部分数据进行实验. 具体来说, 我们按照原始数据集的标签分布进行随机采样, 最终选取了 4 000 个样本作为我们的训练数据.

- 生成提示数据: 参照现有工作的常规做法^[39], 我们将原始数据样本赋予提示模板, 并将融合了提示模版后的数据简称为提示数据. 如图 3(1) 所示, 该提示模板将数据由原始的分类任务形式转化为生成任务形式, 即自然语言问句, 以更加适配生成式大语言模型的预训练目标, 进而更好地激发模型预训练阶段学到的知识. 以自然语言推理任务为例, 该任务的目的是预测输入的句子对中假设句 (hypothesis) 和前提句 (premise) 之间的语义关系为蕴含 (entailment)、中立 (neutral) 还是矛盾 (contradiction). 通过使用特定提示模板, 例如“Premise: [前提句]”. Question: Does this imply that “Hypothesis: [假设句]”? Yes, no, or maybe?>, 可以使原始的假设句和前提句结合成一个自然的问题, 模型随后生成答案. 答案集合 {yes, maybe, no} 与原始标签集合 {entailment, neutral, contradiction} 一一对应. 在本文实验中, 我们从人为构造的高质量的提示资源库 (PromptSource package)^[40]中挑选相应模板.

- 生成解释数据: 得益于当前大语言模型的强大生成能力, 我们仅需设计合适的提示模板以指导模型生成高质量的解释数据. 已有研究表明, 提示模板对于大语言模型的生成质量至关重要^[41]. 据此, 我们基于以下原则设计

提示模板: 1) 简洁性: 提示模板须在明确表达任务要求的同时, 保持简洁和清晰; 2) 普适性: 提示应适用于各种自然语言理解任务, 尽可能保持模板形式的一致性; 3) 可控性: 提示模板应引导模型对于正确的答案生成合理的解释, 避免出现错误的解释. 我们将带有真实标签引导的提示工程称为可控性提示方法. 如图 3(2) 所示, 在实际实验中, 我们使用了<Why the answer is “[label]”?>的提示模板, 以指导生成式大模型 ChatGPT 自动生成可控性解释数据.

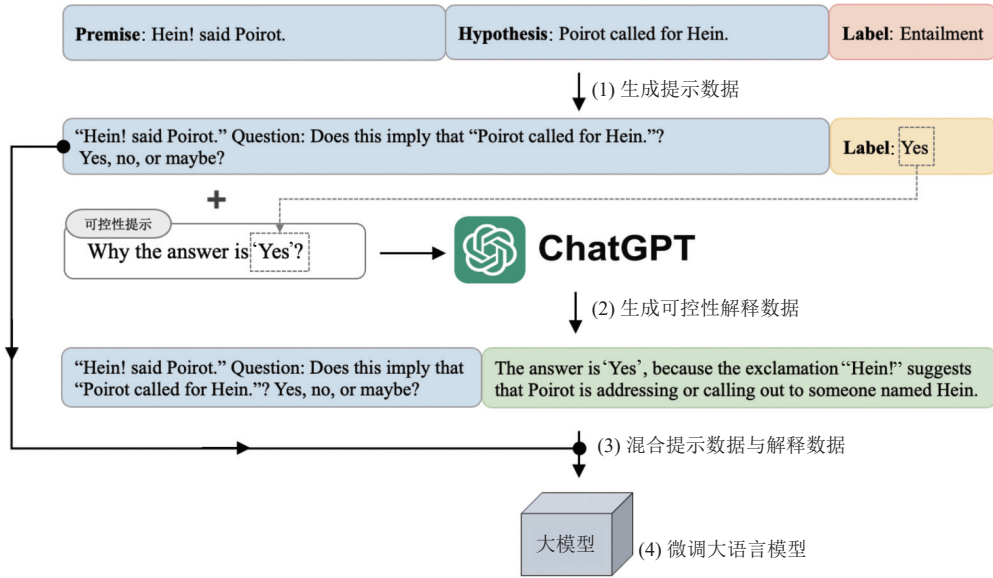


图 3 数据集具体构造过程

- 构造混合数据集: 已有工作完全使用解释性数据集微调生成式语言模型来提高模型的鲁棒性^[37,42], 经过训练的模型将在生成标签的同时严格生成对于输入数据与该标签之间的因果解释. 然而我们认为, 完全使用解释数据进行训练并非最佳选择. 由于样本中解释文本在长度上远大于标签文本, 全解释数据训练可能使得模型过分专注于解释部分, 偏离了原始推理任务的目标, 进而影响最终结果. 因此, 我们选择将一部分解释数据与原始提示数据混合进行训练. 具体来说, 在获得提示数据集之后, 我们先从提示数据集中抽取一部分数据用于生成解释数据, 随后将其与剩余未含解释内容的提示数据混合, 并用该混合数据集进行模型的最终训练. 实验中, 我们采用了 1:1 的混合比例, 并在后续的分析实验中验证了这一比例为最佳设置.

2.1.2 微调模型

我们基于上述获得的带有可控性解释的混合数据集, 对生成式自回归模型进行微调. 以仅解码器结构的模型 LLaMA 为例, 该模型是一个基于 Transformer 架构的大规模预训练语言模型, 由多个 Transformer 的解码器模块组成, 并通过自回归方法生成文本序列. 在预训练阶段, 该模型使用了大量无标签文本数据进行训练, 任务是在给定上文后预测下一个词, 目标是最大化给定上文下一词的对数似然度. 具体来说, 给定一个包含 N 个训练词符的训练文本实例 $U = \{U_1, U_2, \dots, U_N\}$, 训练目标为最大化如下损失函数 L_{pre} :

$$L_{pre} = \sum_{U_i \in U} \log P(U_i | U_{<i}); \theta,$$

其中, U_i 为训练文本 U 中的每一个词符, $U_{<i}$ 为训练文本 U 中在 U_i 之前的所有词符, $P(U_i | U_{<i}; \theta)$ 为模型在参数在 θ 下给定上下文 $U_{<i}$ 预测 U_i 的概率.

微调过程与预训练过程类似, 给定一个经过特定提示模板处理后的训练实例, 该实例包含了提示语句序列 S 以及目标标签序列 Y , 训练目标为最大化如下损失函数 L_{ft} :

$$L_{ft} = \sum_{Y_i \in Y} \log P(Y_i | S \oplus Y_{<i}); \theta,$$

其中, Y_i 为目标标签序列 Y 中的每一个词符, \oplus 为拼接操作, 其他符号的含义与上述一致. 注意, 在微调过程中, 尽管提示语句序列和目标标签序列需要拼接起来一起输入模型, 模型仅学习预测目标标签序列中的词符, 这是因为在推理预测阶段, 提示语句序列是给定的, 模型仅需要预测目标标签部分.

在具体的测试过程中, 我们首先利用相应的提示模板将原始样本整体作为输入, 然后要求模型在句子最后生成对应的标签. 鉴于提示模板可能对模型的预测性能产生一定影响^[41], 我们在所有实验中均报告了使用 3 种不同提示模板的平均结果作为最终实验结果. 更多细节将在第 3 节呈现.

2.2 方法对比与分析

2.2.1 解释数据生成方案对比

以往相关工作获得的解释数据主要来源于人工标注与情景学习方法 (in-context learning, ICL)^[43]. 人工标注方法对于人力资源要求较大, 且众包标注的过程可能为数据集引入偏差^[11]. Ludan 等人^[42]在少样本 (few-shot) 环境下使用情境学习方法通过模型获得解释数据, 但这同样需要部分人工标注的解释数据样本作为示例. 与这些方法相比, 使用生成式模型根据提示模板自动生成解释数据可以极大地节省人力资源.

受到思维链 (chain-of-thought, COT)^[34]方法的启发, 该方法通过在模型完成提示模板的末尾加入 <Let's think step by step>, 引导模型进行逐步思考并最终输出答案和推理过程, 我们可以获取模型的思考过程作为对应的解释数据. 然而, 使用 COT 方法生成的解释性数据可能会存在误导性, 因为原始 COT 方法没有正确答案的引导, 可能会产生错误的答案及其对应的解释. 与之对比, 如图 4(a) 和 (b) 所示, 可控性提示方法借助正确答案的引导, 可确保解释数据的准确性.

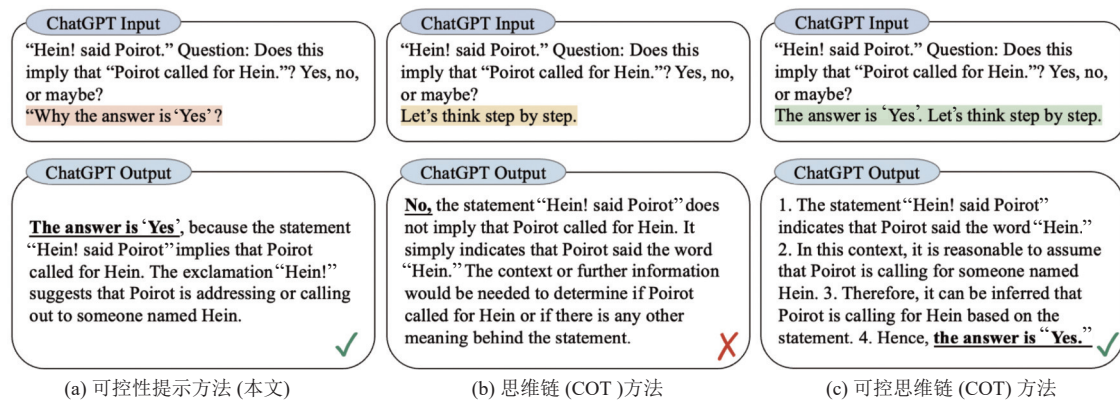


图 4 不同方法生成解释数据示例对比

值得一提的是, 如图 4(c) 所示, 我们也可以将可控性提示与 COT 方法相结合, 即使用带有真实标签引导的 COT 方法生成解释数据. 具体操作为在提示模板的末尾加入 <The answer is "[label]". Let's think step by step.>的提示. 案例分析显示, 这种组合方法同样可以保证解释数据的准确性, 但与单独的可控性提示方法相比, 结合 COT 的方法可能会使解释数据更侧重于逐步解释, 即生成的是详尽的推理路径. 而通过我们设计的可控性提示诱导模型, 能够保证模型在已知正确答案的前提下生成可靠且可控的解释, 并且解释的内容更为简洁明了. 在第 3.5 节的实验分析中, 我们将对比这 3 种不同方法所得到的解释数据构造的混合数据集对缓解模型捷径学习的效果.

2.2.2 模型微调与预测方法对比

在传统 BERT 类别式预训练模型的标准微调与预测过程中^[1], 给定特定句子对 $\langle s_1, s_2 \rangle$, 研究者首先使用特殊词符 [CLS] 与 [SEP] 将句子对接为 "[CLS] s_1 [SEP] s_2 " 的形式, 之后输入模型. 经过多个模型层之后, 该方法将模型最后一层输出中对于 [CLS] 词符的嵌入 (embedding) 通过特定分类器头 (specific classifier head) 进行标签预测. 该分类器头将 [CLS] 词符的嵌入经过线性分类层, 输出对于每一个特定标签的概率, 最后将概率最大的标签类别作为推理结果.

随后, Lester 等人^[44]提出了使用提示的方法微调 BERT 类模型的技术. 该类方法将特定句子对<s1, s2>转换为例如“[CLS] s1 ? [MASK] s2 [SEP]”的形式输入模型. 与上述使用 [CLS] 词符的嵌入预测标签的方法不同, 该类方法使模型直接在 [MASK] 词符对应的位置生成特定标签. 测试过程同样如此, 在模型输出 [MASK] 词符位置对应生成概率分布之后, 将标签集合中对应生成概率最大的标签类别作为结果. 这种提示方法与 BERT 模型在预训练阶段的掩码语言模型 (masked language model, MLM)^[1]任务相符, 即预测被 [MASK] 标记的输入句子位置的词符, 从而在多个下游任务上取得了效果提升.

然而, 当我们使用生成式预训练大模型进行微调时, 同样需要与模型预训练阶段的学习任务相匹配. 该类模型在预训练阶段中的任务是基于上下文预测下一个词, 同时不使用 [MASK] 标签. 所以我们需要利用特定提示模板将句子对转换为问答的方式, 这样使得模型可以在提示输入之后生成预测的标签. 在测试过程中, 我们将通过特定提示模板转换后的句子输入模型, 然后模型将会输出对应的回答作为类别标签.

此外, 仅解码器结构的模型和编码器-解码器结构的模型在训练和推理过程中存在细微差别. 对于仅解码器结构的模型来说, 其流程正如第 2.1.2 节中所提, 该模型将问题直接输入解码器随后逐字生成答案. 而对于编码器-解码器结构的模型来说, 问题部分将被输入编码器, 随后解码器将逐字生成答案.

3 实验

3.1 实验数据

我们选取 3 个经过充分研究且广为采纳的自然语言理解任务展开实验, 分别是自然语言推理任务 (natural language inference, NLI)、事实验证任务 (fact verification) 以及释义识别任务 (paraphrase identification). 参照先前的研究思路^[6], 我们通过比较模型在与训练数据同分布和不同分布的测试场景中的性能来探究模型的捷径学习程度. 与训练数据同分布的测试场景通常为该数据集本身的验证集或测试集, 也被称为分布内数据集 (in-domain data, ID), 与其具有不同分布的测试场景通常包含较多不能通过捷径做出正确预测的测试样本, 这类数据集被称为分布外数据集 (out-of-domain data, OOD)^[45]. 表 1 中同时给出了各个任务对应数据集的统计信息, 下面将对各个数据集进行详细的介绍.

表 1 实验所用数据信息

任务	名称	分割	数量
自然语言推理	MNLI	Train	392 702
	MNLI-m dev	ID Test	9 815
	HANS		30 000
	MNLI-Hard-m	OOD Test	4 573
	MNLI-Hard-mm		4 530
事实验证	FEVER	Train	242 911
	FEVER dev	ID Test	16 664
	FEVER Symm. v1		956
	FEVER Symm. v2	OOD Test	712
释义识别	QQP	Train	363 870
	QQP dev	ID Test	40 430
	PAWS_QQP	OOD Test	677

(1) 自然语言推理任务

- MNLI (the multi-genre natural language inference corpus, 多类型自然语言推理数据库)^[3]数据集是一个被广泛使用的自然语言推理任务数据集, 包含由前提和假设组成的句子对以及二者之间的语义关系 (即标签). 具体来说, 如果假设能够从前提中引申出来, 则标签为蕴含 (entailment); 如果假设与前提无关, 则标签为中立 (neutral); 如果假设与前提相矛盾, 则标签为矛盾 (contradiction). MNLI 数据集涵盖各种领域和文本风格, 所以其测试集分为匹配的 (matched) 和不匹配的 (mismatched) 两个版本, 下面简称 MNLI-m 与 MNLI-mm. 由于原始测试集未提供真实标

签, 我们使用该数据集的验证集作为对应的分布内测试集.

- HANS^[7]是一个专门用于评估自然语言推理任务中模型对表层句法偏差的依赖性问题的受控数据集, 可作为 MNLI 的分布外测试集. HANS 数据集与 MNLI 的格式类似, 同样也由假设句和前提句组成, 但是其标签仅包含蕴含 (entailment) 与非蕴含 (non-entailment) 两个类别. 值得注意的是, 该数据集中每一个样本的前提句和假设句之间均有大量的词汇重叠, 按照由大到小的递进关系可分为词汇重叠 (lexical-overlap)、子序列 (subsequence) 和成分 (constituent) 这 3 种类型, 而其真实标签为蕴含和非蕴含的样本各为 15 000 条. 以往的研究表明, 模型极易倾向于通过前提与假设之间的词汇重叠, 即捷径, 将其关系预测为蕴含. 因此, 如果模型存在捷径学习问题, 在 HANS 数据集中的具体表现为: 在真实标签为蕴含的测试样本中的预测准确率极高, 而非蕴含类别的准确率极低, 最终总体性能不佳.

- MNLI-Hard^[8]的构建主要用于检测自然语言推理任务中假设句偏差 (hypothesis-only bias) 导致的捷径学习现象, 即模型可能仅依赖于假设句来进行预测, 忽略了假设和前提之间的语义关系. 该现象的主要原因通常是假设句中的一些单词与标签之间存在较强的虚假指向性, 例如否定词通常与矛盾标签强相关. 因此研究者使用仅由假设句训练的模型来预测 MNLI 的测试样本, 并使用预测错误的样本集合构建了 MNLI-Hard 测试集. 同样, MNLI-Hard 测试集也被分为匹配的与不匹配的两个版本, 下面简称 MNLI-Hard-m 与 MNLI-Hard-mm, 均作为 MNLI 的分布外测试集. 实验中, 如果模型在 MNLI-Hard 中的性能与在 MNLI 的分布内测试集的性能差异越大, 则反映该模型的捷径学习问题越严重.

(2) 事实验证任务

- FEVER (fact extraction and verification)^[46]数据集是当前在事实提取和验证任务上研究最充分、结构最正式的数据集, 也是使用最广泛的数据集之一. 该任务旨在为在给定证据 (evidence) 的背景下评估声明 (claim) 的有效性, 其标签包括支持 (support)、反驳 (refutes) 或者信息不足 (not enough information). 由于该数据集的测试集真实标签未公开, 因此我们使用 FEVER 训练集微调模型, 并且使用其验证集作为分布内数据集以测试模型性能.

- Fever-Symmetric 由 Schuster 等人^[47]提出, 目的是验证模型对仅声明偏差 (claim-only bias) 这类捷径的依赖程度. 仅声明偏差与自然语言推理任务中的仅假设偏差类似, 来自声明中出现的某些单词和短语与某类标签之间存在着虚假关联. 因此如果模型未能真实理解证据与声明之间的语义关系, 而是仅依赖声明中的虚假关联作为捷径, 在分布外的测试场景中极易出现预测错误的现象. 该数据集的两个版本, 简称 Symm. v1 与 Symm. v2, 均可作为事实验证任务的分布外测试集以测试模型对捷径的依赖程度.

(3) 释义识别任务

- QQP (Quora question pair) 数据集由社区问答网站 Quora 中的问题对的集合组成, 任务是判断一个问题对中的两个问题是否能够相互转述. 举例来说, 在一个样本中, 如果问题 1 能够被问题 2 所转述, 则该样本的标签为重复 (duplicate), 否则即为非重复 (not-duplicate). 我们使用 QQP 的训练集对模型进行释义识别任务的微调, 并使用其验证集作为对应的分布内测试集.

- PAWS (paraphrase adversaries from word scrambling)^[48]的 QQP 子集同样由问题对的集合组成, 然而两个问题之间的单词高度重叠, 但其语义关系大多不可相互转述, 即真实标签为非重复. 先前的实验结果指出, 使用 QQP 数据集训练的模型在该测试场景中的表现甚至不如随机结果 (random baseline), 这表明模型严重依赖于词汇重叠捷径进行推理. 我们同样使用 PAWS 数据集作为释义识别任务的分布外测试集以测试模型是否存在捷径学习现象.

3.2 评价指标与基准模型

在本文中, 我们采用了常用的评价指标准确率 (accuracy, %) 来评估模型性能, 准确率越高, 表明模型的性能越好. 由于捷径学习主要影响模型在分布外测试场景中的鲁棒性, 使模型在分布内外的测试场景中表现出较为明显的性能差异. 因此, 我们使用分布外平均准确率 Avg. (average accuracy in OOD) 表示模型在各个分布外测试场景中的平均表现, 以此反映模型的鲁棒性. 同时我们使用性能的下降 Dec. (decline) 表示模型在分布内测试场景的准

准确率与分布外平均准确率 Avg. 的差值, 这有助于直观地说明模型捷径学习的程度. 具体来说, Avg. 值越高, 表明模型在分布外测试场景中的鲁棒性越强; Dec. 值越低, 则表明模型的捷径学习问题程度越低. 注意, 由于 Dec. 值同时与模型在分布内外测试场景中的准确率均有关, 如果模型在分布内外数据中的准确率均较低, 那么即使 Dec. 值较小, 我们也不认为模型的捷径学习问题得到缓解. 因此, 研究模型的捷径学习问题需要根据 Avg. 值和 Dec. 值共同评估. 我们的目标是在不牺牲模型分布内准确率的前提下, 提升模型在分布外测试场景中的准确率 (Avg. 值), 进而缩小 Dec. 值.

为了全面地评估生成式模型的捷径学习程度, 我们选取了两种最广泛使用的判别式基准模型与 3 种最近具有代表性的生成式基准模型作为对比, 分别是 BERT-base (110M)^[11]、RoBERTa-base (125M)^[2]、LLaMA (7B)^[12]、LLaMA 2 (7B)^[13] 以及 FLAN-T5-xl (3B) 模型^[14]. 同时, 为了验证我们的方法对于缓解生成式预训练模型的捷径学习问题的有效性, 我们分别使用 LLaMA 系列模型与 FLAN-T5-xl 模型作为主干进行实验. 关于模型的简要介绍如下.

- BERT-base (110M) 模型是一个典型的双向编码模型. 其在预训练过程中, 使用了掩码语言模型和下一句预测任务分别捕捉词语和句子级别的表示, 拥有强大的语言表征能力和特征提取能力.

- RoBERTa-base (125M) 模型基于 BERT 模型进一步改进, 拥有更多的模型参数与预训练数据, 同时在预训练过程中删除了下一句预测任务. 这些改进增强了其泛化能力, 使之在众多下游任务中达到更好的性能.

- LLaMA (7B) 模型作为近期开源的具有代表性的生成式大语言模型之一, 该模型在众多基准测试中表现出色. 该模型网络基于 Transformer 架构进一步优化, 使用预归一化 (pre-normalization)、SwiGLU 激活函数和 RoPE 旋转嵌入等先进的预训练策略, 最终表现出了更强大的鲁棒性与泛化能力. LLaMA 2 (7B) 是 LLaMA 系列中的一个升级版本, 采用了更先进的训练技术和数据处理方法, 在处理复杂语言任务时展现出了更高的效率和更准确的输出.

- FLAN-T5-xl (3B) 模型是基于 T5 预训练模型进行多任务指令微调得来的模型. 该模型使用少量的指令数据使其具备完成多种任务的能力, 并且能够充分发挥在其预训练阶段学到的知识. 该模型通过模仿多种任务格式的示例, 增强了其在特定任务上的表现, 如文本摘要、问答和翻译等.

3.3 实验方法

对于判别式预训练模型 BERT 与 RoBERTa 的微调实验, 设置如下: 我们使用 AdamW 优化器, 并将学习率 (learning rate) 设置为 $1E-5$, 最大训练轮次 (epoch) 设置为 10, 并根据验证集的表现挑选最终模型用于推理. 我们将所有实验在不同的种子 (seed) 上面运行 5 次, 最终报告 5 次实验的平均结果.

生成式模型 LLaMA 的微调实验使用 Fairseq library^[49], 以设置并调节参数. 对于训练过程的超参数, 我们采用 Adam 优化器, β 设置为 (0.9, 0.98), 权重衰减设置为 0.01. 我们将学习率设置为 $2E-4$, 模型将在 200 次更新的预热 (warm up) 后进行多轮微调. 我们微调 3 个轮次并根据验证集的表现挑选最终模型用于推理. 在推理阶段, 我们将生成温度 (generation temperature) 设置为 0.2. 所有训练与推理实验均使用 2 张 NVIDIA 3090 显卡进行. 我们使用高效微调方案 LoRA^[50] 代替全参数微调来减小硬件需求.

对于模型 FLAN-T5 的实验, 我们使用 Huggingface 库进行. 具体来说, 我们使用 AdamW 优化器并设置学习率为 $5E-5$. 我们同样微调 3 个轮次并根据验证集的表现挑选最终模型用于推理. 所有训练和推理的实验均使用 4 张 NVIDIA 3090 显卡进行.

对于各个任务, 我们均从提示资源库 (PromptSource package)^[40] 中选择特定任务的 3 个提示模板, 进行特定任务的微调实验, 最终报告模型在 3 个不同提示下的平均性能.

3.4 实验结果与分析

为了探究生成式预训练大模型的捷径学习问题, 并证明本文所提出的缓解框架的有效性, 我们重点研究了以下两个问题.

- RQ1: 生成式预训练大模型是否仍然存在捷径学习问题, 其程度如何?
- RQ2: 基于可控性解释的混合数据增强框架能否有效缓解生成式模型的捷径学习问题?

RQ1: 生成式预训练大模型是否仍然存在捷径学习问题, 其程度如何?

为了研究生成式大模型在自然语言理解任务中的捷径学习问题, 我们将 LLaMA 系列模型与 FLAN-T5 模型与两个被广泛使用的 BERT 类别式模型进行对比, 并在 3 个常见的自然语言理解任务中进行了实验. 实验结果如表 2 所示, 表中以粗体显示最佳性能.

表 2 捷径学习问题的研究结果 (%)

Method	自然语言推理任务						事实验证任务						释义识别任务		
	ID		OOD				ID		OOD				ID	OOD	
	MNLI-m	HANS	MNLI-Hard-m	MNLI-Hard-mm	Avg.	Dec.	FEVER	Symm. v1	Symm. v2	Avg.	Dec.	QQP	PAWS	Dec.	
BERT-base	84.3	61.1	75.9	76.8	71.3	13.0	85.8	57.9	64.4	61.2	24.7	90.7	49.5	41.2	
RoBERTa-base	87.3	71.2	80.6	81.4	77.7	9.6	86.4	68.0	68.2	68.1	18.3	91.1	51.0	40.1	
LLaMA (7B)	89.9	72.3	85.8	86.0	81.4	8.5	88.8	71.5	72.8	72.2	16.6	92.8	53.8	39.0	
LLaMA 2 (7B)	90.6	69.8	86.2	86.4	80.8	9.8	90.2	72.6	73.0	72.8	17.4	89.4	57.0	32.4	
FLAN-T5-xl	87.7	66.9	83.8	84.8	78.5	9.2	94.0	79.4	80.8	80.1	13.9	90.2	50.3	39.9	

实验结果表明: (1) 生成式大模型 LLaMA 系列模型与 FLAN-T5 模型在分布内和分布外测试场景中的性能仍然表现出显著的差距 (其中, LLaMA 模型在 3 个任务中的 Dec. 值分别为 8.5%、16.6%、39.0%; LLaMA 2 模型的 Dec. 值分别为 9.8%、17.4%、32.4%; FLAN-T5 模型的 Dec. 值分别为 9.2%、13.9%、39.9%), 表明该类模型仍然存在严重的捷径学习问题. (2) 与 BERT 类模型相比, 当前的生成式大模型受到捷径学习的影响程度略低, 并且表现出了更强的理解能力与泛化能力, 尤其是在分布外数据场景中 (其中, 与 RoBERTa 相比, LLaMA 模型在 3 个任务中的 Avg. 值分别提升了 3.7%、4.1%、2.8%, Dec. 值分别降低了 1.4%、1.7%、1.1%). 这可能得益于近期的生成式大模型具有更大的参数规模与更先进的预训练策略. (3) 不同的生成式大模型之间的表现也略有差异, LLaMA 系列模型在 NLI 任务与释义识别任务中的表现相对较好, 而 FLAN-T5 模型在事实验证任务中表现较好. 这可能 LLaMA 系列模型与 FLAN-T5 模型具有不同的架构、参数规模以及预训练技术有关.

此外, 从捷径类型的角度分析, 我们使用的分布外测试集主要关注两种类型的捷径: (1) HANS 数据集与 PAWS 数据集主要关注模型对词汇重叠捷径的利用, 即模型倾向于通过句子对中的重叠单词判断二者之间的关系; (2) MNLI-Hard 数据集与 Fever-Symmetric 数据集主要关注模型对于单句捷径的利用, 即模型倾向于仅通过句子对中的单个句子中的一些单词进行预测. 从表中可以看出, 各模型在关于词汇重叠的测试集上 (类型 1) 相对于单句捷径 (类型 2) 的准确率明显较低 (在 LLaMA 模型中, HANS 和 MNLI-Hard-m 之间的准确率差异达到 13.5%). 同时, 模型在分布内测试集与 (1) 类重叠捷径之间的性能差异也更加显著 (LLaMA 中, HANS 与 MNLI-m 之间的准确率相差 17.6%). 实验结果表明, 相比于单句捷径, 词汇重叠捷径将对模型产生更大的影响, 针对词汇重叠捷径的测试场景对模型更具有挑战性.

综上所述, 生成式预训练大模型确实表现出了具有更强的应对捷径学习问题的潜力, 同时表现出了更强的泛化性与鲁棒性. 尽管如此, 模型在分布内外测试场景的显著性能差异表明了捷径学习问题依旧影响着模型的鲁棒性和泛化能力. 因此, 我们的探索对未来关于捷径学习的研究工作提供了有价值的指导. 并且, 现有的缓解捷径学习的工作主要集中于 BERT 类别式预训练模型, 并没有针对生成式预训练大模型所设计的缓解方法, 这进一步鼓励我们探索适用于缓解生成式大模型捷径学习问题的方案. 因此, 我们提出了基于可控性解释的混合数据增强框架, 并在 RQ2 中证明了我们所提缓解框架的有效性.

RQ2: 使用本文提出的基于可控性解释的混合数据增强框架, 能否有效缓解模型的捷径学习问题?

为了解释生成式预训练大模型在自然语言理解任务中的捷径学习问题, 我们从数据的角度出发, 提出了基于可控性解释的混合数据增强框架, 生成混合训练集 D_{mix} 用于微调生成式模型. 在实验中, 我们比较了使用 D_{mix} 微调后的模型与使用原始提示性数据 D_{pro} 微调后的模型在各个场景中的性能, 具体实验结果见表 3. 需要说明的是, 表 3 中的 w/ D_{pro} (all) 表示使用带有提示性数据的完整训练集 (即大规模提示性数据集) 微调模型; w/ D_{pro} (4k) 表

示使用仅抽取的 4 000 条带有提示性数据的训练集 (即小规模提示性数据集) 微调模型; w/ $D_{\text{mix}}(4\text{k})$ 表示使用本文所提出的基于可控性解释的混合数据增强框架生成的数据集微调模型, 该数据集同样仅包含 4 000 条训练样本. 表 3 中最佳结果以加粗形式突出显示.

表 3 生成式大模型捷径学习问题的缓解结果 (%)

模型	方法	自然语言推理任务					事实验证任务					释义识别任务			
		ID		OOD			ID		OOD			ID	OOD	Dec.	
		MNLI-m	HANS	MNLI-Hard-m	MNLI-Hard-mm	Avg.	Dec.	FEVER	Symm.v1	Symm.v2	Avg.	Dec.	QQP		PAWS
LLaMA (7B)	w/ $D_{\text{pro}}(\text{all})$	89.9	72.3	85.8	86.0	81.4	8.5	88.8	71.5	72.8	72.2	16.6	92.8	53.8	39.0
	w/ $D_{\text{pro}}(4\text{k})$	84.8	69.2	81.5	80.5	77.1	7.7	94.9	81.4	82.4	81.9	13.1	84.1	49.6	34.5
	w/ $D_{\text{mix}}(4\text{k})(\text{ours})$	85.8	74.8	81.3	82.2	79.4	6.3	95.4	84.7	86.1	85.4	10.0	84.5	59.3	25.2
LLaMA 2 (7B)	w/ $D_{\text{pro}}(\text{all})$	90.6	69.8	86.2	86.4	80.8	9.8	90.2	72.6	73.0	72.8	17.4	89.4	57.0	32.4
	w/ $D_{\text{pro}}(4\text{k})$	86.0	67.3	81.7	82.3	77.1	8.9	95.3	77.9	79.6	78.8	16.7	84.5	64.8	19.7
	w/ $D_{\text{mix}}(4\text{k})(\text{ours})$	86.2	70.1	82.4	83.2	78.6	7.6	95.4	82.9	83.4	83.2	12.3	84.6	65.4	19.2
FLAN- T5-xl (3B)	w/ $D_{\text{pro}}(\text{all})$	87.7	66.9	83.8	84.8	78.5	9.2	94.0	79.4	80.8	80.1	13.9	90.2	50.3	39.9
	w/ $D_{\text{pro}}(4\text{k})$	88.8	68.7	84.6	85.4	79.6	9.2	90.9	84.3	85.3	84.8	6.1	85.8	61.2	24.6
	w/ $D_{\text{mix}}(4\text{k})(\text{ours})$	89.3	71.0	85.7	86.2	81.0	8.4	95.7	90.5	89.9	90.2	5.5	86.9	66.2	20.7

实验结果表明, (1) 在训练样本数量相同的情况下, 使用 $D_{\text{mix}}(4\text{k})$ 训练后的 LLaMA 系列模型在所有分布内与分布外测试场景中的表现均优于使用 $D_{\text{pro}}(4\text{k})$ 微调后的模型, 并且在分布外测试场景的提升更为显著. 对于 FLAN-T5 模型来说同样如此, 使用 $D_{\text{mix}}(4\text{k})$ 训练后的模型在各个场景下的性能均有所提升. 这表明我们的方法有效缓解了不同架构的生成式模型的捷径学习问题, 显著提升了模型在分布外测试场景中的鲁棒性, 同时未牺牲甚至略有提升其在分布内测试场景中的性能表现. (2) 与使用大规模提示性数据集 $D_{\text{pro}}(\text{all})$ 微调后的模型相比, 使用 $D_{\text{mix}}(4\text{k})$ 训练后的模型在多个分布外测试场景中的性能同样得到了显著提升 (其中, LLaMA 模型在事实验证任务的 Avg. 提升 13.2%, 释义识别任务的 Avg. 提升 5.5%), 并且具有较低的 Dec. 值. 这表明即使在训练数据的规模相差较大的情况下, $D_{\text{mix}}(4\text{k})$ 依然有效缓解了模型的捷径学习问题. 值得注意的是, 在自然语言推理任务中, 虽然 LLaMA 系列模型在 $D_{\text{pro}}(\text{all})$ 微调下的分布内准确率与 Avg. 值均略微高于 $D_{\text{mix}}(4\text{k})$ 微调的结果, 但是由于 $D_{\text{pro}}(\text{all})$ 对模型分布内性能的增益大于其对模型分布外性能的增益, 因此在 $D_{\text{pro}}(\text{all})$ 微调下的 Dec. 值仍低于 $D_{\text{mix}}(4\text{k})$ 下的 Dec. 值, 这可能与任务的难度也有所关系. 总之, 对于生成式预训练大模型来说, 使用小规模数据进行微调, 足以在众多任务中激发模型的理解与推理能力. 对于难度较低的任务, 大规模数据下的过度训练反而可能会加重模型的捷径学习行为, 损害其鲁棒性.

此外, 无论是对于何种类型的捷径, 我们的方法均体现出了巨大的优势, 即缓解了模型对于多种捷径的依赖问题. 相比于单句捷径, 使用 $D_{\text{mix}}(4\text{k})$ 微调后的模型在含有更具挑战性的词汇重叠捷径的测试场景中的提升更为明显. 这进一步表明, 我们所提出的框架不需要关于捷径的先验知识, 不受捷径类型的限制, 能够有效应用于缓解多种类型的捷径.

因此, 我们提出的基于可控性解释的混合数据增强框架能够有效缓解不同架构的生成式预训练大模型的捷径学习问题, 显著提升模型在分布外测试场景中的鲁棒性与泛化性, 同时尽量保持甚至提升模型在分布内数据中的性能. 并且, 我们的方法不受限于任务与捷径类型, 具有更广泛的应用前景和实用价值.

3.5 探讨与分析

在我们的框架中, 训练数据的规模以及可控性解释数据与原始提示性数据的混合比例对缓解模型的捷径学习均有所影响. 因此我们以使用较为广泛的自然语言推理任务为代表, 基于 LLaMA 模型进行了大量的分析实验来

展示不同设置对模型的影响. 同时, 我们在最后分析了解释性数据的获取方案对模型缓解捷径学习问题的影响, 以便进行更为全面的分析和探讨.

(1) 我们在表 4 中展示了采样不同训练样本规模对模型缓解捷径学习问题的影响. 实验结果表明, 使用全数据集微调模型时的准确率在各个数据集中均高于使用 4 000 条样本微调模型. 然而, 随着训练样本的增加, 模型在分布内测试场景中的准确率提升小于模型在分布外测试场景中的准确率提升, 进而导致全数据集微调下的 Dec. 值较高. 训练数据的增加而对模型分布内外性能的提升不一致, 同样属于捷径学习的表现之一. 并且, 从表 4 中可以看出, 仅使用规模大小为 2 000 的数据集微调模型, 即可使模型在分布内场景表现出较为良好的性能 (84.3%), 这表明即使是小规模数据集也能够有效地激发生成式预训练大模型完成特定任务的能力. 同时, 增加训练样本的规模反而可能损害模型的泛化能力: 当训练样本规模由 10 000 提升至 50 000 时, 分布外平均准确率 Avg. 值由 78.8% 下降到 78.4%. 因此, 结合对训练开销等因素的综合考虑, 我们最终选择基于样本数量为 4 000 的小规模数据集来缓解模型的捷径学习.

(2) 关于可控性解释数据与原始提示性数据的混合比例对模型的影响, 我们在表 5 中报告了不同的混合比例设置下的实验结果. 结果表明, 本文使用的等比例混合方案在分布内外测试场景中均具有最高的准确率 (分布内准确率 85.8%, 分布外 Avg. 79.4%) 以及次低的捷径学习问题表现 (Dec. 6.3%), 这表明等比例混合方案为最优设置. 其次, 无论是使用完全的提示性数据进行微调还是使用完全的解釋性数据进行微调, 模型均表现出较为严重的捷径学习问题, 后者甚至 Avg. 值更低且 Dec. 值更高 (Avg. 77.1% vs. 74.4%; Dec. 7.7% vs. 8.2%), 这表明训练数据形式的多样性对缓解模型的捷径学习问题至关重要, 更加验证了我们所提出的混合策略的有效性. 正如第 2.1.1 节所述, 如果完全使用解释性数据进行微调, 过多的解释数据可能使模型在训练时过度关注解释而非目标标签, 从而损害模型的总体性能. 最后, 尽管在提示性与解释性数据混合比例为 1:3 的配置下得到了最小的 Dec. 值 (2.8%), 但这主要是因为在该比例下模型在分布内测试场景中的性能过低 (准确率为 74.2%, 显著低于我们设置的 85.8%), 在分布外测试场景中的性能也是最差的 (平均准确率 Avg. 也只有 71.4%, 远低于我们设置的 79.4%), 因此这并不能说明模型在此设置下的捷径学习程度较低.

表 4 训练数据规模对自然语言推理任务下 LLaMA 模型的影响结果 (%)

Data_size	ID		OOD			Dec.
	MNLI-m	HANS	MNLI-Hard-m	MNLI-Hard-mm	Avg.	
2 000	84.3	66.2	79.4	80.8	75.5	8.8
4 000	84.8	69.2	81.5	80.5	77.1	7.7
10 000	87.3	70.1	82.5	83.9	78.8	8.5
50 000	88.9	66.0	83.9	85.3	78.4	10.5
392 702	89.9	72.3	85.8	86.0	81.4	8.5

表 5 提示性数据与可控性解释数据的混合比例对自然语言推理任务下 LLaMA 模型的影响结果 (%)

$D_{\text{pro}}:D_{\text{exp}}$	ID		OOD			Dec.
	MNLI-m	HANS	MNLI-Hard-m	MNLI-Hard-mm	Avg.	
1:0	84.8	69.2	81.5	80.5	77.1	7.7
3:1	82.9	72.9	77.8	78.4	76.4	6.5
1:1	85.8	74.8	81.3	82.2	79.4	6.3
1:3	74.2	74.7	69.1	70.5	71.4	2.8
0:1	82.6	67.0	77.3	78.9	74.4	8.2

综合 (1) 和 (2) 分析实验发现, 训练样本的规模与提示性数据和可控性数据的混合的比例对模型的性能均有一定影响, 并且后者的影响更为显著. 因此, 选择合适的混合比例对缓解模型的捷径学习问题至关重要. 在合适的实验设置下, 通过本文所提框架构造的数据集微调模型, 能够显著帮助模型缓解捷径学习问题, 增强模型在分布外测试场景中的鲁棒性与泛化性, 甚至能够同时提升模型在分布内测试集中的性能.

(3) 我们进一步探究了多种获得解释性数据的方法对缓解模型捷径学习问题的影响, 如使用传统的 COT 方法获得解释性数据, 以及将传统的 COT 方法与本文提出的可控性提示相结合以获得可控的推理路径. 实验结果如表 6 所示, 我们同时在表 7 中展示了使用不同方法获得解释性数据的具体提示模版. 实验结果表明: 1) 与使用不含解释的提示数据集 D_{pro} 相比, 使用由传统 COT 方法生成的解释所构成的混合数据集 (D_{COT}) 微调模型, 能够同时提示模型在分布内外测试场景中的性能 (MNLI-m 84.8% vs. 85.8%; Avg. 77.1% vs. 77.3%). 表明由传统 COT 方法得到的解释性数据对模型的鲁棒性有所帮助. 2) 使用本文提出的基于可控性解释所构造的混合数据集 (D_{mix}) 微调后的模型, 其鲁棒性与缓解捷径学习的能力均显著优于使用 D_{COT} 微调的模型 (Avg. 79.4% vs. 77.3%; Dec. 6.3%

vs. 8.5%). 实验结果同时印证了第 2.2.1 节的描述: 由于生成 COT 解释的过程中缺乏真实标签的引导, 因此获得的 COT 解释并不完全准确, 而可控性解释则不存在这一问题, 更有利于提升模型的理解和推理能力. 3) 带有真实标签引导的 COT 方法生成解释所构造的混合数据集 (D_{C-COT1} 与 D_{C-COT2}) 同样能够提升模型的性能, 有利于缓解捷径学习. 并且, 相比于传统 COT 方法构造的 D_{COT} , D_{C-COT1} 与 D_{C-COT2} 更有利于缓解模型的捷径学习 (Dec. 8.5% vs. 7.5%), 进一步表明了真实标签的引导对于生成高质量的解释数据具有很大作用. 然而, 使用我们提出的 D_{mix} 数据集微调后的模型在分布外测试场景中的性能仍然优于使用 D_{C-COT1} 与 D_{C-COT2} 微调后的模型 (Avg. 79.4% vs. 78.5%), 并且与分布内准确率的差值更低 (Dec. 6.3% vs. 7.5%). 这表明基于可控性解释构造的数据集 D_{mix} 更有利于缓解模型的捷径学习. 可能是因为 COT 方法更加注重逐步推理, 更适用于需要明确推理步骤的任务, 如数学推理问题. 而对于此类自然语言理解任务来说, 通过 COT 逐步推理所获得的解释性数据可能较为复杂冗长. 总体而言, 由本文提出框架所获得可控性解释数据对于缓解模型捷径学习与提升模型鲁棒性的效果优于使用传统 COT 方法或者带有真实标签引导的 COT 方法所获得的解释数据, 进一步证明了我们构造的数据集 D_{mix} 的有效性.

表 6 解释性数据的获取方案对自然语言推理任务下

w/ Data	LLaMA 模型的影响结果 (%)					
	ID		OOD			Dec.
	MNLI-m	HANS	MNLI-Hard-m	MNLI-Hard-mm	Avg.	
w/ D_{pro}	84.8	69.2	81.5	80.5	77.1	7.7
w/ D_{COT}	85.8	69.6	81.0	81.2	77.3	8.5
w/ D_{mix} (ours)	85.8	74.8	81.3	82.2	79.4	6.3
w/ D_{C-COT1}	86.1	72.1	81.5	82.0	78.5	7.6
w/ D_{C-COT2}	85.8	70.9	81.3	81.3	77.8	7.5

表 7 解释性数据的获取方案的提示模板示例

Dataset	Prompt template
D_{COT}	Let's think step by step.
D_{mix}	Why the answer is <gold-label>?
D_{C-COT1}	The answer is <gold-label>. Let's think step by step.
D_{C-COT2}	Why the answer is <gold-label> ? Let's think step by step.

4 总结

本文首次探究了多种生成式预训练大语言模型在自然语言理解任务中的捷径学习现象. 研究显示: 捷径学习现象仍然存在于最近流行的生成式预训练大模型之中, 并且损害了模型在分布外测试场景中的鲁棒性. 基于这一发现, 本文进一步探索了适用于生成式预训练模型捷径学习问题的缓解方案: 基于可控性解释的混合数据增强框架. 该框架从数据的角度出发, 首先抽取部分样本赋予提示, 其次通过提示工程并利用 ChatGPT 自动生成可控性解释; 最后设计混合策略将可控性解释数据与部分提示性数据相结合, 构造混合数据集用于微调模型. 最终有效缓解了生成式大模型的捷径学习问题, 显著提升了模型在分布外测试场景中的鲁棒性与泛化性, 同时并未牺牲甚至略有提升模型在分布内场景中的表现. 本文提出的解决方案不仅适用于各种类型的任务, 并且能够用于缓解多种已知甚至未知类型的捷径, 具有广泛的适用性与较高的实践价值.

References:

- [1] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: ACL, 2019. 4171–4186. [doi: [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423)]
- [2] Liu YH, Ott M, Goyal N, Du JF, Joshi M, Chen DQ, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.
- [3] Williams A, Nangia N, Bowman S. A broad-coverage challenge corpus for sentence understanding through inference. In: Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long Papers). New Orleans: Association for Computational Linguistics, 2018. 1112–1122. [doi: [10.18653/v1/N18-1101](https://doi.org/10.18653/v1/N18-1101)]
- [4] Utama PA, Moosavi NS, Gurevych I. Mind the trade-off: Debiasing NLU models without degrading the in-distribution performance. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 8717–8729. [doi: [10.18653/v1/2020.acl-main.770](https://doi.org/10.18653/v1/2020.acl-main.770)]
- [5] Utama PA, Moosavi NS, Gurevych I. Towards Debiasing NLU models from unknown biases. In: Proc. of the 2020 Conf. on Empirical

- Methods in Natural Language Processing (EMNLP). Association for Computational Linguistics, 2020. 7597–7610. [doi: [10.18653/v1/2020.emnlp-main.613](https://doi.org/10.18653/v1/2020.emnlp-main.613)]
- [6] Tu LF, Lalwani G, Gella S, He H. An empirical study on robustness to spurious correlations using pre-trained language models. *Trans. of the Association for Computational Linguistics*, 2020, 8: 621–633. [doi: [10.1162/tacl_a_00335](https://doi.org/10.1162/tacl_a_00335)]
- [7] McCoy RT, Pavlick E, Linzen T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In: *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, 2020. 3428–3448. [doi: [10.18653/v1/P19-1334](https://doi.org/10.18653/v1/P19-1334)]
- [8] Gururangan S, Swamydipta S, Levy O, Schwartz R, Bowman S, Smith NA. Annotation artifacts in natural language inference data. In: *Proc. of the 2018 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 2 (Short Papers)*. New Orleans: Association for Computational Linguistics, 2018. 107–112. [doi: [10.18653/v1/N18-2017](https://doi.org/10.18653/v1/N18-2017)]
- [9] Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann FA. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020, 2(11): 665–673. [doi: [10.1038/s42256-020-00257-z](https://doi.org/10.1038/s42256-020-00257-z)]
- [10] Schwartz R, Stanovsky G. On the limitations of dataset balancing: The lost battle against spurious correlations. In: *Proc. of the 2022 Findings of the Association for Computational Linguistics*. Seattle: Association for Computational Linguistics, 2022. 2182–2194. [doi: [10.18653/v1/2022.findings-naacl.168](https://doi.org/10.18653/v1/2022.findings-naacl.168)]
- [11] Du MN, He FX, Zou N, Tao DC, Hu X. Shortcut learning of large language models in natural language understanding. *Communications of the ACM*, 2023, 67(1): 110–120. [doi: [10.1145/3596490](https://doi.org/10.1145/3596490)]
- [12] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. LLaMA: Open and efficient foundation language models. *arXiv:2302.13971*, 2023.
- [13] Touvron H, Martin L, Stone K, *et al.* LLaMA 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*, 2023.
- [14] Chung HW, Hou L, Longpre S, *et al.* Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 2024, 25(70): 1–53.
- [15] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, 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, Lowe R. Training language models to follow instructions with human feedback. In: *Proc. of the 36th Int'l Conf. on Neural Information Processing Systems*. New Orleans: Curran Associates Inc., 2022. 27730–27744.
- [16] Niven T, Kao HY. Probing neural network comprehension of natural language arguments. In: *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, 2019. 4658–4664. [doi: [10.18653/v1/P19-1459](https://doi.org/10.18653/v1/P19-1459)]
- [17] Lai YX, Zhang C, Feng YS, Huang QZ, Zhao DY. Why machine reading comprehension models learn shortcuts? In: *Proc. of the 2021 Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 2021. 989–1002. [doi: [10.18653/v1/2021.findings-acl.85](https://doi.org/10.18653/v1/2021.findings-acl.85)]
- [18] Liu F, Avci B. Incorporating priors with feature attribution on text classification. In: *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence: Association for Computational Linguistics, 2019. 6274–6283. [doi: [10.18653/v1/P19-1631](https://doi.org/10.18653/v1/P19-1631)]
- [19] Han XC, Tsvetkov Y. Influence tuning: Demoting spurious correlations via instance attribution and instance-driven updates. In: *Proc. of the 2021 Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana: Association for Computational Linguistics, 2021. 4398–4409. [doi: [10.18653/v1/2021.findings-emnlp.374](https://doi.org/10.18653/v1/2021.findings-emnlp.374)]
- [20] Clark C, Yatskar M, Zettlemoyer L. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In: *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong: Association for Computational Linguistics, 2019. 4069–4082. [doi: [10.18653/v1/D19-1418](https://doi.org/10.18653/v1/D19-1418)]
- [21] He H, Zha S, Wang HH. Unlearn dataset bias in natural language inference by fitting the residual. In: *Proc. of the 2nd Workshop on Deep Learning Approaches for Low-Resource Natural Language Processing*. Hong Kong: Association for Computational Linguistics, 2019. 132–142. [doi: [10.18653/v1/D19-6115](https://doi.org/10.18653/v1/D19-6115)]
- [22] Sanh V, Wolf T, Belinkov Y, Rush AM. Learning from others' mistakes: Avoiding dataset biases without modeling them. *arXiv:2012.01300*, 2020.
- [23] Zhang DC, Zhang K, Wu L, Wang M. Causal-based debiased reasoning method for grounded textual entailment. *Journal of Computer Research and Development*, 2023, 60(8): 1768–1779 (in Chinese with English abstract). [doi: [10.7544/issn1000-1239.202330248](https://doi.org/10.7544/issn1000-1239.202330248)]
- [24] Nam J, Cha H, Ahn S, Lee J, Shin J. Learning from failure: Training debiased classifier from biased classifier. In: *Proc. of the 34th Int'l Conf. on Neural Information Processing Systems*. Vancouver: Curran Associates Inc., 2020. 20673–20684.

- [25] Liu EZ, Haghgoo B, Chen AS, Raghunathan A, Koh PW, Sagawa S, Liang P, Finn C. Just train twice: Improving group robustness without training group information. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 6781–6792.
- [26] Wu YX, Gardner M, Stenetorp P, Dasigi P. Generating data to mitigate spurious correlations in natural language inference datasets. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Dublin: Association for Computational Linguistics, 2022. 2660–2676. [doi: [10.18653/v1/2022.acl-long.190](https://doi.org/10.18653/v1/2022.acl-long.190)]
- [27] Gardner M, Merrill W, Dodge J, Peters M, Ross A, Singh S, Smith NA. Competency problems: On finding and removing artifacts in language data. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: Association for Computational Linguistics, 2021. 1801–1813. [doi: [10.18653/v1/2021.emnlp-main.135](https://doi.org/10.18653/v1/2021.emnlp-main.135)]
- [28] Kaushik D, Hovy E, Lipton ZC. Learning the difference that makes a difference with counterfactually-augmented data. arXiv:1909.12434, 2020.
- [29] Si CL, Zhang ZY, Qi FC, Liu ZY, Wang YS, Liu Q, Sun MS. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning. In: Proc. of the 2021 Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021. 1569–1576. [doi: [10.18653/v1/2021.findings-acl.137](https://doi.org/10.18653/v1/2021.findings-acl.137)]
- [30] Nie YX, Williams A, Dinan E, Bansal M, Weston J, Kiela D. Adversarial NLI: A new benchmark for natural language understanding. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 4885–4901. [doi: [10.18653/v1/2020.acl-main.441](https://doi.org/10.18653/v1/2020.acl-main.441)]
- [31] Lu KJ, Mardziel P, Wu FJ, Amancharla P, Datta A. Gender bias in neural natural language processing. In: Nigam V, Kirigin TB, Talcott C, Guttman J, Kuznetsov S, Loo BT, Okada M, eds. Logic, Language, and Security: Essays Dedicated to Andre Scedrov on the Occasion of His 65th Birthday. Cham: Springer, 2020. 189–202. [doi: [10.1007/978-3-030-62077-6_14](https://doi.org/10.1007/978-3-030-62077-6_14)]
- [32] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1(8): 9.
- [33] Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, Zhou YQ, Li W, Liu PJ. Exploring the limits of transfer learning with a unified text-to-text Transformer. Journal of Machine Learning Research, 2020, 21(1): 5485–5551.
- [34] Wei J, Wang XZ, Schuurmans D, Bosma M, Ichter B, Xia F, Chi EH, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 24824–24837.
- [35] Lampinen A, Dasgupta I, Chan S, Mathewson K, Tessler M, Creswell A, McClelland J, Wang J, Hill F. Can language models learn from explanations in context? In: Proc. of the 2022 Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi: Association for Computational Linguistics, 2022. 537–563. [doi: [10.18653/v1/2022.findings-emnlp.38](https://doi.org/10.18653/v1/2022.findings-emnlp.38)]
- [36] Stacey J, Belinkov Y, Rei M. Supervising model attention with human explanations for robust natural language inference. In: Proc. of the 36th AAAI Conf. on Artificial Intelligence. AAAI, 2022. 11349–11357. [doi: [10.1609/aaai.v36i10.21386](https://doi.org/10.1609/aaai.v36i10.21386)]
- [37] Chen H, He J, Narasimhan K, Chen DQ. Can rationalization improve robustness? In: Proc. of the 2022 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: Association for Computational Linguistics, 2022. 3792–3805. [doi: [10.18653/v1/2022.naacl-main.278](https://doi.org/10.18653/v1/2022.naacl-main.278)]
- [38] Schuster T, Fisch A, Barzilay R. Get your vitamin C! Robust fact verification with contrastive evidence. In: Proc. of the 2021 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online: Association for Computational Linguistics, 2021. 624–643. [doi: [10.18653/v1/2021.naacl-main.52](https://doi.org/10.18653/v1/2021.naacl-main.52)]
- [39] Min S, Lyu XX, Holtzman A, Artetxe M, Lewis M, Hajishirzi H, Zettlemoyer L. Rethinking the role of demonstrations: What makes in-context learning work? In: Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing. Abu Dhabi: Association for Computational Linguistics, 2022. 11048–11064. [doi: [10.18653/v1/2022.emnlp-main.759](https://doi.org/10.18653/v1/2022.emnlp-main.759)]
- [40] Bach S, Sanh V, Yong ZX, Webson A, Raffel C, Nayak NV, Sharma A, Kim T, Bari MS, Fevry T, Alyafeai Z, Dey M, Santilli A, Sun ZQ, Ben-David S, Xu CW, Chhablani G, Wang H, Fries J, Al-Shaibani M, Sharma S, Thakker U, Almubarak K, Tang XR, Radev D, Jiang MTJ, Rush A. PromptSource: An integrated development environment and repository for natural language prompts. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Dublin: Association for Computational Linguistics, 2022. 93–104. [doi: [10.18653/v1/2022.acl-demo.9](https://doi.org/10.18653/v1/2022.acl-demo.9)]
- [41] Sclar M, Choi Y, Tsvetkov Y, Suhr A. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. arXiv:2310.11324, 2023.
- [42] Ludan JM, Meng YX, Nguyen T, Shah S, Lyu Q, Apidianaki M, Callison-Burch C. Explanation-based finetuning makes models more robust to spurious cues. In: Proc. of the 61st Annual Meeting of the Association for Computational Linguistics (Vol. 1: Long Papers). Toronto: Association for Computational Linguistics, 2023. 4420–4441. [doi: [10.18653/v1/2023.acl-long.242](https://doi.org/10.18653/v1/2023.acl-long.242)]

- [43] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, Agarwal S, Herbert-Voss A, Krueger G, Henighan T, Child R, Ramesh A, Ziegler DM, 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, Amodei D. Language models are few-shot learners. In: Proc. of the 34th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2020. 1877–1901.
- [44] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: Association for Computational Linguistics, 2021. 3045–3059. [doi: [10.18653/v1/2021.emnlp-main.243](https://doi.org/10.18653/v1/2021.emnlp-main.243)]
- [45] Dou SH, Zheng R, Wu T, Gao SY, Shan JJ, Zhang Q, Wu YM, Huang XJ. Decorrelate irrelevant, purify relevant: Overcome textual spurious correlations from a feature perspective. In: Proc. of the 29th Int'l Conf. on Computational Linguistics. Gyeongju: Int'l Committee on Computational Linguistics, 2022. 2278–2287. [<https://aclanthology.org/2022.coling-1.199>]
- [46] Thorne J, Vlachos A, Cocarascu O, Christodoulopoulos C, Mittal A. The fact extraction and verification (FEVER) shared task. In: Proc. of the 1st Workshop on Fact Extraction and Verification (FEVER). Brussels: Association for Computational Linguistics, 2018. 1–9. [doi: [10.18653/v1/W18-5501](https://doi.org/10.18653/v1/W18-5501)]
- [47] Schuster T, Shah D, Yeo YJS, Ortiz DRF, Santus E, Barzilay R. Towards debiasing fact verification models. In: Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int'l Joint Conf. on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2019. 3419–3425. [doi: [10.18653/v1/D19-1341](https://doi.org/10.18653/v1/D19-1341)]
- [48] Zhang Y, Baldridge J, He LH. PAWS: Paraphrase adversaries from word scrambling. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 1298–1308. [doi: [10.18653/v1/N19-1131](https://doi.org/10.18653/v1/N19-1131)]
- [49] Ott M, Edunov S, Baeveski A, Fan A, Gross S, Ng N, Grangier D, Auli M. Fairseq: A fast, extensible toolkit for sequence modeling. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics (Demonstrations). Minneapolis: Association for Computational Linguistics, 2019. 48–53. [doi: [10.18653/v1/N19-4009](https://doi.org/10.18653/v1/N19-4009)]
- [50] Hu EJ, Shen YL, Wallis P, Allen-Zhu Z, Li YZ, Wang SA, Wang L, Chen WZ. LoRA: Low-rank adaptation of large language models. arXiv:2106.09685, 2021.

附中文参考文献:

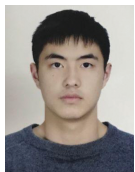
- [23] 张大操, 张琨, 吴乐, 汪萌. 针对情境感知的自然语言的因果去偏推理方法. 计算机研究与发展, 2023, 60(8): 1768–1779. [doi: [10.7544/issn1000-1239.202330248](https://doi.org/10.7544/issn1000-1239.202330248)]



孙泽辰(2000—), 女, 硕士生, 主要研究领域为自然语言处理.



张民(1970—), 男, 博士, 教授, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理, 机器翻译, 人工智能.



肖义胜(1999—), 男, 博士生, 主要研究领域为自然语言处理.



周国栋(1967—), 男, 博士, 教授, 博士生导师, CCF 杰出会员, 主要研究领域为自然语言处理.



李俊涛(1993—), 男, 博士, 副教授, CCF 专业会员, 主要研究领域为自然语言处理.