

分类号 _____

编号 _____

U D C _____

密级 _____



南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

本科生毕业设计（论文）

题 目： 基于大模型思维链的
校园欺凌专家知识检索技术研究

姓 名： 李博翱

学 号： 12011407

系 别： 计算机科学与工程系

专 业： 计算机科学与技术

指导教师： 宋轩 副教授

2024 年 6 月 7 日

诚信承诺书

1. 本人郑重承诺所呈交的毕业设计（论文），是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。
2. 除文中已经注明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。
3. 本人承诺在毕业论文（设计）选题和研究内容过程中没有抄袭他人研究成果和伪造相关数据等行为。
4. 在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

作者签名: 李博翔

2024 年 6 月 7 日

基于大模型思维链的 校园欺凌专家知识检索技术研究

李博翱

(计算机科学与工程系 指导教师: 宋轩)

[摘要]: 随着社会的进步和教育的普及, 校园欺凌已经成为一个严重的社会问题, 威胁着学生的身心健康和教育环境的和谐。尽管国内外已开发出多种防控校园欺凌的技术, 搭建了多个相关教育资源, 但由于网络限制和语言障碍, 这些资源对于国内的教师和学生来说可能难以获取。而想要从根本上解决问题, 更多地需要一些心理辅导和人文关怀。针对校园欺凌问题, 本论文提出了一种结合大模型思维链技术和因果推断的新型提示方法, 旨在提升大模型回答校园欺凌相关问题的能力。该方法通过使用思维链技术, 提升大模型的逻辑推理能力, 在此基础上结合了因果推断技术, 使得大模型能够在回答问题时结合因果推断的结果进行思考, 从而实现对校园欺凌问题的深度分析并提出有效的解决方案。本论文首先介绍了校园霸凌问题的背景和现状, 然后详细阐述了思维链相关的技术以及因果推断技术, 接着提出了使用思维链技术提升大模型在校园欺凌相关垂直领域能力的框架和方法。通过实验验证, 展示了本方法相较于传统提示方法的优势, 并对未来在校园欺凌问题解决中大模型的应用前景提出了建议。

[关键词]: 校园欺凌防控, 大语言模型, 思维链, 因果推断

[ABSTRACT]: As society progresses and education becomes more widespread, school bullying has emerged as a serious social issue, threatening students' mental and physical health and the harmony of educational environments. Despite the development of various anti-bullying technologies and the establishment of numerous educational resources internationally, access to these resources can be challenging for domestic teachers and students due to internet restrictions and language barriers. To fundamentally address these issues, there is a greater need for psychological counseling and humanistic care. Addressing school bullying, this thesis introduces a novel prompting method that combines large model Chain of Thought (CoT) techniques and causal inference, aimed at enhancing the capability of large models to respond to school bullying-related queries. This method leverages CoT to enhance the logical reasoning capabilities of large models, and integrates causal inference to allow the model to consider the results of causal analysis while responding, thus facilitating deep analysis of school bullying issues and proposing effective solutions. The thesis begins by outlining the background and current state of school bullying, followed by a detailed discussion of Chain of Thought and causal inference techniques. It then proposes a framework and method for enhancing large models' capabilities in the domain of school bullying using CoT. Experimental validations demonstrate the advantages of this method over traditional prompting methods, and suggestions are made for the future application of large models in addressing school bullying.

[Key words]: Campus Anti-bullying, Large Language Model, Chain of Thought, Casual Inference

目录

1. 引言	1
1.1 背景	1
1.2 国内外校园欺凌技术调研	2
1.2.1 视频识别技术	2
1.2.2 对话机器人教学系统	3
1.2.3 语音识别报警系统	3
1.3 国内外校园欺凌网站调研	3
1.4 研究意义	4
1.5 挑战	5
1.6 章节安排	5
2. 相关工作	6
2.1 基础提示工程	6
2.1.1 零样本提示技术	6
2.1.2 少样本提示技术	6
2.2 思维链技术	7
2.2.1 思维链简介	7
2.2.2 自洽性	8
2.2.3 Least to Most	8
2.3 因果推断	9
3. 整体方案	10

3.1 方法设计	10
3.2 评估指标	11
3.3 实验设计	15
3.3.1 实验任务及数据	15
3.3.2 大模型及采样选择	15
4. 实验结果	17
4.1 基础实验分析	17
4.2 分类实验分析	20
5. 总结与展望	26
5.1 总结	26
5.2 展望	26
参考文献	27
致谢	29

1. 引言

1.1 背景

在全球以及中国范围内，校园欺凌一直是一个难以解决且严重影响学生心理健康与学习环境的问题。随着信息技术的迅速发展，尤其是互联网和移动设备的普及，校园欺凌的事件越来越多地被曝光于公众视野。尽管这种曝光有助于提高社会对校园欺凌问题的关注，但公开报道的案例只是冰山一角。在中国，根据中国青少年研究中心 2015 年对多个省市的 5864 名学生的调查，表示自己在校时会“偶尔被欺凌”的学生占 32.5%，在校“经常被欺凌”的学生占 6.1%^[1]。同样的，联合国教科文组织发布的《校园暴力和欺凌全球数据报告》指出，全球范围内有近 2.46 亿儿童及青少年每年遭受校园欺凌^[2]。最高人民法院统计数据更显现了校园暴力案件的严峻性：2015 至 2017 年，全国法院审结的校园暴力案件中，有 57.5% 涉及故意伤害，其中 11.59% 的案件导致了受害人死亡^[3]。校园欺凌目前的危害从这些数据中可见一斑。

尽管校园欺凌普遍存在，但由于其隐蔽性强和发生在不易监控的场所，如宿舍和厕所等，使得这一行为难以被及时发现和处理。受害者往往因为害怕报复或羞于开口，而不愿意向教师或家长揭露自己的遭遇^[4]。此外，校园欺凌受害者自身的胆怯也会增长其隐蔽性。受害人在校园欺凌中处于弱势一方，常常会出现面对欺凌不敢反抗，忍气吞声的情况。而依照《预防未成年人犯罪法》、《治安管理处罚法》的相关规定，校园欺凌行为属于学生的不良行为。此种不良行为由于界定不明确，校园欺凌行为就成了漏网之鱼得不到规制^[5]。这样的种种原因使得校园欺凌难以依靠传统的检测等方法进行防控和解决。

为了应对校园欺凌带来的挑战，近年来随着大模型的发展，一个新的思路已然出现，即开发一个基于大模型的校园欺凌专家系统。该专家系统能以对话机器人的形式与教师和学生进行互动，回答有关校园欺凌的咨询，并进一步提供与心理辅导相关的内容。目前已有研究指出^[6]，校园欺凌的发展分为四个阶段：“未萌欺凌”、“萌芽欺凌”、“标准欺凌”、“严重欺凌”，而对于中小学生来说，前两个阶段是与他人接触中的主要形式，及时的防控和教育可以帮助加害人了解玩笑打闹的分寸，从而将欺凌行为扼杀在萌芽中。这一点也正是大模型在其心理辅导和人文关怀方面所擅长的。然

而，校园欺凌及其相关的垂直领域是一个非常专业而具体化的领域，需要具备教育，心理辅导等等相关知识才能对其进行深入了解。要回答这些专业性很强的问题，并给出有效而具有人文关怀的回答，需要具备深厚的专业知识，这对现有的大模型是一个挑战。

1.2 国内外校园欺凌技术调研

近年来，国内外同样有很多校园欺凌相关的技术正在发展。在进行搜集资料和调研后，现有的国内外校园欺凌相关的技术大概可以分为如下三种：

- (1) 基于深度学习和神经网络的视频识别技术
- (2) 基于角色模拟的对话机器人教学系统
- (3) 基于语音识别的报警系统

1.2.1 视频识别技术

该技术主要利用了人工智能和远程传感技术来检测校园暴力事件。文章作者通过分析监控视频序列中的图像特征和声音特征，使用 C3D 神经网络进行特征提取和分类，提高了识别校园暴力事件的准确性^[7]。为了解决证据冲突的问题，文章还提出了一种改进的 Dempster-Shafer (D-S) 融合算法，这种算法比传统的 D-S 理论有更高的识别准确率。如图1所示，该算法能对视频或监控录像中的人物暴力行为进行识别。通过这些技术手段，该研究能够有效地提高校园暴力检测的准确率和实时性，为防止和应对校园暴力事件提供了一种新的技术途径。



图 1 视频识别暴力行为示意图^[7]

1.2.2 对话机器人教学系统

该技术主要研究了通过使用虚拟对话代理（对话机器人）进行反欺凌教育活动对 K-12 学生对欺凌问题态度的影响^[8]。研究使用了实验前后测试设计，并设置了三个群体，分别与扮演欺凌者、受害者和教师角色的虚拟代理进行对话。研究结果显示，与对话机器人进行场景对话后，学生对欺凌问题的态度发生了积极的变化，特别是扮演欺凌者和受害者的小组在反欺凌态度上有显著提升。该技术体现了对话机器人技术在校园防欺凌教育中的作用，并指出这种方法可能成为校园防欺凌教育的一种补充方法，帮助学生通过角色扮演和情感体验以改变对欺凌的态度。

1.2.3 语音识别报警系统

基于 AI 语音识别的校园欺凌一键报警式系统是国内目前比较主流的校园欺凌防控系统。其系统由 AI 语音一键识别报警终端，APP，云平台，电脑处理软件等构成。防欺凌 AI 语音识别一键报警终端部署在学校卫生间、宿舍、教室、楼道、操场等区域，实时 AI 分析设备区域内的音频信息，若识别出欺凌常发的“救命、打架、着火”等关键词时，则现场触发预警语音自动报警干预，同时报警消息可通过手机 APP、云平台、电脑端等实时通知相关管理人员，相关人员收到预警后，可通过收听实时视频监控、语音监听对讲等方式来判断现场情况快速出警。该系统同时支持一键报警、语音广播、远程开闸、报警录像、电子地图定位、联动探测器防盗报警、联防报警等多功能扩展，对于校园欺凌行为的及时发现与制止有很大的帮助。

1.3 国内外校园欺凌网站调研

目前，国内外也有很多有关校园欺凌的网站或者论坛等。在这些网站上可以找到有关校园欺凌的简介，应对措施，以及一些相关故事等。这些网站能为孩子，家长和教师提供大量帮助，并且网站上的相关材料和教育视频等资料体现了很好的人文关怀。另外，这些网站能提供很多数据和问题，是本论文比较重要的数据来源。

- stopbullying.gov: 这是美国政府支持的一个网站，专门提供关于欺凌预防的信息。网站种的内容十分丰富，包括欺凌以及网络欺凌的定义、预防策略、法律政策，以及教育工具。同时在该网站上也配备了全套的关怀体系，比如使用者

可以查询到各种求助电话或在网站上求助等。这个平台也分享了许多关于校园欺凌的研究和案例分析。同时网站上大部分内容都有中文翻译，非常便于阅读。

- **PACER 全国欺凌预防中心:** 该网站由 PACER 组织搭建。PACER 组织提供了广泛的资源，用于教育和动员社区成员参与防止欺凌的活动。该网站主要提供关于如何识别、应对和预防校园欺凌的实用建议和策略。同时可以在网站上找到大量有关校园欺凌的教育读物，如学生手册，家长手册等。同时还有校园欺凌相关的视频供学生观看。
- **Cyberbullying Research Center(网络欺凌研究中心):** 该网站专注于网络欺凌的研究，提供了大量的数据、资源和策略，帮助学生、家长、教育者和政策制定者理解和应对网络欺凌问题。网站上的功能非常齐全，浏览者可以快速了解网络欺凌的概念，预防策略，以及应对网络欺凌的措施等。同时该网站鼓励遭受网络欺凌的人大胆分享自己的故事，以让他们知道自己并不孤单。同时该网站还有大量的博客与书籍的链接，供浏览者阅览。
- **- 中国反欺凌网:** 国内同样有反欺凌的网站，主要提供有关欺凌的信息，包括定义、类型、案例研究以及预防和干预措施。但是该网站上的相关资料以及案例介绍等相比于国外网站少了很多，只有寥寥几个案情通报。

1.4 研究意义

从以上国内外校园欺凌技术的调研结果来看，尽管当前已经有了很多相关的技术手段和教育网站，然而当前国内外的校园欺凌问题尚未得到有效解决，尤其是对于具体化的欺凌场景来说，很多校园欺凌的受害者难以获得深入且具体的解答和指导。此外，尽管国外有丰富的校园欺凌资源，语言和网络的限制使得这些资源对国内学生来说难以接触和利用。在这种背景下，利用大模型的高级文本生成和问答能力构建一个校园欺凌专家系统显得尤为重要。大模型可以通过自己的思维推断能力提供针对具体情况的专业建议和解决方案，帮助学生、家长以及教师更有效地理解和应对校园欺凌，并且提供一定的心理辅导作用，从而在心理层面为学生提供保护。通过结合思维链技术，该系统还能在逻辑推理方面得到加强，从而更精准地分析校园欺凌的特征

和应对措施，提升其在处理校园欺凌相关问题时的可靠程度。通过此技术，该专家系统主要面向教师和家长提供有关校园欺凌的建议，以更好地帮助他们和他们的孩子免受校园欺凌的困扰，增进对校园欺凌问题的理解与防范。

1.5 挑战

从上述调研中能发现国内外已有了一些校园欺凌相关防控技术，其中一部分是基于校园欺凌的检测和报警，这些技术在校园欺凌的防控方面会有一定价值，但是却很难彻底根除校园欺凌，也并不能在校园欺凌发生后对学生进行相关的心理安慰和辅导。另一部分则是使用对话机器人进行教育，但这一研究的对话都是预设定好的，很难满足当前校园欺凌的复杂场景的需求。其次，大模型如何判断当前问题对应的具体欺凌场景也是一个问题。最后，这些技术都没有和大模型所结合过。同样的，大模型在回答这类问题上的研究，并没有太多可以借鉴的论文。本论文在研究方向上比较新颖，使用大模型在回答诸如校园欺凌场景问题等偏开放性问题时，让大模型生成更好的回答需要更好的提示词设计和技术手段，这是一个比较大的挑战。

1.6 章节安排

本论文主要分为五个章节。

第一章为引言部分。介绍了校园欺凌在近年来的严重性以及其对学生的危害，并调研和讨论了目前国内外比较主流的三种校园防欺凌相关的技术。引出了本论文的研究内容以及对应的挑战。

第二章为相关工作介绍。主要介绍思维链和因果推断及其衍生技术。

第三章为整体工作框架。主要介绍思维链相关技术在回答问题方面的相关实验，以及一个整体的工作流程，并介绍相关的评测指标。

第四章为实验结果分析。通过分析实验结果，来验证思维链等提示工程技术在校园防欺凌方面的作用。

第五章为研究总结与未来展望。对该研究的整体结果进行总结，思考研究存在的缺陷和不足以及未来可能的改进和可研究的方向。

2. 相关工作

本论文旨在研究大模型思维链技术对回答校园欺凌相关问题的影响。因此本文的大部分相关方法都来源于提示工程 Prompt Engineering，如 few-shot, 思维链技术等。另有一部分相关方法源于因果推断技术。本论文将在这部分对其进行相关介绍。

2.1 基础提示工程

提示工程 (Prompt Engineering) 主要通过精心，严谨地设计特定任务的指令来引导模型输出，而不改变模型参数。这种方法与不同于传统的训练方法，在传统方法中，为了解决特定需求的人物，往往需要对模型重新训练或进行大量微调。相比之下，提示工程使得模型能够快速适应多样化的任务和领域，在大语言模型 (LLM) 和视觉-语言模型 (VLM) 方面引起了变革性的影响^[9]。

在众多提示工程的方法中，有两种是最基础的，同时他们也是其他的提示工程的起源。这两种方法不需要进行额外的训练，只需要给大模型提出问题，其即可进行回答。这两种方法分别是零样本提示技术 (Zero-shot Prompting) 和少样本提示技术 (Few-shot Prompting)。

2.1.1 零样本提示技术

目前认为零样本提示技术最早于 2019 年，由 A.Radford 等人提出^[10]。这是一项具有开创性的技术，因为它使得大模型不再需要大量的数据，只需要设计好的提示词即可回答问题，消除了其在回答指定问题是对大量训练数据的刚性需求。这对于大语言模型是第一个变革性的转变。具体来说，模型在提示中接收任务描述，但在此之前，并不会给模型提供一个标签数据 (labeled data)。然后，模型利用其已有的知识，即在发布之前进行训练过的数据，基于给定的新提示生成预测。目前来说，Zero-shot Prompting 已经十分简化，即直接向大语言模型 (如 ChatGPT) 提供一个问题描述，之后等待其回答即可。

2.1.2 少样本提示技术

少样本提示技术由 Brown 等人于 2020 年提出^[11]。该技术在让大模型回答具体问题之前，回先向大模型提供少量的输入-输出示例，使得大模型根据提示来回答问题，

并引导大模型给出与例子更加相似的答案。这与零样本提示不同，因为零样本提示不提供任何示例。在 Brown 的论文中和大量的用户反馈中已经证明，即使只用很少的例子 (比如不超过 5 个) 就已经足够让大模型根据提示的例子进行回答，而且回答的质量普遍高于零样本提示的结果。但是，对于少样本提示技术来说，提示示例的选择以及结构会对模型行为产生非常显著的影响。因而，对于提示词的构建以及例子的选择共同影响着这项提示技术表现的好坏，这也是其变得并不太可控。目前来说，Few-shot Prompting 的用法是在向大语言模型提问之前，先给他一个例子 (One-shot)，这样大模型的回答内容和结构就会更加贴近用户所给定的例子。

2.2 思维链技术

2.2.1 思维链简介

即使有了前面的提示技术，大模型仍然在需要复杂推理的任务上 (比如数学题) 表现相对不佳。思维链技术应运而生。思维链 (Chain of Thought) 是一种能够引导大语言模型在执行复杂任务时进行逻辑推理的技术。该技术由 Jason Wei 等人在 2022 年提出，它依赖于一系列叙述性的中间步骤，这些步骤模拟了人类解决问题时的思考过程。^[12]在大型语言模型的上下文中，思维链帮助模型在给出最终答案之前，先行构建和表述一系列逻辑上的推理步骤。这种方法不仅增强了模型的解释能力，即能够清晰地展示其达到某一结论的推理过程，而且还提高了处理复杂问题的准确性。通过鼓励模型展现其思考路径，思维链技术为理解和提升人工智能在多步骤推理任务中的表现提供了新的可能。该技术在教育、医疗和金融等领域的实际应用中展示了其潜力，在需要模型解释其决策过程的场合表现尤其突出。思维链提示技术相比于标准提示技术 (即 Zero-shot)，在给定例子的时候会给大模型提供思考逻辑，即思维链部分。在此之后，大模型在回答后续问题的时候明显表现要相对好于标准提示技术的回答，不仅给出了正确的回答，同时还给出了自己的推理过程。这项技术对于大模型的在较复杂推理任务上的提升是非常显著的。

由此可以发现，思维链 CoT 技术其实是基于少样本提示技术 Few-shot 而做出的改进，因此也被称为 Few-shot-CoT。在之后的研究中，由 Kojima 等人于 2022 年发现一个简单的提示词就可以增强模型的推理能力^[13]。这项技术被称为 Zero-shot-CoT，

只需要一句简单的“Let think step by step.”即可增强大模型的推理能力。用中文即为“让我们一步步思考。”这项技术也同样被 AutoCoT 所使用^[14]，证明其在推理任务上相比于 ChatGPT3 有一定提升。

在之后，Zhou 等人又进一步提出了自动提示工程 (APE) 框架，同时在其中发现了一个效果更好的 Zero-shot-CoT 提示词^[15]：“Let’s work this out in a step by step way to be sure we have the right answer.”这一提示词的效果相比于其他的提示词得到了进一步提升，翻译成中文为“让我们以一步一步的方式来解决它，以确保我们得到正确的答案。”本论文中，也将使用该提示词作为零样本思维链的额外提示词。

2.2.2 自洽性

自洽性 (Self-Consistency, SC) 由 Wang 等人于 2022 年提出^[16]。该技术是对思维链的一种提升和补全。一般来说，一个比较复杂的推理任务往往推理多样性更高，也即可能会有多种尝试去解决该问题的思考过程，但这些思考过程并不全都能解决这个问题。而普通的 CoT 采用比较贪婪的解码策略，这使得有些时候思维链会沿着错误的思维路径推理下去，从而得出一个不正确的答案。自洽性原理通过从语言模型的解码器中采样生成不同的思维链，之后通过比较多个思维链的推理逻辑，最终选择最一致的答案作为输出结果。自洽性技术与思维链技术的结合，Cot-SC 技术在各种基准测试上显著提高了准确性。

2.2.3 Least to Most

在思维链技术的基础上，Zhou 等人提出了一种名为 Least-to-Most（由简到繁）的 prompt 技术^[17]，进一步提升大语言模型在解决问题时的效率和准确性。该技术通过将问题分解为更易管理的子问题，并按照由简到繁的顺序逐步解答，引导模型沿着更清晰的思路前进。通过这种分步骤的方法，模型能够先解决基础部分，再逐渐组合和处理复杂的问题，从而在保持高准确率的同时，提升了问题解决的系统性。这一方法比较适用于需要细致推理和逐步构建答案的复杂任务。但同样的，这种方法受子问题构建的影响更大。能否构造符合逻辑推理过程的子问题来提示大模型对大模型最后的输出和整体思维过程有着举足轻重的影响。

2.3 因果推断

因果推断技术最早用于统计学研究，而在近些年，由 **Pearl** 教授先提出了利用图形模型进行因果关系分析的方法，并定义了因果推断中三种基础的因果图结果。同时，也是由他来提出和完善了潜在结果框架与因果图理论，并于 2009 年做出了总结^[18]。在最近几年，随着机器学习的发展，也出现了很多开源的 **Python** 库用于因果推断。目前的因果推断技术主要用于衡量一个行为的价值。即判断一个原因能导致多少结果。而因果推断往往会衍生出因果发现，即判断一个行为是否和一个事件互为因果（推断是判断带来了多少果，发现是判断是否为因果）。另外，因果推断还会关注当一件事情未发生时如何，称为反事实推断。

随着大模型在近期的普及，有越来越多的研究人员尝试将大模型应用到因果推断上。比如 **Abdali** 等人提出可以结合少样本思维链技术^[19]，让大模型直接帮助进行因果推断，从而从用户反馈的问题中寻找到可能导致问题的原因并绘制因果图。而 **Paul** 等人则提出使用自然语言的因果推断^[20]，来判断思维链技术产生的逻辑推理是否符合预期并加以调整。在本论文中，将会尝试结合前一种方法，尝试使用其简化版本后的技术路线，并最终结合思维链提示技术，以探究其对于大模型回答校园欺凌相关问题能力的影响。

3. 整体方案

3.1 方法设计

本论文拟使用思维链技术提升大模型在校园欺凌相关垂直领域回答问题的能力和可靠性。在此之前，我所在的综合设计小组构建了相关的校园防欺凌知识库，将各地与校园欺凌和未成年人保护相关的法律法规，相关论文以及新闻，报纸摘要等数据进行搜集处理，转换为 pdf 文件后进行向量化，以构建全面的校园防欺凌专家知识库。前期工作中已完成对于大模型的部署，并成功与知识库相结合，构建了一个基于检索增强生成 (RAG) 的对话机器人，已经可以回答校园欺凌相关的问题。

检索增强生成是一种消除大模型幻觉 (Hallucination) 的良好方法，但需要依赖知识库进行检索。而考虑到思维链技术只需要设计提示词，相比于检索增强生成更加轻量级，在该方向上有一定的探索空间。在之后，我的研究分支转向于研究思维链技术在回答校园欺凌相关问题的增益，并进行相关实验，以及探索最终如何将思维链重新结合回项目中。因此，本论文除了会探究思维链技术的影响，同时会探究大模型种类，参数量，运行时间等方面的要求，以此更好地与实际产品进行兼容，并更容易将研究成果进一步推广到工程方面。

本论文在实验过程中主要采用与大模型进行 API 问答的方式。在 Zero-shot 部分，直接对大模型进行提问。在 Few-shot 部分，会先根据当前问题类别进行人工例子设计，并且为大模型提供思考过程，以作为例子传给大模型进行测试。对于 Zero-shot-CoT 部分，则是增加了上文中提到的关键提示词“让我们以一步一步的方式来解决它，以确保我们得到正确的答案。”而 Few-shot-CoT 方面，则同样会进行人工例子设计。在这个部分本论文会尝试先使用因果推断技术对问题进行分析，先提出最可能的几个因素，之后基于此进行有针对性的例子设计，尝试进一步提升大模型回答的能力。在进行有关 Few-shot 相关的测试时，由于提供的例子的数量也会对模型的回答产生较大的影响，因此在实验中也将分别采用不同数量的例子进行测试，以对该因素对大模型回答的影响进行探究。

在引言部分中提到过，目前在校园欺凌相关的网站上，能搜集到的资料和数据主要都是校园欺凌的定义，特点，以及解决措施，防范措施等。这些问题我们可以定义

为概念类问题。而另一方面，在实际生活和实际场景中，大多数情况下孩子们都是在一个特定的环境和情景下遭受了校园欺凌。如果想回答这一方面的问题，目前需要教师，家长或者这一方面的专家综合考虑实际的情况，再参考防范以及解决措施给出相应的回答，回答难度较大。这一类问题我们定义为场景类问题，是本论文提出的一种新方法需要主要解决的一类问题。

具体来说，本论文提出的一种新的实验方式流程大概如图2所示。在数据集中获取输入一个问题之后，首先会对其进行预分类，这里对上文提到的 Abdeli 等人的做法^[19]进行了简化。在本次实验中，会先设定好数据集是由较多的场景类问题和较少的概念类问题所组成。而由于区分这两类问题比较简单，大模型并不需要进行预训练即可通过题目的文字进行正则化匹配等方法的识别。因此，问题的预分类并不在需要大模型进行一个预训练，而是直接使用 zero-shot 提示方法即可让大模型进行分类。由于本论文主要探索大模型在回答场景类问题上的表现，因此在分类后如果是概念类问题就直接让大模型进行回答，同时并不将其计入统计范围。而场景类问题则会尝试仿照上文中提到的方法，使用 Few-Shot 方法对该问题进行一个初步的因果推断分析。如果大模型能对该问题提取出相对应的原因 (treatment)，结果 (outcome) 和混淆因素 (confounders) 之后基于生成的 feedback 列表，根据里面的关键原因设计相对应的提示词，并且引导大模型更多地进行相关的思考，以此来优化提示词并生成答案，最后进行测试和比较。而如果大模型无法进行因果推断分析，或该问题的提问方式很难进行因果推断分析，则大模型将根据之前已经进行因果推断而生成提示词的例子，进行 Few-Shot 方法来回答问题。同样的，这种方法也是基于 Few-shot Prompting 的方法，在开始提问之前需要给大模型一些相关的例子。本论文称这种方法为 Causal Prompting，即初步结合了大模型的因果推断技术，并以此来反馈到提示词的增强上，从而引导大模型进行更好地逻辑推理和回答，以对提示工程技术进行可能的提升。

3.2 评估指标

本段将介绍一些被广泛应用于机器翻译和文本生成评估的相关指标，并阐述选择或不选择某些指标以及在使用该指标时，实验所关注的侧重点的说明。

1) BLEU

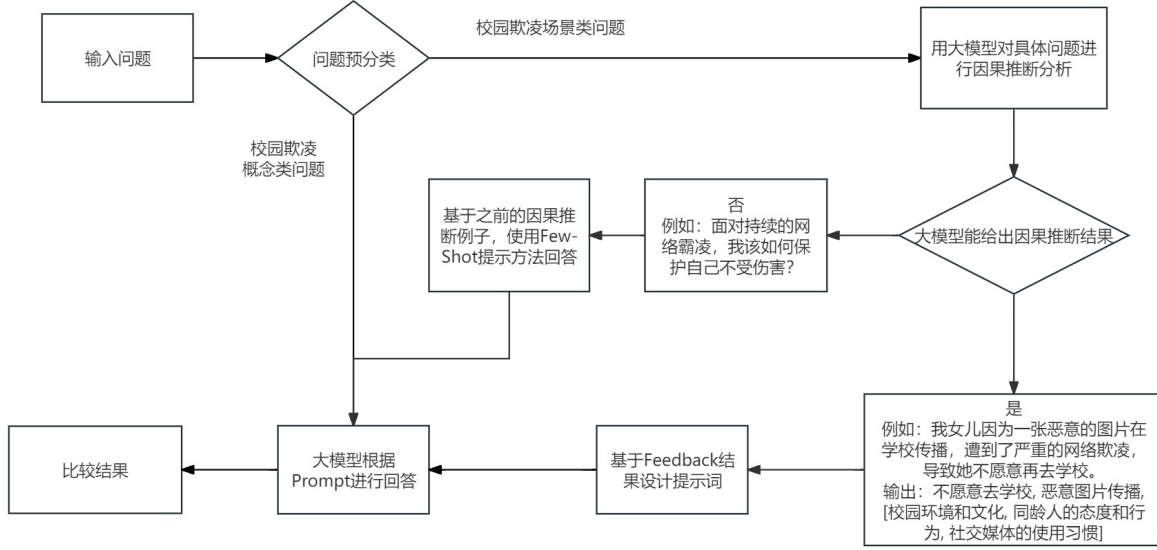


图 2 实验方法流程图展示

BLEU 是在 2002 年由 Papineni 所提出的机器翻译指标^[21]，并且应用非常广泛，直到今天仍然被大量选择为机器翻译和文本生成的评估指标，其公式如下：

$$BLEU = BP \cdot \exp \left(\sum_{n=1}^N w_n \log p_n \right) \quad (1)$$

其中， BP 是长度惩罚因子，用于惩罚过短的生成文本； p_n 是 n -gram 精确度，即匹配的 n -gram 在候选翻译中出现的频率； w_n 是对不同长度 n -gram 精确度的权重。

但是 BLEU 指标在实际测试时会出现一些问题，由于所评价的回答是由大模型生成的校园欺凌相关回答，这一类回答往往长度比较长，同时很难保证在用词上的准确性以及和答案的一致性，很难做到像机器翻译一样准确。实验过程中，BLEU 及其改进版指标 GLEU 在多项测试项目中的输出结果都为 0，一方面是因为生成的语义过于复杂，另一方面是由于长度相差过大，导致长度惩罚因子发挥过多的作用，导致输出结果变成 0。因此，在本论文中选择舍弃该指标，使用其他指标进行测试。

2) ROUGE

ROUGE 指标同样是一个非常传统的指标，在 2004 年由 Lin 等人提出^[22]。同样被广泛用于机器翻译和文本生成评估，其公式如下：

$$\text{ROUGE-N} = \frac{\sum_{s \in \{\text{Ref}\}} \sum_{\text{gram}_n \in s} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum_{s \in \{\text{Ref}\}} \sum_{\text{gram}_n \in s} \text{Count}(\text{gram}_n)} \quad (2)$$

其中 $\text{Count}_{\text{match}}(\text{gram}_n)$ 是在参考摘要中与生成文本匹配的 n -gram 数量, Ref 表示参考摘要文本, $\text{Count}(\text{gram}_n)$ 是参考摘要中 n -gram 的数量, 用于计算召回率。

Rouge 指标同样存在一些问题。由于本论文使用的是中文, 而 Rouge 原本是基于英文进行设计的, 因此本论文采用了其中文改进版 Rouge-Chinese 库进行实验。该库使用 jieba 库进行分词。其相比于传统的 Rouge 指标, 改进了中文的分句机制, 优化了计算 Rouge 分数时的内存占用, 从而提供更准确和官方的中文 Rouge 分数。该库会同时返回 Rouge 的三个值, 准确率 (precision), 召回率 (Recall) 和 F1 调和值 (F1)。考虑到在评测校园欺凌相关问题的回答时, 往往既要关注一个回答是否更多的涵盖了标准答案的内容, 即提供更多有价值, 能够帮助到使用的教师和学生内容, 又要关注其是否精确。这两者对答案评估都很重要, 因此本论文在使用 Rouge 指标时更加关注其 F1 调和值。

3) Distinct

Distinct 指标是一个相对较新的指标, 由 Li 等人于 2015 年提出^[23]。该指标并非广泛用于机器翻译, 而是用于通过计算生成文本中的唯一单词和二元组的数量来评价生成文本的多样性。其公式如下:

$$\text{Distinct-1} = \frac{\text{Unique Unigrams}}{\text{Total Unigrams}} \quad (3)$$

$$\text{Distinct-2} = \frac{\text{Unique Bigrams}}{\text{Total Bigrams}} \quad (4)$$

其中, Unique Unigrams 和 Unique Bigrams 分别表示生成文本中唯一的单词和二元组的数量, Total Unigrams 和 Total Bigrams 表示生成文本中总的单词和二元组的数量。Distinct-1 和 Distinct-2 用于衡量生成文本的多样性, 值越高表示文本的用词更丰富。但是该指标同样会受到长度以及文本内容影响, 并不能非常全面的反应一个生成文本的质量好坏, 因此主要作为辅助判断, 以用来加强说明实验结果。

4) BERT-Score

BERT-Score 指标由 Zhang 等人在 2019 年提出，并在 2020 年发表^[24]。不同于传统的指标，BERT-Score 使用了预训练好的 BERT 模型进行文本生成辅助评测，是一种基于预训练的 BERT 上下文嵌入的语言生成评估指标。BERT-Score 将两个句子的相似度计算为其标记嵌入之间的余弦相似度之和。解决了匹配释义和无法捕获遥远的依赖关系并惩罚语义上关键的顺序变化这两个弊端。其公式如下：

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^\top \hat{x}_j \quad (5)$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^\top \hat{x}_j \quad (6)$$

$$F_{\text{BERT}} = \frac{2 \cdot P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \quad (7)$$

其中 R_{BERT} 表示召回率, P_{BERT} 表示准确率, F_{BERT} 表示二者的调和平均值。 x_i 表示参考句子, \hat{x}_j 表示候选句子。

BERT-Score 基于预训练好的 Bert 模型，可以比较准确地评估候选文本和参考文本之间的相似度。在本论文中，因为所有的文本都是中文，因此要选择相关的 bert-base-chinese 模型进行运算。在模型最优层数选择上，参考了如图3的其他 Bert 模型默认层数值。由于该默认层数是由官方在英文数据集 WMT16 上测试后得出的结果，而中文暂无相关数据集，因此参考多语言 Bert 模型的默认数据，将评测时选择的模型层数定为 9。BERT-Score 基于模型的评测使其相对可靠，但问题在于，由于其衡量的是语义相近性，因此如果使用的提示方法并没有太大的提升，BERT-Score 所评测出来的结果相差就会很小，有些时候会并不利于数据分析。

Model	Best Layer	WMT16 To-English Pearson Correlation	Rank	Max Length
bert-base-uncased	9	0.6925	51	510
bert-large-uncased	18	0.7210	28	510
bert-base-cased-finetuned-mrpc	9	0.6722	74	510
bert-base-multilingual-cased	9	0.6681	81	510

图 3 基础 BERT 模型的默认层数示例^[24]

3.3 实验设计

3.3.1 实验任务及数据

在本论文中，预计一共进行两大项实验，第一项实验在一个较大的校园欺凌数据集上，分别测试不同参数量的大模型在 Zero-shot, Few-shot, Zero-shot-CoT, Few-shot-CoT 这四种提示方法下的回答，使用上文中提到的评测指标评测以后进行对比。该实验的主要目的在于探究不同参数量的大模型对于思维链技术的敏感程度，即思维链技术对该模型的影响程度有多大。另一方面，该实验同时还是对 CoT 和 Few-shot 技术对于普通的 Zero-shot 提示方法的提升性进行一个初步探究，以明确下一步的探究任务。在该任务中，使用的数据集由 Self-Instruct 生成。该数据生成框架主要服务于大模型微调，在已有少量数据下可以通过大模型的辅助来快速生成大量相似数据进行大模型微调，从而无需大量而繁琐的人工标注，极大减少了研究成本和人力资源消耗。最初的 40 余道题目是从校园欺凌相关的论文和新闻，研究报告等文献中提取的，并且用文章相关内容作为答案，之后使用 Self-Instruct 框架生成了 5000 道题目，并从中随机抽取出 1000 道题目进行测试。

在第二项实验中，数据主要依赖于手动构造。本论文主要的数据来源在引言中所介绍的几个校园欺凌相关的网站上，其中能找到大量的定义类题目。对于场景类题目，Cyberbullying research center 上会提供很多优质的数据和示例，但在其他网站上却很难找到相关题目。因此，本论文同样在 Reddit, Quora 或者知乎等论坛和线上交流平台进行了问题和回答的搜集，以构成测试数据。之后，本论文会使用在第一项实验中选定的模型进行进一步测试，同样会先使用 Zero-shot, Few-shot, Zero-shot-CoT, Few-shot-CoT 这四种提示方法进行测试，之后一方面会改变 Few-shot 技术中例子的数量，另一方面会添加在上文的方法中所提到过的结合因果推断技术的 Casual Prompting 提示方法进行测试和比对，以探究其在问题分类后在各数据集上的有效性。

3.3.2 大模型及采样选择

在大模型的选择上，本论文选择了通义千问 (Qwen) 系列的开源模型。在最开始的测试时会选择 Qwen1.5-7B-Chat, Qwen1.5-14B-Chat 和 Qwen1.5-72B-Chat 这三种模

型进行测试比较。选择 Qwen 系列模型的原因是，基于其模型技术报告^[25]，如图4所示，通义千问系列模型相比于同参数规模的其他模型在大多数任务上表现都更为出色，而且通义千问模型是由阿里团队进行开发，因此在面对中文的文本生成任务上的表现相比于其他外国开源模型也更加值得信赖。在所有实验过程中，本论文不对大模型进行任何微调和预训练，只靠 Few-shot 的人工生成的例子对其进行提示，并最终比较结果。在对于多个推理路径的采样处理上，本论文参考了 Radford^[10],Holtzman^[26],Wang^[16]等人的论文中的配置并进行了一定的调整，将模型的 Temperature 定为 0.85。

Model	Params	MMLU 5-shot	C-Eval 5-shot	GSM8K 8-shot	MATH 4-shot	HumanEval 0-shot	MBPP 3-shot	BBH 3-shot
MPT	7B	30.8	23.5	9.1	3.0	18.3	22.8	35.6
	30B	47.9	-	15.2	3.1	25.0	32.8	38.0
Falcon	7B	27.8	-	6.8	2.3	-	11.2	28.0
	40B	57.0	-	19.6	5.5	-	29.8	37.1
ChatGLM2	6B	47.9	51.7	32.4	6.5	-	-	33.7
InternLM	7B	51.0	53.4	31.2	6.3	10.4	14.0	37.0
	20B	62.1	58.8	52.6	7.9	25.6	35.6	52.5
Baichuan2	7B	54.7	56.3	24.6	5.6	18.3	24.2	41.6
	13B	59.5	59.0	52.8	10.1	17.1	30.2	49.0
LLaMA	7B	35.6	27.3	11.0	2.9	12.8	17.7	33.5
	13B	47.7	31.8	20.3	4.2	15.8	22.0	37.9
	33B	58.7	37.5	42.3	7.1	21.7	30.2	50.0
	65B	63.7	40.4	54.4	10.6	23.7	37.7	58.4
LLAMA 2	7B	46.8	32.5	16.7	3.3	12.8	20.8	38.2
	13B	55.0	41.4	29.6	5.0	18.9	30.3	45.6
	34B	62.6	-	42.2	6.2	22.6	33.0	44.1
	70B	69.8	50.1	63.3	13.5	29.9	45.0	64.9
StableBeluga2	70B	68.6	51.4	69.6	14.6	28.0	11.4	69.3
QWEN	1.8B	44.6	54.7	21.2	5.6	17.1	14.8	28.2
	7B	58.2	63.5	51.7	11.6	29.9	31.6	45.0
	14B	66.3	72.1	61.3	24.8	32.3	40.8	53.4

图 4 Qwen 系列模型与其他开源模型比较^[25]

4. 实验结果

本论文将在这部分展示并分析主要进行的两项实验的结果，以及上文中提到的 Causal Prompting 因果提示方法其他提示的效果比较。在本章节的所有实验数据展示中，所展示的评测指标数据是原值放大一百倍后保留两位小数的结果，以方便展示和对比。

4.1 基础实验分析

表 1 不同模型在不同 Prompt 方法下的测试结果

Model	Prompting	Rouge1	Rouge2	RougeL	Dist1	Dist2	Pbert	Rbert	Fbert
Qwen-7B	Zero-Shot	48.15	19.84	31.03	46.46	81.91	75.22	73.77	74.48
	Zero-Shot-CoT	47.90	19.63	30.68	45.80	80.97	75.01	73.59	74.28
	Few-Shot	52.89	22.43	35.23	49.90	84.13	74.28	74.14	74.19
	Few-Shot-CoT	54.85	21.48	35.16	46.10	83.18	72.59	74.85	73.69
Qwen-14B	Zero-Shot	51.19	21.46	33.18	48.57	83.55	74.72	74.13	74.40
	Zero-Shot-CoT	49.54	20.58	31.63	46.30	80.96	74.98	73.63	74.28
	Few-Shot	56.28	24.61	38.27	52.53	86.22	73.68	74.98	74.30
	Few-Shot-CoT	56.95	22.56	36.07	44.97	81.80	72.82	75.23	73.99
Qwen-72B	Zero-Shot	50.99	22.15	32.93	45.24	82.10	75.95	75.28	75.60
	Zero-Shot-CoT	50.38	21.68	32.06	43.74	79.92	75.93	74.74	75.32
	Few-Shot	55.60	24.75	37.93	50.19	85.17	74.70	75.40	75.03
	Few-Shot-CoT	58.22	21.33	36.90	45.50	81.73	70.42	75.17	72.70

该实验中，在数据集中的 1000 个问题上分别用不同的大模型和不同的提示方法，提示词进行了测试并统计了大模型的回答情况，其实验的结果如表1所示。首先可以对不同参数量的大模型之间的结果进行一个对比。如图5所示，在图中展示了不同模型在这四种提示方法下的 Rouge-L 指标值，以及一个其平均值的比较，每一种大模型对应的数据是靠在一起的。从图中可以看出，在所有提示方法下，Qwen-14B-Chat 和 Qwen-72B-Chat 这两种大模型在表现上都要明显好于 Qwen-7B-Chat，在平均值上这两种大模型也是要明显高于 Qwen-7B-Chat 的。这可以说明在参数较小时，的确模型的回答质量是要相对较低的。而对于 Qwen-14B-Chat 和 Qwen-72B-Chat 这两种模型之间的对比，两种模型在不同的提示方法下各有优劣。在不使用思维链技术的两种提示方法中，可以看出 Qwen-14B 的表现要相对较好，但在使用思维链的两种提示方法

中，则是 Qwen-72B 的表现更佳。同样的，在评测结果的平均值上，也是 Qwen-72B 要稍高一点。而这四种提示方法下的 F1-Bert 值如图6所示。从这项指标上，Qwen-14B 相对于 Qwen-7B 的优势和上升就相对较小，而 Qwen-72B-Chat 在前三个任务上的表现就很明显的好于另外两种模型，但是在最后的 Few-Shot-CoT 任务上却有着非常显著的下降。但即便如此，Qwen-72B 模型的平均 F1-bert 值仍然是要明显高于其他两种模型的。

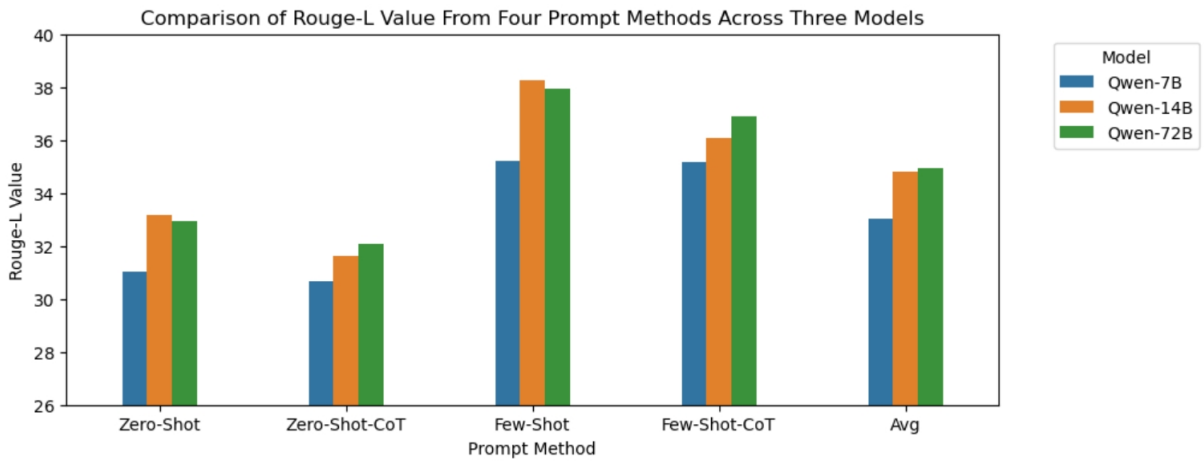


图 5 不同模型对于不同 Prompt 的 Rouge-L 结果

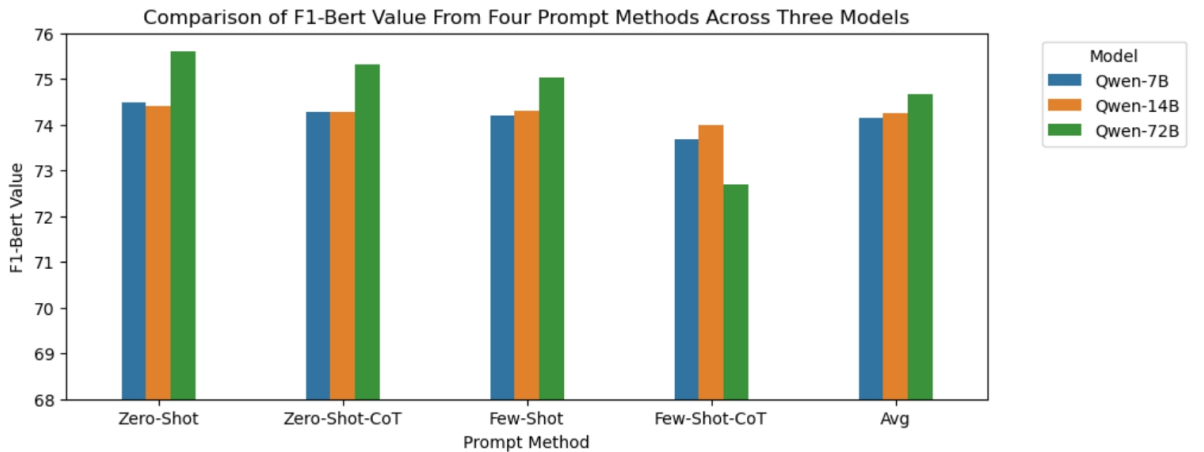


图 6 不同模型对于不同 Prompt 的 F1-Bert 结果

另外，前面提到在选择模型时除了考虑参数对模型回答质量的影响，同时也要考虑到模型回答问题的时间长短，如果模型回答问题时间过长，其易用性将会有一定的下降。因此，本研究统计了在相同时间内不同模型在不同提示方法下所生成题目的数量，如图7所示。从图中可以明显看出，Qwen-7B 模型由于其参数量较小，因此生成

答案所需的时间也相对较短，在相同时间内生成的题目数量远多于其他两种模型。而对于 Qwen-14B 和 Qwen-72B 这两种模型，其在生成题目数量多少上各有优劣，在平均值上 Qwen-72B 要略少于 Qwen-14B，但是综合考虑到 Qwen-72B 模型在不同任务上的回答质量较好，评测指标值都要高于其他两种模型，且在生成题目的效率上并没有慢特别多，因此认为 Qwen-72B 模型更适用于探究思维链相关的研究任务，在接下来的实验中将继续使用 Qwen-72B 模型进行测试。这一实验结果也符合之前 Wei. 等人在论文中的发现，即思维链技术在参数更大的模型上表现会相对更好。^[12]

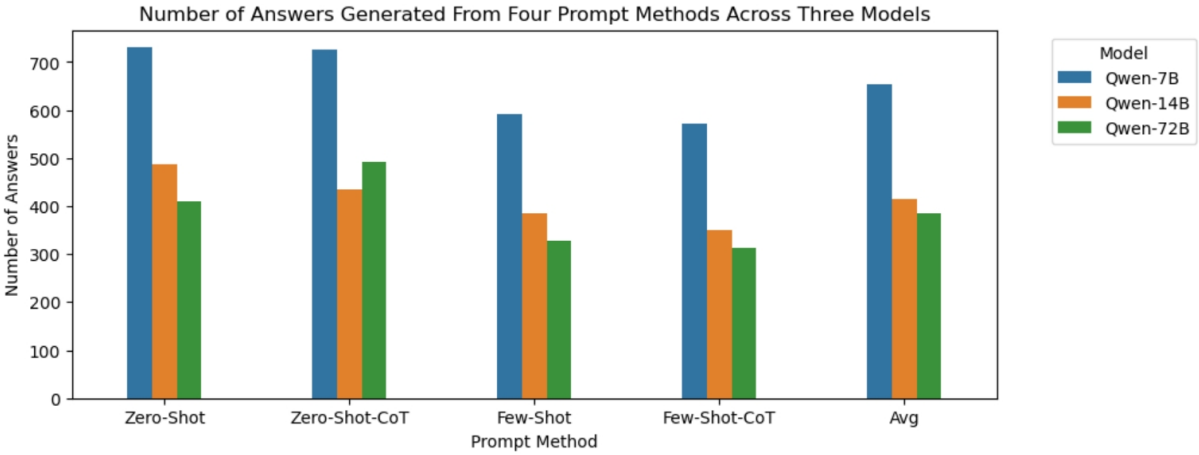


图 7 不同模型对于不同 Prompt 在相同时间内回答的题目数量

之后对于每种模型的不同指标值在不同 Prompt 方法下的变化趋势进行分析，其结果如图8所示。从这三张图的对比中可以看出，不同模型之间的三种指标的变化趋势基本相同，可以表明使用 Few-shot 和思维链提示技术对于不同模型的生成回答的影响是类似的。从 Rouge-L 指标上分析，在使用 Few-shot 技术替代 Zero-Shot 之后，无论是否使用思维链技术，其结果都有一个比较明显的提升，但是对于这两种提示技术，在使用思维链技术辅助之后，其结果却反而下降了。对于 F1-Bert 指标，除了在 Qwen-72B 模型上使用 Few-shot-CoT 时数据有一个明显下滑，其他值基本呈现一个比较轻微的下降趋势。而对于 Distinct-2 值，在使用 Few-shot 之后有一个明显的提升，但是在使用思维链技术后这个值又会有一定的下滑。

从整体上看，该实验展示出的结果与预期并不大相符。经过分析原因大概如下：该测试所使用的数据集少部分由各种校园欺凌相关的论文，资料，案情通报，法律法规等收集，之后由 Self Instruct 框架进行生成。在初期搜集到的数据，其问题质量并

不高，大多数是与定义等相关的题目，且有很多需要参考相关法律法规或者是某些特定的案件情况才能做出精准的回答，这超出了大模型的处理能力，使得大模型在回答这些问题时产生了一定的幻觉。另外，由于思维链技术和 Few-shot 技术本质上都十分依赖提示词与例子的设计，如果例子设计的不好，将导致模型输出的答案质量也下降。在本实验中，例子与题目可能并非在语义和类型上十分相符，因此可能引导大模型在后续的回答上出现方向性偏差，从而导致回答的 Rouge 和 bert 值都出现了下降。而在 Distinct 值上的结果比较反直觉，在使用思维链技术后其值出现下降，是因为提供思维链的思维过程后，大模型会仿照着着重于思维过程中的几个点回答问题，其回答的广泛度会降低。而大模型使用 Few-shot 之后 Distinct 值上升，则是因为大模型效仿例子进行问题分析，使大模型更广泛地分析不同的点来回答问题，从而回答的多样性有所上升。

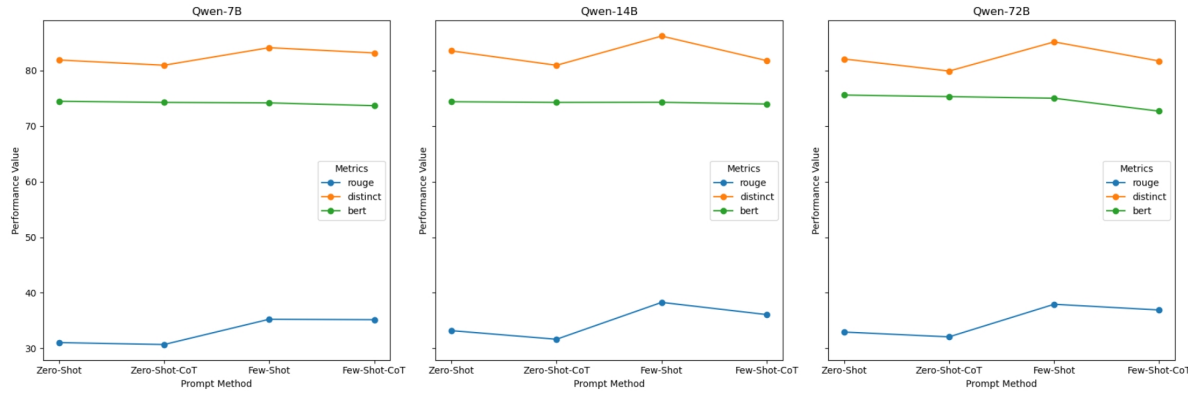


图 8 不同模型的三种指标变化趋势图

4.2 分类实验分析

在本实验中，根据上一步实验的结论继续使用 Qwen-72B-Chat 模型进行测试。在该测试中，主要探究使用不同数量的 Few-Shot 技术的回答质量比较以及本论文之前提出的 Casual Prompting 提示方法在回答场景类问题上的效果。

首先是使用不同数量的例子下，使用 Few-Shot 的几种提示方法的结果对比。其结果如图9和图10所示。这两种指标反应的结果变化趋势其实比较相似。可以看出，在不使用思维链技术，只进行 Few-Shot 的时候和使用 Casual Prompting 方法的时候，随着例子数量的增多，两种指标的值都出现了上升的情况。对于不使用 CoT 的情况来说，其上升相对较缓，每增加两个例子时其两种指标的提升并不多。而在使用 Casual

表 2 Qwen1.5-72B-Chat 测试结果

Prompting	Rouge1	Rouge2	RougeL	Dist1	Dist2	Pbert	Rbert	Fbert
Zero-Shot	44.14	16.75	27.90	46.73	81.70	71.88	73.28	72.55
Zero-Shot-CoT	44.09	17.03	29.00	45.59	80.29	71.84	73.61	72.69
One-Shot	43.68	17.14	29.45	50.26	83.07	72.34	73.50	72.88
One-Shot-CoT	42.73	16.29	28.04	55.09	87.26	72.67	71.76	72.16
One-Shot-Causal	42.38	15.97	27.83	54.46	85.62	72.96	71.47	72.15
Few-Shot(3)	42.97	17.27	29.57	53.89	86.65	73.16	72.60	72.85
Few-Shot-CoT(3)	42.36	16.47	28.82	56.86	89.67	73.52	71.39	72.41
Few-Shot-Causal(3)	44.51	18.48	30.09	49.80	82.62	73.15	73.87	73.47
Few-Shot(5)	43.96	17.48	29.81	53.49	86.28	73.85	72.39	73.08
Few-Shot-CoT(5)	41.14	14.89	25.87	56.84	89.83	72.43	70.26	71.28
Few-Shot-Causal(5)	44.50	18.39	30.27	49.14	81.80	72.96	73.97	73.43

Prompting 方法时，例子数量从 1 提升到 3 时，两种指标的值都出现了明显的上升，然而再从 3 提升到 5 时，Rouge-L 指标的提升就变得不明显，而 F1-Bert 更是出现了轻微的下降。而对于 CoT 技术，例子从 1 提升到 3 时，两种指标都有上升，但是当例子数量再提升到 5 时，这两种指标的值就都出现了很明显的下滑，效果不如只使用一个例子的 One-Shot 方法。同样地，从图中也可以简单的比较三种方法的情况。可以看出 Casual Prompting 方法在使用同样数量的例子时，在 One-Shot 上表现一般，但是在例子较多时表现明显好于其他两种方法。总体来说，随着例子的增加，使用不同提示技术的大模型的表现大多都会变好。而在本实验中，不使用 CoT 以及使用 Casual Prompting 这两种方法的结果相对符合预期，而使用 CoT 方法的结果却与预期出现了一定的偏差。

之后进行不同方法的指标对比。由于从上文的图表中可以发现，例子数量为 5 时，各项指标的表现相对更好，因此将其与最初的 Zero-Shot, Zero-Shot-CoT 两种方法进行一个对比，以探究不同方法是否对大模型在场景类问题上的回答能力带来了提升。其结果如图11所示。可以发现，在使用 Casual Prompting 之后，模型在三个方面的值都要高于模型在使用 Zero-Shot 和 Zero-Shot-CoT 时的值，说明该方法对大模型在回答场景类问题方面的能力有一定提升。同样也可以看出，使用该方法时在 Rouge-L 和 F1-Bert 这两个指标上的表现均要好于其他方法。同样也可以发现使用 CoT 时在这

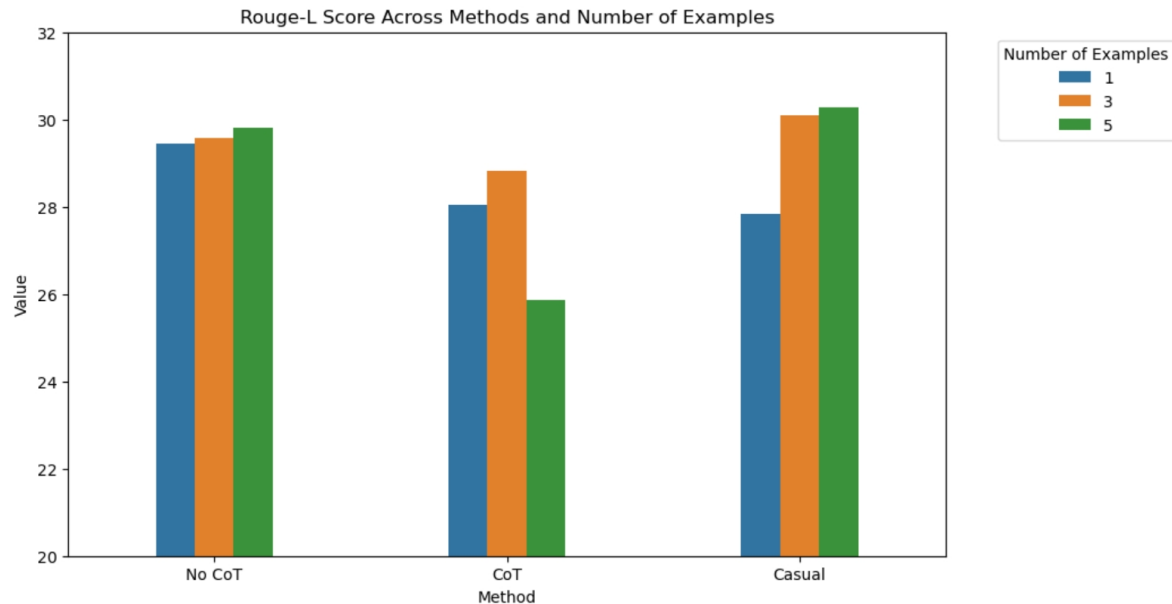


图 9 不同例子数量下不同方法的 Rouge-L 值

两个指标上的表现最差，而其他方式则随着提示工程的变化这两个指标的值都有上升。对于多样性则比较复杂，可以看出 Casual Prompting 方法的多样性值相比于使用 Few-Shot 的两种都下降了很多，稍高于 Zero-Shot，可以辅助说明使用该方法后，大模型的回答会更加注重于因果推断以后的几个影响因素，从而使其多样性有相对降低，但是相比于 Zero-Shot，则仍然考虑的比较全面，所以会有一个小幅度的上升。在使用 Casual Prompting 方法的测试上，可以发现大模型使用该方法之后对于回答这类场景类问题的能力是有一定的提升的。如表3所示，可以看出对于同一个问题的回答，使用 Zero-shot 时大模型回答的已经不错，基本覆盖了标准答案的内容，但是有少部分欠缺，而且并没有很好地体现对遭受欺凌者的心理辅导，而使用 Casual 提示方法生成的答案不仅覆盖了标准答案的内容，而且在思考的电上更丰富，也更加重视遭受欺凌者的自我关爱环节，体现了较好的人文关怀。

而本次实验中与预期不符的情况主要在于 CoT 技术表现不佳，Bert 指标变动较小等情况，对此，我在分析数据和测试结果后认为主要有以下影响因素和导致的结果：

- - 大模型的生成范式：大模型在回答这一类题目时，无论使用提示词与否，使用例子与否，大模型都会比较偏向于生成格式形如：1. 关键点 1。解释 1。2. 关键

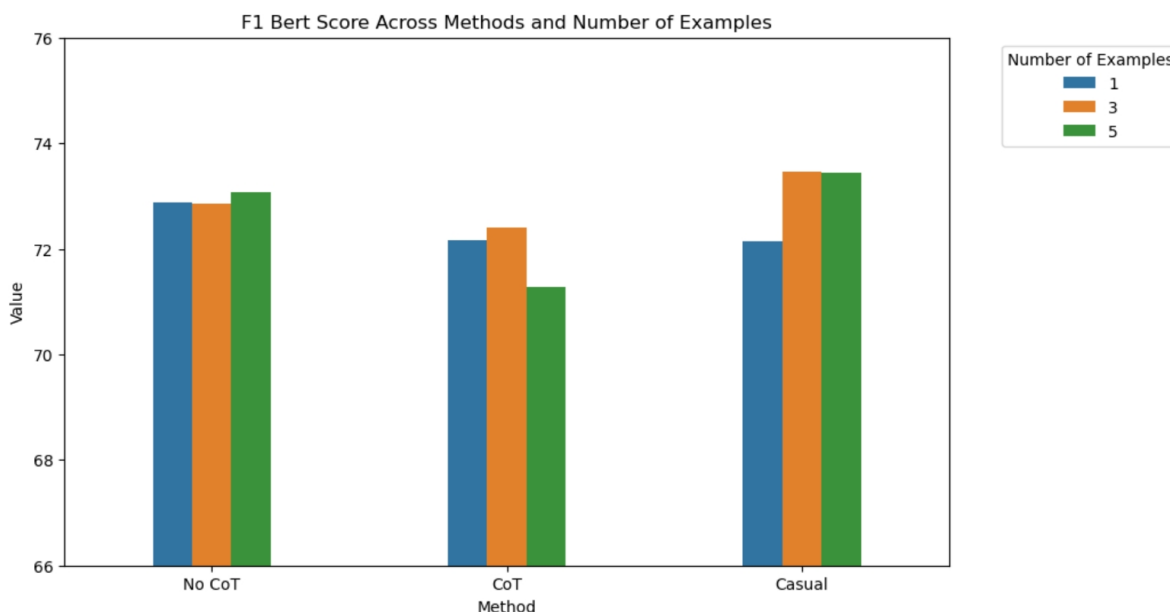


图 10 不同例子数量下不同方法的 F1 Bert 值

点 2。解释 2。这样一种完全由多个回答问题的关键点以及相应的解释所构成的答案。这种答案当然是有效的，但是他会限制大模型的输出格式，使其在格式上与答案有着本质的差异。很多问题的参考答案并不会使用这种格式，而是使用如“首先，其次”等连接词，虽然在意思上和作用上是相同的，但是在格式上有一定差异，另一部分则完全不使用连接词。在格式上的偏差导致使用 Rouge 指标评估的时候，其整体值会相对较低。而由于 Bert Score 指标主要评价的是语义上的相似性，因此其值是相对较高的，但是在不同方法上其显现的差距就会较小。

- - 大模型的思考点数量：在测试中以及浏览大模型生成的答案之后可以发现，大模型生成答案主要遵循其多个思考点。大模型是以思考点为基础继续进行推理，这其实与思维链以及 Casual Prompting 的想法是一致的，即发现问题关键点(推理的逻辑思维流程中的关键点)之后加以解释。由此推断，在大模型的思考点更多的时候，其生成内容就更丰富，多样性更高，而思考点少的时候，则需要有例子进行引导，使大模型能够生成符合情况而又不会过度发散的答案。因此在本实验中，可能是因为使用的例子质量不高导致大模型在回答后续的问题中产生偏离，因此指标的得分相对较低，从使用 CoT 时多样性最高而指标值最低也

可以看出，即大模型进行了很多点的回答，但这些点有可能偏离了答案。而这也从侧面印证了使用 Casual Prompting 时，大模型的回答能够基于因果推断比较准确地找到解决问题的关键点，从而生成一个质量比较高的回答。

- - 大模型的本身能力：其实从本研究的数据中可以看出，虽然在某些方法的实验上模型的指标值有所提升，但是其实并没有非常显著的提升。即使是在第一个实验中，使用 70 亿参数的 Qwen-7B 模型，其生成的答案质量也并不算差，其回答仍然对于一部分情景问题有着不错的质量。大模型目前本身回答这些问题的能力已经并不算弱，只使用基于思维链的提示技术和手段，或许已经不能让大模型获得太多的提升。

总之，该实验展示出了使用 Causal prompting 这一方法后，大模型在回答这类有关校园欺凌的场景类问题上的能力得到了一定的提升。而在使用思维链及其相关技术提升大模型回答问题这一方面，仍然有许多可以尝试的方向。

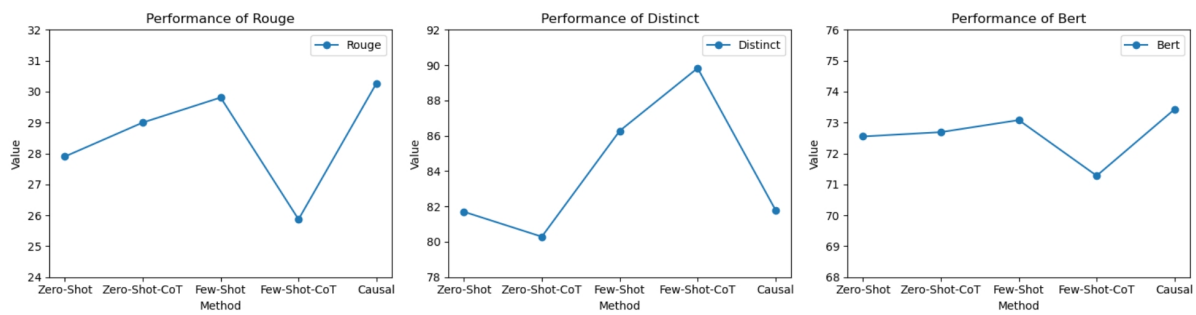


图 11 例子数量为 5 时不同方法下三种指标的比较

表 3 实验例子展示

问题	参考答案	Zero-Shot 答案	Casual 提示答案
问题:当我遭受网络欺凌时,我感到非常抑郁,不想与任何人交流。即使欺凌者认为这很有趣,我却受到了长期的影响。面对网络欺凌,我应该如何寻求帮助,避免长期受到心理影响?	面对严重的抑郁情绪,首先寻求专业帮助是非常重要的。你可以联系心理医生或心理咨询师,他们能提供专业的心理治疗,如认知行为疗法,帮助你理解和应对网络欺凌导致的心理困扰。同时,不要独自承受,告诉家人和朋友你的困扰,他们的支持对你的康复至关重要。参与支持小组也是个好办法,这里有很多经历过类似情况的人,可以互相鼓励,分享经验。许多学校和社区都有这样的资源,你可以寻求当地机构的帮助。保持健康的生活习惯,如规律作息、均衡饮食和适度运动,对心理健康也有积极影响。此外,学习和实践一些放松技巧,如深呼吸、冥想或者渐进性肌肉松弛,可以帮助你保持冷静。记住,改变你对网络欺凌的看法,学会说“不”,并设定明确的界限,这有助于减少未来可能的伤害。别忘记,你有权保护自己的心理和情感健康。	面对网络欺凌,你可以采取以下步骤来寻求帮助并减轻其对你的心理影响:1. 保存证据:将欺凌的聊天记录、邮件、评论或截图保存下来,这将有助于你在需要时证明发生了什么。2. 报告和阻止:大多数社交媒体平台都有举报欺凌行为的功能。立即使用这些工具来报告欺凌者,并阻止他们无法再联系到你。3. 告诉信任的人:跟你的朋友、家人或者老师分享你的经历,让他们知道你正在遭受网络欺凌。他们的支持和理解会对你有所帮助。4. 寻求专业帮助:如果感觉情绪难以控制,不要犹豫寻求心理咨询师或心理医生的帮助。他们可以提供专业的应对策略和治疗建议。5. 加入支持团体:有许多在线和线下的支持团体,专门为遭受欺凌的人提供帮助。在这里,你可以找到共鸣,学习他人如何应对类似的情况。6. 保持健康的生活方式:保持良好的饮食、睡眠和运动习惯,这些都有助于你的心理健康。	面对网络欺凌,首先要认识到这不是你的错,你有权保护自己不受伤害。以下是一些寻求帮助和避免长期心理影响的建议:1. 与信任的人分享:告诉你的家人、朋友或亲密的成年人你正在经历的情况,他们的支持和理解非常重要。2. 记录证据:保存欺凌的证据,包括消息、截图、邮件等,这可能在举报或寻求法律援助时有用。3. 向平台举报:利用社交媒体平台的举报功能,将欺凌行为报告给网站管理员,请求删除不当的内容。4. 寻求专业帮助:联系心理咨询师或心理卫生专业人士,他们可以提供应对策略和治疗,帮助你处理抑郁情绪。5. 加入支持团体:寻找网络欺凌的受害者支持小组,与有类似经历的人交流,共享应对技巧。6. 了解你的权利:研究当地的相关法律法规,了解如何保护自己免受网络欺凌。7. 培养自我关爱:进行正念练习、运动或其他放松活动,以帮助缓解压力和恢复自信。8. 设定网络边界:调整社交媒体设置,限制与欺凌者的互动,或者暂时远离网络环境,给自己一些时间恢复。记住,寻求帮助是恢复的第一步,不要独自承受痛苦。

5. 总结与展望

5.1 总结

本论文通过设计实验，证明了思维链及其相关提示工程技术在协助大模型回答校园欺凌相关方面问题有一定效果，能够使得大模型在不需要微调和预训练的情况下，仅通过提示工程和构建的相关例子即可生成更高质量的回答。然而本论文同样存在一些不足，比如数据量并非十分充足，以及大模型生成回答的相关变量和印象因素过多，并不能完全探究出一个具体而泛用性强的提升方法等。综上所述，本论文比较开创性地提出了使用大模型和思维链技术来回答校园欺凌相关问题，是防控和解决当前的校园欺凌问题的一种新思路，通过为学生和老师提供高质量的校园欺凌相关问题的解答以及辅导，从而实现校园欺凌相关教育，以此来尝试防止，根除校园欺凌现象。

5.2 展望

本研究属于校园欺凌相关研究中比较具有创新性的研究，提出了使用大模型技术来回答问题并结合了思维链技术，属于该领域中比较新颖的研究。因而，该研究还有很多相关方向探究可以进行尝试。比如探索大模型在其他校园欺凌相关方面的回答能力，以及大模型在这个领域的整体应用，比如能够进行分类任务，生成人物，识别任务，问答任务，推断任务等^[27]，从而在校园欺凌相关领域构建一个完善的大模型解决问题的系统。另外，如果不仅局限于使用提示工程技术，在工程方面也可以尝试进行进一步的微调，从而构建类似 MathGPT^[28], ComputeGPT^[29]等能解决专业领域知识的大模型。总而言之，该研究可以探索的方向还很多，在大模型得到普遍应用的今天，用其来解决校园欺凌相关的问题是一个很有价值的尝试方向。

参考文献

- [1] 陈晓英. 校园欺凌谁来解围. 法制日报[Z]. 2015.
- [2] UNESCO. School Violence and Bullying Global Status Report[Z]. 2020.
- [3] 澎湃新闻. 半数受害者为中学生, 揭开校园霸凌的隐秘角落 [N]. 2023.
- [4] 开金英. 我国反校园欺凌立法困境与路径研究[J]. 陕西青年职业学院学报, 2018: 63-67+76.
- [5] 李亚宁. 中小学校园欺凌中受害人权益保护研究[D]. 西北民族大学, 2020.
- [6] 晁磊. 小学班主任防治校园欺凌现状的研究[D]. 曲阜师范大学, 2021.
- [7] YE L, LIU T, HAN T, et al. Campus violence detection based on artificial intelligent interpretation of surveillance video sequences[J]. Remote Sensing, 2021, 13(4): 628.
- [8] YOUNG OH E, SONG D, HONG H. Interactive computing technology in anti-bullying education: The effects of conversation-bot's role on K-12 students' attitude change toward bullying problems [J]. Journal of Educational Computing Research, 2020, 58(1): 200-219.
- [9] SAHOO P, SINGH A K, SAHA S, et al. A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications[J]. arXiv preprint arXiv:2402.07927, 2024.
- [10] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI blog, 2019, 1(8): 9.
- [11] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
- [12] WEI J, WANG X, SCHUURMANS D, et al. Chain-of-thought prompting elicits reasoning in large language models[J]. Advances in neural information processing systems, 2022, 35: 24824-24837.
- [13] KOJIMA T, GU S (, REID M, et al. Large Language Models are Zero-Shot Reasoners[C/OL]. in: KOYEJO S, MOHAMED S, AGARWAL A, et al. Advances in Neural Information Processing Systems: vol. 35. Curran Associates, Inc., 2022: 22199-22213. https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.
- [14] ZHANG Z, ZHANG A, LIM S, et al. Automatic chain of thought prompting in large language models [J]. arXiv preprint arXiv:2210.03493, 2022.
- [15] ZHOU Y, MURESANU A I, HAN Z, et al. Large language models are human-level prompt engineers[J]. arXiv preprint arXiv:2211.01910, 2022.
- [16] WANG X, WEI J, SCHUURMANS D, et al. Self-consistency improves chain of thought reasoning in language models[J]. arXiv preprint arXiv:2203.11171, 2022.
- [17] ZHOU D, SCHÄRLI N, HOU L, et al. Least-to-most prompting enables complex reasoning in large language models[J]. arXiv preprint arXiv:2205.10625, 2022.
- [18] PEARL J. Causal inference in statistics: An overview[J]. 2009.
- [19] ABDALI S, PARIKH A, LIM S, et al. Extracting Self-Consistent Causal Insights from Users Feedback with LLMs and In-context Learning[J]. arXiv preprint arXiv:2312.06820, 2023.

- [20] PAUL D, WEST R, BOSSELUT A, et al. Making Reasoning Matter: Measuring and Improving Faithfulness of Chain-of-Thought Reasoning[J]. arXiv preprint arXiv:2402.13950, 2024.
- [21] PAPINENI K, ROUKOS S, WARD T, et al. Bleu: a method for automatic evaluation of machine translation[C]. in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002: 311-318.
- [22] LIN C Y. Rouge: A package for automatic evaluation of summaries[C]. in: Text summarization branches out. 2004: 74-81.
- [23] LI J, GALLEY M, BROCKETT C, et al. A diversity-promoting objective function for neural conversation models[J]. arXiv preprint arXiv:1510.03055, 2015.
- [24] ZHANG* T, KISHORE* V, WU* F, et al. BERTScore: Evaluating Text Generation with BERT [C/OL]. in: International Conference on Learning Representations. 2020. <https://openreview.net/forum?id=SkeHuCVFDr>.
- [25] BAI J, BAI S, CHU Y, et al. Qwen technical report[J]. arXiv preprint arXiv:2309.16609, 2023.
- [26] HOLTZMAN A, BUYS J, DU L, et al. The curious case of neural text degeneration[J]. arXiv preprint arXiv:1904.09751, 2019.
- [27] WANG J, SHI E, YU S, et al. Prompt engineering for healthcare: Methodologies and applications [J]. arXiv preprint arXiv:2304.14670, 2023.
- [28] SCARLATOS A, LAN A. Tree-Based Representation and Generation of Natural and Mathematical Language[C/OL]. in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, 2023: 3714-3730. <https://aclanthology.org/2023.acl-long.205>.
- [29] LEWIS R H, JIAO J. ComputeGPT: A computational chat model for numerical problems[J]. arXiv preprint arXiv:2305.06223, 2023.

致谢

首先，感谢我的指导宋轩老师，感谢老师提出了这个项目和选题，并且一直关注项目的进度并提供支持。其次，感谢负责项目的谢洪彬学长。谢博一直负责组织项目的各种讨论和主持每周的组会，与我们交流目前的情况并解答问题。学长的组织在项目的进展中起到了很大的作用，非常感谢学长在项目和毕设的进展中对我提供的帮助和支持。同时，也要感谢深圳铠硕达科技有限公司的温鹏先生，作为企业方代表，每周参与组会进行讨论，关切我们的进展，并且给予各种宝贵的指导意见与帮助，提供技术方面的支持。

同样的，我也要感谢我的父母，没有他们的支持，关爱和陪伴我将无法走到今天；感谢所有对我言传身教的老师，在我的求学之路上引导着我；感谢我的朋友和同学们，感谢你们一直以来的陪伴，丰富了我生活的色彩；感谢所有曾经帮助过我的人，在我困难时为我指引方向。感谢他们对我一直以来的支持和理解。