硕士学位论文

面向社交媒体的居高临下言论检测

Patronizing and Condescending Language Detection in Social Media Contexts

作 者 姓 名:XXX

学 号:XXXXXXXX

指 导 教 师:XXX

学科、 专业:XXX

答 辩 日 期:2025 年 X 月

大连理工大学

Dalian University of Technology

学位论文原创性申明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经指明引用的内容外，学位论文不含任何其他个人、法人或者非法人组织已经发表或尚未发表的作品，且学位论文中已经指明作者姓名或者名称、作品名称的内容，不影响该作品的正常使用，也不存在不合理地损害相关权利人的合法权益的任何情形。对学位论文研究做出重要贡献的个人和法人或者非法人组织，均已在论文中以明确方式标明，且不存在任何著作权纠纷。

若因声明不实，本人愿意为此承当相应的法律责任。

学位论文题目：面向社交媒体的居高临下言论检测

作 者 签 名 ： _____ 日期：2025 年 ____ 月 ____ 日

大连理工大学学位论文版权使用授权书

本人完全了解大连理工大学有关学位论文知识产权的规定，在校攻读学位期间论文工作的知识产权属于大连理工大学，允许论文被查阅和借阅。学校有权保留论文并向国家有关部门或机构送交论文的复印件和电子版，可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印、或扫描等复制手段保存和汇编本学位论文。

学位论文题目：面向社交媒体的居高临下言论检测

作 者 签 名 ： _____ 日期：2025 年 ____ 月 ____ 日

导 师 签 名 ： _____ 日期：2025 年 ____ 月 ____ 日

硕士学位论文答辩委员会

答辩人:

答辩委员会委员:

大连理工大学教授: _____（注: 主席）

大连理工大学教授: _____

大连理工大学教授: _____

大连理工大学教授: _____

大连理工大学副教授: _____

答辩日期: 2025 年 X 月 X 日

答辩地点: 大连理工大学

摘要

随着社交媒体的快速发展，在线交流愈加高效便捷，促进了全球信息共享。然而，网络言语暴力和有害言论也随之蔓延。居高临下言论（Patronizing and Condescending Language, 简称 PCL）作为典型的隐性有害言论，以伪善、虚假同情等方式呈现，不仅强化对弱势群体的歧视，还加剧社会分裂，带来潜在风险。传统方法依赖模糊的居高临下定义和小规模检测器，难以有效识别隐性攻击，使该任务面临挑战。本文围绕居高临下言论检测，探讨高质量数据构建、大型语言模型应用及多模态框架引入，对于社交媒体中隐性有害言论的防控与弱势群体保护具有重要理论与应用价值。

（1）针对居高临下研究数据稀缺和中英文数据存在模糊注释的问题，本文提出 CondescendCN 框架，并构建 CCPC，这是首个中文居高临下语料库。本文从新浪微博和知乎收集 11k 条三级居高临下数据，并引入毒性强度（TS）作为关键融合信息，以提升识别的精准度。实验结果表明，TS 信息有效增强了模型的泛化能力，即使迁移到 120k 条社交媒体数据，模型仍能保持较强的检测能力。本研究填补了中文居高临下数据的空白，并为后续中英双语研究提供了坚实的语料支持。

（2）针对传统预训练模型检测能力存在瓶颈的问题，本文构建了 PclGPT，这是首类专门针对居高临下言论开发的大型语言模型组。大模型时代大规模参数有助于隐性有害言论检测能力的提升，因此本文提出基于大模型的预训练-指令微调方法和组合高效微调范式，并整合 Pcl-PT/SFT 数据集，以此为基础开发 PclGPT-EN/CN 中英双语大模型以支持居高临下言论识别。实验表明，PclGPT 在不同弱势群体的居高临下检测中展现出显著优势，能够更精准地区分居高临下的细粒度表现形式，从而提升了隐性有害言论的检测能力。

（3）针对居高临下言论还涉及非语言歧视性信号特征（如面部表情、神态），而目前仅有单一模态工作的问题，本文提出该领域首个多模态框架，包括 PCLMM 数据集及MultiPCL 多模态检测器。本文从 Bilibili 平台收集 715 条带有标注的高质量视频，并引入面部表情检测模块以辅助识别。实验表明，多模态融合有效提升了居高临下言论识别的准确性，弥补了单一文本模式的缺陷。本研究为隐性有害言论的多模态检测提供了全新视角，并拓展了其在视频审核、社交媒体治理等领域的应用潜力。

关键词：社交媒体；居高临下言论；细粒度标注；大型语言模型；多模态检测

ABSTRACT

With the rapid development of social media, online communication has become increasingly efficient and convenient, facilitating global information sharing. However, online verbal violence and harmful speech have also proliferated. Patronizing and Condescending Language (PCL), as a typical form of implicit harmful speech, is often expressed through hypocrisy and false sympathy. It not only reinforces discrimination against vulnerable groups but also exacerbates social division and poses potential risks. Traditional methods rely on vague definitions and small-scale detectors, making it difficult to effectively identify implicit attacks. This study focuses on PCL detection, exploring high-quality dataset construction, the application of large language models (LLMs), and the introduction of multimodal frameworks. The work holds significant theoretical and practical value for controlling implicit harmful speech on social media and protecting vulnerable communities.

(1) To address the scarcity of PCL research data and the ambiguity of annotations in both Chinese and English datasets, this study proposes the CondescendCN framework and constructs CCPC, the first Chinese PCL corpus. A total of 11k three-level annotated samples were collected from Sina Weibo and Zhihu. Toxicity Strength (TS) is introduced as a key fusion feature to improve detection precision. Experimental results demonstrate that TS effectively enhances the model's generalization ability. Even when transferred to a 120k-scale social media dataset, the model still maintains strong detection performance. This work fills the gap in Chinese PCL data and provides a solid corpus foundation for subsequent bilingual research.

(2) To address the limitations of traditional pre-trained models in PCL detection, this study constructs PclGPT, the first group of large language models specifically developed for patronizing and condescending language. In the era of large models, the massive parameters contribute to improved detection of implicit harmful speech. Therefore, this study proposes a pre-training and instruction-tuning approach based on LLMs, combined with an efficient fine-tuning paradigm. By integrating the Pcl-PT/SFT dataset, we develop the bilingual LLMs PclGPT-EN/CN to support PCL recognition. Experimental results demonstrate that PclGPT shows significant advantages in detecting PCL across different vulnerable groups, enabling more accurate identification of fine-grained PCL expressions and enhancing the detection of implicit harmful speech.

(3) To address the issue that PCL also involves non-verbal discriminatory cues (such as facial expressions and demeanor) while current research is limited to single-modality ap-

proaches, this study proposes the first multimodal framework in this field, including the PCLMM dataset and the MultiPCL multimodal detector. A total of 715 high-quality annotated videos were collected from the Bilibili platform, and a facial expression detection module was introduced to assist recognition. Experimental results show that multimodal fusion significantly improves the accuracy of PCL detection, compensating for the limitations of text-only approaches. This study offers a new perspective for multimodal detection of implicit harmful speech and expands its application potential in video moderation and social media governance.

Key Words: Social Media; Patronizing and Condescending Language; Fine-grained Annotation; Large Language Model; Multimodal Detection

目　　录

TABLE OF CONTENTS

图 目 录

表 目 录

1 绪论

1.1 研究背景与意义

有害（毒性）言论治理是近年来自然语言处理领域的重要研究方向，因其在保障互联网安全、促进健康在线交流方面的关键作用而受到广泛关注。有害言论的存在对日益发展的互联网环境造成了深远的负面影响，不仅直接危害用户的心理健康，还可能在潜移默化中助长社会不平等，加剧线上线下的对立情绪。因此，针对有害言论的检测和治理已成为迫在眉睫的研究任务，其中对隐性有害言论的识别尤为重要。

作为隐性有害言论的一个重要分支，居高临下言论（Patronizing and Condescending Language，简称 PCL），也称为屈尊或鄙视言论，是自然语言处理领域中一个尚未充分开发、开放且充满挑战的研究方向[1, 2]。Pérez-Almendros 等人[1] 的研究表明，当个体在言语表达中表现出对他人的优越感，或以一种看似同情的方式描述对方时，该表达可被归类为居高临下言论。此类言论主要针对社会中的弱势群体，往往以隐性、难以察觉的方式表达歧视信息，因而其具有更强的欺骗性和危害性。不同于网络中具有明确攻击性、易于检测的传统仇恨言论和冒犯性言论，居高临下言论通常由优势地位群体对弱势群体发起，且言论本身通常是发出者无意识的，由表面上的看似“善意”或“关怀”驱动，并通过精心包装的措辞进行表达[2, 3]。

针对居高临下言论的研究在隐性有害言论检测中具有开创性的贡献。现有的有害言论研究主要集中于显性有害言论的识别，如仇恨言论（Hate Speech）、网络欺凌（Cyberbullying）以及冒犯性言论（Offensive Language）。这些言论的主要特点是具有明显的攻击对象、直接的攻击性语言表达，且在情感倾向上通常呈现为负面情绪。因此，在相关领域已经形成了较为成熟的检测方法，如 Zampieri 等人[4] 针对仇恨检测的研究。然而，当前对于隐性有害言论的关注仍然较少，因为这类言论往往更加复杂且难以界定。居高临下言论便是典型的隐性有害言论之一，其表达方式更加隐晦，攻击性不易察觉，同时对目标群体的心理和社会影响可能更加深远。例如，在针对单亲家庭这一弱势群体的讨论中，**“单亲家庭的孩子就应该多了解一下现在的社会规则”**这句话并未直接包含攻击或侮辱性词汇，但却隐含了对该群体的偏见，即默认单亲家庭的孩子不懂社会规则，从而形成了一种带有优越感的刻板印象。这种“无形的网络暴力”虽然不会直接引发冲突，却可能对受害者造成长期的心理创伤。因此，研究居高临下言论对于隐性有害言论的检测至关重要，不仅能填补当前研究空白，还能为更广泛的有害言论检测提供基线和范式。

到目前为止，针对于居高临下言论的研究仍存在严重不足，主要面临以下挑战：首先，传统方法在识别这类隐含性表达时存在显著局限，难以有效检测其中的隐性有

害特征。由于其表达方式隐晦且依赖语境，基于关键词匹配、规则检测或浅层机器学习的方法难以适用。即便是当前流行的预训练语言模型（Pre-trained Language Models，简称 PLMs），在这一任务上仍然存在诸多困难。其一，PLMs 对隐含语义的理解能力有限，居高临下言论往往缺乏明显的负面情绪或攻击性词汇，使得仅依赖词汇层面进行分类的模型难以识别。其二，这类言论的核心特征是社会权力的不对等，而这一信息通常需要结合更广泛的世界背景知识才能被准确捕捉，不同的国家和地区在不同语境下会对其有不同的理解。单纯依赖文本建模的 PLMs 在理解这一世界知识时存在瓶颈，这也意味着需要进一步面向多语种社交媒体平台开展居高临下检测研究。在大型语言模型（又称大模型，简称 LLMs）出现的今天，使用和引导大模型所具有的更加广泛的有害言论预训练领域知识，并结合高效微调范式进行面向多平台的居高临下言论检测研究，在理论上已成为可能。

其次，多模态技术在歧视性言论检测中的潜力尚未得到充分挖掘。即结合文本、图像、音频等信息进行联合建模，以提升模型的理解能力。多模态方法已在图文匹配、情感分析等领域取得了突破性进展。然而，在涉及社会公平性问题的研究，尤其是居高临下言论等隐性有害言论的检测方面，探索仍相对有限。居高临下言论不仅限于文本表达，还可以通过面部表情、语调、语音模式等非语言信号传递信息，比如在视频和多媒体内容中，优势群体成员可能通过微妙的面部表情、不屑的语气或特定的肢体语言来向弱势群体传达居高临下的态度，而这些非语言特征往往比单纯的文本更能揭示隐含的权力不对等关系。然而，现有有害言论检测方法主要依赖文本分析，缺乏对多模态特征的有效利用，使得居高临下言论的隐性表达难以被精准识别。此外，多模态数据的标注成本高、数据格式复杂，也进一步限制了该方向的发展。因此，如何充分利用多模态信息，提高对隐性歧视性表达的检测能力，仍然是当前有害言论检测研究中亟待解决的问题。

1.2 研究现状

1.2.1 有害言论检测

有害言论（又称毒性言论）通常被定义为包含不礼貌、缺乏尊重或非理性的陈述。Dixon 等人[5] 指出这类语言可能对对话的氛围产生负面影响，并导致参与者因受到攻击或冒犯而退出讨论。这类言论不仅损害了在线交流的质量，还可能对特定个体或群体造成心理上的压力或伤害。因此，对有害言论的检测成为自然语言处理和计算社会科学领域的一个重要研究方向。在过去的十多年中，研究者对有害言论的不同子类别进行了深入探讨，特别是在仇恨言论检测方面取得了显著进展[6–8]。仇恨言论通常涉及针对特定群体的直接攻击，例如基于宗教、种族或性别的歧视性言辞，这些言论往往具有强烈的负面情感，并可能煽动仇恨或暴力。因此，许多研究致力于识别和抑制

这类明显带有敌意的语言，以减少其在社交媒体和在线论坛上的传播。然而，当前的大多数研究主要集中在显性和直接的攻击性言论，而较少关注那些更加隐晦但同样具有负面影响的有害的语言形式，例如刻板印象（Stereotypes）、讽刺（Sarcasm）和居高临下言论。这些隐性有害言论往往更具迷惑性，因为它们在表面上可能显得无害甚至积极，但实际上可能会加深社会偏见或导致受众群体的不适。这一现象促使研究者近年来进一步关注这些更具隐蔽性和复杂性的有害言论类型。

1.2.2 居高临下言论

居高临下言论作为隐性有害言论的重要子类别，在多个学术领域均有深入研究，包括 Margic 等人[9] 在语言学领域对居高临下的研究、Giles 等人[10] 在社会语言学领域对其的研究、Huckin 等人[11] 在政治学领域对其的研究以及 Komrad 等人[12] 在医学领域对其的研究等。在自然语言处理领域，尽管有大量关于有害语言的研究，但大多聚焦于显性、直接攻击性的语言现象，如 Conrot 等人[13] 对于虚假新闻检测的研究、Atanasova 等人[14] 进行针对可信度预测与事实核查的工作、Zampieri 等人[4] 和 Basile 等人[15] 使用 BERT 等相关深度学习模型进行攻击性语言建模、Derczynski 等人[16] 使用神经网络进行谣言传播的有效检测等。迄今为止的这些工作都忽视了对潜在的隐性有害言论的识别。

相比之下，居高临下言论作为一种更具隐蔽性和语义复杂性的有害表达形式，直到近年才逐渐受到关注。Pérez-Almendros 等人[1] 在整合弱势群体类别的基础上正式引入了居高临下言论的正式定义，指出这类语言往往以优越态度呈现，或将弱势群体描绘为值得怜悯或需要援助的对象。不同于传统仇恨言论的直接攻击，居高临下言论更关注隐性危害特征，常以委婉语气面向边缘化群体表达贬抑性态度，因此在毒性评分上通常偏低，也更难被现有检测系统识别。Wong 等人[2] 指出，居高临下的表意往往是在“善意”驱动下无意识产生的，并通过精心修饰的语言进行表达。Xu 等人[17] 进一步发现，居高临下言论对弱势群体的刻板描绘可能加剧不同社区的排斥与不平等，降低用户的在线参与意愿，甚至促使其退出现有社区。Mendelsohn 等人[18] 则从计算语言学角度分析了媒体语言如何随时间推移逐步“非人化”少数群体。

在对这一领域使用 NLP 技术的实质工作中，Wang 等人[19] 最早提出了用系统方法建模居高临下的表达，并构建了一个带有社交媒体标注的数据集 Talkdown，推动了相关评估框架的发展。Pérez-Almendros 等人[20] 提出了基于深度学习的改进模型，提升了对隐性有害语言的识别能力，但由于预训练资源和任务建模的限制，该工作指出当前模型在捕捉居高临下的隐含语义特征方面仍面临挑战。

居高临下的核心特征在于其隐含的社会权力不对等关系，语言表达常通过怜悯、施舍或“善意”话语构建不平衡的社会结构，从而进一步边缘化弱势群体。这类表达的长期影响包括降低受众的社会参与度、削弱自我认同，并加深社会排斥。在社交媒

体语境中，居高临下甚至可能导致用户减少互动，在严重情境下还可能引发抑郁、焦虑等心理健康问题。因此，其潜在的社会危害性显著增加了其自动检测的挑战性，也凸显了该领域对具有更强理解能力的大型语言模型和多模态模型的迫切需求。

1.2.3 基于大型语言模型的有害言论检测

近年来，基于解码器的大模型（如 ChatGPT[21]、GPT-4[22] 以及 LLaMA[23]）在文本生成领域引发了变革的同时，大模型也被越来越多地应用于有害语言的检测与防护任务中。大模型通过大规模预训练，能够在少样本甚至零样本的情况下，通过提示（Prompt）完成文本分类任务，减少了对大量标注数据的依赖。Shaikh 等人[24] 证明了零样本链式推理（Zero-shot CoT）显著提升了 LLM 生成有害内容的概率；Wen 等人[25] 进一步证明，监督微调和强化学习也会加剧模型生成有害输出的风险。Zhu 等人[26] 则通过提示工程（Prompt Engineering）使用 ChatGPT 将答案映射为二分类标签，用于仇恨检测任务。Roy 等人[27] 通过引入受害群体信息，有效提升了仇恨言论分类的准确率。得益于大规模参数和训练数据，大模型能够捕捉更复杂的上下文关系，提高了对文本的理解深度和分类准确性。然而，目前尚无系统性的大模型工程化方案被用于检测居高临下言论或其他歧视性文本。

1.2.4 多模态有害言论检测

近年来，伴随社交媒体平台内容形态的不断多样化，有害言论的表达方式已从传统的文本扩展至图像、视频、音频等多种模态，单一模态的检测方法逐渐暴露出覆盖面受限、误判率高等问题。为更全面地识别丰富的有害言论表达方式，研究者开始尝试将多模态建模引入有害言论检测任务，推动了跨模态语义理解与多通道特征融合的发展。

早期有害言论多模态研究以图文模因等静态内容为切入点。kiela 等人[28] 提出了结合视觉卷积特征与注意力机制的多模态架构，用于识别图文模因中的仇恨表达，强调视觉模态在仇恨识别中的作用。Thapa 等人[29] 在 CASE 2023 研讨会上组织了首个多模态仇恨事件检测共享任务，促进了标准化数据构建与评估方法的形成。次年，Thapa 等人[30] 在其组织的 CASE 2024 中进一步推出聚焦俄乌冲突语境的多模态仇恨事件检测挑战，提供了视频、图像、文本等复合信息的对齐数据集。Ganguly 等人[31] 提出了用于多模态的 XLM-RoBERTa-large 和 BERTweet-large 的联合仇恨检测，实现了文本与图像信息的深度融合，并在挑战赛中取得优异成绩。

随后，有害言论的研究进一步扩展至视频模态。2024 年，Das 等人[32] 率先提出了 HateMM 数据集，这是第一个专注于仇恨视频领域的高质量数据集。Maity 等人[33] 提出了 ToxVidLM 框架，这是一个面向社交平台短视频的多模态多任务学习系统，融合了视频帧、文本等模态，通过跨模态同步与联合训练机制提升了有害言论检测效果。

多模态有害言论检测从早期图文模因处理逐步发展至面向视频、语音的统一检测系统，并逐渐涵盖多语言、多文化场景。然而，关于隐性有害言论的多模态检测，尤其是居高临下言论的多模态工作尚处于起步阶段，由于隐性有害言论往往更依赖于非语言特征传达攻击意图，因此对于居高临下言论的多模态工作是当前亟需解决的重点方向，尤其是构建其多维数据和相关范式的工作。

1.3 本文工作

本文针对于居高临下言论面临的上述主要挑战提出了相应的解决办法，整体研究可以分为三个部分，分别是基于细粒度标注的居高临下分层语料框架、基于大模型的中英文居高临下言论双语检测、基于歧视性表情特征的多模态居高临下言论检测，以促进居高临下言论的系统性、多语言对比研究。以下将详细阐述这三部分研究内容。

第一部分是基于细粒度标注的居高临下分层语料框架。该部分主要解决了中文居高临下研究中长期面临的语料匮乏问题，以及英文数据注释存在的模糊性挑战。该部分提出了 CondescendCN 框架，系统构建了首个中文居高临下语料库—CCPC。该部分工作以新浪微博和知乎为数据来源，收集了共计 11k 条涵盖三级标注的居高临下样本，并引入“毒性强度”（Toxicity Strength，简称 TS）作为核心融合特征，旨在提升模型对细粒度隐性有害言论的识别能力。实验结果显示，TS 特征显著增强了模型的泛化性能，即便在迁移至 12 万条社交媒体数据后，其仍能维持稳定的检测效果。该研究不仅弥补了中文居高临下数据的空白，也为多种社交媒体场景下的后续研究提供了坚实的语料基础。

第二部分是基于大模型的中英双语居高临下言论双语检测研究。该部分主要解决了现有语言模型在隐性有害言论识别方面能力不足和跨语言分析不足的问题。该部分提出了 PclGPT，一种专门面向居高临下检测的双语种大型语言模型。传统预训练模型在面对如伪善、虚假同情等隐含有害特征时识别效果不佳，而大模型所具备的丰富语言知识则为提升辨别能力提供了可能。因此，本研究设计了一套基于大模型的训练方案，包括双段高效微调方法组合，以及预训练、监督微调的整体范式，并构建了相应的 Pcl-PT/SFT 数据集。在此基础上，开发了支持中英双语任务的 PclGPT-EN/CN 模型体系，用于提升模型对居高临下表达的理解与分类能力。实验结果显示，PclGPT 在不同弱势群体语境下的居高临下言论检测中表现优异，能更有效地区分多样化的隐性危害表达，从而显著增强整体的检测性能。该部分实验同时系统性的对于中英双语社交媒体平台歧视性言论的异同进行了对比研究，证明了不同社交媒体平台对应人群的居高临下差异性和共同研究可行性。

第三部分是基于歧视性表情特征的多模态居高临下言论检测。该部分主要解决了现有方法在单一文本模态下识别能力的局限性问题。该部分构建了首个面向居高临下

言论的多模态检测框架，包括 PCLMM 数据集与 MultiPCL 多模态检测器。由于居高临下常通过非语言线索（如面部表情、语音语调等）传递歧视意图，仅依赖文本无法充分建模其隐性特征。为此，该部分从 Bilibili 平台收集了 715 条高质量标注视频，并引入表情识别模块以捕捉关键面部线索，辅助隐性有害言论的判别。实验结果表明，融合视觉、语音与文本信息的多模态系统显著优于当前所有基线模型，在识别准确性和泛化能力方面均表现突出。该部分工作不仅为 PCL 检测引入多模态视角，也为视频审核、平台治理等实际居高临下的应用提供了重要技术支撑。

1.4 本文结构

本文主要面向社交媒体，针对于有害言论的重要领域-居高临下言论进行了一系列研究，并解决了目前居高临下研究所面临的一系列挑战。本文的总体框架如图 1.1 所示。具体而言，本文的论文结构如下：

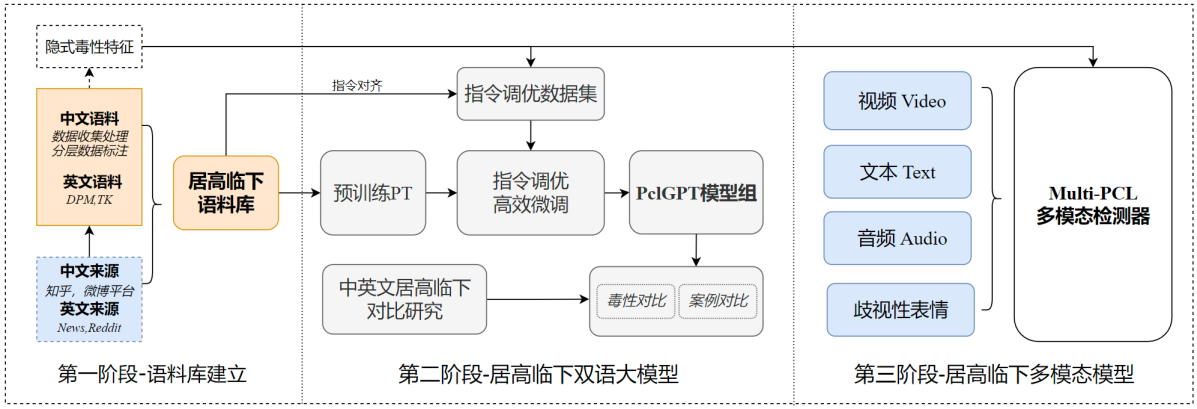

图 1.1　本文的三阶段“推进式”总体框架

Fig. 1.1　The ”progressive” overall framework of the three parts in this paper

第一章，绪论。对于居高临下言论的研究背景、意义以及国内外研究现状进行了详细的介绍，分析了目前居高临下言论研究所面临的挑战，同时阐述了本文的研究内容和整体结构安排。

第二章，相关理论与技术。主要介绍了与本文相关的各项理论与技术，包括神经网络模型、大型语言模型、高效微调与指令调优以及多模态技术，最后列举了本文所使用的部分公开数据集和主要评价指标。

第三章（图 1.1 第一阶段），基于细粒度标注的居高临下分层语料框架。针对于中英文居高临下言论语料资源匮乏，语义表示模糊等问题，本章提出了全新的居高临下言论细粒度分层框架，并提出了第一个中文领域的居高临下语料库，通过系列实验证明了该语料库的质量和本文的方法对于减少模糊分歧的有效性。

第四章（图 1.1 第二阶段），基于大模型的中英文居高临下言论双语检测。针对于传统模型对于居高临下群体检测存在显著误差这一问题，结合大模型时代的到来，本

章提出了有效利用和引导大模型的世界知识以强化居高临下言论的检测效果，包括联合高效微调机制的设计、模型预训练与指令微调的构造，并在四个数据集上证明了本章节所述方法的有效性。本章节在弥补中文领域研究空白的同时对中英文居高临下的异同和协同研究的必要性进行了详细的分析和阐述，实现了高效的面向多维社交媒体平台的居高临下检测研究。

第五章（图 1.1 第三阶段），基于歧视性表情特征的多模态居高临下言论检测。针对于传统的居高临下言论检测只局限于单一文本模态而忽略面部表情等潜在的非文本歧视性特征，本章提出了该领域第一个歧视性多模态数据集以及对应的多模态基线，证明了融合视频帧和面部表情特征的方法对居高临下言论的识别能力具有显著增益。

最后，对本文的贡献和创新点进行总结，并对居高临下言论的未来可能研究方向和交叉研究潜力进行了展望。

2 相关理论与技术

2.1 神经网络模型

2.1.1 注意力机制

注意力机制是序列建模中用于捕捉不同位置间依赖关系的核心模块，其核心思想是为输入序列中的各个位置分配动态权重，从而使模型聚焦于关键信息。该机制最初在神经机器翻译中得到广泛应用，其基本形式如下：

$$\text{Attention}(q,k,v)=\sum_{i=1}^{n}\alpha_i v_i,\quad \alpha_i=\frac{\exp(\text{score}(q,k_i))}{\sum_{j=1}^{n}\exp(\text{score}(q,k_j))} \tag{2.1}$$

其中 α_i 表示 query 与第 i 个 key 之间的相关性，通常通过打分函数（如点积、双线性函数等）计算。

根据设计目的与实现方式，注意力机制主要分为以下三类：

（a）硬注意力（Hard Attention）：通过对位置进行采样，仅选择一个位置参与表示计算。由于采样操作不可导，训练通常依赖强化学习方法进行优化。

（b）软注意力（Soft Attention）：采用 softmax 函数为所有位置分配连续权重，并对所有值向量加权求和，具有可导性。

（c）自注意力（Self-Attention）：一种特殊的软注意力形式，query、key 与 value 均来自同一序列，用于建模序列中各位置之间的全局依赖关系。

在 Transformer 中，自注意力机制被进一步优化为缩放点积注意力（Scaled Dot-Product Attention），具体表达式如下：

$$\text{Attention}(Q,K,V)=\text{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{2.2}$$

其中 Q,K,V 分别表示查询、键和值组成的矩阵，d_k 为键的维度。该结构显著提升了计算效率与训练稳定性，是现代预训练语言模型的关键模块。

2.1.2 Transformer

Transformer 架构由 Vaswani 等人[34] 于 2017 年提出，其核心创新在于完全摒弃传统的循环结构[35–37]，改以多层堆叠的自注意力机制与前馈网络，实现对输入序列的全局建模。

标准 Transformer 包含编码器（Encoder）与解码器（Decoder）两部分。编码器中的每一层由两个子模块组成：多头注意力机制（Multi-Head Attention）和前馈神经网

络（Feed-Forward Network）。多头注意力机制的计算形式如下：

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,\ldots,\text{head}_h)W^O \tag{2.3}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{2.4}$$

其中，W_i^Q、W_i^K、W_i^V 为第 i 个注意力头的投影参数，head_i 表示每个注意力头的输出。多个注意力头的输出通过拼接并乘以输出投影矩阵 W^O 得到最终结果。该机制允许模型在多个表示子空间中并行学习不同的注意力分布，从而提升建模能力与表达多样性。

由于注意力机制本身不具备位置信息建模能力，Transformer 引入位置编码（Positional Encoding）以显式注入顺序信息，其经典实现形式如下：

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right), \quad PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right) \tag{2.5}$$

其中，pos 表示当前词的位置，i 表示维度索引，d 是位置编码的维度。该设计使得不同位置的编码具有可区分性，且具有一定的相对位置信息表达能力，能够通过三角函数周期性变化为序列建模提供顺序感知能力。

此外，后续研究还提出了可学习的位置嵌入[38]，以增强任务适应性与建模灵活性。基于 Transformer 框架，衍生出一系列性能卓越的模型，包括 BERT、GPT、T5[39]、XLNet[40] 与 DeBERTa[41] 等。

2.1.3 BERT 模型

BERT 模型（Bidirectional Encoder Representations from Transformers）是 Devlin 等人[42] 于 2018 年提出的预训练语言模型，开创性地引入了深度双向 Transformer 结构，以捕捉更丰富的上下文语义信息。与传统基于左到右或右到左的单向语言模型不同，BERT 通过掩码语言建模（Masked Language Modeling，简称 MLM）实现了对句子中任意位置词语的预测，使模型能够同时考虑左侧与右侧的上下文，从而获得更强的语言理解能力。此外，BERT 还通过句子关系预测（Next Sentence Prediction，简称 NSP）建模句子之间的逻辑与语义联系，为诸如问答、自然语言推理等句子级任务提供了结构支持。

BERT 的模型结构完全基于 Transformer 编码器，输入由词向量、句段标记与位置编码三部分组成。在输入序列中，模型在最前添加 [CLS] 标记，用于捕获整句信息；当输入包含两个句子时，用 [SEP] 标记进行分隔。整个输入经过多层堆叠的 Transformer 编码器后生成每个词的上下文表示。其中，自注意力机制在不同位置间动态分配注意力权重，有效建模了全局依赖关系。

在训练过程中，BERT 采用大规模语料（如 BooksCorpus[43] 和英文 Wikipedia[44]）进行无监督预训练。MLM 任务随机遮盖输入文本中的部分词语，训练模型根据上下文进行复原；NSP 任务则通过句子对构建二分类任务，预测句子 B 是否为句子 A 的真实后续。预训练完成后，BERT 可通过少量微调样本快速适配至各种下游任务，如文本分类、命名实体识别和问答系统等，表现远超同期方法。

BERT 的提出不仅刷新了多项自然语言处理任务的性能指标，也为后续模型奠定了技术基础。其核心思想——预训练再微调的范式，成为一个时代深度学习语言模型发展的主流方向。通过双向建模和统一的 Transformer 架构，BERT 实现了对语义特征的高效提取，为构建通用语言理解系统提供了重要的技术支撑。

2.1.4 GPT 模型

GPT 系列模型（Generative Pre-trained Transformer）是 OpenAI 于 2018 年提出的一种基于 Transformer 解码器结构的自回归预训练语言模型[45]。不同于 BERT 使用的双向编码器结构，GPT 通过因果自注意力机制实现从左到右的单向语言建模，核心任务是预测下一个词。该设计使其在生成自然语言文本方面具备天然优势。GPT 模型采用“预训练-微调”范式，先在大规模未标注文本上进行语言建模训练，再在具体任务上进行微调，在多个 NLP 任务中展现出良好的性能和通用性。

GPT-2[46] 和 GPT-3[47] 分别是该系列的第二代和第三代模型，它们在模型规模和数据量上大幅提升，显著推动了大模型能力的边界。GPT-2 包含最多 15 亿参数，取消了任务微调环节，转而采用零样本、少样本提示方式完成下游任务，展现了语言模型的通用迁移能力。GPT-3 则将参数规模扩展至 1750 亿，并在 Common Crawl[48] 等海量语料上训练，进一步增强了少样本学习与多任务泛化能力。GPT-3 的推出标志着语言模型由“预训练 + 微调”正式过渡到“预训练 + 提示”的新范式，也验证了“大规模 + 简单目标函数”在语言建模中的巨大潜力。

GPT-3 在架构上并无本质创新，但其能力边界的拓展引发了大模型研究的爆发式增长。后续模型如 ChatGPT 以及国内的 ChatGLM[49]、通义千问（Qwen）[50] 等，均在相关的技术路线基础上发展而来。这些模型普遍采用更大的参数规模、更长的上下文窗口、更复杂的训练数据处理流程，并引入基于人类反馈的强化学习（RLHF[51]）等技术提升对齐能力。GPT-3 作为现代大模型的“起点”，不仅定义了大语言模型的基本范式，也奠定了生成式 AI 系统的技术基础，对整个自然语言处理与人工智能领域产生了深远影响。

2.2 大型语言模型

大型语言模型由 GPT 进一步发展而来，通常指基于 Transformer 架构、具备十亿级以上参数规模，并通过大规模语料学习语言规律的深度神经网络模型。自 GPT-3

展现出强大的零样本与少样本学习能力以来，LLM 迅速成为自然语言处理各类任务（如问答、摘要、代码生成、多轮对话等）的统一建模框架。当前主流 LLM 多采用 Decoder-only 架构，即基于自回归语言建模，仅通过上文预测下一个词，并以此训练出高度泛化的语言能力。在训练流程上，已形成“预训练—指令微调—对齐优化”三阶段范式：预训练使用数万亿 token 的高质量语料进行大规模自监督学习；指令微调阶段基于任务指令-响应样本强化模型的任务执行能力；最后的对齐阶段则使用人类反馈或偏好数据对生成质量进行优化。该范式已在 ChatGPT、Gemini[52] 等闭源系统中得到验证，也为 LLaMA、Qwen 等开源模型提供了训练参考。

2022 年底发布的 ChatGPT 被广泛认为是 LLM 实用化的转折点，其成功不仅源于语言模型本身能力的提升，更关键在于构建了一套有效的对齐机制，使模型行为更加可控、符合用户偏好。ChatGPT 基于前述 GPT 主干模型，首先进行指令微调，使其适应“问-答”结构的任务交互；随后引入人类反馈强化学习，构建由人类评分驱动的奖励模型，对输出进行优化，极大提升了模型的响应合理性与安全性。在此基础上，GPT-4[22] 继续拓展多模态输入能力，引入更长上下文窗口（最高可达 128k tokens）、函数调用等机制，显著增强了模型的任务管理与状态感知能力。同时，随着 GPT-4 的发布，RLHF 的高成本和不稳定性也引发了对于替代方案的探索，全新的技术手段如 DPO 算法（Direct Preference Optimization）[53] 通过直接拟合偏好排序，规避了强化学习中的高方差问题，成为当前对齐阶段的重要研究方向。在此背景下，国内的大型语言模型体系也迅速建立，并持续赶超国际水平。通义千问基于 Qwen 模型族构建了千亿参数规模的通用模型，并开放部分中小参数量版本如 Qwen-1.5B/7B/14B 支持研究使用；智谱 AI 的 ChatGLM 系列采用中英双语预训练，并通过多阶段指令优化强化中文交互能力；近期横空出世的 DeepSeek 大模型[54] 聚焦于多任务融合、长文本生成与 Agent 场景落地，更是在多项指标上达到全球一流水准。

尽管大型语言模型取得了突破性成果，但仍面临若干具体挑战。模型幻觉问题在复杂问答和事实性任务中仍频繁出现；对齐过程成本高、依赖人工标注，限制了高质量对齐的可扩展性；此外，当前 LLM 主要依赖英语语料预训练，在跨语言泛化（尤其是中文）上仍存在性能断层。本文的第四章节将基于大模型进行有害言论跨语言检测与多平台对比研究，针对上述问题进行系统性的工作。

2.3 指令微调与高效微调

2.3.1 指令微调技术

监督指令微调（Supervised Fine-Tuning，简称 SFT）[55] 是将预训练语言模型适配为任务执行系统的第一步，旨在使模型学会“根据明确人类指令完成指定任务”。与预训练阶段不同，SFT 通过有监督的数据对模型进行调整，使其能根据输入指令生成

对应的期望输出。具体而言，SFT 的数据形式通常为一组指令-响应对 (x_i, y_i)，其中 x_i 为自然语言形式的任务描述，y_i 为期望输出。训练时，将指令和目标输出拼接为完整序列，模型在此基础上学习任务执行能力，重点在于从“语言建模”过渡到“指令理解与响应控制”。

在实际应用中，指令微调的数据构建不仅要涵盖问答、翻译等各类主流任务，还需覆盖非结构化指令（如“写一段鼓励的话”）、主观指令（如“评价下面这部电影”）以及边界模糊任务（如“帮我润色这段话”）。这些任务往往缺乏统一的输出格式，具有高度的不确定性，对模型的指令理解和生成控制能力提出更高要求。为提升覆盖面和泛化能力，现有方法通常结合人工构建与自动生成两种策略：一方面，人工设计典型任务指令与目标输出，保证基本语义清晰与结构合理；另一方面，采用现有预训练语言模型自动生成更多任务-响应样本，即“自监督指令扩展”，其中较为代表的方法是 Self-Instruct[56]。该方法以少量人工种子指令为起点，驱动 LLM 生成新任务类型，并筛选生成样本用于再次训练，从而提升任务分布的多样性与指令风格的泛化能力。

为了增强模型应对“非标准指令”的能力，Self-Instruct 在任务生成过程中引入三类机制：

（1）结构多样化：将输出格式从纯文本扩展到有序列表、JSON、嵌套结构等，训练模型识别并控制输出结构；

（2）风格扰动：调整指令表达方式，使同一任务对应多种提问形式，从而提升模型对语义等价变体的理解能力；

（3）任务目标转变：将任务从“问答”改写为“解释、总结、分类、转述”等，模拟真实指令不确定性。这类策略可系统性地拓展模型在复杂任务分布下的鲁棒性，而不是局限在理想指令环境中。

训练时，SFT 需特别处理指令与目标输出之间的边界问题。为防止模型在非监督区域学习错误行为，通常在输入格式中显式标注指令与响应段落，如使用 `<|user|>`、`<|assistant|>` 或 `[Instruction]/[Response]` 等 token 进行分隔。在多轮对话任务中，为确保梯度仅在当前响应上回传，需构造对应的 label mask，对历史对话及指令部分屏蔽损失计算。例如，对于拼接后的序列 $s = [\text{prompt}, \text{response}]$，配套的 mask 向量 $m = [0, 0, \ldots, 1, \ldots, 1]$ 用于控制目标区域，使优化聚焦于本轮响应。该机制在指令重写、角色扮演、多轮问答等任务中是确保模型行为准确的重要手段。

2.3.2 高效微调技术

在工程实现上，由于计算资源和算力的限制，指令微调通常采用参数高效微调方法对大模型进行训练，避免全参数微调更新带来的计算资源压力，同时，使用高效微调技术能够在不多使用算力的情况下接近甚至达到全参数微调的效果，具有很高的性价比。在这些技术中，Prefix-Tuning[57] 和 LoRA[58] 是两种主流的高效微调策略。

（1）Prefix-Tuning 机制

Prefix-Tuning 通过向 Transformer 注意力层的 Key 和 Value 矩阵注入可训练的前缀向量，实现对注意力计算的引导。考虑 Transformer 的输入为 $\boldsymbol{H} \in \mathbb{R}^{n \times d_h}$，其中 n 为输入 token 序列长度，d_h 为隐藏状态维度。

首先，标准的多头注意力机制中的 Q、K、V 计算为：

$$\boldsymbol{Q} = \boldsymbol{H}\boldsymbol{W}_q, \quad \boldsymbol{K} = \boldsymbol{H}\boldsymbol{W}_k, \quad \boldsymbol{V} = \boldsymbol{H}\boldsymbol{W}_v \tag{2.6}$$

其中，$\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v \in \mathbb{R}^{d_h \times d_h}$ 为原始可训练权重。

Prefix-Tuning 在此基础上，针对每一层 Attention 模块注入前缀向量：

$$\boldsymbol{P_k}, \boldsymbol{P_v} \in \mathbb{R}^{l \times d_h} \tag{2.7}$$

其中，l 为前缀 token 的长度（远小于 n）。

通过拼接操作，扩展后的 $\boldsymbol{K}$ 与 $\boldsymbol{V}$ 矩阵为：

$$\boldsymbol{K'} = [\boldsymbol{P_k}; \boldsymbol{K}] \in \mathbb{R}^{(l+n) \times d_h}, \quad \boldsymbol{V'} = [\boldsymbol{P_v}; \boldsymbol{V}] \in \mathbb{R}^{(l+n) \times d_h} \tag{2.8}$$

最终注意力计算公式为：

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K'}, \boldsymbol{V'}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K'}^{\top}}{\sqrt{d_h}}\right)\boldsymbol{V'} \tag{2.9}$$

Prefix-Tuning 无需更新 Transformer 中的主干参数（$\boldsymbol{W}_q, \boldsymbol{W}_k, \boldsymbol{W}_v$），仅对 $\boldsymbol{P_k}, \boldsymbol{P_v}$ 进行优化，因而极大降低了参数量。

（2）LoRA 机制

LoRA 采用低秩分解（Low-Rank Decomposition）策略，将 Transformer 中的投影矩阵分解为两组可学习的低秩矩阵，减少了微调过程中的有效参数量。

LoRA 应用于 $\boldsymbol{W}_q$ 与 $\boldsymbol{W}_k$，重参数化形式为：

$$\boldsymbol{W}_q^{\text{LoRA}} = \boldsymbol{W}_q + \boldsymbol{B}_q\boldsymbol{A}_q, \quad \boldsymbol{W}_k^{\text{LoRA}} = \boldsymbol{W}_k + \boldsymbol{B}_k\boldsymbol{A}_k \tag{2.10}$$

其中，$\boldsymbol{A}_q, \boldsymbol{A}_k \in \mathbb{R}^{r \times d_h}$，$\boldsymbol{B}_q, \boldsymbol{B}_k \in \mathbb{R}^{d_h \times r}$，$r$ 为低秩维度，通常满足 $r \ll d_h$。

实际计算时，查询与键矩阵被替换为：

$$\boldsymbol{Q} = \boldsymbol{H}(\boldsymbol{W}_q + \boldsymbol{B}_q\boldsymbol{A}_q), \quad \boldsymbol{K} = \boldsymbol{H}(\boldsymbol{W}_k + \boldsymbol{B}_k\boldsymbol{A}_k) \tag{2.11}$$

LoRA 冻结了 $\boldsymbol{W}_q$ 与 $\boldsymbol{W}_k$，仅训练新增的低秩矩阵 $\boldsymbol{A}$ 与 $\boldsymbol{B}$，进而极大的减少了训练成本。

2.4 多模态技术

多模态技术旨在联合建模语言与图像、语音等感知模态，以弥补语言单模态在语义理解和现实场景建模中的不足。传统架构多采用双编码结构，即分别使用 Transformer 或 ViT[59] 对文本和图像进行表征提取，再在中间层通过 Cross-Attention 机制[60]实现模态融合。该机制使语言向量可动态聚焦于图像区域，增强语义对应关系。部分工作进一步引入门控或共注意力模块，在模态间建立更强的交互通道，实现信息流的动态调节。这类架构既保留模态独立性，便于模块化训练，又具备一定的跨模态理解能力，在图文匹配、VQA 等任务中被广泛采用。

模态对齐是多模态建模中的关键难点，尤其体现在语言与图像之间表征粒度的不一致。语言通常以词或短语为单位，具有清晰的语义边界，而图像则以视觉 patch 或区域表示为主，缺乏与语言单元的一一对应关系。为缓解这种跨模态语义错位，BLIP-2[61]引入跨模态对比学习，通过正负样本构造和对比损失，在共享嵌入空间中拉近语义匹配的图文对、拉远不匹配对，从而实现弱监督的模态对齐。该模型采用冻结的 ViT-G 编码器和预训练语言模型，利用 Querying Transformer 完成模态间的桥接，仅训练轻量模块即可获得较强的对齐能力，广泛应用于图文检索和多模态指令理解等任务。

另一类对齐路径则侧重于结构统一，通过共享 Transformer 架构实现模态间的隐式融合。Wang 等人[62] 将图像 patch 与文本 token 并列输入编码器-解码器框架，依靠多层自注意力机制自动建模图文间的语义对应，无需显式标注区域与词语的对齐关系。该方法以统一的生成式目标为驱动，同时覆盖图像描述、视觉问答、图文补全等任务，有效引导模型在跨模态语境中学习稳定的对齐策略，具备较强的迁移与泛化能力。

2.5 本文使用的数据集（公开数据集）

我们在这里列举了本文中使用的两个著名公开数据集，由于本文工作的一个重点方向是对于居高临下语言学定义和数据空白的补足，因此自建数据集的部分将放到后续章节详细阐明。

2.5.1 Don’t Patronize Me!

Don’t Patronize Me! (DPM) 数据集[1] 包含来自 News on the Web (NoW) 的 10,469 条涉及潜在弱势群体的英文段落。该数据集采用层次化标注，标签为 0 到 4，其中 0 表示非居高临下标签，1-3 分别表示居高临下的毒性强度从低到高。同时该数据集在

子类别的归属中使用多标签分类将每一条社区评论映射为一至多个子类别。DPM 数据集的使用主要集中于本文第四章的指令微调阶段，我们使用其社区文本及其对应的强度标签来引导大模型学习相应的范式。

2.5.2 TalkDown

TalkDown (TD)[19] 是一个包含 74,000 对英文评论/回复对的 Reddit 社区数据集，数据来源于 2006 至 2018 年的弱势群体社区的讨论。每对样本被标注为 PCL、非 PCL 或不确定。在本文第四章的指令微调工作中，我们拼接评论与回复，手动筛选部分数据作为训练集，并利用攻击性词典去除偏激样本，以符合 PCL 的低攻击性特征。同时，为保证公平性，超出模型输入长度的长文本将被过滤。

2.6 本文主要评价指标

为评估多分类模型的性能，本文采用准确率（Accuracy）、精确率（Precision）、召回率（Recall）以及宏平均 F1 值（Macro-F1）作为主要指标[63]。

准确率表示预测正确的样本数在总样本中的占比，定义如下：

$$\text{Accuracy} = \frac{\sum_{i=1}^{K} TP_i}{\sum_{i=1}^{K} (TP_i + FP_i + FN_i + TN_i)} \tag{2.12}$$

精确率衡量预测为某类的样本中，实际为该类的比例：

$$\text{Precision} = \frac{1}{K} \sum_{i=1}^{K} \frac{TP_i}{TP_i + FP_i} \tag{2.13}$$

召回率表示实际属于某类的样本中，被正确预测的比例：

$$\text{Recall} = \frac{1}{K} \sum_{i=1}^{K} \frac{TP_i}{TP_i + FN_i} \tag{2.14}$$

宏平均 F1 值是精确率与召回率的调和平均数，计算公式如下：

$$\text{Macro-F1} = \frac{1}{K} \sum_{i=1}^{K} \frac{2 \cdot P_i \cdot R_i}{P_i + R_i} \tag{2.15}$$

其中 P_i 和 R_i 分别为第 i 类的精确率与召回率。

3 基于细粒度标注的居高临下分层语料框架

3.1 引言

当一个实体的语言使用展现出对他人的优越态度或以同情的方式描述他人时，我们称之为居高临下的语言（Patronizing and Condescending Language，PCL）。居高临下言论的发出者的意图不一定总是有害的，甚至有时是为了帮助弱势群体（例如呼吁社会关注这些群体和筹措资金的慈善活动）。然而，这种优越态度和怜悯的话语可能会使歧视行为隐藏在看似积极的活动中，进而使其更难被察觉。我们列举了居高临下工作中的典型中英双语样例各一例，如表 3.1 所示，无论是中文还是英文，表中列举的文本都没有针对于弱势群体的明确攻击性词汇和表述，但是从结果而言都造成了隐性的歧视和伤害。由于攻击性较弱，因此使用常规的预训练语言模型如 LSTM、BERT 甚至当前的开源 LLMs 都不能进行有效检测，这种模糊的语言学表述是居高临下领域最为明显的特征之一，也是有害言论检测领域的巨大挑战之一。

表 3.1 居高临下言论检测实例
Tab. 3.1 PCL Detection Example

英文数据	类别	LSTM	BERT	GPT-4o
These poor children! It's truly admirable how they keep striving despite their humble beginnings.	Unbalanced-Power-Relations	✗	✗	✔
中文数据	类别	LSTM	BERT	GPT-4o
单亲家庭的孩子确实没有人爱，真是可怜。*Children from single-parent families are indeed unloved, how pitiful.*	Compassion	✗	✗	✗

尽管近年来已有大量关于明确攻击性言论的研究，例如仇恨言论，但对于居高临下的定义和建模仍是有害言论检测领域中的新兴研究方向, 特别是基础语料和细分语种仍存在滞后性。由于居高临下言论的检测往往需要特定的知识或熟悉某些文化典故，因此推动该领域进展的关键因素之一是需要专家标注的高质量数据集，以应对这些隐晦且具有危害性的特征。在英文领域，Wang 等人[19] 提出了 Talk Down 数据集，专注于社交媒体中的居高临下语言；Pérez-Almendros 等人[1] 提出了 Don't Patronize Me! 数据集，关注新闻报道中对弱势群体的描述。然而，据我们所知，居高临下的相关研究几乎集中于英文领域，而在中文领域的研究仍然处于起步阶段，无论是中文领域的语言学定义，还是数据集和借助相关语料训练的基线都存在空白；此外，针对中文社交媒体弱势群体的特殊性尚未受到充分关注。这些都成为阻碍进一步研究的瓶颈。

在针对于有害言论的数据构建任务中，目前最有效的方法是 Zampieri 等人[4] 提出的 OLID (Offensive Language Identification Dataset) 分层注释框架，其专为 SemEval-2019 Task 6 设计。OLID 数据集采用了三层分级结构，将攻击性语言检测任务逐步细化，细粒度的定义不同层次的语言学危害特征。

Level A - Offensive Language Detection：判定评论是否包含攻击性语言（Offensive vs. Not Offensive）。

Level B - Categorization of Offensive Language：对于被判定为攻击性的样本，进一步划分为针对性攻击（Targeted Insult）与非针对性攻击（Untargeted），前者通常指向特定对象，后者则表现为一般性冒犯或粗鲁用语。

Level C - Offensive Language Target Identification：对于有针对性的攻击，进一步标注攻击的具体目标，包括个人、群体和其他类别，提升模型对攻击目标类型的识别能力。

虽然这一框架能有效的解决仇恨言论的分析和标注模糊性问题，但目前尚未有效的应用于居高临下言论等隐性有害言论的检测工作中，同时居高临下言论和仇恨言论在语义学特征上的明显区别也需要更有针对性的框架区分。为填补这些研究空白，本章节首先提出了 CondescendCN 框架，这是中国互联网首个居高临下语言框架。与传统的单一分类方法相比，该框架具有更加细粒度的分布，采用了三级层次结构：（a）是否有害，（b）有害言论类型，（c）居高临下毒性强度（TS 等级）、居高临下子类别以及目标群体。基于该框架，我们构建了中文领域第一个针对于隐性有害言论的语料库——CCPC，包含来自中国两大主流社交媒体：新浪微博和知乎的超过 1.1 万条评论。质量测试方面，本章节设计了泛化性实验，将训练后的模型迁移至微博平台，应用于事件检测与群体识别任务。最后，我们对 CCPC 进行了详细的统计学分析。结果表明，居高临下言论更倾向于针对女性和儿童群体，迫切需要社会的更多关注。综上，本章的主要贡献为：

（1）提出了首个中文领域的居高临下层次框架 CondescendCN。该框架首次将中文隐性有害言论检测划分为多个层级。这种细粒度的判别方式能更高效的捕捉到中文隐性有害语言的语义复杂性。

（2）构建了 CCPC。这是首个中文居高临下语料库。CCPC 语料库包含了来自新浪微博和知乎的 1.1 万条高质量的社交媒体评论的结构化标注数据，涵盖了多种弱势群体（如残疾人、女性、老年人、儿童等）。该语料库的构建为中文居高临下检测的研究提供了足量的基准数据，填补了中文领域在这一研究方向上的空白。

（3）构建并验证了毒性强度（TS）融合框架对减少居高临下判别主观分歧的作用，并在更广泛的迁移测试中验证了数据集的泛化能力。

（4）揭示了居高临下的主观模糊性并强调了背景知识的重要性。例如难以仅从简

短文本中判断说话者的身份和阶级，以及需要区分伪装成关心的虚假同情与真正的同情。本章节指出，后续工作需要进一步扩展背景知识与对居高临下上下文信息的利用。

3.2 CondescendCN 框架

由于居高临下文本的隐性特征，依靠单一视角和粗粒度特征难以准确捕捉其真实语义。对自由文本的系统性解读将有助于提升机器预测的可信度，这对于缓解注释设计与构建过程中的偏差具有重要意义。基于此，本章节首先提出了用于细粒度居高临下注释的分层框架。在多轮试验中，我们通过小规模标注及边缘案例测试，不断完善该注释框架及其标注指南。CondescendCN 框架细节如图 3.1 所示，具体结构如下：

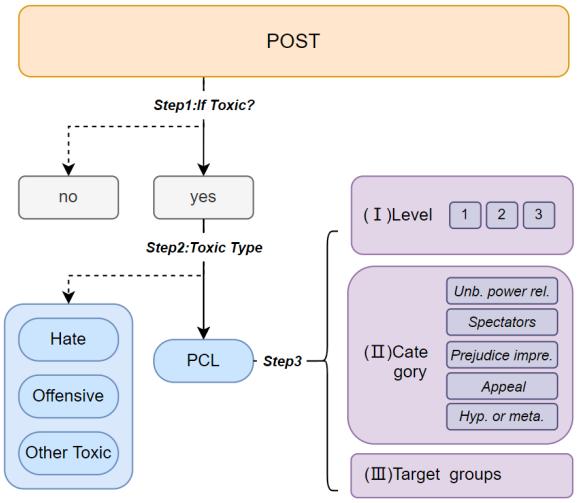

图 3.1 如何识别居高临下言论及其类型？CondescendCN 框架示例

Fig. 3.1 How to identify PCL and its type? Here is an example of our CondescendCN framework

3.2.1 有害判定

结合图 3.1，我们首先判断给定文本（POST）是否为有害言论。有害言论通常被定义为粗鲁且令人不适的言论。第一阶段有害言论的判定将作为后续标注判别的基础，这被视为一个二分类任务（有害或无害）。

3.2.2 有害类别

第二步是进一步确定有害言论的类型（仇恨言论、攻击性语言、居高临下言论）。居高临下言论由于其隐晦的有害表达，常常被用于看似正面的言论场景中，其展示的攻击性较弱。在本章中，我们将无害语言、仇恨言论、攻击性语言等统一归类为非居高临下言论，其余则归类为居高临下言论。

3.2.3 居高临下毒性强度

在本章节中，我们创新性地将毒性强度 TS 作为细粒度居高临下言论的重要分类维度，这有助于我们更好地理解有害信息，并在细粒度层面上减少对边缘案例的主观误判[64]。我们将居高临下言论的毒性强度 TS 按照语义强弱依次划分为三个等级：

（1）弱：通常以“伪善”的方式对弱势群体进行鼓励，例如称赞弱势群体中值得学习的品质。语气最为温和，但隐含歧视性态度。

（2）中等：尽管发言者与弱势群体之间保持一定的阶层距离，但语气中会流露出更多的鄙视态度与虚伪，并以“客观”视角向弱势群体提供建议，这些浅层意见往往无法改善弱势群体的现状。

（3）强：发言者与弱势群体之间阶层差距明显，语气更为尖锐，通常伴随明显的歧视性语言、优越态度以及讽刺性的表达。

3.2.4 居高临下子类别

最后，借鉴 Pérez-Almendros 等人[1] 的研究成果，并结合中文居高临下语言的独特特征，我们提出了针对中文领域的居高临下言论的详细语言学分类：

不平等的权力关系（Unbalanced Power Relations，简称 unb）：与弱势群体保持阶层和权力上的距离，以“救世主”的姿态宣称帮助他们摆脱困境。

旁观者（Spectators，简称 spe）：作为看客，缺乏深入思考，凭一时兴起提出浅层意见，无法从根本上解决问题。

偏见印象（Prejudice，简称 pre）：在给予帮助或建议时，带有对弱势群体的刻板印象或歧视性看法，但作者的优越感被“友善”或“同情”所掩盖，这种刻板印象在表面上并不明显。

呼吁（Appeal，简称 appe）：代表专家或倡导者的声音，呼吁弱势群体应当改变自身现状。

引起同情（Elicit Sympathy，简称 es）：作者直接表达对弱势群体的怜悯与关切，或通过隐喻等手法，将弱势群体描述为“需要帮助的人”，引发读者的同情。

3.2.5 居高临下群体检测

居高临下言论主要针对弱势群体。在语料收集阶段，我们对所采集评论中涉及的弱势群体进行分类，以便后续研究。弱势群体包括残障人士、女性、老年人、儿童、普通大众、单亲家庭以及其他弱势群体（如工人阶级、农民阶级等）。在后续多模态的研究工作中，这些类别将被进一步归纳。

3.3 CCPC 语料库

3.3.1 数据收集

本章从中国两大主流社交媒体平台：知乎和新浪微博中收集了热门帖文下的评论，作为我们的数据来源。我们将数据收集的范围限定在与居高临下相关的热门话题和事件中涉及的七类主要弱势群体中。随后，针对每类群体，通过人工统计整理了一组频度最高的关键词，并在该关键词列表的范围内进行检索。

为确保语料库的质量，我们剔除了评论数少于 20 条的帖子，并进行了额外的人工筛选。最终，我们从包含这些关键词的 1082 条热门博文中共收集了超过 1.4 万条评论。我们删除了重复和无关的样本（包括微博中的常见固定标签，如“回复”和“转发微博”），以及少于 5 个汉字或完全不含中文字符的样本。同时，我们保留了样本中的表情符号，并将其转换为平台上规定的对应中文文本对，以尽可能保留表情中所映射的情绪语义信息。最终，CCPC 数据集保留了 11397 条评论。数据集示例如表 3.2 所示。

表 3.2　CCPC 语料库示例
Tab. 3.2　A sample description for the CCPC corpus

Exp.	评论	是否有害	有害类别	毒性强度	PCL 类别
1	现在离婚率本来就很高很高了。	no	-	-	-
2	我接触过的单亲家庭出来的孩子，性格多少都不太好，孩子也是受害者啊。	yes	PCL	1	pre, es
3	残疾人就业是个严重的问题，应该给他们更多的岗位。	yes	PCL	2	unb, appe
4	与其担心别的，农民工朋友倒不如想想如果自己被欠薪了，应该如何合法讨薪才对。	yes	PCL	3	unb
5	笑拉了，我就是瞧不起你们小仙女怎么了。	yes	HATE	-	-

数据集的详细统计信息如表 3.3 所示。本章将会在 3.3.2 节中进一步阐述我们的标注工作。统计结果表明，在微博平台上，居高临下言论的比例显著高于知乎平台。此外，在两个平台上，女性和儿童群体遭受居高临下言论的比例均较高，且知乎平台对儿童的攻击率最高，微博平台则对女性的攻击率最高。这一发现为后续聚焦于特定弱势群体提供了数据支撑。

我们在数据收集过程中认真考虑了相关的伦理和道德问题。为保护用户隐私，所有用户名均被替换为统一的“用户”占位符。我们确保 CCPC 语料库仅用于本领域的学术研究，并将全部采集数据向公众开放。

表 3.3　来自不同平台的 CCPC 语料库统计结果
Tab. 3.3　Statistical Results of the CCPC corpus from different platforms

	残障人士	女性	老年人	儿童	普通大众	单亲家庭	其他	总计
zhihu	838	735	656	858	922	628	815	5452
zhihu$_p$	66	110	72	177	72	87	123	700
prop.(%)	7.9	15.0	11.0	**20.6**	7.8	13.8	15.1	12.8
weibo	920	760	747	950	864	754	950	5945
weibo$_p$	167	263	142	323	78	247	226	1446
prop.(%)	18.2	**34.6**	19.0	34.1	9.0	32.8	23.8	**24.3**
Total	1758	1495	1403	1808	1786	1382	1765	11397

3.3.2　数据标注

（1）标注过程

基于我们提出的框架，我们对数据集进行了如下标注：将无害言论、仇恨言论与攻击性言论统一标注为非居高临下标签 N-PCL（又称标签 0）；对于居高临下标签 PCL（又称标签 1-3），我们引入了毒性强度，基于当前 PCL 标签的居高临下毒性强度将其细分为 1 至 3 级，毒性强度的定义如章节 3.2.3 所述。

每条评论由两名标注员独立标注，第三名标注员负责校对，以提升边界样本的标注质量，特别是针对更细粒度的边缘样本评估[1]。本章节提出了毒性强度融合（Toxic Strength Fusion，简称 TS）方法，整合两位标注员的结果，融合的过程如图 3.2 所示。在得到融合标签后，我们基于评分函数对最终标签进行区间函数映射，得到我们的最终标签结果。

毒性强度融合过程的评分函数定义如下：

$$S = \text{ANN1} + \text{ANN2}, \quad S \in \{0, 1, 2, 3, 4, 5, 6\} \tag{3.1}$$

最终标签通过以下区间函数映射：

$$F(S) = \begin{cases} 0, & S \in \{0, 1, 2\} \quad \text{(N-PCL)} \\ 1, & S \in \{3, 4\} \quad \text{(P-Weak)} \\ 2, & S = 5 \quad \text{(P-Middle)} \\ 3, & S = 6 \quad \text{(P-Strong)} \end{cases} \tag{3.2}$$

最终标签被划分为四大类：非居高临下 N-PCL、弱居高临下 PCL-Weak、中居高临下 PCL-Middle 和强居高临下 PCL-Strong，图 3.2 展示了标注判断的示例。我们执行了两

轮类似的标注任务，并通过投票机制获得最终数据集。值得注意的是，部分仇恨与攻击性文本可能会干扰居高临下毒性强度的判定，因此我们在汇总过程中进行了人工校对。

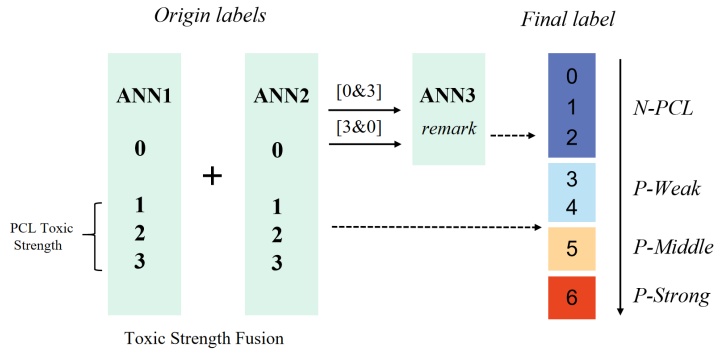

图 3.2　居高临下语言的毒性强度融合 TS 示意图，用于生成最终标签

Fig. 3.2　Toxic Strength Fusion of PCL to produce final label

（2）注释人员基本情况

为了保证数据集的质量，我们首先确保注释人员的多样性，我们招募了六位注释人员（四名主要标注员，两名校对员），他们在性别、年龄和教育背景上各不相同（男女比例各占 50%；年龄在 25±5 岁；学历包括三名硕士、两名博士和一名本科生）。

（3）注释一致性

我们计算了二分类任务和多类别标注任务中的 Kappa 注释者间一致性（IAA）[65]。当我们剔除所有被至少一名注释人员标注为边界性类别（如 Label1/Label2）的评论时，IAA 指标有所提升。详细信息请见表 3.4 所示。

表 3.4　标注一致性（IAA）测试结果

Tab. 3.4　Inter-annotator agreement (IAA) results

标签比例	IAA	PCL 子类别	IAA
All labels	0.62	Unb. power relations	0.65
		Spectators	0.42
Remove label1	0.64	Pre. impre.	0.59
		Appeal	0.48
Remove label1,2	0.69	Hyp. or meta.	0.71
		Others	0.66

3.3.3　数据统计分析

在数据收集与标注完成后，我们的 CCPC 语料库共包含 11397 条评论，其中包括 9251 条负样本（N-PCL）和 2146 条正样本（居高临下标签根据 TS 强度被划分为

P-Week、P-Middle 和 P-Strong）。这些数据覆盖了中文主流论坛中的弱势群体。具体的毒性分布情况如表 3.5（左）所示，表 3.5（右）展示了七类弱势群体中三种毒性强度的居高临下标签分布情况。

表 3.5　CCPC 语料库基本统计与不同弱势群体下的毒性强度分布，右表为毒性强度分布比例（%）
Tab. 3.5　Basic statistics of the CCPC corpus and toxicity strength distribution across vulnerable groups. The right table shows the toxicity strength distribution (%)

毒性类别	数量
非居高临下	9251
弱毒性	1167
中等毒性	439
强毒性	540
Total	11397

	弱毒性	中等毒性	强毒性
Disabled	61.8	16.4	21.8
Women	29.7	29.7	40.6
Elderly	54.6	16.9	28.5
Children	63.7	17.7	18.6
Commons	59.8	17.1	23.1
Single.	62.2	14.7	23.1
Disadv.	57.6	19.7	22.7

3.4　实验结果与分析

3.4.1　基线设计和结果测试

这里我们展示了本章节使用的主要基线模型。数据集按照 8:1:1 的比例划分训练集、验证集、测试集。本章节将训练 epoch 设置为 15，batch size 设置为 32，并使用相同的随机种子，同时使用了 Precision、Recall 和 F1 值作为评估指标。实验结果展示在表 3.6 和表 3.7 中。

BERT：我们基于 BERT 及其相关变体进行了预训练语言模型（PLMs）实验。我们使用了 BERT、BERT_M 和 BERT_C 对 CCPC 语料库进行实验。这些 PLMs 被用作编码器，在居高临下任务中采用全连接层作为分类器。我们分别对原始标签和通过毒性强度融合 TS 方法获得的标签进行了评估。

BiLSTM：我们使用双向 LSTM 对单词级别的 glove 词向量进行建模。LSTM 层和分类器层的 dropout 比例均为 0.5%。

我们首先进行居高临下言论的二分类检测，如表 3.6 所示。结果表明，包含 TS 方法的模型普遍获得了更好的 F1 结果。具体来看，$\text{BERT}_\text{C} \wedge \text{TS}$ 模型在各项指标中表现最佳，F1 值达到 0.714，显示出中文 BERT 模型在此任务中的优越性。同时，BiLSTM 依然具有一定的竞争力，F1 值为 0.656，说明在特定场景尤其是短语言文本的条件下 LSTM 依然有自身的序列化优势。

随后，我们将居高临下分类任务细化为为句子级的多标签分类问题，对每段文本分配一个或多个居高临下类别标签，如表 3.7 所示。结果显示，BERT_C 模型在大部分

表 3.6　PCL 二分类检测中的实验结果
Tab. 3.6　Experimental results of PCL binary classification

	融合		Precision	Recall	F1
BERT_C	∧	TS	**0.709**	**0.719**	**0.714**
BERT_C			0.682	0.693	0.687
BERT_M	∧	TS	0.653	0.646	0.649
BERT_M			0.637	0.659	0.643
BERT	∧	TS	0.579	0.590	0.584
BERT			0.586	0.600	0.589
BiLSTM			0.677	0.645	0.656

类别上的表现优于其他模型，特别是在不平衡的权力关系（unb）和旁观者（spe）类别上，分别获得了 95.80 和 71.19 的 F1 值，但由于隐式特征的存在，在部分子类别上仍难以获得有效的检测效果（如呼吁和同情类别），这些模糊子类别也是居高临下言论检测效果不佳的关键症结所在。

表 3.7　PCL 多标签分类实验结果。该任务被视为多标签分类任务
Tab. 3.7　Results of categorizing PCL, which is regarded as a multi-label classification task

	BERT			BERT_M			BERT_C		
(%)	P	R	F1	P	R	F1	P	R	F1
unb	86.81	98.75	92.40	95.12	92.34	93.71	97.21	94.43	**95.80**
spe	33.33	22.58	26.92	55.56	64.52	59.70	75.01	67.74	**71.19**
pre	63.49	78.43	70.18	72.73	62.75	67.37	73.08	74.51	**73.79**
appe	21.35	23.10	22.18	22.12	24.35	**23.18**	22.22	21.98	22.10
es	17.14	31.58	22.22	34.48	52.63	41.67	38.46	54.63	**45.11**

3.4.2　迁移性测试

本章节希望使用 CCPC 所训练的基线模型能够在中国各大主流舆论平台上有效应用于居高临下言论检测，这依赖于模型在大规模外部数据中对居高临下言论的准确识别能力。我们注意到，微博作为一个注重传统文本领域的优秀中文媒体平台，被广泛用于评估现有模型的可迁移性，其用户主要通过 # 关键词参与社区互动。从常识来看，弱势群体社区的居高临下比例一定高于中立和非弱势群体社区。因此，我们通过观察经 CCPC 训练后的模型是否在微博等平台具有对于不同社区的居高临下言论的不同比例的准确判定能力，并判断模型在基于常识的识别任务中的表现，对我们模型的泛化性能进行了验证。

在迁移性测试中，我们选取了三个面向弱势群体的不同类别社区：群体 A 为弱势群体社区。该社区被广泛认为具有较高的居高临下言论比例，例如 # 妇女（Women）、

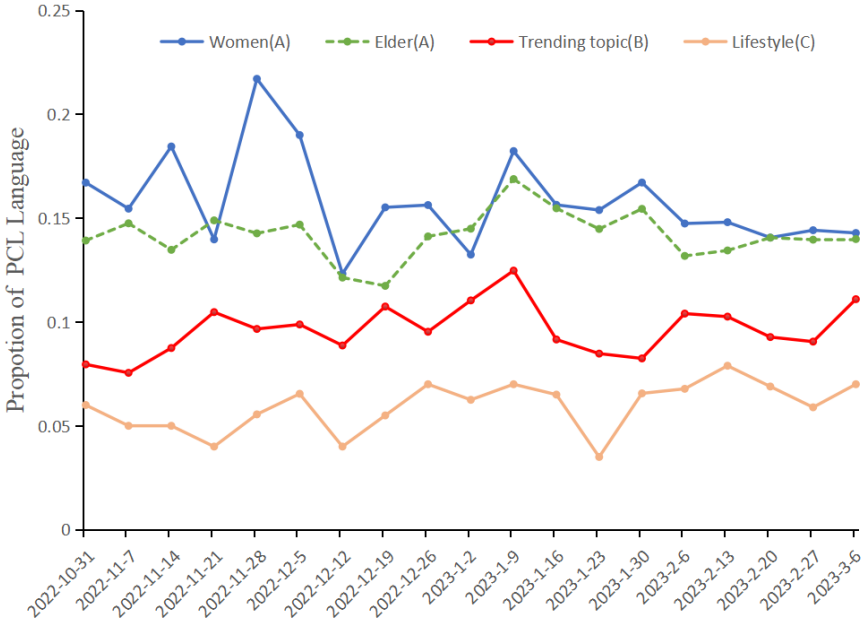

图 3.3　不同弱势群体社区的居高临下言论比例。A 类社区为妇女（蓝色）和 老人（绿色）；B 类为 微博热搜（红色）；C 类为 生活方式（橙色）

Fig. 3.3　Condescension rates for different disadvantaged communities. Community A contains *Women* (Blue) and *Elder* (Green); B contains *Trending topics* (Red); C contains *Lifestyle* (Orange)

老人（Elder）；群体 B 为综合性社区，其居高临下言论比例处于正常水平，例如 # 微博热搜（Trending topics）；群体 C 为非弱势群体社区，这些社区中的居高临下言论发生率较低，例如 # 生活方式（Lifestyle）、# 娱乐（Entertainment）。在 2022 年 11 月至 2023 年 3 月期间，我们按周从上述社区中获取了超过 12 万条评论。三类数据的对比与验证结果如图 3.3 所示。我们观察到，# 妇女和 # 老人社区中的居高临下言论比例较高，而 # 生活方式中的比例最低。显然，模型更倾向于在弱势群体社区中识别居高临下言论。此外，我们还可推断，作为弱势群体的女性群体，其所遭受的居高临下言论比例远高于男性群体。这些定性分析进一步佐证了我们工作的有效性。

3.4.3　居高临下的模糊性

由于居高临下言论的定义较为宽泛，其判断过程存在较大的主观性。首先，需要明确居高临下话语中是否存在清晰的阶层划分，这通常依赖于更丰富的上下文知识。仅凭简短的文本句子，难以准确识别说话者的身份和所属阶层。例如，**“我认为仅有最低生活保障是不够的”**，该评论无法确定“我”的阶层归属，也无法判断其是否针对弱势群体发表；其次，居高临下言论中存在“同情”这一明确类别，但应将其与真诚关怀加以区分，因为虚伪关怀与真实的同情在语言表达上极为相似，但背后的意图却不同。例如，**“我很同情你，我们都在经历很多”**这句话就不属于居高临下表述，因为它并未通过虚伪的关怀反映出不同阶层的优越感，而是真正出于对相同群体的关心。为了有效划分这些模糊的判断，模型需要具备更多的世界知识和更为精准的定义。

3.5 本章小结

针对弱势群体的居高临下言论识别是一项具有潜在社会价值的研究任务，能够帮助自然语言领域有害言论检测的研究者更好地理解隐性有害表达，并为关爱弱势社区和多语言群体联合研究提供更多理论支撑。然而，当前中文领域的相关研究仍相对滞后，关于中文领域的 PCL 定义、数据、基线方法都存在大量空白，这严重阻碍了中文隐性有害言论研究的推进和面向社交媒体的居高临下检测研究。基于以上问题，本章节首先提出了 CondescendCN 框架，这是中文互联网上首个系统性用于居高临下言论识别的框架，能够对评论内容进行更细粒度的划分。在此基础上，本章构建了首个中文 PCL 数据集 CCPC，并验证了引入毒性强度 TS 特征对居高临下言论检测的有效增益。训练后的模型能够在更大规模的平台上执行检测任务，进一步证明了 CCPC 数据集的可靠性。此外，本章节还进行了详细的数据统计分析，发现居高临下言论主要指向女性和儿童等弱势群体，这些群体亟需更多的人文关怀。实验结果表明，居高临下言论检测依然具有较强的主观性，其科学定义也尚不明晰，模型在该任务上亟需更多的上下文信息与领域知识扩展。本章节为后续章节，尤其是多语言、多平台的居高临下群体检测打下了坚实的数据、基线基础。

4 基于大模型的中英双语居高临下检测与研究

4.1 引言

本文第三章针对于现有居高临下言论定义、判别标准具有模糊性以及语料呈现稀缺性的特点，开创性的提出了 CondescendCN 分层框架，并在此基础上提出了面向社交媒体的 CCPC 数据集和相关基线，在权威的社交平台进行了有效的迁移性测试。第三章从多维度、细粒度的角度填补了中英文在隐性有害言论检测领域的空白。然而，由于居高临下言论本身判别的主观性和潜在分歧，使用传统的预训练模型（如 BERT）在准确的判别意图上并不能取得最佳的效果；同时随着大型语言模型（LLM）时代的到来，世界知识和相关算法有大一统的趋势，这使得我们能够借助大模型，利用更广泛的预训练语言知识，使用指令微调和对比研究的范式，以增强居高临下的双语检测性能和多平台研究能力。

大型语言模型 LLM，是指参数量通常在数十亿甚至千亿以上的深度神经网络，基于海量数据进行预训练，具备强大的语言理解和生成能力。它们广泛应用于对话系统、文本生成等各类自然语言处理任务，甚至拓展至多模态（图文、语音等）领域。大模型的核心优势在于其超大的预训练参数，和规模化带来的强泛化能力，能够通过“预训练-微调”或“零样本/少样本学习”的方式，处理复杂多样的任务，而不必为每一个任务专门设计特定的模型结构或大量收集标注数据。

近年来，解码器架构驱动的大型语言模型不仅在文本生成领域展现出颠覆性的进展，也日益成为应对有害语言检测与防控任务的重要工具。得益于其在大规模语料上的强大预训练能力，这类模型能够在极少甚至无需标注数据的情况下，通过提示引导完成文本分类任务，显著降低了对人工标注资源的依赖。实践中，人们观察到，不同的训练方式会对模型的输出行为产生显著影响。例如，零样本的链式推理提示可能增加模型生成有害内容的倾向，而某些监督式调优或强化学习方法亦可能放大这一风险。为提升检测准确性，研究人员也在不断尝试通过优化提示构造或融合弱势群体语义信息来增强模型的判断能力。凭借其庞大的参数量与深层语义建模能力，LLM 在捕捉复杂语境和理解隐含意图方面具备显著优势。然而，目前尚无系统性的 LLM 工程化方案被用于检测居高临下言论或其他歧视性文本。此外，目前对于居高临下言论的多语言、多平台研究尚处于空白阶段。为了利用 LLM 时代的技术进一步解决居高临下言论领域相关的问题，我们首先需要关注以下三个问题：

（1）模型训练成本：我们如何高效的对超大规模参数的模型进行高效微调？有没有节约成本的方式？

（2）数据稀缺：针对于 LLM 预训练和微调的要求，我们如何进一步集成已有数

据集和建立新的数据集，以满足我们居高临下 LLM 的基本要求？

（3）多语言要求：居高临下问题是全球社区共同面临的问题，而不仅限于英文社区。我们如何基于前两点设计一组能够提升隐性有害言论识别能力的多语种（例如中英文双语）LLM 模型基准，以支持中文地区的弱势群体，对比中英文弱势群体的差异，并为其他地区的未来工作设定范式？

为解决上述问题，我们提出了 PclGPT—第一个面向居高临下检测的综合性 LLM 基准，专注于探索 LLM 对隐性有害言论的理解能力。首先，我们从主流互联网平台收集社区数据，分别选取 Reddit 作为英文数据源，新浪微博作为中文数据源，构建面向特定领域自适应预训练的 Pcl-PT 数据集。随后，我们对数据进行标注、重构与筛选，构建高质量的 Pcl-SFT 数据集，构建指令数据范式，在输入和输出两端引入更多约束条件。接着，我们完成了全流程的预训练与指令监督微调，创新性的将两种先进的高效微调方式 Prefix-Tuning 和 LoRA 进行结合，构建了双语大模型组 PclGPT-EN/CN。这一模型组是目前已知的首个专为居高临下检测设计的 LLM 基准。在处理难以区分的模糊样本测试中，该模型在双语任务中均显著优于其他主流 PLM 和 LLM 基线。此外，进一步的群体检测与中英双语社区研究表明，居高临下言论在针对不同弱势群体时存在显著的偏向差异，不同居高临下子类别中的偏向模糊性也不尽相同。上述发现凸显了社会亟需进一步关注居高临下现象，以更有效的保护全球范围的多语种弱势群体。综上，本章节的主要贡献总结如下：

（1）我们创新性的使用 LoRA 结合 Prefix-Tuning 高效微调的方式进行指令微调的工作，以节约训练成本和提高效果。

（2）我们构建了 Pcl-PT 和 Pcl-SFT 数据集，用于提升模型对居高临下言论的领域知识理解。Pcl-PT 涵盖超过 140 万条来自弱势群体社区的数据，用于模型预训练；Pcl-SFT 则包含高质量的双语指令样本，用于模型指令微调。我们在此基础上成功训练出双语模型 PclGPT-EN/CN。作为首个专为居高临下及隐性有害语言检测设计的大模型组，PclGPT 在四个数据集上均超越现有先进的 PLM 和 LLM。

（3）通过群体检测、细粒度有害性分析与中英文双语对比研究，我们在前述章节的基础上进一步揭示了居高临下言论在不同弱势群体中的偏向差异，部分群体所面临的隐性有害言论问题尤为严重，且中英文社区面对的居高临下等隐性威胁同等严峻。PclGPT 为识别和管理这些偏向性群体奠定了基础，有助于更好地保护受影响群体。

4.2 PclGPT 模型设计

4.2.1 模型总体框架

本章节的整体方法如图 4.1 所示。我们的 PclGPT 模型组由两个子模型组成：PclGPT-EN 和 PclGPT-CN，分别基于 LLaMA-2-7B 和 ChatGLM-3-6B[49] 架构。LLaMA

是当前最具代表性的英文开源大语言模型之一，已在超过 20 万亿个 tokens 上进行了预训练。ChatGLM 作为最先进的中文开源大模型之一，基于广义线性模型 GLM 架构，针对中文问答与对话任务进行了大量优化，在中文领域表现出色。LLaMA-2-7B 支持最长 4096 个标记的上下文长度，ChatGLM-3-6B 则支持 8192 个标记，能够更充分地理解上下文信息。我们通过设计组合更高效的微调机制，同时构建全面的预训练-指令微调框架，第一次提出了完整的居高临下检测大模型。

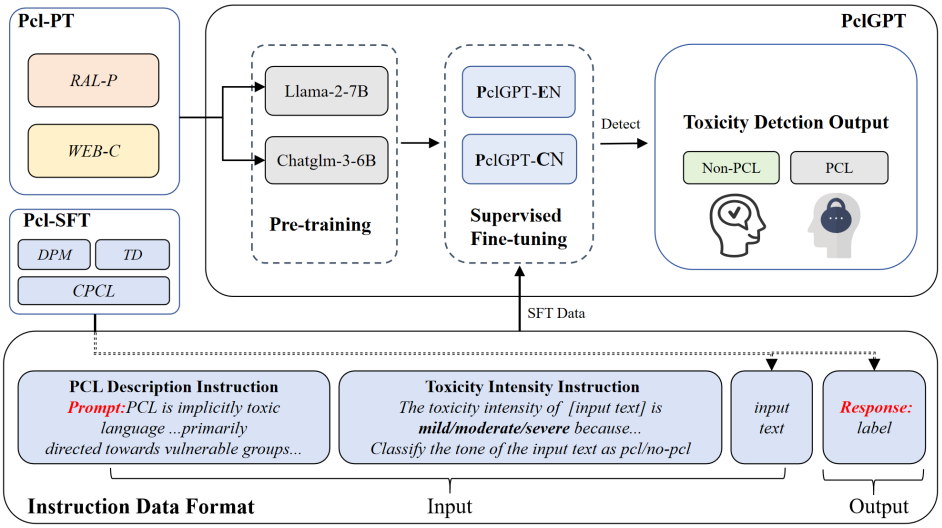

图 4.1　PclGPT 中英双语 LLM 的总体框架

Fig. 4.1　Overall framework of the PclGPT bilingual LLM

4.2.2　Prefix-Tuning 与LoRA 的高效微调协同设计

在大模型的训练中，全参数微调通常需要对数百亿甚至千亿参数进行更新，代价高昂。为此，参数高效微调（PEFT）方法逐渐成为当前研究热点。本章节提出一种结合 Prefix-Tuning 与 LoRA 的高效微调框架，通过在 Transformer 的注意力模块中同时引入可学习前缀向量与低秩重参数化，实现对输入空间与模型参数空间的协同优化，减少微调参数量的同时，保持模型对下游任务的强适应能力。

在 LLM 的高效微调中，Prefix-Tuning 与 LoRA 被设计用于不同的空间：前者作用于输入空间，后者作用于参数空间。这种协同机制能够在保持参数冻结的同时，引入少量任务特定的可训练参数，同时进一步提升高效微调的整体效率。

具体而言，Prefix-Tuning 通过在注意力计算前向 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 矩阵注入前缀向量，引导模型对任务先验信息的关注，而 LoRA 则对 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的线性变换矩阵（即 $\boldsymbol{W}_q$ 和 $\boldsymbol{W}_k$）进行低秩分解，调整局部特征的表示能力。

两者协同后的 Attention 计算表达为：

$$\boldsymbol{Q} = \boldsymbol{H}(\boldsymbol{W}_q + \boldsymbol{B}_q\boldsymbol{A}_q) \tag{4.1}$$

公式 4.1 表示查询向量 $\boldsymbol{Q}$ 的计算方式，其中 $\boldsymbol{H}$ 是输入的隐藏状态，$\boldsymbol{W}_q$ 是原始的查询权重矩阵，$\boldsymbol{B}_q\boldsymbol{A}_q$ 为 LoRA 方法引入的低秩权重调整项。通过在原始权重上叠加一个低秩矩阵乘积，LoRA 能以较小的参数量实现对模型性能的微调。

$$\boldsymbol{K'} = [\boldsymbol{P_k}; \boldsymbol{H}(\boldsymbol{W}_k + \boldsymbol{B}_k\boldsymbol{A}_k)] \tag{4.2}$$

公式 4.2 给出了修改后的键向量 $\boldsymbol{K'}$ 的构成，其中 $\boldsymbol{P_k}$ 是 Prefix-Tuning 引入的前缀键向量，$\boldsymbol{W}_k$ 是原始键权重，$\boldsymbol{B}_k\boldsymbol{A}_k$ 为 LoRA 的低秩调节项。通过将前缀向量与经过 LoRA 调整后的键向量拼接，模型能够同时结合固定的任务信息和上下文相关的动态信息。

$$\boldsymbol{V'} = [\boldsymbol{P_v}; \boldsymbol{H}\boldsymbol{W}_v] \tag{4.3}$$

公式 4.3 表示修改后的值向量 $\boldsymbol{V'}$，由前缀值向量 $\boldsymbol{P_v}$ 与上下文值向量 $\boldsymbol{H}\boldsymbol{W}_v$ 拼接组成。与键向量不同的是，值向量部分仅应用了 Prefix-Tuning 方法，而未引入 LoRA。最终，注意力输出为：

$$\boldsymbol{O} = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K'}^\top}{\sqrt{d_h}}\right)\boldsymbol{V'} \tag{4.4}$$

在此结构中，Prefix-Tuning 提供了“前缀引导机制”，为注意力机制引入显式的“虚拟上下文”，增强了输入信息的上下文可控性；而 LoRA 则以低秩分解的方式，赋予 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 更强的任务特异性表示能力。两者结合，提升了模型对输入空间与参数空间的双重适应能力，进一步提高了微调效率和节约了成本。

考虑模型包含 L 层 Transformer Block，每层均集成 Prefix-Tuning 和 LoRA 模块，整体可训练参数量由两部分组成：

$$\text{Params}_{\text{total}} = L \cdot (2ld_h + 4rd_h) \tag{4.5}$$

其中，$2ld_h$ 表示 Prefix-Tuning 在 $\boldsymbol{K}$ 和 $\boldsymbol{V}$ 中引入的前缀向量 $\boldsymbol{P_k}, \boldsymbol{P_v}$，每个前缀向量的维度为 $l \times d_h$，分别作用于注意力模块中的键和值；$4rd_h$ 表示 LoRA 在 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的线性映射矩阵 $\boldsymbol{W}_q, \boldsymbol{W}_k$ 上引入的 rank 为 r 的双低秩矩阵（$\boldsymbol{A}$ 与 $\boldsymbol{B}$）。

在本工作中，前缀长度 l 与 LoRA 低秩维度 r 采用与输入长度 n 和隐藏维度 d_h 成比例的设置方式，具体为：

$$l = \alpha n, \quad r = \beta d_h \tag{4.6}$$

其中，α 和 β 为控制比例的超参数，取值范围设定如下：

$$\alpha \in \{0.1, 0.2, 0.3\}, \quad \beta \in \{0.05, 0.1\} \tag{4.7}$$

在不同实验场景下，α 与 β 的取值依据硬件资源和具体任务难度进行调整。本节所述参数设置方案有效平衡了模型微调过程中的计算成本与性能表现，能够满足大模型在多种下游任务中的微调需求。

4.2.3 模型预训练

为了促进预训练过程，我们引入了 Pcl-PT 数据集，该数据集由 RAL-P 和 WEB-C 两部分组成。具体而言，为了满足中英文双语社区的对比研究，我们采用了中英文分离的语料库，分别对 PclGPT-EN/CN 模型组进行预训练。预训练遵循标准的自回归范式，即模型根据已有的输入历史预测下一个 token。对于 PclGPT-EN 和 PclGPT-CN，我们均采用了与其基础模型相同的词表，并使用 AdamW 作为优化器，初始学习率设置为 2×10^{-4}，权重衰减为 0.1。此外，我们还采用了高效的训练策略，包括 Micikevicius 等人[66] 提出的基于 bf16 的混合精度训练。

预训练目标遵循经典的自回归语言建模损失函数：

$$\mathcal{L}_{\text{PT}} = -\sum_{t=1}^{T} \log P(x_t | x_{<t}; \theta) \tag{4.8}$$

并在此基础上引入权重衰减正则化，以提升模型的泛化能力：

$$\mathcal{L} = \mathcal{L}_{\text{PT}} + \lambda \|\theta\|_2^2 \tag{4.9}$$

本章的数据工作量巨大，除了已有公开数据集的指令改造，还包括全新的双语预训练和指令微调数据集。下面将详细介绍数据集设计。我们的数据集设计遵循 tian 等人[67] 提出的分层结构，具体细节如下表 4.1 所示。数据集的构造将在下文详细阐述。

表 4.1 各阶段用于训练 PclGPT 的数据集统计。Pcl-PT 用于预训练，Pcl-SFT 用于指令微调阶段
Tab. 4.1 Statistics of the datasets used in training PclGPT under different stages. Pcl-PT is used in the pre-training stage, and Pcl-SFT is used in the SFT stage

训练阶段	数据集	语言	自建/公开	# 总量
Pcl-PT	RAL-P	EN	**Self-built**	1091945
	WEB-C	CN	**Self-built**	315074
Pcl-SFT	Don't Patronize Me (DPM)	EN	Public	10469
	TalkDown (TD)	EN	Public	74865
	CPCL	CN	**Self-built**	18253
Test	DPM/TD/CPCL/CCPC	EN,CN	Public	N/A

（1）RAL-P 英文预训练数据集。该数据集改编自 RAL-E 数据集。Caselli 等人[7] 构建的 RAL-E 数据集收集了 Reddit 社区中包含攻击性、辱骂性和仇恨内容的语料，包

含自 2005 年 12 月至 2017 年 3 月期间的 4300 万个 token。然而，RAL-E 中以明显的仇恨言论为主，这在一定程度上混淆了对居高临下言论的准确识别。因此，我们首先利用 LLM 生成了一个包含 500 余个英文居高临下术语的词典，并由三位人工校对者协同过滤，剔除了与居高临下无关的词条，最终保留了 379 个相关术语。随后，我们使用该词典对 RAL-E 进行选择性匹配，筛选出与居高临下更密切相关的数据，同时保留了 30% 的非居高临下样本，以保证预训练数据的平衡性。最终，RAL-P 数据集共包含 1,091,945 条预训练数据。

（2）WEB-C 中文预训练数据集。中文领域的数据稀缺性限制了居高临下检测尤其是预训练任务的开展。为了解决这一问题，我们设计了一个框架，系统性地从新浪微博这一主流中文媒体平台，收集针对边缘化群体的霸凌、暴力与歧视性内容。我们首先依据中文已有的居高临下标准[68] 将搜索范围限定在主要弱势群体类别，并相应扩展了关键词列表。随后，基于这些关键词，我们爬取了 2022 年 7 月至 2024 年 1 月期间的微博帖子，并对数据进行了有效筛选。最终，共收集到 315,074 条数据。

4.2.4 指令数据格式

近年来的研究强调了监督指令微调在塑造大语言模型认知能力中的关键作用。Chiang 等人[69] 指出适当格式化的指令数据有助于充分发挥大模型的知识潜力。Wang 等人[68] 的研究指出，纳入细粒度的毒性强度可以进一步增强居高临下识别的鲁棒性。我们构建的指令模板同时包括了居高临下描述指令和毒性强度指令，旨在更准确地捕捉居高临下言论的隐晦语义特征，本章节构建的双语指令微调模板细节如图 4.2 所示。

（1）PCL 描述指令。由于居高临下言论属于主观性较强的有害类别，因此首先需要提供完整的居高临下描述，以引导模型以规范的格式进行响应。该描述包括居高临下的定义及其子类别。这部分内容为固定的描述性信息。

（2）毒性强度指令（可选）。接下来，我们关注毒性强度对隐性情绪的潜在影响。我们使用开源的 Perspective API[70] 对文本的危害性进行评分，并基于这些得分，在原始数据中引入了毒性强度标签，具体划分为轻度、中度和重度。

4.2.5 指令微调

按照第 4.2.4 节中制定的指令格式，我们构建了用于 SFT 过程的 Pcl-SFT 数据集，其中包括英文数据集 Don’t Patronize Me! 和 TalkDown，以及中文数据集 CPCL。我们遵循第 3.1 节所述的双语训练规则，以确保 PclGPT 具备多语言的检测能力。我们基于上述构造的指令模板，对 PclGPT 进行系统化的指令微调，训练目标为最小化以下所示损失函数：

$$\mathcal{L}_{\text{SFT}} = -\sum_{t=1}^{T} \log P(y_t|x, y_{<t}; \theta) \tag{4.10}$$

(PCL Description Instruction)
Suppose you are a linguist and you are asked to judge whether a given text is patronizing and condescending. *<definition of PCL>*
Main Subcategories (Optional):
<definition of Subcategories1><definition of Subcategories2>...

(Toxicity Intensity Instruction) (Optional)
The toxicity intensity of this sentence is *mild/moderate/severe*

Your return: Based on the following conversation, make a decision and return your choice.
Here is the text: *input text*

Output: *label*

图 4.2　PclGPT 的详细指令微调模板构建
Fig. 4.2　Instruction-tuning template construction of PclGPT

其中，x 表示输入的指令和上下文，y 为输出的响应序列，θ 为模型参数。

CPCL 指令微调数据集。CPCL 是我们从中文社交媒体平台手工收集和标注的中文数据集，依据居高临下言论的定义对数据进行了分层结构化标注，包含危害存在性、细粒度 PCL 子类别和弱势群体考量。我们在上述章节构造的 CCPC 数据集的基础上，进一步完善了标注准则和质量扩充，CPCL 最终获得了累计超过 18,000 条二级结构标注。标注过程由两名标注员初步标注，一名复审员校对，后续还进行了主观一致性审查，确保标注数据的可靠性。

按照第 4.2.4 节所述的指令微调数据格式，我们结合了 PCL 描述与毒性强度信息，生成不超过 LLM 序列长度上限的长输入文本。在训练过程中，我们采用序列到序列的损失函数，并将生成的输出映射为二分类标签。

SFT 训练在 8 张 RTX 4090 GPU 上进行，采用 4.2.2 所述的 LoRA 与 Prefix-Tuning 协同高效微调策略：Prefix-Tuning 以长度为 l 的前缀向量插入至所有注意力模块中的 $\boldsymbol{K}$ 和 $\boldsymbol{V}$，同时 LoRA 以 rank 为 r 的低秩矩阵作用于 $\boldsymbol{Q}$ 和 $\boldsymbol{K}$ 的权重映射。前缀长度 l 和低秩维度 r 与输入长度 n 和隐藏维度 d_h 成比例缩放，具体设置为 $\alpha = 0.2$ 和 $\beta = 0.05$。

训练过程使用 AdamW 优化器，学习率设为 2e-5，最终共完成了 5 轮的训练，并

表 4.2　PclGPT 在预训练和指令微调阶段的详细参数设置
Tab. 4.2　Detailed configuration parameters for the pre-training and SFT phases of PclGPT

预训练参数	数值	微调参数	数值
Lr	2e-4	Lr	2e-5
Batchsize	32	Batchsize	16
Training Epochs	5	Training Epochs	5
Max Source Len	512	Block Size	1024
Max Target Len	512	-	-
GPUs	RTX 4090*8	GPUs	RTX 4090*8
-	-	GPUs_inference	A100_PCIE*2

在验证集上保持了最佳模型的权重。所有非核心模块参数保持冻结，仅更新前缀向量与 LoRA 注入参数，确保参数高效利用。

4.3　实验结果与分析

4.3.1　基线设计

为验证 PclGPT 的性能，我们在四个中英文数据集（两个公开数据集）上针对多种 PLMs 和 LLMs 以及 PclGPT 模型组进行了全面测试。为确保模型在双语 PCL 检测中的最优表现，我们分别使用 PclGPT-EN 处理英文数据集，PclGPT-CN 处理中文数据集。

（1）PLMs。预训练语言模型一直是传统有害言论检测任务中最核心的模型类型。我们在 PLM 类别中选用了 BERT 及其相关变体，如 RoBERT[71]、ChineseBERT(C-BERT)[72] 和 Multilingual-BERT (M-BERT)[73]。为确保 PLMs 在测试集上的最优表现，我们采用标准的训练与微调流程。具体而言，使用公开数据集的训练集对 PLMs 进行训练（对于 CCPC，继续使用 CPCL 的训练集）。此外，为保证对比公平，PLMs 与 LLMs 共用同一测试集进行评估。

（2）Base-LLMs。LLMs 的使用分为两个部分。一方面，对于 ChatGPT、Claude-3 等先进但非开源的 LLMs，我们通过 API 接口调用；另一方面，我们在 PclGPT 消融实验中使用未经过参数微调的 LLaMA-2-7B 和 ChatGLM-3-6B 原始版本，以评估微调带来的性能提升。为保持实验一致性，其余 LLMs 也统一采用与 PclGPT 相同的指令格式。考虑到居高临下具有隐性有害特征，且 Base LLMs 在 few-shot 场景下表现受限，我们采用 zero-shot 测试方式进行更清晰的对比。下表 4.2 展示了详细的 PclGPT 训练过程的参数配置。

对于 PLMs 与 LLMs 的实验结果，我们采用领域通用的宏平均的 Precision（P）、Recall（R）与 F1-score（F1）三项指标进行评价。

4.3.2 总体实验结果

表 4.3 详细对比了 PclGPT 与 PLMs 及其他 LLMs 在四个测试集上的检测表现。

表 4.3 主实验结果。最佳与次优结果分别以**加粗**与下划线 标注。- *TII* 表示移除毒性强度指令模板的消融试验结果

Tab. 4.3 The main experiment results. Optimal and suboptimal scores are denoted in **bold** and underlined, respectively. - *TII* is the result of removing the Toxicity Intensity Instruction template

		DPM			TD			CPCL			CCPC
LM	**Model**	P	R	F1	P	R	F1	P	R	F1	F1
PLMs	RoBERTa	76.3	78.7	77.4	88.4	86.7	86.5	61.2	61.3	61.3	55.4
	RoBERTa-L	80.2	74.9	77.2	88.1	86.0	85.9	62.5	61.6	62.0	55.3
	C-BERT	71.2	63.5	66.2	76.7	74.7	74.2	66.6	71.0	67.3	57.1
	M-BERT	69.2	76.0	71.8	87.6	87.4	87.4	65.8	67.8	66.6	56.0
Base-LLMs	ChatGPT	50.8	52.3	46.9	59.2	58.1	56.7	53.1	54.2	53.6	53.3
	GPT-4.0	51.5	57.5	54.3	60.8	60.3	60.5	55.4	56.3	55.7	56.3
	Claude-3	52.3	52.5	52.3	61.6	64.1	63.2	57.2	57.7	57.3	57.6
	LLaMA-2	50.9	52.6	51.4	49.9	49.9	49.7	45.2	47.5	46.3	42.5
	ChatGLM-3	N/A	N/A	N/A	N/A	N/A	N/A	51.9	50.2	51.0	49.1
LLMs(Ours)	**PclGPT-EN**	**80.4**	**81.8**	**81.1**	**89.9**	**89.0**	**88.9**	N/A	N/A	N/A	N/A
	- *TII*	79.5	80.3	79.9	88.5	88.0	88.2	N/A	N/A	N/A	N/A
	PclGPT-CN	N/A	N/A	N/A	N/A	N/A	N/A	**69.1**	**72.0**	**70.2**	**60.2**
	- *TII*	N/A	N/A	N/A	N/A	N/A	N/A	68.1	71.0	69.5	57.2

观察表 4.3 我们可以获得如下结论：

（1）PLMs 在有害言论检测领域仍具有重要地位，但其局限性也十分明显。从主观性歧义的角度来看，PLMs 在数据分布均匀、定义清晰的 TalkDown（英文）数据集上表现良好，但在定义更为模糊的 DPM（英文）与 CPCL（中文）数据集上表现欠佳。

（2）PclGPT 在中英文领域均取得了优异的检测效果，尤其在处理存在歧义的数据时表现突出。具体而言，PclGPT 在 DPM 数据集上相较于最佳的 RoBERTa 模型提升了 3.7%，在 CPCL 数据集上相较于最佳的 Chinese-BERT 模型提升了 2.9%。

（3）Base-LLMs 在未进行参数调整的情况下，未能在主观有害检测中展现出潜力。由于缺乏对有害文本的专门优化，未经微调的 LLMs 在检测隐性危害性文本（如 PCL）时表现不佳。与 PLMs 相比，LLMs 的平均精度降低约 20.49%，召回率降低约 18.87%，F1 值下降约 19.66%。值得关注的是，PCL 样本中常包含正面表达与善意语义，这干扰了 LLMs 的预训练特征。PclGPT 有效引导 LLMs 理解 PCL 的有害性定义与子类别，为未来 LLM 安全规范的制定提供了重要参考。

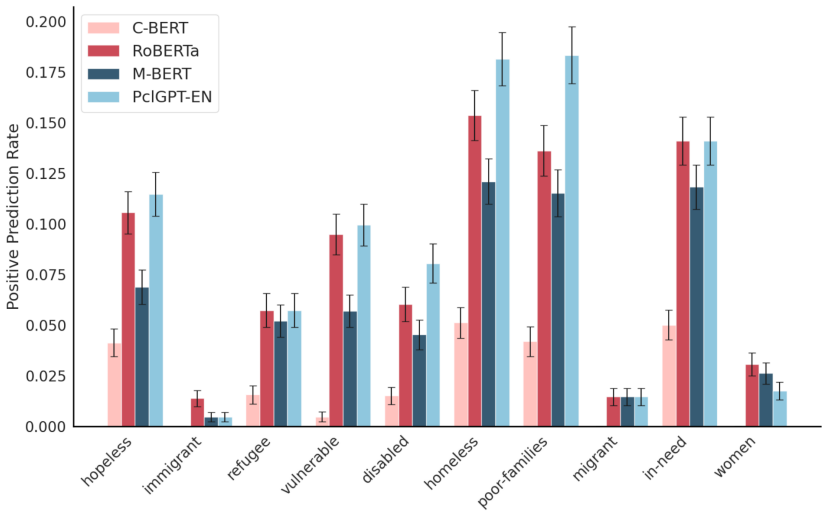

图 4.3　不同模型的群体检测。测试群体包括 10 个不同的弱势群体

Fig. 4.3　Group detection for different models. The test group consists of 10 different disadvantaged communities

4.3.3　居高临下偏差性分析

（1）群体检测。群体检测有助于我们解决模型在不同群体中的偏差问题。我们使用 DPM 数据集进行了实验，该数据集在多个少数群体之间实现了覆盖的平衡。在这些实验中，我们比较了微调后的 BERT 系列模型与 PclGPT-EN 的表现，如图 4.3 所示。测试集中，各类弱势群体与正样本分布均衡。然而，模型在识别贫困家庭和无家可归者时表现出明显的偏好，表明这些群体具有更易识别的语义特征。针对这些群体的同情或怜悯表达更容易被感知为居高临下。PclGPT 进一步提升了对这些群体的检测能力。相比之下，对于移民与外来人口等群体，含糊的歧视性态度依然难以识别，表明有必要采取更多措施以保护这些群体。

（2）细粒度分析。对有害类别的细粒度分析对于理解隐性毒性情感具有重要意义。我们的中文 CPCL 数据集将 PCL 划分为五个子类别。为测试 PclGPT-CN 对不同类型毒性的敏感性，我们依据这五个类别将 CPCL 数据集划分为五个子集。在实验中，我们将 PclGPT-CN 与 Chinese-BERT 和 ChatGLM 进行了对比。表 4.4 展示了我们在细粒度 PCL 检测任务中的实验结果。实验表明，模型在检测不同 PCL 子类别时仍存在不同程度的偏差。在“呼吁”和“同情”这两个子类别中，主观且模糊的表达形式容易干扰模型的识别效果。值得注意的是，我们的 PclGPT-CN 在所有子类别上均取得了性能提升，尤其在模糊性的“呼吁”子类别中提升最为显著。

4.3.4　中英文社区居高临下言论的定性对比分析

在本章节的工作中，我们为建立中英双语的多平台检测大模型进行了大量的对比论证工作，图 4.4 展示了我们的其中一项关键成果。图 4.4（左）为英文语料的居高临

表 4.4　细粒度 PCL 检测实验结果。我们使用宏平均 F1 值作为评价指标
Tab. 4.4　Experimental results for fine-grained PCL Detection. We evaluated our model using the macro-average F1-score as the metric

子类别	ChatGLM	Chinese-BERT	PclGPT-CN
Unb.	52.1	66.5	**69.4** ↑2.9
Spectators	44.3	71.3	**72.1** ↑0.8
Prejudice	49.7	64.3	**67.5** ↑3.2
Appeal	24.5	59.0	**65.0** ↑6.0
Compassion	44.2	52.3	**57.4** ↑5.1

下样例散点抽样（蓝色散点），图 4.4（右）为中文语料的居高临下样例散点抽样（蓝色散点），与之对比的是对应的英文和中文仇恨语料样例的散点抽样（橙色散点），散点图的横坐标为经过 Perspective API 得到的毒性评分，纵坐标为情感得分。为了确保质量，所有的随机采样都进行了 5 次重复实验。观察结果可以发现，相比于强攻击性的仇恨语料，中文和英文的居高临下言论都具有较低的毒性评分，这证明了居高临下言论在多语种领域具有语义共通性，即隐式的有害特征，不通过明显的传统攻击范式传递歧视特征。同时中文的散点相较于英文语料更加发散，这说明了中文语料的部分样例更加具有攻击倾向；在情感得分指标中，无论是中文语料还是英文语料都在全区间呈现发散性特征，这进一步证明了居高临下言论在全球范围内的模糊性语义特点，即我们通过单一的情感分析并不足以有效区分这种弱危害性语言。居高临下言论可能是攻击性的、消极的，另一方面，它也可能是看似积极或中性的、实则“伪善”的样例。这进一步凸显了利用大模型的世界知识，强化领域理解的重要性。

4.3.5　PclGPT 中英居高临下双语样例对比分析

为了进一步说明 PclGPT 的设计原理，并验证模型组是否能够有效解决我们最关切的模糊的居高临下子类别的识别问题，我们分别从中英文测试结果中选取样本进行案例测试，结果详见表 4.5。英文部分选取了 M-BERT、RoBERTa、GPT-4.0、Claude-3、LLaMA-2-7B 和本章的 PclGPT-EN 进行对比分析；中文部分则选用了中文预训练模型 Chinese-BERT、ChatGLM-3-6B 和本章的 PclGPT-CN 进行对比分析。

案例 A 主要选取了 PCL 中带有“不平衡的权力关系-Unbalanced Power Relations”和“偏见-Prejudice”标签的样本。在这些例子中，占优势群体将自己置于更高的社会地位，并对弱势群体表现出强烈的歧视性特征。例如，A(i) 中的“so-called”讽刺了贫困群体不应获得补助，这是一种典型的偏见表达；A(ii) 则体现了对“单亲家庭的孩子难以相处”的刻板印象。这类言论的有害特征非常明显，尽管没有精确的攻击性词汇，模型仍能有效识别。其中 A(i) 被大多数模型准确识别，A(ii) 也取得了类似结果，说明中文模型同样利用了 PCL 的语义信息。

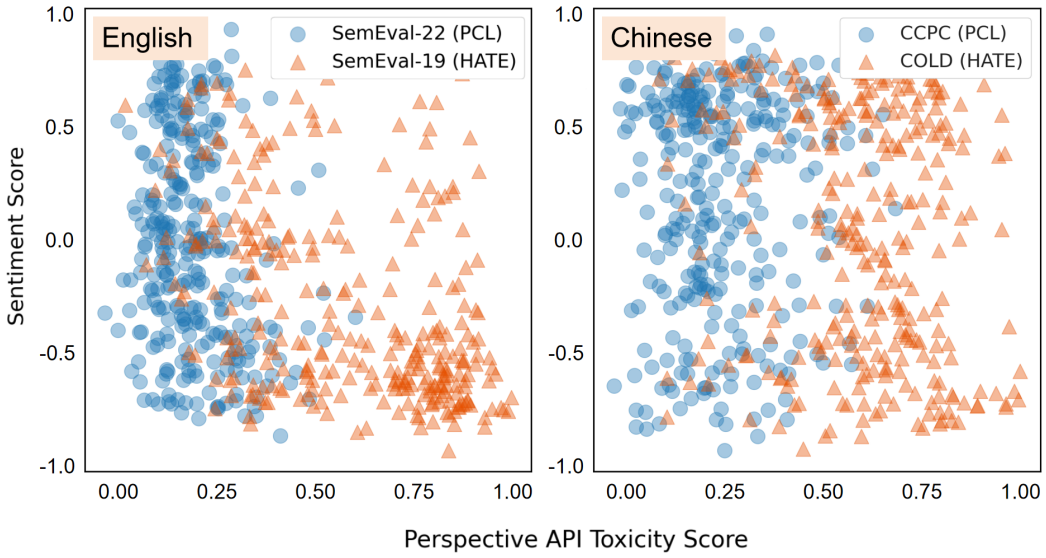

图 4.4　对中英文居高临下语料的对比分析散点图（仇恨语料为参照）
Fig. 4.4　Scatter plot comparing condescending expressions in Chinese and English corpora, with hateful expressions as reference

案例 B 所选样本多属于“旁观者-Spectator”和“引起同情-Elicit Compassion”的子类别，这些语料将优势群体置于旁观者的角色，仅提供表面化的意见来解决问题，或对弱势群体表示同情。例如 B(i) 通过描述性句子激发人们对“客户”的同情，B(ii) 表达了对“农民工”的关切，并呼吁保障工资。这类言论的 PCL 危害性隐藏在模糊的表达中，模型难以识别其隐性有害特征。对于 B(i)，仅 Claude-3 和 PclGPT-EN 正确识别；对于 B(ii)，仅 ChatGLM-3 和 PclGPT-CN 成功识别。这凸显了 PclGPT 在隐性有害言论检测方面的重要性。

4.4　本章小结

在本章节中，我们介绍了 PclGPT，一个专门设计用于检测针对弱势群体的居高临下言论的大型语言模型组。PCL 作为有毒语言的一个子集，通过歧视性语言对弱势群体造成伤害。传统的预训练语言模型在 PCL 检测方面面临困难，因为其难以准确识别隐性的有害特征。PclGPT 通过利用大型语言模型在情感语义方面的能力，显著提高了检测性能。我们将 LoRA 和 Prefix-Tuning 在不同位置的高效微调策略进行整合，提出了联合高效微调框架；接着，我们收集、标注并合并了 Pcl-PT/SFT 数据集，并通过全面的预训练和 SFT 微调过程建立了 PclGPT-EN/CN 模型组，以检测中文和英文社区中的 PCL。PclGPT 在四个数据集上的表现超过现有模型，展示了其在处理隐性有害语言方面的强大能力。此外，群体检测和细粒度有害类别分析揭示了针对各种弱势群体的显著偏见差异，凸显了社会保护的紧迫需求。对比实验表明，无论是中文社区还是英文社区，居高临下言论都存在全球化泛化性的隐性特征，对于双语以及多语言

表 4.5　案例分析示意图。我们分别从英文和中文测试集中选取了代表性的样本
Tab. 4.5　Case study illustration. We select representative samples from the English and Chinese test sets

英文例	案例 A(i)	案例 B(i)
Text	*After already receiving relief funds, what else do these so-called 'poor' families think they deserve?*	*For some of these male prostitutes, the 'clients' they picked up on this corner were their only means of survival.*
Category	"Unbalanced Power Relations","Prejudice"	"Spectator","Elicit Compassion"
Expl.	The phrase "so-called 'poor' families" suggests a condescending attitude towards impoverished households, reflecting an unbalanced power relationship, where those with more resources view those with less through a biased perspective. The tone is dismissive and judgmental.	The phrasing of this sentence suggests a spectator's indifferent attitude towards male prostitutes. It implies that these men have no other choice but to engage in sex work for survival. Spectators elicit compassion for their plight while maintaining a superior stance. The toxicity of such descriptive statements is often complex to detect.
Difficulty	Middle	High
Prediction	M-BERT:✔, RoBERTa:✔, GPT-4.0:✗, Claude-3:✔, LLaMA-2:✗, PclGPT-EN:✔	M-BERT:✗, RoBERTa:✗, GPT-4.0:✗, Claude-3:✔, LLaMA-2:✗, PclGPT-EN:✔

中文例	案例 A(ii)	案例 B(ii)
Text	单亲的小孩大概率很难相处。	农民工挣钱不容易的，确保工资该发就发呗。
Category	"Unbalanced Power Relations","Prejudice"	"Appeal","Elicit Compassion"
Expl.	这句话体现了对单亲家庭的不平等权力关系和偏见。它假设来自这类背景的儿童天生面临社会困境，忽视了个体经历的复杂性以及可能存在的多样化支持系统。	这句话对农民工人表面上的公平呼吁掩盖了发言者内心隐性的偏见。对于他们被欠债的原因轻描淡写，而仅从工资出发，将他们等同于可怜的乞讨者。由于缺乏攻击性意图，这种居高临下的态度在没有深入分析的情况下难以察觉。
Difficulty	Middle	High
Prediction	RoBERTa:✗, Chinese-BERT:✔, GPT-4.0:✗, Claude-3:✔, ChatGLM-3:✔, PclGPT-CN:✔	RoBERTa:✗, Chinese-BERT:✗, GPT-4.0:✗, Claude-3:✗, ChatGLM-3:✔, PclGPT-CN:✔

的联合研究将有效促进隐性有害言论检测领域的进一步发展。

无论是传统的预训练模型还是大型语言模型，目前针对于居高临下的研究集中于文本模态的工作，根据感知上的理解，居高临下言论往往都是由说话者的语言和肢体表情共同作用的，比如发表居高临下言论时对弱势群体的的歧视性神态表情。因此，我们认为在接下来的工作中引入多模态领域的知识和框架是进一步推进隐性有害言论检测的重要潜在手段。

5 基于歧视性表情特征的多模态居高临下言论检测

5.1 引言

社交媒体平台的迅速发展超出了预期，尤其是自 2010 年以来，视频自媒体逐渐在主流平台上传播思想，包括英文平台 YouTube、TikTok 和中文平台 Bilibili[74]。这些视频平台不仅创造了显著的经济效益和社会影响力，还加速了有害内容的传播，如本文第三、四章所述，这些内容通常也被称为有毒言论。尽管近年来社交媒体平台在纯文本的有害言论识别管控方面已经取得了较大的进展，并通过严格的监管措施有效减少了显性有害内容（如仇恨言论）的传播，但在多模态领域，尤其是对微攻击有害视频（如针对弱势群体的居高临下言论）的监控和管理仍然有限，进而造成了新的安全问题。这类隐性有害视频内容同样对网络环境和社会安全产生了深远的影响，是需要重点关注的一个危害子类别。

居高临下言论是针对弱势群体的歧视性有害（毒性）言论，通常表现为对这些群体的优越态度。尽管如前置章节所述，已有一些研究致力于居高临下语料库的构建，并通过深度学习网络对相关问题进行探索，同时本文的三、四章节也为该领域数据和模型组的构建做出了诸多努力，但目前针对于居高临下言论的检测仍主要依赖于基于文本的模式，而居高临下言论的边缘特征表明这种危害性更多的通过说话者的歧视性面部表情变化而非语言来传递至弱势群体社区。因而对视频、语音等多模态融合的忽视严重制约了居高临下检测的进一步发展。尽管在仇恨言论视频检测中已经陆续开始出现多模态框架，但有关居高临下等隐性有害内容的多模态方向仍未被充分探索。此外，目前的研究仅限于英语，而对其他语言社区中的弱势群体关注较少。因此，发展居高临下言论的多模态检测框架可以同时促进隐性有害言论和多模态检测两个领域的协同发展，具有极其重要的潜在研究价值。

在本章节中，我们介绍了第一个多模态居高临下数据集和相应的检测器，对加强视频平台的自动化微攻击监管，尤其是对弱势社区的保护提出了一套完整的解决方案。我们首先构建了 PCLMM 数据集，这是首个用于检测视频中 PCL 的多模态数据集，包含 715 个高质量注释视频，总时长超过 21 小时，来自中国最大的视频社区平台之一——Bilibili，检索范围涵盖几乎所有的中文弱势群体类别。我们将我们的数据集公开，以便于进一步社会研究。我们还提出了 MultiPCL 检测器，该检测器将面部表情特征与视频、文本和音频结合，增强了歧视性语言的检测。对于表情模态，我们锁定视频中的表情帧并捕获表情特征，结合视频，文字，语音模态，我们的融合模型获得了最优的基线测试结果，展示了多模态结合对于居高临下检测的重要价值。我们对于标记的居高临下表情帧跨度进行了系统分析，我们发现当居高临下的表情特征表

现为“积极”时，其具备的毒性强度和“强烈歧视”是相当的，这进一步指出居高临下言论作为隐性有害言论的子集，哪怕是非消极的标签特征，也会对于弱势群体社区构成严重伤害。

我们的贡献总结如下：

（1）我们开发并发布了 PCLMM，这是首个多模态居高临下数据集，包括 715 个 Bilibili 视频（21+ 小时），标注为居高临下（PCL）或非居高临下（非 PCL），并标注了 PCL 面部帧范围。

（2）我们介绍了 MultiPCL 检测器，该检测器整合了面部表情、视频、文本和音频四种模态特征，显著提高了检测准确性。

（3）我们的情感和危害性分析表明，PCL 具有一定的模糊性，而我们的检测器能够进一步有效识别这些边缘特征。

5.2 总体设计

本章的主要设计架构如图 5.1 所示。我们设计的框架是一个全面的、逐步推进的多模态检测范式。我们收集、标注 PCLMM 数据集，使用创新性的方法捕捉居高临下的歧视性面部表情特征，并单独设计表情维度的特征提取和融合模块，将其作为重要的组成部分参与后续我们的多模态融合的工作。我们的多模态融合包括原始视频模态、视频中的音频模态、由音频转录的文本模态以及表情模态。我们将在下文详细阐述各个部分的工作。

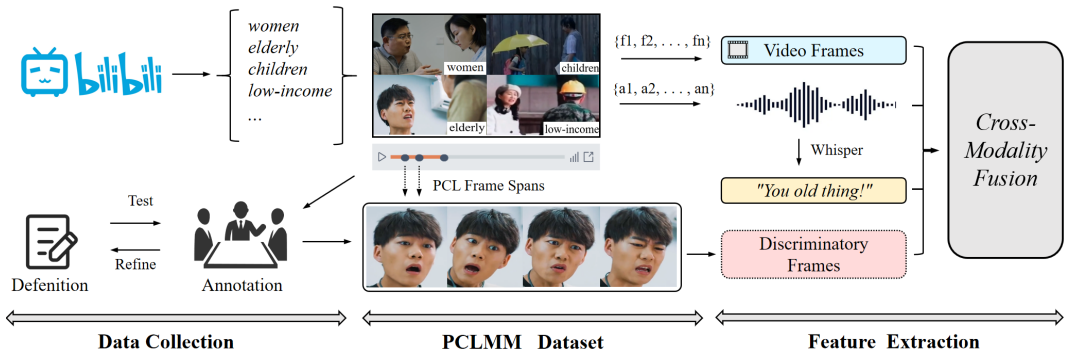

图 5.1 多模态居高临下言论检测的总体框架图

Fig. 5.1 Overall framework for our multimodal PCL detection

5.3 PCLMM 居高临下视频数据集

5.3.1 PCLMM 设计总览

在本节中，我们概述了 PCLMM 数据集的构建过程。我们开发了一个全面的 PCL 中文语义标准，以制定标注指南。通过使用来自中文互联网的六个关键易受攻击群体

类别，我们编制了关键词列表，并通过定向搜索收集了视频。数据集由三名标注员手工标注，随后进行了情感和危害性分析。

5.3.2 中文多模态居高临下言论的标准化定义

居高临下言论通常针对弱势群体，但英文的既定定义在部分语境下与中文互联网的背景并不一致，这是因为我国针对弱势群体的细分类别与英语国家存在一定差异。例如，由于政策和文化原因，我国对外来移民的居高临下言论展示很少。同时，对于多模态模式下的居高临下言论需要进一步的进行概念建立和阐述。本章节中，我们基于先前的工作[1, 68]并进一步加以完善，提出了一种适应中文语义背景的居高临下全面定义，同时融合了多模态检测的知识，作为我们的多模态中文居高临下言论标注指南。

中文多模态居高临下言论指的是对中国社区内六类弱势群体进行的歧视性、虚伪同情和伪善的言论。这些群体包括残疾人、女性、老年人、儿童、单亲家庭和低收入群体。居高临下言论的一个关键特征是说话者居高临下的态度，他们的言论并未根本上改善弱势群体的处境。伴随言语歧视，居高临下言论的表达通常含有轻蔑和歧视性的面部表情和肢体动作。为了减少主观差异，我们明确规定以下情况不应标注为 PCL：

（1）弱势群体描述他们自身遭遇的不公正待遇。

（2）关于歧视事件的客观新闻报道。

（3）含有歧视性内容但缺乏歧视意图的宣传内容，如公益广告。

5.3.3 数据收集

根据进一步的标准化定义，我们在中国互联网中进一步凝练了六类主要的弱势群体。我们找出描述每类群体最常用的 10 个常用关键词，并设计了一个包含攻击性和歧视性术语的词典作为查询键。这些术语查询键与关键词列表进行匹配，生成最终的搜索集（例如，将查询键“歧视”添加到“空巢老人”这一关键词中）。我们的搜索列表包含 1800 个关键词-值对，检索出了 2654 个初步视频。我们保留了时长为 30 秒到 5 分钟的视频，过滤掉了内容损坏和不相关的视频，并隐藏了所有可能暴露用户隐私的水印。最终，我们得到了 715 个高质量的可标注样本。

5.3.4 数据标注

两位经过训练的博士生对视频进行了标注，第三位作为审核员（两位男性，一位女性，年龄在 25 至 30 岁之间，均为计算机科学专业，专注于有害言论检测）。视频根据中文多模态居高临下言论的标准化定义被标注为 PCL 或非 PCL。初步使用 30 个视频（其中 20 个为非 PCL，10 个为 PCL）来达成共识，解决标注中的分歧。由于长期观看居高临下的视频具有负面的毒性危害，为了减少心理影响，标注者每天的标注数量限制为 20 个视频，并报告其心理状态。本章节使用 CVAT 工具进一步记录了 PCL

视频的所有歧视性面部表情帧，而非 PCL 视频则没有居高临下的面部表情帧。Fleiss' Kappa[65] 用于衡量标注者之间的一致性，IAA 值为 0.72，对于分歧严重的视频由第三位标注者进行人工校对。最终，我们获得了 196 个 PCL 视频和 519 个非 PCL 视频。

PCLMM 数据集包含 715 个视频，总时长 21 小时，平均视频长度为 1.80 分钟，帧率为 30 帧每秒，总计 2.3 百万帧。约 27.4% 的视频被标注为居高临下的（PCL），这一比例与互联网上平台的 PCL 数据分布相一致。详细的数据集统计信息见表 5.1。

表 5.1　PCLMM 数据统计
Tab. 5.1　Statistics of PCLMM

	非居高临下	居高临下	PCL 歧视性帧	总计
总数	519	196	330	715
总时长 (hrs)	15.1	6.5	2.3	21.6
总帧数 (M)	1.6	0.7	0.2	2.3
平均视频长度 (min)	1.7	1.9	0.4	1.8
平均字符数 (char)	455	536	158	477

5.3.5　数据情感分析

我们使用先进的开源模型 DeepFace[75] 对 PCLMM 数据集中的面部表情进行了分析。我们从 PCL 和非 PCL 子集的每个群体中各随机抽取了 20 个视频，总计 240 个样本。对于 PCL，表情采样来自已标注的 PCL 片段，每段选取 10 帧；对于 non-PCL，则从一般视频帧中选取 10 个表情。数据的情感分析结果如图 5.2 所示，我们可以观察到非 PCL 标签中的表情主要表现为积极或中性（浅色系维度），而 PCL 中的表情则常常传达出愤怒、悲伤、厌恶等负面情绪（深色系维度）。然而，也有一些 PCL 片段被分类为“快乐”或“中性”，尽管它们表达了歧视或讽刺的态度。这表明 PCL 并不完全等同于负面情绪，而虽然整体上居高临下言论延续了仇恨言论等显性有害言论的情感倾向，但是模糊样例明显增多，这一现象也揭示了情感分类在检测居高临下言论时的巨大局限性。

5.3.6　数据危害性分析

我们使用 Perspective API 对转写文本进行评分，详细结果如图 5.3 所示。橙色柱状图表示居高临下样本的平均毒性得分，而蓝色柱状图表示非居高临下样本的平均毒性得分。观察得到，居高临下样本在所有社区类别中的毒性评分均高于非居高临下样本（0.37 vs 0.24）。尽管如此，居高临下的毒性程度仍远低于传统仇恨言论（仇恨言论的正样本平均毒性得分通常高于 0.7），这凸显了居高临下言论的隐性危害特点以及准确识别所面临的挑战。

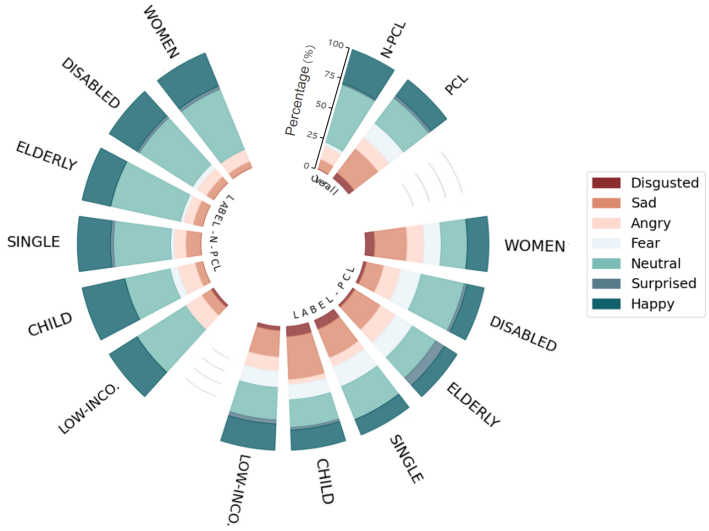

图 5.2　PCLMM 中六类弱势群体的情感分析

Fig. 5.2　Sentiment analysis for the six vulnerable groups in PCLMM

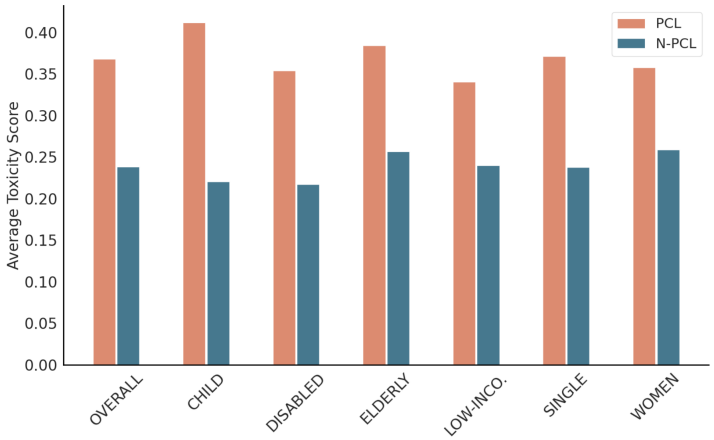

图 5.3　PCLMM 中样本的平均毒性得分

Fig. 5.3　Average toxicity scores in PCLMM

5.4　多模态架构 MultiPCL 设计

5.4.1　问题描述

在得到数据集后，我们着手建立我们的多模态检测框架，首先进行任务的定义。给定一个视频样本集合 V，任务是将针对弱势群体的视频分类为居高临下（$y = 1$）或非居高临下（$y = 0$）。每个视频 V 表示为一系列帧的序列 $F = \{f_1, f_2, \ldots, f_n\}$ 以及一组面部表情帧的子集 $F_v = \{f_{1v}, f_{2v}, \ldots, f_{nv}\}$。若某帧 f_n 不包含面部表情，则对应的 f_{nv} 用零向量填充。音频序列表示为 $A = \{a_1, a_2, \ldots, a_l\}$，转写文本序列表示为 $T = \{w_1, w_2, \ldots, w_m\}$。我们的目标是构建一个基于注意力机制的多模态分类器 $X : X(F; F_v; A; T) \rightarrow y$，其中 $y \in \{0, 1\}$，即最终的任务是多模态二元分类任务。

5.4.2 视频编码

我们使用 Vision Transformer（ViT）[59] 从视频中提取特征。给定一系列帧的序列 $F = \{f_1, f_2, \ldots, f_n\}$，ViT 为每一帧 f_i 提取对应的特征向量。特征向量 $\mathbf{z}_i$ 的计算方式如下：

$$\mathbf{z}_i = \text{ViT}(f_i), \quad \mathbf{z}_i \in \mathbb{R}^{d_v}, \quad i = 1, 2, \ldots, n \tag{5.1}$$

其中，$\mathbf{z}_i \in \mathbb{R}^{d_v}$ 表示 ViT 为每一帧 f_i 编码得到的 d_v 维特征向量。

5.4.3 面部表情编码

为捕捉视频中的居高临下面部表情，我们首先使用 MTCNN（Multi-task Cascaded Convolutional Networks）[76] 进行人脸检测。随后，使用 FER-VT（Facial Expression Recognition using Vision Transformers）[77] 通过基于网格的注意力机制和视觉 Transformer 编码面部特征，从而建模长距离依赖关系。对于每一帧视频帧 f_i，若 MTCNN 检测到人脸，则由 FER-VT 提取面部特征向量 $\mathbf{z}_i^v$；否则分配一个零向量：

$$\mathbf{z}_i^v \begin{cases} \text{FER-VT}(f_i v), & \text{若 MTCNN 在} f_i \text{ 中检测到人脸} \\ f_i v = 0, & \text{若未检测到人脸} \end{cases} \tag{5.2}$$

5.4.4 音频编码

我们使用广泛应用的多媒体处理工具 FFmpeg [78] 从视频中提取高质量音频，随后应用梅尔频率倒谱系数（Mel Frequency Cepstral Coefficient, MFCC）[79] 提取音频特征。提取得到的音频序列 A 被编码为 $\mathbf{z}^a$。

5.4.5 文本编码

我们使用 OpenAI 提出的语音识别模型 Whisper [80] 将音频转录为文本。在文本编码方面，我们采用了 RoBERTa-Chinese [81]，以及在 CCPC 数据集上微调的 RoBERTa 模型（BERT-PCL）[68]，用于识别居高临下语言。这些模型从每段转写文本中提取 CLS 标记，生成特征向量 $\mathbf{z}^t$。

5.4.6 跨模态特征融合

在我们的模型中，采用统一的跨模态多头注意力机制（Cross-Modality Multi-Head Attention, MHCA）来融合不同模态之间的信息。MHCA 的通用形式如下：

$$\text{MHCA}(\mathbf{Q}_i, \mathbf{K}_j, \mathbf{V}_j) = \text{Softmax}\left(\frac{\mathbf{Q}_i \mathbf{K}_j^\top}{\sqrt{d_k}}\right)\mathbf{V}_j \tag{5.3}$$

其中，$\mathbf{Q}_i$ 表示来自模态 i 的查询向量，$\mathbf{K}_j$ 和 $\mathbf{V}_j$ 分别表示来自模态 j 的键和值。通过变换 i 和 j，即可表示不同模态对之间的交互关系：

$$\mathbf{A}_{i,j} = \text{MHCA}(\mathbf{Q}_i, \mathbf{K}_j, \mathbf{V}_j), \quad i,j \in \{\mathbf{z}, \mathbf{z}^v, \mathbf{z}^a, \mathbf{z}^t\} \tag{5.4}$$

最终得到的注意力特征被聚合为统一的多模态表示：

$$\mathbf{Z} = \sum_{i,j} \mathbf{A}_{i,j} \tag{5.5}$$

5.4.7 损失函数

我们采用 BCEWithLogitsLoss 作为损失函数，该函数适用于二分类任务。损失的计算方式如下：

$$\text{Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log \sigma(x_i) + (1-y_i)\log(1-\sigma(x_i))] \tag{5.6}$$

其中，$\sigma(x_i)$ 表示模型原始输出 x_i 的 Sigmoid 函数，y_i 为真实标签。

5.5 实验结果与分析

5.5.1 基线设计

我们的实验在两块 NVIDIA A800-80G GPU 上进行，采用五折交叉验证以确保训练的稳健性。我们训练了 20 个 epoch，取性能最好的前五次结果的平均值。训练过程中使用的批大小为 10，学习率为 1e-4。所有代码均基于 PyTorch 实现。评估指标包括 Precision、Recall、F1-score 和 Accuracy，这些都是有害言论检测中的标准指标。值得注意的是，ViT 架构在多模态模型中表现出较高的效率，适用于多模态融合任务。在短视频分析任务中，其性能可与 VideoMAE [82] 相媲美。因此，我们选择 ViT 作为模态融合的基线模型，而非 VideoMAE。

5.5.2 总体实验结果

我们采用逐步融合多模态的策略，从单一模态开始进行集成。详细实验结果如表 5.2-5.4 所示。

（1）在单模态场景下，文本模态的检测性能最高，而仅使用音频的效果较差，凸显了文本在有害言论检测中的持续重要性。详细数据如表 5.2 所示。

（2）在多模态场景中，引入视频模态通常能显著提升性能。在双模态设置中，包含视频模态的组合平均 F1 分数为 75.46，而不包含视频的组合仅为 66.56。这一趋势

表 5.2　PCL 视频分类任务中的模型表现（单模态）。缩写：MC（MFCC）、RC（RoBERTa-Chinese）、BP（BERT-PCL）、FT（FER-VT）、VM（VideoMAE）、VT（ViT）
Tab. 5.2　Model performance on the classification task of PCL videos (single modality)

M	Model	$\mathbf{P}_p$	$\mathbf{R}_p$	$\mathbf{F1}_p$	$\mathbf{F1}_m$	Acc
A	MC	35.81	56.89	45.21	54.28	64.14
T	RC	54.84	50.00	52.31	69.14	78.32
	BP	58.06	52.94	55.38	71.13	79.72
	GPT4	65.52	55.88	60.32	74.55	82.52
F	FT	65.52	47.50	55.07	70.46	78.47
V	VM	61.76	52.50	56.76	70.90	77.78
	VT	65.62	52.50	58.33	72.22	79.17

在三模态配置中同样明显，说明视频模态在特征理解中发挥着关键的辅助作用。此外，面部表情模态仅在与视频模态结合时才能展现出最佳性能。详细数据如表 5.3 所示。

表 5.3　PCL 视频分类任务中的模型表现（双模态融合）
Tab. 5.3　Model performance on the classification task of PCL videos (bimodal fusion)

M	Model	$\mathbf{P}_p$	$\mathbf{R}_p$	$\mathbf{F1}_p$	$\mathbf{F1}_m$	Acc
A+F	MC+FT	39.13	45.00	41.86	58.55	65.28
A+T	MC+BP	58.82	50.00	54.05	69.08	76.39
T+F	BP+FT	62.89	55.00	58.67	72.06	78.47
A+V	MC+VT	58.00	72.50	64.44	74.14	77.78
V+F	VT+FT	62.79	67.50	65.06	75.46	79.86
V+T	VT+BP	63.04	72.50	67.44	76.79	80.56

（3）我们提出的 MultiPCL 模型融合了四种模态，在性能上显著优于所有基线模型，分别比最好的单模态、双模态和三模态配置提升了 6.51%、4.27% 和 2.22%，进一步验证了我们检测器的有效性。详细数据如表 5.4 所示。

5.5.3　消融试验

我们还对 MultiPCL 检测器进行了消融实验（表 5.5），以验证 MHCA 机制的作用。实验结果显示，将 MHCA 替换为标准的全连接层会导致 F1 分数下降近 4%，这表明 MHCA 在捕捉不同模态之间关系方面具有关键作用。

表 5.4　PCL 视频分类任务中的模型表现（三模态及全模态融合）
Tab. 5.4　Model performance on the classification task of PCL videos (tri-modality and full fusion)

M	Model	$\mathbf{P}_p$	$\mathbf{R}_p$	$\mathbf{F1}_p$	$\mathbf{F1}_m$	Acc
A+T+F	MC+BP+FT	61.90	65.00	63.41	74.43	79.17
V+T+F	VT+BP+FT	64.44	72.50	68.24	77.47	81.25
V+T+A	VT+BP+MC	65.91	72.50	69.05	78.15	81.94
V+A+F	VT+MC+FT	67.44	72.50	69.88	78.84	82.64
V+A+T+F	**MultiPCL**	**68.09**	**80.00**	**73.56**	**81.06**	**84.03**

表 5.5　MHCA 机制的消融实验
Tab. 5.5　Ablation Study of MHCA

Model	$\mathbf{P}_p$	$\mathbf{R}_p$	$\mathbf{F1}_p$	$\mathbf{F1}_m$	Acc
MultiPCL	**68.09**	**80.00**	**73.56**	**81.06**	**84.03**
-MHCA	62.50	75.00	68.18	77.09	80.56

5.6　本章小结

居高临下言论是一类针对弱势群体的歧视性言论，广泛存在于社交媒体中。尽管前述章节已经有大量基于文本领域的工作，但其作为一种具有显著面部歧视性特征的语言，对于非文本模态的忽视导致了现有工作难以突破检测的模糊性困境，因此亟需更加全面的尤其是结合多模态工作的支持。为回应这一挑战，本章首先构建了 PCLMM，这是首个用于居高临下言论检测的多模态视频数据集，共包含 715 个高质量人工标注的视频，总时长超过 21 小时，涵盖全面的中文弱势群体社区。相比以往以英文文本为主的 PCL 数据资源，PCLMM 更加贴近真实的社交媒体使用场景，并拓展了研究范围至中文视频内容，具有重要的实证价值。

本章节同时提出了 MultiPCL 检测器，融合了视频特征、面部歧视性表情帧特征、文本以及音频信息，构建的跨模态特征综合建模在多个基线模型上取得最优表现，验证了多模态融合在识别隐性有害内容中的潜力。实验还发现，当居高临下表达以“积极情绪”呈现时，其有害（毒性）强度依然可能与“强烈歧视”持平，这说明即便缺乏传统负面情感标签，PCL 也可能对弱势群体造成实质性伤害。

本章节的研究为进一步在社交平台上自动化识别微攻击类视频内容提供了重要基础，并为隐性有害言论识别与多模态检测两个方向的交叉研究提供了新的路径。未

来工作将进一步探索居高临下语言与讽刺、刻板印象等微攻击类型之间的关系，并利用本章节提出的数据集与检测器作为评估基准，系统测评现有支持音频输入的多模态大模型在隐性有害言论检测中的表现。

6 结论与展望

6.1 结论

随着社交媒体的快速发展，网络言语暴力和隐性有害言论愈加频繁地出现在公众视野中，尤其是以伪善、虚假同情等形式表现的居高临下言论（Patronizing and Condescending Language, PCL），对弱势群体构成了深远影响。该类言论具有表达方式隐蔽、主观判断困难的特点，传统基于浅层语义和规则的方法难以捕捉其中的歧视性意图。此外，现有研究普遍存在三方面不足：其一，中文语料严重稀缺，英文语料也存在标注粗糙、歧义显著等问题；其二，传统预训练模型在应对复杂隐性攻击时表现不佳，缺乏对语义细节和社会语境的理解能力；其三，居高临下言论常常伴随非语言特征出现，然而现有研究大多仅聚焦文本模态，未能充分利用多模态特征进行辅助识别。

为应对上述挑战，本文系统开展了面向社交媒体的居高临下言论检测研究。本文从数据稀缺性出发，构建了首个中文居高临下数据集 CCPC，引入毒性强度融合以有效降低主观性分歧；本文从大模型时代对隐性有害特征的多维理解出发，提出 PclGPT 中英双语大语言模型，利用预训练-指令微调机制增强模型对隐性表达的泛化与解释能力；本文从多模态辅助歧视性言论识别出发，构建了 PCLMM-首个多模态视频数据集并设计 MultiPCL 多模态检测器，有效利用面部表情等非语言特征增益检测性能。本文的的相关工作显著推动了社交媒体中的隐性有害言论检测，特别是居高临下识别的研究进展。

6.2 创新点

（1）基于中文居高临下数据稀缺的问题，首次提出 CondescendCN 框架并构建中文居高临下语料库 CCPC，填补了中文社交平台（如微博、知乎）上居高临下言论研究的空白。引入三级标签体系和毒性强度（TS）标签，有效捕捉言论的毒性强度与主观性特征，为中英双语研究提供结构完备、覆盖面广的高质量语料支持。

（2）基于传统预训练模型识别隐性表达能力不足的瓶颈，开发首个面向居高临下言论检测的中英双语大型语言模型组 PclGPT。区别于通用预训练语言模型，PclGPT 明确针对隐性有害表达进行优化训练，采用预训练-指令微调双阶段策略，结合高效微调方法（Prefix Tuning 与 LoRA 组合），构建了 PclGPT-EN 与 PclGPT-CN 双语模型，展现出对伪善、同情性歧视等细粒度表达的强感知能力和跨平台迁移能力。

（3）基于居高临下言论中存在非语言歧视信号而现有方法仅关注文本模态的问题，构建首个多模态检测框架。基于 Bilibili 平台采集并标注 715 条高质量视频，构建 PCLMM 数据集，并在此基础上设计 MultiPCL 检测器，集成文本、视觉与面部表情特

征，实现对视频中复杂言论的立体式理解。该框架有效弥补文本模态的表达局限，首次实现在动态场景中对隐性歧视态度的识别，具备良好的扩展性和落地前景。

6.3 展望

本文在隐性有害言论-居高临下言论检测方面取得了一定的进展，然而还有一些研究方向需要在未来的工作中进一步探索。

（1）本文尝试了预训练-指令微调的大模型时代标准范式构建检测模型，然而这可能会导致过度调节（过拟合），进而诱导模型将不均衡数据无差别分类为高比例样本对应的标签。在未来的工作中，我们将尝试进一步引入基于人类反馈的偏好调优，以抑制当前的过度学习现象。

（2）本文聚焦于居高临下的理论层面研究，而居高临下言论的进一步工作有着非常广阔的应用前景，比如在目前多模态基线的基础上使用 API 进一步发展能对接互联网平台的检测模型，从而有效监测互联网中潜在的歧视性冷暴力（尤其是缺少监管的短视频领域），并及时提示相应弱势群体，减少隐性有害言论造成的伤害。

（3）居高临下言论需要更进一步的联合研究。对于和交叉领域的深度结合，尤其是和刻板印象、讽刺言论的共同研究将有助于我们对于隐性歧视性特征拥有根本上的理解，从而为我们健康的网络和舆论环境贡献更多的力量、创造更多的社会价值。

参 考 文 献

[1] Pérez-Almendros C, Anke L E, Schockaert S. Don’t Patronize Me! An Annotated Dataset with Patronizing and Condescending Language towards Vulnerable Communities [C]. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain (Online), 2020: 5891–5902.

[2] Wong G, Derthick A O, David E, et al. The what, the why, and the how: A review of racial microaggressions research in psychology [J]. Race and social problems, 2014, 6: 181–200.

[3] Huckin T. Textual silence and the discourse of homelessness [J]. Discourse & Society, 2002, 13 (3): 347–372.

[4] Zampieri M, Malmasi S, Nakov P, et al. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval) [C]. In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, Minnesota, USA, 2019: 75–86.

[5] Dixon L, Li J, Sorensen J, et al. Measuring and mitigating unintended bias in text classification [C]. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, New Orleans, Louisiana, USA, 2018: 67–73.

[6] Zhou J, Deng J, Mi F, et al. Towards identifying social bias in dialog systems: Framework, dataset, and benchmark [C]. In Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 2022: 3576–3591.

[7] Caselli T, Basile V, Mitrovic J, et al. HateBERT: Retraining BERT for Abusive Language Detection in English [C]. In Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Online, 2021: 17–25.

[8] Mathew B, Saha P, Yimam S M, et al. Hatexplain: A benchmark dataset for explainable hate speech detection [C]. In Proceedings of the AAAI conference on artificial intelligence, Online, 2021: 14867–14875.

[9] Drljača Margić B. Communication courtesy or condescension? Linguistic accommodation of native to non-native speakers of English [J]. Journal of English as a lingua franca, 2017, 6 (1): 29–55.

[10] Giles H, Fox S, Smith E. Patronizing the elderly: Intergenerational evaluations [J]. Research on Language and Social Interaction, 1993, 26 (2): 129–149.

[11] Huckin T. Critical discourse analysis and the discourse of condescension [J]. Discourse studies in composition, 2002, 155 (176): 00002–4.

[12] Komrad M S. A defence of medical paternalism: maximising patients’ autonomy. [J]. Journal of medical ethics, 1983, 9 (1): 38–44.

[13] Conroy N K, Rubin V L, Chen Y. Automatic deception detection: Methods for finding fake news [J]. Proceedings of the association for information science and technology, 2015, 52 (1): 1–4.

[14] Nakov P, Barrón-Cedeno A, Elsayed T, et al. Overview of the CLEF-2018 CheckThat! Lab on

automatic identification and verification of political claims [C]. In Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9, 2018: 372–387.

[15] Basile V, Bosco C, Fersini E, et al. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter [C]. In Proceedings of the 13th international workshop on semantic evaluation, Minneapolis, Minnesota, USA, 2019: 54–63.

[16] Derczynski L, Bontcheva K, Liakata M, et al. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours [C]. In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), Vancouver, Canada, 2017: 69–76.

[17] Xu J. Xu at SemEval-2022 Task 4: Pre-BERT Neural Network Methods vs Post-BERT RoBERTa Approach for Patronizing and Condescending Language Detection [C]. In Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), Seattle, United States, 2022: 479–484.

[18] Mendelsohn J, Tsvetkov Y, Jurafsky D. A framework for the computational linguistic analysis of dehumanization [J]. Frontiers in artificial intelligence, 2020, 3: 55.

[19] Wang Z, Potts C. TalkDown: A Corpus for Condescension Detection in Context [C]. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, 2019: 3711–3719.

[20] Perez-Almendros C, Schockaert S. Identifying condescending language: a tale of two distinct phenomena? [C]. In Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), Abu Dhabi, United Arab Emirates (Hybrid), 2022: 130–141.

[21] Singh S K, Kumar S, Mehra P S. Chat gpt & google bard ai: A review [C]. In 2023 International Conference on IoT, Communication and Automation Technology (ICICAT), Gorakhpur, India, 2023: 1–6.

[22] Achiam J, Adler S, Agarwal S, et al. Gpt-4 technical report [J]. arXiv preprint arXiv:2303.08774, 2023.

[23] Touvron H, Martin L, Stone K, et al. Llama 2: Open foundation and fine-tuned chat models [J]. arXiv preprint arXiv:2307.09288, 2023.

[24] Shaikh O, Zhang H, Held W, et al. On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning [C]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, 2023: 4454–4470.

[25] Wen J, Ke P, Sun H, et al. Unveiling the Implicit Toxicity in Large Language Models [C]. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, 2023: 1322–1338.

[26] Zhu Y, Zhang P, Haq E-U, et al. Can chatgpt reproduce human-generated labels? a study of social computing tasks [J]. arXiv preprint arXiv:2304.10145, 2023.

[27] Roy S, Harshvardhan A, Mukherjee A, et al. Probing LLMs for hate speech detection: strengths and vulnerabilities [C]. In Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, 2023: 6116–6128.

[28] Kiela D, Firooz H, Mohan A, et al. The hateful memes challenge: Detecting hate speech in multimodal memes [J]. Advances in neural information processing systems, 2020, 33: 2611–2624.

[29] Hürriyetoğlu A, Tanev H, Mutlu O, et al. Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2023): Workshop and Shared Task Report [C]. In Proceedings of the 6th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text, Varna, Bulgaria, 2023: 167–175.

[30] Thapa S, Rauniyar K, Jafri F A, et al. Extended multimodal hate speech event detection during russia-ukraine crisis-shared task at case 2024 [C]. In Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), St. Julians, Malta, 2024: 221–228.

[31] Ganguly A, Puspo S S C, Raihan M N, et al. MasonPerplexity at Multimodal Hate Speech Event Detection 2024: Hate Speech and Target Detection Using Transformer Ensembles [C]. In Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024), St. Julians, Malta, 2024: 125–131.

[32] Das M, Raj R, Saha P, et al. Hatemm: A multi-modal dataset for hate video classification [C]. In Proceedings of the International AAAI Conference on Web and Social Media, Limassol, Cyprus, 2023: 1014–1023.

[33] Maity K, Sangeetha P, Saha S, et al. ToxVidLM: A Multimodal Framework for Toxicity Detection in Code-Mixed Videos [C]. In Findings of the Association for Computational Linguistics ACL 2024, Bangkok, Thailand, 2024: 11130–11142.

[34] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J]. Advances in neural information processing systems, 2017, 30.

[35] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model. [C]. In Interspeech, Makuhari, Chiba, Japan, 2010: 1045–1048.

[36] Sak H, Senior A W, Beaufays F, et al. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. [C]. In Interspeech, Singapore, 2014: 338–342.

[37] Lei T, Zhang Y, Wang S I, et al. Simple Recurrent Units for Highly Parallelizable Recurrence [C]. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 2018: 4470–4481.

[38] Press O, Smith N A, Levy O. Improving Transformer Models by Reordering their Sublayers [C]. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020: 2996–3005.

[39] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. Journal of machine learning research, 2020, 21 (140): 1–67.

[40] Yang Z, Dai Z, Yang Y, et al. Xlnet: Generalized autoregressive pretraining for language understanding [J]. Advances in neural information processing systems, 2019, 32.

[41] He P, Liu X, Gao J, et al. Deberta: Decoding-enhanced bert with disentangled attention [J]. arXiv preprint arXiv:2006.03654, 2020.

[42] Devlin J, Chang M-W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [C]. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), Minneapolis, Minnesota, 2019: 4171–4186.

[43] Zhu Y, Kiros R, Zemel R, et al. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books [C]. In Proceedings of the IEEE international conference on computer vision, Santiago, Chile, 2015: 19–27.

[44] Denoyer L, Gallinari P. The wikipedia xml corpus [C]. In ACM SIGIR Forum, New York, NY, USA, 2006: 64–69.

[45] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training [J], 2018.

[46] Radford A, Wu J, Child R, et al. Language models are unsupervised multitask learners [J]. OpenAI blog, 2019, 1 (8): 9.

[47] 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.

[48] Baack S. A critical analysis of the largest source for generative ai training data: Common crawl [C]. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency, Rio de Janeiro, Brazil, 2024: 2199–2208.

[49] GLM T, Zeng A, Xu B, et al. Chatglm: A family of large language models from glm-130b to glm-4 all tools [J]. arXiv preprint arXiv:2406.12793, 2024.

[50] Bai J, Bai S, Chu Y, et al. Qwen technical report [J]. arXiv preprint arXiv:2309.16609, 2023.

[51] Stiennon N, Ouyang L, Wu J, et al. Learning to summarize with human feedback [J]. Advances in neural information processing systems, 2020, 33: 3008–3021.

[52] Team G, Georgiev P, Lei V I, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context [J]. arXiv preprint arXiv:2403.05530, 2024.

[53] Rafailov R, Sharma A, Mitchell E, et al. Direct preference optimization: Your language model is secretly a reward model [J]. Advances in Neural Information Processing Systems, 2023, 36: 53728–53741.

[54] Liu A, Feng B, Xue B, et al. Deepseek-v3 technical report [J]. arXiv preprint arXiv:2412.19437, 2024.

[55] Dong G, Yuan H, Lu K, et al. How Abilities in Large Language Models are Affected by Supervised Fine-tuning Data Composition [C]. In Proceedings of the 62nd Annual Meeting of the Association

for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 2024: 177–198.

[56] Wang Y, Kordi Y, Mishra S, et al. Self-Instruct: Aligning Language Models with Self-Generated Instructions [C]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, 2023: 13484–13508.

[57] Li X L, Liang P. Prefix-Tuning: Optimizing Continuous Prompts for Generation [C]. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Bangkok, Thailand, 2021: 4582–4597.

[58] Hu E J, Shen Y, Wallis P, et al. Lora: Low-rank adaptation of large language models. [J]. ICLR, 2022, 1 (2): 3.

[59] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [J]. arXiv preprint arXiv:2010.11929, 2020.

[60] Gheini M, Ren X, May J. Cross-Attention is All You Need: Adapting Pretrained Transformers for Machine Translation [C]. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Online and Punta Cana, Dominican Republic, 2021: 1754–1765.

[61] Li J, Li D, Savarese S, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models [C]. In International conference on machine learning, Honolulu, Hawaii, USA, 2023: 19730–19742.

[62] Wang P, Yang A, Men R, et al. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework (2022) [J]. arXiv preprint arXiv:2202.03052, 2022.

[63] Grandini M, SpA C, Bagli E, et al. METRICS FOR MULTI-CLASS CLASSIFICATION: AN OVERVIEW [J]. stat, 2020, 1050: 13.

[64] Lu J, Xu B, Zhang X, et al. Facilitating Fine-grained Detection of Chinese Toxic Language: Hierarchical Taxonomy, Resources, and Benchmarks [C]. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Toronto, Canada, 2023: 16235–16250.

[65] Fleiss J L. Measuring nominal scale agreement among many raters. [J]. Psychological bulletin, 1971, 76 (5): 378.

[66] Micikevicius P, Narang S, Alben J, et al. Mixed precision training [J]. arXiv preprint arXiv:1710.03740, 2017.

[67] Tian Y, Gan R, Song Y, et al. ChiMed-GPT: A Chinese Medical Large Language Model with Full Training Regime and Better Alignment to Human Preferences [C]. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 2024: 7156–7173.

[68] Wang H, Li M, Lu J, et al. Ccpc: A hierarchical chinese corpus for patronizing and condescending language detection [C]. In CCF International Conference on Natural Language Processing and Chinese Computing, Foshan, China, 2023: 640–652.

[69] Chiang W-L, Li Z, Lin Z, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality [J]. See https://vicuna. lmsys. org (accessed 14 April 2023), 2023, 2 (3): 6.

[70] Hosseini H, Kannan S, Zhang B, et al. Deceiving google's perspective api built for detecting toxic comments [J]. arXiv preprint arXiv:1702.08138, 2017.

[71] Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized bert pretraining approach [J]. arXiv preprint arXiv:1907.11692, 2019.

[72] Sun Z, Li X, Sun X, et al. ChineseBERT: Chinese Pretraining Enhanced by Glyph and Pinyin Information [C]. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Bangkok, Thailand, 2021: 2065–2075.

[73] Pires T, Schlinger E, Garrette D. How Multilingual is Multilingual BERT? [C]. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 2019: 4996–5001.

[74] Peng H-P. Exploring Symbolic effect of new media: The impact of Bilibili on Gen Z' s cohort identity and aesthetic choices in fashion [C]. In International Conference on Fashion communication: between tradition and future digital developments, Pisa, Italy, 2023: 176–187.

[75] Serengil S, Özpınar A. A benchmark of facial recognition pipelines and co-usability performances of modules [J]. Bilişim Teknolojileri Dergisi, 2024, 17 (2): 95–107.

[76] Zhang K, Zhang Z, Li Z, et al. Joint face detection and alignment using multitask cascaded convolutional networks [J]. IEEE signal processing letters, 2016, 23 (10): 1499–1503.

[77] Huang Q, Huang C, Wang X, et al. Facial expression recognition with grid-wise attention and visual transformer [J]. Information Sciences, 2021, 580: 35–54.

[78] Tomar S. Converting video formats with FFmpeg [J]. Linux journal, 2006, 2006 (146): 10.

[79] Davis S B, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences [J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1980, 28 (4): 357–366.

[80] Radford A, Kim J W, Xu T, et al. Robust speech recognition via large-scale weak supervision [C]. In International conference on machine learning, Honolulu, Hawaii, USA, 2023: 28492–28518.

[81] Cui Y, Che W, Liu T, et al. Revisiting Pre-Trained Models for Chinese Natural Language Processing [C]. In Findings of the Association for Computational Linguistics: EMNLP 2020, Online, 2020: 657–668.

[82] Tong Z, Song Y, Wang J, et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training [J]. Advances in neural information processing systems, 2022, 35: 10078–10093.

附录 A　第四章 PclGPT 模型的实验补充

A.1　Pcl-PT 数据集的详细构建

A.1.1　RAL-P 数据集

在转换 RAL-E 的过程中，我们使用了大语言模型构建了一个 PCL 字典。具体来说，我们让 LLM 基于置信度生成了超过 500 个最能反映居高临下语义的词汇，随后进行了人工验证。按置信度排序的部分词云信息如图 A.1 所示。对于 RAL-E 中不包含任何字典信息的句子，我们仅保留了 30% 作为非居高临下语料库，而所有包含字典信息的句子都被保留。原始文本语料库包含 1,476,472 个句子，过滤后的语料库包含 1,091,945 个句子，这些被用作 RAL-P 的预训练数据。

附录-图 A.1　居高临下字典的词云统计
App.Fig. A.1　Word cloud statistics of the condescending dictionary

A.1.2　WEB-C 数据集

我们从微博平台上统一收集了八个常见弱势群体的数据作为我们的 WEB-C 中文预训练语料库。标注团队为每个群体添加了 20 个最常用的搜索词，最终形成了搜索列表。社区类别的详细信息见表 A.1。在筛选过程中，我们删除了重复和不相关的样本（包括微博上的常见固定标签，如“# 话题内容”和“# 评论日期”），并将用户信息替换为 # 用户，以遵守社区隐私协议。我们保留了样本中的表情符号，并将其转换为平台指定的相应中文文本，以尽可能保留表情符号所传达的情感语义信息。

附录-表 A.1　WEB-C 中不同居高临下社区的最终收集数据
App.Tab. A.1　The final collection status of different PCL communities

Community	*Total*
# Disabled	38981
# Women	40256
# Elderly	39385
# Children	38475
# Single-parent	40689
# Ordinary People	37589
# Disadvantaged	40324
# Others	39375

A.2　Pcl-SFT 数据集的详细构建

我们采用了与附录 A.1 中描述的 WEB-C 相同的方法进行数据选择和过滤，并手动标注了高质量文本。本节提供了我们构建的 CPCL 数据集的标注和统计的详细描述。由于 PCL 言论的主观性，我们放弃了由 LLM 进行的自动标注方法，继续采用手动标注。我们招募了四名标注员，具有不同的性别、年龄和教育背景（两名主标注员和两名校对员）（50% 女性，50% 男性；年龄 25±5 岁；两名硕士学位持有者，两名博士学位持有者）。我们采用了 Wang 等人[68] 提出的标准，并在标注前对测试样本进行了详细培训，以确保标注员理解 PCL 的细微毒性差异。为了确保标注一致性，我们计算了二分类和多分类标注的 Kappa 标注员一致性（IAA）。IAA 结果如表 A.2 所示。如果忽略至少有一名标注员标注为低毒性强度的所有标注，IAA 值有所提高。这表明，毒性强度较弱的 PCL 具有更高的模糊性。

附录-表 A.2　CPCL 二分类和多分类标注的 Kappa IAA 得分
App.Tab. A.2　Kappa IAA scores of CPCL binary and multi-class annotations

Binary-classification	*Kappa IAA*
All labels	0.62
Remove Weak level	0.67
Multi-classification	*Kappa IAA*
Unbalanced Power Rel.	0.65
Spectators	0.54
Prejudice	0.61
Appeal	0.48
Sympathy	0.71

攻读硕士学位期间科研项目及科研成果

已发表论文

[1] **XXX**, XXX, XXX, et al. (题目隐去) [C]//CCF International Conference on Natural Language Processing and Chinese Computing. Cham: Springer Nature Switzerland, 2023: 640-652. (第一作者，**CCF-C** 类会议，本学位论文第三章)

[2] **XXX**, XXX, XXX, et al. (题目隐去) [C]//Findings of the Association for Computational Linguistics: EMNLP 2024. 2024: 6913-6928. (第一作者，**CCF-B** 类会议，本学位论文第四章)

[3] **XXX**, XXX, XXX, et al. (题目隐去) [C]//ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2025: 1-5. (第一作者，**CCF-B** 类会议，本学位论文第五章)

已录用论文

[1] XXX, XXX, XXX, XXX, XXX, XXX *. (题目隐去) (第二作者，**CCF-B** 类期刊)

参与科研项目

[1] 国家自然科学基金，细粒度仇恨言论检测 XXXXXX，No. XXXXXXXX，2024-2027.

[2] 国家自然科学基金，基于认知视角和语义表示的 XXXX，No. XXXXXXXX，2021.1-2024.12

获得奖励

[1] 2022-2023学年浪潮集团企业奖学金

致　谢

三年时光匆匆，转眼间又到了人生的又一个毕业季，回首三年，感慨万千，这三年有很多人在我的前进道路中为我提供了巨大的帮助，谨以此文向所有帮助过我和鼓励过我的人致以崇高的敬意。

最要感谢的是自己的父母和 XXX，受之父母的资助和鼓励，才让我能够有信心继续深造自己的学业，在我人生路遇诸多迷茫的方向时，是他们告诉我路在何方，正确的道路如何选择。人生路漫漫，却唯一没有回头的路，我知道最重要的就是自己选对要走的道路，并为之努力，我感谢父母对我三年以来的鼎立支持，感谢 XXX 一路以来的陪伴，能让我在漫长的迷茫之中仍向往未来的希望。

更要感谢的是我的导师 XXX 老师，在这三年里为我提供了无数的帮助。无论是初入实验室时提供的各种帮助和答疑解惑，还是逐渐成长的过程中对我的教诲，都让我明白了很多做人做学问的道理。老师在学问上有很深的造诣，但老师更是一个和蔼可亲的人，在老师身边会给人一种放心的安全感。从本科到现在，老师一直是我崇拜的人，我感谢老师三年来对我工作的认可以及对于未来选择的支持。我也非常感谢实验室的 XXX 老师、XXX 老师、XXX 老师、XXX 老师、XXX 老师等其他所有老师在学术上对于我的无私指导，感谢老师们的无私付出。

另外我要感谢实验室的各位师兄师姐。在初入实验室的过程中，是 XXX 师兄、XXX 师兄和 XXX 师姐引导我走上的科研之路，在这几年的工作中他们一直在我有困难的时候为我答疑解惑，让我感受到实验室家的温暖。在几年的大小工作中，XXX 师兄给我的帮助最多，他就像是大哥哥一样，走在最前面，用灯火照亮着其他所有人的前进之路。在平日的工作里，非常感谢 XXX 师兄、XXX 师兄为我提供的帮助，感谢自己的室友 XXX、XXX、XXX、XXX、XXX，以及老室友 XXX、XXX、XXX 在生活中的互相分享和共同进步，最后感谢各位同学和师弟师妹们，希望大家都前程似锦、都有光明的未来。