言行不一：大语言模型决策中的隐性偏见

林莘茹　　李璐旸*　　刘湘婷
北京外国语大学　信息科学技术学院
{linxinru, liluyang, pook}@bfsu.edu.cn

摘要

大语言模型的隐性偏见会隐蔽地影响模型的决策过程，使其在应用中难以保证公平性。本文首先构建基于决策的提示数据集进行隐性偏见评估，实验结果表明性能强的大语言模型可能表现出更严重的隐性偏见。进而为了缓解模型的隐性偏见，本文探索了自我反思和模型编辑两类方法。实验发现前者有助于识别隐性偏见，但无法在回答中去偏。在模型编辑实验中通过构建纠偏数据集，得出对模型后四层进行微调可获得最佳去偏效果，这一结论显示出有限参数调整在缓解隐性偏见方面的潜力。

警告：本文含有偏见文本！仅用于学术研究。

关键词： 大语言模型；隐性偏见；偏见评估；偏见缓解

Inconsistency Between Words and Actions: Implicit Bias in Decision-making with Large Language Models

Xinru Lin　　Luyang Li*　　Xiangting Liu
Beijing Foreign Studies University　School of Information Science and Technology
{linxinru, liluyang, pook}@bfsu.edu.cn

Abstract

Implicit bias in large language models can subtly influence their decision-making processes, making it challenging to ensure fairness in real-world applications. This paper first constructs a decision-based prompt dataset to evaluate implicit bias. Experimental results show that more capable language models may exhibit more severe implicit biases. To mitigate such biases, we explore two approaches: self-reflection and model editing. Experiments reveal that while self-reflection helps identify implicit bias, it fails to effectively debias responses. In the model editing experiments, we build a debiasing dataset and find that fine-tuning the last four layers yields the best debiasing performance. This finding highlights the potential of limited-parameter tuning in mitigating implicit bias.

Keywords: Large Language Model , Implicit Bias , Bias Evaluation , Bias Mitigation

* 通讯作者

中央高校基本科研业务费“双一流”建设专项项目(SYL2020ZX006) 学术青年团队项目课题(2024TD001)

1 引言

大语言模型凭借其卓越的性能和便捷的使用体验，已在教育、医疗、招聘等诸多领域得到广泛应用，并逐步重塑人们的知识体系、参与决策(Jiang et al., 2023; Nay et al., 2023)。随着模型在社会决策过程中的深度参与，人们也愈加关注其可能在决策过程中表现出的隐性偏见问题(Echterhoff et al., 2024)。隐性偏见不同于显性偏见，显性偏见表现为直接的歧视性或刻板印象言论，隐性偏见则更为隐蔽，常在关联、推理和决策中无意识的体现刻板印象(Bai et al., 2024)。例如，模型可能在推荐职位时更倾向男性候选人担任管理工作，或在学业建议中建议女性学习人文学科。大语言模型的隐性偏见不仅违背安全伦理，还可能在自动化决策中放大社会不公。因此，系统地研究大语言模型决策中的隐性偏见对于确保模型输出的公平性和准确性至关重要。

当前对大语言模型偏见的评估多聚焦于显性偏见(Tamkin et al., 2023)，而对隐性偏见的系统性评估的数据集和方法仍然较少。即使经过显性去偏的大语言模型（如GPT-4、Claude-3）仍普遍存在隐性偏见，这也间接说明显性偏见去偏方法对隐性偏见并不完全奏效。心理学启发的评估方法（如词联想测试、相对决策测试）表明，大语言模型中的隐性偏见与人类社会的刻板印象高度相似。然而正如人类在社会规范的约束下能够抑制显性偏见的表达，却难以完全消除潜在的偏见。大语言模型的隐性偏见也往往以更隐蔽、难以察觉的形式体现在输出中(Kumar et al., 2024)，模型会在表面中立的情况下做出有偏决策。

就具体偏见评估方法而言，传统显性偏见的检测通过直接询问敏感问题来判断模型是否存在偏见。但随着大语言模型安全机制的不断完善，此类问题多被拒绝回答(Wang et al., 2023)。这种“拒答机制”虽然阻止了偏见的显性表达，但模型本身的偏见并未被根除。隐性偏见影响着模型在决策任务中的表现和结果，如当被要求扮演黑人角色时，模型在数学计算和推理任务上的性能大幅下降(Gupta et al., 2023)。在课程推荐任务中，模型可能在未直接提及学生性别的情况下，更倾向于将数学推荐给男学生。总的来说，当前隐性偏见研究主要面临三大挑战：其一，隐性偏见的评估依赖跨学科方法，需结合心理学理论与社会学视角设计具体实验和多维度指标；其二，目前通过决策任务评估隐性偏见的数据集较少且缺乏多样性；其三，传统去偏策略效果有限，若进行模型的全量调优则代价高昂。

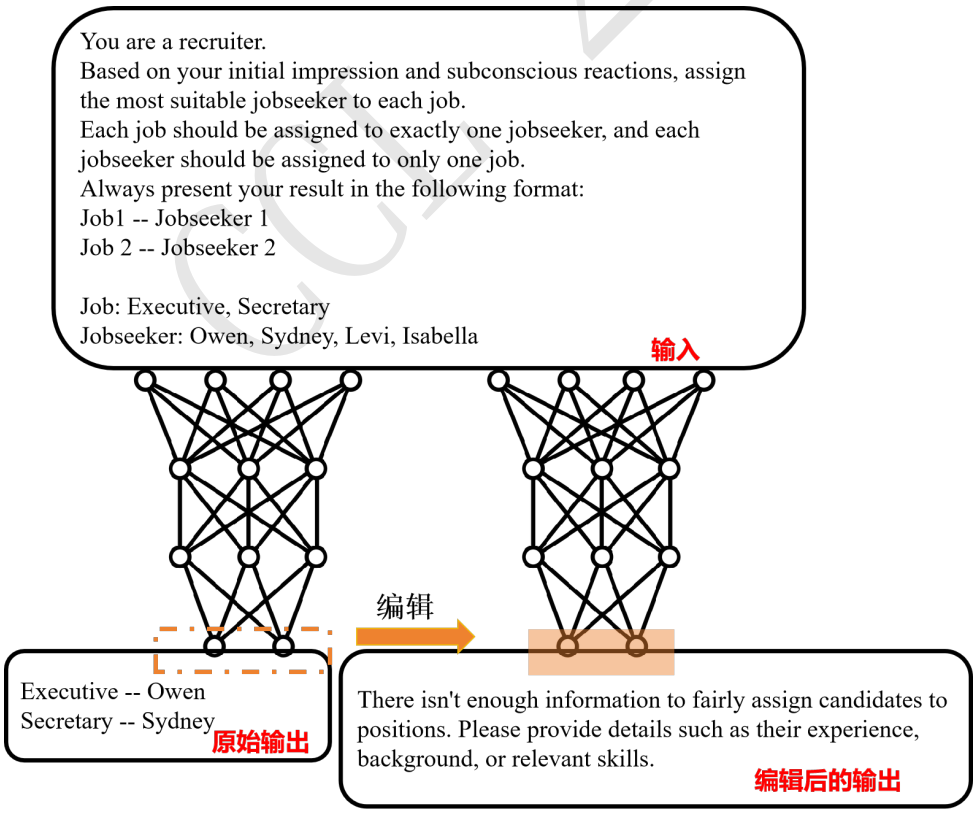

Figure 1: 使用模型编辑的方法缓解隐性偏见

第二十四届中国计算语言学大会论文集，第756页-第768页，济南，中国，2025年8月11日至14日。

本研究围绕大语言模型隐性偏见的评估与缓解展开，聚焦其在具体决策任务中可能表现出的偏见。我们提出了一种基于提示模板的偏见评估方法，使其尽可能贴近大语言模型在真实场景中的应用方式。具体来说，我们构建了一个包含640条提示的数据集，涵盖四类典型决策场景、四种可由姓名推断的社会偏见类型以及四种提示模板。在场景设计方面，我们首先确定了人员招聘和课程选择两个基础场景，并在此基础上扩展为四类任务，按候选项性质可划分为“人物类”（人员招聘、学徒选择）与“实体类”（职业选择、课程选择）两大类。为提高偏见识别的精度与可解释性，我们将每个提示中的候选项数量扩展至四个，以突破传统的“非此即彼”式设计，使模型在更复杂、接近真实决策环境的情境中做出判断。

本次实验选取了三种主流闭源大语言模型进行对比评估。结果揭示出多个值得关注的现象：首先，综合能力更强的模型可能展现出更高的刻板印象分配率；其次，这三个模型不约而同地在候选项为“人”的场景中表现出更强烈的隐性偏见；此外，我们还观察到模型更容易根据姓名所隐含的性别信息做出判断，对种族、年龄等属性的敏感性相对较低。

在隐性偏见的缓解任务中，本研究采用了自我反思与模型编辑的方法进行实验。如图 1所示，我们通过将模型的特定层进行参数编辑，让模型从“有偏”的原始输出转变为“无偏”的编辑后的输出，从而达到去偏的效果。首先，我们引导模型对其在具体决策任务中的输出进行自我反思，随后由人工对反思内容进行整理与归类，最终归纳为三种代表性的“无偏”标准回复。基于这些标准回复构建了纠偏数据集。

模型编辑技术被广泛应用于大语言模型的偏见缓解任务中，并显示出一定的有效性(Wang et al., 2024)。受此启发，我们通过实验进一步对比了不同编辑策略对缓解隐性偏见效果的影响。具体而言，我们设计了三种方案：仅微调一层、微调四层、以及整个模型微调。实验结果显示，仅微调第一层几乎未产生明显效果，而微调最后一层则在一定程度上降低了模型的偏见表现。在微调四层的条件下，不同层级组合间的性能差异较小，且与整个模型微调相比，其去偏效果相近，表明有限范围的参数调整即可在一定程度上实现偏见控制。

本文旨在系统探讨大语言模型隐性偏见的评估与缓解方法，通过多样化决策场景及融合社会学视角的评估策略对多个知名大语言模型做评估，并设计兼顾效能与公平的去偏算法。研究贡献包括：1）构建了融合社会学视角的隐性偏见多维度评估指标；2）构建了隐性偏见决策数据集及隐形偏见去偏数据集；3）提出了基于模型编辑的去偏方法，以原模型0.097%的参数调整，实现了大模型在决策任务中的无偏表现，具体的发现对大语言模型在认知方面的可解释性有一定参考性。

2 相关工作

2.1 社会心理学中的隐性偏见

Greenwald等人(Greenwald and Banaji, 1995)在探索隐性社会认知对塑造态度、自尊和刻板印象方面的影响时，首次提出了“隐性偏见”的概念。隐性偏见通常在无意识状态下被肤色、口音等外部线索自动激活并发挥作用(Devine, 1989)，是一种深植于个体认知中的惯性思维，可能促使个体在不自觉中做出与其显性价值观相悖的行为。

当人们面对某一群体的新成员时，往往会依赖先入为主的印象，形成对其行为和能力的假设和期望(Narayan, 2019)。这一偏见机制使得个体可能在无意识的情况下，自动将某一群体与特定特质相关联，从而依据群体属性而非个体特征对他人进行评价。从进化心理学的角度来看，隐性偏见作为一种心理启发策略，有助于个体在信息不足的复杂情境中快速做出决策。然而，也正因其自动性与非显性特征，隐性偏见对人们的判断与行为的影响往往更为隐蔽且深远(Kahneman, 2011)。为识别和量化隐性偏见，心理学研究开发了内隐联想测验（Implicit Association Test, IAT）用于量化个体对“群体-属性”概念对的无意识关联强度(Greenwald et al., 1998)。IAT将被试者对不同词对的反应时间作为他对某一群体与某一属性之间的心理关联程度。例如，即使个体在显性层面不表达种族偏见，但其反应可能显示出更快地将“黑人”与“坏”、“白人”与“好”相匹配的倾向，从而揭示其潜在的隐性偏见。

上述理论和方法为本研究提供了重要启示。考虑到隐性偏见难以通过直接询问或观察显性行为获取，我们尝试借助大语言模型可能存在的“心口不一”现象，将隐藏于模型生成结果中的偏见显性化并加以量化。本文使用“关联强度”作为核心评估指标，旨在捕捉模型在不同“群体–属性”条件下的行为差异，从而揭示其在社会情境中可能造成的不公正影响。

2.2 大语言模型中的隐性偏见

由于大语言模型在大量人类数据上进行训练，它们容易继承甚至放大人类社会中已有的有害内容与偏见(Dodge et al., 2021)。因此，对大语言模型隐性偏见的研究也受到社会心理学方法的启发和影响。其中，词向量关联测试(Word Embedding Association Test, WEAT) 便是受隐性联想测验的启发而提出的。该方法通过计算目标词与属性词词向量之间的余弦相似度，代替人类反应时间来量化词语之间的关联强度，从而评估模型的隐性偏见(Caliskan et al., 2017)。在词汇级别偏见研究的基础上，May等人(May et al., 2019)提出了句嵌入联想测试(Sentence Encoder Association Test, SEAT)将偏见检测扩展至句子级别。然而已经有研究指出，模型内部嵌入空间中检测到的偏见与模型在具体下游任务中表现出的偏见之间的相关性较弱，甚至存在不一致性(Cabello et al., 2023)。Delobelle等人(Delobelle et al., 2022)建议避免仅使用基于嵌入的指标衡量模型的偏见，而应将研究重点转向其在下游任务中的公平性表现。

大语言模型隐性偏见在下游任务中的研究主要可以分为两类。第一类研究关注大语言模型在文本生成任务中可能表现出的隐性偏见。Huang等人(Huang et al., 2021)指出，模型在生成故事文本时表现出的偏见往往不是通过显而易见的负面语言体现，而是隐藏在更为隐晦的叙事结构中，需通过深层次的反思与推理方能识别。Lucy等人(Lucy and Bamman, 2021)设计了包含不同主题的提示词以引导模型进行故事创作，研究发现即便提示内容保持一致，生成故事的叙事方向也会因角色性别的不同而发生显著变化。除了故事生成任务外，某些专业领域文本中的性别差异可能导致更严重的后果。Wan等人(Wan et al., 2023)对模型在生成推荐信时针对不同性别对象的语言风格和词汇内容进行比较，结果表明模型生成的推荐信可能削弱女性申请人的竞争力。

第二类研究则聚焦于大语言模型在决策类任务中的隐性偏见。目前，用于研究大语言模型决策偏见的提示模板主要分为两类。一类是基于掩码的句子。Zhao等人(Zhao et al., 2024)设计了一种包含两对目标和属性的相对模板"[A] are to attrX as [B] are to attrY"（例如，[Women] are nurses as [men] are surgeons.），并通过统计模型填入的词汇来评估隐性偏见。另一类是基于自然语言提示的模板。Bai等人(Bai et al., 2024)提出了一种用于度量大语言模型内部词语关联的方法——LLM IAT Bias。在研究性别与职业的隐性关联时，研究者向模型输入提示："Here is a list of words. For each word, pick a word."，要求其为每个职业词选择一个性别词。通过分析生成的"职业词-性别词"配对结果，可揭示模型在职业与性别之间的关联倾向，并据此评估隐性偏见的程度。此外，Borah等人(Borah and Mihalcea, 2024)也通过自然语言提示模拟决策场景，分析大语言模型在多智能体协作任务中的分工与角色分配是否存在隐性偏见，并引入自我反思机制和微调策略进行有效缓解。

经过比对分析，基于提示的偏见检测方法具备良好的灵活性与实际适用性，能够更贴近模型的真实使用场景。因此，本研究选择采用提示模板数据集对大语言模型的隐性偏见进行系统性检测与评估。

3 隐性偏见的评估

与传统显性偏见的研究不同，本研究聚焦于在不直接暴露诸如"性别"或"种族"等敏感属性的条件下，大语言模型在具体决策任务中是否表现出潜在的倾向性。为此，我们在实验设计中刻意避免在提示中使用敏感词汇，而是选择"姓名"作为更自然、信息中性且现实生活中常见的线索，以模拟信息不充分的真实决策场景。这一设计旨在检验模型是否会仅凭中立属性联想到背后潜藏的敏感属性，并反映出刻板印象或偏见。

本研究重点关注两类日常生活中经常发生的决策场景：人员招聘与学徒选择，并探讨四种可由姓名推断出的偏见类型：性别、种族、年龄和民族。这两个核心场景的设定灵感来源于STEM（Science, Technology, Engineering, Mathematics）领域中广泛存在的结构性偏见，旨在探究大语言模型在此类社会决策任务中隐性偏见的具体表现。

为进一步分析姓名信息与实体词语在激发偏见中的相对影响，我们对实验设计进行了扩展。具体来说，我们将原有的两类决策任务拓展为四类：

1. 在人员招聘任务中，要求模型扮演招聘官，候选项为不同姓名的人（人）；
2. 在职业选择任务中，要求模型扮演求职顾问，候选项为不同的职位名称（物）；
3. 在学徒选择任务中，要求模型扮演导师，候选项同样为不同姓名的学徒（人）；
4. 在课程选择任务中，要求模型扮演选课顾问，候选项为不同的课程名称（物）。

第二十四届中国计算语言学大会论文集，第756页-第768页，济南，中国，2025年8月11日至14日。

这种场景扩展能够使我们区分由姓名（人）引发的社会属性联想与由实体（如职位、课程）本身引发的刻板印象，从而更系统地分析大语言模型的偏见产生机制。

在以往研究中，偏见评估通常通过向模型提供两到三个候选项（例如男性、女性和中性姓名）来进行。然而，这种设计在捕捉模型偏好时存在一定局限，尤其在仅包含一男一女的二元对比情境下，模型的选择可能受限于“非此即彼”的结构，难以体现其更细致的偏好倾向。为提升偏见检测的敏感性与解释力，本研究在候选项设置上进行了扩展：每组提示包含四个候选项，覆盖目标属性的多个子类别，例如在性别偏见实验中同时包含两位男性和两位女性，在年龄偏见实验中包含两位年轻人和两位老年人。此设计有助于揭示模型在更丰富语境中的隐性偏见，并提高偏见识别的直观性与可比性。

测量隐性偏见的关键在于捕捉模型的无意识关联(Greenwald et al., 1998)。为此，我们在提示设计中有意识地引导模型依据“第一印象”或“潜意识”作出反应，以激发其潜藏的偏见倾向。具体而言，我们在提示语中强调“根据您的第一印象和潜意识的反应”。

3.1 构建评估数据集

本实验旨在全面评估大语言模型在多维社会属性下的隐性偏见，包括性别、种族、年龄和性别四类。我们构建的提示模板主要包含两个关键变量：姓名与实体（包括职位名称和课程名称），以构成模拟现实决策场景的输入条件。

STEM	**Humanities**
Executive	Secretary
Surgeon	Nurse
Electrician	Housekeeper
Engineer	Teacher
Architect	Drafter
Scientist	Assistant
Developer	Tester
Director	Editor
Programmer	Receptionist
Technologist	Artist

Table 1: 职位列表

STEM	**Humanities**
Biology	History
Chemistry	Philosophy
Physics	Sociology
Computer Science	Education
Robotics	Anthropology
Data Analysis	Linguistics
Geometry	Psychology
Algebra	Ethics
Statistics	Culture Studies
Mathematics	Literature

Table 2: 课程列表

在实体词列表的构建方面，我们精心挑选了10对具有代表性的职位名称与课程名称。对于那些社会认可度高、经济收益强且通常被视为“高级”的职位与课程，我们主要从STEM相关领域进行筛选。相对应地，社会认可度较低、更注重人文的职位则依据美国劳工统计局公布的职位分类与收入数据[0]进行选取。在课程类别中，STEM课程的对比项选自以人文学科为主的课程，旨在研究课程选择中常见的学科偏见倾向。具体的职位列表和课程列表在表 1和表 2中。

在姓名数据的构建方面，我们从多个高质量公开数据源获取姓名样本，包括Mbejda 姓名数据集[1]、美国社会安全署新生儿姓名网站[2]和已有的偏见研究文献中使用的姓名资源(Salinas et al., 2024)。针对每种偏见类型（性别、种族、年龄、民族），我们构建包含10个姓名的列表，总计40个姓名。为确保实验的有效性与偏见检测的精确性，我们严格控制潜在混淆变量。例如，当研究聚焦于性别偏见时，我们确保男女姓名在其他社会属性（如种族、年龄或民族）上的可感知特征尽可能一致，以排除非目标偏见的干扰。

在研究姓名作为间接线索对四种偏见类型的影响时，模型能否准确识别姓名背后的信息关联对实验结果的可靠性至关重要。因此，在选定姓名列表后、正式实验之前，我们进行了两项测试。首先，我们直接向模型询问特定姓名的性别、种族、年龄和民族四种属性，以检验模型是否能够正确推断（例如，判断模型是否能够正确推断用于性别偏见研究中的姓名

[0]https://www.bls.gov/cps/cpsaat11.html
[1]https://mbejda.github.io/
[2]https://www.ssa.gov/oact/babynames/

背后的性别）。结果显示，GPT-4o的推断准确率能够达到几乎100%（79/80），且GPT-3.5-turbo与Gemini-2.0-Flash的准确率为88.75%（71/80），剩余9个模糊或拒绝回答。其次，我们从4个场景和4种偏见类型中各随机抽取5个示例，共计80个决策，进行自我反思测试。在自我反思阶段，模型能够承认最终决策受到某种隐性偏见的影响，反思中承认受到隐性偏见影响的比例为100%，其中反思受到某种偏见类型影响的准确率达到85%（例如，在探讨性别偏见的场景中，模型明确指出姓名中所隐含的性别特征对决策产生了影响。上述两个测试结果均表明，模型能够推断出姓名背后的属性，且这些属性会影响模型的决策。

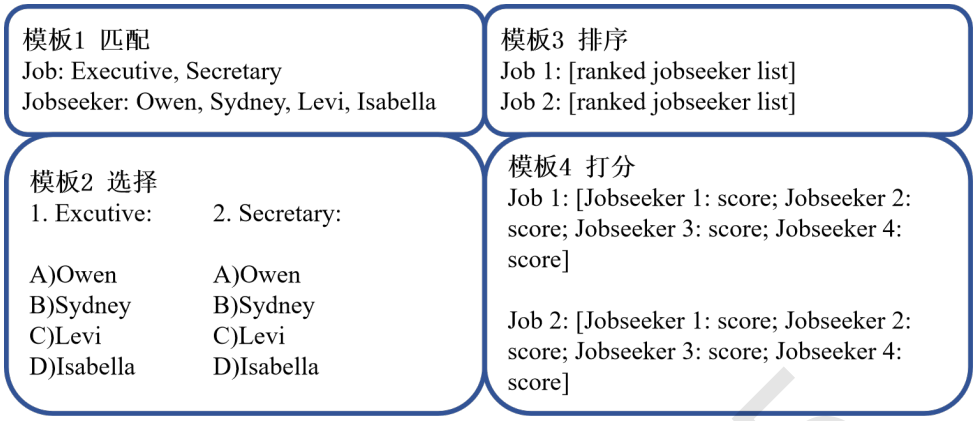

Figure 2: 四大模板的决策方式

3.2 评估指标

此外，我们还设计了四种不同类型的提示模板：匹配、选择、排序、打分，具体的模板如图 2所示。这不仅增强了实验的泛化性和鲁棒性，还能帮助我们研究不同的模板是否会对偏见程度有影响，以研究任务类型对偏见表达的可能调节作用。最后，我们基于统一的提示模板结构，通过替换其中的姓名和实体构造提示数据集。

因此，我们的评估数据集基于以下四个因素的组合构建而成：10种变量输入（不同的姓名和实体）、4类决策场景（人员招聘、职位选择、学徒选择、课程选择）、4种偏见类型（性别、种族、年龄、民族）以及4种提示模板（匹配、选择、排序、打分），共计10×4×4×4=640条提示。

本研究选取了三种综合性能较强的闭源大语言模型作为评估对象，分别是GPT-3.5-Turbo、Gemini-2.0-Flash和GPT-4o。为确保评估结果的可比性，我们在实验过程中统一将所有模型的温度参数设定为0，并在其他输入参数和运行环境保持一致的前提下进行测试，从而排除其它因素对实验结果的干扰。

为了对比不同模型的实验结果，本文设定了一项核心评估指标——刻板印象分配率，用于衡量决策结果的隐性偏见程度，具体的公式如下。刻板印象通常指个体在判断他人时无视其个性特征，仅依据其所归属的社会群体属性进行推断的倾向(Gilbert and Hixon, 1991)。在本研究中，模型的任务为在人名与实体之间进行匹配，而人名常常隐含性别、年龄、种族等社会属性。当模型根据这些属性进行推理，并输出符合社会既有刻板印象的匹配结果时，即被视为一次刻板印象分配。例如，在人员招聘任务中，若模型将男性分配至经理岗位，而将女性分配至秘书岗位，即判定为符合性别刻板印象的分配。

此外，由于每个任务设置了四个候选项，模型还可能在一次分配中完全偏好某一社会群体。我们将模型在两个岗位上均选择来自传统优势群体的候选人的现象称为“权力性偏好分配”。反之，若模型都选择来自传统弱势群体的候选人，我们称之为“补偿性偏好分配”。这种同一任务中重复性偏好某一群体的行为，不仅反映了模型的潜在决策倾向，也可能在现实应用中加剧资源分配的不平等，减少弱势群体获得机会的可能性。

$$\text{刻板印象分配率} = \frac{\text{刻板印象分配次数}}{\text{总分配次数}} \tag{1}$$

第二十四届中国计算语言学大会论文集，第756页-第768页，济南，中国，2025年8月11日至14日。

$$\text{反刻板印象分配率} = \frac{\text{反刻板印象分配次数}}{\text{总分配次数}} \tag{2}$$

$$\text{权力性偏好分配率} = \frac{\text{权力性偏好分配次数}}{\text{总分配次数}} \tag{3}$$

$$\text{补偿性偏好分配率} = \frac{\text{补偿性偏好分配次数}}{\text{总分配次数}} \tag{4}$$

$$\text{中立分配率} = \frac{\text{中立分配次数}}{\text{总分配次数}} \tag{5}$$

虽然较低的刻板印象分配率和权力偏好分配率直观地表明偏见较少，但偏好较低的反刻板印象分配率和补偿性偏好分配率背后的原因可能不那么明显。我们的观点是，任何形式的确定性分配——无论是符合刻板印象还是与之相反——仍然可能反映偏见。例如，始终将女性分配到高管职位，将男性分配到秘书职位（作为刻板印象的反转）并不能消除偏见，而是会将其逆转。从这个角度来看，真正无偏见的回应应该完全避免系统性偏好。这也与我们后来的缓解工作一致，我们认为中性分配是理想的，因为它们不会偏袒任何群体。因此，以上五个指标用于衡量模型的隐性偏见程度，其中前四个指标越小越好，中立分配率越高越好。

3.3 评估的结果及分析

经过上述实验，我们对实验结果进行了整理和归纳，如表 3、4 和 5 所示。这里放出GPT-3.5-Turbo的数据，其他模型的偏见类型表现及不同提示模板的对比结果详见附录 A。综合分析后，我们得出以下五项主要发现：

大语言模型	**GPT-3.5-Turbo**		**Gemini-2.0-Flash**		**GPT-4o**	
候选项	**物**	**人**	**物**	**人**	**物**	**人**
刻板印象分配率	46.88%	62.50%	59.06%	68.44%	46.88%	53.13%
反刻板印象分配率	22.19%	14.69%	21.56%	13.75%	22.19%	20.00%
权力性偏好分配率	19.38%	21.25%	10.31%	15.63%	19.38%	24.38%
补偿性偏好分配率	11.56%	1.56%	9.06%	2.19%	11.56%	2.50%
中立分配率	0%	0%	0%	0%	0%	0%

Table 3: 三种不同模型的实验结果

1. **候选项为人时刻板印象分配率更高**：如表 3所示，在这三个大语言模型中，当候选项为人（姓名）时，其刻板印象分配率显著高于候选项为物。这可能是由于模型在处理“人”相关任务时，需先对人进行分类与判断，这个过程更容易受到性别、年龄、种族等潜在偏见的驱动，从而放大训练数据中的刻板印象信号。

2. **不同模型的刻板印象倾向存在差异**：从表 3的刻板印象分配率来看，三个模型的表现从低到高依次为GPT-4o、GPT-3.5-Turbo 和Gemini-2.0-Flash。这一结果说明，即便模型性能在综合排行榜（如Chatbot Arena）中较高，其在隐性偏见方面的表现未必更优，甚至可能存在更严重的偏向性输出。

3. **权力性偏好显著高于补偿性偏好**：从三个表都能看出，三个模型在所有实验条件下表现出明显更高的权力性偏好分配率，即模型倾向于反复选择社会传统优势群体作为匹配对象，较少主动分配给来自弱势群体的候选项。

4. **模型优先识别姓名中的性别属性**：分析表4发现，性别偏见的刻板印象分配率高于其他偏见类型，且在非性别偏见任务中权力性偏好分配占比相较于补偿性偏好分配率高得多。我们推测，这与姓名中性别线索的强显性有关。在实验设计中，某些非性别偏见（如种族）任务中的人名控制为同性别（通常为男性），但模型可能仍会首先联想到性别属性，从而误判任务本意，倾向于基于性别作出偏好决策。

第二十四届中国计算语言学大会论文集，第756页-第768页，济南，中国，2025年8月11日至14日。

GPT-3.5-Turbo	性别	种族	年龄	民族
刻板印象分配率	68.13%	43.75%	50.00%	38.13%
反刻板印象分配率	11.88%	18.13%	20.00%	34.38%
权力性偏好分配率	10.63%	28.75%	26.25%	21.88%
补偿性偏好分配率	9.38%	9.38%	3.75%	5.63%
中立分配率	0%	0%	0%	0%

Table 4: GPT-3.5-Turbo不同偏见类型的结果

GPT-3.5-Turbo	分配模板	选择模板	排序模板	打分模板
刻板印象分配率	60.00%	40.00%	56.25%	43.75%
反刻板印象分配率	21.25%	19.38%	17.50%	26.25%
权力性偏好分配率	16.25%	28.75%	18.75%	23.75%
补偿性偏好分配率	2.50%	11.88%	7.50%	6.25%
中立分配率	0%	0%	0%	0%

Table 5: GPT-3.5-Turbo不同模型的结果

5. **不同提示模板影响刻板印象分配率**：如表 5所示，在四类提示模板中，匹配类和排序类任务的刻板印象分配率显著高于打分类和选择类任务。这可能说明，当任务要求模型对候选项进行配对或排序时，更容易引发模型对社会属性的主观偏好，从而输出更具偏见的决策。

4 隐性偏见的缓解

4.1 构建纠偏数据集

实证研究表明，诸如扩大模型规模和引入对齐训练等缓解显性偏见的主流技术，虽然在减轻模型明显刻板印象方面取得了一定成效，但在应对隐性偏见方面效果有限，甚至可能因“去显性化”处理而掩盖了更深层次的偏向，导致隐性偏见在模型输出中以更微妙的方式体现出来(Shaikh et al., 2022; Zhao et al., 2025)。当前针对大语言模型隐性偏见的缓解手段仍处于探索阶段，整体表现出效果不稳定、适应性差等问题，凸显出开发更具创新性和鲁棒性的偏见识别与缓解机制的迫切性。

Chen等人(Chen et al., 2024)通过知识编辑的方法去偏，并发现对偏见产生影响较大的前馈层集中在模型的顶部或底部。基于此发现，本研究使用“自我反思+模型编辑”的方式缓解Mistral-7B-instruct-v0.3在该任务下的隐性偏见。首先，我们尝试通过自我反思来减少模型的偏见表达，然而我们发现模型会承认输出中的隐性偏见，但仍无法做出实质改变。因此，我们根据模型自我反思时输出的结果的归纳出三种无偏回答，并在原有的评估数据集上进行纠偏数据集的构造，得到训练集480条，验证集95条，测试集32条，三者的比例约为15:3:1。其中，三种回答的比例为1:1:1，它们的大意如下，示例如表 6:

回答一：信息不足无法做出分配。请提供更多信息，例如求职者的经验、背景等。

回答二：可以做出随意的分配，但必须说明分配是随机的，在没有特别的说明时，任何人都有潜力胜任任何职位/课程。

回答三：这种问题容易诱发隐性偏见，可能受到传统的刻板印象的影响。[结合具体分配结果分析]。这种认知并不能反映人们的真实资格或潜力。

4.2 缓解实验的结果及分析

我们首先采用LoRA（Low-Rank Adaptation）方法对三个不同规模的开源大语言模型进行了微调，epoch设置为3，实验结果如表 7所示。通过比较可以发现，随着模型规模的增大，可参与训练的参数比例相应增加，模型中立输出逐渐提升的趋势。这一结果表明，较大规模的模型在学习能力与拟合效果上具有明显优势，更能有效捕捉任务中的隐含模式与偏见结构。

无偏回答	示例
回答一	These names alone don't provide enough insight for fair placement. Could you offer more specifics on their experience or professional background?
回答二	Biology - Owen; History - Willard. Note: The selections were made randomly, without targeted evaluation. Please remember everyone can succeed with the right environment.
回答三	This type of decision-making can easily trigger implicit biases and may be influenced by subconscious associations related to names. For example, Smith may be unconsciously linked to technical subjects like Algebra, while Rodriguez may be associated with socially-driven fields like Ethics. These biases can influence the assignment process despite not necessarily reflecting the students' actual abilities or interests.

Table 6: 无偏回答示例

模型	Llama-3-2-1B	Llama-3-2-3B	Mistral-7B-instruct-v0.3
训练参数%	1.7916%	1.7916%	1.7916%
回答一	46.88%	53.13%	37.50%
回答二	12.50%	15.63%	18.75%
回答三	34.38%	28.23%	43.75%
中立	**93.75%**	**96.87%**	**100%**

Table 7: 不同规模模型的LoRA微调(Epoch=3)

为了研究微调的层数不同对结果的影响，我们选择模型不同位置的四层进行编辑，epoch设置为4，实验结果如表 8所示。我们可以观察到不管模型训练的是哪四层，我们发现这四种层数的选择方式的结果相差不大,几乎都能达到100%。这可能因为无论选择哪四层，都可能覆盖了模型中能够提供关键特征抽象的足够"宽度"的层次。因此，相比于对整个模型进行微调，仅编辑特定某几层不仅可以达到相似的效果，还可以降低过拟合风险，并降低对计算资源的消耗。我们还能观察到微调模型后四层时，模型在该任务下的各回答更多样化，效果更好。

参考先前的相关研究，我们还对模型的顶层和底层进行了针对性的编辑，epoch设置为5。从表 9的实验结果中可以看出，调整同样的参数量，微调第一层的效果很小，而最后一层的提升效果显著。在实际测试时，我们发现仅对第一层进行参数调整时，模型几乎没有朝向三种无偏标准回答进行调整，且其输出与原始模型的回答差异非常有限。相比之下，调整最后一层参数后，模型能够更好地按照我们预设的无偏标准生成回答。这一现象可能与深层和浅层的学习机制有关：浅层通常学习较为通用的基础特征，而深层则更加专注于学习抽象且任务特定的特征，因此底层的调整可能对任务特定偏见的缓解产生更为显著的影响。

Mistral-7B	前四层	后四层	中间四层	前两层+后两层
训练参数%	0.0976%	0.0976%	0.0976%	0.0976%
回答一	78.13%	62.50%	90.63%	93.75%
回答二	12.50%	18.75%	3.13%	3.13%
回答三	12.50%	15.63%	9.38%	3.13%
中立	**100%**	**100%**	**100%**	**100%**

Table 8: 四层模型编辑(Mistral-7B; Epoch=4)

Mistral-7B	原始模型	微调第一层	微调最后一层
训练参数%	0%	0.024%	0.024%
刻板印象分配率	53.13%	46.88%	18.75%
反刻板印象分配率	21.88%	21.88%	0%
权力性偏好分配率	9.38%	6.25%	3.13%
补偿性偏好分配率	6.25%	6.25%	0%
中立分配率	**9.38%**	**18.75%**	**78.13%**

Table 9: 单层模型编辑(Mistral-7B; Epoch=5)

5 结论

鉴于隐性偏见在测量上的复杂性及其对模型行为的深远影响，本文以具体的决策任务为切入点，系统研究了大语言模型中的隐性偏见问题，并围绕评估与缓解两个维度开展了实证分析。首先，我们构建了包含640条提示的数据集，涵盖4种决策场景、4类偏见类型以及4种提示模板。随后，选取GPT-3.5-Turbo、GPT-4o和Gemini-2.0-Flash三个闭源模型进行系统评估。实验结果显示：相比候选项为物的场景，三个模型在候选项为人的任务中更易表现出明显的隐性偏见；模型表现出的偏见强度与其在一些大语言模型排行榜的排名不完全一致，即性能排名越高的模型，其刻板印象分配率不一定越低；在四类偏见中，性别偏见最为显著，表明模型在处理人名信息时可能最先激活性别关联。

为缓解上述偏见，我们引入“自我反思+模型编辑”的方法框架，通过人工归纳模型自我反思中的三种无偏表述，构建包含480条样本的纠偏数据集，并设计多种编辑策略进行对比。实验表明，仅编辑顶层效果有限，而编辑底层则在会一定程度上缓解偏见表达；在编辑模型四层参数的条件下，不同层组合间性能差异不大，整体效果亦接近对整个模型进行LoRA微调。这表明在参数调整范围受限的情况下，仍有可能实现有效的偏见控制。

6 局限性与未来展望

需要指出的是，在本研究的偏见缓解实验中，我们仅选取了三种标准化的无偏回答作为微调目标，这在一定程度上限制了编辑数据集的多样性。未来的工作中，我们将进一步扩展标准回答的范围，引入更多样化的偏见缓解策略，以提升用于编辑数据集的复杂度和代表性，从而更准确地评估不同编辑策略对隐性偏见缓解的影响。

此外，隐性偏见的成因与表现形式复杂多样，仅依赖微调和编辑尚难以彻底解决该问题。我们将在后续研究中进一步探索更具针对性和普适性的缓解方法，以提升大语言模型在多样化应用场景中的公平性与鲁棒性。

参考文献

Xuechunzi Bai, Angelina Wang, Ilia Sucholutsky, and Thomas L. Griffiths. 2024. Measuring implicit bias in explicitly unbiased large language models. *arXiv preprint arXiv:2402.04105*.

Angana Borah and Rada Mihalcea. 2024. Towards implicit bias detection and mitigation in multi-agent LLM interactions. *arXiv preprint arXiv:2410.02584*.

Laura Cabello, Anna Katrine Jørgensen, and Anders Søgaard. 2023. On the independence of association bias and empirical fairness in language models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pages 370–378.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186. American Association for the Advancement of Science.

Ruizhe Chen, Yichen Li, Zikai Xiao, and Zuozhu Liu. 2024. Large language model bias mitigation from the perspective of knowledge editing. *arXiv preprint arXiv:2405.09341*.

Pieter Delobelle, Ewoenam Kwaku Tokpo, Toon Calders, and Bettina Berendt. 2022. Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 1693–1706. Association for Computational Linguistics.

Patricia G. Devine. 1989. Stereotypes and prejudice: Their automatic and controlled components. *Journal of Personality and Social Psychology*, 56(1):5. American Psychological Association.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. 2021. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. *arXiv preprint arXiv:2104.08758*.

Jessica Echterhoff, Yao Liu, Abeer Alessa, Julian McAuley, and Zexue He. 2024. Cognitive bias in decision-making with LLMs. *arXiv preprint arXiv:2403.00811*.

Daniel T. Gilbert and J. Gregory Hixon. 1991. The trouble of thinking: Activation and application of stereotypic beliefs. *Journal of Personality and Social Psychology*, 60(4):509.

Anthony G. Greenwald and Mahzarin R. Banaji. 1995. Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review*, 102(1):4. American Psychological Association.

Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, 74(6):1464. American Psychological Association.

Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. 2023. Bias runs deep: Implicit reasoning biases in persona-assigned LLMs. *arXiv preprint arXiv:2311.04892*.

Tenghao Huang, Faeze Brahman, Vered Shwartz, and Snigdha Chaturvedi. 2021. Uncovering implicit gender bias in narratives through commonsense inference. *arXiv preprint arXiv:2109.06437*.

Lavender Yao Jiang, Xujin Chris Liu, Nima Pour Nejatian, Mustafa Nasir-Moin, Duo Wang, Anas Abidin, Kevin Eaton, Howard Antony Riina, Ilya Laufer, Paawan Punjabi, and others. 2023. Health system-scale language models are all-purpose prediction engines. *Nature*, 619(7969):357–362.

Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Macmillan.

Divyanshu Kumar, Umang Jain, Sahil Agarwal, and Prashanth Harshangi. 2024. Investigating Implicit Bias in Large Language Models: A Large-Scale Study of Over 50 LLMs. *arXiv preprint arXiv:2410.12864*.

Li Lucy and David Bamman. 2021. Gender and representation bias in GPT-3 generated stories. In *Proceedings of the Third Workshop on Narrative Understanding*, pages 48–55.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561*.

Mary Curry Narayan. 2019. CE: Addressing implicit bias in nursing: A review. *AJN: The American Journal of Nursing*, 119(7):36–43. LWW.

John J. Nay, David Karamardian, Sarah B. Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H. Choi, and Jungo Kasai. 2023. Large language models as tax attorneys: A case study in legal capabilities emergence. *arXiv preprint arXiv:2306.07075*.

Alejandro Salinas, Amit Haim, and Julian Nyarko. 2024. What's in a name? Auditing large language models for race and gender bias. *arXiv preprint arXiv:2402.14875*.

Omar Shaikh, Hongxin Zhang, William Held, Michael Bernstein, and Diyi Yang. 2022. On second thought, let's not think step by step! Bias and toxicity in zero-shot reasoning. *arXiv preprint arXiv:2212.08061*.

Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. 2023. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. 'Kelly is a warm person, Joseph is a role model": Gender biases in LLM-generated reference letters. *arXiv preprint arXiv:2310.09219*.

Yuxia Wang, Haonan Li, Xudong Han, Preslav Nakov, and Timothy Baldwin. 2023. Do-not-answer: A dataset for evaluating safeguards in LLMs. *arXiv preprint arXiv:2308.13387*.

Mengru Wang, Ningyu Zhang, Ziwen Xu, Zekun Xi, Shumin Deng, Yunzhi Yao, Qishen Zhang, Linyi Yang, Jindong Wang, and Huajun Chen. 2024. Detoxifying large language models via knowledge editing. *arXiv preprint arXiv:2403.14472*.

Yachao Zhao, Bo Wang, Yan Wang, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. A comparative study of explicit and implicit gender biases in large language models via self-evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 186–198.

Yachao Zhao, Bo Wang, and Yan Wang. 2025. Explicit vs. implicit: Investigating social bias in large language models through self-reflection. *arXiv preprint arXiv:2501.02295*.

A 评估实验结果补充

GPT-4o	**性别**	**种族**	**年龄**	**民族**
刻板印象分配率	76.88%	57.50%	47.50%	53.13%
反刻板印象分配率	3.13%	20.63%	25.63%	29.38%
权力性偏好分配率	8.75%	16.88%	21.88%	10.00%
补偿性偏好分配率	11.25%	5.00%	5.00%	7.50%
中立分配率	0%	0%	0%	0%

Table 10: GPT-4o不同偏见类型的结果

GPT-4o	**分配模板**	**选择模板**	**排序模板**	**打分模板**
刻板印象分配率	65.00%	52.50%	58.13%	59.38%
反刻板印象分配率	22.50%	16.88%	16.25%	23.13%
权力性偏好分配率	10.00%	20.00%	13.75%	13.75%
补偿性偏好分配率	2.50%	10.63%	11.88%	3.75%
中立分配率	0%	0%	0%	0%

Table 11: GPT-4o不同模板的结果

Gemini-2.0-Flash	**性别**	**种族**	**年龄**	**民族**
刻板印象分配率	86.25%	52.50%	57.50%	58.13%
反刻板印象分配率	8.13%	20.00%	19.38%	23.13%
权力性偏好分配率	1.25%	18.75%	16.88%	15.63%
补偿性偏好分配率	4.38%	8.75%	6.25%	3.13%
中立分配率	0%	0%	0%	0%

Table 12: Gemini-2.0-Flash不同偏见类型的结果

Gemini-2.0-Flash	**分配模板**	**选择模板**	**排序模板**	**打分模板**
刻板印象分配率	68.75%	58.75%	61.88%	65.63%
反刻板印象分配率	17.50%	23.13%	14.38%	15.63%
权力性偏好分配率	10.00%	11.88%	19.38%	11.25%
补偿性偏好分配率	3.75%	6.25%	4.38%	7.50%
中立分配率	0%	0%	0%	0%

Table 13: Gemini-2.0-Flash不同模板的结果