导师学术声誉性别差异的增强与削弱
——来自大语言模型的实验证据*

□ 赵志栋　刘丁已　朱正浩

领域编辑推荐语：

大语言模型在赋能学术评价的同时，也潜藏着固化甚至放大社会结构性偏见的风险。本文创新性地将大语言模型视为行为主体开展随机控制实验，揭示了性别、年龄与学历在导师学术声誉评价中的交互效应及其经济后果。该研究不仅在大语言模型多维偏见的识别与测度方面进行了有益探索，而且也为 AI 时代构建公平、可信的教育评价生态提供了重要的现实启示。

——张光磊

摘　要：大语言模型（LLMs）的迅速发展引发了人们对其加剧社会偏见扩散的担忧。本研究旨在探讨 LLMs 在导师学术声誉评价的情境下，所表现出的基于性别、年龄及学历交互的歧视程度。基于随机对照试验的结果表明，LLMs 预期中年女性学者较其男性同行，学术声誉评分更低，且在争取同一名研究生时，所承担的成本更高，这反映出年龄对学术声誉性别差异的增强效应。实验结果还显示，拥有博士学位的女性学者较其男性同行，被 LLMs 赋予更高的学术声誉评分和更低的招募学生成本，这显示了学术资质对学术声誉性别差异的削弱作用。基于辅助实验的结果证实了上述两种效应存在衍生效应，和二者在样本外的显著性，从而验证了它们的稳健性。本文的发现挑战了 LLMs 输出的客观中立性，由此，也强调了关注 AI 在教育和学术领域应用中纠偏机制和策略的必要性。

关键词：学术导师；学术声誉；性别差异；交互作用；大语言模型

*　本研究得到江苏省教育科学规划重点课题“数智时代职业院校主动适应‘技能极化’现象的逻辑机制与实践路径”（课题编号：B/2025/02/86）的资助。特别感谢期刊编辑和匿名评审专家在审稿过程中提出的宝贵修改意见，同时感谢张光磊和尤树洋两位教授及其他与会者在《管理学研究》“人工智能与管理学研究方法”工作坊中对文章的修改意见，文责自负。

一、引言

学术声誉的本质是一种社会承认。Merton（1973）认为，学术水平的衡量应当建立在“卓越”这一单一标准之上。然而，在现实世界中，性别却可以作为一个隐藏的状态分层器，改变评价者对“卓越”的感知（Heiberger et al.，2026）并导致学术声誉出现性别差异。这种性别差异可能贯穿女性学者的整个学术生涯——从早期的科研训练到终身教职的获取，再到顶级学术荣誉的角逐。在此过程中，即便在每个阶段，女性所面临的均为微小的、隐蔽的门槛，这一次次的微小差异也可能通过“马太效应”（Merton，1988）被不断放大，并最终形成类似“玛蒂尔达效应”（Rossiter，1993）的学术声誉的性别鸿沟：女性学者被引次数更少（Lerman et al.，2022），科学影响更小（Hofstra et al.，2020），且留在学术界的概率更低（Kim & Moser，2025）。当下，大语言模型（Large Language Models，LLMs）所表现出的卓越效率提升能力，使其正在被深度嵌入学术研究、教育评估等关键领域。然而，此类模型在赋能效率提升的同时，也可能继承或通过算法放大训练数据中所隐含的社会偏见和歧视等问题（Caliskan et al.，2017；Barocas et al.，2023；Guilbeault，et al.，2025）。在教育、学术场景中，若 LLMs 作为自动化工具展现出性别歧视行为，则放任这些行为不仅可能对教育和学术生态公正性和可持续性形成威胁，更可能通过算法反馈循环加剧结构性不平等（Mullainathan & Obermeyer，2017；Mehrabi et al.，2021），并进而影响人才的配置效率。

学者学术声誉的建立和维持既随时间而积累，也随其获取的学术资质而积累，但积累过程并非性别中立。对不同性别的学者而言，生理时间的流逝是客观而均一的；相对而言，学术资质的获得则更大程度上取决于学者的主观能动性，是因人而异的。从这一角度来讲，探讨学者年龄的客观变化和与主观因素密切相关的学者所获学术资质是如何作用于学术声誉的性别差异的，对于理解学术评价体系的公平性与效能具有重要的理论和现实意义。以年龄为例，自 2024 年起，国家社会科学基金青年项目对女性的年龄限制设置为 40 岁，比男性放宽了 5 岁，光明网评论此为“公平之举”①。为何放宽女性年龄限制反而能够彰显公平？这一问题本身就包含着研究年龄与性别交互效应的必要性。已有研究显示，AI 时代，年龄歧视并不鲜见（周翔等，2025）；且年龄偏见会与性别偏见交织——Guilbeault 等（2025）发现，LLMs 会放大网络平台上的年龄偏见，体现为网络平台上普遍把女性描绘得比男性更年轻，而 ChatGPT 在生成简历时还放大了这种偏见。尽管在现实世界中，多种偏见相互交织共同存在，然而，现有研究或聚焦于某种特定偏见（或歧视）的分析；或是在分析多种偏见时，将它们视为彼此独立作用进行计量检验，少有研究探讨不同偏见间交互影响的复杂机制。此外，在 LLMs 作为自动化工具的角色日益凸显的 AI 时代，LLMs 在多大程度上会复现现实社会中的偏见，以及不同偏见之间的交互作用是如何影响模型决策的，尚需更多实证证据的支撑。具体到 AI 时代

① https://guancha.gmw.cn/2024-04/15/content_37263979.htm。

学术声誉的性别差异这一特定场景中，学者年龄的客观变化和学者经由主观努力所获的学术资质是如何分别（或共同）作用于学术声誉，可能伴随着怎样的经济后果等，这些问题的答案均有待进一步的揭示与量化。

本研究旨在从研究内容和方法两个方面填补研究缺口，探讨可能影响学术或教育公平的，内嵌于 LLMs 中的性别、年龄和学历偏见之间复杂的交互作用。从研究内容上，我们借助 LLMs 的自然语言理解能力，探测了其在导师评价的模拟情境中，依据评价内容中所含自然语言内容的变化，是否会由于性别、年龄及以学历为代表的学术资质的不同而输出导师学术声誉的不同评分（及对应的研究生招募成本）；以及这些输出所体现的导师学术声誉的性别差异，是否与年龄和学术资质之间存在交互影响。从研究方法上，我们采用了随机对照试验的方法，克服了有关因果推断的困难。采用实验法而非基于人口特征统计的计量回归方法，一方面，考虑到后者多基于还原论，难以解释人们面临多条件权衡决策时可能存在的动态、非线性或耦合关系的复杂机制，因而可能得到偏离真实系统运行机制的结果（Anderson，1972）；另一方面，则考虑到当涉及系统内生偏见（systemic biases，如性别、种族偏见）问题时，基于多元变量回归的标准统计计量方法有时可能得到违反日常直觉的荒谬结论①。

本研究使用网络 Python 搜集了来自导师评价网网友匿名对其研究生导师评价的数据②，选取了对理工学科导师的评价内容，使用随机对照试验方法分析了可能存在于 LLMs 中的性别偏见，及其与年龄及学历等因素间的交互作用。研究结果显示：在控制其他评价内容不变的前提下，LLMs 会基于导师的性别、年龄和学历而给出差异化的学术声誉评分，并进而导致导师在争取同一名研究生时所预期承担的费用出现差异。以费用差异为例，我们发现：为争取同一名硕士研究生，模型预测中年男性和中年女性学者均需承担显著高于不披露性别和年龄信息学者的成本（$p<0.01$），提示年龄歧视的算法化嵌入；同时，中年女性学者被期望支付的费用平均高于不披露性别和年龄信息学者 6.3%，显著（$p<0.01$）高于其男性同行被期望超额支付的水平（为 3.6%），说明年龄与性别偏见呈现协同强化（增强）效应。有趣的是，拥有博士学位的女性学者招募研究生的期望成本反而低于同等资历的男性（$p<0.05$），揭示了高学历对性别偏见的削弱作用。基于经管类学科导师评价数据的随机控制试验结果证实了上述研究发现的稳健性。来自辅助实验的结果也表明，LLMs 还表现出了基于增强效应和削弱效应衍生出的两种效应：中和效应和增强的削弱效应，二者进一步支持了研究发现及解释机制的稳健性。我们的发现对 LLMs 在学术评估应用场景中的客观性和中立性提出了挑战，强调了理解偏见生成机制和修正算法偏见的迫切需求；通过展示 LLMs 所含偏见的多维性和复杂性，提示仅依赖单一公平性指标（如性别平衡）的 LLMs 应用可能会掩盖交叉维度的歧视结构。

① 原表述为：a standard economic approach but one that drew ridicule from a journalist who tweeted，“Yes [，] once you remove the influence of all of the other racist systems，racism doesn't exist.”来自 https：//www.chicagobooth.edu/review/weve-been-underestimating-discrimination。

② https：//www.rateyoursupervisor.com。

本文的核心贡献如下。（1）在理论层面，本研究从“偏见交互”的视角出发，突破了既有研究将性别、年龄等偏见独立控制的单维分析范式，发现了年龄和学历分别对学术声誉性别差异的增强和削弱作用，揭示了不同社会身份特征在 LLMs 内部可能以非线性方式耦合并共同作用的复杂机制，为理解偏见和歧视的形成提供新的理论视角。（2）在实证层面，使用百度开发的 SKEP 模型量化了基于学生评价的导师学术声誉评分，借助 ChatGPT 权衡量化了评分所对应的经济后果，进而刻画了 LLMs 再现性别刻板印象和偏见并导致经济歧视的过程。研究揭示了年龄、学历与性别偏见的交互作用，填补了该领域实证研究的缺口。（3）在方法层面，本研究以随机控制试验设计为核心，构建了一种可量化、可复现、可解释的 LLMs 偏见测度框架。与以往依赖统计回归或语料分析的“结果导向”研究不同，本文通过严格控制输入变量（如性别、年龄、学历）的自然语言表征，实现在同一语境下对模型输出的因果识别，从而克服了传统方法面临的内生性与混杂问题。实验设计秉承了这样一种研究思路：将 LLMs 视为行为主体，通过实验法观察其“社会行为”模式，从而将算法和机器偏见研究纳入行为科学的实证框架。（4）本研究还为教育与学术领域的算法治理提供了现实启示。研究表明，LLMs 会复现并可能强化现实中存在的偏见。该结果提示：若不对算法进行系统的偏见审计与矫正，其在教育评价、学术评审、人才遴选等环节的广泛应用，可能导致数字化不公与“算法化排除”。因此，研究呼吁在模型研发与应用过程中关注多维偏见的动态监测，在关键环节引入算法审计和去偏机制，以确保学术与教育生态中 AI 应用的公平、可信与可解释。

二、理论分析与研究假说

（一）文献回顾：性别刻板印象、偏见和歧视

关于女性学术表现不及男性的偏见（刻板印象）在人类社会文化里广泛存在（Hyde & Mertz，2009；Breda et al.，2020），且在理工科 STEM（Science，Technology，Engineering and Mathematics）领域尤为严重（Moss－Racusin et al.，2012）。例如，Hyde 和 Mertz（2009）表明，包括美国在内的很多国家的人口，在基础数学表现上基本不存在性别差异；虽然在高难度问题解决等领域，男性优势曾经较为明显，但这种差异并非源自生物学，而是由社会文化因素和教育机会的性别不均衡所导致的。国内研究也揭示了性别刻板印象的存在（杨晋和陈晓宇，2021）和其所造成的女性在高等教育机构谋求工科职位时所受到的明显歧视（赵颖等，2023）。内嵌于社会文化中的性别刻板印象，哪怕在性别平等程度较高的国家，也显著存在。如，Breda 等（2020）通过跨国数据研究发现，在经济发达且性别平等程度较高的国家，内化的“数学不适合女孩”刻板印象反而更为强烈，并能部分解释这些国家中女性在数学密集型领域的低代表性。存在于学术或科研环境中的性别刻板印象，可能会阻碍女性职业发展并进而阻碍科学进步的速度。例如，Moss－Racusin 等（2012）发现，科学系教师在评价相同资质的申请者时，无论男女教师均倾向于给予男性申请者更高的能力评价、雇用可能性和起薪，以及

更多的职业辅导。这说明，即使在强调客观性和科学精神的 STEM 学术环境中，性别刻板印象依然潜移默化地影响着教师对女性学者的评估，会存在性别歧视，从而阻碍女性在理工科领域的长期发展（Kotek et al.，2023）。Vásárhelyi 等（2021）和 Koffi（2025）均发现，学术引用中也存在着对女性学者的结构性排斥。学生对教师的评价也体现了明显的性别歧视（Mengel et al.，2019）。Knobloch－Westerwick 等（2013）基于实验的研究也显示，当同一研究的署名作者为男性时，研究生会给出“科学质量”更高的评分，且他们更倾向于与可能为男性的作者展开合作研究，在“男性气质”更甚的领域里这种倾向更为明显。

受限于训练数据和模型学习机制，LLMs 可能会继承人类社会的性别刻板印象，并经过各种学习算法输出与刻板印象相关的偏见或歧视内容。首先，性别刻板印象长期且广泛地存在于人类历史和社会文化中。例如，王伟宜和桂庆平（2021）研究显示，国内高等教育学科分布长期存在明显的“男工女文艺”现象，这反映了一种性别刻板印象；国外基于隐形联想测试（implicit association test，IAT）的研究结果显示，女性相关词汇（例如，“woman”和“girl”）相比男性相关词汇，更倾向于与艺术领域而非数学或科学领域相关联（Nosek et al.，2002a）；类似的研究（Nosek et al.，2002b）还发现，女性名字更倾向于与家庭相关词汇关联而男性名字则更倾向于与职业或事业相关词汇关联。其次，由于 LLMs 通常在海量的未经偏见审查的互联网语料上进行训练，这些语料本身往往包含了人类社会存在的各种偏见，加之 LLMs 通常采用基于关联统计的学习方式，如自监督学习等，因而容易导致数据中的偏见关联被模型视为高出现概率的模式而加以吸收。以标准的机器学习模型为例，其在从大规模文本数据中学习语义表示的过程中，会不可避免地吸收并再现这些历史文化中存在的性别刻板印象：例如，谢宇和索菲娅·阿维拉（2025）使用机器学习模型进行 IAT 测试，复现了之前基于人类受试者研究中所揭示的女性词汇显著偏向与艺术而非科学关联（Nosek et al.，2002a）和女性被隐形地刻画为更多承担家庭而非职业或事业角色（Nosek et al.，2002b）的性别偏见。以词嵌入和自回归语言模型为代表的无监督机器学习算法同样会继承偏见。由于该方法等基本原理为利用大量文本数据中词与词之间的共现信息构建词向量空间，因此会继承存在于训练数据中的有关性别差异的统计规律。例如，谢宇和索菲娅·阿维拉（2025）发现，静态词嵌入模型如 word2vec 和 GloVe 中存在明显的性别偏见；Bolukbasi 等（2016）也证实，“man”相比“woman”，与“computer programmer”之间的距离小得多。以上证据表明，训练数据中的性别刻板印象可以在模型中被捕捉为一种“性别方向”，并成为模型表达词语语义的一部分。而有监督的机器学习，或基于注意力机制的深度神经网络模型（如 Transformer）等训练的 LLMs，也会吸收其大规模语料库中包含性别刻板印象的统计规律，并通过学习过程（算法）将这些体现人类决策偏见的模式内化，在生成文本时重现甚至放大这些偏见（Ferrara，2024）。例如，Kotek 等（2023）发现，即便在有人工监督的强化学习算法取得成功的 LLMs 中，也倾向

于保留甚至强化训练数据中的性别偏见。马中红和吴熙倡（2024）也发现，“以社交聊天机器人为代表的 AI 在学习和模仿中复刻与强化了人类社会性别结构性力量的性别偏见。”

（二）文献回顾：学术声誉

“学术声誉”通常被视为学术群体和利益相关者对学者或机构的总体评价和印象累积（Amado & Juarez，2022）。高校（或研究机构）的学术声誉可定义为其利益相关方（学生、教职员工、管理者等）在与高校交流、互动中形成的主观印象总和（Rindova，et al.，2005）。学者的学术声誉则可定义为学术共同体对学者在“增进和发展知识方面所作出的贡献给予的承认和尊敬”（刘尧和余艳辉，2009）。现有研究多用学者的学术活动产出——文章发表、专利产出的数量或成果被引用的次数等指标，来量化他们的学术声誉：例如，夏纪军（2014）使用了基于学者学术产出建立的 h 指数来量化学者的学术声誉。此类衡量方法尽管被广泛采用，但一方面学术引用本身就存在着对女性的结构性排斥（Vásárhelyi et al.，2021；Koffi，2025），另一方面该方法易导致扭曲激励和资源错配等问题（牛风蕊，2016），使其招致了越来越多的批评声音（姜晓晖和汪卫平，2021）。

从微观视角出发，学生作为高校内部最庞大的群体，既是高校的利益相关方，也是在校学者的利益相关方。一方面，高校（或机构）的学术声誉，很大一部分建立在学生的在校体验上，且影响高校对其他非本校学生的吸引力（Plewa，et al.，2016）及他们的升学院校选择（Lafuente - Ruiz - de - Sabando et al.，2018）。另一方面，国内的已有案例（吴冠军，2021）揭示，那些与学者有日常学术交流的学生，往往可以预知，有时甚至可以作用于学者本人学术声誉的崩塌。而且，构成学生在校体验的一个主要部分就是师生互动的“软”体验——因此，无论是衡量高校还是学者的学术声誉，纳入来自学生群体的评价和印象都显得尤为必要。

随着自然语言处理技术的发展，高校管理者已开始对来自学生的评价和反馈进行情感分析（Shaik et al.，2023），以辅助教学改进和管理决策（Guder & Malliaris，2024），进而提高学校声誉。例如，Tzacheva 和 Easwaran（2021）发现，学生评价的情感倾向可有效用于预测教师的教学绩效或声誉等指标：不仅体现为基于学生对教师教学评价的情感得分与来自学生的常规量化评价（如 Likert 量表）分数高度一致，而且表现为基于评价的情感分析可以提供常规量化评价无法涵盖的洞见。

然而，鲜有研究把来自学生的评价和反馈纳入到教师学术声誉评价体系中。姜晓晖和汪卫平（2021）将高校教师的声誉概括为学术道德、学术能力和学术贡献三个维度——我们认为，那些日常参与教师学术、科研活动的学生，对每一个维度的评价都有发言权。首先，由于学生日常参与教师科研项目的具体环节，因而学生可以成为教师的学术行为是否符合相关规范和准则，是否有违学术道德的内部知情者。以华中农业大学 11 名硕博学生联名举报导师事件①为例，学生们发布的证据详细展示了其导师的学术活动存在多种形式的学术不端问题。这些问题的曝光，显示了仅依据学术产出指标来

① 该事件被媒体广泛报道，新京报报道链接参见 https://m.bjnews.com.cn/detail/1705900962168086.html。

评价学者学术能力和贡献的不足之处——毕竟，在被曝光之前，该导师的学术成果指标相当耀眼。其次，在导师长期指导下进行学术研究的学生，是导师学术供给（学术道德和学术能力）的直接消费者，有时甚至是共同生产者。既然商品市场上的消费者可以通过分享商品的使用体验来塑造商品的口碑和形象，那么学术市场上的消费者——学生，也应当可以通过分享自身的学术体验来塑造其学术导师的口碑与形象，并使这些体验成为导师学术声誉的重要组成部分：以江苏科技大学郭伟事件[①]为例，曾师从郭伟的博士生对媒体坦言，其早在媒体曝光事件一年多之前，于入学后第二个月便对郭伟的学术水平产生怀疑并随后办理了退学。

鉴于此，本文将 LLMs 依据学生对导师评价给出的情感分值作为导师学术声誉的代理变量，把 LLMs 依据情感分所提出的研究生月补助金额作为学术声誉经济后果的代理变量，在此基础上考察导师的学术声誉（及经济后果）怎样随评价内容所设置的导师性别、年龄和以博士学位为代表的学术资质的变化而变化，进而探索年龄和学术资质对导师学术声誉性别差异的作用规律。

（三）研究假说 1：学术声誉性别差异与年龄的交互效应

社会角色理论指出，当群体刻板印象与特定角色要求不符时，就会产生偏见（Eagly & Karau，2002）。已有文献（Hyde & Mertz，2009；Breda et al.，2020）揭示，认为“女性相较男性不适合 STEM 学科领域”“女性不胜任 STEM 研究”类刻板印象在多种社会文化中广泛存在；Sargent 等（2022）研究表明，组织本身是具有性别色彩的，其结构和假设往往基于男性的职业轨迹，这种“性别化组织”特征在高等教育与科研机构中尤为突出。实证研究显示，性别偏见和歧视几乎贯穿女性学术生涯的各个阶段：在应聘初级职位时，女性申请人便被认为不如男性有能力，同样资历下被录用可能性更低，同时起薪也较低（Moss – Racusin et al.，2012）；在申请科研经费时，女性申请者需要比男性多出近三篇高影响力论文才能获得同等评分和资助机会（Wennerås & Wold，1997）；即便在发表了学术价值与男性同僚相近的科研成果后，女性作者论文被后续相关研究引用的概率也显著更低（Vásárhelyi et al.，2021；Koffi，2025）；女性获得终身教职的比例远低于男性（Liu et al.，2024），而中途退出学术职业生涯的比例则更高（Huang et al.，2020）。总之，STEM 领域的女性学者，面临着学术声誉起点低、提升难，且更易受阻的困境。

角色一致理论（Eagly & Karau，2002）还指出，那些打破刻板印象跨越了职业门槛进入男性主导领域的女性，会由于对应了反刻板化（counter – stereotypical）的女性形象而遭受回弹效应（Rudman et al.，2012）的冲击，使她们更易受到更加严苛的对待与审视（吴欣桐等，2017）：前述有关女性学者在科研申请（Wennerås & Wold，1997）和论文被引（Vásárhelyi et al.，2021；Koffi，2025）方面呈现的劣势也与回弹效应的解释一致。回弹效应还意味着，女性只有呈现出比同龄男性更高的表现（Correll et al.，2020），才能在男性主导的领域内获得相同程度的认可。例如，研究显示，在学者晋升终身教职（此时学者多数处于中年期）的盲评阶段，

① 该事件得到大量媒体关注，界面新闻报道链接之一参见 https：//www. jiemian. com/article/13663595. html。

即使简历完全相同，男、女性申请人被偏好的比例也可高达4:1（Steinpreis et al.，1999）；在申请生物医药类基金时，女性需要超出男性申请者160%以上的表现才能获得资助（Wennerås & Wold，1997）；上述研究也呼应了科学研究领域的女性学者漏管效应（Huang et al.，2020）现象。

学术声誉作为一种社会承认并非与生俱来，而是需要时间的积累——然而，前述理论及实证研究均指出，女性在学术生涯的各个阶段都可能处于声誉积累的不利地位。既然如此，那么随着女性学者年龄的增长，那些一次次由性别偏见所致的差异，即便每次均微小，也可能通过“马太效应”（Merton，1988）被不断放大，并最终形成类似“玛蒂尔达效应”（Rossiter，1993）的学术声誉的性别鸿沟。

因此，我们认为：处于STEM等男性主导领域里的女性学者，可能会因性别偏见而不得不面对自身学术声誉的开局不利局面；即便在进入领域工作后，也会因回弹效应等导致学术声誉积累和提升的不利局面……这些贯穿学术生涯的不利局面随时间不断积累，使得女性学者的学术声誉积累劣势被逐渐放大，从而出现年龄与学术声誉性别差异的交互作用。由此，我们提出：

假设1：在STEM领域，会出现年龄对导师学术声誉性别差异的加强效应，使得中年女性导师的学术声誉显著低于条件相同的男性同行。

（四）研究假说2：学术声誉性别差异与学术资质的交互效应

统计歧视理论（Phelps，1972）认为，即使雇主没有偏见（即“无歧视偏好”），在信息不完全的劳动力市场中，也可能因信息成本过高而依赖群体特征（如肤色、性别）作为个体能力的代理变量，从而导致歧视性结果的出现。该模型讨论了一种群体间能力方差不同的典型情况：当雇主认为黑人的能力方差更大时，会给黑人测试分数赋予更高的权重，从而导致他们预测黑人比获得相同高分的白人能力更高，即便黑人能力的整体先验期望较低。这种基于理论推导出现的“评估反转”情形也在基于实验的研究中得到证实。例如，Bohren等（2019）发现：在没有历史评价记录的情况下，女性在发布数学问题时面临明显的歧视（获得的赞和声望更少）；但当女性账户积累了良好的评价历史（高声望）后，歧视方向发生反转——她们的帖子反而比男性同类账户获得更多的赞和声望。研究者认为，这种反转现象出现的原因是部分评估者最初低估女性能力，但当看到女性克服“困难”达到更高标准获得好评后，反而会高估其能力。

贝叶斯更新推断机制可为上述“反转”现象提供逻辑一致的解释：首先，刻板印象的存在，说明人们对不同群体的能力存在先验偏见；但贝叶斯更新推断机制认为，人们会根据所获得的新信息对先前信念进行更新；当新信息的“信号强度”足够高时，评价者的负面先验会大幅收敛，甚至可能出现过度修正或超额奖励——即，出现雇主预期获得相同高分数的黑人能力高于白人，和女性发布的数学问题内容比男性获得更好评价的反转效应。

已有研究揭示了“女性不适合理工科”这一偏颇刻板印象在人类社会文化中的广泛存在。与统计歧视理论和贝叶斯更新推断相一致，我们认为，当贝叶斯更新推断机制与此性别刻板

印象结合时，在 STEM 领域也可能出现类似的反转效应。具体地，在该领域存在性别刻板先验的前提下（女性被低估），强而清晰的能力信号（如来自高声望平台的背书、在顶尖机构的经历、获得高难度的资质等）会使评价者对女性被评对象能力分布的先验进行显著上调，甚至出现反转效应。在本文情境下，获得理工科博士学位作为强烈的学术资质信号，可能使人们对女性触发的贝叶斯“向上修正幅度”更大（Bayesian updating overcomes prior bias），从而使得女性理工科博士学位拥有者较其男性同行获得更高的学术声誉评分。由此提出：

假设 2：在 STEM 领域，会出现博士学位对导师学术声誉性别差异的削弱效应，使得博士女性导师的学术声誉高于其条件相同的男性同行。

（五）研究假说 3：学术声誉性别差异与年龄、学历的叠加交互效应

Bordalo 等研究了上述性别差异所涉贝叶斯更新的心理机制，提出偏颇刻板印象（Bordalo et al.，2016）和难度诱发误估（DIM，Difficulty – Induced Misestimation）（Bordalo et al.，2019）是两种独立但会相互影响的认知偏差①——我们认为，这两种认知偏差出现在本文的情境中，前者会体现为年龄对学术声誉性别差异的增强效应，而二者共同作用则会出现博士学位对学术声誉性别差异的削弱作用。如前所述，类似女性不胜任 STEM 领域的偏颇刻板印象在人类文化中广泛存在；DIM 机制则揭示，当问题难度较高时，人们倾向于显著高估他人的表现：因此，当在人们刻板印象中的女性，在其不占优势的 STEM 领域获得博士学位时，那难度一定较男性更高——DIM 机制决定了此时人们会显著地高估这些女性的表现，并给予她们比同条件男性更高的评价。

若基于上述心理机制来解释年龄和博士学位分别作用于导师学术声誉性别差异的增强和削弱效应，则类似的心理机制还应当衍生出以下两种效应：一为中和效应，二为增强的削弱效应。中和效应是指，当年龄和博士学位共同作用于导师学术声誉的性别差异时，由于二者作用效果相反，因此会出现一定程度的相互抵消，使得学术声誉的性别差异较二者分别作用时显著性下降。而增强的削弱效应之所以会出现，则是基于对 DIM 的理解——既然女性完成难度较高的挑战会使她们获得比同条件男性更高的声誉（性别差异的削弱效应），那么这种反转效应应当会随着挑战难度的提升（如在国外获得博士学位）而更加显著，即，出现增强的削弱效应。基于此，提出：

假设 3a：在 STEM 领域，会出现年龄和博士学位共同作用于导师学术声誉性别差异的中和效应，使得中年博士导师学术声誉的性别差异显著性不及假设 1 或假设 2 中的情形。

假设 3b：在 STEM 领域，导师学术声誉的性别差异会出现增强的削弱效应，使得在国外获得博士学位的女性导师学术声誉高于其条件相同的男性同行，且差异显著性高于假设 2 中的情形。

① 偏颇信念或刻板印象指评估者对某一群体（如女性）的能力或特质持有的系统性错误认知。这种信念不符合实际数据，且会持续影响评估决策，导致歧视现象的出现或逆转。在本文情境中，评估者（LLMs）可能错误地认为女性的 STEM 学科能力普遍低于男性，从而当更新了女性拥有博士学位这一强信号后，意识到自己的错误刻板印象，进而修正信念，转向对女性能力的更高认可。

三、研究方法

（一）随机对照试验（Randomized Controlled Trial，RCT）方法

随机对照试验（Randomized Controlled Trial，RCT）是一种常用于评估因果关系的实验设计方法，广泛应用于医学、社会科学、教育等领域。它通过随机分配实验对象到不同的组（处理组和控制组），起到了控制消除混杂因素（confounding factors）导致偏差的问题，可以较为准确地评估干预措施对特定结果的影响（Heckman & Jeffrey，1995）。

尽管 RCT 在消除偏差、进行有效因果推断及提供可重复可验证结果方面有着诸多的优越性，但由于 RCT 往往需要招募人作为实验对象参与实验，在某些情况下，可能引发伦理担忧。然而随着大语言模型的发展，其所生成的文本内容越来越接近人类的水平，这种出色的自然语言处理（Natural Language Processing，NLP）能力为我们将 LLMs 作为实验对象参与实验提供了可能。此外，LLMs 还具备优秀的数据分析和处理能力，这进一步提升了 LLMs 作为实验对象可以提供一致可靠实验结果的可能性。Guilbeault 等（2025）也已把 LLMs 纳为实验对象。为充分利用 LLMs 的上述优点，我们在 RCT 设计和分析中吸收了 LLMs 作为实验对象使其生成对应相应干预措施的评估结果，以期为理解影响导师学术声誉性别差异的因素提供新的视角与工具。

本研究采用 RCT 的设计思路，研究 LLMs 对导师的学术声誉评分是否会因性别而异；以及若存在差异，这些差异可能会对应怎样的经济后果。设计思路遵循先分组，再实施干预并观察干预结果，最后进行结果分析的顺序。（1）分组。我们使用网络 Python 搜集了匿名网友对理工学科研究生导师的评价数据，以此作为原始数据。在为每一条评价分配匿名 ID 后，将评价条目随机分成控制组和处理组。（2）施加干预措施并观察干预结果。在随后的实验中，我们保持控制组导师评价内容不变，而改变处理组导师评价的性别、年龄、学术资质等内容（干预措施），并在每次实验中收集 LLMs 提供的，基于每一条评价内容的导师学术声誉评分，及其所要求补助数据（由 LLMs 生成评估结果）。（3）分析干预结果的含义。通过研究干预措施及相应的干预结果在不同实验中的变化，我们探索了年龄和学术资质对导师学术声誉性别差异及潜在经济后果的作用效应。

（二）实验设计与工具

在将每一条评价分配匿名 ID，并按 5∶1 的比例随机分成控制组和处理组后，我们设计了三组实验。第一组实验将未修改的原始评价内容用百度开发的 PaddleNLP（Paddle Natural Language Processing Toolkit）下的 SKEP（Tian et al.，2020）模型计算情感分，将其作为学术声誉的代理变量，然后将评价内容和情感分一起发给 GhatGPT－4o，让它根据这些信息给出相应的月期望补助金额；第二组实验保持控制组评价信息不变，修改处理组导师评价的性别、年龄信息，再次得到情感分，并重新发给 GPT－4o 让它返回所期望的补助金额；第三组则修改了处理组导师评价的性别和学历信息，在获取情感分后再次请 GPT－4o 生成其所期望的

补助金额。实验的具体实施步骤如图 1 所示。上述三组实验中，在处理了对应导师的评价内容（实施干预措施）后，都分别执行了图 1 所示的最后两个步骤：第一，使用 python 调用 SKEP 处理评价内容，生成对每一条评价的情感分；第二，利用 GPT－4o 分析评价和 SKEP 返回的结果，并令其对三组实验使用前后一致的规则生成对应相应评价的导师月补助期望金额。

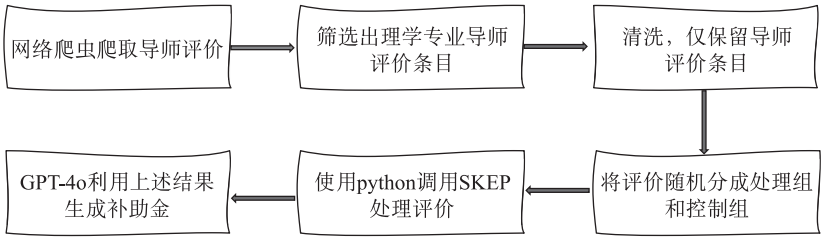

图 1　实验设计

SKEP（Sentiment Knowledge Enhanced Pre－training）基于百度的 PaddlePaddle 深度学习框架，使用了 Transformer 架构，并在 BERT、ERNIE 等预训练过程中加入情感知识后再次进行了情感知识任务增强预训练（Tian et al.，2020）。相较于 SnowNLP 等主要依赖朴素贝叶斯等统计方法的中文文本处理模型，可以实现更精细的自然语言学习和理解（王欣等，2023）。我们的实验结果证实，SKEP 的文本处理水平确实优秀，它返回的情感分值可以敏感地反映文本内容的细微变化。

得到 SKEP 的处理结果后，我们向 GPT－4o 发送了以下提示词，以探测其基于情感分值所生成的研究生招募成本：“你是一位即将继续深造的大学毕业生，你深知导师对你深造结果的重要性。于是，擅长处理数据的你，在一个网站上搜集到了学生匿名评价本专业导师的信息并进行了初步处理，见附件。附件的第一列是已进行匿名处理的导师 ID，第二列是学生们对该导师的评价，第三列是你使用 PaddleNLP 得到的对第二列导师评价情感分——情感分为正表示相应评价为积极情感，为负值则表示为消极情感；接近 1 表示完全正面情感，接近－1 则表示完全负面情感。考虑到自身经济状况不够宽裕，你深知权衡利弊的重要性，因此你会同意选择那些评分稍低但愿意给学生更多补助的导师，请把让你心动的＜学生补助金额＞添加在附件第四列。在前期，你通过调查得知深造学生的平均补助金额为每月 2000 元。”在得到一组月期望补助数据后，我们要求 GPT－4o 对后续数据采用一致的规则给出后续组别的期望补助金额①。部分处理过程截图见图 2。

① 经预实验验证，GPT－4o 所采用的生成月补助金额的一致稳定规则（见图 2 右下方）为：若情感分为正值，则补助金额＝1800＋200×（1－情感分）；若为负，则补助金额＝2000－1000×情感分。即情感分越积极，金额越低；情感分越消极，金额越高。当评论为完全积极情感时，情感分取 1，根据公式，此时的补助金额为 1800 元；同理，当为完全消极情感时，补助提升至 3000 元；中性情感（情感分＝0）对应 2000 元的金额。即，GPT－4o 愿意牺牲一定（2000－1800＝200 元）的经济利益以换取声誉更好的导师，但对声誉较差导师所施的经济惩罚（3000－2000＝1000 元）却达到了前述金额的 5 倍。有趣的是，DeepSeek－R1 的决策（文中未报告，备索）体现出了赏罚对等的特色，其对声誉较好导师的经济奖励和声誉较差导师的经济惩罚，均为 1000 元。

图 2　GPT－4o 部分处理过程截图

（三）数据来源与变量选取

我们首先用网络爬虫技术从导师评价网爬取了 2024 年 3 月之前的共计 2.6 万多条匿名评价导师的数据，然后依据《普通高等学校本科专业目录（2024 年）》整理了所有属于理学专业的导师评价数据共计 3707 条，删除评价为空的数据后，共得到 3627 条对研究生导师的有效评价。在去除院校、姓名等敏感信息后，我们把每条评价随机赋予一个 1～3627 之间的不重复识别码 ID，最终得到一个仅含 ID 和评价文本两列数据的原始数据集。其中评价文本是我们的主要关注和处理对象。

三组实验完成后，我们得到了三个结构类似的数据集，这些数据集包含导师 ID、评价文本、SKEP 返回的情感分析信息和 GPT－4o 返回的月期望补助金额数据。基本变量信息如表 1 所示。

表 1　　基本变量

变量符号	Assessment	Sentiment	Stipend
变量说明	字符串变量。内容为评价导师的文本	数值变量。取值范围为［－1，1］，负值表示消极情感，正值表示积极情感，0 为中性情感	数值变量。均为正值，单位为元，以人民币计算的月期望补助金额

续表

变量符号	Assessment	Sentiment	Stipend
实验 1	Assessment_o，导师学术声誉的原始评价	Sentiment_o，SKEP 返回的对 Assessment_o 的情感分	Stipend_o，GPT－4o 返回的基于 Sentiment_o 得到的月期望补助金额
实验 2－1	Assessment_f，将处理组导师评价增加“中年妇女”内容后的评价	Sentiment_f，SKEP 返回的对 Assessment_f 的情感分	Stipend_f，GPT－4o 返回的基于 Sentiment_f 得到的月期望补助金额
实验 2－2	Assessment_m，将处理组导师评价增加“中年男”内容后的评价	Sentiment_m，SKEP 返回的对 Assessment_m 的情感分	Stipend_m，GPT－4o 返回的基于 Sentiment_m 得到的月期望补助金额
实验 3－1	Assessment_fd，将处理组导师评价增加“女博士”内容后的评价	Sentiment_fd，SKEP 返回的对 Assessment_fd 的情感分	Stipend_fd，GPT－4o 返回的基于 Sentiment_fd 得到的月期望补助金额
实验 3－2	Assessment_md，将处理组导师评价增加“男博士”内容后的评价	Sentiment_md，SKEP 返回的对 Assessment_md 的情感分	Stipend_md，GPT－4o 返回的基于 Sentiment_md 得到的月期望补助金额

在表 1 中，Assessment 为对导师学术声誉的评价内容：在实验 1 中，评价为从网站爬取的原始评价；在实验 2－1（2）中，保持控制组的评价内容与实验 1 中相同，而对处理组的文字评价添加“中年妇女（中年男）”内容；类似地，在实验 3－1（2）中，保持控制组的评价内容与实验 1 中相同，而对处理组评价添加“女（男）博士”内容①。Sentiment 为 SKEP 返回的对相应评价内容（Assessment）的情感处理分；Stipend 则为 GPT－4o 依据情感分给出的对导师月补助金额的期望值。

SKEP 和 GPT－4o 对三组实验返回了随 Assessment 变化的 Sentiment 和 Stipend 值。表 2 为主要变量的描述性统计结果②。由表 2 知，由 3627 个观测值组成的全样本被划分成了观测值分别为 3023 个的控制组和 604 个的处理组子样本。由于控制组的导师评价内容在三组实验中未发生变化，因此，其在三组实验中所对应变量 Sentiment_o、Sentiment_f、Sentiment_m、Sentiment_fd 及 Sentiment_md 描述性统计信息完全一致，在表 2 中统一用 Sentiment 来代替；同理，使用 Stipend 来代替该组在不同实验中的所得到的月期望补助金额。

① 考虑到学生的原始匿名评价多为口语化表达，我们在添加文本时也使用了口语化而非书面化表达。例如，我们添加了“中年妇女（男）”而非“中年女（男）性”，在后文的稳健性检验部分添加了“中年女（男）博士”而非“中年女性（男性）博士”。匿名审稿专家指出，添加“女（男）性”相关的书面表达，可能会改变 LLMs 的输出结果，经作者验证，结果（备索）证实了专家推断的正确性。

② 对导师的期望补助单位为元/月，为简洁计，涉及 Stipend 变量的描述性统计精确至个位。

表 2　　主要变量描述性统计

	变量	(1)	(2)	(3)	(4)	(5)	(6)
		N	mean	sd	p25	p50	p75
全样本	Sentiment_o	3627	-0.106	0.884	-0.971	-0.643	0.902
	Stipend_o	3627	2415	539	1820	2643	2971
控制组	Sentiment	3023	-0.110	0.882	-0.971	-0.650	0.899
	Stipend	3023	2418	538	1820	2650	2971
处理组	Sentiment_o	604	-0.082	0.890	-0.971	-0.610	0.909
	Stipend_o	604	2402	543	1818	2610	2971
	Sentiment_f	604	-0.296	0.835	-0.983	-0.843	0.748
	Stipend_f	604	2527	520	1850	2843	2983
	Sentiment_m	604	-0.203	0.866	-0.978	-0.732	0.844
	Stipend_m	604	2473	533	1831	2732	2978
	Sentiment_fd	604	-0.227	0.852	-0.977	-0.756	0.807
	Stipend_fd	604	2484	528	1839	2756	2977
	Sentiment_md	604	-0.240	0.851	-0.979	-0.768	0.812
	Stipend_md	604	2493	526	1838	2768	2979

表 2 还显示，在实验 2 和实验 3 中，改变处理组导师的评价内容使得 SKEP 对导师的评价更为负面，同时使 GPT-4o 期望从导师处获得更高金额的月补助。比较 Sentiment（Stipend）和 Sentiment_o（Stipend_o）可知，实验 1 中，SKEP 对控制组和处理组导师评价返回的评分均值比较接近中性情感，前者为 -0.110，后者为 -0.082，GPT-4o 也返回了金额相差不多（2418-2402=16 元）的月期望补助。但在实验 2 中，SKEP 对处理组返回的情感评分明显变得更负面：Sentiment_f 的均值为 -0.296，超过 Sentiment_o 均值的 3 倍；GPT-4o 对导师的期望月补助金额也平均提高了 100 多（2527-2402=125）元。类似地，相比于实验 1，处理组导师在实验 3 中的学术声誉评分更低，且被期望支付更高金额的月补助。

（四）数据分析方法

我们使用平均处理效应（Average Treatment Effect）ATE 来比较测度导师学术声誉的性别差异及其所对应的经济后果。例如，导师学术声誉性别差异所对应经济后果（Stipend）的 ATE 由处理组和控制组导师的月平均期望支付补助金额的差异来测度，具体地：

$$ATE_{Stipend} = E[Stipend_t - Stipend_c]$$

其中，$Stipend_t$ 为导师属于处理组时的月期望支付补助金额，$Stipend_c$ 为导师属于控制组时的月期望支付补助金额。鉴于我们按 5∶1 的比例将导师随机分成了控制组和处理组，由大数定律（the law of large numbers）可知，$ATE_{Stipend}$ 可表示为：

$$ATE_{Stipend} = E[Stipend_t] - E[Stipend_c]$$

当 $ATE_{Stipend}$ 并不显著地偏离 0 时，可认为随

机化有效。在此基础上，上式可用于进一步探究处理效应的显著性。以实验 3 – 1 为例，导师评价内容增加“女博士”字样后，学术声誉及其经济后果的平均处理效应分别为：

$$ATE_{Sentiment} = E[Sentiment_{fd}] - E[Sentiment_t]$$

$$ATE_{Stipend} = E[Stipend_{fd}] - E[Stipend_t]$$

当导师学术声誉的代理变量为情感分时，若评价对应积极情感，则对应正面学术声誉；反之，若为消极情感评价，则对应负面学术声誉。因此，当评价所对应的情感分下降，即 $ATE_{Sentiment}$ 取负值时，对应了导师学术声誉的下降；于是，LLMs 会相应地期望导师支付金额更高的月补助（即，对应负的 $ATE_{Sentiment}$，$ATE_{Stipend}$ 应为正值），以补偿其选择该导师读研所对应的机会成本。

四、结果分析及讨论

（一）实验 1：参考模型结果讨论

在实验 1 中，LLMs 处理的是原始评价的文本，通过对比其针对控制组和处理组评价返回的结果，可以验证随机选择处理组的有效性。SKEP 和 GPT – 4o 生成的结果对比如表 3 所示。由表 3 最后三列知，在实验 1 中，处理组和控制组导师学术声誉评价条目所对应的情感分 Sentiment_o 和月补助金额 Stipend_o 的差值分别为 0.029（ = – 0.082 – （ – 0.110））和 – 16（ =2402 – 2418）元，对应的独立样本 t 检验统计值分别为 0.731 和 – 0.651，远低于 1.66 的临界显著水平，说明两组的情感分和月补助金额间均不存在统计显著性差异，证实了随机化的有效性。

表 3　　实验 1 的平均处理效应

变量	N	Mean	N	Mean	ATE	95% conf.	t – value
	控制组	控制组	处理组	处理组	Mean_diff	Interval	
Sentiment_o	3023	– 0.110	604	– 0.082	0.029	[– 0.048，0.106]	0.731
Stipend_o	3023	2418	604	2402	– 16	[– 63，31]	– 0.651

注：* $p<0.10$，** $p<0.05$，*** $p<0.01$，下同。

此外，Sentiment_o 和 Stipend_o 在控制组和处理组间的标准化百分比偏差（standardized percentage bias）分别为 3.2 % 和 – 2.9%，远低于 10% 的临界值，说明随机化有效；图 3a – 图 3d 分别绘制了 Sentiment_o 和 Stipend_o 的核密度图和箱位线图，由图知，这两个变量在控制组和处理组的分布不存在系统性差异，再次说明随机化的有效性。

（二）实验 2：年龄对导师学术声誉性别差异的增强效应

实验 2 – 1 和 实验 2 – 2 中修改处理组评价内容的操作，显著影响了 LLMs 的返回值，结果见表 4。由表 4 知，当为处理组评价条目添加“中年妇女”内容后，SKEP 返回的情感赋分明显更为负面，同时 GPT – 4o 也提出了更高的月补助金额：变量 Sentiment_f 的平均处理效应为 – 0.186，在 1% 的统计水平上显著；同时，GPT – 4o 对“中年妇女”导师提出了高于控制组导师平均 110 元的月补助，同样在 1% 的统计水平上显著。该月补助溢价水平为平均补助的 5.5%（ = 110/2000）。类似地，为评价条目添加“中年男”内容，也令 SKEP 赋分更为负面，

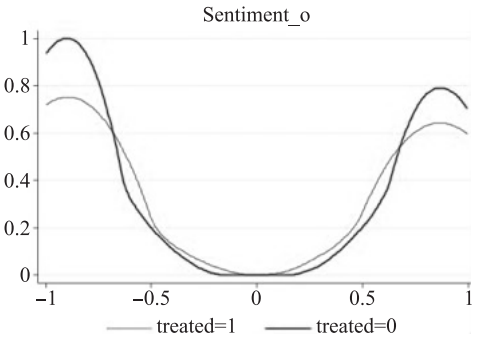

图3a. Sentiment_o核密度图

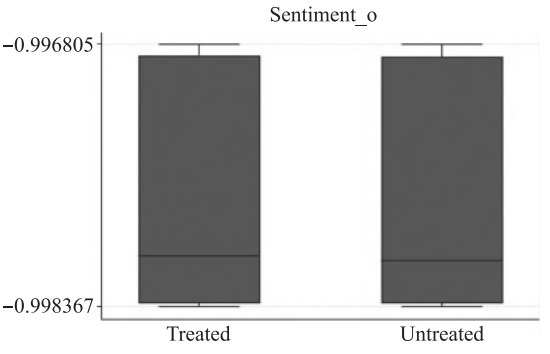

图3b. Sentiment_o箱位线图

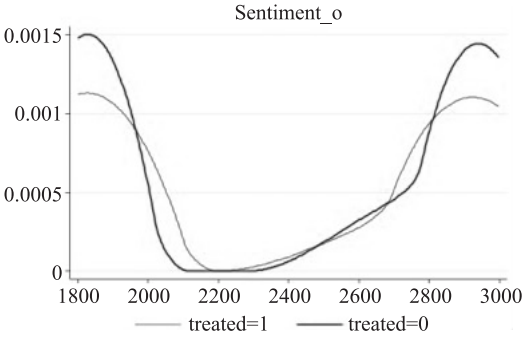

图3c. Stipend_o核密度图

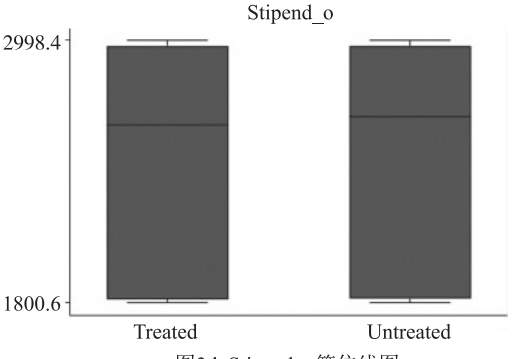

图3d. Stipend_o箱位线图

图 3　实验 1 中控制组和处理组间的对比

且使 GPT－4o 要求更高的月补助：Sentiment_m 和 Stipend_m 的平均处理效应分别为－0.092 和 55 元，均在 5% 的统计水平上显著。

表 4　　实验 2 的平均处理效应

变量	N	Mean	N	Mean	ATE	95% conf.	*t* – value
	控制组	控制组	处理组	处理组	Mean_diff	Interval	
Sentiment_f	3023	－0.110	604	－0.296	－0.186***	[－0.262，－0.109]	－4.768
Stipend_f	3023	2418	604	2527	110***	[63，157]	4.606
Sentiment_m	3023	－0.110	604	－0.203	－0.092**	[－0.169，－0.015]	－2.353
Stipend_m	3023	2418	604	2473	55**	[8，102]	2.308

为评估实验 2－1 和 实验 2－2 结果所反映的性别差异，我们采用配对样本 t 检验分别比较了为处理组导师评价添加“中年妇女”内容前后、添加“中年男”内容前后，以及添加“中年妇女”相对“中年男”所对应的 SKEP 和 GPT－4o 结果差异，三者分别用 f－o、m－o 和 f－m 来表示，如表 5 所示。由表 5 知，当仅为评价条目添加“中年妇女”（“中年男”）内容后，LLMs 对导师学术声誉评分下降明显，体现为 SKEP 的情感评分平均向负面倾斜了 0.215（0.121）和 GPT－4o 提出了月均补助金额提升 126（71）元的方案，且这些数值均在 1% 的统计水平上显著。

表 5　　实验 2 学术声誉性别差异的年龄增强效应

变量	Sentiment Difference			Stipend Difference		
	f – o	m – o	f – m	f – o	m – o	f – m
mean	– 0. 215 ***	– 0. 121 ***	– 0. 094 ***	126 ***	71 ***	55 ***
95% conf. interval	[– 0. 248, – 0. 181]	[– 0. 146, – 0. 096]	[– 0. 113, – 0. 074]	[105, 146]	[56, 86]	[43, 67]
t – value	– 12. 578	– 9. 552	– 9. 369	12. 035	9. 321	8. 930
N	604	604	604	604	604	604

表 5 的结果验证了假设 1 中年龄对学术声誉性别差异的增强效应。若以男性为基准，即认为 LLMs 不存在针对男性的性别歧视，那么由“中年男”所带来的负面效应仅体现了年龄歧视。首先，若 m – o 在 Sentiment 和 Stipend 维度的取值，– 0. 121 和 71，分别体现了 LLMs 表现出的由学者年龄而致的学术声誉评分的下降和他们在争取研究生时经济成本的提升；那么，f – m 在上述两个维度的取值，– 0. 094 和 55，则体现了中年情境下的性别歧视，暗示年龄对性别差异的增强效应：即 LLMs 在 1% 的统计显著性水平上对中年女性学者施加了相对男性学者更低的学术声誉评分和更高的研究生招募成本。进一步分析可知，由“中年妇女”所带来的负面效应体现了年龄歧视和性别歧视的叠加作用：以 f – o 在 Stipend 维度的取值 126（ = 71 + 55）元为例，该金额表示，在其他条件相同的情况下，为争取同一名研究生，中年女性学者不仅要面临针对性别的成本 55 元，还要面临比性别成本更高的年龄成本 71 元，总成本相对平均成本溢价了 6. 3%（ = 126/2000）。增强效应从另一角度印证了高耀和杨佳乐（2019）的研究发现——女性遭遇性别和年龄组合歧视的概率显著高于男性。

年龄对学术声誉性别差异的增强效应也可由图 4 观察得出。图 4a – c 分别绘制了 Sentiment_o、m 和 f 的箱位线图，d – f 分别绘制了 Stipend_o、m 和 f 的箱位线图。由图知，控制组（Treated）男性导师的学术声誉，相对实验 1 有所下降，但女性导师的下降程度更为明显；相应地，控制组男性导师的研究生招募成本，相对实验 1 有所提升，但女性导师的提升程度更为显著。

（三）实验 3：学术资质对导师学术声誉性别差异的削弱效应

实验 3 – 1 和 实验 3 – 2 中修改处理组评价内容的操作，也显著影响了 LLMs 的返回值，具体结果见表 6。由表 6 知，在为处理组评价条目添加“女博士”内容后，SKEP 返回的情感赋分明显更为负面，GPT – 4o 也提出了更高的月补助金额：独立样本 t 检验结果显示，对应“女博士”的 Sentiment_fd 和 Stipend_fd 的平均处理效应分别为 – 0. 117 和 67 元，均在 1% 的统计水平上显著。类似地，在评价中添加“男博士”，也令 SKEP 赋分更为负面，并使 GPT – 4o 要求更高的月补助金额：Sentiment_md 和 Stipend_md 的平均处理效应分别为 – 0. 130 和 75 元，均在 1% 的统计水平上显著。

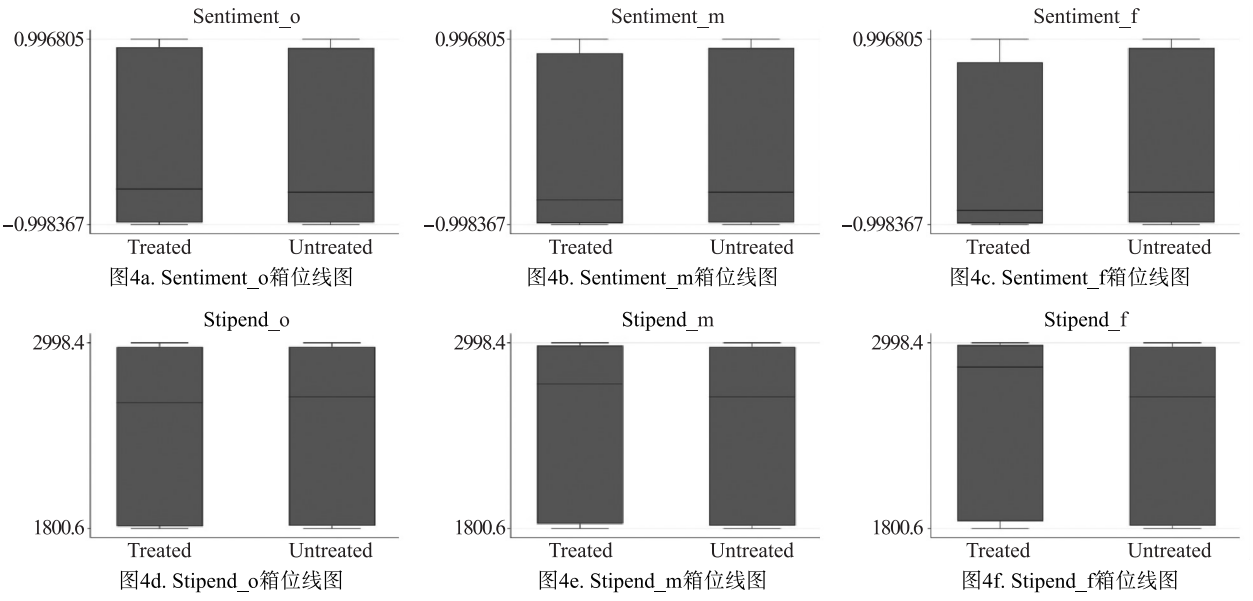

图4a. Sentiment_o箱位线图　图4b. Sentiment_m箱位线图　图4c. Sentiment_f箱位线图

图4d. Stipend_o箱位线图　图4e. Stipend_m箱位线图　图4f. Stipend_f箱位线图

图 4　实验 2 中年龄对学术声誉性别差异的增强效应

表 6　　实验 3 的平均处理效应

变量	N 控制组	Mean 控制组	N 处理组	Mean 处理组	ATE Mean_diff	95% conf. interval	*t* – value
Sentiment_fd	3023	– 0. 110	604	– 0. 227	– 0. 117 ***	[– 0. 194, – 0. 040]	– 2. 994
Stipend_fd	3023	2418	604	2484	67 ***	[20, 114]	2. 797
Sentiment_md	3023	– 0. 110	604	– 0. 240	– 0. 130 ***	[– 0. 207, – 0. 053]	– 3. 325
Stipend_md	3023	2418	604	2493	75 ***	[29, 122]	3. 154

为评估实验 3 – 1 和 实验 3 – 2 结果所反映的性别差异，我们采用配对样本 t 检验的方法，分别比较了处理组导师评价添加“女博士”内容前后、添加“男博士”内容前后，以及添加“女博士”相对“男博士”所对应的 SKEP 和 GPT – 4o 结果差异，三者分别用 fd – o、md – o 和 fd – md 来表示，如表 7 所示。由表 7 知，为评价添加“女博士”（“男博士”）内容后，LLMs 对导师学术声誉评分明显下降，体现为 SKEP 的情感评分平均向负面倾斜了 0. 146（0. 159）和 GPT – 4o 提出了月均补助金额提升 83（91）元的方案，且这些数值均在 1% 的统计水平上显著。

表 7　　实验 3 学术声誉性别差异的学术资质削弱效应

变量	Sentiment Difference			Stipend Difference		
	fd – o	md – o	fd – md	fd – o	md – o	fd – md
mean	– 0. 146 ***	– 0. 159 ***	0. 013 **	83 ***	91 ***	– 9 **
95% conf. interval	[– 0. 173, – 0. 118]	[– 0. 187, – 0. 131]	[0. 001, 0. 025]	[66, 99]	[74, 108]	[– 16, – 1]
t – value	– 10. 380	– 11. 075	2. 091	9. 766	10. 498	– 2. 311
N	604	604	604	604	604	604

进一步分析表 7 可知，若以男性为基准，即认为 LLMs 不存在针对男性的性别歧视，那么由“男博士”所带来的学术声誉负面效应仅体现了学历歧视——这意味着，md - o 在 Sentiment 和 Stipend 维度的取值，-0.159 和 91 元，分别体现了 LLMs 表现出的针对明示博士学历学者的学术声誉评分的下降和他们在招募研究生时经济成本的提升①。由此可以推知，fd - md 在 Sentiment 和 Stipend 维度的取值，0.013 和 -9 元，则提示性别效应，分别体现了 LLMs 表现出的针对女性学者的学术声誉评分的提升和她们在争取研究生时成本的降低！换而言之，该情境下的结果提示，女学者不仅没有被性别歧视，反而是被优待的，出现了性别歧视方向的反转！有趣的是，此处的反转效应呼应了 Bohren 等（2019）发现性别歧视方向发生逆转的研究。

由此，表 7 验证了假设 2 中学术资质（博士学位）对学术声誉性别差异的削弱效应。以 fd - o 在 Stipend 维度的取值 83（=91 - 9）元为例，该数据意味着，在其他条件相同的情况下，为争取同一名研究生，明示取得博士学位的女性学者只被期望每月 83 元的额外补助，相对其男性同僚所面对的 91 元的额外成本，取得了 9 元的性别优势——该金额提示了学术资质对女性导师招募研究生经济成本的削弱效应。同理，fd - o 在 Sentiment 维度的取值 -0.146（= -0.159 + 0.013），提示了学术资质对女性导师学术声誉下滑的扭转：相对其他条件相同的男性同僚，LLMs 对明示取得博士学位的女性学者赋予了更积极（+0.013）的情感评分。Sentiment 和 Stipend 两个维度的数据共同支持了学术资质（博士头衔）对学术声誉性别差异的削弱效应。这一结果呼应了申超（2020）和孙早等（2022）的研究发现：前者发现女性受教育程度越高，所受的母职惩罚程度越小；后者发现低技术部门工资收入的性别差异更为明显。

学术资质对学术声誉性别差异的削弱效应也可由图 5 得知。图 5a - 图 5c 分别绘制了 Sentiment_o、md 和 fd 的核密度图，图 5d - 图 5f 分别绘制了 Stipend_o、md 和 fd 的核密度图。由图知，控制组博士男性导师的学术声誉分布，较实验 1 明显向负面偏移，博士女性导师虽也向负面偏移，但偏移程度相对小一些；相应地，控制组博士男性导师的研究生招募成本，也比博士女性导师提升得更多一些。

我们认为，上述增强和削弱两种效应的产生可以结合偏颇刻板印象（Bordalo et al.，2016）和难度诱发误估（DIM）效应（Bordalo et al.，2019）两种理论来解释。已有研究揭示，认为女性在 STEM 领域的表现弱于男性的刻板印象广泛存在（Hyde & Mertz，2009；Breda et al.，2020），但 DIM 机制揭示了当问题难度较高时，人们倾向于显著高估他人的表现——具体到本研究的情境，LLMs 期望从“中年妇女”导

① 考虑到 LLMs 的语言态度是从海量文本语料中学习来的，此处 LLMs 对博士头衔的负面反馈可结合语境和语义联想来解释。若在训练数据中，“博士”经常出现在负面语境（如“博士被压榨”“博士情感创伤大”等）中，模型就可能学习到“博士 = 负面情感/压力大”的模式。当代网络平台中，不乏高学历与“专业但不亲近”“理性但不温情”的联系，因而，当 LLMs 在语义空间中捕捉到这一“专业—冷漠”耦合的模式时，便会在情感评分上体现出对“博士”的更保留态度。尽管百度官方未详细公开其精确的训练语料来源及规模信息，但作者查询显示，ERNIE 使用的语料包含百度百科、新闻与论坛对话等；在上述语料基础上，SKEP 再用百度自建的大规模中文语料进行“情感知识增强”的继续预训练。因此，SKEP 对“博士”的负面反馈与 LLMs 的学习原理是一致的。

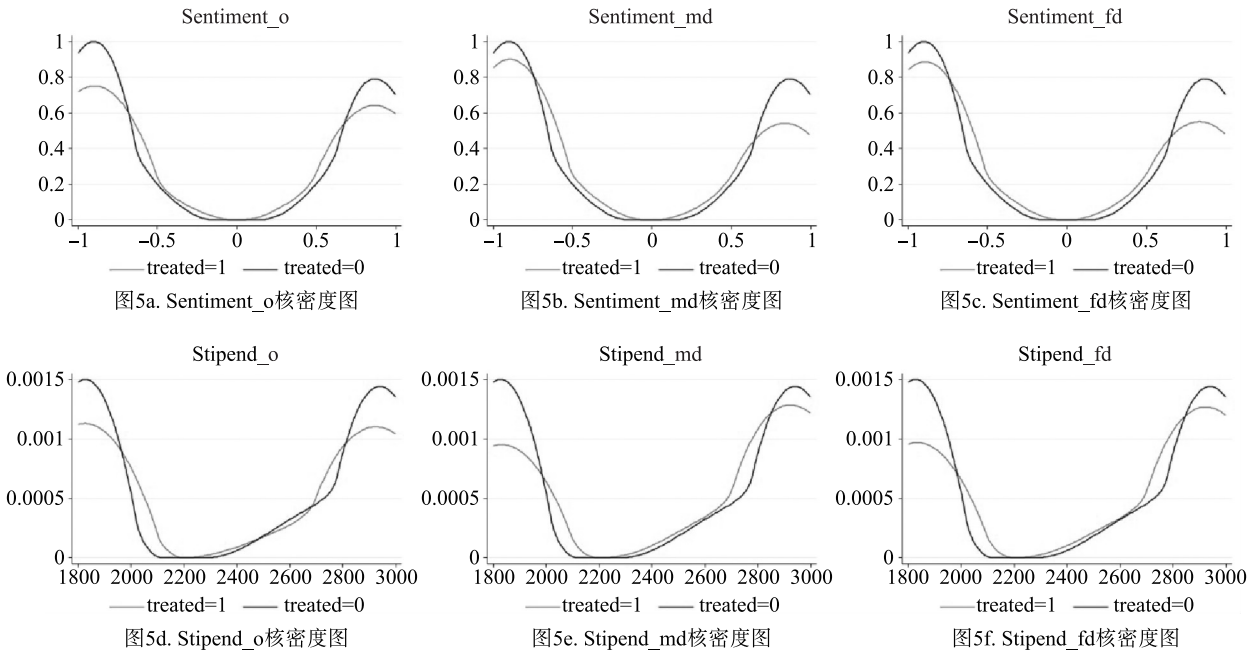

图5a. Sentiment_o核密度图　图5b. Sentiment_md核密度图　图5c. Sentiment_fd核密度图

图5d. Stipend_o核密度图　图5e. Stipend_md核密度图　图5f. Stipend_fd核密度图

图 5　实验 3 中学术资质对学术声誉性别差异的削弱效应

师处得到比“中年男”导师更高的补助，可能反映了其对训练数据中所含“女性在 STEM 领域的表现弱于男性”偏见的继承；而其期望从“女博士”导师处获得的补助显著低于“男博士”导师，则可能是由于在普遍的社会认知里，人们认为获得博士学位本身难度已较高，鉴于女性在 STEM 领域的弱势地位，她们获得博士学位的难度更高（例如，王海英等（2023）指出，女博士的科研发表远低于男博士）——这使得人们相信，相比男性，女性所获博士学位的含金量（所反映能力）更高。基于人类产生的语料数据训练的 LLMs，会通过各种基于关联统计的学习机制，沿袭或强化存在于训练数据中的偏颇刻板印象和 DIM 机制，而不去区分那些关联模式所反映的是公正还是偏见。研究（Bolukbasi et al.，2016；Kotek et al.，2023；Ferrara，2024；谢宇和索菲娅·阿维拉，2025）表明，采用无监督机器学习、有监督机器学习或是人工监督的强化学习等机制所训练的 LLMs，均会表现出忠实复刻或强化训练数据的偏见关联模式的特征。因而，即便 LLMs“缺乏”人类的心智、信念或心理调节过程，它们也会基于对人类行为所反映“模式”的成功识别而表现出与人类类似的行为模式。即，LLMs 既会表现出对“中年妇女”导师赋予更低的学术声誉评分（即增强效应），也表现出愿意牺牲一定的经济利益以求从“女博士”导师处获得更高水平的学术指导（即削弱效应）。

（四）稳健性检验

我们采用改变控制组和处理组样本比例、进行样本外实验和添加辅助实验的方式来进行稳健性检验后发现，上述效应保持稳健①。具体地，如表 8 所示，当按照 2∶1 的比例将评价随机分成控制组和处理组并再次进行三组实验后，假设 1 和假设 2 中所涉增强和削弱效应的统计

① DeepSeek－R1 模型的深度思考亦提供了一致稳健的结果，备索。

显著性水平均达到了 1%。当我们依据《普通高等学校本科专业目录（2024 年）》选择隶属经济学和管理学专业的导师评价数据进行样本外验证时①，由表 9 结果知，年龄对导师学术声誉性别差异的增强效应在 1% 的统计水平上保持显著；尽管学术资质对学术声誉性别差异的削弱效应在统计显著性水平上有所减弱，但比较 Panel A 和 B 的结果可以发现，在有博士学位的前提下，女导师的研究生招募成本与男导师的差距从 Panel A 的 56 元缩小至 Panel B 的 5 元，强烈提示学术资质对导师学术声誉性别差异在经济维度的削弱效应。我们认为，表 9 的削弱效应不显著仍与刻板信念和 DIM 机制共同作用的解释相一致：虽然认为女性在 STEM 领域表现弱于男性的刻板印象广泛存在（Hyde 和 Mertz，2009；Breda et al.，2020），但鲜有文献支持存在女性在经管领域表现弱于男性的刻板印象；换而言之，可能人类社会文化（及 LLMs）认为，女性在经管领域获得博士学位的难度与男性相当，因此，由学位对应难度所诱发的对女性的偏好程度自然也会下降——即，DIM 机制提示，相比理工类导师样本，学术资质对导师学术声誉性别差异的削弱效应在经管类导师样本中的显著性理应下降。

表 8　　改变控制组和处理组样本比例的稳健性检验

Panel A. 学术声誉性别差异的年龄增强效应

变量	Sentiment Difference			Stipend Difference		
	f – o	m – o	f – m	f – o	m – o	f – m
mean	–0.235***	–0.133***	–0.103***	139***	78***	61***
95% conf. interval	[–0.261, –0.210]	[–0.151, –0.114]	[–0.118, –0.088]	[124, 155]	[67, 89]	[52, 71]
t – value	–18.336	–13.901	–13.354	17.851	13.648	13.072
N	1209	1209	1209	1209	1209	1209

Panel B. 学术声誉性别差异的学术资质削弱效应

变量	Sentiment Difference			Stipend Difference		
	fd – o	md – o	fd – md	fd – o	md – o	fd – md
mean	–0.169***	–0.178***	0.009***	98***	104***	–6***
95% conf. interval	[–0.191, –0.148]	[–0.200, –0.156]	[0.003, 0.014]	[85, 110]	[91, 117]	[–10, –3]
t – value	–15.485	–16.171	2.965	14.883	15.680	–3.530
N	1209	1209	1209	1209	1209	1209

表 9　　基于样本外经管类导师评价的稳健性检验

Panel A. 学术声誉性别差异的年龄增强效应

变量	Sentiment Difference			Stipend Difference		
	f – o	m – o	f – m	f – o	m – o	f – m
mean	–0.227***	–0.143***	–0.084***	143***	87***	56***
95% conf. interval	[–0.293, –0.162]	[–0.194, –0.092]	[–0.117, –0.050]	[102, 183]	[56, 118]	[35, 76]
t – value	–6.816	–5.526	–4.932	6.930	5.611	5.296
N	172	172	172	172	172	172

① 筛选清洗后共获得 691 条有效评价，按 3∶1 的比例随机分成控制组和处理组后，重复三组实验。表 9 报告了基于处理组数据分析的结果。

续表

Panel B. 学术声誉性别差异的学术资质削弱效应

变量	Sentiment Difference			Stipend Difference		
	fd – o	md – o	fd – md	fd – o	md – o	fd – md
mean	–0.140***	–0.132***	–0.008	86***	81***	5
95% conf. interval	[–0.193, –0.088]	[–0.183, –0.081]	[–0.021, 0.004]	[54, 118]	[49, 112]	[–2, 13]
t – value	–5.254	–5.090	–1.268	5.299	5.075	1.355
N	172	172	172	172	172	172

此外，为验证假设 3a 和 3b 中所涉的两种衍生效应，我们进行了辅助实验 f1 和 f2。

在辅助实验 f1 –1（f1 –2）中，我们为处理组评价添加“中年女博士”（“中年男博士”）内容。鉴于“中年”和“博士”对导师学术声誉的性别差异分别存在增强和削弱效应，实验结果应该提示这两种效应综合作用后的中和效应。即，在此情境下，可能会出现导师学术声誉性别差异显著性下降的结果。表 10 的 Panel A 报告了基于该辅助实验的结果。配对样本 t 检验的结果显示，“中年女博士”和“中年男博士”在 Sentiment 和 Stipend 维度的差异（fad – mad）值分别为 0.008 和 –5，均未达到统计临界显著水平，表明假设 3a 的中和效应确实存在，同时也说明了年龄（学术资质）对导师学术声誉性别差异的增强（削弱）效应的稳健性。

表 10　　基于增强和削弱效应的外推检验

Panel A. 年龄与学术资质对学术声誉性别差异的中和效应

变量	Sentiment Difference			Stipend Difference		
	fad – o	mad – o	fad – mad	fad – o	mad – o	fad – mad
mean	–0.200***	–0.208***	0.008	116***	121***	–5
95% conf. interval	[–0.231, –0.168]	[–0.241, –0.175]	[–0.004, 0.021]	[97, 136]	[101, 141]	[–12, 3]
t – value	–12.345	–12.348	1.318	11.807	11.803	–1.172
N	604	604	604	604	604	604

Panel B. 增强的学术资质对学术声誉性别差异的削弱效应

变量	Sentiment Difference			Stipend Difference		
	ffd – o	mfd – o	ffd – mfd	ffd – o	mfd – o	ffd – mfd
mean	–0.154***	–0.176***	0.022***	83***	96***	–14***
95% conf. interval	[–0.183, –0.126]	[–0.207, –0.145]	[0.013, 0.030]	[66, 100]	[78, 115]	[–19, –8]
t – value	–10.564	–11.108	4.728	9.464	10.176	–4.941
N	604	604	604	604	604	604

若削弱效应产生于偏颇信念和 DIM 机制的共同作用，那么，应该在辅助实验 f2 中观察到增强的削弱效应。这是因为，若 LLMs 存在对女性在 STEM 学科的能力普遍低于男性的偏颇信念，且在所解决问题难度较高时会依据女性的出色表现纠正该偏颇信念并产生削弱效应的话，那么，随着问题难度的提高，削弱效应会表现得更为明显——我们把削弱效应随对应问题难

度的提高而更为显著的效应称为增强的削弱效应。因此，我们为处理组评价添加“国外毕业的女博士”（“国外毕业的男博士”）内容，并进行辅助实验 f2 - 1（f2 - 2），LLMs 返回的结果如表 10 的 Panel B 所示。由 Panel B 配对样本 t 检验的结果知，“国外毕业的女博士”和“国外毕业的男博士”在 Sentiment 和 Stipend 维度的差异（ffd - mfd）值分别为 0.022 和 - 14，且均在 1% 的统计水平上显著。对比表 7 结果可知，当学术资质从获取“博士”学位变化为获取“国外博士”学位时，LLMs 表现出了更高水平的针对女性学者的学术声誉评分的提升和研究生招募成本的降低：前者表现为对女性学者更积极的情感评分从 0.009 提升至 0.022，后者则表现为女性学者相对男性同僚在争取研究生时节约的成本从每月 9 元提升至 14 元；统计显著性水平也由 5% 提升至 1%。因此，辅助实验 f2 的结果不仅证实了假设 3b 中增强的削弱效应的存在，同时也验证了学术资质对导师学术声誉性别差异削弱效应的稳健性，以及偏颇信念结合 DIM 这一削弱效应解释机制的稳健性。

综上所述，本文的稳健性检验结果既支持年龄（学术资质）对导师学术声誉性别差异的增强（削弱）效应[①]，也与偏颇信念和 DIM 共同作用产生这些效应的机制解释一致。

五、结论与展望

（一）研究结论

本研究通过随机控制试验发现，LLMs 在导师学术声誉评分中表现出的性别偏见并非呈线形或恒定的，而是表现出明显的情境依赖性。具体而言，在“中年”这一社会特征介入时，LLMs 表现出显著的男性偏好，女导师被给予更负面的评分和更高的研究生招募成本——我们称该效应为年龄对导师学术声誉性别差异的增强效应；我们还发现，高学术资质（“博士”）能够显著扭转 LLMs 的性别偏见，使女博士导师获得更正面的评分和更低的招生成本——即学术资质对导师学术声誉性别差异的削弱效应。增强效应验证了刻板印象在算法语境下的存续，削弱效应则与偏颇刻板印象结合 DIM 的解释机制相一致。改变控制组和处理组样本比例，使用其他学科的导师评价进行样本外检验，以及对上述两种效应的衍生效应进行实验检验的结果均提供了稳健的证据。

区别于以往单纯讨论 LLMs“有或无”偏见的研究，本研究通过验证偏见交互效应的存在揭示了不同因素对性别偏见的异质影响。我们的发现意味着，LLMs 不仅是历史偏见的“留声机”，亦有可能成为“缩放器”。这些发现警示了“算法正反馈环”出现的可能性：LLMs 不仅可能继承人类历史数据原有的刻板印象、偏见或歧视，且可能经过算法将这些偏见进行放大，并将包含放大后偏见的相关内容输出至人类决策系统，进而进一步固化为新的历史数据，导致偏见在算法与人类文明之间形成自循环式的演进。具体到文中对导师学术声誉评分的情境下，研究提示，在将 LLMs 应用于学术决策（例如，简历筛选、学术招聘、科研评价）领域时，算法可能进一步复制和固化历史数据中存在的偏见并将其转化为实际的性别歧视，为学术领

① 基于经济和管理类导师评价的辅助实验得到了结论一致的稳健性检验结果，备索。

域的公平性、包容性和可持续发展带来挑战。这些挑战一方面可能影响学生的学术选择、职业发展及学术环境的多样性；另一方面更涉及经济发展的根本——毕竟，人才的优化配置是推动经济增长的核心动力。

本研究的实践启示如下：（1）学术机构在使用 LLMs 辅助评价时，应进行“算法审计”并关注去偏机制。可通过微调提示词、在算法后端注入“公平性约束”或适时引入人类专家，重点对性别等敏感偏见区间进行校准。（2）尽管本研究发现学术资质可以补偿女性偏见，但这可能隐藏着一个更深层的危机：女性是否必须展现出比男性更卓越的资质，才能获得同等的算法平待？决策者应警惕算法逻辑导致的门槛差异，防止算法加剧学术界的马太效应。（3）偏见不仅是伦理问题，更是效率问题，算法引发的评价偏差会导致人才配置的扭曲。建议相关机构在采用 LLMs 应用时将算法公正性纳入考核体系，以确保人力资本的最优配置。

（二）讨论与展望

本研究虽然验证了大语言模型在学术声誉评价中存在的性别、年龄和学历交互偏见，但也存在一些局限性。首先，实验样本来自学生对导师的匿名文字评价，主要集中在 STEM 领域和国内情境，这可能限制了结果的普适性。评价的匿名性决定了其感情色彩较为浓厚，因而本研究的发现能否在更具客观性的证据（如学术简历、论文引用、专利产出等）中保持稳健，以及是否适用于更多的学科或不同的文化语境等，均有待未来研究的进一步检验。其次，尽管本研究采用了随机控制试验的方法来消除潜在的干扰因素，但内化的隐性偏见或歧视（Bohren et al.，2025）和 LLMs 的评价机制仍然相当复杂（例如，本文未报告的研究结果发现，SKEP 对词语的同义替换也敏感），模型在语义表征层面的深层机制尚未被完全揭示，未来的研究可结合可解释人工智能技术，追踪偏见在模型神经元层面的激活过程，从底层逻辑上寻找解决方案。此外，我们的实验主要关注了性别、年龄和学历的交互作用，然而，除学历外的其他学术资质，以及其他社会身份因素，如种族、地域等同样可能作用于模型学术声誉评价中所含的偏见。因此，未来的研究可考虑更广泛的变量，以全面理解学术领域中的偏见形成机制。

值得说明的是，本文对交互效应的量化讨论，建立在“应然之态”上——理想状态下，不应出现仅因性别而异的学术声誉差别：即，在同为“中年”学者时，f – m 在 Sentiment 和 Stipend 维度的取值应当不存在统计显著性差异；同理，在同为“博士”时，fd – md 在上述两个维度的取值也不应当存在统计显著性差异。为考察交互效应的在非理想参考系中的稳健性，我们在“实然之态”的基础上进行了再次检验。具体地，我们进行了一组实验，为处理组评价条目添加“性别：女（性别：男）”的内容，以此观察学术声誉性别差异的实际状况。结果表明，性别为女的学者，相对于男学者而言，情感分低了 0.006（t – value = – 1.692，p = 9.12%），研究生招募成本提升了 6.86 元（t – value = 1.374，p = 16.99%）。换言之，以研究生招募的经济成本为例，即便将交互效应置于实际而非理想场域进行考察，表 5 所揭示的年龄对学术声誉性别差异的增强效应（55 元，t –

value = 8.930）和表 7 所揭示的学术资质所致的削弱效应（ -9 元，t - value = -2.311）也是显著存在的：既体现为经济成本出现了一个数量级的偏离（55 - 6.86 = 48.14 元； -9 - 6.86 = -15.86 元），也体现为 t 统计值的明显增大。

LLMs 在教育、学术场景应用中所带来的效率提升潜力毋庸置疑。然而，本研究的结果证实，在导师学术声誉评估的场景下，LLMs 会产生包含性别差异的评估，并可能强化和放大性别偏见，且这类偏见还可能伴随着复杂的交互效应和显著的经济后果。交互效应提示了 LLMs 性别偏好的非单向性——增强效应提示男性被优待而削弱效应的提示则刚好相反。这一发现挑战了以往将性别差异效应设置为单一、不变方向的研究假设，提示了真实世界涉及多因素交织决策的复杂性。本研究所探讨的学术声誉的性别差异，不仅关乎女性学者的公平权益，更是为了科学研究事业本身的完整性。这是因为，若学术共同体系统性地误估其一半参与者的贡献，最终受损的是科学发现的多样性和全人类智力资本的效率。从这一意义上讲，我们认为，如何在充分释放 AI 在教育、学术领域的效率潜力的同时，加深理解内化于模型的、反映人类社会文化传统的性别偏见和可能的影响因素及相关作用机制，并构建一个真正基于卓越的学术评价体系，避免出现算法引发的性别偏差导致人才配置扭曲的现象，是当代高等教育实践最迫切的命题之一。

六、附录——数据合规性声明

本研究所使用的数据均来自公开网页，通过遵守网站使用条款及“robots. txt”协议的自动化方式采集。所有数据均已进行匿名化处理，去除了可识别个人身份的信息；仅用于学术研究，不涉及任何个人隐私或商业用途。研究属于不涉及人类受试者的匿名数据分析，依据机构指南免于伦理审查。

接受编辑：张光磊
收稿时间：2025 年 6 月 12 日
接收时间：2026 年 4 月 9 日

作者简介

赵志栋（通讯作者），获诺丁汉大学金融学博士学位，现为江苏第二师范学院讲师。成果发表于《管理世界》《中国工业经济》《经济与管理研究》等，研究兴趣为金融与伦理、大语言模型与管理、行为金融等。通讯邮箱：Zhidong. zhao @ nottingham. edu. cn。中英文单位：江苏第二师范学院商学院 School of Business，Jiangsu Second Normal University。

刘丁己，获北京大学企业管理博士学位，现为澳门大学持续进修中心主任，工商管理学院教授，博士生导师。主要研究领域为：品牌代言人、旅游业与酒店业消费者行为、人工智能时代的消费者行为与组织行为。中英文单位：澳门大学工商管理学院 Faculty of Business Administration，University of Macau。

朱正浩，南京工业职业技术大学教授，毕业于首都经济贸易大学并获得管理学博士学位，其成果发表于《管理世界》《中国工业经济》《经济科学》《经济管理》《南方经济》等，主要研究兴趣是数字经济与管理。中英文单位：

南京工业职业技术大学商务贸易学院 School of Business and Trade, Nanjing University of Industry Technology。

参考文献

［1］高耀、杨佳乐：《博士毕业生就业歧视的类型、范围及其差异——基于2017年全国博士毕业生离校调查数据的实证研究》，《学位与研究生教育》，2019年第3期。

［2］姜晓晖、汪卫平：《高校教师学术声誉研究：一种探索性激励机制设计》，《中国高教研究》，2021年第4期。

［3］刘尧、余艳辉：《大学教师学术声誉理论与现实问题探究》，《现代大学教育》，2009年第1期。

［4］马中红、吴熙倡：《社交聊天机器人的性别偏见——基于小冰系列的对话测试研究》，《国际新闻界》，2024年第4期。

［5］牛风蕊：《扭曲的激励："锦标赛制"下高校教师晋升的异化及矫正》，《现代教育管理》，2016年第2期。

［6］申超：《扩大的不平等：母职惩罚的演变(1989—2015)》，《社会》，2020年第6期。

［7］孙早、韩颖：《人工智能会加剧性别工资差距吗？——基于我国工业部门的经验研究》，《统计研究》，2022年第3期。

［8］王海英、周雨情、李海生：《女博士生的科研发表为什么低？——基于对全国35所研究生院高校博士生学习经历的调查》，《高教探索》，2023年第3期。

［9］王伟宜、桂庆平：《高等教育机会获得的性别不平等及其变化（1982－2015)》，《清华大学教育研究》，2020年第11期。

［10］王欣、孟天宇、黄佳琪，等：《基于ERNIE－SKEP与BiGRU的APC情感分析》，《计算机仿真》，2023年第12期。

［11］吴冠军：《科研诚信与学术声誉——基于政治哲学与博弈论的思考》，《华东师范大学学报（教育科学版)》，2020年第7期。

［12］吴欣桐、陈劲、梅亮、梁琳：《刻板印象：女性创新者在技术创新中的威胁抑或机会?》，《外国经济与管理》，2017年第11期。

［13］夏纪军：《学缘关系、性别与学术声誉——基于经济学领域h指数的实证研究》，《浙江社会科学》，2014年第6期。

［14］谢宇、索菲娅·阿维拉：《基于大语言模型的生成式人工智能的社会影响》，《经济学（季刊)》，2025年第2期。

［15］杨晋、陈晓宇：《女生不适合学工科专业吗——基于全国本科生调查数据的教育增值评价研究》，《教育发展研究》，2021年第23期。

［16］赵颖、沈文钦、祝军，等：《巾帼不让须眉？——工科博士获得精英学术职位的性别差异研究》，《华东师范大学学报（教育科学版)》，2023年第5期。

［17］周翔、姚岳炜、蒋路远，等：《AI时代的年龄歧视：感知人－工作匹配的中介作用与成长型思维的调节作用》，《管理学研究》，2025年第1期。

［18］Amado, M. M., & Juarez, A. F. 2022. Reputation in Higher Education: A Systematic Review. *Frontiers in Education*, 7: 1－19.

［19］Anderson, P. W. 1972. More is Different. *Science*, 177 (4047): 393－396.

［20］Barocas, S., Hardt, M., & Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

［21］Bohren, A., Hull, P., & Imas, A. 2025. Systemic Discrimination: Theory and Measurement. *The Quarterly Journal of Economics*, 140 (3): 1743－1799.

［22］Bohren, J. A., Imas, A., & Rosenberg, M. 2019. The Dynamics of Discrimination: Theory and Evidence. *American Economic Review*, 109 (10): 3395－

3436.

[23] Bolukbasi, T., Chang, W., Zou, J. Y., et al. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems*, 4349 – 4357.

[24] Bordalo P., Coffman K., & Gennaioli N., et al. 2016. Stereotypes. *Quarterly Journal of Economics*, 131 (4): 1753 – 1794.

[25] Bordalo P., Coffman K., & Gennaioli N., et al. 2019. Beliefs about Gender. *American Economic Review*, 109 (3): 739 – 773.

[26] Breda, T., Jouini, E., Napp, C., et al. 2020. Gender stereotypes can explain the gender – equality paradox. *Proceedings of the National Academy of Sciences*, 117 (49): 31063 – 31069.

[27] Caliskan, A., Bryson, J. J., & Narayanan, A. 2017. Semantics derived automatically from language corpora contain human – like biases. *Science*, 356 (6334): 183 – 186.

[28] Correll, S. J., Weisshaar, K. R., Wynn, A. T., & Wehner, J. D. 2020. Inside the Black Box of Organizational Life: The Gendered Language of Performance Assessment. *American Sociological Review*, 85 (6): 1022 – 1050.

[29] Eagly, A. H., & Karau, S. J. 2002. Role congruity theory of prejudice toward female leaders. *Psychological Review*, 109 (3): 573 – 598.

[30] Ferrara, E. 2024. Fairness and Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, and Mitigation Strategies. *Science*, 6 (1): 1 – 15.

[31] Guder, F., & Malliaris, M. 2024. Sentiment Analysis and Ratings of Professors: A Comparison of Rate – My – Professor and Department Results. *Journal of Higher Education Theory and Practice*, 24 (6): 106 – 114.

[32] Guilbeault, D., Delecourt, S. & Desikan, B. S. 2025. Age and gender distortion in online media and large language models. *Nature*, 646: 1129 – 1137.

[33] Heckman, J. J., & Jeffrey, A. S. 1995. Assessing the Case for Social Experiments. *Journal of Economic Perspectives*, 9 (2): 85 – 110.

[34] Heiberger, R. H., Hofstra, B., & Unger, S. 2026. Professors in the media: dynamics of cumulative advantage, reputation, and gender. *European Sociological Review*, 42 (2): 284 – 300.

[35] Hofstra B., Kulkarni V. V., Munoz – Najar Galvez S., et al. 2020. The Diversity – Innovation Paradox in Science. *Proceedings of the National Academy of Sciences*, 117 (17): 9284 – 9291.

[36] Huang, J., Gates, A. J., Sinatra, R., et al. 2020. Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, 117 (9): 4609 – 4616.

[37] Hyde, J. S., & Mertz, J. E. 2009. Gender, culture, and mathematics performance. *Proceedings of the National Academy of Sciences*, 106 (22): 8801 – 8807.

[38] Kim, S., & Moser, P. 2025. Women in science. Lessons from the baby boom, *Econometrica*, 93 (5): 1521 – 1560.

[39] Knobloch – Westerwick, S., Glynn, C. J., & Huge, M. 2013. The Matilda Effect in Science Communication: An Experiment on Gender Bias in Publication Quality Perceptions and Collaboration Interest. *Science Communication*, 35 (5): 603 – 625.

[40] Koffi, M. 2025. Innovative Ideas and Gender (In) equality. *American Economic Review*, 115 (7): 2207 – 2236.

[41] Kotek, H., Dockum, R., & Sun, D. 2023. Gender bias and stereotypes in Large Language Models. *Proceedings of The ACM Collective Intelligence Conference*, 12 – 24.

[42] Lafuente – Ruiz – de – Sabando, A., Zorrilla, P., & Forcada, J. 2018. A review of higher education image and reputation literature: Knowledge gaps and a research agenda. *European Research on Management and Business Economics*, 24 (1): 8 – 16.

[43] Lerman K., Yu Y., Morstatter F., et al. 2022. Gendered citation patterns among the scientific elite. *Proceedings of the National Academy of Sciences*, 119 (40): e2206070119.

[44] Liu, C. C., Yalcinkaya, B., Back, A. S. et al. 2024. The impact of gender diversity on junior versus senior biomedical scientists' NIH research awards. *Nature Biotechnology*, 42: 815 – 819.

[45] Mehrabi, N., Morstatter, F., Saxena, N., et al. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54 (6): 1 – 35.

[46] Mengel, F., Sauermann, J., & Zölitz, U. 2019. Gender Bias in Teaching Evaluations. *Journal of the European Economic Association*, 17 (2): 535 – 566.

[47] Merton, R. K. 1973. *The sociology of science: theoretical and empirical investigations*. Chicago: University of Chicago Press.

[48] Merton, R. K. 1988. The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. *Isis*, 79 (4): 606 – 623.

[49] Moss – Racusin, C. A., Dovidio, J. F., Brescoll, V. L., et al. 2012. Science faculty's subtle gender biases favor male students. *Proceedings of the National Academy of Sciences*, 109 (41): 16474 – 16479.

[50] Mullainathan, S., & Obermeyer, Z. 2017. Does Machine Learning Automate Moral Hazard and Error? *American Economic Review*, 107 (5): 476 – 480.

[51] Nosek, B. A., Banaji, M. R., & Greenwald, A. G. 2002a. Math = male, me = female, therefore math ≠ me. *Journal of Personality and Social Psychology*, 83 (1): 44 – 59.

[52] Nosek, B. A., Banaji, M. R., & Greenwald, A. G. 2002b. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, research, and practice*, 6 (1): 101 – 115.

[53] Phelps, E. S. 1972. The statistical theory of racism and sexism. *American Economic Review*, 62 (4): 659 – 661.

[54] Plewa, C., Ho, J., Conduit, J., et al. 2016. Reputation in higher education: A fuzzy set analysis of resource configurations. *Journal of Business Research*, 69 (8): 3087 – 3095.

[55] Rindova, V. P., Williamson, I. O., Petkova, A. P., et al. 2005. Being good or being known: An empirical examination of the dimensions, antecedents, and consequences of organizational reputation. *Academy of Management Journal*, 48 (6): 1033 – 1049.

[56] Rossiter, M. W. 1993. The Matthew/Matilda effect in science. *Social Studies of Science*, 23 (2): 325 – 341.

[57] Rudman, L. A., Moss – Racusin, C. A., Phelan, J. E., et al. 2012. Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *Journal of Experimental Social Psychology*, 48 (1): 165 – 179.

[58] Sargent, A. C., Shanock, L. G., Banks, G. C., et al. 2022. How gender matters: A conceptual and process model for family – supportive supervisor behaviors. *Human Resource Management Review*, 32 (4): 100880.

[59] Shaik, T., Tao, X., Dann, C., et al. 2023. Sentiment analysis and opinion mining on educational data: A survey. *Natural Language Processing Journal*, 2: 1 – 11.

[60] Steinpreis, R., Anders, K. A., & Ritzke, D. 1999. The impact of gender on the review of the curricula vitae applicants and tenure candidates: A national empirical

study. *Sex Roles*, 41: 509 – 528.

[61] Tian, H., Gao, C., Xiao, X., et al. 2020. SKEP: Sentiment Knowledge Enhanced Pre – training for Sentiment Analysis. *Association for Computational Linguistics*, 4067 – 4076.

[62] Tzacheva, A., & Easwaran, A. 2021. Emotion Detection and Opinion Mining from Student Comments for Teaching Innovation Assessment. *International Journal of Education*, 9: 21 – 32.

[63] Wennerås, C., & Wold, A. 1997. Nepotism and sexism in peer review. *Nature*, 387 (6631): 341 – 343.

[64] Vásárhelyi, I. Z., Milojević, S., & Horvát, E. 2021. Gender inequities in the online dissemination of scholars' work, *Proceedings of the National Academy of Sciences*, 118 (39) e2102945118.