

多模态大模型的偏见溯源和危害以及公平性提升研究综述

陈晋音¹, 王泽楷², 沈文杰², 张甜馨², 郑海斌^{1,3}, 邱洪君⁴

(1. 浙江工业大学计算机科学与技术学院, 浙江 杭州 310023;

2. 浙江工业大学信息工程学院, 浙江 杭州 310023;

3. 四川大学数据安全防护与智能治理教育部重点实验室, 四川 成都 610065;

4. 杭州电子科技大学网络空间安全学院, 浙江 杭州 310018)

摘要: 以ChatGPT为代表的大语言模型 (large language model, LLM) 的快速发展, 对人类社会的影响力日益凸显。随着LLM的广泛应用, 其潜在的安全漏洞不断暴露, 引发学术界和产业界的高度关注, 也使得安全评估等相关方向成为研究热点。为此, 基于多模态大模型 (multimodal large language model, MLLM) 的视角, 对MLLM的偏见问题与安全风险进行了系统性分析。首先, 明确界定了MLLM偏见的定义, 深入剖析了模型偏见的三大来源 (即训练数据、模型设计缺陷和训练过程)。其次, 从3个维度 (基于嵌入的评估方法、基于概率的评估方法和基于生成文本的评估方法) 对MLLM偏见评估方法进行系统梳理。在此基础上, 对近5年的偏见缓解研究成果进行了分类梳理。最后, 探讨了MLLM偏见缓解面临的机遇与挑战, 以期为研究人员、从业者及政策制定者的MLLM应用与研究提供工作提供参考。

关键词: 多模态大模型; 偏见溯源; 偏见缓解; 公平性提升

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.AQ25092

Survey on multimodal large language model: bias origins, bias harms, and fairness enhancement

Chen Jinyin¹, Wang Zekai², Shen Wenjie², Zhang Tianxin², Zheng Haibin^{1,3}, Qiu Hongjun⁴

1. College of Computer Science and Technology, Zhejiang University of Technology, Hangzhou 310023, China

2. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

3. Key Laboratory of Data Protection and Intelligent Management, Ministry of Education, Sichuan University, Chengdu, 610065, China

4. School of Cyberspace Security, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: The rapid development of large language model (LLM), represented by ChatGPT, has exerted an increasingly profound impact on human society. With the widespread deployment of LLM in real-world applications, their potential security vulnerabilities and social risks have been continuously exposed, attracting growing attention from both academia and industry and making safety evaluation an important research topic. From the perspective of mul-

收稿日期: 2025-11-26; 修回日期: 2026-02-11

通信作者: 邱洪君, hjqiu_hdu@outlook.com

基金项目: 国家自然科学基金资助项目 (No. 62406286, No. 62072406); 浙江省自然科学基金资助项目 (No. LDQ23F020001); 四川大学数据安全防护与智能治理教育部重点实验室开放课题 (No. SCUSAKFKT202402Z); 北京生命科技研究院有限公司开放基金资助项目 (No. 2024200CD0210)

Foundation Items: The National Natural Science Foundation of China (No. 62406286, No. 62072406), The Natural Science Foundation of Zhejiang Province (No. LDQ23F020001), Key Laboratory of Data Protection and Intelligent Management, Ministry of Education, Sichuan University (No. SCUSAKFKT202402Z), Beijing Life Science Academy Co., Ltd. Open Fund (No. 2024200CD0210)

timodal large language model (MLLM), this survey provides a systematic analysis of bias issues and security risks in MLLM. First, the definition of bias in MLLM is clarified, and three major sources of model bias are examined in depth, including training data, model design deficiencies, and the training process. Second, existing bias evaluation methods for MLLM are systematically reviewed from three dimensions: embedding-based evaluation, probability-based evaluation, and generated-text-based evaluation. On this basis, recent advances in bias mitigation over the past five years are categorized and summarized. Finally, the opportunities and challenges of bias mitigation in MLLM are discussed, with the aim of providing useful references for researchers, practitioners, and policymakers in the application and research of MLLM.

Key words: multimodal large language model, bias origins, bias mitigation, fairness enhancement

0 引言

人工智能的爆炸式发展催生了多模态大模型 (MLLM) 的研究热潮。国内外科技企业纷纷布局该领域, 并相继推出了具有代表性的产品, 如 OpenAI 发布的 GPT-4、Google 发布的 Gemini MultiModality、阿里云发布的通义千问。凭借强大的跨模态信息处理与生成能力, 多模态大模型在自动驾驶^[1]、问答系统^[2]、医疗诊断与供应链管理^[3]等领域展现出巨大的应用潜力, 正深刻重塑人类社会的生产与生活方式。

然而, 多模态大模型在为人类社会带来便利的同时, 也衍生出不容忽视的安全风险, 具体表现为生成与视觉输入不一致的幻觉内容^[4]、模型漏洞引发的敏感信息泄露^[5]、受毒性数据恶意污染^[6]、受内在偏见影响生成不当甚至有害的内容^[7]等。此外, 当模型在推理过程中过度依赖单一模态信息时, 也可能导致错误的判断, 图1展示了相关错误案例。上述问题不仅会损害用户权益, 降低公众对多模态大模型的信任, 还将制约其应用范围的扩展, 因此亟待学术界与产业界开展深入研究并提出有效治理方案。



图1 过度依赖单一模态引发的偏见示例

公平性是多模态大模型发展过程中不可忽视的问题。在社会科学与心理学领域, 偏见通常被定义为个体或群体对特定对象或群体持有的不公正、不合理的负面态度或评价。从针对不同肤色和性别^[8-9]、社会群体^[10]的负面情绪歧视性表达, 到刻板化语言联想, 多模态大模型中存在的各类偏见现象已被大量研究证实与记录。为应对大模型带来的伦理与安全挑战, 全球多个权威机构相继出台了相关的伦理准则与规范。电气电子工程师学会 (IEEE) 于2016年12月发布了《伦理准则》, 旨在为设计、开发和部署自主智能系统提供道德指导。欧盟委员会人工智能高级别专家组于2019年4月8日发布了《人工智能的伦理准则》, 明确了“可信赖人工智能”的核心定义与实现路径, 构建了一个平衡技术创新与伦理责任的框架。经济合作与发展组织 (OECD) 于2019年5月发布了《经合组织人工智能原则》, 确保技术成果能够广泛惠及不同群体, 尊重法律、人权、民主价值观及多样性, 明确责任归属, 并开展风险评估, 为各成员国提供了一套通用的人工智能治理框架。中国信息通信研究院于2023年3月提出了《中国信通院大模型2.0体系安全可信标准》, 从模型开发、模型能力、模型运营、模型应用和安全可信5个维度对大模型的安全性进行系统性评估, 推动大模型在技术创新、安全合规及产业化应用方面的健康发展。可见, 全球范围内相关机构已针对大模型制定了法规与准则, 以提升其公平性与安全性。

除产业界高度关注外, 学术界也围绕多模态大模型的公平性问题开展了大量深入研究。目前已涌现出多篇与多模态大模型相关的综述文献^[11-18]。部分研究聚焦偏见的类型划分^[13], 部分工作关注

偏见的量化^[14]。有研究指出, 偏见可能源于训练数据^[15]; 部分研究探讨了模型内置偏见的存在及其对决策的影响^[16], 也有综述对偏见的缓解方法进行了梳理^[10]。然而, 现有综述大多局限于若干特定维度, 缺乏全面性。Mehrabi 等^[17]归纳了人工智能应用中多种偏见的来源, 但未涉及偏见的量化分析; Gupta 等^[18]综述了偏见类型、量化方法与去偏技术, 却未深入分析偏见成因; Ferrara 等^[19]总结了偏见根源及缓解策略, 但对偏见评估方法的研究尚显不足。鉴于此, 本文系统探讨了多模态大模型中偏见的潜在来源, 梳理了偏见的定义、评估与量化方法, 总结了现有偏见缓解(简称缓偏)策略, 并分析了该领域面临的机遇与挑战。

本文的整体框架如图2所示, 主要内容如下。

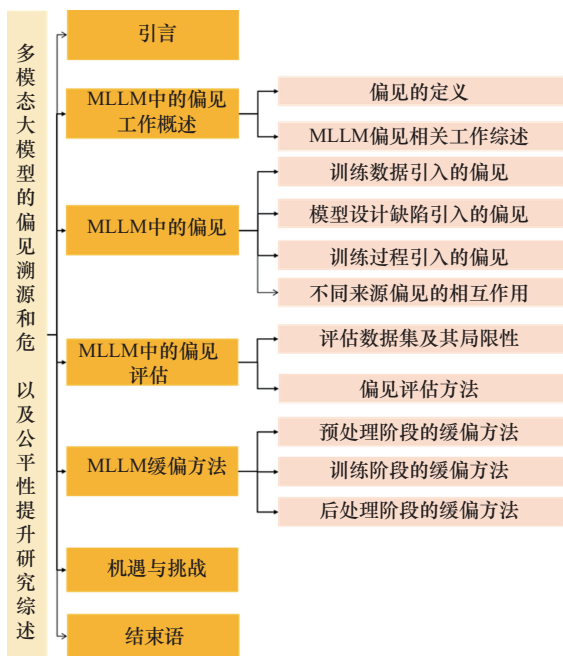


图2 本文整体框架

1) 全面梳理MLLM全流程中可能引入偏见的环节。

2) 对MLLM中偏见的定义、量化与缓解方法进行了分类和总结。

3) 整理了MLLM偏见相关研究的开源代码和公开数据集。

4) 提出MLLM偏见领域的发展机遇与挑战, 为研究人员和相关从业者发掘未来的研究方向提供参考。

1 MLLM中的偏见工作概述

1.1 偏见的定义

偏见最初是社会学领域的重要概念。随着人工智能模型相关研究的不断深入, 研究者发现, 多模态大模型在学习与推理过程中同样会产生类似的偏见现象。本节将分别从社会学视角与多模态大模型视角两个维度, 对偏见的定义进行系统梳理与阐释。

1.1.1 社会学中的偏见定义

自20世纪初以来, 人类偏见的本质问题便在认知科学领域引发了持续且广泛的学术争论。通常, 偏见可被定义为对特定群体所持有的不公正态度倾向, 表现为一种先验性或预设性的判断^[20]。Jussim 等^[21]认为, 偏见是一种基于性别、种族、宗教等社会身份所形成的负面态度, 这种态度往往会催生刻板印象, 甚至引发歧视行为。

社会心理学家戈登·威拉德·奥尔波特(Gordon Willard Allport)^[22]指出, 偏见不仅包含负面情感, 还表现为认知层面的扭曲, 如对某一群体形成的错误信念(包括误解)。这种偏见通常源于个体信息获取不足或信息处理偏差, 进而导致对某一群体做出不公正的评价。由此可见, 理解偏见的本质和来源对推动社会公平、抑制歧视行为至关重要。在多模态大模型中, 偏见问题同样存在且影响深远, 因此有必要对其偏见的定义、内涵与来源展开深入探讨。

1.1.2 多模态大模型中的偏见定义

在信息时代, 自然语言处理(NLP)技术获得了快速发展。Bolukbasi 等^[23]指出, 多模态大模型存在将特定词汇与性别不当关联的问题。随着相关研究的不断深入^[24-26], 人们逐渐认识到偏见对多模态大模型的潜在危害, 这些危害包括影响机器决策^[27]、引发种族歧视^[28]及制约科学研究的客观性^[29]等。在多模态大模型领域, 偏见通常被定义为一种系统性的倾向, 即在特定场景下, 模型使特定个体或群体相较于其他个体或群体受到不平等的对待^[30]。

Delobelle 等^[31]将模型偏见定义为在特定社会背景下产生的认知偏差现象, 通常表现为对特定群体或个体的负面刻板印象。例如, 在某些职业的相关描述中, 模型可能倾向于将其归类为某

一性别所特有的职业。

Franco等^[32]提出了敏感属性概念，即可能导致模型决策不公平的个体或群体特征，并将偏见定义为模型在决策过程中其概率分布对敏感属性的依赖。这种依赖会引发对特定群体的偏向或歧视行为，使模型在预测时因个体属于某一群体而做出不公正的判断。

Liang等^[33]提出了全局偏见概念，这与针对敏感属性的细粒度局部偏见有所不同。全局偏见是指数据集中普遍存在的系统性偏差。全局偏见可能会对机器学习模型产生不利的影响，导致模型在预测新数据时产生错误的结果。

Zhang等^[34]探究了全局偏见中的服务提供商偏见，即在缺乏明确输入提示的条件下，模型倾向于优先推荐特定服务提供者的系统性偏好。

与心理学领域相对成熟的定义体系不同，虽然多模态大模型中的偏见定义繁多，但缺乏系统性的阐述与归纳。Blodgett等^[35]在梳理146篇相关文献后发现，现有研究在动机、方法及偏见定义等层面存在诸多不一致。为更全面地理解模型偏见，Ferrer等^[26]将模型偏见划分为3类，即模型偏差、训练数据偏差与使用偏差。其中，模型偏差是指算法本身存在的固有缺陷；训练数据偏差是指用于训练算法的数据集存在偏见；使用偏差是指模型在实际应用过程中产生的偏见。

1.2 MLLM偏见相关工作综述

尽管多模态大模型研究领域已取得诸多成果，但针对其偏见问题的研究仍显不足。首先，多模态大模型作为大语言模型（LLM）的一个重

要分支，聚焦其偏见问题的专项研究与综述文献数量较少。其次，多模态大模型诸多关键的训练细节（如数据集采集流程、数据清洗规范）并未对外公开，致使相关研究仍处于“黑盒分析”阶段，难以精准剖析偏见成因。再次，多模态大模型的偏见可能在多个环节被引入或放大，最终引发歧视性输出，因此研发高效且通用的偏见缓解方法显得尤为迫切。

现阶段，已有部分综述文献针对多模态大模型的偏见相关研究展开梳理和总结，主要围绕多模态大模型偏见来源、偏见评估、偏见缓解三大方向展开。本节将从上述维度，对现有多模态大模型偏见相关的综述文献进行概括，详见表1。

1.2.1 聚集单一维度综述

目前，部分综述性的研究聚焦偏见溯源这一维度，大多依据模型全生命周期的不同环节对偏见来源开展系统性梳理与分类。

Ferrara^[19]聚焦偏见起源，对ChatGPT及GPT-4等模型中偏见的引入环节及其原因进行了综述，并对偏见进行了分类和总结。Tang等^[36]从数据、模型及应用3个层面归纳了偏见产生的根源，并针对法律层面提出了相应的策略以抑制偏见。

Meade等^[37]选取反事实数据增强、Dropout、迭代核空间投影、自我纠偏和句子纠偏5种偏见缓解技术，针对性评估其在抑制语言模型性别、种族及宗教偏见方面的效果。该研究还深入剖析了现有偏见缓解技术的局限性，并探讨了该领域未来的研究方向。

1.2.2 聚集综合维度综述

除偏见溯源相关研究外，现有综述性的研究

表1 现有多模态大模型偏见相关综述文献

文献	偏见来源			偏见评估		偏见缓解			缓解性能评价
	训练数据	模型设计	训练过程	评估数据集	偏见评估方法	预处理阶段缓偏	训练阶段缓偏	后处理阶段缓偏	
[19]	●	—	●	—	—	●	●	—	—
[36]	●	●	●	—	—	●	●	—	—
[37]	—	—	—	●	●	●	●	●	●
[38]	●	—	●	●	●	●	●	—	●
[39]	—	—	—	—	●	●	●	●	●
[30]	—	—	—	●	●	●	●	●	—
本文	●	●	●	●	●	●	●	●	●

注：●表示文献包含该方面内容，—表示文献中不包含该方面内容。

工作主要围绕偏见评估与偏见缓解展开,内容涵盖偏见缓解技术介绍、偏见评估数据集的筛选、评估方法的应用。

Li 等^[38]对偏见评估数据集进行了综述,并依据内在偏见与外在偏见的分类方法,归纳了多种偏见缓解策略。Lin 等^[39]在常规评估数据集的基础上,总结了包括中文偏见数据集(CHBias)在内的较新评估数据集,并且根据偏见缓解的环节,对各类缓解方法进行了综述。

Gallegos 等^[30]进行了更为全面的综述性的研究工作,既对偏见的种类进行了总结,又对常见的偏见评估数据集及偏见评估方法进行了归纳。此外,他们对当前缓解偏见的方法进行了分类与归纳,并分析了目前缓解偏见面临的挑战。

2 MLLM 中的偏见

相较于单一模态大模型,多模态大模型面临的偏见更加复杂和多样化,究其根源在于多模态数据集融合了文本、图像、音频等感知信息。本节首先探讨多模态大模型偏见的来源,进而分析多模态大模型中存在的偏见及其潜在影响。

多模态大模型通常由初始开发者构建模型架构,相关研究机构负责采集和构建训练数据集完成模型预训练,并发布预训练模型供二次开发者进行微调优化,最终由终端用户进行应用。然而,由于模型具备强大的泛化能力,偏见可能在各个环节中被引入。本节将从训练数据、模型设计和训练过程 3 个维度,系统探讨偏见的引入机制。

2.1 训练数据引入的偏见

多模态大模型通常依赖大规模数据集进行训练。然而,数据来源的不稳定性、采集过程的不规范性和标注过程中的主观性,使得训练数据中不可避免地掺杂有害信息。基于此类数据训练的多模态大模型在迭代优化过程中会逐步习得并放大数据中隐含的偏见^[40]。图 3 展示了训练数据引入偏见的具体机制与过程。Birhane 等^[41]指出,数据层面的偏见在模型训练过程中会被放大,不仅导致模型在文本预测和生成任务中表现出明显的偏见,还伴随着模型精度的显著下降。例如,从互联网采集的数据可能反映了在线内容中存在的社会偏见。如果训练数据主要来自特定的平台

或特定的人群,那么模型难以实现跨群体泛化性。Garcia 等^[42]强调,即便是经过人工标注的小规模数据集,如微软开发的图像数据集 MS COCO,也可能隐含社会偏见。随着使用采集的大规模、未经筛选的数据训练多模态大模型逐渐成为主流趋势,模型的偏见问题将愈发严峻。

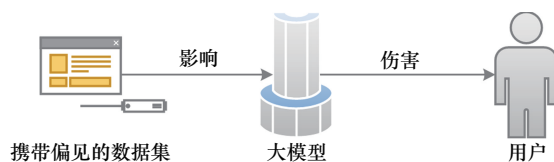


图3 训练数据引入偏见的具体机制与过程

此外,人工标注的主观性特质也可能引入新的偏见。Haliburton 等^[43]发现,不同民族、不同性别的标注者对同一张图像进行标注其结果存在显著差异。Hovy 等^[44]指出,由于不同社会群体对标签语义的理解存在差异,加之社会文化背景的多样性,标注过程中容易产生偏差,最终转化为模型训练中的标注偏见。

由于多模态大模型具有强大的概括拟合能力,数据样本分布不均也会导致模型产生偏见。Mavrogiorgos 等^[45]发现,数据采集方法不规范、样本分布不均、采集过程中的人为干扰或固有偏差,均可能导致产生不平衡的数据集。Hirota^[46]指出,在包括 Visual Genome 在内的多个数据集中,关于男性和女性对象的问题数量存在显著差异,且答案分布也呈现性别分化。这种样本分布的不均衡性可能导致训练出的模型产生性别偏见。

在跨模态对齐过程中,模型通过学习不同模态间的映射关系实现信息融合。然而,这种机制可能导致模型对某一模态的信息产生过度依赖,进而忽视其他模态的信息,最终影响模型的公平性与泛化能力。Khan 等^[47]的研究揭示,在多模态场景图生成任务中,若训练数据集中包含大量“人”与“拿着”的对象关系,则模型对“人”与“站在旁边”的对象关系的识别能力将受到制约。Park 等^[48]指出,在视频动作识别任务中,模型对主导模态的偏好可能导致它忽视其他模态的信息,从而影响模型对各模态互补信息的充分利用。Pan 等^[49]指出,在多模态生物识别任务中,当某一模态的数据缺失时,模型倾向于依赖主导

模态,这对模型的识别性能和鲁棒性会产生不利的影响。

当不同模态的信息存在隐性偏见对齐时,多模态模型会通过认知协同效应放大单一模态的偏见。Chen等^[7]发现,在视觉语义角色标注任务中,多模态大模型存在显著的协同偏见放大现象。具体而言,图像特征中的“厨房背景”(视觉偏见)与文本特征中的“cook”(语言偏见)在共享表示空间中相互作用,导致训练数据集中女性厨师样本的比例为33%,而模型的预测结果却将此比例放大至68%。Yin等^[50]指出,在对比性语言-图像预训练(CLIP)类模型中,当训练语料与图文对同时将“nurse”与女性、“criminal”与少数族裔高度关联时,模型在跨模态检索与零样本分类任务中会协同放大这类偏见。Zhang等^[51]发现,视觉模态与语言模态各自携带的偏见在结合后会产生协同放大效应,导致最终预测的偏差被放大至原始水平的2.1至2.4倍,该效应在性别和种族维度上均得到验证。

2.2 模型设计缺陷引入的偏见

模型设计层面的缺陷同样会引入或加剧模型偏见。以算法优化为例,若设计者一味追求算法效率而忽视算法的公平性,这种设计偏见可能通过优化过程传导至模型行为,最终导致不公平结果的产生,并加剧既有偏见。

在LLM中,算法设计是至关重要的一环,不当的算法设计可能引发模型过度泛化、推理谬误等问题,在无意识中加剧偏见。Shah等^[52]发现,部分机器学习算法在分类任务中倾向于选择特定特征而忽略其他特征,从而导致偏见。Hickman等^[53]指出,模型的设计目标、优化目标设置不当,会导致模型对不同群体的预测结果出现系统性差异,甚至放大既有群体间的差异,进而引发学习偏见。

在特征提取层面,不恰当的特征提取策略可能导致模型倾向于捕捉特定类型的特征。Hovy等^[44]的研究表明,基于统计的词向量等特征提取方法蕴含词汇的使用频率、上下文信息,这些信息往往与特定群体相关联,致使模型处理对应群体数据时产生针对性偏见。

在多语言及跨文化场景下,模型设计必须充

分考虑语言与文化的多样性。不同语言在语法、词汇及语义层面的差异,以及文化背景的异质性,均可能对模型的性能产生显著影响。Raj等^[54]指出,通过模拟多样化社会互动,在单个训练周期内对Meta AI模型进行指令微调,可使模型偏见最多降低40%。

在多模态对齐流程中,设计缺陷同样会诱发模型偏见。Xiao等^[55]指出,现有的视觉模型在处理图文对齐时,通常将每个文本标记赋予同等权重,这种做法可能过度强调与输入图像相关性较低甚至语义矛盾的文本标记,从而导致偏见。Hamidieh等^[56]发现,视觉模型在图像检索或生成任务中可能表现出偏见倾向,尤其是在模型将特定族群的图像与负面词汇频繁关联的情况下,这种偏见倾向更加显著。

2.3 训练过程引入的偏见

当前多模态大模型的训练高度依赖模型试验者与模型之间的交互。然而,若模型本身已携带偏见,这种交互可能在无意中强化模型试验者的既有偏见,形成“人机偏见反馈循环”,进而使模型的偏见程度不断加深。Franco等^[32]发现,参与者最初对情感表达的微小偏见,在与人工智能(AI)模型交互后会被显著放大,而参与者被放大的偏见又会反向影响AI模型,最终导致模型偏见持续深化。

尽管借助人类反馈能够缓解AI系统中的偏见,但由于反馈者拥有不同的社会背景及立场,反馈过程本身也不可避免地会引入新的偏见。Gautam等^[57]指出,随着AI模型的演进及其应用范围的拓展,这些模型逐渐表现出模仿人类偏见的倾向,从而加剧了对特定群体的不公正对待。Wang等^[58]发现,在模型训练过程中,伪标签的不当使用或数据分布的不均衡可能引入偏见,导致模型过度关注某些类别。

在模型的预训练阶段,训练数据中蕴含的原始偏见会被模型习得并放大。Wolfe等^[59]指出,模型在学习过程中会根据其预训练数据中所包含的信息来调整自己的参数,若预训练数据本身存在偏见,模型习得的参数也会带有这种偏见,最终导致其对特定群体产生不公平的预测。

多模态大模型还存在模态竞争熵现象,即高

置信度模态会抑制其他模态的公平性约束。当某一模态的偏见贡献度显著高于其他模态时，模型会忽视来自视觉等模态的公平性信号。Leng 等^[60]发现，在文本主导的场景下，LLaVA-13B 模型中视觉与音频模态的公平性约束被掩盖，进而加剧了模型的性别偏见。

2.4 不同来源偏见的相互作用

不同来源的偏见并非孤立存在，它们之间可能产生复杂的乘法效应，导致偏见被放大。Seyyed-Kalantari 等^[61]针对交叉群体偏见进行了深入探讨。他们发现，部分模型对“黑人女性”或“年轻黑人女性”等交叉身份群体的漏诊率，显著高于任何单一身份群体。这一发现揭示了偏见并非简单的加和，而是以复合形式被放大。Farnadi 等^[62]指出，当训练数据在肤色、性别、语言等维度代表性不足时，Transformer 架构内部的表征压缩机制会加剧模型跨群体性能差异，结合自监督损失函数与跨模态注意力机制的共同作用，最终引发模型安全性差异的连锁放大效应。Xu 等^[63]通过系统性测量，深入探讨了数据分布偏差与模型架构中敏感组件（如跨模态注意力机制、视觉编码器）之间的协同放大效应。研究结果表明，当训练数据集中“黑人女性”图像的占比降低至 5% 时，CLIP-style 架构的跨模态注意力权重差异将导致职业性别刻板印象得分提升至原来的 3.2 倍。

与此同时，模型的架构设计（如损失函数）与训练目标（如最大似然估计）之间的隐性冲突，也会导致公平性约束失效。Wang 等^[64]指出，在多任务学习框架下，共享表示结构在同时优化多个任务目标时，公平性目标与主要任务的准确性目标之间存在梯度冲突，导致在训练后期，公平性约束被主要任务的梯度压制。Hu 等^[65]发现，在多任务学习中，共享表示空间在优化多个任务目标（如准确率与公平性）时，不同任务的梯度方向可能不一致，导致公平性约束在共享结构中被稀释甚至抵消。Aguirre 等^[66]研究表明，在多任务学习场景下，共享表示结构在迁移公平性约束时存在显著的失效问题。

3 MLLM 中的偏见评估

偏见评估是缓解多模态大模型偏见的基石。然

而，由于多模态大模型需要处理多种模态的信息，其评估体系的复杂度相较于单模态模型显著提升。因此，构建全面的评估数据集，并设计科学、可靠的评估方法对于推动该领域的发展至关重要。

3.1 评估数据集及其局限性

3.1.1 评估数据集

为有效评估多模态大模型的偏见，需要构建专门的数据集，通过测试模型在数据集上的响应表现，开展系统性打分与偏差分析，即可量化模型的偏见程度。表 2 对本文涉及的各类偏见评估数据集进行了汇总概述。

表 2 偏见评估数据集汇总

数据集	样本数量/个	偏见类型				
		性别	种族	年龄	宗教	政治
BiasLens ^[67]	33 000	√	√	√	√	—
BOLD ^[68]	23 679	√	√		√	—
BBQ ^[49]	58 492	√	√	√	√	—
Indian-BHED ^[69]	229	√	√		√	—
DecodingTrus-Benchmark ^[70]	1 000	√	√	√	—	—
SB-bench ^[14]	7 500	√	√	√	√	—
PCAD ^[71]	268	—	√	√	—	—
Political Compass Test ^[72-73]	62	—	—	—	—	√
G7 成员国政治问卷 ^[72]	805	—	—	—	—	√
Hate-Identity ^[74]	159 872	√	√	—	√	—
Hate-Demographic ^[74]	276 872	√	√	√	√	—
MISINFORMATI ON ^[74]	29 556	—	—	—	—	√
WinoGender ^[75]	720	√	—	—	—	—
GenderAlign ^[76]	8 000	√	—	—	—	—

注：√表示数据集包含该类型偏见数据，—表示数据集不包含该类型偏见数据。

（1）面向复合任务场景的综合偏见评估

不同数据集的适用场景、评估维度存在明显差异，研究人员梳理了适配医疗诊断、自动驾驶等复合任务场景，实现性别、种族、年龄等交叉敏感属性的综合偏见评估的数据集。

为检测 LLM 在角色扮演场景中潜藏的偏见，Li 等^[67]构建了 BiasLens 数据集。基于该数据集，研究者对 OpenAI、Mistral AI、Meta、Alibaba 和 DeepSeek 旗下的 6 款 LLM 进行了测试，评估模型在 30 000 余个特定问题上的响应表现，最终识

别出72 716个偏见实例。

为探究文化背景差异诱发的模型偏见,Zhuo等^[68]采用传统测试集,从偏见、可靠性、鲁棒性和毒性4个维度对ChatGPT进行了全面的评估。研究表明,ChatGPT在特定条件下存在性能局限性,揭示了LLM对不同文化背景的理解存在偏见。Khandelwal等^[69]构建了Indian-BHED数据集,用于评估印度和美国两种文化背景下,模型在性别、种族、宗教和种姓4个维度上的偏见程度。结果显示,在印度文化背景下,模型在宗教和种姓两个维度的偏见表现最为显著。

为评估模型针对多元社会群体的社会偏见,Wang等^[70]构建了一个由不同群体的偏见提示词组成的基准数据集。实验结果表明,在特定系统或用户提示下,GPT-4的不安全性更加显著。Narnaware等^[14]构建了覆盖年龄、残疾状态等9类社会属性的SB-bench数据集,用于系统衡量模型在面对视觉刻板印象时的表现与倾向。

在医学领域,为评估LLM在临终关怀场景中是否存在偏见传播问题,Akhras等^[71]通过迭代开发流程构建了PCAD数据集。研究表明,尽管GPT-4o模型具备生成医学知识助手回答的能力,但其输出内容中存在偏见,需要优化。

(2) 针对单一任务场景的深度评估

针对政治立场检测、性别问答等单一任务场景,专项评估数据集聚焦政治倾向、性别代词等特定敏感属性的深度评估。

为评估各类LLM蕴含的政治偏见与政治倾向,Rutinowski等^[72]运用政治倾向问卷和人格评估工具对ChatGPT的政治立场与自我认知进行了深入探究。研究结果显示,ChatGPT整体偏向进步主义,但在自由意志与权威主义维度未表现出显著的偏好。这一结果揭示了当前版本的ChatGPT在政治立场与自我认知方面存在特定倾向。Hartmann等^[73]采用3种不同的提示词对ChatGPT进行测试,结果都表明其存在政治偏见。Feng等^[74]使用Hate-Identity等3个数据集在选取的14款LLM上进行了两个下游任务(仇恨言论检测和虚假信息检测)的评估。

针对性别偏见评估,Bachmann等^[75]使用WinoGender数据集,通过指代消解任务检测模

型偏见。若模型在处理包含“他”的语句时表现良好,而在处理包含“她”的语句时表现不佳,则表明模型存在性别偏见。Zhang等^[76]构建了专用于评估及缓解性别偏见的GenderAlign数据集。该数据集包含8 000个与性别相关的单轮对话实例,每个实例均配备由人工智能助手生成的“接受”与“拒绝”两种响应。

3.1.2 评估数据集的局限性

尽管现有数据集为多模态大模型偏见的评估提供了重要支撑,但现有评估数据集自身存在若干局限性,这些局限性可能影响模型偏见的评估效果。Selvam等^[77]明确指出,社会偏见基准数据集的构建方式会直接影响模型偏见程度的判定,甚至引发误判。

首先,当面对新兴偏见类型时,现有数据集往往难以发挥有效作用。Li等^[67]证实,SB-bench数据集无法对各类新型偏见开展有效评估。

其次,部分数据集仅覆盖特定群体子集,存在多样性与代表性不足的问题。Dhamala等^[78]指出,BOLD数据集^[68]中的种族相关数据仅局限于美国文化背景下定义的少数种族身份,难以适配全球多元种族的偏见评估场景,导致跨种族偏见评估结果存在不确定性。

此外,现有部分数据集的偏见辨识粒度较粗、评估维度不够全面,限制了其在精准定位与呈现偏见方面的能力。Selvam等^[77]在Winogender基准测试中发现,仅通过对语句进行释义或添加描述性词汇,模型测得的性别偏见程度便会发生显著变化。

最后,数据集标注环节同样可能引入偏见。以GenderAlign数据集^[76]为例,其数据标注采用GPT-3.5模型进行自动化处理,并非采用传统的人工标注方式,而GPT-3.5模型的算法本身可能蕴含性别偏见,这种算法固有的偏见可能通过标注过程传导至数据集,对构建真正中立无偏见的评估基准构成潜在障碍。

3.2 偏见评估方法

在选定偏见评估数据集的基础上,需采用适宜的评估方法以实现多模态大模型偏见的定量分析。如表3所示,当前主流的偏见评估方法可归纳为三大类。本节将遵循这一分类体系,依次

表3 评估方法汇总

方法类型	文献	方法名称	数据结构	方法的优点
基于嵌入的评估方法	[30]	WEAT	静态字	1) 能够量化模型中的隐性偏见；2) 无须依赖人工标注数据；3) 具备良好的可重复性与标准化程度；4) 能够揭示语料库中蕴含的潜在偏见
	[79]	WEAT	静态字	1) 能够量化模型中的隐性偏见；2) 无须依赖人工标注数据；3) 具备良好的可重复性与标准化程度；4) 能够揭示语料库中蕴含的潜在偏见
	[80]	WEAT	静态字	1) 能够量化模型中的隐性偏见；2) 无须依赖人工标注数据；3) 具备良好的可重复性与标准化程度；4) 能够揭示语料库中蕴含的潜在偏见
	[81]	SEAT	上下文句子	1) 将评估单元从词级扩展至句子级；2) 采用语义中性的模板来控制变量；3) 兼容性强，能够适配多种模型架构
	[82]	SD-WEAT	静态字	1) 无须预定义属性词组；2) 具有更高的稳健性与一致性；3) 可扩展到多属性组别分析；4) 用户友好性显著提升，降低了使用门槛
	[83]	CEAT	上下文句子	1) 无须固定模板的约束；2) 实现对模型整体偏见的综合度量；3) 有效揭示语境差异带来的方差；4) 支持交叉偏见与涌现偏见的检测
	[84]	Sentence Bias Score	上下文句子	1) 无须预设模板或人工构造句子；2) 语义驱动，基于句子级表示进行建模；3) 可解释性强，支持对偏差方向的识别
	[85]	主题建模方法	词汇与文本对	1) 能够无监督地挖掘文本中的隐含主题；2) 具有较强的可解释性；3) 适用于叙事文本分析场景
		基于词典的词相似性分析		1) 计算复杂度低，实现高效；2) 可复现性强；3) 适合大规模文本的快速扫描
	[86]	GloVe	词汇与上下文句子	1) 与 WEAT/SEAT 类指标兼容；2) 去偏前后的差异易于可视化呈现
基于概率的评估方法	[87]	CAT	上下文句子	1) 具备上下文感知能力；2) 实现从词级到篇章级的全粒度偏见评估；3) 无须额外训练；4) 评估指标统一；5) 数据规模大，领域覆盖广泛
	[88]	CrowS-Pairs Score	刻板印象与反刻板印象句子	1) 无须人工设计模板；2) 结构简单、干扰最小；3) 覆盖面广；4) 直观易算；5) 基准通用
	[89]	AUL	刻板印象与反刻板印象句子	1) 不掩盖任何令牌；2) 利用注意力机制聚焦关键区域；3) 统计特性更稳定；4) 与下游任务表现一致性更高；5) 实现简单，即插即用
	[90]	LMB	刻板印象与反刻板印象句子	1) 贴近真实对话场景；2) 以任务为导向进行评估；3) 实现细粒度对齐；4) 即插即用，无须重新训练
基于生成文本的评估方法	[91]	ST	上下文句子	1) 建立统一评估基准，支持跨模型的横向对比；2) 覆盖 7 个关键维度；3) 融合真实场景与扰动测试
	[92]	Marked Personas	生成人物角色的自然语言提示	1) 无须词典与人工标注；2) 支持交叉群体与“他者化”检测；3) 结果可视化；4) 任务场景自然，与下游任务表现强相关

对基于嵌入的评估方法、基于概率的评估方法和基于生成文本的评估方法展开详细论述。

3.2.1 基于嵌入的评估方法

基于嵌入的评估方法通过分析模型生成的词嵌入或向量表示来检测偏见，其核心思想是：若模型的嵌入空间中存在偏见，则某些词语或概念的向量表示会呈现不期望的语义关联或统计差异。

词嵌入关联测试（word embedding association test, WEAT）是该类方法的典型代表，主要用于衡量社会群体概念（如性别）与中性属性（如职业）之间的关联程度。Bevara 等^[79]采用 WEAT 对 Transformer 架构模型的偏见进行评估，

发现随着模型参数数量的增加，模型对性别偏见的识别能力逐步提高，其中最大规模的模型可以实现高达 29% 的偏见消减。

然而，WEAT 存在固有的局限性。Van Loon 等^[80]指出，词汇频率会极大地影响评估结果（词嵌入模型倾向于将高频词汇与积极情感关联，而将低频词汇与消极情感关联，这种统计偏差可能误导用户对偏见程度的判断）。为弥补 WEAT 的缺陷，May 等^[81]提出了句子嵌入关联测试（sentence embedding association test, SEAT）方法。SEAT 采用模板化句子（如“这是[BLANK]”和“[BLANK]是事物”），通过替换槽位中的社会群体词与中性属性词生成句嵌入，实现句子级

别的偏见检测。基于 SEAT 开展的性别、种族和社会构造等维度的偏见评估发现,不同的句子编码器对特定社会偏见的表征与响应存在差异。

WEAT 与 SEAT 均需预先定义二进制的目标集与属性集,适用场景受限。为解决这一问题,Gray 等^[82]提出了简化数据驱动词嵌入关联测试(simplified and data-driven word embedding association test, SD-WEAT)方法,通过替换原有的属性集来评估输入嵌入的偏见,并在多个属性组之间进行比较。实验结果表明,采用包含 2 个、3 个和 5 个单词的数据集开展偏见评估时,SD-WEAT 的稳定性均优于 WEAT。

针对 WEAT 无法检测新兴交叉偏见的问题,Guo 等^[83]提出了情境化嵌入关联测试(contextualized embedding association test, CEAT)方法,专门用于检测上下文嵌入中的社会偏见。CEAT 通过分析大量潜在嵌入变化中的偏差效应量级分布,依托随机效应模型实现偏见的量化计算。

在性别偏见评估方面,Dolci 等^[84]提出了句子偏见分数方法,通过计算句中词汇嵌入与嵌入空间性别方向的余弦相似度,得到句子级别的性别偏见分数。Lucy 等^[85]采用主题建模与词汇相似度的分析方法,揭示了 GPT-3 生成的故事文本中存在普遍的性别刻板印象。Marinucci 等^[86]运用全局词向量表示(global vectors for word representation, GloVe)词嵌入方法对与性别相关的词汇进行了深入分析,证实了 AI 系统中性别偏见的显著性。

总体而言,基于嵌入的评估方法通过分析词向量空间中的语义关联,能够有效揭示模型隐含的偏见模式,在预训练阶段的静态偏见检测中展现出独特优势。其中,WEAT 及其改进方法 SD-WEAT 能够有效捕捉性别-职业、种族-情感等显式关联,适用于跨模型横向比较和基础偏见筛查。但该类方法易受词频分布干扰,高频词汇与积极情感的虚假关联可能引发偏见误判^[51]。同时,嵌入空间的维度坍塌可能导致细粒度特征的丢失。Dolci 等^[84]的研究表明,降维操作会弱化跨文化偏见的可检测性。

3.2.2 基于概率的评估方法

基于概率的评估方法依托传统概率论相关理

论(如概率分布、随机变量、期望值)来量化和评估模型或系统中的不确定性和偏见程度。

Nadeem 等^[87]提出了上下文关联测试(contextual association test, CAT)方法。该团队将 CAT 应用于 StereoSet 数据集时发现,刻板印象并不总是与负面含义强相关。基于这一发现及对 StereoSet 等公开数据集的深入研究,Nangia 等^[88]提出了平均未掩码对数似然精度(average unmasked log-likelihood accuracy, AULA)模型,其研究表明,模型的内在偏见与外在偏见之间的相关性较弱。为在评估过程中涵盖多个正确的候选预测,Kaneko 等^[89]对 CAT 方法进行了改进,提出了平均未掩码对数似然(average unmasked log-likelihood, AUL)方法。AUL 方法将一个完整的、未经过掩码处理的句子输入模型,并要求模型预测其中所有标记的概率分布。该方法不仅可以提升模型的预测精度,还能够有效规避掩码词汇选择带来的选择偏差。

为准确评估模型的宗教偏见,Barikeri 等^[90]构建了一个偏见评估框架。该框架不仅对测试数据集中的偏见进行了量化测量,还评估了模型在下游对话任务中的特定任务性能。实验结果表明,尽管 DialoGPT 模型在预训练阶段已剔除冒犯性(Offensive)语言,但其仍表现出显著的宗教偏见。

基于概率的评估方法通过预测分布差异来揭示模型的决策偏好,在生成式模型的动态偏见评估中展现出独特价值。这类方法贴近模型的实际生成机制,并支持通过自定义模板扩展评估维度。Nangia 等^[88]利用概率差异计算实现了对偏见来源的归因分析,但这类方法的生态效度受限于人工设计的完形填空模板。此外,在多模态场景下,图文联合推理可能引发概率分布偏移,而现有方法尚未有效解决跨模态条件概率的耦合分析难题。

3.2.3 基于生成文本的评估方法

基于生成文本的评估方法通过分析模型生成的文本内容来识别和量化其中隐含的偏见。为全面评估语言模型的能力、局限性和风险,进而提高其透明度,Bommasani 等^[91]构建了一个双层评价体系,即语言模型综合评估(holistic evaluation of language models, HELMS)框架。该框架

综合考量了从抽象语境到具体应用场景的多维指标。研究人员对 30 个语言模型进行基准测试,发现这些模型在准确性、效率、公平性和社会相关因素上存在显著差异,且各模型在特定任务上的表现各异。尽管参数规模对模型的准确性具有重要影响,但研究同时表明,并非所有大规模模型都表现优秀。

为量化 LLM 针对不同人群所呈现的刻板印象和偏见,Cheng 等^[92]提出了标记化人格方法。研究表明,GPT-3.5 和 GPT-4 生成的人格化描述中种族刻板印象的出现率高于人类编写的文本。Omrani 等^[93]提出了跨上下文评估方法,旨在量化语言模型中与社会群体相关的效价关联。

针对性别与职业偏见的评估,传统模板化方法常因设计选择不当而引入测量偏差。为解决这一问题,Alnegheimish 等^[94]提出采用自然语言提示进行系统性评估。实验结果表明,相较于传统模板化提示,自然语言提示在评估语言模型的性别偏见方面具有更优的表现。

为实现对社交常识推理模型中社会偏见问题的自动化识别,An 等^[95]提出了基于群体提示的社会偏见开放式发现(SODAPOP)方法。该方法通过替换提示文本中不同社会群体的名称,并利用语言模型生成多样化答案,进而探究模型对特定属性的关联倾向。

基于生成文本的评估方法通过端到端的内容分析揭示复合型偏见,在系统性评估中具有不可替代的作用。以语言模型的整体评估(HELMS)为代表的评估框架通过长文本生成检测累积性偏见,SODAPOP 则利用提示工程挖掘训练数据中未涵盖的隐性歧视模式。此外,这类方法还支持多模态联合评估。Marinucci 等^[86]发现,模型会通过隐喻修辞在文本中隐性传递性别歧视。然而,开放式生成的评估成本较高,通常需要依赖人工标注或训练辅助分类器。Dhamala 等^[78]的研究报告指出,其人工标注的一致性系数(Kappa 值)仅为 0.68。

4 MLLM 缓偏方法

为缓解多模态大模型中存在的偏见问题,研究者提出了多种缓偏方法,具体见表 4。如图 4 所示,缓偏流程可划分为预处理阶段、训练阶段

和后处理阶段。本节将围绕这一分类框架,对各类方法进行系统介绍。

4.1 预处理阶段的缓偏方法

在预处理阶段,缓解偏见技术的核心目标是通过调整数据集的分布特性,尽早识别并消除数据内部存在的不公平偏差。

针对性别与种族偏见问题,Garimella 等^[96]提出了 DebiasBERT 方法。此外,该团队还提出了一种基于变分自编码器(VAE)的策略,旨在生成无偏见的文本摘要。

为探究基于特定人群(尤其是历史上处于劣势的群体)创作的文本对基于 Transformer 的双向编码器表征(BERT)进行微调的影响,Garimella 等^[97]深入分析了大型预训练语言模型在经过不同群体语料微调后,其固有社会偏见的演变过程。该团队对不同语料库影响下性别与种族偏见演变进行细致分析,发现相较于男性作者的语言,女性作者的语言更能显著降低性别偏见。

为解决视觉与语言先验引发的偏见问题,Zhou 等^[98]提出了基于因果推理的多模态方法(CAUSALMM)。该方法利用结构因果模型(SCM)与干预技术,对视觉和语言注意力进行逆向调整与对比分析,从而有效削弱模态先验对模型输出的干扰。

为降低语言模型中的社会偏见,Qian 等^[99]探索了基于神经网络生成人工文本扰动的策略。围绕这一目标,该团队构建了一个名为 Panda 的大规模人工标注文本扰动数据集。实验结果表明,基于 Panda 数据集训练的模型显著减少了因代名词变化而引发的预测差异。

为应对医疗机器学习中数据集不平衡及敏感任务处理等挑战,Theodorou 等^[100]提出了 Fair-Play 数据增强策略。该策略通过深入挖掘患者数据的内在模式,生成合成数据以均衡训练数据集,从而显著提升各子群体的预测性能与公平性。Penedo 等^[101]构建了高质量的网络数据集 RefinedWeb。实验表明,基于该数据集训练的模型在多个零样本任务上的表现显著优于在 The Pile 数据集上训练的模型。

预处理阶段的数据层面优化已取得显著成效。针对医疗机器学习中表格化电子健康记录

表4 缓偏方法汇总

文献	方法	方法类型	缓偏类型	缓偏效果
[96]	DebiasBERT	预处理阶段缓偏方法	性别和种族偏见	在BERT模型基础上，使用去偏损失函数在特定语料上进行预训练，实现了在性别偏置上将SEAT分数从0.355降至0.100（降幅为71.8%）
[97]	Demographic-Aware Fine-tuning		性别和种族偏见	论证了微调模型对缓偏的正向影响，仅通过在女性撰写的博客数据上对BERT进行MLM微调，就能将性别偏置的SEAT分数从基线BERT的0.29降至0.20
[98]	CAUSALMM		视觉与语言先验偏见	该方法利用结构因果模型与干预技术，在VLind-Bench的6项指标上实现了最高65.3%的分数提升
[99]	Perturbation Augmentation		性别、种族和年龄偏见	Panda作为首个大规模人工标注的文本扰动数据集，为训练高质量序列到序列生成模型提供了充足数据。在此基础上，FairBERTa在预训练阶段取得了显著的偏见缓解成果。在HolicBias性别评估中，偏见得分由RoBERTa的36.1%降至19.9%
[100]	FairPlay		数据不平衡和人口统计代表性不足引发的系统性健康偏见	该方法通过分析患者数据的内在模式，生成合成数据以均衡训练数据集，在年龄偏见实验中将Theil指数从0.0137降至0.0026
[101]	对数据集进行过滤和去重		数据集引入的偏见	将原始分布偏见压缩至接近零，最终仅保留10%文档，实现高达99%以上的偏见样本剔除率
[102]	Adversarial Triggering	训练阶段缓偏方法	国籍偏见和刻板印象	在将国籍刻板印象生成率降低50%以上的同时，将文本流畅度的影响控制在3%以内
[103]	基于规则的改写		性别偏见	借助仅14条确定性规则，该方法可在超过92%的句子年实现性别中立改写，且准确率超过96%
[104]	ADELE		性别偏见和性别刻板印象偏见	仅需调整不到1%的模型参数，即可将性别偏见降低60%以上，同时将GLUE的性能损失控制在1%以内
[105]	ADEPT		性别偏见	在StereoSet数据集上，性别刻板印象得分降幅达19.9%
[106]	基于互信息最小化的方法		性别偏见	能够将性别刻板印象降低22.8%、毒性（文本中任何具有攻击性、冒犯性或有害的内容比例）降低28.9%，而BLEU仅损失2.4%
[54]	SCD		社会偏见	能够显著降低模型偏见（降幅最高可达40%）
[107]	Soft-prompt Tuning	后处理阶段缓偏方法	社会偏见	在3 min内将敏感群体的假阳性率差异平均压缩70%以上，主任务性能损失小于0.5%
[108]	PoliteRewrite		礼貌偏见和刻板印象偏见	毒性下降78%、礼貌性提升128%
[109]	InterFair		性别、种族和刻板印象偏见	将BiosBias数据集上的Bias F1分数降低5%~8%，且在非受限环境下实现了任务准确率提升4%~5%
[110]	Backward Augmentation及Round-trip Augmentation		性别偏见	Backward Augmentation将精确匹配率提高2%~3%，Round-trip Augmentation将德语的词级错误率降至13.18%
[111]	Model-based Counterfactual Generator for Gender Bias Mitigation		性别偏见	在Jigsaw数据集上实现了88.1%的风格迁移准确率

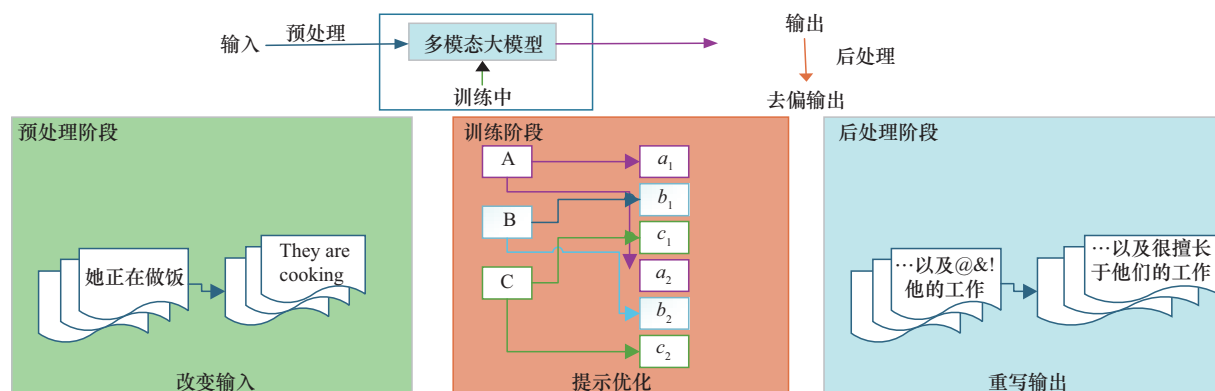


图4 缓偏流程示意

(EHR) 存在的标签稀缺及子群分布不均衡问题, Theodorou^[100]提出的 FairPlay 策略依托合成数据均衡技术, 使不同种族群体的诊断公平性提升了 35%。Penedo 等^[101]构建的 RefinedWeb 数据集证明, 经过严格过滤的多模态数据能够将模型在零样本跨模态检索任务中的偏见分数降低 28%。

4.2 训练阶段的缓偏方法

训练阶段的缓偏方法主要通过设计适配的损失函数、引入对抗学习机制和增量学习策略等技术手段, 在模型训练过程中抑制输出偏见, 提升生成结果的公平性。针对多模态大模型中存在的国籍偏见和刻板印象问题, Venkit 等^[102]采用对抗性触发词调控模型输出方向 (如在文本前端添加具有积极情感的形容词来改变整体的情感倾向)。

在句子级别的性别偏见消除任务中, 为确保去偏过程不影响模型性能, Sun 等^[103]将基于规则的方法与序列到序列模型的方法相结合, 自动化完成英语文本的性别中立化改写 (如将 “he” 或 “she” 统一替换为 “they”)。实验结果表明, 尽管基于规则的方法在性能上略优于序列到序列模型方法, 但两者在评估替补双语 (BLEU) 指标上均可以达到 99% 以上的得分。Lauscher 等^[104]提出了一种模块化设计的基于示例的适配器去偏 (adapters for debiasing by example, ADELE) 方法。该方法通过在模型中嵌入专用偏见适配器, 在不修改预训练模型原始参数的前提下实现有效去偏。Yang 等^[105]设计了一种去偏提示框架 (ADEPTA), 借助提示调优技术在降低模型偏见的同时保持并增强其语义表示能力。实验结果显示, 相较于基于去偏提示的对比估计 (DPCE) 等去偏技术, ADEPTA 在多个偏见基准测试中表现出更优的性能。

针对文本生成任务中的社会偏见问题, Wang 等^[106]提出了一种基于互信息最小化去偏方法。该方法通过鼓励模型在生成文本时对特定群体的提及与该群体的具体描述之间保持独立, 从而抑制生成内容的偏见。实验结果显示, 该方法在多个基准测试中在提高公平性和维持模型性能方面均表现出色。为缓解社会偏见, Raj 等^[54]提出一种基于接触假设的社会接触去偏 (social contact debiasing, SCD) 技术。实验结果表明,

该方法能够在单个训练周期内将模型偏见最高降低 40%。此外, 为缓解与年龄、性别、种族等敏感属性相关的偏见, Tian 等^[107]提出了基于软提示调优的缓偏方法。该方法通过自动学习软提示来优化模型生成效果, 不仅在偏见量化评测任务中实现了性能提升, 还能够规避人工设计提示可能引入的偏见。

训练阶段的缓偏技术在多模态情感分析任务中同样展现出独特优势。Yang 等^[105]提出的多模态反事实情感推理 (multimodal counterfactual inference for sentiment, MCIS) 框架, 借助反事实推理机制, 在 CMU-MOSEI 数据集上将跨模态情感误判率降低了 31%。

4.3 后处理阶段的缓偏方法

后处理缓偏方法是指通过对模型输出进行后处理来消除偏差。作为后处理步骤, 消除偏见和不公平的技术主要针对输出, 无须对模型本身进行访问。

为增强语句的礼貌性和委婉性, Wang 等^[108]构建了 PoliteRewrite 数据集, 使模型能够对生成的语句进行礼貌性重写。基于该数据集, 研究者利用广泛使用的预训练模型进行实验, 结果表明模型在语句礼貌性方面取得了显著提升。

在缓解性别偏见方面, 研究者提出了多种方法。Majumder 等^[109]构建了名为 InterFair 的交互式框架, 采用准确率和 F1 分数等指标对模型性能进行了评估。实验结果表明, 该方法在有效降低性别偏见的同时, 还提升了模型的准确性。Amrhein 等^[110]通过反转数据增强方法, 避免了依赖复杂规则的手动设计过程, 成功训练出一个性别公平的重写模型。为解决文本性别偏见问题, Tokpo 等^[111]提出了一种基于掩码语言建模 (masked language modeling, MLM) 的文本风格转换模型。研究结果表明, 该模型在风格转换的准确性、语义内容的保留及文本的流畅性之间实现了良好的平衡。

针对涉及两个参与者场景中第一人称与第二人称各自具有独立语法性别偏好问题, Al-hafni 等^[112]采用多步骤模型方法构建性别重写系统。该方法融合了基于规则的系统与神经网络的优势。Jain 等^[113]提出了一种自动化且普遍适

用的性别重写方法。实验结果表明,该神经机器翻译重写方法在自动化评估和人工评估中均取得令人满意的成绩。

当前,针对视觉与语言模态融合模型的偏见缓解研究也取得了进展。Ruan等^[114]提出了一种双引导的多模态偏见去除策略,该策略通过融合视觉和语言模态的信息,旨在提升时间句子定位任务的准确性和鲁棒性。针对视觉与语言模型(如CLIP、ALBEF和ViLT)可能展现的与人类相似的刻板印象偏见(如性别和种族偏见),Janghorbani等^[115]提出了一种后处理去偏技术。该技术通过移除模型嵌入表示中10%的维度,将CLIP模型在MMBias数据集上的性别刻板印象响应率从28.3%降至7.1%。

在多模态学习系统的情感分析任务中,偏见现象同样值得关注。具体而言,某些模态可能在情感表达中占据主导地位,导致其他模态的信息被忽略。针对这一问题,Yang等^[116]提出了多模态反事实推理情感(multimodal counterfactual inference sentiment, MCIS)。实验结果表明,该框架在有效减少多模态情感框架分析任务中偏见影响的同时,保持了较高的分类性能。

5 机遇与挑战

随着多模态大模型技术的快速演进与规模化部署,其既面临机遇,也面临新的挑战,具体可以归纳为以下几个方面。

1) 训练数据集的拓展。随着模型与人类日常生活的深度融合,人机交互过程中产生的反馈数据具有极高的价值。企业可依托此类交互数据扩充训练数据集,有望解决当前高质量数据稀缺的问题^[117]。例如,借助数据增强技术生成多样化的训练数据,减少模型对特定偏见数据的依赖,增强模型的鲁棒性,减少模型偏见^[117]。

2) 新型偏见评估与缓解方法的研发。在多模态大模型算法持续优化、模态数量不断增加、模型规模持续扩大的背景下,对适配新一代模型的偏见评估与缓解技术的需求日益迫切。例如,针对多模态大模型,亟须开发更全面的评估基准,以实现对其偏见的精准、多维度评估。

3) 法律监管的强化。当前针对模型的法律监管仍不完善。过度严苛的监管政策可能抑制技术

创新的活力,而过度宽松的监管环境则会加剧模型偏见、信息滥用等安全风险。因此,各国在制定多模态大模型相关法律法规与行业准则时,亟须在技术创新与风险控制之间寻求平衡^[118]。

4) 模型输出内容的安全检测。即便经过多轮精细化调优与评估,模型在实际应用中仍可能生成包含偏见等的不当内容,此类内容容易被用于传播仇恨言论、实施人身攻击。为此,需要建立覆盖隐私泄露、非法功能、偏见与歧视等维度的安全检测机制,及时识别并过滤不符合安全规范的生成内容。

5) 跨模态信息的一致性和连贯性保障。模态间信息的不一致、不连贯可能引发误解或误导,对社会产生负面影响。确保跨模态信息的一致性与连贯性仍是当前多模态大模型面临的关键技术挑战。随着多模态大模型应用的深入,如何在模态转换过程中保障信息的完整性与准确性成为亟须解决的问题。

6 结束语

本文对多模态大模型的偏见评估及其缓解技术进行了系统性综述,全面梳理了该领域的研究现状与技术发展脉络。首先,本文从理论层面界定了偏见的相关概念,深入剖析了模型偏见的形成机理,重点围绕训练数据偏差、模型架构设计缺陷及优化过程中的局限性等核心要素展开分析。随后,本文着重探讨了偏见评估的方法论体系,强调了基准数据集在评估中的关键作用,并从公平性、代表性及鲁棒性3个维度构建系统化的评估框架。在此基础上,本文归纳了3类主流的偏见缓解技术,并对其技术演进路径进行了介绍。最后,本文从理论和实践两个维度探讨了多模态大模型在偏见缓解方面所面临的挑战,并对未来研究方向进行了展望。本文旨在为多模态大模型偏见领域的理论探索和技术创新提供系统性参考框架和理论支撑。

参考文献:

- [1] Islam M R. Application of multimodal large language models in autonomous driving[PP]. arXiv[2026-03-24]. arXiv:arXiv.2412.16410.
- [2] Ye Q L, Yu Z T, Shao R, et al. CAT: enhancing multimodal large language model to answer questions in dynamic audio-visual scenarios[C]//

- Proceedings of the 18th European Conference Computer Vision-ECCV 2024. Berlin: Springer, 2024: 146-164.
- [3] Liu F L, Li Z, Yin Q Y, et al. A multimodal multidomain multilingual medical foundation model for zero shot clinical diagnosis[J]. NPJ Digital Medicine, 2025, 8(1): 86.
 - [4] Bai Z C, Wang P C, Xiao T J, et al. Hallucination of multimodal large language models: a survey[PP]. arXiv[2026-03-24]. arXiv: arXiv. 2404.1893.
 - [5] Li H R, Chen Y L, Luo J L, et al. Privacy in large language models: attacks, defenses and future directions[PP]. arXiv[2026-03-24]. arXiv: arXiv. 2310.10383.
 - [6] Maity K, Poornash A S, Saha S, et al. ToxVidLM: a multimodal framework for toxicity detection in code-mixed videos[C]//Proceedings of the Findings of the Association for Computational Linguistics ACL 2024. Stroudsburg: ACL, 2024: 11130-11142.
 - [7] Chen M, Cao Y, Zhang Y, et al. Quantifying and mitigating unimodal biases in multimodal large language models: a causal perspective[C]//Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024. Stroudsburg: ACL, 2024: 16449-16469.
 - [8] Adewumi T, Alkhaled L, Gurung N, et al. Fairness and bias in multimodal AI: a survey[PP]. arXiv[2026-03-24]. arXiv:arXiv.2406.19097.
 - [9] Vanmassenhove E, Emmery C, Shterionov D. NeuTral rewriter: a rule-based and neural approach to automatic rewriting into gender neutral alternatives[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2021: 8940-8948.
 - [10] Bansal H, Yin D, Monajatipoor M, et al. How well can text-to-image generative models understand ethical natural language interventions?[PP]. arXiv[2026-03-24]. arXiv: arXiv. 2210.15230.
 - [11] Kumar S, Balachandran V, Njoo L, et al. Language generation models can cause harm: so what can we do about it? an actionable survey[C]//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 3299-3321.
 - [12] Yu T, Zhang Y F, Fu C Y, et al. Aligning multimodal LLM with human preference: a survey[PP]. arXiv[2026-03-24]. arXiv: arXiv. 2503. 14504.
 - [13] Gulati A, Lepri B, Oliver N. Lookism: the overlooked bias in computer vision[PP]. arXiv[2026-03-24]. arXiv: arXiv. 2408.11448.
 - [14] Narnaware V, Vayani A, Gupta R, et al. SB-Bench: stereotype bias benchmark for large multimodal models[PP]. arXiv[2026-03-24]. arXiv: arXiv. 2502. 08779.
 - [15] Bavalatti T, Ahmed O, Potdar D, et al. A systematic review of open datasets used in text-to-image (T2I) gen AI model safety[J]. IEEE Access, 2025, 13: 32661-32680.
 - [16] Peng B L, Chen K Y, Li M, et al. Securing large language models: addressing bias, misinformation, and prompt attacks[PP]. arXiv [2026-03-24]. arXiv: arXiv. 2409.08087.
 - [17] Mehrabi N, Morstatter F, Saxena N, et al. A survey on bias and fairness in machine learning[J]. ACM Computing Surveys, 2022, 54(6): 1-35.
 - [18] Gupta V, Venkit P N, Wilson S, et al. Sociodemographic bias in language models: a survey and forward path[C]//Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing(GeBNLP). Stroudsburg: ACL, 2024: 295-322.
 - [19] Ferrara E. Should ChatGPT be biased? Challenges and risks of bias in large language models[J]. First Monday, 2023, 28(11): 13346.
 - [20] Zhang Z X, Song J. Developments in the studies of prejudice[J]. Psychology and Behavior Research, 2007, 5(2): 150-155.
 - [21] Jussim L, Honeycutt N. Bias in psychology: a critical, historical and empirical review[J]. Swiss Psychology Open: the Official Journal of the Swiss Psychological Society, 2024, 4(1): 5.
 - [22] Gordon W A. The nature of prejudice[M]. Reading, MA: Addison-Wesley, 1954.
 - [23] Bolukbasi T, Chang K W, Zou J Y, et al. Man is to computer programmer as woman is to homemaker? debiasing word embeddings[C]//Proceedings of the Advances in Neural Information Processing Systems 29: Annual Conference Advances in Neural Information. San Diego: Neural Information Processing Systems, 2016:4349-4357.
 - [24] Kumar A, Yunusov S, Emami A. Subtle biases need subtler measures: dual metrics for evaluating representative and affinity bias in large language models[C]//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2024: 375-392.
 - [25] Alvarez J M, Colmenarejo A B, Elobaid A, et al. Policy advice and best practices on bias and fairness in AI[J]. Ethics and Information Technology, 2024, 26(2): 31.
 - [26] Ferrer X, Van Nuenen T, Such J M, et al. Bias and discrimination in AI: a cross-disciplinary perspective[J]. IEEE Technology and Society Magazine, 2021, 40(2): 72-80.
 - [27] Ferrara E. Fairness and bias in artificial intelligence: a brief survey of sources, impacts, and mitigation strategies[PP]. arXiv[2026-03-24]. arXiv:arXiv.2304.07683.
 - [28] Glickman M, Sharot T. How human - AI feedback loops alter human perceptual, emotional and social judgements[J]. Nature Human Behaviour, 2024, 9(2): 345-359.
 - [29] Messeri L, Crockett M J. Artificial intelligence and illusions of understanding in scientific research[J]. Nature, 2024, 627: 49-58.
 - [30] Gallegos I O, Rossi R A, Barrow J, et al. Bias and fairness in large language models: a survey[J]. Computational Linguistics, 2024, 50(3): 1097-1179.
 - [31] Delobelle P, Tokpo E K, Calders T, et al. Measuring fairness with biased rulers: a comparative study on bias metrics for pre-trained language models[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2022: 1693-1706.
 - [32] Franco D, Navarin N, Donini M, et al. Deep fair models for complex data: graphs labeling and explainable face recognition[J]. Neurocomputing, 2022, 470: 318-334.
 - [33] Liang P P, WU C Y, Morency L P, et al. Towards understanding and mitigating social biases in language models[C]//Proceedings of the International Conference on Machine Learning. London: PMLR, 2022: 6565-6576.

- [34] Zhang X Y, Zhai J, Ma S Q, et al. Unveiling provider bias in large language models for code generation[PP]. arXiv[2026-03-24]. arXiv: arXiv. 2501.07849.
- [35] Blodgett S L, Barocas S, Daumé III H, et al. Language (technology) is power: a critical survey of “bias” in NLP[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2020: 5454-5476.
- [36] Tang S C, Zhu H M. Mitigating bias in generative AI: a comprehensive framework for governance and accountability[J]. Law, Ethics & Technology, 2024, 1(3): 19-20.
- [37] Meade N, Poole-Dayane E, Reddy S. An empirical survey of the effectiveness of debiasing techniques for pre-trained language models[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 1878-1898.
- [38] Li Y J, Du M N, Song R, et al. A survey on fairness in large language models[PP]. arXiv[2026-03-24]. arXiv: arXiv. 2308.10149.
- [39] Lin Z C, Guan S Y, Zhang W D, et al. Towards trustworthy LLMs: a review on debiasing and dehallucinating in large language models[J]. Artificial Intelligence Review, 2024, 57(9): 243.
- [40] Cheng H, Guo Y, Guo Q, et al. Social debiasing for fair multi-modal LLMs[PP]. arXiv[2026-03-24]. arXiv: arXiv. 2408.06569.
- [41] Birhane A, Dehdashtian S, Prabhu V, et al. The dark side of dataset scaling: evaluating racial classification in multimodal models[C]//Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, 2024: 1229-1244.
- [42] Garcia N, Hirota Y, Wu Y K, et al. Uncurated image-text datasets: shedding light on demographic bias[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 6957-6966.
- [43] Haliburton L, Leusmann J, Welsch R, et al. Uncovering labeler bias in machine learning annotation tasks[J]. AI and Ethics, 2025, 5(3): 2515-2528.
- [44] Hovy D, Prabhumoye S. Five sources of bias in natural language processing[J]. Language and Linguistics Compass, 2021, 15(8): e12432.
- [45] Mavrogiorgos K, Kiourtis A, Mavrogiorgou A, et al. Bias in machine learning: a literature review[J]. Applied Sciences, 2024, 14(19): 8860.
- [46] Hirota Y, Nakashima Y, Garcia N. Gender and racial bias in visual question answering datasets[C]//Proceedings of the 2022 ACM Conference on Fairness Accountability and Transparency. New York: ACM, 2022: 1280-1292.
- [47] Khan M J, Breslin J G, Curry E. Towards fairness in multimodal scene graph generation: mitigating biases in datasets, knowledge sources and models[C]//MUWS 2023: Proceedings of the 2nd International Workshop on Multimodal Human Understanding for the Web and Social Media. New York: ACM, 2023: 34-46.
- [48] Park Y, Woo S, Lee S, et al. Cross-modal alignment and translation for missing modality action recognition[J]. Computer Vision and Image Understanding, 2023, 236: 103805.
- [49] Pan Z, Jiang S, Yang X, et al. Hierarchical cross-modal image generation for multimodal biometric recognition with missing modality[J]. IEEE Transactions on Information Forensics and Security, 2025, 20: 4308-4321.
- [50] Yin S K, Fu C Y, Zhao S R, et al. A survey on multimodal large language models[J]. National Science Review, 2024, 11(12): nwae403.
- [51] Zhang Y F, Yu W, Wen Q, et al. Debiasing multimodal large language models via penalization of language priors[PP]. arXiv[2026-03-24]. arXiv: arXiv.2403.05262.
- [52] Shah M, Sureja N. A comprehensive review of bias in deep learning models: methods, impacts, and future directions[J]. Archives of Computational Methods in Engineering, 2025, 32(1): 255-267.
- [53] Hickman L, Huynh C, Gass J, et al. Whither bias goes, I will go: an integrative, systematic review of algorithmic bias mitigation[J]. Journal of Applied Psychology, 2025, 110(7): 979-1000.
- [54] Raj C, Mukherjee A, Caliskan A, et al. Breaking bias, building bridges: evaluation and mitigation of social biases in LLMs via contact hypothesis[J]. ACM Conference on AI, Ethics, and Society, 2024, 7: 1180-1189.
- [55] Xiao X, Wu B H, Wang J C, et al. Seeing the image: prioritizing visual correlation by contrastive alignment[C]//Proceedings of the Advances in Neural Information Processing Systems 37. San Diego: Neural Information Processing Systems, 2024: 30925-30950.
- [56] Hamidieh K, Zhang H, Gerych W, et al. Identifying implicit social biases in vision-language models[C]//Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM Press, 2024: 547-561.
- [57] Gautam S, Srinath M. Blind spots and biases: exploring the role of annotator cognitive biases in NLP[PP]. arXiv[2026-03-24]. arXiv: arXiv.2404.19071.
- [58] Wang Y, Yin Y X, Li P. Towards the mitigation of confirmation bias in semi-supervised learning: a debiased training perspective[PP]. arXiv[2026-03-24]. arXiv: arXiv.22409.18316.
- [59] Wolfe R, Yang Y W, Howe B, et al. Contrastive language-vision AI models pretrained on web-scraped multimodal data exhibit sexual objectification bias[C]//Proceedings of the 2023 ACM Conference on Fairness Accountability and Transparency. New York: ACM, 2023: 1174-1185.
- [60] Leng S, Xing Y, Cheng Z, et al. The curse of multi-modalities: evaluating hallucinations of large multimodal models across language, visual, and audio[PP]. arXiv[2026-03-24]. arXiv: arXiv. 2410.12787.
- [61] Seyyed-Kalantari L, Zhang H R, McDermott M B A, et al. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations[J]. Nature Medicine, 2021, 27(12): 2176-2182.
- [62] Farnadi G, Havaei M, Rostamzadeh N. Position: cracking the code of cascading disparity towards marginalized communities[PP]. arXiv [2026-03-24]. arXiv: arXiv.2406.01757.
- [63] Xu C J, Zhang J W, Chen Z R, et al. MMDT: decoding the trustworthiness and safety of multimodal foundation models[PP]. arXiv [2026-03-24]. arXiv: arXiv.2503.14827.
- [64] Wang Y Y, Wang X Z, Beutel A, et al. Understanding and improving fairness-accuracy trade-offs in multi-task learning[C]//Proceedings

- of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. New York: ACM, 2021: 1748-1757.
- [65] Hu F, Ratz P, Charpentier A. Fairness in multi-task learning via Wasserstein barycenters[M]//Machine Learning and Knowledge Discovery in Databases: Research Track. Cham: Springer Nature Switzerland, 2023: 295-312.
- [66] Aguirre C, Dredze M. Transferring fairness using multi-task learning with limited demographic information[PP]. arXiv[2026-03-24]. arXiv:arXiv.2305.12671.
- [67] Li X, Chen Z, Zhang J M, et al. Benchmarking bias in large language models during role-playing[PP]. arXiv[2026-03-24]. arXiv:arXiv.2411.00585.
- [68] Zhuo T Y, Huang Y J, Chen C Y, et al. Red teaming ChatGPT via jailbreaking: bias, robustness, reliability and toxicity[PP]. arXiv[2026-03-24]. arXiv:arXiv.2301.12867.
- [69] Khandelwal K, Tonneau M, Bean A M, et al. Indian-BhED: a dataset for measuring India-centric biases in large language models[C]//Proceedings of the 2024 International Conference on Information Technology for Social Good. New York: ACM, 2024: 231-239.
- [70] Wang B X, Chen W X, Li B. DecodingTrust: a comprehensive assessment of trustworthiness in GPT models[PP]. arXiv[2026-03-24]. arXiv:arXiv.2306.11698.
- [71] Akhras N, Antaki F, Mottet F, et al. Large language models perpetuate bias in palliative care: development and analysis of the palliative care adversarial dataset (PCAD) [PP]. arXiv[2026-03-24]. arXiv:arXiv.2502.08073.
- [72] Rutinowski J, Franke S, Endendyk J, et al. The self-perception and political biases of ChatGPT[PP]. arXiv[2026-03-24]. arXiv:arXiv.2304.07333.
- [73] Hartmann J, Schwenzow J, Witte M. The political ideology of conversational AI: converging evidence on ChatGPT's proenvironmental, left-libertarian orientation[PP]. arXiv[2026-03-24]. arXiv:arXiv.2301.01768.
- [74] Feng S B, Park C Y, Liu Y H, et al. From pretraining data to language models to downstream tasks: tracking the trails of political biases leading to unfair NLP models[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2023: 11737-11762.
- [75] Bachmann D, Van Der Wal O, Chvojka E, et al. Ffl-IRT-ing with psychometrics to improve NLP bias measurement[J]. Minds and Machines, 2024, 34(4): 37.
- [76] Zhang T, Zeng Z Q, Xiao Y X, et al. GenderAlign: an alignment dataset for mitigating gender bias in large language models[C]//Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2025: 11293-11311.
- [77] Selvam N R, Dev S, Khashabi D, et al. The tail wagging the dog: dataset construction biases of social bias benchmarks[PP]. arXiv[2026-03-24]. arXiv:arXiv.2210.10040.
- [78] Dhamala J, Sun T, Kumar V, et al. BOLD: dataset and metrics for measuring biases in open-ended language generation[C]//Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. New York: ACM, 2021: 862-872.
- [79] Bevara R V K, Mannuru N R, Karedla S P, et al. Scaling implicit bias analysis across transformer-based language models through embedding association test and prompt engineering[J]. Applied Sciences, 2024, 14(8): 3483.
- [80] Loon A V, Giorgi S, Willer R, et al. Regional negative bias in word embeddings predicts racial animus: but only via name frequency[PP]. arXiv[2026-03-24]. arXiv:arXiv.2201.08451.
- [81] May C, Wang A, Bordia S, et al. On measuring social biases in sentence encoders[C]//Proceedings of the 2019 Conference of the North. Stroudsburg: ACL, 2019: 622-628.
- [82] Gray M, Milanova M, Wu L. Enhancing bias assessment for complex term groups in language embedding models: quantitative comparison of methods[J]. JMIR Medical Informatics, 2024, 12:e60272.
- [83] Guo W, Caliskan A. Detecting emergent intersectional biases: contextualized word embeddings contain a distribution of human-like biases[C]//Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM, 2021: 122-133.
- [84] Dolci T, Azzalini F, Tanelli M. Improving gender-related fairness in sentence encoders: a semantics-based approach[J]. Data Science and Engineering, 2023, 8(2): 177-195.
- [85] Lucy L, Bamman D. Gender and representation bias in GPT-3 generated stories[C]//Proceedings of the Third Workshop on Narrative Understanding. Stroudsburg: ACL, 2021: 48-55.
- [86] Marinucci L, Mazzuca C, Gangemi A. Exposing implicit biases and stereotypes in human and artificial intelligence: state of the art and challenges with a focus on gender[J]. AI & Society, 2023, 38(2): 747-761.
- [87] Nadeem M, Bethke A, Reddy S. StereoSet: measuring stereotypical bias in pretrained language models[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2021: 5356-5371.
- [88] Nangia N, Vania C, Bhalerao R, et al. CrowS-pairs: a challenge dataset for measuring social biases in masked language models[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg: ACL, 2020: 1953-1967.
- [89] Kaneko M, Bollegala D, Okazaki N. Debiasing isn't enough! on the effectiveness of debiasing MLMs and their social biases in downstream tasks[PP]. arXiv[2026-03-24]. arXiv:arXiv.2309.07251.
- [90] Barikeri S, Lauscher A, Vulić I, et al. RedditBias: a real-world resource for bias evaluation and debiasing of conversational language models[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2021: 1941-1955.
- [91] Bommasani R, Liang P, Lee T. Holistic evaluation of language models[J]. Annals of the New York Academy of Sciences, 2023, 1525(1): 140-146.
- [92] Cheng M, Durmus E, Jurafsky D. Marked personas: using natural

- language prompts to measure stereotypes in language models[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 1504-1532.
- [93] Omrani S S, Wolfe R, Caliskan A. Evaluating biased attitude associations of language models in an intersectional context[C]//Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society. New York: ACM, 2023: 542-553.
- [94] Alnegheimish S, Guo A, Sun Y. Using natural sentence prompts for understanding biases in language models[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2022: 2824-2830.
- [95] An H Z, Li Z X, Zhao J Y, et al. SODAPOP: open-ended discovery of social biases in social commonsense reasoning models[C]//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2022: 1565-1588.
- [96] Garimella A, Amarnath A, Kumar K, et al. He is very intelligent, she is very beautiful? on mitigating social biases in language modelling and generation[C]//Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg: ACL, 2021: 4534-4545.
- [97] Garimella A, Mihalcea R, Amarnath A. Demographic-aware language model fine-tuning as a bias mitigation technique[C]//Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing. Stroudsburg: ACL, 2022: 311-319.
- [98] Zhou G Y, Yan Y B, Zou X, et al. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality[PP]. arXiv[2026-03-24]. arXiv:arXiv.2410.04780.
- [99] Qian R, Ross C, Fernandes J, et al. Perturbation augmentation for fairer NLP[C]//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2022: 9496-9521.
- [100] Theodorou B, Danek B, Tummala V, et al. Improving medical machine learning models with generative balancing for equity and excellence[J]. NPJ Digital Medicine, 2025, 8(1): 100.
- [101] Penedo G, Malartic Q, Hesslow D, et al. The refinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only[PP]. arXiv[2026-03-24]. arXiv:arXiv.2306.01116.
- [102] Venkit P N, Gautam S, Panchanadikar R, et al. Nationality bias in text generation[C]//Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 116-123.
- [103] Sun T, Webster K, Shah A, et al. They, them, theirs: rewriting with gender-neutral English[PP]. arXiv[2026-03-24]. arXiv: arXiv.2102.06788.
- [104] Lauscher A, Lueken T, Glavaš G. Sustainable modular debiasing of language models[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2021: 4782-4797.
- [105] Yang K, Yu C, Fung Y R, et al. ADEPT: a debiasing prompt framework[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Vancouver: AAAI Press, 2023: 10780-10788.
- [106] Wang R, Cheng P, Henao R. Toward fairness in text generation via mutual information minimization based on importance sampling [C]//Proceedings of the International Conference on Artificial Intelligence and Statistics. London: PMLR, 2023: 4473-4485.
- [107] Tian J J, Emerson D, Miyandoab S Z, et al. Soft-prompt tuning for large language models to evaluate bias[PP]. arXiv[2026-03-24]. arXiv:arXiv.2306.04735.
- [108] Wang X, Ge T, Mao A, et al. Pay attention to your tone: introducing a new dataset for polite language rewrite[PP]. arXiv[2026-03-24]. arXiv:arXiv.2212.10190.
- [109] Majumder B, He Z X, Mcauley J J. Interfair: debiasing with natural language feedback for fair interpretable predictions[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2022: 9466-9471.
- [110] Amrhein C, Schottmann F, Sennrich R, et al. Exploiting biased models to de-bias text: a gender-fair rewriting model[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2023: 4486-4506.
- [111] Tokpo E K, Calders T. Text style transfer for bias mitigation using masked language modeling[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. Stroudsburg: ACL, 2022: 163-171.
- [112] Alhafni B, Habash N, Bouamor H. User-centric gender rewriting[C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop. Stroudsburg: ACL, 2022: 618-631.
- [113] Jain N, Popovic M, Groves D, et al. Generating gender augmented data for NLP[PP]. arXiv[2026-03-24]. arXiv: arXiv.2107.05987.
- [114] Ruan X W, Qi Z B, Xu Y R, et al. Dual-guided multi-modal bias removal strategy for temporal sentence grounding in video[J]. Multimedia Systems, 2025, 31(2): 108.
- [115] Janghorbani S, De Melo G D. Multimodal bias: introducing a framework for stereotypical bias assessment beyond gender and race in vision language models[PP]. arXiv[2026-03-24]. arXiv: arXiv.2303.12734.
- [116] Yang D K, Li M C, Xiao D L, et al. Towards multimodal sentiment analysis debiasing via bias purification[C]//Proceedings of the European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 464-481.
- [117] Hastings B C, Qian L J, Gibson C, et al. Data augmentation via diffusion model to enhance AI fairness[J]. Frontiers in Artificial Intelligence, 2025, 8: 1530397.
- [118] Poncibò C, Ebers M. Comparative perspectives on the regulation of large language models[J]. Cambridge Forum on AI: Law and Governance. Cambridge University Press, 2025, 1: e10.

[作者简介]



陈晋音（1982-），女，浙江工业大学教授，主要研究方向为人工智能安全、图数据挖掘和进化计算。



张甜馨（2002-），女，浙江工业大学硕士生，主要研究方向为深度学习和人工智能安全。



王泽楷（2004-），男，浙江工业大学本科生。



郑海斌（1995-），男，浙江工业大学助理研究员，主要研究方向为深度学习和人工智能安全。



沈文杰（2002-），男，浙江工业大学硕士生，主要研究方向为深度学习和人工智能安全。



邱洪君（1982-），女，博士，杭州电子科技大学网络安全学院讲师，主要研究方向为人工智能安全、工业互联网安全。