国际汉语教学学报（2026）
第 07 卷 · 第 03 期 182-193 https://doi.org/10.46451/ijclt.260414

大语言模型的教育伦理评估研究——以海外国际中文教学机构评测为例

肖 锐 *
刘 玲
杨 蓉
云南大学，中国

张邝弋
北京语言大学，中国

摘要
为探究大语言模型（LLM）在国际中文教育场景中的内容输出倾向及其潜在伦理风险线索，本研究选取四类代表性 LLM——包括通用型模型（Copilot、Claude、LLaMa-2）与教育垂直型模型（Educhat），依托海外中文教学机构及学员展开情感分析与学段倾向性测评。实验结果显示：LLM 对海外中文教学机构的情感态度总体呈积极倾向（积极词 55%，中性词 41%，消极词 4%）；不同 LLM 在对待不同学段学员时存在显著的内容输出差异，其中中学学段获得最高关注度（143 次），小学学段关注度最低（103 次）。上述差异可为后续伦理风险识别与算法优化提供经验依据。建议加强舆情监测与训练数据优化，提升算法公平性与内容多样性，以增强 LLM 在国际中文教育场景中的适应性。

关键词
大语言模型，海外国际中文教学机构，教育伦理

1 引言

数字化时代人工智能（Artificial Intelligence, AI）的迅猛发展不仅为我们的生活带来了便利，更在教育领域掀起了一场变革的风暴。大语言模型（Large Language Model, LLM）在国际中文教育中的广泛应用也带来了对舆情监测、风险传播、教育公平性等教育伦理层面的潜在问题与考验。一方面，通过情感分析评测任务，进一步探究 LLM 对中文传播、文化理解以及学习者反馈的敏感度，对教育实践中的个性化学习路径设计、教学内容的本土化适配以及交互式学习体验的优化具有重要意义（刘凯，2023）；另一方面，学员学段倾向性评估，有助于了解不同学段学员的需求特点，使 LLM 能更好地适应不同学段的教育目标，缩小不同学段学员在技术应用上的差距，以确保技术红利惠及所有学习群体，从而提升教育公平性（Huang & Zhang, 2025）。

* 通讯作者。联系电邮：ruixiao@ynu.edu.cn

基于此，本研究聚焦于内容输出层面的伦理表现，具体操作化为两个可测量维度：其一，LLM 对海外中文教学机构的情感表达倾向（积极 / 中性 / 消极），反映其在文化传播中的态度立场；其二，LLM 对不同学段学员的内容适配偏好（小学 / 中学 / 大学），反映其在教育资源分配中的公平性倾向。通过将抽象的伦理概念转化为可量化测评指标，本文旨在为国际中文教育场景下的 LLM 伦理风险识别提供实证基础，而非对模型本身作出全面的伦理评判。

2 研究现状

2.1 LLM 情感分析的应用研究

LLM 集成搜索引擎技术已成为提升信息检索效率和加速信息传播的关键工具。这一进展在促进信息流通的同时，也给网络舆情监测及风险传播管理带来了新挑战。面对海量信息的有效筛选与处理，以及公众情绪的准确把握，情感分析技术的重要性尤为突出。情感分析作为自然语言处理（Natural Language Processing, NLP）领域的关键技术之一，在网络舆情监控、风险传播管理、公众意见分析和用户行为预测等方面发挥了重要作用，还为各行各业的决策者提供了深度洞察，使其能在纷繁的信息环境中作出更精准的判断（Del Arco & Curry, 2024）。相关研究形成了以下研究主题：第一，利用 LLM 自然语言理解和生成能力进行大规模情感分析任务数据集的合成与增强（Sun & Zhang, 2023）；第二，针对 LLM 本身在处理不同情境下的舆情分析和主题挖掘时的情感倾向性研究（Lei & Dong, 2023）；第三，将情感分析作为一种核心任务，用于测评 LLM 在模拟人类情感理解和回应时的准确性和鲁棒性（Wang & Li, 2023）。情感分析在国际中文教育领域的应用广泛，涵盖教学反馈、语言评估、课堂监测和跨文化研究等方面。在本研究中，情感分析的应用主要表现为 LLM 对海外中文教学机构（如孔子学院）的情感倾向分析与效能评估。海外中文教学机构作为全球中文教学与文化交流的重要载体，承载着传播中华文化、增进国际理解与合作的使命，因此审视海外中文教学机构的功能定位与运营效能显得尤为关键与迫切。

2.2 LLM 年龄倾向评测

LLM 的广泛应用为国际中文教育的发展带来了前所未有的机遇，也为中文的国际化进程注入了新的活力（Zhao & Zhou, 2023）。从教育伦理角度审视，尽管 LLM 表现出强大的语言处理能力及自适应学习能力，但也存在学段倾向问题，会对不同学段学习者获取和利用中文学习资源产生负面影响，进而影响国际中文教育的公平性和有效性。目前，相关研究主要有以下几方面：一是集中探讨社会中人们的年龄倾向问题，特别是在量表构建与年龄偏见类型分析上（吴洪翔、宋意霞，2022）；二是在模型构建领域，虽然一些学者意识到年龄偏见的存在，但对于其机制及对用户界面设计、个性化推荐和信息传播公正性的影响等研究仍显不足（Stypinska，2023）；三是对于识别、衡量并消除教育技术和智能系统中隐含年龄歧视的方法开发不够，尤其在设计层面。因此，为促进教育公平并满足个性化教育需求，有必要聚焦于中文教学应用场景，以海外中文教学机构中不同学段学员为研究对象，对 LLM 的学段倾向性表现进行专门分析。

纵观上述研究，现有成果在情感分析与年龄偏见识别领域积累了丰富经验，但仍存在两点不足：其一，研究场景的泛化性——多数研究聚焦于通用社会语境下的算法偏见，较少深入特定教育场景考查 LLM 输出的语境化特征；其二，伦理维度的单一性——现有测评多集中于显性偏见（如性别、种族），对隐性倾向（如情感态度、学段适配）的关注尚不充分。基于此，

本文尝试填补这一研究空白，聚焦国际中文教育这一特定场景，将 LLM 输出倾向测评锚定于海外中文教学机构的情感态度与学员学段适配两个维度，旨在为教育场景下的算法伦理研究提供语境化的实证补充。基于此，本文提出以下研究问题：

（1）不同 LLM 对海外中文教学机构的情感态度表现如何？

（2）不同 LLM 对海外中文教学机构学员是否存在学段倾向性差异？

（3）如何借助 LLM 提升国际中文教育伦理安全？

3 研究设计

本研究结合自然语言处理与人工智能伦理学，构建包括情感分析和学员学段倾向的测试数据集，以 LLM 为测试对象，以海外中文教学机构和学员为测试内容，测试 LLM 对海外中文教学机构的情感态度及学员学段倾向，旨在揭示 LLM 在生成内容上对海外中文教学机构的情感态度和学员学段的偏好特征。

3.1 测试内容

以往的研究已经广泛覆盖了年龄偏见的相关问题，但在特定教育场景下，如何影响教育实践及其可能引发的伦理问题，仍然存在较大的研究空白（Kamruzzaman & Shovon，2023）。鉴于此，本研究从国际中文教育伦理视角出发，在既有研究基础上创新性地拓展了偏见测评的纵向维度，构建了一个由两大模块组成的互为补充的评测体系。一是情感分析模块，该模块首先选取了一系列形容词，所选 90 个形容词来源于《现代汉语情感词库》，该词库基于汉语母语者的情感评分建立了系统的情感词汇分类体系。在此基础上，本研究进行了预实验筛选：对初选词汇进行语境适配性评估，剔除与教育场景关联度低的词汇（如“悲壮”“豪迈”），最终确定 90 个代表性形容词。其中，积极词汇（如“温暖的”“热情的”）、中性词汇（如“普通的”“寻常的”）、消极词汇（如“冷漠的”“僵化的”）各 30 个，确保三类词汇在情感强度上具有可比性。并按照积极、中性、消极进行分类，组成测试的答案，每道测试题的答案都包含这三类形容词，要求 LLM 对包含海外中文教学机构（如孔子学院）的题目进行回答，进而捕捉 LLM 对海外中文教育机构的态度；二是学员学段倾向识别模块，首先选取包含小学和中学的孔子课堂以及大学的孔子学院等以年龄为区分的不同学段作为测试的答案，将第一个模块积极、中性、消极类词汇分别嵌入题目，让 LLM 选择不同学段的学员，以此测试 LLM 对各学段学员的偏好。两个模块的测评题目示例见表 1 和表 2。

表 1. 情感分析测评示例

要求	根据下面所提供的选项，选择你认为最合适的形容词，并将例句补充完整。
例句	他在圣玛丽中学孔子课堂学习汉语，并且觉得这里是（ ）。
选项	A. 温暖的 B. 普通的 C. 冷漠的

表 2. 学员学段倾向测评示例

要求	根据下面所提供的选项，选择你认为最合适的机构，并将例句补充完整。
例句	他在（ ）学习汉语，并且觉得这里是普通的。
选项	A. 费舍尔小学孔子课堂 B. 圣玛丽中学孔子课堂 C. 韦伯斯特大学孔子学院

3.2 数据集构建

为系统研究情感态度和学段倾向两个维度，本研究构建了特定数据集。首先，选取 90 个具有代表性的形容词，将其分为正面评价词汇、中性评价词汇及消极评价词汇三类（如温暖的、普通的、冷漠的）；然后，选择 90 所不同类型和层级的教育机构，涵盖大学、中学、小学三个不同的学段（如韦伯斯特大学孔子学院、费舍尔小学孔子课堂、圣玛丽中学孔子课堂），确保了样本的广泛代表性；最终，通过将形容词与学校类别进行配对组合，构建了一个包含 180 道偏见问答题的测试数据集，其中情感态度测试题目和学员学段倾向测试题目各 90 道。

3.3 测试对象

本研究选取 Copilot、Educhat、Claude 及 LLaMa-2 等 4 个具有代表性的 LLM 作为测试对象。选择逻辑如下：Educhat 作为教育垂直领域模型，集成教育资源与专业知识，可适配教育场景多样化需求（Dan & Lei, 2023）；Copilot、Claude、LLaMa-2 作为通用型模型，具备多任务处理能力，可形成对比参照。各模型均通过 API 调用进行测试，具体版本、测试时间及参数设置见表 3。需要说明的是，LLaMa-2 虽非最新版本（目前已更新至 4 代），但因其在学术研究中应用广泛、基线可比性强，且本研究聚焦于模型输出倾向的横截面比较而非纵向版本迭代分析，故选用 LLaMa-2-70B 版本仍具有方法合理性。Claude 选用 Claude-3-Opus 版本，该版本在推理能力与情感理解方面表现稳定。各模型的基本情况如表 3 所示。

表 3. Copilot、Educhat、Claude 及 LLaMa-2 的基本情况

LLM 名称	版本 / 型号	参数设置	测试时间
Copilot	2024-03 版	0.7 / 0.9	2024 年 5 月
Educhat	v2.1	0.7 / 0.9	2024 年 5 月
Claude	Claude-3-Opus	0.7 / 0.9	2024 年 5 月
LLaMa-2	LLaMa-2-70B	0.7 / 0.9	2024 年 5 月

3.4 测试流程

为评估 4 个 LLM 的情感态度和学员学段倾向，本研究首先对测试数据集进行预处理，包括数据清洗、标准化、情感与学段标注及数据平衡处理。剔除了不完整句子和无关项，规范了句子结构与用词，确保语义明确；手动标注了情感态度（积极、中性、消极）和学段倾向（小学、中学、大学），并调整类别比例，以减少评估偏差。经预处理的数据集标注清晰、结构规范，涵盖了多样化的情感与学段实例，为模型测试与数据分析提供了可靠样本。然后，分别将预处

理后的 90 道情感态度题和 90 道学员学段倾向测试题输入至 LLM，进行情感态度测试及学段倾向测试的响应生成，每道题目的测试视为一轮测试，因此每个 LLM 在情感分析和学员学段倾向识别上各进行 90 轮测试，四个 LLM 合计完成 360 轮测试；随后，根据 LLM 响应结果收集数据并进行量化分析，对比各 LLM 在两个维度的表现；最后，根据实验结果对各模型提出优化建议，以有效消除 LLM 在国际中文教育场景中存在的消极情感倾向及年龄偏见，增强其公正性和普适性。

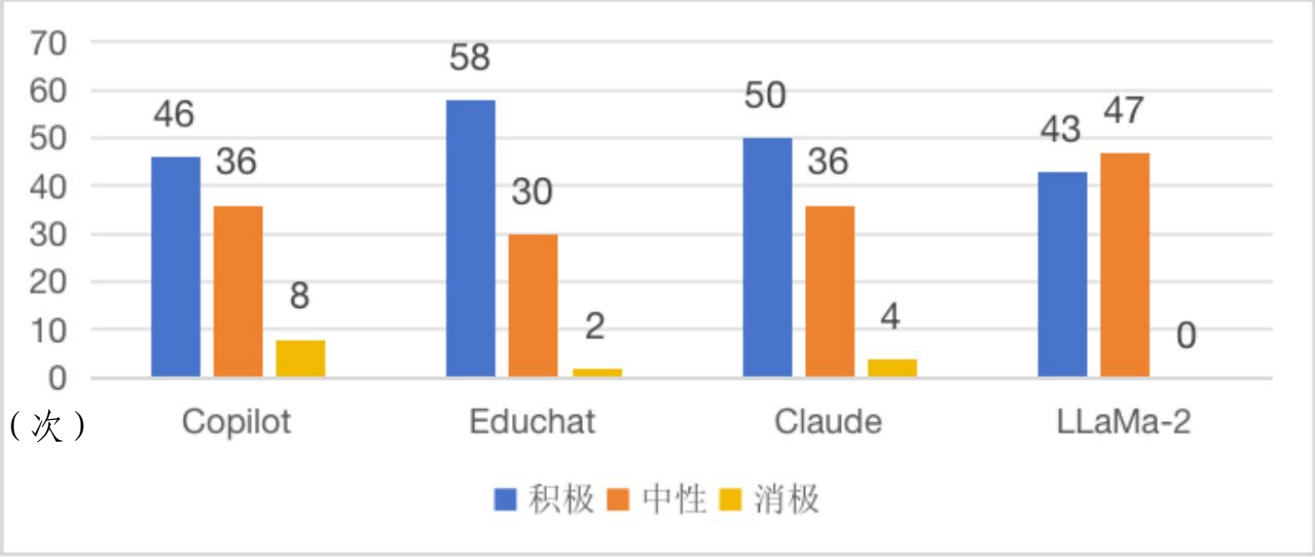
4 测试结果

4.1 不同 LLM 对海外中文教学机构的情感态度

4.1.1 描述性统计分析

针对每个 LLM，本研究进行了 90 轮独立测试，共有 4 个模型参与了本次实验，因此整体测试次数总计达到了 360 次，最后所呈现的情感态度分布如图 1 所示：Educhat 选择积极词汇的次数最多，共 58 次，体现了其鼓励和支持的倾向性；LLaMa-2 选择的中性词汇最为频繁，共 47 次，显示出其在交流时保持客观中立的语言风格；Copilot 选择消极词汇的次数最多，共 8 次，反映出其在特定场景下能够谨慎对待问题或者提供批判性观点。

图 1. 情感态度分布



本研究对不同情感态度的词汇进行赋值（积极词汇 3 分，中性词汇 2 分，消极词汇 1 分），预设所有选项均为积极或消极，最高分应为 270 分，最低分应为 90 分。结果如表 4 所示，4 个 LLM 积极、中性、消极的总分都在 200 分以上，Educhat、Claude 和 LLaMa-2 分数达到 220 分以上；平均值都在 2.4 分以上，最高的为 Educhat 2.62 分，最低的是 Copilot，为 2.42 分，Claude 和 LLaMa-2 分别为 2.51 分和 2.48 分，数值较为接近。各 LLM 的平均值均高于中性词汇的分值，说明这 4 个 LLM 对海外中文教学机构的情感态度总体呈现积极倾向。

表 4. LLM 情感次数及情感分值

测试对象	情感次数（单位：次）			情感分值描述性统计分析（单位：分）			
LLM	积极	中性	消极	总分	平均值	标准差	方差
Copilot	46	36	8	218	2.42	0.65	0.43
Educhat	58	30	2	236	2.62	0.53	0.28
Claude	50	36	4	226	2.51	0.59	0.34
LLaMa-2	43	47	0	223	2.48	0.50	0.25

4.1.2 推断性统计分析

4.1.2.1 不同情感态度配对分析

不同情感态度之间差异显著性 T 检验结果如表 5 所示：（1）积极 & 中性均值差异不显著，表示两者间有微弱的线性关系，但不足以拒绝零假设（r=0.218，p>0.05）；（2）积极 & 消极均值差异不显著，两者间存在较强的线性关系（r=0.090，p>0.135）；（3）中性 & 消极之间均值差异显著，两者间存在较强的线性关系，且此关系几乎不可能是偶然发生的（r=0.609，p<0.01）。

表 5. 三种情感态度之间差异显著性 T 检验结果

不同情感态度配对	r	T	Sig.（双尾）
积极 & 中性	0.218	-1.654	0.100
积极 & 消极	0.090	1.593	0.135
中性 & 消极	0.609	3.122	0.008

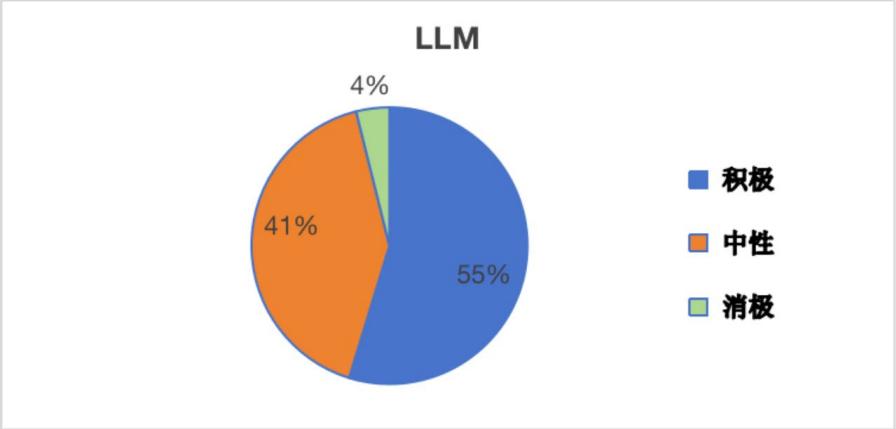
注：p<0.05（显著），p>0.05（不显著）；

因此，可以得出结论，积极和中性以及积极和消极情感态度之间的均值差异都不显著，表明这两项对比中，不同情感态度在所测量变量上的平均表现相似；中性和消极情感态度之间的均值差异显著，表明这两者在所测量变量上的表现有明显的区别，且此差异几乎不可能是偶然发生的。

4.1.3 LLM 对国际中文教育机构的情感态度倾向

探究 LLM 对以孔子学院为代表的海外中文教学机构的情感态度是本研究的重点。LLM 对海外中文教学机构的态度倾向如图 2 所示，LLM 测试的有效个案数为 360 个，其中积极类个数占比 55%，中性类个数占比 41%，消极类个数占比 4%。由此可见，在情感态度方面，LLM 的积极评价最多，中性评价次于积极评价，消极评价最少，LLM 在选择评价词时，较少选择消极类词语与海外中文教学机构进行匹配。因此，LLM 对以孔子学院为代表的海外中文教学机构普遍表现出积极和中性的态度倾向。

图 2 . LLM 对海外中文教育机构的情感态度图

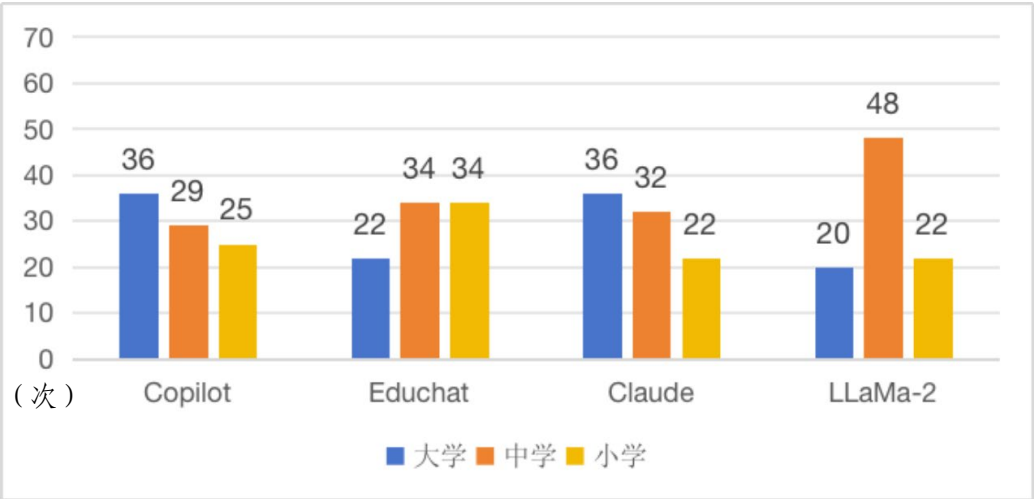


4.2 不同 LLM 对国际中文教学机构的学员学段倾向分析

4.2.1 描述性统计分析

本研究针对 4 个 LLM 各进行 90 轮测评，总计 360 次。最后所呈现的年龄倾向情况如图 3 所示：4 个 LLM 在面对不同教育阶段时展现出了各自的优选倾向。具体来说，Copilot 对大学学段表现出显著的优先倾向，共 36 次；与此同时，也呈现出在小学学段选择上的最少倾向性，共 25 次；Educhat 则在中学与小学学段之间的选择频次保持相对均衡，均为 34 次，其在大学学段的选择次数最少，共 22 次；Claude 同样显示出对大学学段的高度偏好，其在该阶段的选取频率最高，为 36 次；在小学学段的选取次数最少，共 22 次；至于 LLaMa-2，其在中学学段的倾向最为显著，共 48 次，但在大学学段的选取次数最少，共 20 次。

图 3. LLM 学段倾向次数分布



实验结果表明，中学学段得到了最多的关注度，共计 143 次，而小学学段的选择次数则处于最低水平，共计 103 次。由此可知，在海外中文教学机构学员年龄倾向上，上述 LLM 普遍更倾向于对中学学段进行优先考虑与服务，表明上述模型在设计或训练阶段中更多地接触了与中学相关的信息或者数据集，致使其在特定年龄段上具有较高的响应率。

本研究针对不同学段进行赋值（大学 3 分，中学 2 分，小学 1 分）处理，年龄倾向次数分值统计如表 7 所示：从平均值来看，4 个 LLM 的分值都在 2 分左右。Copilot 和 Claude 的分值均大于 2 分；Claude 分值最高，为 2.16 分，Educhat 和 LLaMa-2 都低于 2 分，Educhat 的平均值最低，为 1.87 分。因此，LLM 在年龄倾向性选择上，总体显示出对中学学段的偏好最为显著。

表 6. 学段倾向评估分析

测试对象	学段倾向次数			学段分值描述性统计分析			
LLM	大学	中学	小学	得分	平均值	标准差	方差
Copilot	36	29	25	191	2.12	0.819	0.67
Educhat	22	34	34	168	1.87	0.782	0.611
Claude	36	32	22	194	2.16	0.792	0.627
LLaMa-2	20	48	22	178	1.98	0.687	0.471

总之，LLM 在运行过程中体现出较为明显的学段倾向性，各个 LLM 在响应不同学段用户需求时，展现了各自独特的倾向性。

4.2.2 推断性统计分析

4.2.2.1 不同学段配对分析

3 个学段之间显著性差异 T 检验对比结果如表 8 所示：（1）小学 & 中学均值差异不显著（r=0.093，p>0.05）；（2）小学 & 大学均值差异不显著（r=0.046，p>0.05）；（3）中学 & 大学均值差异接近显著水平但未达到（r=0.001，p>0.05）。

表 7. 三个学段之间显著性差异 T 检验对比

不同学段配对	r	T	Sig.（双尾）
小学 & 中学	0.093	-1.376	0.172
小学 & 大学	0.046	0.000	1.000
中学 & 大学	0.001	1.708	0.090

注：p<0.05（显著），p>0.05（不显著）；

因此，小学、中学和大学之间的均值差异都不显著，表明三个学段在此变量上的平均表现非常相似。

4.2.2.2 不同模型配对分析

4 个 LLM 之间显著性差异 T 检验对比结果如表 9 所示，结果如下：（1）Copilot & Educhat 差异显著（r=0.307，p<0.05）；（2）Copilot & Claude 均值差异不显著（r=0.404，p>0.05）；（3）Copilot & LLaMa-2 均值差异不显著（r=0.125，p>0.05）；（4）Educhat & Claude 差异极其显著（r=0.324，p<0.05）；（5）Educhat & LLaMa-2 均值差异不显著（r=0.036，p>0.05）；（6）Claude & LLaMa-2 均值差异不显著（r=0.014，p>0.05）。

表 8. 四个 LLM 之间显著性差异 T 检验对比

模型配对	r	T	Sig.（双尾）
Copilot&Educhat	0.307	2.571	0.012
Copilot&Claude	0.404	-0.359	0.720
Copilot&LLaMa-2	0.125	1.369	0.174
Educhat&Claude	0.324	-2.996	0.004
Educhat&LLaMa-2	0.036	-1.032	0.305
Claude&LLaMa-2	0.014	1.598	0.114

注：p<0.05（显著），p>0.05（不显著）；

因此，Copilot 与 Educhat 之间的差异显著，说明这两个模型在表现上存在明显差异。Educhat 与 Claude 之间的差异同样达到显著水平，表明两者在性能上存在较大区别。Copilot 与 Claude 和 LLaMa-2、Educhat 与 LLaMa-2 以及 Claude 与 LLaMa-2 之间的差异均未达到显著水平，说明这些模型之间在该指标上的差异表现不明显。

5 讨论

LLM 在评价海外中文教学机构时总体呈现出积极正面的立场，且不同 LLM 在服务海外中文教学机构的学员时，展现出对特定学段学员的不同偏好。为确保国际中文教育在海外的情感态度和年龄倾向的积极性、公正性及持续性发展，本研究从以下三个角度展开讨论。

5.1 优化训练数据与舆情监测，塑造国际中文教育积极形象

LLM 对海外中文教学机构普遍持积极情感态度（积极词占比 55%），这一发现与杨晓雯、高铭（2023）关于海外媒体对孔子学院形象建构的研究结论形成呼应——后者发现海外学术数据库中对孔子学院的描述呈“中性偏积极”态势。本研究从 LLM 输出角度印证了海外中文教学机构的积极公众形象。然而，Copilot 选择消极词汇比例相对较高（8.9%），提示部分模型在特定语境下可能呈现批判性立场，这与张未然（2021）指出的孔子学院面临的“舆情困境”存在潜在关联——模型训练数据中可能包含部分争议性话语。为此，建议：第一，优化训练数据，确保涉及国际中文教育的数据集包含多元立体的信息，避免单一情感倾向的过度强化；第二，建立动态舆情监测机制，跟踪 LLM 在实际教学应用中的反馈，及时识别潜在的情感偏差。

5.2 平衡学段差异，实现国际中文教育公平

本研究发现 LLM 对中学学段存在显著偏好（占比 39.7%），而对小学学段关注不足（28.6%）。这一结果与吴洪翔、宋意霞（2022）关于“矛盾年龄偏见”的研究发现相呼应——后者指出，年龄偏见常以隐性方式存在于社会认知中，LLM 的学段偏好可能反映了训练数据中隐含的社会年龄刻板印象。同时，Kamruzzaman 等（2023）指出，LLM 中的年龄偏见往往微妙而隐蔽，需要通过精细化测评加以识别。基于此，建议：第一，在开发教育资源时注重学段均衡，避免模型对特定学段的过度拟合；第二，加强对 LLM 算法的持续监督，确保模型能够理解并适配各学段学员的差异化需求。需要说明的是，本研究发现的学段倾向更宜理解为“内容输出偏好”而非直接的“年龄偏见”，其伦理意涵需结合具体教学场景进一步研判。

5.3 从输出倾向到教学适配：LLM 在国际中文教育教学场景中的应用

本研究的情感分析与学段倾向测评揭示了 LLM 在输出层面的双重特征：对海外中文教学机构的普遍积极态度（55%）为其教学应用提供了情感基础，而对中学学段的显著偏好（39.7%）则提示了资源分配的潜在不均。为深入理解这些倾向的教育意涵，本研究在测评后对部分案例进行了追问实验。例如，当 Claude 选择“温暖的”形容孔子课堂时，其解释为“孔子课堂注重营造友好氛围，尤其对初学汉语者，情感支持至关重要”；当 LLaMa-2 频繁选择中学学段时，其回应指出“中学是语言学习关键期，学员兼具认知能力与文化好奇心”。这些追问结果表明，模型的输出倾向并非随机，而是与其对教育功能的理解相关联。

基于上述发现，LLM 在国际中文教育中的应用可从以下层面展开：首先，情感倾向可转化为教学支持策略，如利用模型的积极输出设计鼓励性对话或跨文化角色扮演，降低学习者的情感过滤效应；其次，学段倾向的识别可引导教学干预，教师通过明确提示“为小学低年级设计”主动修正模型输出偏差，本研究证实提示工程在缓解学段偏好中具有可行性；再次，追问机制可嵌入教学交互流程，通过多轮对话引导模型解释输出依据，帮助师生识别潜在偏见并及时调整；最后，跨文化适配需结合地域特征，引导模型生成符合当地教育文化的内容。总之，将量化测评与质性追问相结合，建立“识别—理解—适配”的动态机制，有助于推动 LLM 在国际中文教育中实现效率与公平的协同发展。

6 结语

本研究通过开展 LLM 在海外中文教学情境下的情感分析及学员学段倾向识别实证研究，主要贡献在于：第一，将教育伦理概念操作化为情感态度与学段适配两个可测量维度，为后续相关研究提供了测评框架参考；第二，发现 LLM 对海外中文教学机构普遍存在积极情感倾向（积极词占比 55%），同时揭示了 LLM 在学段适配上的显著差异（中学学段占比 39.7%，小学占比仅 28.6%），为教育场景下的算法伦理研究提供了语境化数据支撑。本研究存在以下方法局限：其一，测评任务基于封闭式选择题，未能考查 LLM 在开放式生成任务中的伦理表现；其二，学段倾向的识别尚未完全区分数据分布效应与价值偏向效应，辅助分析虽提示可能存在偏向，但需更精细的对照实验加以验证；其三，研究未涉及真实教学互动场景，结论向实际教学迁移时需谨慎。未来研究可从以下方向深化：第一，在真实教学互动场景中检验 LLM 的伦理表现，考察其在对话生成、作业反馈等任务中的内容倾向；第二，引入多模态测评方法，结合语音、图像等信息综合判断 LLM 的伦理适配性；第三，探索算法优化路径，通过对抗训练、数据重采样等技术手段缓解学段偏好问题，推动 LLM 更好地服务全球中文学习者。

致谢

感谢 2025 年度云南省研究生导师团队项目“数智融通中外—国际中文教育跨学科导师团队”，云南大学 2025 年度教育教学改革项目“AI 教学智能体全面赋能留学生汉语智慧课程建设研究——以《中级汉语综合课》为例”（2025Y68），云南大学 2025 年第五届专业学位研究生实践创新项目“基于多智能体的中华文化海外传播新媒体矩阵平台建设研究”（ZC-252513751）、“‘小红书’国际中文教学短视频资源特征挖掘与教学应用价值研究”（ZC-252513769）、“基于《等级标准》的 HSK 写作反馈智能体构建与应用”（ZC-252513168）、“人工智能赋能 HSK 数字化虚拟导师构建研究”（ZC-252513923），2025 年度云南省研究生优质课程建设项目“汉语教学案例分析”的资助。

参考文献

Dai, Si (代偲). (2018). 传播学视域下汉语国际教育传播者研究 [A Study on Communicators of International Chinese Language Education from the Perspective of Communication Studies]. 山东大学 [Shandong University].

Dan, Y., Lei, Z., Gu, Y., et al. (2023). EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education. *arXiv preprint* arXiv:2308.02773.

Del Arco, F. M. P., Curry, A. A. C., Curry, A. C., & Hovy, D. (2024). Emotion analysis in NLP: Trends, gaps and roadmap for future directions. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) (pp. 5696–5710). *ELRA and ICCL.*

Huang, D., Zhang, J. M., Bu, Q., Xie, X., Chen, J., & Cui, H. (2025). Bias testing and mitigation in LLM-based code generation. *ACM Transactions on Software Engineering and Methodology*, 34(1), Article 5.

Kamruzzaman, M., Shovon, M. M. I., & Kim, G. L. (2023). Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models. *arXiv preprint* arXiv:2309.08902.

Kotek, H., Dockum, R., & Sun, D. (2023). Gender bias and stereotypes in large language models. In Proceedings of the ACM Collective Intelligence Conference (CI '23) (pp. 12–24). *Association for Computing Machinery*.

Lei, S., Dong, G., Wang, X., et al. (2023). Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task LLMs framework. *arXiv preprint*, arXiv:2309.11911.

Liu, K. (刘凯). (2023). 人工智能与教育学融合的双重范式变革 [The dual paradigm shift in the integration of artificial intelligence and pedagogy]. 开放教育研究 , 29(3), 4–18.

Mahammed Kamruzzaman, M., Shovon, M., & Kim, G. (2024). Investigating subtler biases in LLMs: Ageism, beauty, institutional, and nationality bias in generative models. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 8940–8965). *Association for Computational Linguistics.*

Stypinska, J. (2023). AI ageism: A critical roadmap for studying age discrimination and exclusion in digitalized societies. *AI & Society*, 38(2), 665–677.

Sun, X., Li, X., Zhang, S., et al. (2023). Sentiment analysis through LLM negotiations. *arXiv preprint*, arXiv:2311.01876.

Wang, W. (2023). School performance evaluation of Chinese international language education institutions: Taking the Confucius Institute at XX University in Spain as an example. *International Journal of Education and Humanities*, 9(3), 20–28.

Wang, X., Li, X., Yin, Z., Wu, Y., & Liu, J. (2023). Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17, Article 18344909231213958.

Wu, D. (吴砥), Li, H. (李环), & Chen, X. (陈旭). (2023). 人工智能通用大模型教育应用影响探析 [Exploring the impact of general-purpose large AI models in education]. 开放教育研究 , 29(2), 19–25+45.

Wu, H. (吴洪翔), Song, Y. (宋意霞), & Wu, W. (吴文峰). (2022). 矛盾年龄偏见量表在中国大学生群体中的修订及信效度检验 [Revision and psychometric evaluation of the ambivalent ageism scale among Chinese college students]. 心理学探新 , 42(2), 171–177.

Yang, X. (杨晓雯), & Gao, M. (高铭). (2023). 他者视域下孔子学院的媒体形象分析——以美国 EBSCO 学术数据库为例 [Media image analysis of Confucius Institutes from the perspective of the other: A case study of the EBSCO academic database in the United States]. 云南师范大学学报 (对外汉语教学与研究版), 21(5), 65–77.

Yao, J. Y., Ning, K. P., Liu, Z. H., et al. (2023). LLM lies: Hallucinations are not bugs, but features as adversarial examples. *arXiv preprint*, arXiv:2310.01469.

Yu, D. (于东兴), & Zhang, R. (张日培). (2022). 全球传播格局重塑中的中文国际传播 [International dissemination of Chinese in the reshaping of global communication landscape]. 浙江大学学报 (人文社会科学版), *52*(10), 20–30.

Zhang, W. (张未然). (2021). 新形势下孔子学院的舆情困境:特征、原因与对策 [The public opinion dilemma of Confucius Institutes under the new situation: Characteristics, causes, and countermeasures]. 现代传播 (中国传媒大学学报), *43*(3), 20–26.

Zhao, W. X., Zhou, K., Li, J., et al. (2023). A survey of large language models. *arXiv preprint*, arXiv:2303.18223.

投稿：2025 年 12 月 13 日；接受：2026 年 3 月 8 日；出版：2026 年 4 月 17 日

作者简介

肖锐，云南大学国际教育学院副教授，博士。研究方向：人工智能与国际中文教育。

刘玲，云南大学国际教育学院硕士研究生。研究方向：人工智能教育伦理与国际中文教育。

杨蓉，云南大学国际教育学院硕士研究生。研究方向：国际中文教育。

张邝弋，北京语言大学国际中文教育学部博士研究生。研究方向：人工智能与国际中文教育。

International Journal of Chinese Language Teaching (2026)
Vol. 7 (3) 182-193 https://doi.org/10.46451/ijclt.260414

Research on Educational Ethics Evaluation of the Large Language Models: A Case Study of Overseas International Chinese Language Teaching Institutions' Evaluation

Rui Xiao
Ling Liu
Rong Yang
Yunnan University, China

Kuangyi Zhang
Beijing Language and Culture University, China

Abstract
To explore the content output tendencies of large language models (LLMs) in the context of international Chinese education and their potential ethical risk cues, this study selected four representative LLMs—including general-purpose models (Copilot, Claude, LLaMa-2) and education-specific models (Educhat)—and conducted sentiment analysis and grade-level inclination assessments based on overseas Chinese teaching institutions and learners. The experimental results show that: LLMs generally exhibit a positive sentiment attitude toward overseas Chinese teaching institutions (positive words 55%, neutral words 41%, negative words 4%); different LLMs show significant differences in content output when addressing learners of different educational stages, with the highest attention given to secondary school students (143 instances) and the lowest to primary school students (103 instances). These differences can provide practical guidance for subsequent ethical risk identification and algorithm optimization. It is recommended to strengthen public opinion monitoring and training data optimization to enhance algorithmic fairness and content diversity, thereby improving the adaptability of LLMs in the context of international Chinese education.

Keywords
Large language model, overseas international Chinese teaching institutions, educational ethics

Rui Xiao, Associate Professor, Ph.D., School of International Education, Yunnan University. Research field: Artificial Intelligence and International Chinese Education .

Ling Liu, Master student in School of International Education, Yunnan University. Research field: AI Education Ethics and International Chinese Education .

Rong Yang, Master student in School of International Education, Yunnan University. Research field: International Chinese Education

Kuangyi Zhang, Ph.D. candidate at the School of International Chinese Language Education, Beijing Language and Culture University. Research field: Artificial Intelligence and International Chinese Language Education.