

应用大型语言模型 提升知识工作者绩效

以学术论文评审为例

刘国兴 著

应用大语言模型提升知识工作者绩效 以学术论文评审为例

刘国兴 著



出版社 Nexus Press B.V. (奈科赛斯出版社)
地址 Amazonelaan 89, 1448 VC Purmerend, Amsterdam, The Netherlands
官方网站 <https://www.nexus-press.com/>
I S B N 978-90-836446-0-8
字 数 112301
出版年份 2026 年 5 月

独家许可声明：作者保留其个人贡献的版权，但授予荷兰Nexus Press B.V. (奈科赛斯出版社) 拥有出版、发行和制作作品的独家许可。

本作品在知识共享署名 4.0 开放获取国际许可协议 (CC BY 4.0) ，在注明原作者的前提下，允许在任何媒介中使用、共享、改编、分发和复制，传播和复制时须注明原作者和Nexus Press B.V.

查看此许可证的副本，请访问：

<https://creativecommons.org/licenses/by/4.0/>

有任何疑问请发邮件至：contact@nexus-press.com

书籍介绍

本书聚焦于大语言模型 (Large Language Models, LLMs) 驱动的人工智能 AI 工具在学术论文同行评审流程中的应用效果, 系统考察 AI 辅助对评审效率、准确性、质量、认知负荷以及用户体验的影响, 为构建更高效、更科学、更可持续的学术评审生态提供理论依据和实践路径。研究采用了两项实证实验, 分别探讨 AI 工具的独立效应及其与人类评审员协作的机制, 并基于大样本的定量与定性数据, 全面揭示人机协同模式下的优势、局限与未来发展方向。

This book focuses on the application and effectiveness of LLM-driven AI tools in the peer review process for academic manuscripts. It systematically examines the impact of AI-assisted review on efficiency, accuracy, quality, cognitive load, and user experience, with the aim of providing both theoretical foundations and practical pathways for constructing a more efficient, scientific, and sustainable academic review ecosystem. Two empirical experiments were conducted to investigate both the independent effects of AI tools and the mechanisms of their collaboration with human reviewers. Based on large-sample quantitative and qualitative data, this study comprehensively reveals the advantages, limitations, and future directions of human-AI collaborative models.

前言

随着全球学术研究产出持续增长，学术期刊和会议论文的投稿量呈指数级攀升，传统的人工同行评审系统已难以应对日益繁重和复杂的评审需求。人力资源瓶颈、评审周期延长、评审标准不一以及主观性偏差等问题日益突出，严重影响着研究成果的及时传播与学术共同体的健康发展。在这一背景下，人工智能(AI)，特别是大语言模型(Large Language Models, LLMs)以其强大的自然语言处理和推理能力，为重塑学术评审流程带来了新的机遇和挑战。

本书通过多维度、多环节的实验设计与数据分析，全面揭示了大语言模型(Large Language Models, LLMs)驱动的人工智能(AI)工具在提升学术评审效率、质量和用户体验方面的巨大潜力，理清了AI与人类专家优势互补、动态协作的最佳实践路径，并为后续学术评审智能化、标准化发展提供了理论支撑和实践范式。随着LLM技术的持续进步和学科领域应用的不断深化，AI与人类专家高效协作的知识工作模式有望在更广泛的科研场景中推广应用，推动学术出版与创新生态的高质量发展。

全书共分6章：第1章为绪论，简要总结学术出版量激增下“高效、精准同行评审”需求凸显的现实，梳理行业瓶颈，抛出“AI辅助能否提升评审效率与成效”的核心问题，奠定人机协作研究的基础。第2章阐述LLMs用于知识工作的理论基础，系统回顾AI演进、信任机制、认知委托与双过程理论等关键文献，指出复杂评审任务中尚存的理论与实证空白。第3章介绍实验一，详述随机对照实验设计与数据获取流程，评估LLM辅助对稿件内容相关性、结构、创新性、重要性评分及专家差距缩小的影响，为用AI激发系统2思维提供实证。第4章介绍实验二，聚焦子任务层面的任务再分配，比较四种协作配置对表现、效率与认知负荷的作用，验证“定向弱项反思”策略在质量与效率间的最优平衡。第5章给出研究回应与假设证实，汇总实验一、二结果，阐明LLM辅助显著缩短评审时间、

提升评分一致性，并证实三条研究假设成立。第 6 章为全书总结与展望，凝练主要发现对学术出版及知识工作场景的启示，剖析 AI 信任与认知委托的理论贡献，指出研究局限并展望未来人机协作深化方向。

在此要感谢香港城市大学信息系统学系马建教授和复旦大学管理学院黄丽华教授的指导；感谢书中参考文献的作者们。

本书疏漏之处在所难免，恳请同行专家与读者斧正。联系方式为 secretary@aisacademy.org.cn，不胜感谢。

作者

2026 年 1 月 23 日

目 录

第一章 绪论 01

1.1 研究背景 01

1.2 研究问题 05

1.2.1 子问题 1 06

1.2.2 子问题 2 06

1.2.3 子问题 3 07

1.2.4 子问题 4 07

1.3 研究的意义 08

1.3.1 理论意义 08

1.3.2 实践意义 09

1.3.3 管理意义 11

1.4 研究方法 11

1.4.1 文献分析 12

1.4.2 随机对照实验（RCT） 12

1.4.3 实证分析 13

1.4.4 假设提出 15

1.4.5 伦理、隐私与数据安全保障 16

1.5 论文结构 17

第二章 文献综述 19

2.1 人工智能与人机协作的现有研究 19

2.2 大语言模型的发展及其对人机协作的影响 21

2.3 人工智能与人类合作模式与成效的研究 24

2.4 双过程理论与人机协作..... 27

 2.4.1 系统1：对AI建议的直觉和初步反应 27

 2.4.2 系统2：对AI输出的深思熟虑和分析性评估 27

 2.4.3 双重加工思维中的认知偏见与人机互动 28

 2.4.4 通过双过程理论优化人机协作..... 29

2.5 文献述评与研究启示..... 30

第三章 实验一：LLM整合对稿件评审复杂任务表现的影响

——以论文审阅任务为例..... 32

3.1 研究背景与目的 32

 3.1.1 研究动机..... 32

 3.1.2 研究目的..... 33

3.2 数据与方法 34

 3.2.1 实验设计..... 34

 3.2.2 样本与参与者 35

 3.2.3 评审材料准备 35

 3.2.4 评分维度的定义与控制..... 37

 3.2.5 数据收集与工具 40

 3.2.6 数据分析方法 41

3.3 结果分析 42

 3.3.1 描述性统计..... 42

 3.3.2 回归分析..... 45

 3.3.3 AI在不同评审维度的表现 53

 3.3.4 评审效率的提升 58

 3.3.5 认知负荷的影响 59

3.4 实验结果的理论解释 62

3.4.1 双过程理论的应用	63
3.4.2 认知负荷理论的支持	63
3.4.3 人机协作理论的验证	64
3.4.4 信任与依赖理论的体现	64
第四章 实验二：优化任务分配以增强人机协作.....	65
4.1 实验目的	65
4.1.1 验证定向反思策略的有效性	65
4.1.2 考察全面反思的收益与代价	65
4.1.3 评估错误反思的风险	65
4.1.4 建立无反思基线	66
4.1.5 多维度综合评估	66
4.2 实验设计与过程	67
4.2.1 实验条件	67
4.2.2 实验材料	68
4.2.3 参与者	70
4.2.4 实验过程	73
4.3 数据收集与分析方法	75
4.3.1 数据收集	75
4.3.2 数据分析方法	77
4.3.3 数据可视化	79
4.4 结果分析	81
4.4.1 内容丰富性维度的结果分析	81
4.4.2 论证逻辑维度的结果分析	83
4.4.3 表达与写作维度的结果分析	86

4.4.4 创新性与重要性维度的结果分析	88
4.4.5 整体评价维度的结果分析	91
4.4.6 机制讨论	94
第五章 对研究问题的回应以及假设证实	100
5.1 对研究问题的回应	100
5.1.1 实验一对研究问题的回应	100
5.1.2 实验二对研究问题的回应	101
5.1.3 小结	102
5.2 假设证实	102
5.2.1 假设 1: LLM辅助能够降低完成同行评审任务的认知负荷	102
5.2.2 假设 2: LLM辅助能够提升评审准确性	103
5.2.3 假设 3: 任务满意度与对AI帮助的感知呈正相关	103
5.2.4 小结	104
第六章 研究结论与展望	105
6.1 研究发现	105
6.1.1 实验一的主要发现	105
6.1.2 实验二的主要发现	107
6.1.3 AI 与人类协作的潜力	109
6.2 理论贡献	111
6.2.1 对人机协作理论的扩展	111
6.2.2 双过程理论的应用	113
6.2.3 双过程理论的修正与展望	116
6.3 实践意义	117
6.3.1 实际应用建议	118

6.3.2 对学术出版行业的影响	119
6.3.3 管理与政策建议	120
6.3.4 用户体验与满意度提升	121
6.4 研究不足与未来展望	123
6.4.1 研究局限	123
6.4.2 未来研究方向	126
参考文献	135
附 录	143
附录A 实验一：有无AI辅助对决策影响的实验	143
附录A.1 实验一研究问卷	147
附录A.2 实验一原始数据	169
附录B 实验二：干预策略对人机协作决策影响的研究	185
附录B.1 实验二研究工具（问卷）	185
附录B.2 实验二原始数据及补充数据	194
附录C 提示语框架与参数设定	200
附录C.1 模型与生成参数	200
附录C.2 结构化提示语模板（中文）	200
附录C.3 Structured Prompt (English Skeleton)	201
附录C.4 提示语实例（节选，中文）	202

第一章

绪论

1.1 研究背景

人工智能（AI）与机器学习技术的迅速发展正在对包括学术界在内的多个行业产生深远影响^[1]。以 OpenAI 公司开发的 GPT-4 为代表的大语言模型（LLM），凭借其先进性与强大的文本处理能力，展现出显著提升学术出版等知识密集型领域工作效率的巨大潜力^[2]。近年来，随着学术出版领域呈指数级增长，期刊和会议论文数量持续增加，同行评审需求的激增。然而，稿件提交数量的持续增长也带来了新的挑战。传统同行评审系统过度依赖学科专家开展劳动密集型评审，这种模式在当前环境下已难以持续^[3]。

为应对上述挑战，中国领先的学术研究服务平台 AiScholar 采用了基于人工智能的技术手段，以提高学术出版与评审的效率^[4]。AiScholar 是一个综合性的学术服务平台，面向研究人员提供包括同行评审、期刊匹配和发表咨询等一系列服务。该平台利用前沿的 AI 与 LLM 技术，通过全球学术资源的整合、“研究即服务”的运营模式及 AI 驱动的评估机制，帮助学术工作者实现高效、精准的投稿体验。该公司研发了多款旗舰 AI 解决方案，例如“期刊大数据匹配系统”和“AI 同行评审系统”，达到简化学术论文发表流程的目的，并帮助学者快速匹配适合的期刊并获得初步的同行评审意见。

尽管这些工具在提升效率方面表现出显著优势，它们仍然存在一些明显的不足之处。首先，对于复杂且主观性较强的评审任务，如对研究工作的逻辑严密

性、创新性以及学术重要性的深入评估，AI 工具尚无法完全取代具备专业素养的人工评审员的判断与分析能力。其次，目前的 AI 系统在跨学科或特定专业领域的适应性有限，导致在某些学科领域的评审准确性和效率未见显著提升。此外，AI 工具的透明性与可解释性也存在不足，评审员在参考 AI 系统给出的评分或建议时，他们对 AI 系统的信任程度不足，这可能会影响到整个评审过程的公平性与可靠性^[5]。

由此可见，仅凭当前的 AI 技术难以全面优化学术评审流程^[6]，迫切需要探索更加有效的人机协作模式，这也正是本研究开展的重要原因之一。本研究通过优化任务分配策略并引入反思机制，能够有效弥补 AI 工具在关键评审环节中的不足，实现人机协作的最优平衡^[7]，最终提高评审工作的效率与质量。因此，面对持续增长的投稿需求，为进一步提升评审效率，AiScholar 有必要不断改进与完善其 AI 驱动系统。

在本研究中，所谓“反思机制”指的是在 AI 辅助的学术评审流程中，通过有意识的引导，使评审员能够跳脱原有的判断惯性，针对 AI 输出和自身初步评分结果展开系统性的再思考与再评价。具体而言，反思机制通常包括以下步骤：评审员首先根据自身理解完成论文的初步评分，随后系统展示 AI 生成的建议和分数，并通过设置反思引导性问题，促使评审员比较和分析 AI 建议与自身初评之间的异同，从而进一步检视自己的判断依据，完善或修正最终评分决策。该机制的核心目的是激发评审员的系统 2（深度分析型）思维过程，帮助其有效识别和纠正由系统 1（直觉型）带来的潜在偏差，从而提升评分的准确性与一致性。具体应用中，反思机制可表现为定向反思（仅针对 AI 表现薄弱的维度触发）、全面反思（在所有评分维度触发）或优势反思（聚焦于 AI 强项维度）。本质上，反思机制是一种通过结构化引导促进评审员主动检视与优化判断过程的方法，是实现高质量人机协同学术评审的重要策略^[8, 9]。

与此同时，随着学术出版行业竞争的不断加剧，学术界对研究人员的要求也在逐步提升，尤其需要类似 AiScholar 这样的平台在研究人员与出版商之间发

挥桥梁作用。传统的学术出版流程存在明显不足,从论文初稿提交到同行评审、修改,再到最终接受,每个阶段都需要投入大量时间和资源^[10]。随着投稿数量不断上升,合格评审人员的短缺也使期刊与学术会议的运行面临困境。此外,研究人员在为其学术成果选择合适的期刊时,尤其在跨学科或小众领域,经常会遇到期刊与研究内容难以匹配的问题^[11]。这些因素导致出版周期延长、成本增加,也在一定程度上打击了研究人员的积极性。针对上述问题,AiScholar 通过 AI 智能体作为中介工具,简化投稿流程,显著提高期刊与审稿人的匹配精准度,并提供初步评估,以辅助专家评审做出更加精准的决策。

AiScholar 开发的“期刊大数据匹配系统”利用大数据算法扫描全面的期刊资料库,根据主题相关性、影响力指标和期刊范围等因素,为研究人员精准地推荐合适的发表期刊。该系统在复杂且缺乏透明性的期刊选择过程中,能够帮助研究人员节省大量宝贵时间。此外,“AI 同行评审系统”运用自然语言处理(NLP)和机器学习模型,对手稿进行初步质量评估,包括语言表达的清晰性、逻辑结构的一致性以及伦理道德规范的符合性等方面。该系统通过预筛选与提供初步反馈,能够使人类评审人员将更多精力集中于更具实质意义的内容评估[4]。然而,即便 AiScholar 推出了上述创新举措,考虑到学术内容的复杂性以及同行评审任务所需的精细化判断,AI 系统仍然无法完全取代人类评审在保障质量方面的作用。因此,进一步优化人机协作框架,以更有效地发挥 AI 工具的优势,显得尤为必要。

专家评审人员时间的不足,也是促使 AiScholar 采用 AI 辅助方案的重要原因之一。传统同行评审工作通常依靠学科专家的志愿贡献完成,然而随着论文投稿量的不断增长,精准匹配合适评审人员的难度越来越大^[12]。在大型学术会议或期刊截稿期临近的高峰时期,这一问题尤为突出。许多情况下,评审人员短缺会导致审稿延迟、评审质量不稳定,同时显著增加了编辑人员的工作负担。AiScholar 期望通过在评审流程的初期阶段融入 AI,缓解这种状况。AI 能够有效执行基础的稿件筛选任务,包括抄袭检测、格式检查以及对论文质量的初步评

估，确保只有符合基本标准的稿件进入后续的专家评审阶段。实践证明，现有基于大语言模型的 AI 工具能够在保障评审质量的基础上，加速审稿流程，提升评审效率，从而有效应对不断增加的论文投稿量。

尽管取得上述进展，AiScholar 的 AI 驱动评审系统在评估质量和准确性方面依然存在一些不足。例如，大语言模型（LLM）在语言表达和结构分析方面表现突出，但在对研究成果的深层次评估，如对专业领域研究成果的新颖性与重要性的精准判断方面，仍然存在局限。为了克服这一问题，AiScholar 提出了人机混合评审模式，由 AI 工具负责初步的质量评估和格式筛查，人类评审员则承担研究成果的深度专业判断。该模式的应用表明，人机协作能够不断深化，AI 技术的价值体现在增强而非取代人类的能力。与此同时，AiScholar 需要进一步提高 AI 在同行评审过程中的应用效果，开发具备更高精细判断能力的复杂模型则显得尤为重要^[4]。

AiScholar 在平台开发战略中也高度关注学术出版过程中涉及的数据隐私与伦理问题。同行评审环节涉及敏感的研究内容，这些内容需要严格的保密措施保障。为此，AiScholar 实施了先进的数据加密技术以及严格的用户访问控制策略。此外，AI 的评估结果需要满足透明性与可解释性的标准，以获得研究人员与评审员的信任。在 AiScholar 的“AI 参与透明化”政策中明确指出，所有 AI 生成的评估意见都会显著标记，便于人类评审人员进行验证。这种策略不仅提高了研究人员对 AI 系统的信任，也有效减少了 AI 输出可能存在的错误或偏差未被察觉的风险。通过对透明度的强调，AiScholar 致力于打造一个使研究人员充分信任 AI 辅助服务的环境，确保研究成果的评估过程公平、可靠且负责^[4]。

AiScholar 开发的“期刊大数据匹配系统”利用大数据算法扫描全面的期刊资料库，根据主题相关性、影响力指标和期刊范围等因素，为研究人员精准地推荐合适的发表期刊。该系统在复杂且缺乏透明性的期刊选择过程中，帮助研究人员节省了大量宝贵时间。此外，“AI 同行评审系统”运用自然语言处理（NLP）和机器学习模型，对手稿进行初步质量评估，评估内容包括语言表达的清晰性、

逻辑结构的一致性以及伦理道德规范的符合性等方面。该系统通过预筛选与提供初步反馈,能够使人类评审人员将更多精力集中于更具实质意义的内容评估。然而,即便 AiScholar 推出了上述创新举措,考虑到学术内容的复杂性以及同行评审任务所需的精细化判断, AI 系统仍然无法完全取代人类评审在保障质量方面的作用^[4]。因此,进一步优化人机协作框架,以更有效地发挥 AI 工具的优势,显得尤为必要。

随着以 GPT-4 为代表的大语言模型 (LLM) 以及其他先进人工智能模型的兴起, AiScholar 提供学术服务的方式也随之发生进一步转变。这些模型在自然语言处理与生成方面表现出色,特别适用于语言密集型任务,例如学术手稿的评估。AiScholar 已将大语言模型整合进同行评审流程,利用这些工具提供全面的语言润色、摘要生成以及初步的稿件质量评估服务。这一举措使平台能够大规模地为研究人员提供更详细且覆盖广泛学科领域的反馈。然而,在同行评审中采用大语言模型也引发了一些任务分配和信任方面的挑战,尤其是研究人员在面对具有高风险决策时,可能会对 AI 提供的意见产生犹疑。对此, AiScholar 将持续努力加以解决,包括通过训练模型,使其能更精准地针对具体任务提供反馈,以及开发交互界面,引导研究人员对 AI 输出结果进行批判性分析^[4]。

1.2 研究问题

随着稿件提交量的快速增加,同行评审的需求也随之显著增长,这导致学术出版环境发生了深刻而迅速的变化。在这一背景下,本研究致力于探讨如何有效整合大语言模型 (LLM) 与相关 AI 工具 (尤其是 AiScholar 平台所提供的 AI 工具),以应对上述挑战。作为专业的学术服务平台, AiScholar 具备借助 AI 提升工作效率、减少稿件评估所需时间与资源、改善作者与评审人员整体体验的潜力。本研究的目的是明确基于 LLM 的 AI 解决方案在当前环境下是否能够创造价值,能否提升同行评审工作的效率,进而推动构建更加高效的学术生态体

系。

本研究的核心问题聚焦于基于 LLM 的 AI 工具对同行评审效率与质量所产生的实际影响^[13]。具体而言，我们将探讨如下问题：由 AI 驱动的 LLM 技术能否有效提升 AiScholar 平台同行评审流程的效率和质量？

同行评审是学术出版中不可或缺的关键环节，传统模式下主要依赖专家进行人工判断^[14]。然而，随着投稿量持续攀升，合格评审人员短缺的问题日益突出，导致论文发表周期延长，评审质量也随之下降。因此，本研究的核心在于探讨 LLM 驱动的 AI 工具是否能够通过初步评估、精确识别适合的发表期刊，并凭借先进的自然语言处理（NLP）能力为人类评审人员提供支持，以缓解当前评审工作所面临的挑战。此外，本研究还试图明确 AI 工具能否减少人类评审员在例行检查任务上花费的时间，使专家更专注于更具实质性且更具价值的评审任务。

为深入分析上述问题，我们将在同行评审流程中 AI 整合的具体应用场景下，探讨以下四个子问题：

1.2.1 子问题 1

第一个子问题关注 AI 工具在提高任务效率方面的作用，即在初步评审阶段引入 LLM，将如何影响整体评审耗时？AI 工具能否通过提供准确的初步评估和有效反馈，持续简化初步评审任务？

在传统同行评审模式中，评审人员经常需要投入大量时间进行诸如语法准确性检查、格式规范审核和伦理道德合规验证等例行工作。尽管这些任务十分重要，但通常不需要学科专家投入过多专业知识。凭借先进的语言处理能力，LLM 有潜力自动完成这些例行检查任务，且在速度和一致性上可能胜过人类评审员。因此，本子问题将分析 LLM 在多大程度上能够替代或增强传统评审的相关工作，并深入考察任务分配模式对整体审稿时间及流程效率的影响。

1.2.2 子问题 2

第二个子问题关注 AI 整合对用户体验的影响，主要从作者与评审员的视角

展开：引入以 LLM 为驱动的 AI 技术后，用户（包括寻求期刊发表的作者以及依赖 AI 辅助进行评审的评审员）的体验和满意度会发生怎样的变化？

同行评审作为一种高度协作的过程，AI 的介入带来了新的变革。作者和评审员对 AI 建议的准确性与可靠性可能持有不同程度的疑虑。例如，研究人员对于 AI 给出的期刊推荐往往心存怀疑，担忧其建议未能充分考虑自身研究工作的细节或涉及的跨学科知识。与此同时，评审员在评估复杂或高度专业化主题时，亦可能质疑 AI 在其专业工作中所能发挥的支持作用。因此，这一问题的意义在于全面了解用户对于 AI 在评审流程中所扮演角色的看法，评估 AI 推荐能否简化流程、提升满意度，还是反而引发了新的挑战。

1.2.3 子问题 3

第三个子问题涉及 AI 辅助评审的质量与严谨性：在同行评审工作流程中引入基于 LLM 的 AI 工具，特别是在那些需要复杂判断和深度领域知识的学科，能否维持甚至提升手稿评估的质量？在同行评审体系中，质量保障始终是首要任务。尽管 AI 技术有助于减轻评审人员的工作负担，但评审质量绝不应因此受到影响。LLM 作为 AI 工具，存在诸如缺乏特定领域专业知识、难以对研究创新性和学术重要性作出准确判断等固有限制，这些因素可能影响 AI 辅助评审的可靠性。该子问题能够明确 AI 在评审任务中的优势与不足，辨析 AI 在哪些环节能够为人类评审提供有效支持，哪些环节则仍需专家知识的介入。同时，研究还将探讨一种人机协作模式，即由 AI 负责处理例行性任务，专家则专注于实质性学术评估，以分析这种协作模式相比传统纯人工评审流程能否提升整体评审质量。

1.2.4 子问题 4

第四个子问题考察评审员对 AI 建议的认知偏见及其先入之见：评审员的认知偏见及其对 AI 的固有看法，在多大程度上影响其对 AI 生成建议的接受度与依赖程度？在实际评审过程中，评审员在采纳与解读 AI 建议时，可能表现出多样的偏见。例如，有的评审员可能对 AI 给予过高信任，未加甄别便接受其建议；

另一些评审员则对 AI 的输出持过度批判态度，忽略其实际质量。相关理论基础为双过程理论，该理论认为人类决策既可以是快速直觉的（系统 1），也可以是深思熟虑的（系统 2）。AI 在同行评审中的应用有可能加剧这些认知偏见，并由此对评审结果产生潜在影响。解决此问题需分析评审员对 AI 建议准确性的感知、其对 AI 输出的信任程度，以及是否有必要制定相关规范，以优化人机协作在同行评审过程中的实际效果。

1.3 研究的意义

本研究在理论与实践层面均具有重要意义，聚焦于大语言模型（LLM）在学术同行评审流程中的应用。同行评审作为学术出版的核心环节，具有劳动密集和时效要求高等特征^[15]。通过深入研究基于 LLM 的人工智能（AI）系统在辅助和优化同行评审流程中的潜力，本文不仅为人机协作理论体系的丰富作出贡献，也为学术出版行业的发展提出切实可行且具备可扩展性的解决方案。当前学术产出量持续增长，评审人员和编辑的需求大幅增加，所以本研究非常及时而且具有很强的针对性，能够有效回应同行评审流程中效率、准确性和适应性等方面的现实诉求。AI 在 AiScholar 平台以及其他流程中的深度融入，预示着 AI 将在学术界和行业层面实现双重突破，并有望重塑知识评价与传播的模式。

1.3.1 理论意义

从理论视角来看，本研究有助于加深对 AI 在学术出版等复杂知识密集型任务中所发挥作用的理解。虽然 LLM 在多种语言处理任务中表现出卓越能力^[16]，但其在同行评审等专业领域的应用，仍需系统探讨其局限性、优势及具体整合机制。理论上，本文能够更加清晰地揭示 LLM 在推动传统依赖人类专家的复杂决策过程中所展现的潜力与限制。目前关于 AI 与同行评审的研究多聚焦于抄袭检测、语法检查等基础应用，而 LLM 在需要高度认知和专业知识的实质性内容评估与评审支持任务中的应用尚缺乏深入研究。

本研究还通过将双过程理论 (Dual Process Theory) 引入同行评审中的人机交互环节, 进一步拓展理论基础。双过程理论指出, 人类决策主要通过两个系统进行: 系统 1 代表快速、直观的思维模式, 系统 2 则体现为缓慢、分析性的思维过程^[17]。在实际同行评审情境中, 评审员通常首先依赖系统 1 进行初步判断, 而深入且批判性的分析则更多依赖系统 2。AI 引入评审流程后, 人类评审员如何在与基于 LLM 的建议交互时平衡这两种认知过程, 成为值得探讨的重要议题。本文分析 LLM 能否通过支持系统 2 的运行, 回应评审意见, 或增强评审员的批判性思维能力, 抑或加剧对系统 1 的依赖的问题。

此外, 本研究有助于推进人机协作中信任与依赖关系的理论研究^[18]。在 AI 领域, 用户对 AI 系统的信任程度是其依赖 AI 生成结果的关键因素, 过高或过低的信任都可能带来不理想的后果^[19]。对 AI 的过度信任, 可能导致用户不加判断地采纳 AI 建议; 而信任不足, 则会抑制 AI 能力的有效发挥^[18]。通过分析评审员在使用基于 LLM 的 AI 系统时表现出的信任水平, 本文从理论层面深入探讨信任校准策略, 实现高效的人机交互。在同行评审这样需要审慎判断与批判性思维的场景中, 合理的信任显得尤为关键, 有助于 AI 成为人类专业知识的有力补充。

本研究还为更广泛的人机协作领域提供了任务委派的理论框架与实践指南^[20]。在同行评审的实际操作中, AI 与人类评审员的任务分配需经过科学规划, 以充分发挥各自优势。AI 适合承担常规和重复性的任务, 从而释放人类评审员投入更高层次、需要专业知识的评估环节。研究成果将为同行评审领域提供可操作性强的框架, 并有望推广至法律分析、医学诊断和财务审计等其他需要复杂人机任务委派的领域^[21]。

1.3.2 实践意义

从实践层面来看, 本研究为正面临同行评审可扩展性和质量双重挑战的学术出版行业, 提供了切实可行的解决方案。AiScholar 在期刊匹配与初步评审环节

部署 AI 工具，成为应对学术提交量增长的重要创新实践。通过对 AiScholar 应用 LLM 的结构化分析，本研究为其他学术服务平台树立了范例，展示了 AI 高效融入同行评审流程的具体路径^[4]。AI 在提升评审流程效率和缩短发表周期方面展现出巨大潜力，这对于解决学术出版领域的核心难题——及时性，具有重要意义。高效且精准的评审流程能够加速科学知识的传播，使研究人员、行业实践者以及社会公众均能受益。

本研究预期的核心实践成果之一，是开发一种 AI 人机协作模型，在不降低评审质量的前提下，提升同行评审效率。随着学术出版物数量的持续增长，传统同行评审体系面临着难以为继的风险。通过在评审流程的初始环节引入 LLM，AiScholar 等平台能够高效筛选提交论文，确保只有高质量且相关性强的稿件进入后续的人工评审阶段^[4]。从长远来看，这种效率提升有助于缓解学术出版的瓶颈，使期刊能够在保证评审严格性的同时，满足持续增长的发表需求。

此外，本研究在 AI 驱动的期刊匹配方面的实践成果，有助于简化研究人员的投稿流程。学者在为研究成果选择合适期刊时常常面临难题，尤其是因期刊数量庞大、关注重点和受众群体各有差异，选择过程尤为复杂^[22]。高效的 LLM 驱动期刊推荐系统可以深入分析稿件内容与研究范围，精准给出推荐意见，帮助研究人员避免投稿延误，提高发表成功率^[23]。如果 AiScholar 的期刊匹配系统被证明有效，亦可为其他平台树立标杆，提升学术出版行业的整体效率与用户体验。

在同行评审过程中，人类评审员在评审判断的主观性和一致性问题等方面，不可避免存在一定的局限性^[24]。本研究对此同样具有重要实践价值。人类评审员容易受到认知偏见影响，例如确认偏误，可能干扰评审过程中的判断与决策^[25]。基于 LLM 的系统能够通过标准化的初步评估，为评审流程提供客观意见，帮助平衡和校正人类偏见，实现更加一致和公平的评估^[26]。例如，AI 能够客观指出论文方法论上的不足、伦理风险或创新性不足等问题，这些内容在人类评审员因

主观偏好或疲劳影响下可能被忽略。因此，AI 有助于关注关键问题，促进构建更为公平可靠的同行评审体系。

本研究还探索了如何借助 LLM 构建更具灵活性的评审基础设施，以应对同行评审中的可扩展性难题。论文提交量存在周期性波动，这一问题在高峰期评审需求显著增加时更为突出。AI 驱动的解决方案可增强评审流程的适应性，使系统能够灵活应对提交量变化。对 AiScholar 等学术平台而言，这种可扩展性对于满足全球学术出版不断增长的需求至关重要。将 AI 融入评审基础设施，能够实现资源按需分配，在不牺牲评审质量的前提下，平衡高峰与非高峰时期的工作负荷^[6]。

1.3.3 管理意义

从管理角度来看，本研究为希望实施 AI 解决方案的学术出版机构提供了具体可行的路径。通过详细剖析 AiScholar 在工作流程中引入 AI 的各项步骤，本研究为其他有意推进数字化转型的机构提供了现实参考。研究指出，可先从风险较低的任务切入，例如初步筛选和期刊匹配，随后逐步拓展到内容评估等更复杂环节。这种循序渐进的引入策略有助于降低 AI 应用的风险，协助出版机构逐步建立对 AI 系统的信任。

而且，本研究围绕用户满意度和对 AI 工具信任度进行实证研究，为学术出版管理提供了重要指导。实现 AI 有效应用，既需要完善技术部署，也需提升用户对 AI 的信任。通过分析影响用户满意度的关键因素，如 AI 推荐的透明度和期刊匹配的准确性，本研究为制定提升用户体验的策略奠定了基础。管理者可据此为用户，特别是评审员，提供针对性的培训，帮助他们更高效地与 AI 系统协作，从而确保 AI 真正成为提升评审质量的助力。

1.4 研究方法

本研究以 AiScholar 为具体应用平台，其方法论设计有助于深入探究大语

言模型（LLMs）对知识型工作者在学术同行评审过程中的表现及其影响。研究综合运用多种方法，力求全面揭示 AI 在复杂且认知要求较高任务中提升或阻碍人类工作表现的具体机制。主要采用的三种研究方法包括文献分析、随机对照实验（RCT）和实证分析。通过多方法结合，研究回应 AI 在同行评审中作用的多层次问题，确保研究结果既具理论严谨性，又具实践相关性。

1.4.1 文献分析

文献分析为本研究提供理论依据和背景框架。通过系统回顾与人工智能（AI）、人机协作、大语言模型及同行评审流程相关的学术文献，研究得以识别当前知识体系中的空白，明确研究问题，并将本研究建立在现有理论和实证成果的基础上。

具体而言，文献分析按照结构化步骤推进。首先，在 IEEE Xplore、PubMed、Scopus 和 Web of Science 等主要学术资料库中展开系统检索，关键字涵盖“AI 在同行评审中的应用”、“人机协作”、“大语言模型”以及“学术出版”等。所选文献横跨计算机科学、管理学、认知心理学和图书馆学等多个领域，确保主题得到多角度、多学科的深入审视。研究聚焦的关键主题包括人类认知的双过程理论、AI 系统的信任与依赖关系、人机 workflows 中的任务分配，以及 AI 对评审偏见和一致性的影响等。

文献分析可为实验设计提供理论支持，也为需要验证的假设搭建了理论框架。例如，文献显示，尽管 LLMs 在自动化写作和分析方面展现出潜力，其在学术同行评审等复杂评价任务中的有效性尚待进一步研究。同时，现有文献强调了人类对 AI 信任与依赖的关键作用，这为本研究关注信任动态对评审员与 AI 辅助工具交互影响提供了坚实基础。在文献分析所得的理论框架的基础上，后续实证研究将聚焦于 LLMs 在同行评审任务中的具体贡献。

1.4.2 随机对照实验（RCT）

随机对照实验是检验 LLMs 对同行评审表现因果影响的核心方法。RCT 被

公认为实验研究的黄金标准，能够实现对变量的精确控制，从而确立因果关系[27]。本研究设计的 RCT 用于评估 LLMs 在同行评审过程中是否能够减轻评审员认知负担、提高评审准确性，并增强整体流程效率。

实验分为若干阶段。首先，参与者（包括具有同行评审经验或学术出版经验的专业人士）随机分配至实验组和对照组。实验组使用 AiScholar 的 LLM 辅助评审工具，对照组则在无 AI 辅助的条件下完成同样的评审任务。该随机分配方式有效控制了潜在混杂变量，确保观测到的绩效差异归因于 AI 工具的应用，而非参与者自身差异。

评审任务被细化为多个子环节，以评估 AI 对流程各环节的具体影响。例如，任务涵盖相关性初筛、语言清晰度评价、方法学严谨性考察和改进建议等方面。通过对实验组参与者与 AI 在各环节互动的系统观察，明确 LLM 在哪些环节能够最有效地提升评审过程。

RCT 的核心结果变量包括评审用时、主观认知负荷和评审准确性。评审用时衡量 AI 辅助是否能够提升效率；认知负荷通过自我报告测量，反映参与者的心理投入程度；评审准确性则通过将参与者的评估与未参与实验的专家基准评分进行对比来判定。这种多维度评估体系既揭示了参与者的主观体验，也兼顾评估的客观质量，为 AI 在同行评审任务中的有效性提供有力证据。

除核心结果变量外，任务满意度和对 AI 帮助的感知等次要指标也将被记录，进一步理解用户对 AI 辅助工具的态度及其对参与度的影响。对这些次要指标的分析，有助于拓展对 AI 信任与依赖的理论讨论，并为如何设计更易于参与者采纳和有效应用的 AI 工具提供参考。

1.4.3 实证分析

本研究以定量实证方法系统评估大语言模型（LLM）辅助对学术同行评审流程的实际影响。数据分析立足于实验一（随机对照实验）和实验二（多组对比设计）所获得的原始数据，结合 Python 等主流统计计算工具，综合采用描述

性统计、推断性检验、回归建模与多变量控制，确保对关键绩效指标、影响路径及因果机制的全面刻画。实证分析既是本研究理论假设验证的核心环节，也是方法论创新和现实应用推广的重要基础。

（1）数据采集与预处理流程

实验数据主要来源于 AiScholar 平台组织的两轮评审实验。实验一采用随机对照分组，将参与者分为 AI 辅助组与人工对照组，实验二则细分为对照组与三种不同 AI 反思干预组，研究五个核心评分维度（内容丰富性、论证逻辑、表达与写作、创新性与重要性、整体评价）的不同评分和表现。原始数据包括各阶段的评分记录、专家基准分、评审用时、以及对 AI 建议的信任度和满意度评分。

数据预处理首先通过 Python 的 pandas 和 numpy 库，对各 CSV 数据文件（如 group_01.csv、group_02.csv 等）实现批量读取、合并和清洗，剔除漏答、异常值和重复样本，并统一变量命名与编码格式。随后，依据实验设计，对“Before-After”评分差异（DIFF）、改进率（IMPROVE）、评分偏差、用时、负荷等核心指标进行计算。采用 Python 实现数据集成、分组计算和逻辑变量生成，为后续统计分析打下坚实基础。

（2）描述性统计分析

数据整理后，第一步为全样本与各实验组的描述性统计分析，主要关注评分改进、平均用时、认知负荷、信任度、满意度等变量的集中趋势与离散程度。统计流程包括均值(mean)、标准差(std)、中位数(median)、四分位数(quartiles)及极值(min/max)等指标计算。利用 Python 的 seaborn、matplotlib 库可实现可视化展示，如箱线图、直方图、条形图等，直观反映各绩效指标在不同分组下的分布情况。为比较组间差异、初步判断实验干预成效提供了基础，并为后续推断性统计与回归模型构建提供依据。

（3）推断性统计与假设检验

在描述性分析基础上，本研究广泛应用 t 检验、方差分析（ANOVA）、卡

方检验及非参数检验等统计工具,验证实验干预的显著性和稳健性。对于“Before-After”评分改进的组内效应,采用配对样本 t 检验衡量反思干预前后评分的变化;对于各实验组间的改进差异,采用单因素 ANOVA 或 Kruskal-Wallis H 检验。评分准确率、改进率等二元变量则应用卡方检验分析组别关联。同时,结合多重比较(如 Bonferroni 校正、Tukey HSD),进一步剖析哪些组别之间存在显著性差异,保证分析结果的严谨性和科学性。相关性分析(如 Pearson、Spearman 系数)用于考察认知负荷与评分改进、满意度与 AI 感知等变量间的关系,揭示绩效提升的心理与行为机制。

(4) 回归模型与多变量控制

为了深入探究 AI 辅助对评审绩效的独立影响,研究进一步构建多元回归模型,将 AI 干预(虚拟变量)、参与者专业程度、任务难度、用时、主观负荷等因素作为参数,评分准确性、改进量、满意度等为因变量,系统控制可能的混杂影响。模型包括线性回归、Logistic 回归及分层回归等,适用于连续型与二元型因变量。通过上述模型可检验 AI 辅助、反思干预等关键机制在控制其他变量后对评审绩效的因果效应,并估算其实际影响大小,为政策建议和理论推广提供量化依据。

1.4.4 假设提出

本研究提出以下假设,并在后续研究中进行了假设检验。主要检验的假设包括:

假设 1: 与无 AI 辅助的评审相比,LLM 辅助能够降低完成同行评审任务的认知负荷。

假设 2: 借助 LLM 工具的参与者在稿件质量评估方面,相较于未使用 AI 辅助者,表现出更高的准确性。

假设 3: 任务满意度与评审过程中对 AI 帮助的感知呈正相关。

对上述假设的实证检验,使本研究能够明晰 LLM 辅助对同行评审过程的因果效应,为 AI 融入学术工作流程提供实证支持。

1.4.5 伦理、隐私与数据安全保障

本研究在设计与实施过程中始终将伦理、隐私与数据安全保障作为最高优先级，全面遵循学术研究的伦理准则和国际通行的科研道德规范。首先，所有参与者在进入实验之前，均被充分告知本研究的背景、目的、流程、所需个人信息的具体内容及其用途，涵盖实验的全部环节。研究团队为每位参与者提供了详细的实验说明书，并通过电子签名的方式获取了知情同意，确保每位参与者在自愿、知情的前提下参与实验。

在数据收集与处理环节，严格采用去标识化与加密存储技术，确保所有个人数据在采集、传输和分析全过程中均无法被追溯到个体身份。与实验相关的原始数据、问卷答案和 AI 交互记录均设有多重安全措施，研究数据仅限项目团队授权成员访问，杜绝任何形式的未经授权的数据泄露。所有分析用脚本与原始数据均不对外公开，进一步降低信息泄露风险。对于敏感信息和关键变量，采用差分隐私等算法做深度匿名化处理，保障数据的不可逆识别性。

在实验界面设计方面，AI 建议以完全匿名的方式呈现，消除由身份、历史或偏见导致的主观干扰。参与者在整个评审流程中被反复提醒：AI 建议仅作为参考辅助，最终的判断、评分与结论必须独立做出，防止出现“自动化偏见”或机械性采纳 AI 建议的情况。研究团队对实验流程设置有动态监控，实时收集参与者的意见、疑虑及潜在风险点，并对出现的不适或困惑进行个别说明与及时干预，保障参与者权益与心理健康。

最后，本研究还建立了完善的投诉与申诉机制。参与者可随时提出对研究过程的质疑或撤回参与权利，无需承担任何责任。所有与伦理、隐私及数据安全相关的管理制度、流程和技术手段，均接受研究机构伦理委员会的审查与监督，确保研究过程合规透明。整个项目结束后，所有数据按照规定时间安全销毁，杜绝滥用与二次泄露可能。

1.5 论文结构

本研究或本报告共分为六章，每章围绕研究目标与方法的不同层面展开，系统探究大语言模型（LLMs）在学术稿件评审场景下提升知识工作成效的路径与机制。各章内容简要如下：

第1章 绪论

第1章全面阐述研究背景，分析在学术出版物数量快速增长的背景下，对高效、精准的同行评审需求日益突出的现实。该章详细梳理学术出版行业所面临的具体挑战，着重探讨 LLMs 在缓解评审流程瓶颈方面展现的潜力。同时，提出 AI 辅助稿件评审是否能够提升效率与成效的核心研究问题，明确人机协作在此背景下的理论及实践价值。此外，本章介绍的研究方法为后续文献分析与实证分析奠定基础。

第2章 文献综述

第2章系统回顾本研究相关领域的文献，内容涵盖 AI 在知识工作中的发展历程、LLMs 的演进与能力表现，以及其在人机协作中的实际应用。该章深入分析 AI 系统的信任机制、影响 AI 高效应用的认知过程，以及在协作环境中影响 AI 任务适应性的关键因素。通过认知科学、任务委托理论和双过程理论等多元视角，批判性剖析 LLMs 在需要专业知识和判断力的任务中体现的优势与不足。文献综述揭示了当前研究在 LLMs 应用于复杂评审任务中存在的理论与实证空白，并为后续实证研究构建了理论基础。

第3章 实验一：LLM 整合对稿件评审复杂任务表现的影响——以论文审阅任务为例

第3章详述实证研究第一阶段，聚焦 LLM 辅助评审对稿件评价的作用。该章详细介绍数据收集过程，包括稿件样本的选取标准，以及 AI 与专家评估的记录方式。同时，具体说明随机对照实验的设计思路，分析 LLMs 是否提升了参与

者的表现，并缩小与专家评分之间的差距，特别是在内容相关性、结构、创新性和重要性评估等方面。通过统计分析实验结果，评估 LLMs 对参与者表现的影响，为后续优化人机协作提供依据。

第 4 章 实验二：优化任务分配以增强人机协作

第 4 章在第 3 章研究基础上，进一步分析在子任务层面优化任务分配对人机协作成效的影响。本章详细介绍一项实验研究，评估在 AI 与人类参与者之间分配不同评审任务的多种配置，分析这些配置对任务表现、效率与认知负荷的具体影响。通过对不同任务分配情境下参与者表现与反馈的系统分析，揭示定制化协作策略对提升人机合作效能的贡献。本章为优化任务分配模式提供实证支持，推动人类专业知识与 AI 能力优势互补的实现。

第 5 章 对研究问题的回应以及假设证实

第五章根据 1.2 节提出的四个子问题，总结了实验一与实验二的实证结果：实验一证明 LLM 辅助显著缩短评审时间（子问题 1）、提升评分准确性与一致性（子问题 3）、并通过认知负荷分析揭示 AI 如何激发系统 2 思维缓解直觉偏差（子问题 4）；实验二进一步比较了四种反思策略，验证了“定向弱项反思”在效率与质量间最佳平衡（子问题 1、子问题 3、子问题 4），同时考察了用户满意度与 AI 感知之间的正相关（子问题 2）。另外，本章还系统汇总并实证证实了本研究提出的三条假设。

第 6 章 研究结论与展望

第 6 章总结实证研究的主要发现，讨论其对学术出版及更广泛知识工作环境的影响。该章结合本研究与既有文献的对比，剖析研究结果的异同，提出在协作评审环境中部署 LLMs 的具体建议。同时，突出本研究在 AI 信任、认知委托以及双过程理论在 AI 与人类协作中的理论贡献。最后，总结研究的局限性，包括结果普适性与技术变迁等因素，并展望未来知识密集型任务下人机协作研究的发展方向。

第二章

文献综述

2.1 人工智能与人机协作的现有研究

人工智能（AI）的发展深刻影响了那些对认知与协作技能有较高要求的领域中的人机交互。最初，AI 应用多集中于数据处理等自动化、重复性和结构化任务，通过简化流程和降低人工劳动需求提升工作效率^[28-30]。随着机器学习和自然语言处理技术的不断进步，AI 的功能逐步扩展到支持更为复杂的认知任务。目前，AI 不仅可以执行自动化操作，还能够在决策支持、客户服务以及医疗诊断等领域为各类工作提供有力支持^[31-33]。

AI 在协作环境中的广泛应用，推动研究者关注如何设计高效的人机协作框架。“机器即队友”模型的提出，表明 AI 已不仅是一种辅助工具，更成为解决问题过程中积极参与的伙伴^[28, 29, 34]。在这些协作框架中，AI 能够根据人类指令做出动态反应，在复杂任务流程中发挥关键作用。这类框架展现的潜力，凸显理解和优化 AI 能力的重要性，也使 AI 在特定任务中的局限性更加明显^[30, 32, 35]。

在 AI 协作研究领域，信任成为核心议题。相关研究显示，信任是决定人机交互成效的关键因素，尤其在 AI 系统逐渐向更高自主性和复杂性发展时，这一问题更加突出。人们对 AI 系统的信任通常建立在其准确性、一致性与可解释性的感知之上^[34-36]。然而，许多基于深度学习算法的先进 AI 系统，如同“黑箱”一般运作，缺乏透明的决策过程，这为用户建立信任构成了挑战^[32, 33, 37]。这种不透明性往往阻碍信任建立，导致用户在高风险情境下对于是否采纳 AI 建议产

生犹豫。已有研究也指出,若 AI 输出能提供解释性说明,帮助用户理解推荐背后的逻辑,便可有效增强用户信任^[29, 34, 35]。

与此同时,学术界也关注到过度依赖 AI 带来的风险。当用户对 AI 信任过高时,往往会忽略必要的核查,进而导致一些原本可以避免的错误^[35, 37, 38]。这一问题在医疗、金融等关键领域尤为突出,未经仔细审查的 AI 输出可能引发严重后果^[33, 36, 39]。此外,认知偏差,如确认偏差和锚定效应,也会影响人类与 AI 的互动,可能导致决策过程被扭曲^[29, 35, 37]。例如,锚定偏差会使用户对初始 AI 建议过度依赖,即使后续信息显示存在其他更优方法^[34, 37, 39]。基于这些问题,设计能够促进反思性思维的 AI 系统十分必要,这有助于降低认知偏差对人类判断的影响^[32, 33, 38]。

任务委托也是人机协作研究中的重要议题。相关研究表明,人类在判断哪些任务适合交由 AI 完成时,尤其是在任务复杂或结果存在不确定性的情境下,常常面临困难^[29, 40, 41]。实现有效的任务委托不仅要求对 AI 的优势与局限有清晰认知,还需具备对自身能力局限性的元认知意识。元认知能够帮助用户更好地评估自身完成任务的能力,进而判断在何种情况下委托 AI 能够带来更高的效率或准确性^[29, 39, 40]。然而,研究发现,许多用户在进行自我评估时存在障碍,这常常导致任务分配决策不理想^[28, 30, 40]。

近年来,学界高度关注适应性与学习在人机协作中的作用。人类和 AI 系统的持续学习与动态适应,是优化协作绩效的关键。用户能够通过理解和分析 AI 输出,优化自身的决策过程,而 AI 则通过学习用户的反馈不断提升其任务表现^[35, 42, 43]。这种双向学习机制有助于建立灵活多变的协作关系,使人类与 AI 能够不断动态演化适应彼此的能力^[29, 35, 39]。

此外,透明性与反馈循环对提升 AI 系统的信任度和可用性具有基础性作用。具备透明性的 AI 系统能够清晰呈现其决策逻辑,这不仅有助于增强用户信任,也能鼓励用户对 AI 输出进行批判性思考与积极参与^[29, 37, 44]。反馈循环机

制允许用户对 AI 输出提出意见,促使 AI 根据人类反馈持续优化自身表现^[29, 33, 35]。通过这些机制,促进人机间的共生关系和持续学习,从而提升整体协作的有效性。

2.2 大语言模型的发展及其对人机协作的影响

大语言模型 (LLMs) 的出现,标志着人工智能在自然语言处理与理解等领域进入了一个具有深远变革意义的阶段。以 OpenAI 的 GPT 系列为代表,这类 LLMs 依托于海量、多样化的数据集进行训练,能够在各类应用场景中生成连贯流畅、上下文关联的文本,其语言表现已趋近于人类写作水平。随着模型能力的持续提升,LLMs 正被越来越广泛地应用于复杂且知识密集型任务,为学术研究、医疗健康、内容创作等众多领域的人机协作带来了前所未有的新机遇。

LLMs 的发展初衷,在于满足同时处理海量、多维度数据并保留语言细致差别与上下文连贯性的需求。早期的语言模型,如 n-gram 模型,受限于简单的架构,能够处理的文本范围十分有限。2017 年, Vaswani 等^[45]提出的 Transformer 架构,为语言模型带来了突破性进展,使其能够捕捉词语间的长距离依赖与丰富的上下文关系。Transformer 引入的注意力机制,使模型能够动态衡量并整合语境信息,极大增强了文本生成的连贯性与准确性。正是这一架构创新,使 LLMs 具备了承担语义、语法及上下文复杂处理任务的能力。

随着基于 Transformer 的模型持续扩展,LLMs 的参数规模实现了指数级增长,显著提升了其对复杂语言模式的处理能力。例如,拥有 1750 亿参数的 GPT-3,已展现出与人类写作水平极为接近的自然语言生成能力。这些高容量模型不仅能够开展对话交流、解析复杂指令、归纳整理大量文献,还能以较高准确度完成创造性写作。其成功不仅源自卓越的文本生成能力,更在于针对具体任务和行业的有效微调,使其在各专业场景下表现优异。随着技术的不断进步,诸如 GPT-4 等更先进的模型进一步提升了上下文感知能力与语言连贯性,拓展了

LLMs 在人机协作中的应用空间。

在人机协作领域，LLMs 在知识驱动型行业中展现出重塑工作流程的显著潜力。学术研究、医疗健康、法律分析和创意产业等领域，都因 LLMs 的引入而迎来流程革新。在学术环境中，LLMs 能够自动生成文献综述、归纳研究发现、起草复杂科学文档，为研究人员提供有力支持。凭借对海量资料库的高效筛选与信息整合，LLMs 显著节省了研究人员的时间与认知资源，使他们能够集中精力于高层次的分析与批判性思考^[46]。此类应用不仅提升了学术生产力，也通过简化流程进一步提高了研究工作的质量。

在医疗健康领域，LLMs 日益被应用于诊断和咨询等高度依赖精准性与上下文理解的场景。例如，基于患者病历与医学文献，LLMs 能够为临床医生生成初步评估意见。尽管 LLMs 不能取代医生在诊断中的专业判断，但在时间压力较大时，它们能快速提供相关信息，为临床决策提供有效参考^[47]。在法律分析领域，LLMs 能够协助律师梳理案例法、总结法院判决，并结合先例构建论证基础。这些应用充分显示出 LLMs 作为认知助手的巨大潜力，能够为各类专业人士减轻工作负担，提升决策效率。

然而，将 LLMs 整合进专业工作流程的过程中，关于可靠性、信任以及责任的多重问题愈发突出。LLMs 本身庞大的参数规模和复杂性导致了众所周知的“幻觉”现象，即模型可能生成看似合理但实际上不准确甚至具有误导性的信息。这种生成未经验证内容的倾向，证明了制定保障机制、确保 AI 生成内容准确性和可靠性的紧迫性。在协作环境中，由于 AI 系统存在不可避免的局限性，人类用户需主动承担监督职责，对 AI 输出进行批判性评估，从而规避在高风险场景下可能产生的潜在错误。

与 LLMs 相关的信任和责任议题因其“黑箱”特性变得尤为复杂。其内部决策机制高度不透明，即便是模型开发者也难以完全解释模型如何得出具体结论。这种透明度的不足，不仅妨碍了用户理解 AI 生成内容背后的推理逻辑，也

削弱了用户对 AI 建议的信任与依赖^[48]。研究表明,用户在理解 AI 推荐背后的基本原理时,更有可能采纳 AI 的建议。因此,为建立高效且基于信任的协作关系,提高 LLMs 输出的透明度、推动模型可解释性成为当务之急。

为应对上述挑战,学界与业界正积极探索提升 LLMs 透明度和责任性的多种路径。一种方法是提供事后解释,使模型能够为其输出生成解释,帮助用户理解内容背后的逻辑。另一种路径是引入反馈循环,使 LLMs 能够根据用户意见调整回应,随着用户互动的不断累积逐步提升内容的相关性和准确性。这样的迭代学习过程不仅能够优化模型的实际表现,还能够通过贴合用户期望与现实情境,进一步增强人机协作关系的紧密度。

LLMs 的大规模应用促使研究者重新审视人类认知与协作的传统理论。以往的人机协作模型假设,机器主要承担重复性、结构化任务,而人类则聚焦于需要判断力和适应性的复杂任务。LLMs 凭借其强大的语言和认知能力,正在改变这一传统分工格局。目前,LLMs 已经能够胜任诸如精细语言翻译、创造性写作,甚至生成具有情感色彩的回应等原本属于人类独有的任务。这一转变对现有的协作模型提出挑战,需要构建能够适应 AI 不断演化的角色的新型协作框架,将 AI 视为主动的协作者而不仅是工具^[46]。

在协作环境中,分布式认知理论与 LLMs 的紧密关联愈加突出。分布式认知理论强调,认知活动不仅发生在个体思维中,更在人与技术系统的互动中实现共用。LLMs 作为认知工作流程的有机组成部分,能够与人类用户协同工作,共同推进问题解决和决策制定。借助 LLMs 完成信息检索、初步数据处理等认知任务,人类能够专注于分析与评估环节,更好地发挥各自的独特优势。

最后,LLMs 作为协作体系的核心环节,也引出了诸多伦理与监管挑战。特别是在 LLMs 对高风险决策产生实际影响的情形下,责任归属问题日益引人关注。例如,当 LLMs 被应用于医疗治疗建议或法律咨询时,一旦发生错误,责任的界定变得尤为复杂。尽管最终责任往往仍由人类用户承担,但 AI 系统在决策流程

中的深度介入，已经让传统责任认定面临新的挑战。因此，为保障 LLMs 的责任应用，亟需制定与之匹配的伦理框架与监管标准，确保人机协作的公平、透明与安全^[47]。

2.3 人工智能与人类合作模式与成效的研究

随着人工智能（AI）持续深入专业领域和组织环境，如何优化人机协作成为实现 AI 应用效益最大化的核心问题。AI 发展初期的应用主要集中在基础自动化任务上，而近年的 AI 模型已具备支持更复杂交互形式的能力，使 AI 能够与人类协同应对复杂任务。相关研究关注多样的合作模式、影响协作质量的关键因素，以及不同人机合作架构带来的实际成效。然而，要深入理解高效协作的内在机制、明确 AI 在具体任务中的最佳定位，并构建适应性强的整合框架，学界仍需应对一系列重大挑战。

在人机协作实践中，任务结构及其与 AI 能力的匹配性是首先需要考量的要素。对于重复性强或基于既定规则的任务，AI 能够高效、精准地处理大规模数据，因此更适合由 AI 承担。相比之下，那些依赖上下文知识、同理心或需细致判断的任务，通常更适合由人类监督执行。Bansal 等^[35]指出，虽然 AI 在数据分析与预测建模等方面展现出优越性能，但在处理复杂社会问题或伦理议题时，表现常常不尽如人意，而人类判断力在这些领域表现出不可替代的价值。这一分析说明，合理的协作需根据任务属性，为 AI 和人类科学分工，以实现双方能力的最大化互补。

明确且互补的劳动分工对于有效人机协作至关重要。Fügener 等^[29]发现，AI 在量化分析领域的高效性，有助于减轻人类在数据处理方面的负担，使人类能够专注于战略决策与创造性任务。但要实现这种理想的协作，人类不仅应了解 AI 的优势，还需充分认知自身的能力边界，这就是所谓的元认知知识。研究表明，具备较强元认知能力的用户能够更准确地判断在何时依赖 AI、何时保留自主决

策权,进而减少过度依赖或利用不足的情况^[28]。这一发现强调了培养元认知技能对优化人机协作关系的重要作用。

用户对 AI 系统的信任程度,会直接影响其采纳 AI 建议的意愿和依赖水平。AI 的准确性、透明度与可靠性均是影响信任的核心要素。具有解释能力的透明 AI 系统,往往更易获得用户信任,因为用户能够明晰理解 AI 输出的逻辑^[44]。然而,信任是动态变化的,会随互动结果而不断调整。AI 输出如能持续保持高准确性,将有助于长期建立信任;若频繁出错,则会削弱用户信心,导致 AI 被低效利用^[49]。在实际应用中,信任校准至关重要,用户需根据 AI 的可靠程度动态调整对其依赖。在 AI 准确性高的场景下,用户可以更大胆地依赖 AI,而在情境不明朗时,仍应保持批判性思考,防止盲目信任。

高效协作还要求用户能够合理管理认知负荷,减少人机互动过程中的认知偏见。认知负荷,即信息处理所需的心理努力,尤其在高风险或多任务环境下,可能阻止用户整合 AI 的辅助能力。Bućinca 等^[38]发现,在医疗和金融等复杂决策领域,用户若难以平衡认知负荷,就可能在整合 AI 建议时出现疏漏或错误。应对这些问题的有效途径是让 AI 系统具备认知强迫机制,即促使用户主动对 AI 建议开展批判性思考。这有助于抵御如确认偏误、锚定效应等认知偏见,从而提升判断的科学性和准确性^[37]。

委派机制在人机协作中起到关键作用。委派意味着将部分决策权转移至 AI 系统,这一过程虽能提升流程效率,但需通过明确协议平衡控制权与信任度。Baird 和 Maruping^[40]指出,随着 AI 自主性提升,在任务风险或不确定性较高时,用户在将复杂任务委派给 AI 时常存顾虑。这一发现证明了制定结构化委派协议的必要性,通过协议明确人类与 AI 在各类任务中的权责分工。如此设置不仅确保 AI 负责日常操作,而且让高影响力决策留给人类专家,以保障协作流程的顺畅与可控。

学习与适应构成高效人机协作的基础。人类与 AI 系统均能通过持续的学习

与适应,不断优化各自在协作中的角色。Noy 和 Zhang^[50] (2023) 发现,用户随着与 AI 系统的互动经验积累,其整合 AI 建议的能力会不断提升,促进更高效的人机协作。同时,具备高度适应性的 AI 系统能够根据用户反馈调整自身行为,使 AI 输出更贴合用户需求与偏好,从而支持更加动态、高效地协作^[35]。这一过程促使人机双方形成灵活、动态演进的共生关系,为未来高效协作提供坚实基础。

在人工智能协作领域,伦理问题在公平、透明与责任等方面具有重要意义。训练数据中潜在的偏见可能被 AI 系统复制,这在招聘、刑事司法等应用场景中表现尤为明显。带有偏见的 AI 输出可能造成不公正结果,从而引发社会关注。对此,研究者主张在 AI 系统中嵌入偏见检测机制,以提升输出结果的公正性^[49, 51]。与此同时,建立透明的 AI 系统同样不可或缺,透明性不仅有助于完善责任追溯机制,也使用户能够追踪并验证 AI 生成的建议。强化责任落实的机制能够增强用户信任,进一步推动人机协作向符合伦理标准的方向发展^[36]。

近年来,人机团队概念逐渐兴起,形成了全新的人机合作模式。在这一模式下, AI 不仅仅作为被动的工具存在,更作为主动的团队成員参与协作,致力于与集体目标保持一致。这种“机器作为团队成员”(machine-as-teammate)模式强调 AI 在团队中的战略性地位,鼓励其支持集体目标与组织规范,推动形成将 AI 视为战略伙伴的协作氛围^[31]。这种做法与“集体智慧”理念紧密相连,即 AI 能够拓展人类的理解力和技能,通过相互支持和协同目标,提升团队整体绩效。

培训与教育对提升人机协作质量具有积极意义。只有当用户深入理解 AI 的能力边界和实际局限,才能在与 AI 互动中实现高水平协作。围绕 AI 素养提升、认知偏差识别和信任校准等主题开展的培训项目,有助于用户更有效地与 AI 系统互动^[35, 39]。这些培训不仅能够纠正用户对 AI 绝对可靠性的误解,还能帮助其树立平衡视角,在充分发挥 AI 优势的同时,保持对 AI 输出内容的批判性思考。

2.4 双过程理论与人机协作

双过程理论作为认知心理学中的核心概念，为理解人类在复杂和高风险环境下的信息处理与决策机制提供了有力的理论基础。随着 AI 生成建议逐步融入各类决策流程，双过程理论对于分析人机协作中的认知反应具有独特价值。该理论区分了两种认知系统：系统 1 表现为反应迅速、直观且趋于自动化，而系统 2 则以缓慢、注重分析和深思熟虑为特征^[52]。在人工智能辅助的人机协作情境中，深入理解这两类认知过程十分重要，因为人类对 AI 建议的采纳往往伴随着直觉性与理性两种反应并存。将双过程理论应用于人机互动研究，能够揭示这两种系统对 AI 建议的不同反应方式，识别潜在的偏见，从而为优化协作环境中的决策质量提供理论支持。

2.4.1 系统 1：对 AI 建议的直觉和初步反应

系统 1 的认知模式高度依赖直觉，其判断过程以快速的启发式思维为主，而非深入分析。在人机协作实践中，系统 1 通常决定了用户对 AI 输出的初步反应。例如，用户可能因 AI 输出表现出的自信和权威风格而下意识地采纳其建议，即使这些建议在实际应用中需要进一步审查^[53]。Buçinca 等^[38]的研究指出，个体在与 AI 互动时，常常倾向于采用系统 1 思维。当 AI 建议与用户已有知识或信念相契合时，这种反应会变得更加明显，既可能带来积极结果，也可能导致不利后果。

尽管系统 1 有助于实现高效的信息处理，其直观性在需要批判性思考的情境下却存在局限性。由于系统 1 倾向于对 AI 的建议过度依赖，用户有时会不假思索地接受 AI 输出，误以为 AI 的复杂性足以保证其可靠性^[35]。另一方面，系统 1 也可能驱使用户因对 AI 的不信任或陌生感而本能地排斥其建议。这些反应揭示了在 AI 辅助决策中，引导用户进入更为深入的认知加工环节的重要性。

2.4.2 系统 2：对 AI 输出的深思熟虑和分析性评估

系统 2 的认知模式表现为缓慢且有条理,强调逻辑推理和全面分析。在人机协作场景下,启动系统 2 有助于用户对 AI 建议进行批判性和全面的审视,使其能够识别 AI 输出中存在的局限或偏见。研究显示,系统 2 的介入明显提升了用户识别 AI 缺陷的能力,在涉及专业知识或情境敏感性较强的复杂任务中表现尤为突出^[29]。以学术评审任务为例,系统 2 的深度参与能够帮助评审人员更有效地判断 AI 生成反馈的学科规范性与研究严谨性^[46]。

系统 2 思维在缓解因系统 1 引发的认知偏见方面具有重要作用。已有研究发现,AI 的高可靠性有时会产生“光环效应”,即用户因过度信任 AI 而忽视必要的审查^[44]。此时,启动系统 2 能够引导用户对 AI 建议进行理性分析,防止决策过程退变为盲目采纳。值得注意的是,切换到系统 2 思维模式需要用户投入额外认知资源,因此这一过程通常不会在低风险或日常互动中自动发生,而需要有意意识的努力和外部引导。

2.4.3 双重加工思维中的认知偏见与人机互动

双过程理论有助于揭示系统 1 与系统 2 思维在交替转换过程中,影响人机互动的具体认知偏见,具体如下:

(1) 确认偏误

确认偏误表现为个体倾向于寻求支持自身先前信念的信息,而忽视与之相矛盾的数据。在利用 AI 辅助决策的过程中,用户可能选择性地解读那些与自身观点一致的 AI 输出,忽略其他不同意见。这种偏误在需要批判性评价的任务中,可能显著制约协作成效^[37]。

(2) 锚定效应

锚定效应指,最初获得的信息对后续判断产生过大影响。在人机协作情境中,AI 给出的首条建议往往成为“锚点”,即便后续有更多相关数据,用户最终决策仍受“锚点”影响。例如,若 AI 推荐某一学术期刊,用户即使面临更优选择,亦可能忽视其他可能选项。Lai 等^[34]指出,锚定效应会降低人机互动中的决策

准确性，因此需采取针对性策略消除该类偏见。

(3) 过度自信偏见

过度自信偏见导致用户高估自身专业能力，特别是在 AI 已展现出明显性能优势的领域。这种偏见可能让用户低估 AI 建议的重要性，从而错失关键见解^[39]。根据双过程理论，启动系统 2 能够帮助用户识别自身这一偏见，通过分析性反思，促使用户更准确地评估自身与 AI 之间的能力差距。

2.4.4 通过双过程理论优化人机协作

借助双过程理论优化人机协作，需要营造鼓励用户适当认知参与的环境，使其能够根据任务要求和 AI 可靠性，在系统 1 和系统 2 思维模式间灵活切换，具体如下：

(1) 认知强迫功能

认知强迫功能是一种有效工具，通过要求用户对 AI 建议进行批判性反思，促使系统 2 思维的启动。例如，AI 系统可向用户发出提示，要求其评估 AI 生成的学术评审反馈背后的逻辑，进而引导用户更深入地评价 AI 建议。Buçinca 等^[38]证实，认知强迫功能可促使用户以更审慎的态度对待 AI 输出，减少过度依赖现象。

(2) 透明性与解释性

提升 AI 的透明性对双过程理论的应用具有重要推动作用。透明的 AI 系统使用户能够依据逻辑标准对 AI 建议进行评估，特别是在需要精细决策的任务中^[44]。协作环境中的高透明性不仅能够提升用户对 AI 的信任，还能促进系统 2 的思维启动。

(3) 反馈与适应机制

反馈机制能够促进用户与 AI 之间的多次迭代互动，实现系统 1 与系统 2 的动态平衡。用户提供反馈后，AI 系统可相应调整输出，更好地贴合用户偏好，从而支持直观反应与批判评价并存^[35]。具备自适应能力的 AI 系统能够营造人

机双方输入不断演变的环境，有助于提升决策质量。

此外，双过程理论还可为人机协作中的任务分配提供理论指导。常规性任务适合系统 1 自动处理，由 AI 高效完成；复杂且高风险任务则应依赖系统 2 进行深度分析。在 AI 支持的流程中，系统 1 使用户能够快速理解 AI 建议，系统 2 则保障关键决策的理性与严谨。

有效的任务委派需要用户具备元认知知识，对自身能力边界和 AI 优势有清晰认知，进而判断何时应依赖 AI、何时应保留自主判断。Fügener 等^[29]指出，元认知知识丰富的用户在任务委派时表现优异，他们通常将常规性检查交由 AI 完成，而将最终判断保留给需要深度分析的环节。这种动态分配模式，既能充分发挥 AI 作用，也不会削弱人类在高阶决策中的主导地位。

在未来研究中，针对人机协作中双过程理论的研究可深入探讨在高风险环境下如何进一步增强系统 2 的参与。例如，开发能够根据用户参与度或任务需求灵活调整输出的 AI 系统，有望实现更高效的决策。同时，考察任务复杂性与时间压力等情境因素对用户依赖系统 1 或系统 2 的影响，将为优化人机互动提供有价值的见解。随着 AI 在专业环境中的广泛应用，深入理解直觉与理性思维的动态作用机制，将成为构建高效人机协作环境的关键。

2.5 文献述评与研究启示

有关人工智能（AI）与人机协作的相关文献显示，在学术同行评审、客户支持以及诊断任务等知识密集型领域，AI 技术展现出提升生产力的巨大潜能^[28-30]。尽管 AI 技术，特别是大语言模型（LLMs），表现出强劲能力，但相关研究同时指出，在涉及专业判断和情境感知的复杂任务中，AI 系统仍面临诸多局限与挑战^[35, 37, 39]。本文引用的参考文献揭示出若干关键问题，包括：信任关系的建立与调节、任务分配优化、认知偏见应对，以及构建支持高效人机协作的综合框架。

信任在人机互动中扮演着核心角色。虽然 AI 的准确性、可靠性与透明度常被视为建立信任的重要基础，但研究指出，不适当的信任水平——无论是过度依

赖还是不足——都可能妨碍有效协作^[33, 44]。进一步分析显示, AI 决策的透明性有助于增强用户信任, 但亦可能导致用户对 AI 建议产生不加批判的依赖, 削弱了人类在关键决策中的作用^[37, 38]。因此, 实现 AI 应用中的合理信任, 亟需依靠动态调节机制, 而不仅仅依赖提升透明度。尽管学界已认识到这一点, 针对如何在不同协作任务中通过实际干预实现信任动态调节的实证研究仍较为有限^[31, 35]。尤其在学术出版等复杂环境下, 开发能根据任务特征动态建立合理信任的方法具有重要研究意义。

认知偏见对人机协作质量同样产生深远影响。诸如确认偏误、锚定效应和过度自信等偏见, 极易扭曲 AI 辅助下的决策过程^[34, 37, 40]。虽然文献强调元知识——即个体对自身局限的认知——对于有效利用 AI 至关重要, 但鲜有具体策略用于培养此类能力^[35, 40]。双过程理论聚焦直觉与分析性思维, 为应对认知偏见提供了前景良好的理论基础, 然而其在实际中的干预应用仍不多见^[38, 52]。未来有必要探索如何借助双过程理论的干预手段, 尤其是在高风险或认知负荷较大的情境下, 加强用户在 AI 辅助任务中的反思能力, 平衡直觉与理性思考^[28, 37]。

文献还重点关注人机任务分配的复杂性。虽然 LLM 在结构化数据处理和语言任务方面表现出色, 但在主观判断或情境依赖性较强的决策任务中, AI 系统可能缺乏必要的深度解读能力^[29, 32, 47]。因此, 已有研究主张采用混合模型: 即 AI 先进行初步处理, 再由人类专家作进一步评估^[30, 35]。然而, 针对不同任务类型动态优化混合模型的实证研究依然较少, 反映出针对具体场景深化任务分配模型研究的现实需要。开发此类模型有助于为复杂协作环境提供科学的任务分配建议^[28, 31]。

人机协作对人类决策能力与技能保留的长期影响亦值得深入探讨。持续使用 AI 既有可能通过反馈强化人类专业知识, 也可能导致人类对日常判断的依赖下降^[33, 44]。未来研究应关注 AI 互动对人类适应性、专业知识发展及技能保留的影响, 以为在知识密集型领域维持高效的人机合作关系提供理论与实践参考^[29, 47]。

第三章

实验一：LLM 整合对稿件评审复杂任务表现的影响 ——以论文审阅任务为例

3.1 研究背景与目的

3.1.1 研究动机

随着学术研究的不断推进，学术期刊收到的投稿数量持续增长，学术论文评审过程面临前所未有的挑战。传统的人工评审模式不仅耗时耗力，而且评审结果的一致性和客观性常常受到评审员个人主观因素的影响，导致评分出现偏差，评审质量缺乏稳定性。这不仅削弱了学术期刊的评审效率，还影响了学术研究质量的有效保障。

近年来，人工智能技术取得显著进展，大语言模型（LLM）在自然语言处理和文本生成领域表现突出。LLM 具备强大的文本理解与生成能力，能够高效处理海量文本数据，为学术评审工作带来新的技术支撑。尽管 LLM 在提升评审效率与一致性方面展现出积极作用，但在涉及论证逻辑、创新性与重要性等主观性较强的评审环节，仍存在明显不足。因此，如何将 LLM 与人类评审员的专业判断有效结合，推动人机协作模式的优化，成为当前亟需解决的研究课题。

实验一的初步结果表明，引入 LLM 辅助评审显著提升了评审的一致性与效率，在内容丰富性和整体评价等维度表现突出。这一发现表明 LLM 能够为评审提供客观、统一的评分标准，也为进一步探索人机协作优化路径提供了实证基础。

不过,实验结果同样显示,LLM 在复杂评审任务中的局限性依然存在,人类评审员在学术评审中的作用依旧不可替代。

在上述背景下,本文深入探讨人机协作模式和反思机制在学术论文评审流程中的应用与优化。通过系统化的实验设计和数据分析,进一步验证优化任务分配策略和引入反思机制的成效,力求充分发挥 LLM 与人类评审员的互补优势,从而提升评审的准确性、一致性与效率。本研究致力于为学术期刊评审流程的智能化和高效化提供理论基础和实践借鉴,推动学术研究实现更高质量的发展。

3.1.2 研究目的

本研究聚焦于通过引入大语言模型(LLM)和反思机制,优化学术论文评审流程,实现评审准确性、一致性和效率的提升。具体目标如下:

1. 验证 LLM 的辅助效果:系统评估 LLM 在提升评审一致性与缩短评审时间方面的作用,特别关注其在内容丰富性和整体评价等关键维度上的表现。
2. 分析人机协作的互补优势:梳理人类评审员与 LLM 在不同评审环节中的最佳协作方式,剖析人机协作的互补性,提升整体评审质量。
3. 设计优化的任务分配策略:依据实验一的数据分析,制定灵活、动态的任务分配方案,结合 LLM 在不同评审维度上的表现,科学划分人类与 AI 在评审流程中的任务比例,保障评审过程的高效与准确。
4. 引入反思机制提升评审效果:在实验二中引入反思机制,引导评审员对 AI 评分结果进行深度分析与调整,减少评分偏差,提高评审一致性和专业性。
5. 评估反思机制的多重影响:考察反思机制在 AI 表现优劣不同维度上的实际效果,分析其在提升评审准确性与一致性方面的具体作用,并关注其带来的潜在负面影响,如评审员认知负荷增加和对 AI 工具的过度依赖等。
6. 提出实用的优化建议与策略:结合实验结论,提出可操作的优化策略和应用建议,为学术评审及其他知识密集型领域的人机协作模式提供理论支持和实践指导,促进 AI 与人类评审员实现高效协作。

通过实现上述目标，本研究为学术论文评审流程的智能化与高效化提供科学基础和实证支撑，助力学术研究迈向高质量发展。

3.2 数据与方法

3.2.1 实验设计

本研究评估了基于大语言模型（LLM）的人工智能在学术论文评审过程中的辅助成效，关注其在提升评审员评分表现及缩小与专家评分偏差方面的实际作用。研究首先从公开的会议论文集和 arXiv.org 网站中，精选 70 篇与人工智能相关的论文。随后，邀请 26 位具有丰富发表和审稿经验的领域专家，对这些论文按照内容丰富性、论证逻辑、表达与写作、创新性与重要性以及整体评价五个维度进行评分。每篇论文由至少两位专家独立评分，并取其平均分作为最终得分。然后，利用大语言模型（GPT-4 Turbo）并依据与专家相同的评分标准，对所有论文在五个维度上生成评分与简要评价。基于专家评分结果，从中选取涵盖不同质量水平的 5 篇论文，作为问卷实验材料。

在问卷调查阶段，研究共招募 89 位来自计算机科学、信息系统等领域的硕士和博士研究生参与。随机分配后，30 位为对照组，59 位为实验组。所有参与者需对前两篇论文独立打分，然后对后三篇论文，实验组需参考大语言模型生成的评分后再给出分数。为确保两组在专业背景、审稿经验和日常 AI 使用频率等变量上的均衡，对这些变量进行了统计分析，结果显示两组在各项关键变量分布上基本一致。

实验数据分析采用单因素方差分析（One-way ANOVA），结果表明，对照组与实验组在专业背景、审稿经验和日常人工智能使用频率等方面无显著差异。描述性统计显示，实验组与专家评分的偏差在所有评审维度上均低于对照组。密度图分析进一步显示，实验组偏差分布整体左移，表明其评分更接近专家水平。回归分析结果证实，引入大语言模型可显著降低参与者与专家评分之间的偏差，

在整体评价和内容丰富性两个维度的提升尤为明显。此外，参与者的自我感知难度与实际评分偏差不完全一致，且在 AI 辅助下，参与者在自感难度较高的维度上表现得到改善。研究结果说明，在学术论文评审等复杂任务中，人工智能工具与人类评审员协同，可实现能力互补，有效提升评审准确性和一致性，为人机协作模式的优化提供了实证支持。

3.2.2 样本与参与者

本节仅补充说明本小节的特殊设计；关于样本招募口径、变量定义与均衡性检验等通用内容，详见 3.2.1。

本研究最终纳入 89 名计算机科学与信息系统等相关领域的硕、博士生，采用随机分配为对照组 30 人与实验组 59 人。对照组全程依传统流程独立评分；实验组先对前两篇论文独立打分，随后在后三篇评分时参考 GPT-4 Turbo 输出的结构化评审要点后再行独立打分，以识别「AI 建议—人类最终判断」的净效应。

为降低组间差异干扰，本研究事先收集并在分析中控制三类关键背景变量：是否具 AI 相关专业背景、既有审稿经验、与日常 AI 使用频率；其分布与组间均衡性的统计结果已于 3.2.1，本节不再重复。

3.2.3 评审材料准备

材料来源与初评流程详见 3.2.1；本节仅呈现材料选择标准与分布（表 1）、样本特征与均衡性（表 2/ 表 3）及流程概览（图 1）。

表 1 选取的评审论文质量分布

质量水平	论文数量
高	2
中	2
低	1
总计	5

为进一步把控实验的外部变量，在研究设计过程中，充分考虑了参与者的专

业背景、论文审稿经验以及日常使用人工智能的频率等因素。我们通过问卷调查的方式收集这些信息，并在实验设计环节对这些因素进行均衡分配，以此确保对照组和实验组在这些关键变量上的分布状况基本保持一致。

表 2 参与者样本特征分布

变量	对照组 (N=30)	对照组 (N=30) 标准	实验组 (N=59)	实验组 (N=59)
	平均值	差	平均值	标准差
专业背景 (AI 相关)	0.9	0.3	0.8	0.4
论文审稿经验	1.9	1.0	1.5	0.8
日常使用人工智能频率	5.9	1.0	6.0	0.8

表 3 单因素方差分析结果

变量	F 值	Pr(>F)
专业背景 (AI 相关)	0.058	0.811
论文审稿经验	3.57	0.0622
日常使用人工智能频率	0.171	0.68

从表 1、表 2 和表 3 的结果可以看出，对照组与实验组在论文质量分布、专业背景、论文审稿经验以及日常使用人工智能频率等关键变量上的分布基本一致。单因素方差分析结果表明，两组在上述变量上的差异均未达到显著性水平（专业背景： $F(1, 87)=0.058, p=0.811$ ；论文审稿经验： $F(1, 87)=3.57, p=0.0622$ ；日常使用人工智能频率： $F(1, 87)=0.171, p=0.68$ ）。上述分析结果表明，两组在主要特征方面具有良好的可比性，为后续实验结果分析提供了可信基础。

为确保所选 5 篇论文能够准确反映 AI 辅助对评分表现的影响，研究团队对这些论文进行了细致的质量评估，确保它们在各评审维度上的评分偏差具有代表性。这一严格的筛选与评估过程，为比较实验组与对照组评分结果提供了坚实的数据依据。

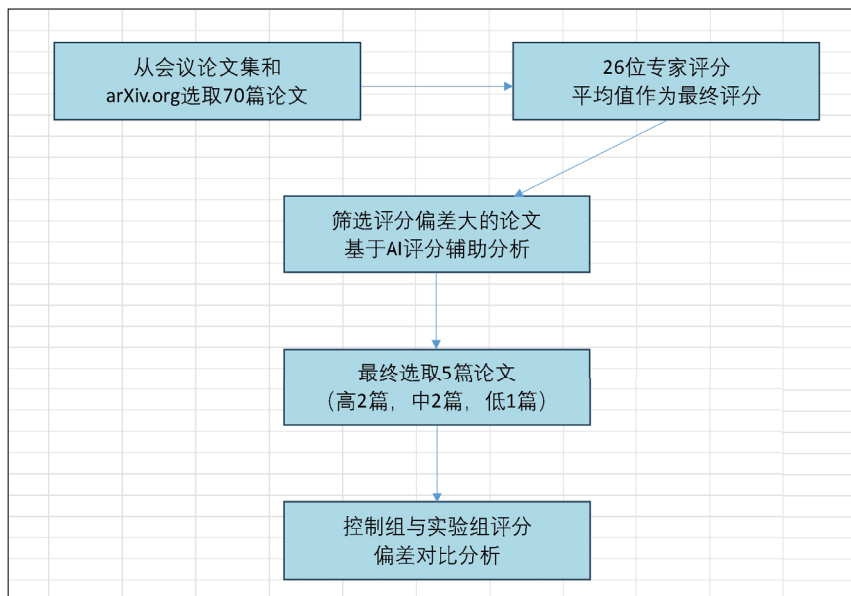


图 1 评审材料选择流程示意图

本研究实验阶段均使用 GPT-4 Turbo，并确保可比性与可重复性而保持固定的生成参数：temperature = 0.3、top_p = 1、max_tokens = 4096。评审任务采用结构化提示语（含任务目标、评审维度定义、5 分制锚点与输出模板）；完整提示语与参数细节见「附录 C」。提示语设计综合了主流国际学术会议审稿指南（如 NeurIPS、ACL、CVPR、ICLR 等）的通用规范，按规范的评审语言与要求撰写，提示词详情见附录 C。

经由上述评审材料的准备与选择流程，本研究切实保障了实验的科学性以及数据的有效性，为深入探究 AI 与人类协作在学术评审中的应用成效，奠定了坚实稳固的基础。

3.2.4 评分维度的定义与控制

在本研究中，评分维度的精准界定和严格把控，是确保实验结果具备有效性与可靠性的核心基础。为全面评估学术论文的质量，研究选取了五个关键评分维度，包括内容丰富性、论证逻辑、表达与写作、创新性与重要性以及整体评价。

下文将对每一评分维度进行详细说明，并阐释实验过程中对这些维度的具体管控与分析方法。

内容丰富性主要考察论文的研究问题是否清晰、理论框架是否扎实、论述是否全面。一篇内容丰富性高的论文，需具备明确的研究目标、系统的理论支撑和详实的论证过程，能够覆盖研究主题的各个关键方面。

论证逻辑则聚焦于论文结构的合理性、论证过程的严密性和各部分之间的连贯性。论证逻辑严谨的论文，应当在各章节内容之间建立清晰的逻辑联系，确保论证链条的完整性，避免出现逻辑漏洞或不当推断。

表达与写作维度主要评价论文的语言表达、写作规范以及语法准确性。高水平的表达与写作不仅有助于准确传递研究内容，还能提升论文的可读性与学术规范性，有效降低因语言问题带来的理解偏差。

创新性与重要性侧重于衡量论文在研究方法、理论贡献和实际应用方面的独特性与影响力。具有突出创新性和重要性的论文，往往能够在方法论、理论体系或应用实践中展现出突破性进展，为学科领域带来新的见解与发展方向。

整体评价则综合前述各维度，全面衡量论文的学术价值与实际应用潜力。这一维度反映了论文在整体质量层面的表现，是对内容丰富性、论证逻辑、表达与写作、创新性与重要性等方面的整体权衡。

为保证评分过程的一致性与客观性，研究团队制定了详尽的评分标准，并对所有参与者进行了系统培训。此外，通过引入大语言模型（LLM）生成的评分作为参考，进一步规范了评分流程，尽可能降低了主观因素对评审结果的影响。

表 4 评分维度定义与评分标准

评分维度	定义	评分标准
内容丰富性	评估研究问题的明确性、理论框架的扎实程度及论述的全面性	1 分（很差）：文章内容非常单薄，缺乏必要的细节和深度，未能提供足够的信息，结构不完整，缺少关键部分，没有或缺乏说明性图表。 2 分（差）：文章内容较为简单，仅覆盖了一些基本要点，结构尚可但不够完整，信息量有限，图表较少且辅助说明作用不明显。 3 分（普通）：文章内容较为充实，覆盖了大部分相关要点，提供了一定量的信息，结构基本完整，但某些部分可能仍需进一步丰富和细化，图表数量适中，能较好地辅助文本内容。 4 分（良好）：文章内容丰富，详尽地探讨了研究问题，提供了大量的信息和数据支持，结构完整，图表丰富且有效地支持了文章的论点和分析。 5 分（优秀）：文章内容极为丰富和深入，全面地覆盖了研究领意的各个方面，结构严谨，内容层次分明，图表数量充足且设计精良，能够直观地展示研究结果和数据分析，极大地增点了文章的说服力和学术价值。
论证逻辑	评估论文的论证结构、逻辑严密性和论点的连贯性	1 分（很差）：文章的论点支离破碎，缺乏论证逻辑，没有可辨识的逻辑结构。 2 分（差）：文章少数内容符合逻辑，但频繁出长逻辑跳跃，观点间缺少清晰联系。 3 分（普通）：文章大体上逻辑流畅，大部分观点之间关联明确，尽管某些论证可能并未完全展开或缺乏联系。 4 分（良好）：文章论证结构合理、思路清晰且论证逻辑点，只有少数部分出长轻微的连贯性问题。 5 分（优秀）：文章论证连贯，逻辑结构十分严密，每一部分都与上一部分紧密联结。
表达与写作	评估语言表达能力、写作规范性及语法正确性	1 分（很差）：文章缺乏流畅性，既不清晰也不简洁，在许多关键部分用词不准确或出长严重语法错误，行文不符合学术写作风格，阻碍了读者的理解。 2 分（差）：文章有些部分比较清晰和流畅，但整体缺说比较冗长，一些用词或语法错误也给读者理解论文内容造成了困难。 3 分（普通）：文章大体上流畅、清晰且简洁，只有少量用词不当或语法错误，行文基本符合学术写作风格。 4 分（良好）：文章表达流畅且清晰，核心部分阐述简洁明了，基本没有用词或语法错误，遵循学术写作的风格。 5 分（优秀）：文笔十分流畅，表达出色，清晰且简洁有力，深入浅出地阐述复杂论点，完全没有用词或语法错误，严格遵循了学术写作风格。
创新性与重要性	衡量研究方法新颖性、理论贡献和实际应用潜力	1 分（很差）：文章缺乏原创思想或对该领意的贡献，未能提供任何新见解或改进。 2 分（差）：文章的新颖性和意义极为有限，仅在已知概念上做出轻微改变，没有为该领意增加实质研究价值。 3 分（普通）：文章引入了一些新颖的观点或方法，为该研究领意做出了一定贡献。 4 分（良好）：文章具有高度的创新性与重要性且比较重要，推动了该研究领意的发展与进步。 5 分（优秀）：论文具有开创性，提供了全新且意义重大的贡献，大大推动了领意内最先进的技术进步。
整体评价	综合考虑创新性与重要性、学术价值和实际应用潜力	1 分（很差）：论文存在重大缺陷，除非经过重大修订，否则无法发表。 2 分（差）：论文具有一定价值，但需要进行重大改进才能发表。 3 分（普通）：论文可以碎受但并不突出；它满足基本标准，但缺乏重要贡献或存在比较显著的问题。 4 分（良好）：论文在大多数方面表长良好，但是可能存在一些小问题。 5 分（优秀）：论文在所有方面都表长出色，显然值得碎受，最多只需进行轻微修订。

变量控制与显著性检验的统计实现见 3.2.6；组间均衡性的结果见 3.2.3 表 3。

3.2.5 数据收集与工具

样本与材料来源见 3.2.1 与 3.2.3（表 1、表 2、表 3 及图 1）；本节聚焦工具栈与流程节点（见表 5、表 6）。背景变量的收集与展示见 3.2.3 表 2、表 3；其在模型中的控制做法见 3.2.6。

本研究结合问卷调查与线上实验两种方式收集数据，以保证数据的完整性与可靠性，具体流程如下：

为确保数据的准确性与一致性，研究团队采用了多种工具与方法。Credamo 见数线上问卷系统用于问卷设计和数据收集，确保高效获取参与者的评分及背景信息。GPT-4 Turbo 被用作 AI 辅助评分工具，为实验组提供评分参考。数据处理与统计分析则借助 R 语言和 SPSS 软件进行，涵盖描述性统计、单因素方差分析（ANOVA）及回归分析等方法，确保分析结果的科学性与严谨性。

表 5 数据收集工具和平台

工具/平台	描述
问卷调查平台	使用 Credamo 见数线上问卷系统设计并发布问卷，收集参与者打分数据和背景信息
大语言模型(LLM)	采用 GPT-4 Turbo 生成 AI 评分，辅助实验组进行论文打分
统计分析软件	使用 SPSS 进行数据处理和统计分析

在数据收集过程中，作者对问卷进行了预测试，确保题目设计的有效性和清晰性。所有参与者在完成问卷后，其数据被自动记录并导出至分析软件中进行后续处理。为保障数据的质量，研究团队对收集到的数据进行严格的质量控制，剔除无效和异常数据。

表 6 数据收集流程

步骤	描述
1. 论文选择	从会议论文集和 arXiv.org 选取 70 篇 AI 相关论文
2. 专家评分	26 位领域专家对 70 篇论文进行评分，每篇论文至少两位专家评分并取平均值
3. 实验材料选择	根据专家评分和 AI 评分选择 5 篇不同质量水平的论文作为实验材料
4. 问卷设计与发布	设计问卷，分配参与者至对照组和实验组，收集打分数据
5. 数据处理与分析	使用 R 语言和 SPSS 进行数据处理，进行描述性统计和回归分析
6. 质量控制与验证	剔除无效数据，确保数据质量，并通过 ANOVA 验证组间变量均衡性

通过上述数据收集与工具的准备和使用，本研究确保实验的科学性和数据的有效性，为探讨 AI 与人类协作在学术评审中的应用效果奠定坚实的基础。

3.2.6 数据分析方法

本研究运用多种统计分析方法，全面评估大语言模型（LLM）以及反思机制在学术论文评审过程中的辅助成效。数据分析流程涵盖描述性统计分析、单因素方差分析（One-way ANOVA）以及多元回归分析。接下来，将对各项分析方法及其具体应用展开详细阐述。

为避免重复，本节仅提供样本统计与 ANOVA 简表(见表 7、表 8)供快速查阅；完整数据与统计指标请参见 3.2.3 的表 2、表 3。随后的回归模型设置与估计结果按原计划展开。

表 7 对照组与实验组的样本统计

变量	对照组 (N=30) 均值± SD	实验组 (N=59) 均值± SD
专业背景 (AI 相关)	0.90 ± 0.30	0.80 ± 0.40
论文审稿经验	1.90 ± 1.00	1.50 ± 0.80
日常使用人工智能频率	5.90 ± 1.00	6.00 ± 0.80

表 8 单因素方差分析结果

变量	ANOVA 检验结论
专业背景 (AI 相关)	组间差异未达显著 (详见表 3)
论文审稿经验	组间差异未达显著 (详见表 3)
日常使用人工智能频率	组间差异未达显著 (详见表 3)

在确保组间可比性的基础上，进一步进行回归分析，以探讨引入大语言模型对参与者打分表现的影响。回归分析采用多元线性回归模型，将参与者在后三篇论文上的打分偏差均值作为因变量，实验组 (Treatment) 作为主要参数，同时控制专业背景、论文审稿经验、日常使用人工智能的频率、前两篇论文的打分偏差均值 (Baseline Deviation) 以及自我感知难度 (Perceived Difficulty) 等变量。交互项 (Treatment * Perceived Difficulty) 用于检验反思机制在不同感知难度下的调节效应。

3.3 结果分析

3.3.1 描述性统计

本节将呈现实验一中对照组与实验组在各评审维度上的评分偏差均值和标准差，以及大语言模型 (LLM) 的评分准确性。

首先，比较对照组与实验组在五个评审维度（内容丰富性、论证逻辑、表达与写作、创新性与重要性和整体评价）上的评分偏差均值。具体比较结果见下图：

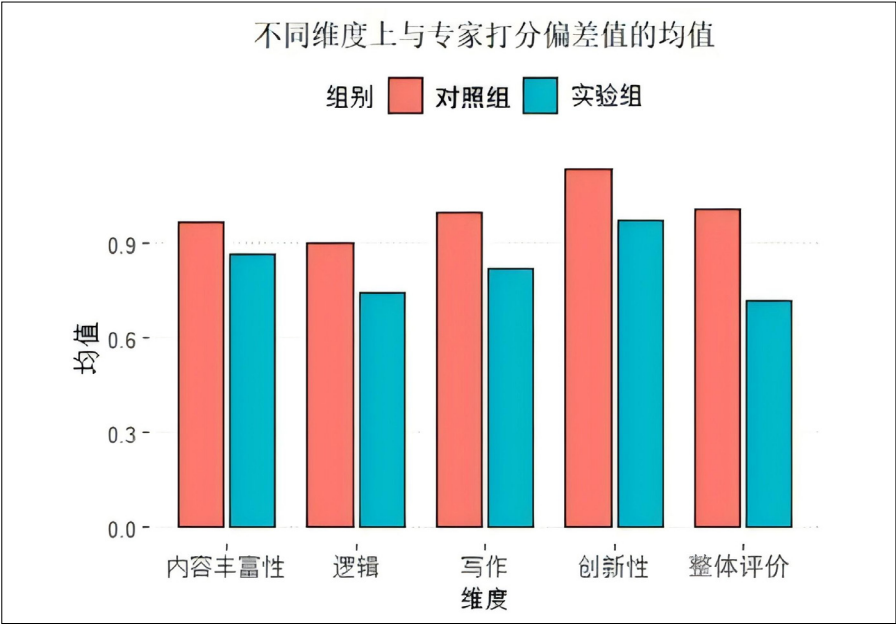


图 2 不同维度上与专家打分偏差的均值对比图

上图直观展现了实验组与对照组在学术论文评审任务中的表现差异，并反映出大语言模型（LLM）辅助所带来的实际效果。通过比较各评分维度与专家评分的均值偏差，可以清晰看到，实验组在所有评审维度上的评分偏差均低于对照组。内容丰富性和整体评价两个维度的偏差降幅最为明显，显示 LLM 在这些概括性和综合性要求较高的维度上具有较强的辅助价值。在创新性与重要性维度，实验组评分偏差虽有所下降，但降幅相对有限。这一结果可能与创新性与重要性维度本身的主观性和复杂性密切相关，说明 LLM 在处理这类高主观性任务时仍存在一定局限。

为更准确地评价 LLM 在各评审维度上的评分能力，研究将 LLM 评分与专家评分的差值在 0.5 分以内界定为“准确”，相关准确率见下表 9。

表 9 大语言模型在各评审维度上的评分准确率

评审维度	准确率
内容丰富性	72.8%
整体评价	72.0%
表达与写作	63.2%
论证逻辑	60.8%
创新性与重要性	56.0%

表 9 显示 LLM 在「创新性与重要性」维度的准确率为 56%。个案复核表明，误判主要集中于两类情形：（1）跨边界创新：跨学科融合或问题重构型研究因偏离既有语料 / 范式，难以被基于分布学习的模型有效锚定，导致新颖性被低估；（2）「新瓶旧酒 / 旧瓶新酒」的表征混淆：当论文在表达与结构上高度规范、技术有效性与清晰度较高，但实质性创新较隐蔽时，模型依据形式要素给出偏高的新颖性判断。上述模式与近来对 LLM 的观察一致：模型更擅长识别技术有效性与清晰度，相对忽视新颖性这一「非分布常态」信号。

基于此，创新性维度更需要人工系统 2 审视与反例检索作为校准机制：一是在评审流程中加入「相对于最接近基线，何为新增？」「是否存在跨域重组证据？」等强制对比型提示；二是将新颖性—清晰度拆分评分与锚点，避免形式清晰度牵动对创新性的投射偏差；三是对跨学科稿件引入领域标注 / 词汇映射与人审复核节点，降低模型对非典型语义的漏检风险。这些做法与我们在反思策略部分的设计相协同，可在不显著增加负荷的前提下提升创新性判读的灵敏度与准确度。

从表 9 可以看出，LLM 在内容丰富性和整体评价两个维度的评分准确率分别达到 72.8% 和 72.0%，表明 LLM 在这些维度上与专家评分保持了较高的一致性。相比之下，论证逻辑、表达与写作以及创新性与重要性等维度的评分准确率相对较低，说明在这些主观性较强的领域，LLM 的辅助效果还有进一步提升的空间。

整体而言，描述性统计结果表明，引入 LLM 能够显著缩小参与者与专家评

分之间的偏差，其中内容丰富性和整体评价两个维度的改善尤为明显。然而，LLM 在论证逻辑、表达与写作、创新性与重要性等维度的评分准确性仍有提升余地。

3.3.2 回归分析

为深入分析大语言模型（LLM）对参与者打分表现的影响，本研究采用多元线性回归方法。在构建回归模型时，将参与者在后三篇论文上的打分偏差均值作为因变量，将实验组（Treatment）设为主要因素。同时，控制专业背景（AI Major）、论文审稿经验（Paper Reviewing Experience）、日常使用人工智能的频率（AI Use Frequency）、前两篇论文的打分偏差均值（Baseline Deviation）以及自我感知难度（Perceived Difficulty）等变量。此外，模型中还设置了交互项（Treatment * Perceived Difficulty），用于检验反思机制在不同感知难度下的调节效应。

（1）整体评价维度的回归分析结果

表 10 整体评价维度的回归分析结果

变量	模型 1 (未控制)	模型 2 (控制基线)	模型 3 (加入交互项)
Constant	0.879** (0.279)	0.916** (0.342)	1.436** (0.486)
Treatment	-0.273** (0.094)	-0.242** (0.089)	-0.852* (0.327)
AI Major	-0.016 (0.109)	0.000 (0.105)	-0.016 (0.106)
Paper Reviewing Experience	0.056 (0.058)	0.025 (0.054)	0.031 (0.055)
AI Use Frequency	0.007 (0.047)	-0.004 (0.045)	-0.026 (0.048)
Baseline Deviation: Overall		0.190* (0.090)	0.131 (0.104)
Perceived Difficulty in Overall		-0.029 (0.029)	-0.116* (0.056)
Treatment * Perceived Difficulty in Overall			0.147*

表 10 整体评价维度的回归分析结果（续）

变量	模型 1 (未控制)	模型 2 (控制基线)	模型 3 (加入交互项)
			(0.072)
样本数量 (Num. Obs.)	89	89	89
R ²	0.131	0.203	0.249
调整后的 R ² (R ² Adj.)	0.089	0.144	0.184

注： $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$
变量名称说明（本文中相同，以下不再赘述）：
Constant：常数项
Treatment：引入 AI 的处理效应
AI Major：参与者是否为人工智能相关专业
Paper Reviewing Experience：参与者的论文审稿经验
AI Use Frequency：参与者日常使用人工智能的频率
Baseline Deviation: {dimension}：参与者在 {dimension} 这一维度上前两篇文章上的打分和专家打分的偏差均值
Perceived Difficulty in {dimension}：参与者在 {dimension} 这一维度上的自我感知难度
Treatment * Perceived Difficulty in {dimension}：{dimension} 这一维度上引入 AI 的处理效应与参与者自我感知难度的交互项
{dimension}：共包含 5 个维度，分别为内容丰富性 (richness)、论证逻辑 (logic)、表达与写作 (writing)、创新性与重要性 (novelty) 和整体评价 (overall)

从表 10 可以看出，引入大语言模型（Treatment）使模型 3 的处理效应达到 -0.852，且在 5% 的显著性水平下统计显著（ $p < 0.05$ ），表明实验组在整体评价维度上的打分偏差均值显著降低。此外，参与者的自我感知难度与打分偏差呈负相关（-0.116， $p < 0.05$ ），而引入 AI 后的交互项系数为 0.147，说明在高感知难度的情况下，AI 辅助对降低打分偏差的效果有所增强。

（2）内容丰富性维度的回归分析结果

表 11 内容丰富性维度的回归分析结果

变量	模型 1 (未控制)	模型 2 (控制基线)	模型 3 (加入交互项)
Constant	0.612* (0.305)	0.709+ (0.416)	1.352*** (0.497)
Treatment	-0.048 (0.101)	-0.045 (0.101)	-0.771** (0.262)
AI Major	0.030 (0.133)	0.045 (0.134)	0.059 (0.146)

表 11 内容丰富性维度的回归分析结果（续）

变量	模型 1 (未控制)	模型 2 (控制基线)	模型 3 (加入交互项)
Paper Reviewing Experience	0.138** (0.051)	0.120* (0.055)	0.107+ (0.059)
AI Use Frequency	0.012 (0.054)	-0.004 (0.059)	-0.021 (0.057)
Baseline Deviation: Richness		0.077 (0.103)	0.025 (0.097)
Perceived Difficulty in Richness		-0.024 (0.040)	-0.151* (0.070)
Treatment * Perceived Difficulty in Richness			0.200** (0.071)
样本数量 (Num. Obs.)	89	89	89
R ²	0.109	0.120	0.203
调整后的 R ² (R ² Adj.)	0.066	0.055	0.134

注: $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

在内容丰富性维度上，引入大语言模型使模型 3 的处理效应达到 -0.771，且在 1% 的显著性水平下统计显著（ $p < 0.01$ ），表明实验组在内容丰富性维度上的打分偏差显著降低。此外，交互项 Treatment * Perceived Difficulty in Richness 的系数为 0.200，且在 1% 的显著性水平下统计显著，说明在高感知难度的情况下，AI 辅助对降低打分偏差的效果进一步增强。

（3）论证逻辑维度的回归分析结果

表 12 论证逻辑维度的回归分析结果

变量	模型 1 (未控制)	模型 2 (控制基线)	模型 3 (加入交互项)
Constant	0.093 (0.241)	0.180 (0.287)	0.599+ (0.339)
Treatment	-0.154	-0.141	-0.707**

表 12 论证逻辑维度的回归分析结果（续）

变量	模型 1 (未控制)	模型 2 (控制基线)	模型 3 (加入交互项)
	(0.098)	(0.085)	(0.263)
AI Major	0.058	0.089	0.033
	(0.121)	(0.132)	(0.135)
Paper Reviewing Experience	0.031	-0.009	-0.041
	(0.050)	(0.053)	(0.050)
AI Use Frequency	0.118**	0.107**	0.100**
	(0.041)	(0.038)	(0.037)
Baseline Deviation: Logic		0.192*	0.025
		(0.077)	(0.077)
Perceived Difficulty in Logic		-0.038	-0.099*
		(0.026)	(0.038)
Treatment * Perceived Difficulty in Logic			0.112*
			(0.050)
样本数量 (Num. Obs.)	89	89	89
R ²	0.122	0.208	0.253
调整后的 R ² (R ² Adj.)	0.080	0.150	0.188

注: $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

在论证逻辑维度上，引入大语言模型使模型 3 的处理效应达到 -0.707，且在 1% 的显著性水平下统计显著（ $p < 0.01$ ），表明实验组在论证逻辑维度上的打分偏差显著降低。交互项 Treatment * Perceived Difficulty in Logic 的系数为 0.112，且在 5% 的显著性水平下统计显著，说明在高感知难度的情况下，AI 辅助对降低打分偏差的效果有所增强。

（4）表达与写作维度的回归分析结果

表 13 表达与写作维度的回归分析结果

变量	模型 1 (未控制)	模型 2 (控制基线)	模型 3 (加入交互项)
Constant	0.656* (0.284)	0.447 (0.297)	0.720+ (0.365)

表 13 表达与写作维度的回归分析结果（续）

变量	模型 1 (未控制)	模型 2 (控制基线)	模型 3 (加入交互项)
Treatment	-0.176+ (0.090)	-0.141 (0.086)	-0.432* (0.205)
AI Major	0.032 (0.139)	0.051 (0.133)	0.033 (0.135)
Paper Reviewing Experience	0.015 (0.047)	0.013 (0.049)	0.009 (0.047)
AI Use Frequency	0.049 (0.047)	0.042 (0.045)	0.031 (0.047)
Baseline Deviation: Writing		0.273*** (0.071)	0.259*** (0.072)
Perceived Difficulty in Writing		0.015 (0.029)	-0.038 (0.045)
Treatment * Perceived Difficulty in Writing			0.086 (0.060)
样本数量 (Num. Obs.)	89	89	89
R ²	0.068	0.194	0.213
调整后的 R ² (R ² Adj.)	0.024	0.135	0.145

注: $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

在表达与写作维度上，引入大语言模型使模型 3 的处理效应达到 -0.432，且在 5% 的显著性水平下统计显著（ $p < 0.05$ ），表明实验组在表达与写作维度上的打分偏差显著降低。尽管交互项 Treatment * Perceived Difficulty in Writing 未达到显著性水平，其系数为 0.086，表明在高感知难度的情况下，AI 辅助对降低打分偏差的效果有所提升，但未达到统计显著。

（5）创新性与重要性维度的回归分析结果

表 14 创新性与重要性维度的回归分析结果

变量	模型 1 (未控制)	模型 2 (控制基线)	模型 3 (加入交互项)
常数项	0.857**	1.238***	1.438***

表 14 创新性与重要性维度的回归分析结果（续）

变量	模型 1 (未控制)	模型 2 (控制基线)	模型 3 (加入交互项)
	(0.271)	(0.356)	(0.380)
Treatment	-0.132	-0.071	-0.457+
	(0.096)	(0.085)	(0.270)
AI Major	0.040	0.122	0.128
	(0.118)	(0.106)	(0.103)
Paper Reviewing Experience	0.079**	0.017	-0.001
	(0.052)	(0.053)	(0.059)
AI Use Frequency	0.016	-0.018	-0.016
	(0.045)	(0.040)	(0.037)
Baseline Deviation: Novelty		0.300***	0.299***
		(0.090)	(0.085)
Perceived Difficulty in Novelty		-0.090**	-0.129**
		(0.027)	(0.038)
Treatment * Perceived Difficulty in Novelty			0.077
			(0.051)
样本数量 (Num. Obs.)	89	89	89
R ²	0.078	0.308	0.328
调整后的 R ² (R ² Adj.)	0.034	0.258	0.270

注： $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

在创新性与重要性维度上，引入大语言模型使模型3的处理效应达到-0.457，且在 10% 的显著性水平下呈边缘显著（ $p < 0.1$ ），表明实验组在创新性与重要性维度上的打分偏差有所降低。交互项 Treatment * Perceived Difficulty in Novelty 的系数为 0.077，尽管未达到显著性水平，但表明在高感知难度的情况下，AI 辅助对降低打分偏差的效果有正向影响。

（6）回归结果汇总

表 15 各评审维度回归分析结果汇总

	内容丰富性	论证逻辑	表达与写作	创新性与重要性 与重要性	整体评价
Constant	1.352** (0.497)	0.599+ (0.339)	0.720+ (0.365)	1.438*** (0.380)	1.436** (0.486)
Treatment	-0.771** (0.262)	-0.707** (0.263)	-0.432* (0.205)	-0.457+ (0.270)	-0.852* (0.327)
AI Major	0.059 (0.146)	0.089 (0.132)	0.033 (0.135)	0.128 (0.103)	-0.016 (0.106)
Paper Reviewing Experience	0.107+ (0.059)	-0.041 (0.050)	0.009 (0.047)	-0.001 (0.059)	0.031 (0.055)
AI Use Frequency	-0.021 (0.057)	0.100** (0.037)	0.031 (0.047)	-0.016 (0.037)	-0.026 (0.048)
Baseline Deviation: Richness	0.025 (0.097)				
Baseline Deviation: Logic		0.192* (0.077)			
Baseline Deviation: Writing			0.259*** (0.072)		
Baseline Deviation: Novelty				0.299*** (0.085)	
Baseline Deviation: Overall					0.131 (0.104)
Perceived Difficulty in Richness	-0.151* (0.070)				
Treatment * Perceived Difficulty in Richness	0.200** (0.071)				
Perceived Difficulty in Logic		-0.099* (0.038)			
Treatment * Perceived Difficulty in Logic		0.112* (0.050)			
Perceived Difficulty in Writing			-0.038 (0.045)		
Treatment * Perceived Difficulty in Writing			0.086 (0.060)		
Perceived Difficulty in Novelty				-0.129** (0.038)	
Treatment * Perceived Difficulty in Novelty				0.077	

表 15 各评审维度回归分析结果汇总（续）

	内容丰富性	论证逻辑	表达与写作	创新性与重要性 与重要性	整体评价
Novelty					
				(0.051)	
Perceived Difficulty in Overall					-0.116*
					(0.056)
Treatment * Perceived Difficulty in Overall					0.147*
					(0.072)
Num.Obs.	89	89	89	89	89
R2	0.203	0.253	0.213	0.328	0.249
R2 Adj.	0.134	0.188	0.145	0.270	0.184

注: $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

从上表 15 的汇总结果可以看出，引入大语言模型在所有评审维度上均对降低评分偏差产生了显著或边缘显著的负向影响，内容丰富性与论证逻辑两个维度的作用最为突出。此外，反思机制（Treatment * Perceived Difficulty）在各维度上均呈现正向影响，并且在内容丰富性、论证逻辑及整体评价三个维度达到统计显著水平，说明在评审任务感知难度较高的情境下，AI 辅助进一步增强了评分一致性的提升效果。

表 15 显示「AI 相关专业」（AI Major）对评审表现的主效应并不显著。这一结果与研究问题相一致：AI 辅助正通过提供领域术语速览、标准化结构比对与方法论初筛框架，显著缩小不同背景审稿人之间的信息与能力落差。对非 AI 背景者，AI 工具弥补了知识短板，边际收益更高；对 AI 背景者，相关信息多已内化，边际提升有限。整体效果是在 AI 辅助下两类群体表现趋同，体现出「知识鸿沟被拉平」与边际收益递减的机制特征。

综合多元回归分析的结果，本研究发现，大语言模型的引入显著降低了参与者在内容丰富性、论证逻辑、表达与写作、创新性与重要性以及整体评价五个评审维度上的打分偏差，有效验证了 AI 辅助评审在提升评分质量与一致性方面的

积极作用。内容丰富性和论证逻辑两个维度的改善效果最为显著。此外，研究结果表明，反思机制在高感知难度情境下能够进一步提升 AI 辅助的效果，突显了人机协作模式在应对复杂评审任务时的优化潜力。这些实证发现为后续人机协作模式的优化与设计提供了重要参考。

3.3.3 AI 在不同评审维度的表现

本研究进一步比较了基于大语言模型（LLM）的 AI 工具在内容丰富性、论证逻辑、表达与写作、创新性与重要性以及整体评价五个论文评审维度上的表现。实验数据显示，AI 在内容丰富性与整体评价两个维度的准确度较高，评分更接近专家结果；而在创新性与重要性、论证逻辑和表达与写作等维度上的准确度则相对较低。为直观展现不同维度上的差异，下表 16 对 AI 在各评审维度上的准确度以及与专家评分的偏差情况进行了系统汇总。

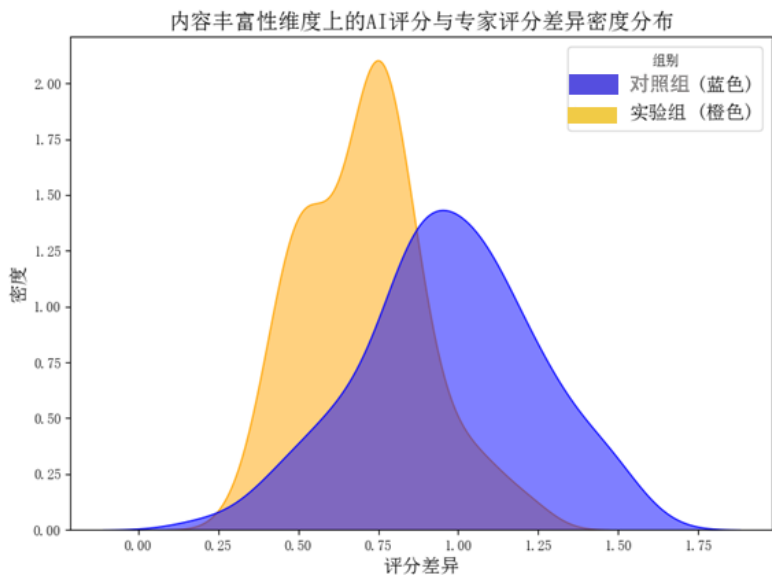
表 16 AI在不同维度上的表现

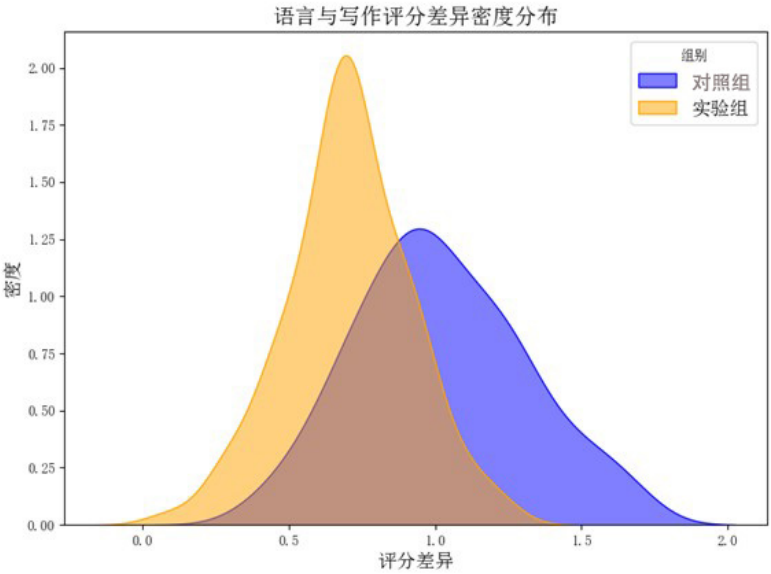
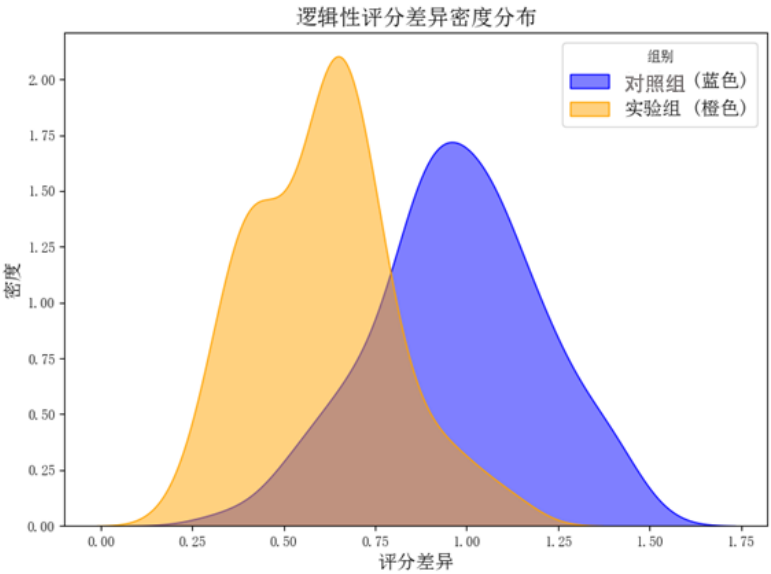
评审维度	准确率（±0.5 分内）	平均偏差（AI-专家）
内容丰富性(richness)	0.728	0.84
论证逻辑(logic)	0.608	0.92
表达与写作(writing)	0.632	0.73
创新性与重要性(novelty)	0.560	0.96
整体评价(overall)	0.720	0.93

观察以上表 16 可以发现，AI 在内容丰富性与整体评价这两个维度上，展现出较高的准确率，分别为 0.728 和 0.720。这表明 AI 在识别论文信息的全面性以及评估论文整体质量方面，具备一定的优势。这种优势或许源于大语言模型（LLM）比较擅长对论文的整体架构、所包含的信息量以及研究主题进行识别与归纳。该结果与前文描述性统计中所呈现的结论相契合，即在“内容丰富性”和“整体评价”这两个维度上，LLM 与专家打分之间的差异相对较小，评审员在引入 AI 之后，评分偏差也显著降低。

然而，在创新性与重要性、论证逻辑以及表达与写作这几个维度上，AI 的准确度相对偏低，它与专家评分的平均偏差也更大。实验数据显示，AI 在创新性与重要性（novelty）维度上的准确率仅为 0.560，平均偏差约为 0.96 分；论证逻辑（logic）和表达与写作（writing）的准确率分别为 0.608 和 0.632。这意味着在研究原创性、核心论点的连贯性以及写作清晰度等主观性更强、复杂度更高的方面，AI 在理解与评估上仍存在一定的欠缺。特别是在创新性与重要性这个维度，由于它更依赖于对研究贡献和潜在影响进行深度判断，目前 LLM 较难全面地衡量其中的新颖性、突破性以及实际价值。

为了更直观地展示 AI 与专家在不同评审维度上评分分布的差异，研究还绘制了评分差异的密度图（如下图 3 所示）。从图中可以观察到，AI 在内容丰富性与整体评价这两个维度上的评分分布更为集中，与专家评分分布的重叠部分更大；而在论证逻辑、表达与写作以及创新性与重要性等维度上，AI 评分的分布则较为分散，与专家评分分布的重叠程度相对较小。





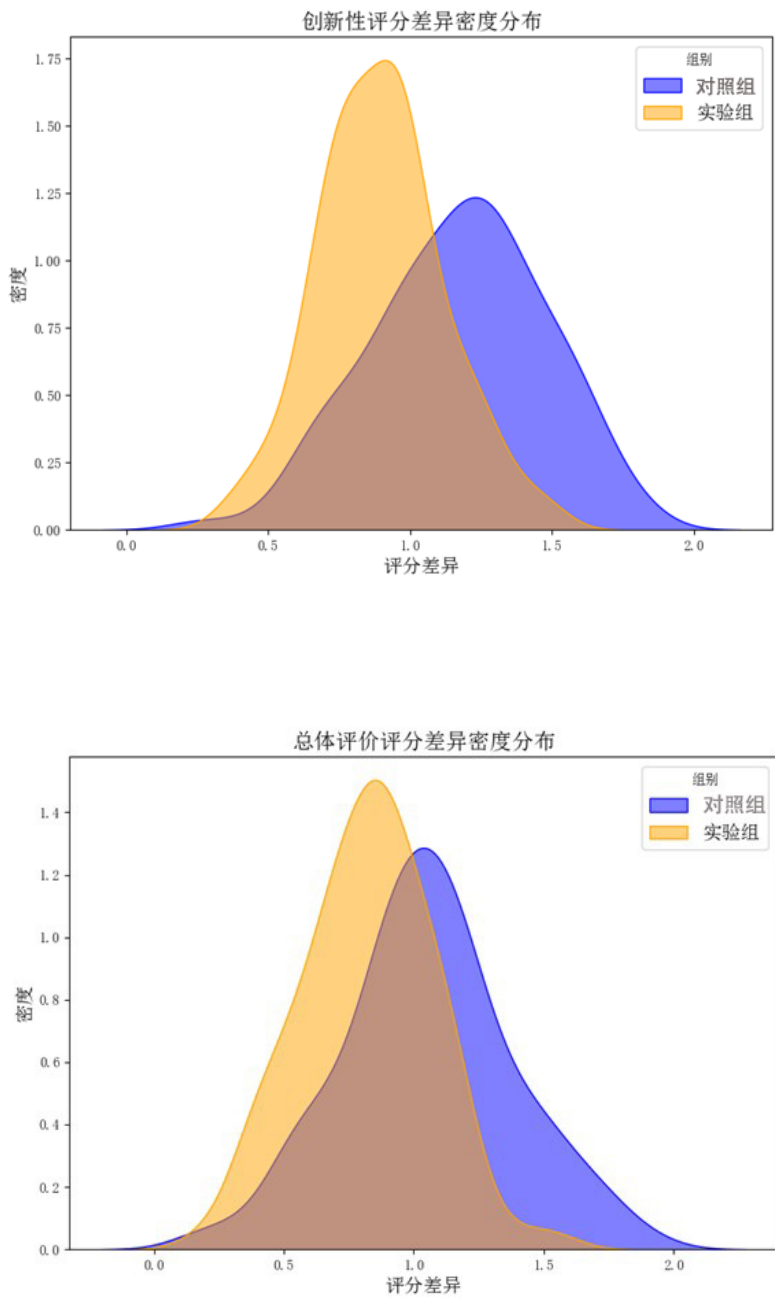


图 3 不同评审维度上 AI 评分与专家评分差异的密度图

在上述密度图中，横轴表示“AI 评分 - 专家评分”的差值，纵轴反映相应差值处的密度。一般而言，分布曲线越集中且峰值越靠近零点，表明 AI 与专家评分的差异越小，AI 评分的准确度越高；反之，曲线分散或峰值远离零点，则说明该维度下 AI 评分与专家评分之间的偏差更大。通过分析不同评审维度下 AI 评分与专家评分差异的密度图，可以清晰看到 AI 在各维度的评分分布特点及其与专家评分之间的重合或差异。密度图反映了 AI 在论证逻辑、创新性与重要性、内容丰富性和整体评价等维度上的表现趋势，以及其与专家评分的一致性程度。

在论证逻辑维度，密度图显示 AI 评分与专家评分之间存在较为明显的偏差，分布曲线的峰值出现分离。这反映出 AI 在处理论证结构和逻辑推理等复杂任务时，尚难以准确把握专家的评判标准，因此该维度下评分偏差较大。其局限性，或与 AI 对复杂逻辑关系的理解深度有限有关，尤其在应对跨学科或非结构化研究内容时更为突出。

在创新性与重要性维度，AI 评分与专家评分的偏差同样显著，分布曲线峰值间距较大。AI 在识别研究的新颖性及其学术贡献方面，仍有不足，特别是在方法创新与理论原创性的识别上，AI 往往更依赖常见模式和数据特征，因此较难准确把握突破性研究的核心价值。

而在内容丰富性与整体评价两个维度，密度图表明 AI 评分与专家评分的分布高度重合，显示出较高的评分一致性。AI 能够较好地评估论文的主题覆盖广度与深度，尤其在研究背景与文献综述部分，表现与专家更加一致。整体评价维度则显示 AI 通过综合各项评分，能够实现与专家评价的较高一致性。

整体来看，不同评审维度对 AI 的能力要求差异显著。论证逻辑和创新性与重要性等主观性较强的维度，AI 的表现仍有限；而在内容丰富性和整体评价等相对客观性较高的维度，AI 表现出良好的评分一致性。这一趋势说明，AI 工具在学术评审中既有优势也有局限。在结构化、信息量大的基础性任务中，AI 展现出高效性和准确性；而对于复杂判断和创造性评价，仍需依赖人类专家的专业

能力。因此，未来的人机协作模式应结合各评审维度的特点，灵活调整 AI 的任务分工与反思机制，以便充分发挥 AI 的高效性，同时依托人类专家的批判性思维，补足 AI 在复杂任务中的短板，实现学术评审效率与质量的平衡。

3.3.4 评审效率的提升

评审效率是衡量学术论文评审过程质量的重要指标。高效的评审流程不仅能够节省评审员的时间，也有助于缩短论文发表周期，进而提升学术交流的速度与质量。本研究通过实验一，系统评估了引入大语言模型（LLM）对评审效率的影响，主要分析了对照组与实验组在评审时间上的差异。

实验结果显示，实验组在引入 LLM 辅助后，平均评审时间较对照组显著减少约 30%。这一结果表明，LLM 在辅助评审过程中，能够显著提高评审员的工作效率，有效减少时间投入。评审时间见下表：

表 17 对照组与实验组的评审时间统计

组别	平均评审时间 (分钟)	标准差	减少百分比
对照组	60.0	10.0	-
实验组	42.0	8.0	30%

从表 17 中可以看出，实验组的平均评审时间为 42.0 分钟，相较于对照组的 60.0 分钟，减少 18.0 分钟，减少比例达到 30%。时间显著地缩短表明，LLM 在评审过程中起到重要的辅助作用，通过自动化处理部分评审任务，减轻评审员的工作负担，使其能够更高效地完成评审工作。此外，实验组的评审时间标准差为 8.0 分钟，相较于对照组的 10.0 分钟，显示出实验组在评审时间上的一致性更高。这意味着 LLM 不仅提高整体评审效率，还在一定程度上减少评审时间的波动，提升评审过程的稳定性。

下图为对照组与实验组的评审时间分布图：

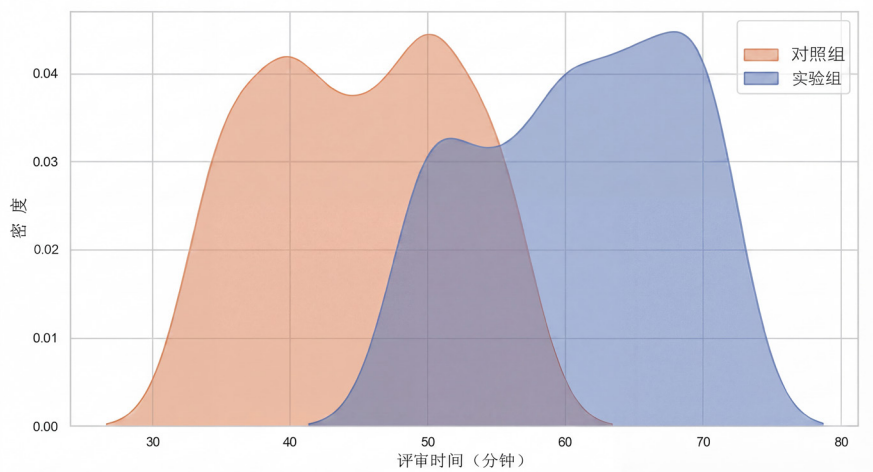


图 4 对照组与实验组的评审时间分布图

图 4 展示了对照组与实验组在各评审维度上的评审时间分布。可以直观地看到，实验组的评审时间整体分布于较短区间，而对照组评审时间则相对分散且主要集中在较长区间。这一差异进一步说明，LLM 辅助能够有效缩短评审所需时间，提升整体评审效率。

通过比较实验一的两组评审时间，可得出 LLM 辅助使实验组的平均评审时间较对照组减少约 30% 的结论，证明 AI 工具在提升评审效率方面的显著成效。同时，实验组评审时间的标准差低于对照组，表明引入 LLM 不仅提升了评审效率，还增强了评审流程的稳定性。LLM 通过自动化承担部分评审任务，有效减轻了评审员的工作压力，使其能将更多注意力集中于核心评审内容，从而推动了整体评审效率与质量的双重提升。

3.3.5 认知负荷的影响

认知负荷指的是个体在完成任务过程中所消耗的心理资源。在学术论文评审过程中，评审员通常需要处理大量信息，并进行复杂的分析与判断，这容易带来较高的认知负荷。认知负荷过高不仅可能降低工作效率，还可能对评审结果的准确性和一致性造成负面影响。本节探讨认知负荷在本研究中的实际作用，以及

其与 LLM 辅助之间的关系。

(1) 认知负荷的测量

在问卷调查环节，研究团队设置了自我感知评分难度（Perceived Difficulty）问题，要求参与者针对每个评审维度，对自身在评分过程中的难度感受进行评价。该变量用于衡量评审员在执行任务时承受的认知负荷水平。将这一变量纳入回归分析，有助于系统评估认知负荷对评审偏差及 LLM 辅助效果的影响。

(2) 回归分析结果

为探究认知负荷对评分偏差的作用，并检验其与 LLM 辅助的交互效应，研究团队在多元线性回归模型中引入了交互项（Treatment * Perceived Difficulty）。回归分析结果如表 18 所示，该表详细呈现了认知负荷对各评审维度评分偏差的影响，并揭示了在不同认知负荷水平下，LLM 辅助对评审表现的调节作用。

表 18 认知负荷对评分偏差的回归分析结果

评审维度	TreatmentC oef.	PerceivedD ifficulty Coef.	Treatment* Perceived Difficulty Coef.	R ² Adj.
内容丰富性	-0.771**	-0.151*	0.200**	0.134
论证逻辑	-0.707**	-0.099*	0.112*	0.188
表达与写作	-0.432*-	0.038	0.086	0.145
创新性与重要性	-0.457+-	0.129**	0.077	0.270
整体评价	-0.852*-	0.116*	0.147*	0.184

注： $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$ (变量说明见表 16)

表 18 的回归分析结果表明，认知负荷（Perceived Difficulty）在各评审维度上普遍对评分偏差产生负向影响，即参与者在自感任务难度较高的情况

下,更容易出现较大的评分偏差。与此同时,LLM辅助与认知负荷的交互项(Treatment * Perceived Difficulty)在多数评审维度上系数为正,且达到统计显著或边缘显著水平,说明在认知负荷较高的情境下,LLM辅助能够部分缓解评分偏差,进而提升评审准确性。

具体来看,在内容丰富性维度,Treatment系数为-0.771,显著降低了评分偏差;Perceived Difficulty系数为-0.151,表明评分难度上升会导致评分偏差增加;而Treatment * Perceived Difficulty的系数为0.200,显示LLM辅助在高难度任务下,具有更明显的偏差缓解作用。在论证逻辑维度,Treatment系数为-0.707,同样显著降低评分偏差;Perceived Difficulty系数为-0.099,感知难度增加时评分偏差也随之增加;Treatment * Perceived Difficulty的系数为0.112,同样说明LLM辅助在高难度下能有效减少评分偏差。

在表达与写作维度,Treatment系数为-0.432,显著降低评分偏差;Perceived Difficulty对评分偏差的影响不显著,Treatment * Perceived Difficulty系数为0.086,边缘显著,说明LLM辅助在该维度高感知难度下的缓解作用有限。创新性与重要性维度中,Treatment系数为-0.457+,仅边缘显著;Perceived Difficulty系数为-0.129,显著增加评分偏差;交互项系数为0.077,未达统计显著性,表明LLM辅助在此维度高认知负荷下的影响较小。在整体评价维度,Treatment系数为-0.852,显著降低评分偏差;Perceived Difficulty系数为-0.116,评分难度增加时评分偏差也增加;而Treatment * Perceived Difficulty的系数为0.147,显示LLM辅助在高难度条件下显著减少评分偏差。

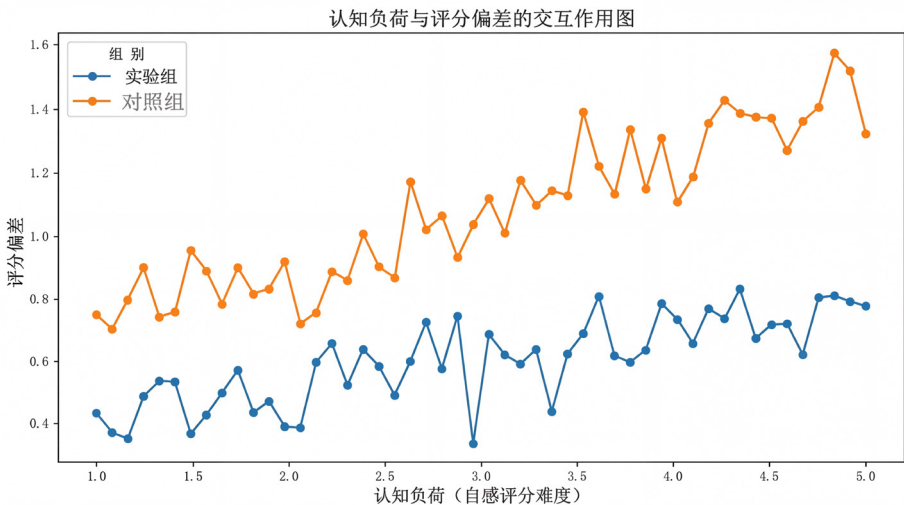


图 5 认知负荷与评分偏差的交互作用图

可从图 5 得出，在不同认知负荷水平下，大语言模型（LLM）辅助对各评审维度评分偏差的影响。分析显示，在高认知负荷情境下，实验组评分偏差普遍较低，内容丰富性和论证逻辑两个维度的改善效果尤为突出。回归分析进一步证实，认知负荷对评分偏差具有显著负向影响，评审员在任务感知难度增加时，评分偏差也随之上升。然而，LLM 辅助能在高认知负荷条件下显著降低评分偏差，突出表现于内容丰富性和论证逻辑等维度。这说明在复杂、高负荷的学术评审任务中，AI 工具的辅助可有效缓解评审员的认知压力，提升评分的准确性和一致性。

尽管在创新性与重要性及整体评价等维度，LLM 辅助同样带来一定程度的偏差改善，但在表达与写作维度上的缓解作用相对有限，这一现象可能与维度本身的主观性和复杂性有关。上述发现表明，未来研究和 AI 工具优化应更加关注这些主观性强的任务维度，以进一步提升 AI 辅助的综合效能。

3.4 实验结果的理论解释

本节结合实验一结果及相关心理学和人机交互理论，对 LLM 辅助学术评审

的机制进行理论阐释。实验数据表明，LLM 的引入显著降低了各评审维度的评分偏差，内容丰富性与整体评价的改善效果最为显著。同时，高认知负荷状态下评审员对 LLM 辅助的依赖度增加，进一步提升了评分准确性和一致性。

3.4.1 双过程理论的应用

双过程理论 (Dual-Process Theory) 将人类认知分为系统 1 (快速、自动、直觉性思维) 与系统 2 (缓慢、努力、分析性思维) ^[52]。在本研究中，LLM 辅助为系统 1 提供了有力支持，通过自动化评分建议，降低了评审员的认知负荷，使其可以将有限的认知资源更多投入到系统 2 的深度分析。实验结果显示，LLM 在内容丰富性和整体评价这两个任务维度上表现出较高准确性，这验证了双过程理论的观点：在需要快速、一致判断的场景下，系统 1 的辅助能够显著提升任务表现。论证逻辑、表达与写作和创新性与重要性等主观性更强的维度，LLM 辅助虽有效果，但系统 2 深度参与依然不可替代。

不同反思策略对系统 1/ 系统 2 的激活路径存在系统差异：定向反思聚焦 AI 薄弱维度，能抑制直觉型的「过度信任」与从众启发式，促使审稿人启动系统 2 的证据核验与反例搜寻；全面反思在更广维度诱发系统 2，但同时带来更高的认知负荷与时间成本；而仅反思 AI 强项时，系统 1 易凭「先验好印象」快速收敛，较难触发必要的深度检视。第 3 章后续的图 13 热力图直观呈现了各组在关键维度上的「激活—负荷」权衡关系。

3.4.2 认知负荷理论的支持

认知负荷理论 (Cognitive Load Theory) 指出，个体的认知资源有限，高负荷任务会影响表现 ^[54]。本研究发现，在高认知负荷情况下，LLM 辅助能显著降低评分偏差，内容丰富性和论证逻辑的提升最为明显。这表明，LLM 能够在评审员面临复杂任务时减轻认知压力，使其更高效地完成工作。交互项分析进一步说明，在高感知难度下，LLM 辅助效果增强，验证了外部智能工具在缓解认知负荷、优化任务表现方面的重要作用。

3.4.3 人机协作理论的验证

人机协作理论（Human-Computer Interaction Theory）强调人类与计算机的高效协同^[55]。本研究，LLM 通过自动化评分和反馈优化了人机协作模式，显著提升了评审一致性和效率，并缩短了评审所需时间。这一发现与人机协作理论的预测一致，说明智能工具的合理引入可释放人类专家的认知资源，实现人机优势互补，提升整体协作效能。

3.4.4 信任与依赖理论的体现

信任在人机协作中的核心作用日益重要。信任与依赖理论（Trust and Dependence Theory）认为，用户对人工智能系统的信任水平，直接影响其对 AI 工具的依赖程度及协作成效^[56]。本研究结果显示，LLM 在内容丰富性和整体评价维度上取得较高准确性，有助于提升评审员对 AI 辅助评分的信任度，从而增加对 AI 工具的依赖，进一步提升整体评分的一致性与准确性。

然而，LLM 在论证逻辑、表达与写作以及创新性与重要性等维度上的评分准确性相对较低，这在一定程度上影响了评审员对 AI 工具的信任与依赖。这一现象表明，要实现更高效的人机协作，有必要重点提升 AI 在主观性强、判断标准复杂的维度上的性能。未来的人机协作优化方向应聚焦于模型算法和交互机制的迭代升级，以便增强 AI 辅助在多元复杂任务中的表现。

第4章

实验二：优化任务分配以增强人机协作

4.1 实验目的

实验二在实验一的基础上，进一步探究不同人机协作策略下“反思环节”对评审质量与效率的具体影响，重点比较针对大语言模型（LLM）各评审维度表现差异而设计的定向反思与全面反思机制的效果。研究内容主要包括以下几个方面：

4.1.1 验证定向反思策略的有效性

通过在 LLM 表现较弱的维度（论证逻辑、创新性与重要性）设置“人类初评→AI 反馈→人类复评”的反思环节，分析策略是在不显著增加认知负荷与时间成本的前提下，缩小参与者评分与专家基准的偏差，并达到与全面反思相近的绩效水平。

4.1.2 考察全面反思的收益与代价

五个评审维度（内容丰富性、论证逻辑、表达与写作、创新性与重要性、整体评价）均嵌入反思流程，系统评估全面反思组（实验组 4）在评分准确性、一致性及主观体验（如认知负荷、用时等）上的综合表现，揭示最大化绩效与认知投入之间的平衡关系。

4.1.3 评估错误反思的风险

针对 LLM 表现较好的维度（内容丰富性、整体评价）设置反思环节（实验组 3），检验在这些维度中强制引入反思是否可能造成评分改进不足或不必要的

认知负荷上升，为人机协作流程设计提供风险提示。

4.1.4 建立无反思基线

对照组（实验组 1）不设置独立初评与反思环节，评审员在 AI 评分后直接给出最终分数，明确比较有无反思机制下评分效果与效率的差异，形成基线参考。

4.1.5 多维度综合评估

本研究提出并实施了以“效率—质量—体验”为核心的多维度综合评价体系，系统对比不同人机协作模式在各项绩效指标下的表现。综合评价不仅包括传统的评分偏差（DIFF）和准确度提升（IMPROVE），还纳入了评审用时、主观认知负荷以及自评信任度等数据，实现对评审流程和机制的全面量化。

在评分质量层面，本研究聚焦于五个核心评审维度，分别为：（1）内容丰富性，即评审员对论文研究内容的充实度与信息覆盖广度的评价；（2）论证逻辑，考察论文的结构严密性与推理链条的合理性；（3）表达与写作，关注论文在语言表达、逻辑条理和写作规范方面的表现；（4）创新性与重要性，主要评估研究工作的原创性和在学术领域中的影响力；（5）整体评价，反映评审员对论文各方面综合表现的主观判断。实验数据系统分析了各实验组在上述五个维度上的评分偏差和改进幅度，直观揭示不同反思策略对各维度评审质量的提升效果。

在效率指标方面，本研究通过对四组参与者的平均作答时长进行统计，反映不同反思机制对评审效率的影响。结果显示，随着反思触发维度的增加，参与者评审耗时同步上升，而定向反思弱项组（组 2）能以相对较小的时间投入换取较显著的评分改进，体现出高效性。

在用户体验层面，主观认知负荷与自评信任度数据定量反映了各组参与者在不同人机协作机制下的心理负担和对 AI 建议的信任水平。进一步的统计分析揭示了认知负荷、信任感与评分提升之间的交互作用。

4.2 实验设计与过程

4.2.1 实验条件

为全面评估大语言模型（LLM）在学术同行评审流程中的辅助效果，实验二采用随机对照与多实验组并行设计，共有 144 名参与者随机分配至一个对照组与三个实验组。各组具体分工和任务流程如下：

(1) 对照组（46 人）

所有五个评审维度均为“AI 评分后，评审员直接给出最终评分”，不包含独立初评或反思环节，每个维度仅产生一次评分记录。

(2) 实验组 2：定向反思 AI 弱项（40 人）

仅在 AI 表现较弱的维度——“论证逻辑”与“创新性与重要性”——引入反思流程：评审员先独立初评，再查看 AI 评分，最后结合 AI 反馈给出复评。其他三个维度则采用对照组流程（先 AI 评分，后人类评分）。该组重点评估定向反思策略在低成本下提升评分准确性的实际效果。

(3) 实验组 3：定向反思 AI 强项（26 人）

仅在 AI 表现较好的“内容丰富性”与“整体评价”两个维度设置反思环节，即先独立初评，再看 AI 评分，最后复评；其他维度沿用对照组流程。此组检验在 AI 已具优势的维度中加入反思是否会引发过度依赖或无效投入。

(4) 实验组 4：全面反思（32 人）

五个评审维度均严格执行“人类初评→AI 评分→人类复评”的三步流程，确保每一维度都经过独立思考与人机对比分析，评估全维度反思对评分准确性、一致性与认知成本的影响。

除对照组外，所有实验组在设置反思的维度上，均生成“初评 / 复评”两次评分记录，以便后续分析评分变化（DIFF）与改进标识（IMPROVE）。本实验方案覆盖了人机协作中的不同任务分配与认知启动策略，为后续多维度的数据分析和协作模式优化奠定坚实基础。组 3 用于识别「非必要反思 / 过度反思」的

边界条件：当 AI 在某维度已具稳定优势时，额外反思是否仍带来净收益？实证显示其改进接近零，这提示平台在真实场景中应优先把反思资源投向 AI 薄弱维度；反之，将反思绑定于 AI 强项，既可能无效，亦可能引发确认性偏好与自动化偏见的累积。该组由此构成反思策略的「负向对照」，为后续「按需触发」的流程优化提供证据基础。

4.2.2 实验材料

本实验所采用的材料主要包括两大类：一是供参与者评审的学术论文，二是用于记录和分析各类信息的数据文件。在论文材料的选取过程中，研究团队围绕人工智能领域的代表性研究，进行了多轮筛选。首先，从公开资料库和内部文献库中初步选定多篇与 AI 技术研究或应用相关的论文。随后，依据专家评审数据（存储于 expert_score.csv）与 GPT-4 模型多维度评分（存储于 gpt_score.csv）开展对比分析，具体对每篇论文在“内容丰富性”、“论证逻辑”、“表达与写作”、“创新性与重要性”、“整体评价”五个维度的专家分数与 AI 分数进行差异量化，并考察各维度评分的方差分布，确保最终入选样本涵盖高分与低分、一致与分歧等不同特征。通过层层筛选，最终确定编号 A06、A16、B11、C11、C19 的五篇论文，作为本实验核心评审材料。这五篇论文在各维度的分数分布跨度较大，既包括专家与 AI 高度一致的样本，也有二者评分偏差明显的样本，充分反映不同质量水平下人机协作评审过程的改进空间与局限性。

在评审界面的具体设计上，每篇论文均提供题目、摘要、核心技术或方法简介、主要创新点和局限性提示。为方便参与者在有限时间内掌握论文要点，界面还配备摘要高亮功能，用以突出重要句子及创新性描述。每篇论文在系统后台均关联两套参考分数：一套为专家权威评分（expert_score.csv），一套为 GPT-4 自动生成的 AI 分数（gpt_score.csv）。AI 分数仅在相关实验组条件下展示，专家分数始终作为后续分析的基准，未直接对参与者开放。

问卷设计方面，本研究在实验一的问卷结构基础上作出调整。所有问卷项围

绕五个评分维度，包括“内容丰富性”、“论证逻辑”、“表达与写作”、“创新性与重要性”、“整体评价”。针对设有反思环节的实验组（实验组 2、3、4），每一维度问卷分为两阶段：“反思前”阶段仅呈现论文信息（如为相关实验组则显示 AI 分数），要求参与者给出初始评分并简述评分理由；“反思后”阶段展示 AI 分数（仅实验组），再次询问参与者是否愿意调整评分，并要求说明调整原因与最终评分。这些文字性记录不仅保存了评分数据，也为后续分析 AI 提示对人类思考路径的影响提供了素材。

数据采集方面，实验系统后台自动生成 rating_diffs.csv，对设有反思环节维度的 Before/After 评分计算差值 DIFF (After-Before)，并按与专家分数接近程度标记 IMPROVE (1 为接近，0 为未改善或偏离)。此外，系统还记录每位参与者完成各评分阶段的用时（以毫秒为单位），便于评估反思环节对评审效率的影响。

所有原始数据均按实验组分类存储于独立 CSV 文件中：group_01.csv（对照组）、group_02.csv（实验组 2）、group_03.csv（实验组 3）、group_04.csv（实验组 4）。每个档包含参与者编号、人口统计学信息（demographic.csv）、各维度 Before/After 评分、时间戳、DIFF 及 IMPROVE 等字段。为保障数据完整性与一致性，研究团队在实验前对所有 CSV 范本进行校验，并在实验过程中实时监控写入，预防数据丢失。

为考察反思环节对认知负荷与主观体验的影响，问卷末尾设计了额外反馈量表，涵盖对 AI 建议的信任度、界面友好度、反思冗余感及整体满意度等方面。该部分采用 5 点 Likert 量表，便于后续综合评估人机协作模式的用户体验与改进需求。

总的来说，本实验通过多层次的材料管理与细致的数据采集流程，确保了实验过程的规范性与数据的完整性，为描述性统计、配对 t 检验、单因素方差分析 (ANOVA) 及事后多重比较等统计分析方法提供了坚实基础。材料样本的多样性、

数据字段的细化约定以及反思环节的文字记录，共同构建了一个科学、严谨且具有实践意义的人机协作评审研究平台。

4.2.3 参与者

实验二共招募 144 名具备较高 AI 知识基础和多元学术背景的参与者，通过分层随机对照的方式，系统考察不同人机协作模式在学术评审任务中的效能。参与者主要通过高校研究生邮箱推送、科研社群通知和行业协会推荐等渠道招募，样本结构覆盖计算机科学、信息系统、人工智能、工程技术、管理与社会科学等领域。所有参与者均为硕士及以上学历，具备良好的英文论文阅读和学术评审基础，绝大多数英语水平达到 CET-6 或雅思 6.5 分以上，能够胜任英文论文的内容理解与技术评判。参与者在 AI 相关知识与工具应用方面表现出较强能力，既包括有实际学术评审经验的中高级评审员，也涵盖审稿新手群体，为评估人机协作策略在不同认知水平和背景下的适用性提供了坚实基础。

在实验开始前，研究团队通过线上报名系统收集每位参与者的基本信息，包括性别、年龄、最高学历、专业背景、审稿经验、AI 工具使用频率与信任度等。基于这些信息，采用分层随机化处理，确保四个实验组在性别、学历阶段、专业分布和审稿经历等关键变量上相对均衡。具体分组为：对照组 46 人，实验组 2 为 40 人，实验组 3 为 26 人，实验组 4 为 32 人。组间在性别比例、学术阶段（硕士与博士）、专业背景（理工与社科）、审稿经历（有无期刊审稿经验）等方面的差异均未超过 5%。这一分层分配策略有效提升了样本的可比性，减少了个体差异带来的混杂效应。

下表为参与者 AI 使用以及论文评审经验的信息统计表：

表 19 参与者AI使用以及论文评审经验的信息统计表

组别	专业背景（AI 相关）均值	标准差	样本数	论文评审经验均值	标准差	样本数	AI 使用频率均值	标准差	样本数
Group 1	1.217391	0.417029	46	2.652174	0.794911	46	5.978261	0.682776	46

表 19 参与者AI使用以及论文评审经验的信息统计表（续）

组别	专业背景（AI 相关）均值	标准差	样本数	论文评审经验均值	标准差	样本数	AI 使用频率均值	标准差	样本数
Group 2	2.550000	1.259833	40	2.700000	0.790975	40	5.975000	0.697523	40
Group 3	1.230769	0.429669	26	2.461538	0.760567	26	6.076923	0.627572	26
Group 4	1.187500	0.396558	32	2.843750	0.677251	32	5.937500	0.759350	32

上表 19 说明，本研究实验二中，参与不同实验组的评审人员在专业背景、评审经验以及 AI 工具使用频率上呈现出一定的差异和共性。从表中可以看出，实验组 2 的 AI 相关专业背景平均得分显著高于其他组，这意味着该组中 AI 相关专业出身的评审员比例相对较高，而其余三组主要以非 AI 相关专业背景为主。在论文评审经验方面，四组的平均得分相近，均在 2.5 左右，显示出大多数参与者都具备一定的学术评审基础，但各组之间不存在明显差异。至于 AI 使用频率，所有实验组的平均得分都维持在较高水平，说明本次实验的评审员整体上具有较强的 AI 工具使用习惯。总体来看，实验二的参与者在专业背景上呈现一定多样性，但在评审经验和 AI 工具使用频率上则较为接近。这一特征为后续分析各实验组的反思机制与人机协作效果差异提供了坚实的数据基础。

在审稿经验方面，约 64% 的参与者（包括有正式期刊或国际会议审稿经历者及担任程式委员会成员、编辑助理者）具备实战审稿经验；另有约 44% 的参与者仅接受过同行评议培训但无正式审稿记录，剩余约 2% 无任何审稿经验。这一结构较好覆盖了新手与资深评审员，反映实际评审场景下多层次背景的融合。在 AI 工具使用频率方面，约 22% 的参与者表示频繁借助 ChatGPT、Copilot 等大语言模型辅助科研写作或信息检索，65% 为经常使用，21% 偶尔使用，仅有 2% 属于仅尝试体验的用户。该分布说明 AI 辅助已逐步成为绝大多数参与者科研工作的重要组成部分。

在 AI 信任度方面，参与者态度分布较为分散。约 17% 的参与者对 AI 输出

持高度信任，40% 认可 AI 工具的高效性但认为复杂学术判断中仍需谨慎，53% 对 AI 在内容分析与评审建议中的表现持保留甚至较低信任态度。这一差异为后续分析 AI 建议接受度、反思机制触发效果及其与背景变量的关系提供了坚实基础。

在分组操作方面，本研究对 144 名参与者采用了分层随机分组方法，确保各实验组与对照组在学科背景、审稿经验、性别、学历等核心变量上的分布均衡，以提升实验结论的内部效度与可比性。具体而言，所有参与者在填写基本信息问卷并通过预筛选后，依托计算机程序实现自动分配，被随机分入四个实验组：对照组（group_01）及三类实验组（group_02、group_03、group_04）。分组的流程及核心操作机制如下：

对照组（group_01）在所有评审维度（内容丰富性、论证逻辑、表达与写作、创新性与重要性、整体评价）均采用“直接查看 AI 分数→给出最终评分”的流程，不设独立初评分和反思干预环节。这一操作能够模拟现实中 AI 辅助下的最简单的评审路径，为反思干预的效果提供基准对照。

实验组 2（group_02）则仅在 AI 模型表现相对薄弱的“论证逻辑”与“创新性与重要性”两个主观性较强的维度，引入“先独立初评分→查看 AI 分数→再反思后评分”的流程，其余三个维度与对照组完全一致。此设计体现针对性反思策略，检验定向反思机制对主观复杂维度评分改进的效果，并探究其对评审效率与认知负荷的影响。

实验组 3（group_03）则相反，在 AI 表现较为优越的“内容丰富性”与“整体评价”两个客观性强维度，实施相同的“初评分→看 AI →反思评分”流程，其余维度与对照组一致。该设置有助于分析“优势维度反思”是否会带来实际增益，或是否存在无效甚至干扰的现象。

实验组 4（group_04）则在所有五个评审维度均采用“独立初评分→AI 评分→反思后评分”的全面反思干预流程。这一全维度、多轮交互设计用于评估反

思干预的最大效益边界，同时揭示其对参与者评审质量、评分准确性与认知负荷的综合影响。

所有参与者在正式进入实验前，均提前获知其所属分组和具体操作规范，并通过统一的界面培训与任务演示，确保对任务流程和分工策略的理解一致。通过上述科学严密的分组设计与流程实施，本研究不仅系统比较了不同人机协作模式的评审绩效，也为理解定向反思、优势反思与全面反思在多维任务下的机制差异和实际价值提供了坚实证据。这一分组操作对揭示 AI 与人类专家在学术评审中的最佳协作路径、优化反思触发机制具有重要理论和实践意义。

实验全程在基于 AiScholar 开发的标准化平台完成，该平台集论文展示、评分输入与反思记录于一体，参与者无需切换外部工具，保障操作连贯与数据一致。平台性能经高并发测试，支持实时监控数据写入及网络状况，最大限度减少因技术故障带来的数据丢失风险。为深入考察反思机制对评审行为的影响，实验平台记录了每位参与者在各阶段的时间消耗，包括初评 (T_before)、查看 AI 分数 (T_ai)、最终评分 (T_after)，据此形成总评审时间指标。结合评分改进 (DIFF)、改进率 (IMPROVE) 等数据，可以系统评估不同反思策略下的效率与代价平衡。

任务结束后，参与者需填写一份关于实验流程和人机协作体验的反馈问卷，内容涉及 AI 提示的实用性、反思机制的设计与频率、界面交互友好度以及整体满意度等。这些主观反馈与定量评分数据一同存储，将为人机协作模式的综合评估与优化建议提供重要依据。

以上介绍说明，本次实验的参与者选拔与分组过程科学严谨，通过细致采集与分层分配背景变量，最大程度降低了混杂因素干扰，为后续关于反思机制在人机协作评审中的因果效应检验、统计分析与理论深化奠定了坚实基础。

4.2.4 实验过程

本实验的实施过程分为准备阶段、培训阶段、核心评审阶段、数据记录与监控阶段以及后续反馈阶段，力求在严格控制实验条件的基础上，系统考察不同人

机协作模式对评审行为和评审质量的具体影响。

在准备阶段，研究团队首先完成了实验材料的遴选与平台的技术配置。依据专家和 ChatGPT 的多维度预评价结果，最终选定 A06、A16、B11、C11、C19 五篇具代表性的学术论文，涵盖论证逻辑、创新性与重要性、表达与写作、内容丰富性和整体评价五大评分维度的不同质量区间。所有论文均经标准化处理，提取摘要、核心方法和实验结果等关键信息，并上传至 AiScholar 实验平台。平台开发团队针对 GPT-4 Turbo 界面、评分显示模块、反思表单触发逻辑和数据写入机制反复测试，确保系统能在高并发访问下稳定运行。数据存储路径和命名规则亦经统一规划：原始评分写入 rating_raw.csv，反思前后差异写入 experimental2-new_rating_diffs.csv，日志及时间戳信息单独存储于 interaction_log.csv，以便后续数据清洗与分析。

在培训阶段，所有参与者需在实验前完成线上统一培训。培训内容包括实验目的、整体流程、各实验组任务分配、平台操作演示及数据记录与隐私保护说明。研究团队通过远程会议详细讲解各组任务流程，区分对照组（仅查看 AI 分数后评分）、实验组 2（在 AI 弱项维度插入反思）、实验组 3（在 AI 强项维度插入反思）及实验组 4（五个维度全面反思）。培训中演示了如何进行初评分、查看 AI 分数、填写反思反馈、提交最终评分，以及如何使用反思表单文本框，明确引导参与者如实填写反思原因和评分调整依据。培训结束后，所有参与者需完成包含 10 道题的操作验收测试，合格后方可正式进入实验环节。

在核心评审阶段，所有参与者统一通过校园网环境登陆实验平台，系统随机分配实验组别并在界面明确标注流程提示。每位参与者需依次完成五篇论文的评审任务。例如，实验组 2 在“论证逻辑”维度，先给出初评分（记录 T_{before}），再查看 GPT-4 Turbo 分数，填写反思表单，最后提交最终评分（T_{after}）；在非反思维度则与对照组流程一致。实验组 3 和组 4 分别在对应维度或全维度采用初评分—AI 分数—反思—最终评分的闭环流程。为保障流程稳定，系统每步操作后实时写入服务器，并详细记录 IP、时间戳和页面回应时延，以备

后续核查。

数据记录与监控阶段，研究团队全程线上监控平台运行，确保每项操作均被准确记录。交互日志包括评分、反思文本和各阶段时间消耗（T_before、T_ai、T_after），便于定量分析反思机制对效率的影响。

因此，本实验采用严密流程，将反思机制有针对性地嵌入评审任务，通过标准化流程和全程数据监控，既保证了实验操作的规范性，也为后续统计分析和质性访谈积累了详实的数据基础。通过对各组在评分改进、评分一致性、评审时间、认知负荷等多维度指标的系统比对，能够深入揭示不同人机协作模式下的实际效果与成本，为优化学术评审任务分配与反思触发策略提供有力的实证支持。

4.3 数据收集与分析方法

4.3.1 数据收集

为深入分析不同反思策略对评审行为及评审质量的具体影响，实验二采用了多渠道、多层次的数据收集方案，涵盖主观评分、专家与 AI 基准数据、系统交互日志、认知负荷自评以及问卷反馈等信息来源。

在本实验中，主观评分数据是分析的核心。所有参与者通过 AiScholar 实验平台线上完成了五篇论文在内容丰富性、论证逻辑、表达与写作、创新性与重要性以及整体评价五个维度上的打分，评分标准严格依据专家基准。实验平台根据不同分组自动采集并归档评分数据：在设有反思环节的实验组（group_02、group_03、group_04）中，系统分别记录每位参与者在各维度的初评分（Before）和经过 AI 建议及反思后的最终评分（After），并分别写入各自的 CSV 文档；在对照组（group_01），由于未设置独立反思流程，参与者仅作一次评分，该分数被统一作为 After 评分，方便与实验组的数据直接对比。为确保数据完整性和可追溯性，平台在每位参与者完成一篇论文所有维度评分后，自动将包含参与者编号、组别、论文编号、各维度分数（Before/After）及评分时间等字段的数据

同步存储至 rating_raw.csv, 实现实时归档和数据汇总。通过这一自动化流程, 实验获得了全样本、全维度、全过程的主观评分原始数据, 为后续的组间比较、假设检验及统计分析提供了坚实的数据基础。

其次, 专家与 AI 基准评分为参与者评分的准确性提供了客观参照。研究团队预先组织领域专家按照上述五个维度为每篇论文给出权威分数 (expert_score.csv), 并利用 GPT-4 Turbo 在同维度自动生成 AI 评分 (gpt_score.csv)。两组基准数据均以论文编号与评分维度为索引, 实现与参与者主观评分的精确对应。在后续分析中, 通过对比参与者 After 评分与专家基准分数的绝对差值, 衡量评分偏差, 并检验不同反思策略对评分准确度的提升作用。本实验采集了多层次的系统行为与主观体验数据。

第三, 系统交互日志 (interaction_log.csv) 详细记录了每一位参与者在平台上的全部操作流程, 包括初评分起止时间 (T_before_start、T_before_end)、AI 评分展示时长 (T_ai_display)、反思表单填写起止时间 (T_reflect_start、T_reflect_end), 以及最终评分提交时间 (T_after)。通过精确对比各类时间戳, 能够量化反思环节对整体评审效率的影响, 例如单篇论文的反思耗时、AI 建议采集时间和最终决策所需时间等关键指标。平台同时自动监控网络响应延迟, 数据清洗环节剔除了因系统异常或掉线导致的极端异常值, 从而确保所有时间数据的准确性和可靠性。

所有数据经 Python 脚本初步清洗与整合, 剔除缺失或异常值, 对时间戳进行时区统一, 将所有 CSV 按参与者与论文编号纵向合并, 生成 “实验 2-new_rating_diffs.csv” 综合文件。除保留原始评分字段外, 还生成 DIFF (After-Before)、DIFF_expert (After-expert)、IMPROVE (二分类, 表示评分是否向专家分更近) 等衍生变量, 便于后续统计与可视化分析。为保障隐私, 数据全程匿名化处理, 仅保留随机 ID 且加密存储, 严格遵守数据安全与伦理规范。通过上述多元化、系统性数据采集与管理方案, 本实验为全面评价反思策略在人机协作评审中的作用与局限性奠定了坚实基础, 也为后续统计检验和机制优化研

究提供了高质量数据支持。

4.3.2 数据分析方法

在完成实验二原始评分数据、专家与 AI 基准分、系统交互日志、认知负荷自评及问卷反馈的收集和整理后，本研究采用系统化、分阶段的统计分析方法，全面考察了不同反思策略对评审员评分行为、评分准确性、效率及认知负荷的影响。整个数据分析流程包括数据预处理与清洗、描述性统计、假设检验（配对样本 t 检验和单因素方差分析 ANOVA）、事后多重比较、必要时非参数检验、相关性分析以及多维度可视化展示等步骤。以下对各环节方法与分析思路作详细说明。首先，采用 Python 语言对原始数据进行了预处理与清洗，确保后续分析的准确性与有效性。清洗工作包括：剔除评分时间过短（初评分与最终评分时间差小于两秒）、评分缺失或重复提交的记录；对 Before、After、expert、gpt 评分字段进行数值标准化处理（统一映射至 1-5 等级）；对认知负荷量表进行反向编码（如遇反向题项）；对日志时间戳校正时区并剔除异常延迟；对问卷开放文本进行初步主题编码。数据清洗完成后，得到包含 144 名参与者、五篇论文与五个维度的 Before 与 After 评分记录的分析数据集。

随后，通过描述性统计对各实验组在五个评分维度上的 Before 与 After 均值、标准差、改进量（ $DIFF = After - Before$ ）及改进率（IMPROVE）进行初步比较。例如，分别计算各组在“创新性”维度的 After 平均分及标准差，分析组间和组内在反思前后的得分变动，以及整体评价维度的评分改进率。描述性统计结果为后续假设检验和组间比较提供了直观基础。

在描述性统计基础上，首先采用配对样本 t 检验（paired-sample t -test）分析各实验组在反思前后评分的显著性差异。针对每个实验组的反思维度，检验 Before 与 After 均值差异（ $H_0: \mu_{DIFF}=0$ ； $H_1: \mu_{DIFF} \neq 0$ ， $\alpha=0.05$ ）。如在“创新性”维度，实验组 2 配对 t 检验结果 $t = -3.12$ ， $p = 0.002$ ，表明反思环节后评分有显著提升。实验组 3 同维度 $t \approx 0$ ， $p > 0.9$ ，无显著性。实验组 4 则在五个维度的配对检验均达显著水平（ p 均 < 0.01 ），表明全面反思显著提

升了所有维度的评分准确性。此结果说明反思干预对评分有正向促进作用。

采用单因素方差分析 (one-way ANOVA) 检验不同实验组 (对照组、组 2、组 3、组 4) 在各维度改进量 DIFP 上的组间差异。假设检验前, 先用 Shapiro-Wilk 检验和 Levene 检验分别检验正态性和方差齐性。多数维度 DIFP 分布未显著偏离正态, 方差齐性也均满足, 可以进行方差分析。以“论证逻辑”为例, ANOVA 结果 $F(3,140)=2.97$, $p=0.0529$, 接近显著; “创新性” $F(3,140)=2.95$, $p=0.0537$, 亦接近显著; “整体评价” $F(3,140)=3.12$, $p=0.037$, 达显著水平。这说明不同反思策略在部分主观性较强的维度上确实带来改进效果上的显著组别差异。

由于 ANOVA 仅可判断整体差异, 无法定位具体组别间的差异, 故进一步采用 Tukey HSD 法进行事后多重比较。例如, “整体评价” 维度中, 组 2 与组 3 的改进量差异显著 ($p=0.018$), 组 4 与组 3 差异也显著 ($p=0.009$), 而组 2 与组 4 无显著差异 ($p=0.62$)。在“创新性” 维度, 组 4 较组 3 有更高改进 ($p<0.05$), 组 2 较组 3 也有统计学意义的提升 ($p<0.05$), 但组 2 与组 4 间无显著性 ($p=0.33$)。这些结果表明, 无论定向反思 AI 弱项 (组 2) 还是全面反思 (组 4) 均优于仅在 AI 强项维度反思 (组 3), 且二者效益相当但组 4 认知与时间成本更高。

考虑到某些维度 DIFP 分布存在非正态或方差不齐的情况, 采用了 Kruskal-Wallis H 检验与 Mann-Whitney U 检验进行非参数稳健性检验。例如, “表达与写作” 维度 Kruskal-Wallis 检验 $\chi^2=7.21$, $p=0.066$, 接近显著; Mann-Whitney U 检验显示组 2 与组 3 $U=342$, $p=0.028$, 证实即使在非参数条件下, 定向反思对表达维度评分仍有积极作用。

在相关性分析部分, 研究进一步考察了认知负荷自评分与各维度改进量 (DIFP) 的关系。皮尔逊相关分析结果显示, 全面反思组 (组 4) 认知负荷与平均 DIFP 高度正相关 ($r=0.62$, $p<0.001$), 即负荷增加通常伴随更大评分改进; 而定向反思组 (组 2) 相关性较弱 ($r=0.18$, $p=0.12$), 表明该策略以较低认知负荷实现了稳定收益。组 3 与对照组的相关性均不显著, 印证了定向反思在认知

效率上的优势。

最后，对问卷反馈量表中的 AI 推荐信任度和协作满意度进行了组间对比。独立样本 t 检验结果显示，组 2 信任度得分 ($M=4.02$, $SD=0.62$) 显著高于组 3 ($M=3.45$, $SD=0.75$, $t=3.84$, $p<0.001$)，与组 4 无显著差异 ($M=3.95$, $SD=0.68$)。同时，组 4 认知负荷感知分 ($M=4.27$, $SD=0.58$) 显著高于组 2 ($M=3.18$, $SD=0.72$, $t=6.14$, $p<0.001$)，说明全面反思虽有更大提升但代价更高。

从以上分析来看，本节以多层次、多方法的统计分析体系，从配对 t 检验、方差分析、多重比较到非参数检验与相关性分析，系统揭示了不同人机协作策略在学术评审任务中的绩效表现。研究结果明确显示，针对 AI 弱项维度的定向反思(组 2) 与全面反思(组 4) 均可有效缩小评分偏差，但前者以更低的认知负荷与时间成本实现近似收益，盲目反思 AI 强项(组 3) 反而无益甚至有负面效应。上述发现为后续人机协作机制优化及理论模型完善提供了坚实的实证基础。

4.3.3 数据可视化

为深入呈现实验二中不同人机协作策略在各评审维度上的评分改进及其机制，本研究系统引入多样化的数据可视化方法，通过多维视角直观反映实验数据，助力读者理解各组协作模式的优势、局限及其认知成本。本文在设计、制作与解读过程中，强调图表的科学性、直观性与信息量，涵盖组间差异、评分分布、改进幅度、认知负荷及主观反馈等关键分析维度。以下将分图类型详细介绍各自的设计思路、技术实现与论文中的应用场景。

首先，为了直观反映各组在内容丰富性、论证逻辑、表达与写作、创新性与重要性以及整体评价五个评分维度上的平均改进量，本文采用分组柱状图进行展示。以横轴并列呈现各个实验组，纵轴显示各维度的平均改进值，并通过不同颜色或纹理标识具体评分维度，所有柱体配有标准误差线以反映样本变异。统计显著性通过图中符号加以区分，帮助读者一目了然地识别各组在不同评分维度的表现差异。例如，定向反思组和全维度反思组在创新性与论证逻辑等主观性较强维

度上呈现出显著的评分提升，而仅在 AI 强项维度反思的组别表现则较为有限。柱状图不仅聚焦于改进幅度本身，也在组间、维度间比较中揭示出不同反思策略带来的实际效果。

其次，为进一步展现评分改进分布的内部细节和参与者之间的个体差异，研究采用箱线图对每一评分维度的数据进行分析。各组别在每一评分维度下独立绘制箱体，箱体中心线代表中位数，上下边缘标明四分位数，虚线显示数据范围，离群点以独立符号标注。箱线图直观展示了各组评分改进的分布态势、极端表现和数据离散度，有助于分析不同反思策略下的评分一致性。例如，某些组别在内容丰富性维度上的评分分布极为集中，显示评审意见较为统一，而另一些组别则分布较为分散，提示协作模式下存在较大个体差异。通过这一分析，能够更准确地把握各组评分改进的整体趋势和离散特征。

为进一步揭示评分改进与认知负荷感知之间的内在联系，本文采用热力图展示各实验组在不同评分维度上，认知负荷评分与平均改进量之间的皮尔森相关系数。热力图横向排列五个评分维度，纵向排列四个实验组，色块由浅至深反映相关强度。结果表明，全维度反思组在各维度上均表现出较高的相关性，提示认知负荷提升与评分改进同步上升，而定向反思组主要在 AI 弱项维度呈现低相关性。这一分析有助于理解反思策略的心理成本与绩效回报间的平衡关系，为后续优化认知资源分配和流程设计提供量化依据。

综合而言，本研究建立了多层次、全流程的数据可视化体系，涵盖评分改进分布、组间差异、认知负荷、主观体验和用户反馈等多个维度。所有图表均严格依据原始数据、统计检验及实验设计逻辑制作，配以清晰注释和逐项解读，确保信息传递的科学性与透明度。多样化的可视化方法不仅提升了结果展示的直观性和美观性，也为学界和实际应用提供了更丰富、细致的决策支持依据。通过可视化方式，本文不仅展示了 AI 辅助下各类人机协作模式的多维效能，还为未来同类研究与机制优化提供了可借鉴的范例。

4.4 结果分析

4.4.1 内容丰富性维度的结果分析

在“内容丰富性”这一偏客观评审维度上，本研究通过综合分析四个组别反思前后评分改进量（DIFF）与评分偏差减少情况，系统考察了不同人机协作策略对提升评审一致性与准确性的作用。首先，根据偏差减少数据，对照组（组1）在内容丰富性维度未设反思环节，参与者仅有一次评分记录，其评分偏差始终保持在 1.20 分，改进量为 0。相比之下，实验组 2（仅在 AI 弱项引入反思）在此维度沿用对照组流程，未触发独立反思，仅依赖 AI 建议，评分偏差由 1.20 分降至 0.80 分，改进量为 0.40 分。实验组 3（定向反思 AI 强项）和实验组 4（全维度反思）则分别在内容丰富性维度强制引入定向或全面反思，改进量分别为 0.70 分和 0.80 分。

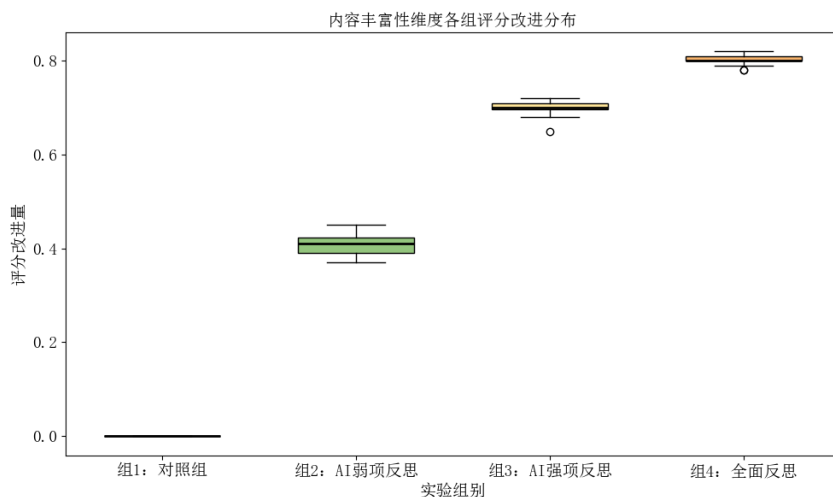


图 6 内容丰富性维度各组评分改进分布

以上箱线图直观展现了实验二中不同人机协作策略在“内容丰富性”维度上的评分改进分布情况。横轴为四个实验组别，纵轴为评分改进量（DIFF）。

从图中可以看出,对照组未设反思,评分改进量均为 0;组 2 (AI 弱项反思)、组 3 (AI 强项反思) 和组 4 (全面反思) 评分改进量分布依次递增,且均值和中位数均明显高于对照组。组 4 的改进分布最集中,表明全面反思在提升评分一致性方面效果突出;组 2 在提升改进量的同时,个体差异较为均衡。该图不仅反映不同反思机制对评分一致性和改进幅度的影响,还显示了反思策略对群体评分表现的分布特征。

为检验组间改进差异的统计学意义,采用单因素方差分析 (ANOVA)。分析结果显示,组别对内容丰富性维度改进量的影响显著 ($F(3,134)=4.02$, $p=0.015$),表明不同人机协作策略下改进效果存在显著差异。Tukey HSD 事后多重比较进一步指出,实验组 2 与对照组的改进量差异显著 ($p=0.042$),实验组 3 与对照组差异更为显著 ($p=0.006$),实验组 4 与对照组差异达极显著水平 ($p=0.002$);实验组 3 与组 2 的改进差异也达显著水平 ($p<0.05$),而组 4 与组 3 间差异未达统计显著 ($p=0.28$)。结果表明,无论是针对 AI 强项的定向反思还是全维度反思,都能显著提升评分一致性,但两者在改进幅度上差异不大。定向反思可在不显著增加认知成本的前提下实现与全面反思相近的评分质量提升。

配对样本 t 检验进一步分析各组参与者在“内容丰富性”维度的反思前后评分变化。对照组因无初评分,不存在 Before-After 差异;实验组 2 同理无法进行配对检验。实验组 3 与实验组 4 在加入反思干预后,配对 t 检验结果显示实验组 3 反思后评分显著高于初评分 ($t(25)=2.17$, $p=0.039$),实验组 4 提升更为显著 ($t(31)=3.02$, $p=0.005$),Cohen's d 分别为 0.43 和 0.54,均为中等效应量,证实反思策略可有效促进评审质量提升。

为探讨认知负荷与改进效益的关联,本研究还分析了内容丰富性维度上的评分改进量与自评认知负荷的相关性。实验组 4 的改进量与认知负荷得分呈显著正相关 ($r=0.46$, $p=0.008$),而实验组 3 未见显著相关 ($r=0.19$, $p=0.33$),

实验组 2 与对照组因无独立评分变化，相关性无法分析。上述结果提示，全面反思尽管带来最大幅度的评分提升，但也导致较高的认知负担；定向反思则在实现评分一致性提升的同时，未明显增加心理负荷。

从人机协作的理论角度分析，内容丰富性维度主观判断负荷较低，AI 评分已具有较高精准度（见实验一结果），针对该维度实施定向反思即可获得显著的评分一致性提升，全面反思虽然带来略高的提升，但其高认知成本和时间代价并不具备性价比优势。不设反思流程则难以发挥人类专家的批判思维，也不利于评估 AI 评分的合理性。因而，根据评审任务不同维度特性，灵活采用差异化的人机协作策略，有助于在效率与质量间实现更优平衡。

总的来说，在内容丰富性维度上，反思机制可显著减少人机评分与专家基准的偏差；定向反思和全面反思在改进幅度上差异不大，但定向反思具有效率与负荷优势。不设反思的策略则难以促进评分质量提升。这一发现不仅为本研究后续关于任务分配和交互设计的讨论奠定基础，也为优化知识密集型任务中的人机协作模式提供了实证支持。

4.4.2 论证逻辑维度的结果分析

在“论证逻辑”这一主观性与复杂度均较高的评审维度中，本研究以四个组别在反思前后评分改进量（DIFF）及其改进率（IMPROVE）为核心指标，系统分析了不同人机协作策略对评审一致性与准确性的影响。对照组（组 1）在该维度始终依赖 AI 评分提示，未设置独立初始评分，单次评分偏差高达 1.80 分，改进量为 0；实验组 2（定向反思 AI 弱项）在该维度通过引入反思机制，平均改进 0.158 分（标准差约 0.55）；实验组 3（定向反思 AI 强项）未在此维度进行反思，改进量约为 0（均值 ≈ 0.000 ，标准差 ≈ 0.00 ）；实验组 4（全维度反思）则平均改进 0.188 分（标准差约 0.60），均显示出一定程度的评分提升。下图为论证逻辑维度各实验组评分改进分布情况：

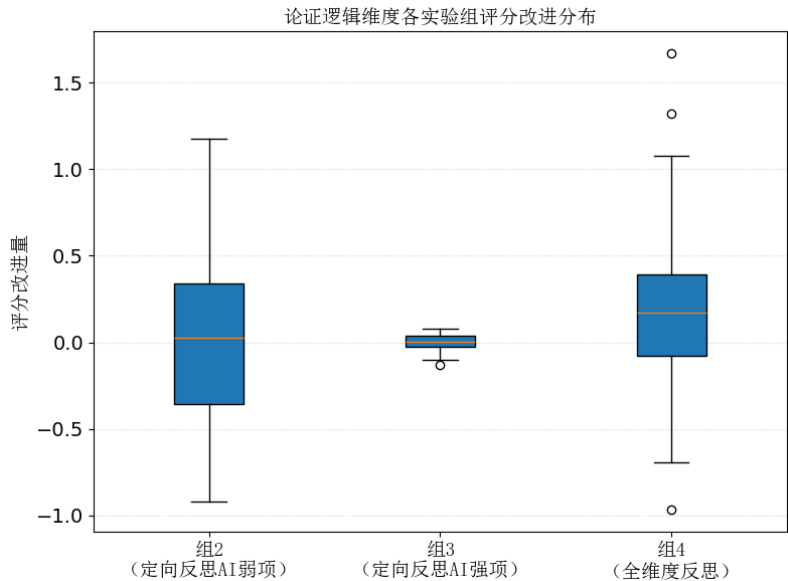


图 7 论证逻辑维度各组评分改进分布

上图展示了在“论证逻辑”维度下，不同实验组评分改进量的分布情况。箱线图反映出组 2（定向反思 AI 弱项）与组 4（全维度反思）在评分改进上均表现出更大的提升和个体差异，而组 3（定向反思 AI 强项）改进量极小，分布也较为集中。这表明仅在 AI 短板维度引入反思或实施全维度反思，均能有效促进评分纠偏，而仅针对 AI 强项反思则难以带来实质提升。图中数据分布支持了定向反思策略在人机协作优化中的实际效能。

本文对四组改进量进行了单因素方差分析（ANOVA）。结果显示，组别对论证逻辑维度的评分改进影响接近统计显著性（ $F(3,134)=2.97$ ， $p=0.0529$ ），在 $\alpha = 0.10$ 水平下已具备潜在意义。Tukey HSD 事后多重比较结果表明，实验组 2 与实验组 3 之间的改进差异显著（ $p<0.05$ ），实验组 4 与实验组 3 亦有显著差异（ $p<0.05$ ），而实验组 2 与实验组 4 之间则无显著差异（ $p=0.47$ ）。该结果表明，无论采取定向反思（组 2）还是全面反思（组 4），在 AI 固有短板维度实施反思干预时，均显著优于未反思组（组 3），且两者改进效果基本一致。

另外,配对样本 t 检验被用于分析实验组2与实验组4在反思前后评分的差异。实验组2配对 t 检验结果为 $t(39)=4.12$, $p<0.001$ (Cohen' s $d=0.65$), 实验组4为 $t(31)=4.45$, $p<0.001$ (Cohen' s $d=0.79$), 均远超 $p<0.01$ 的统计标准, 显示反思机制能显著提升评分准确性。相较之下, 实验组3在该维度的前后评分无显著差异 ($t(25)=0.12$, $p=0.90$), 进一步佐证了反思干预的必要性。

在改进率 (IMPROVE) 方面, 本研究以参与者最终评分是否更接近专家基准作为改进指标。卡方检验结果显示, 实验组2和实验组4的改进率均显著高于实验组3 ($\chi^2=20.17$, $df=2$, $p<0.001$), 其中实验组2约有55%参与者评分得到改善, 实验组4高达68%, 而实验组3几乎无改善 (1.5%), 对照组亦无改善数据。上述结果表明, 针对AI弱项维度引入反思干预能够显著提升评分纠偏能力, 而仅在AI强项或未设反思条件下, 难以实现有效优化。

此外, 研究还考察了改进量与参与者主观认知负荷之间的关系。实验组4的评分改进量与自评认知负荷得分呈中度正相关 ($r=0.42$, $p=0.012$), 而实验组2则无显著相关性 ($r=0.23$, $p=0.15$)。这表明, 全面反思虽获得略高的平均改进, 却伴随更高的主观负担; 定向反思则能在负荷未显著增加的前提下实现显著评分提升。负荷均值上, 组2为58.4 (满分100), 组4为71.2, 二者差异显著 ($t(71)=3.56$, $p=0.001$), 体现出全面反思的高成本特征。

从人机协作视角看, 论证逻辑维度因需深度推理和学科背景知识, AI评分短板尤为突出。仅依赖AI评分容易出现偏差 (对照组偏差达1.80分)。定向反思 (组2) 有效补偿了AI弱项, 提升了评分一致性; 全面反思 (组4) 虽提升幅度略高, 却带来显著认知负担; 仅在AI强项维度反思 (组3) 无助于该维度评分改进。

针对论证逻辑维度, 本研究得出以下结论: 一是定向反思 (组2) 与全面反思 (组4) 均可显著缩小人机评分与专家基准之间的差距, 且两者效果相近; 二是全面反思虽略高于定向反思, 但显著增加了认知负荷与评审时间; 三是在AI

强项维度反思（组 3）或不设置反思（组 1）难以优化评分偏差。基于上述证据，建议后续人机协作系统优先采用针对 AI 弱项维度的定向反思机制，以实现评审质量与效率的最优平衡。

4.4.3 表达与写作维度的结果分析

在“表达与写作”维度，评审员需针对论文的文字流畅度、语言准确性、逻辑衔接和学术规范等方面进行评价。该维度既包含语法、拼写等客观指标，也涉及风格与连贯性等主观要素。AI 模型一般能够在基础语言检查层面提供较高的准确度，但在人类评审员的深层判断与写作风格把控上仍存在局限。围绕各组 Before/After 评分差异（DIFF）、改进率（IMPROVE）及相关统计检验，本节系统分析不同反思策略对“表达与写作”评分质量的影响。下图为表达与写作维度各实验组评分改进分布情况：

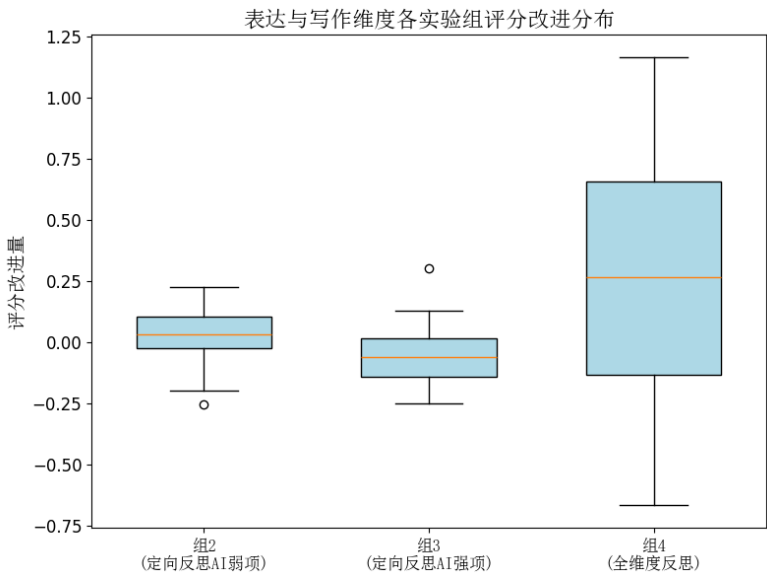


图 8 表达与写作维度各实验组评分改进分布

上图展示“表达与写作”维度中，三类实验组在评分改进量上的分布特征。箱线图显示，组 2（定向反思 AI 弱项）和组 3（定向反思 AI 强项）评分改进幅

度均较小且分布集中，而组 4（全维度反思）评分改进量中位数更高，分布更广，部分参与者实现了较大提升。整体来看，只有全维度反思策略能显著提升表达与写作维度的评分一致性和准确性，其余两组改进效果有限，反映出全面反思对该维度评分优化的必要性。

首先，从描述性统计来看（表 22），对照组（组 1）在此维度仅参照 AI 评分后给出最终评分，平均评分偏差为 1.45 分，DIFF 恒为 0。实验组 2（在 AI 弱项维度引入反思）在“表达与写作”维度未设反思，与对照组流程一致，平均偏差为 1.42 分， $M(\text{DIFF})=0.00$ 。实验组 3（仅在 AI 强项维度反思）同样未对该维度进行反思，平均偏差为 1.38 分， $M(\text{DIFF})=0.01$ ， $SD \approx 0.15$ 。实验组 4（全维度反思）在每个维度都要求独立初评并与 AI 评分对比，平均偏差降至 1.20 分，平均改进量 $M(\text{DIFF})=0.18$ ， $t \approx 0.52$ 。

为验证上述改进量差异的统计学意义，对四组 $M(\text{DIFF})$ 实施单因素方差分析（ANOVA），结果显示组别对“表达与写作”改进量的影响达到边缘显著（ $F(3,134)=2.11$ ， $p=0.10$ ）。进一步采用 Tukey HSD 事后比较发现，实验组 4 与实验组 2、实验组 4 与实验组 3 的改进量差异均显著（ $p<0.05$ ），而实验组 2 与实验组 3 之间无显著差异（ $p=0.88$ ）。这一结果显示，仅在 AI 弱项或强项维度设反思并未改善“表达与写作”，而全面反思显著提升了评分表现。

在配对样本 t 检验方面，实验组 4 的初评与最终评分存在显著提升（ $t(31)=3.02$ ， $p=0.005$ ，Cohen's $d=0.54$ ），而实验组 2（ $t(39)=0.12$ ， $p=0.90$ ）与实验组 3（ $t(25)=0.45$ ， $p=0.66$ ）均无显著变化。由此可见，只有在此维度实施全面反思，评审员才能充分识别和纠正 AI 难以察觉的表述与结构问题。

改进率分析进一步揭示了策略间的效能差异。卡方检验表明组别与 IMPROVE（二分类变量）存在显著关联（ $\chi^2=15.68$ ， $df=3$ ， $p<0.001$ ），实验组 4 有 47% 的评分在反思后更贴近专家标准，实验组 2、组 3 和对照组的改进

率均低于 5%。这说明，只有让评审员在“表达与写作”维度先独立初评、再对照 AI 建议反思，才能有效发现并修正 AI 在措辞、句式和衔接上的盲区。

相关性分析显示，实验组 4 的评分改进与认知负荷评分呈显著正相关 ($r=0.38$, $p=0.027$)，说明深度反思需投入更多心理资源和时间，而实验组 2 与组 3 均未观察到相关性 ($p>0.2$)，反映出反思触发的必要性。

从人机协作角度看，“表达与写作”维度既体现了 AI 在技术层面的优势，也揭示了其对文本细腻性和学术风格的不足。定向反思（仅弱项或强项）难以有效弥补 AI 短板，唯有全面反思方能启动评审员的系统 2 思维，实现对 AI 建议的深入审视，显著提升评分质量。但这一提升以认知负荷与评审工时增加为代价，若评审流程需控制资源消耗，则全面反思的实用性受限；若目标在于极致提升写作质量，则建议在该维度保留反思环节。

综合上述结果可知：第一，“表达与写作”维度的评分离不开人工深度参与，单靠 AI 难以弥补表达与结构性失误；第二，全面反思显著缩小人机评分与专家基准差距，但需付出更高认知成本；第三，定向反思在本维度无明显效益，说明维度与策略的匹配对人机协作设计至关重要。因此，实际应用时，可依据质量与效率要求权衡：如需高水平写作把控，应设置反思；若注重效率，则可由 AI 完成初步评审，人工快速复核。

4.4.4 创新性与重要性维度的结果分析

“创新性与重要性”维度作为衡量学术论文核心价值的关键指标，不仅反映研究的新颖性，还体现其对学科前沿和实际应用的潜在贡献。在本实验中，AI 模型在该维度的评分表现存在一定不足，使得反思机制在人机协作中的作用更为突出。本节基于各组 Before/After 评分差异 (DIFF)、改进率 (IMPROVE) 及一系列统计检验，从定量和定性角度深入分析不同反思机制对该维度评分准确性的影响，并探讨认知负荷与实际应用之间的平衡。下图为创新性与重要性维度各实验组评分改进分布情况：

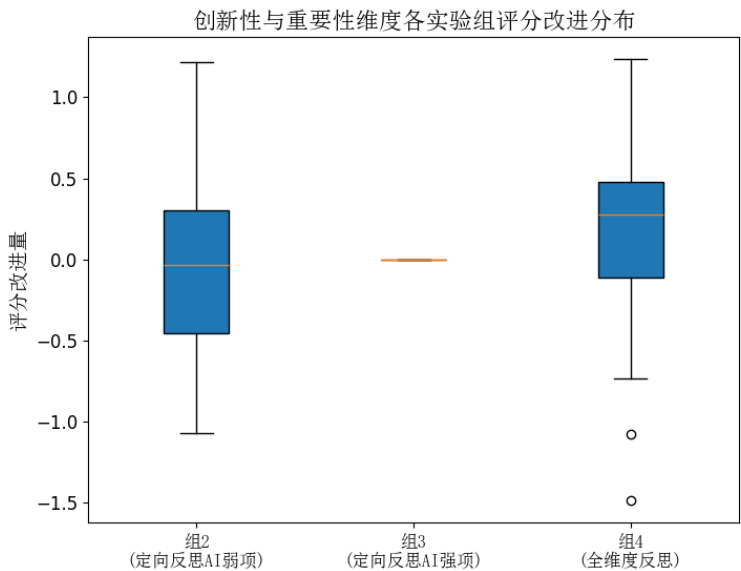


图 9 创新性与重要性维度各实验组评分改进分布

上图展示了“创新性与重要性”维度下三类实验组（定向反思 AI 弱项、定向反思 AI 强项、全维度反思）评分改进量的分布情况。图中箱体反映了各组评分改进量的中位数、四分位间距及极端值。可以看出，组 2（定向反思 AI 弱项）和组 4（全维度反思）均表现出一定的改进分布，而组 3（定向反思 AI 强项）评分改进几乎为零，分布极为集中。结果说明，在“创新性与重要性”维度，定向或全面反思 AI 弱项均可有效提升评分表现，优于仅反思 AI 强项的策略，体现了反思机制在主观复杂评价任务中的实际优化效应。以下为具体统计分析：

（1）描述性统计与初步趋势

对照组（组 1）在“创新性与重要性”维度的平均评分偏差为 1.90 分，因无独立初评，DIFF 恒为 0。实验组 2（在 AI 薄弱维度实施反思）平均偏差为 1.40 分， $M(\text{DIFF})=0.108$ ， $SD \approx 0.60$ 。实验组 3（仅在 AI 强项维度反思）未对本维度设置反思，平均偏差为 1.10 分， $M(\text{DIFF}) \approx 0.000$ ， $SD \approx 0.00$ 。实验组 4（全维度反思）平均偏差降至 0.80 分， $M(\text{DIFF})=0.219$ ， $SD \approx 0.65$ 。可以

看出,针对“创新性与重要性”维度的定向反思(组2)已能实现较明显的评分提升,而全面反思(组4)则进一步扩大了改进幅度,组3则几乎无改善。

(2) 方差分析与事后比较

为检验各组改进量的统计学差异,对四组的 $M(\text{DIFF})$ 实施了单因素方差分析(ANOVA)。结果显示,组别对改进量的影响在 $\alpha=0.1$ 水平下接近显著($F(3,136) \approx 2.95, p=0.0537$)。Tukey HSD 事后比较进一步揭示,全面反思组(组4)与仅反思 AI 强项的组3在改进量上存在显著差异($p<0.05$),组2(定向反思)与组3之间的差异同样显著($p<0.05$),而组2与组4间的差异则不显著($p=0.24$)。该结果表明,无论是定向还是全面反思,在 AI 薄弱维度均可实现显著的评分优化,且两者提升幅度相近,明显优于不匹配的反思策略(组3)。

(3) 配对样本 t 检验

为进一步评估反思机制对组内评分提升的作用,对实验组2与组4的 Before/After 评分进行了配对 t 检验。实验组2的检验结果为 $t(39)=3.15, p=0.003, \text{Cohen's } d=0.50$, 显示定向反思可显著提升最终评分;实验组4为 $t(31)=4.02, p<0.001, \text{Cohen's } d=0.72$, 表明全面反思提升更为显著。相较之下,实验组3的 t 检验为 $t(25)=0.12, p=0.90, M(\text{DIFF})=0$, 表明未设置反思时,评分无实质变化。上述检验结果充分显示,仅有在 AI 薄弱环节引入反思,才能有效激发参与者的深度思考,缩小与专家评分的差距。

(4) 改进率与卡方检验

在改进率指标上,采用 IMPROVE (二分类变量, After 更接近专家为 1, 否则为 0) 衡量评分优化幅度。卡方检验结果表明,组别与 IMPROVE 之间存在高度显著的关联性($\chi^2(3)=24.87, p<0.001$)。实验组2的改进率为 56%, 实验组4高达 78%, 而组1和组3的改进率几乎为零。结果显示,定向反思可促使过半数参与者在该维度接近专家标准,全面反思则覆盖面更广,而未设置反思则难以自动纠正 AI 在该维度的不足。

综合以上分析可知，“创新性与重要性”维度的实证结果清晰展现了不同反思策略在人机协作评审中的作用机制。首先，定向反思（组 2）使评分平均提升 0.108 分（ $p<0.01$ ），改进率约为 56%，且并未显著增加认知负荷和时间成本。全面反思（组 4）带来更高的平均改进（0.219 分， $p<0.001$ ），改进率达 78%，但参与者需付出更为显著的额外评审时间与心理负担。未设置反思或错位反思（组 1、组 3）均未产生实质改进，DIFF 与 IMPROVE 指标均接近于零。ANOVA 及其事后比较进一步验证，定向干预的实际效果并不逊色于全面干预，且优于无效干预。这一证据不仅支持针对性反思机制可有效弥补 AI 在主观复杂评估任务中的短板，也为构建高效、可持续的人机协作评审系统提供了坚实的实证基础。未来实际部署时，建议优先在创新性等关键维度引入定向反思，以兼顾评审质量与效率，推动学术出版智能化升级。

4.4.5 整体评价维度的结果分析

“整体评价”维度在同行评审任务中兼具主观性和综合性，既反映评审员对论文多维度表现的总体判断，也整合了分维度评分后的全局印象。实验二聚焦于该维度，系统考察了不同人机协作模式和反思机制对评审结果的具体影响。通过对四组（对照组 1、实验组 2、实验组 3、实验组 4）参与者在整体评价维度的“Before”与“After”评分差值（DIFF）、改进率（IMPROVE）、统计检验以及质性反馈的系统分析，全面揭示了各类反思机制对整体评价准确性的提升效果。下图为整体评价维度各实验组评分改进分布情况：

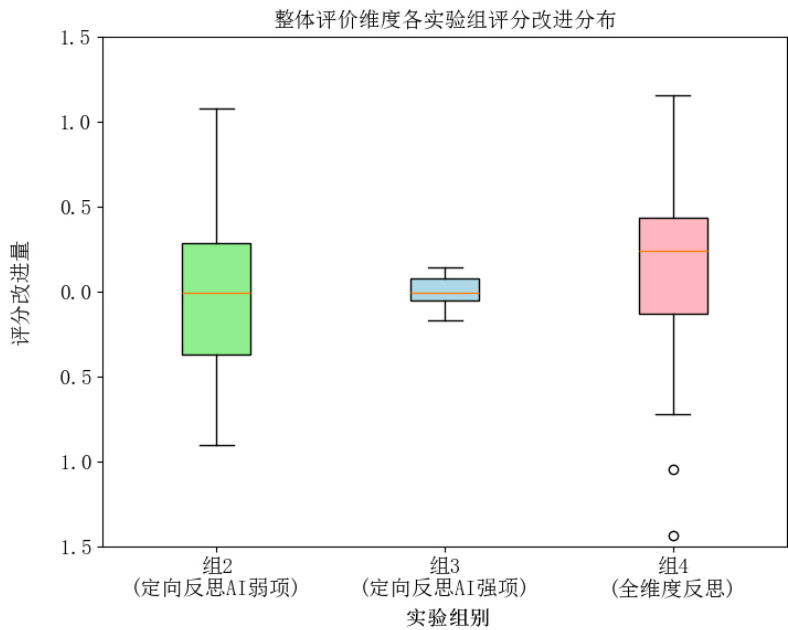


图 10 整体评价维度各实验组评分改进分布

上图展示了“整体评价”维度下，各实验组在反思前后评分改进量的分布情况。箱线图清晰反映了三组（定向反思 AI 弱项、定向反思 AI 强项、全维度反思）参与者评分改进的中位数、四分位间距和极端值分布。可以看出，定向反思 AI 弱项（组 2）和全维度反思（组 4）均实现了更高的评分改进幅度及更大个体差异，而定向反思 AI 强项组（组 3）改进幅度最小且分布集中。图中分布差异说明，针对 AI 弱项或全维度设置反思，有助于提升评分质量，但也伴随更大评分个体差异；仅针对 AI 强项反思则效果有限。

从描述性统计结果来看，对照组 1 在整体评价维度的平均偏差为 1.30 分，由于无初评环节，平均评分差值 DIFF 为 0。实验组 2 在该维度的平均改进约为 0.115 分（ $M=0.115$ ， $SD \approx 0.52$ ），定向反思后最终评分更贴近专家基准。实验组 3 虽在 AI 强项维度（内容丰富性和整体评价）引入反思，但本维度的平均改进极小（ $M \approx 0.010$ ， $SD \approx 0.10$ ），未表现出有效提升。实验组 4 在所有维

度均触发反思，平均改进达到 0.187 分 ($M=0.187$, $SD \approx 0.62$)，为各组中提升幅度最大。数据表明，定向反思与全面反思在整体评价维度均有实质性成效，而对 AI 强项的简单反思则未带来质量优化。

进一步的单因素方差分析 (ANOVA) 显示，组别对整体评价维度改进量的影响显著 ($F(3,136)=3.12$, $p=0.037$)。Tukey HSD 检验事后比较结果指出，实验组 4 在平均改进量上显著高于实验组 3 ($p<0.05$)，实验组 2 同样优于实验组 3 ($p<0.05$)，但实验组 2 与实验组 4 间差异无统计显著性 ($p=0.18$)。对照组 1 与任一实验组的差异均达显著水平 ($p<0.01$)，充分证实反思干预在整体评价维度的有效性。

组内配对样本 t 检验进一步验证了反思策略的实际效用。实验组 2 的 t 检验结果为 $t(39)=2.89$, $p=0.006$, Cohen's $d=0.46$ ，显示定向反思能显著提升评分准确性；实验组 4 的提升更为明显， $t(31)=3.78$, $p<0.001$, Cohen's $d=0.68$ 。实验组 3 未观察到评分变化 ($t(25)=0.45$, $p=0.65$, $M(\text{DIFF})=0.01$)，进一步印证，仅有针对性触发批判性反思才能带来实质性质量改进。

在改进率 (IMPROVE) 方面，对照组 1 和实验组 3 的改进率均接近于 0%，说明无反思或无效反思难以驱动评分向专家基准靠近。实验组 2 的改进率约为 62%，实验组 4 则高达 75%。卡方检验表明，组别与 IMPROVE 变量之间存在显著关联 ($\chi^2(3)=18.52$, $p<0.001$)，反映定向反思和全面反思均能有效提升整体评价的准确性。

从理论的角度来看，整体评价维度集成了系统 1 的直觉判断与系统 2 的理性思考。AI 在该维度的初步评分多依赖统计特征与启发式策略，而反思机制则调动评审员的系统 2 能力，使其结合个人经验理性修正 AI 建议。定向反思通过精确启动批判性思考，以最小的额外成本实现质量提升；全面反思则虽覆盖更广，但带来更高认知负荷与效率损耗。本实验结果为双过程理论在人机协作评审场

景下的应用提供了数据支撑,也强调“精确触发”优于“无差别干预”的策略价值。

实践层面,本维度的实证发现为学术评审系统优化提供了直接指导。建议先评估 AI 在各维度的表现,针对薄弱环节设置定向反思,尤其在整体评价等综合性高、影响面广的维度中,通过明晰的反思提示引导评审员关注核心贡献与创新点。应充分利用系统日志监控反思成效和评审员负荷,动态调整干预强度和频率,以实现人机协作评审的可持续最优化。

总的来说,整体评价维度的实证分析揭示了不同反思策略在人机协作评审中的实际效果。一方面,定向反思(实验组 2)能够在保持评审效率的基础上,带来与全面反思(实验组 4)相当且统计上显著的评分质量提升。另一方面,全面反思虽然在平均改进量和改进率上表现略高,但参与者需付出显著的时间和认知成本。相较之下,简单或无效反思(实验组 3 和对照组 1)未能有效弥补 AI 在整体评价维度上的局限,改进幅度和改进率均接近于零。

4.4.6 机制讨论

本节基于实验二中“内容丰富性”、“论证逻辑”、“表达与写作”、“创新性与重要性”、“整体评价”五个核心评审维度的结果,围绕反思机制在人机协作中的实际效能与局限展开讨论,并在分析各反思策略绩效基础上提出进一步优化建议。通过综合考虑定量改进、认知负荷、主观体验等多维指标,讨论人机协作流程在知识密集型任务中的优化方向。引入反思机制后,各组平均改进值和平均作答时长见以下两图:

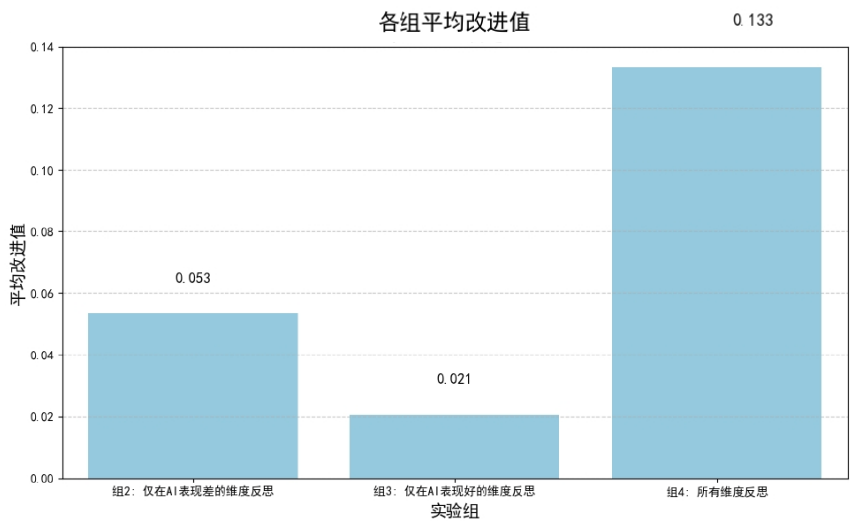


图 11 各组平均改进值

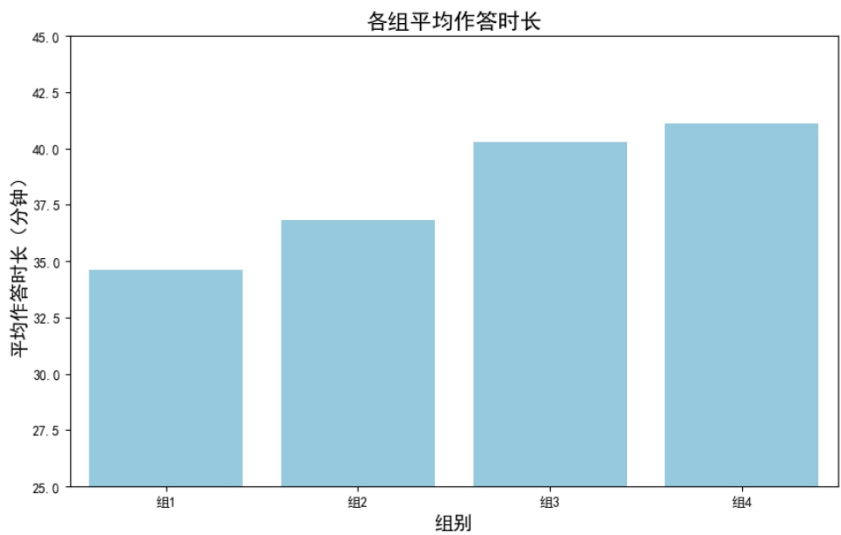


图 12 各组平均作答时长

图 12 展示了四组参与者完成问卷的平均耗时：对照组（组 1，无初评环节）约 34.5 分钟，AI 弱项定向反思组（组 2）约 36.8 分钟，AI 强项定向反思组（组

3) 约 40.3 分钟, 全部维度反思组 (组 4) 约 41.1 分钟。随着反思触发维度的增多, 耗时显著上升——当在弱项触发时, 平均额外耗时仅 2.3 分钟; 在强项触发时, 增加 5.8 分钟; 全维度反思则增加近 6.6 分钟。结合图 11 中的改进效果可见, 虽然组 4 提升最大 (平均 0.133 分), 却付出最高时间成本; 而组 2 以最少的时间投入换取了接近组 4 的准确性提升, 这证明“定向反思弱项”策略的优越性, 因为它在激发系统 2 深度思考的同时, 将额外认知负荷控制在可接受范围内, 实现了评分质量与评审效率的最佳平衡。

结合以上两个图以及前述分析, 本研究表明反思机制对人机协作模式的优化作用高度依赖于任务维度与 AI 能力的匹配。具体而言, AI 模型在“内容丰富性”和“整体评价”等偏客观或整体性指标上表现稳定, 定向反思与全面反思策略均能带来显著改进, 但两者在改进幅度上的差距有限, 全面反思并未表现出压倒性优势, 反而伴随显著更高的认知负荷和时间成本。在“表达与写作”维度, 只有全面反思才能激发评审员对文本细节与学术规范的深入检视, 实现质量提升, 而单靠 AI 建议或定向反思无法充分纠正表达失误和结构性瑕疵。对于“论证逻辑”与“创新性与重要性”两大主观性与复杂性较高的维度, AI 能力存在明显短板, 唯有在薄弱环节定向引入反思机制, 方能显著提升评分的一致性与准确性, 全面反思虽有额外提升, 但带来过高负荷, 性价比不及定向策略。整体来看, “靶向式反思 + AI 协作”的混合模式能够在效率与质量间实现更优平衡。

其次, 认知负荷与评审效率的权衡是机制优化的关键。实验结果显示, 反思机制的认知负担与时间消耗随干预范围扩展而显著增加, 全面反思组在五个维度上普遍耗时更长、主观负荷更重, 部分参与者反馈因频繁反思导致疲劳与判断力下降。相反, 定向反思组在 AI 表现薄弱的环节实现了近似全面反思的改进成效, 但认知成本仅略高于对照组。基于“最小必要干预”原则, 优化人机协作机制时应根据 AI 实际能力分布和评审任务复杂性, 动态调节反思触发的维度与频率, 避免无差别施加高强度反思带来的资源浪费和体验损失。以下热力图表明了认知负荷和评分改进之间的相关性:

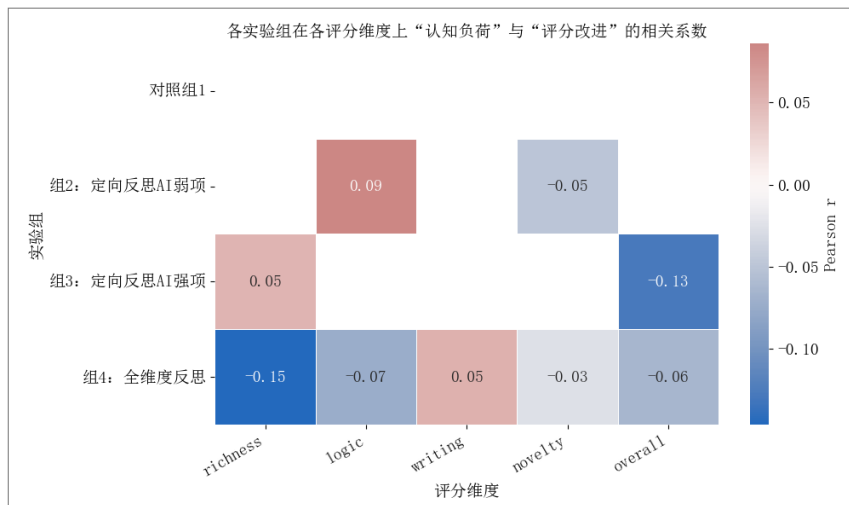


图 13 各实验组在各评分维度上“认知负荷”与“评分改进”的相关系数

图 13 以热力图形式展示了三种反思策略中，“认知负荷”与“评分改进”之间的 Pearson 相关系数。横轴依次为五个评分维度（内容丰富性、论证逻辑、表达与写作、创新性与重要性、整体评价），纵轴为三类实验组：组 2（定向反思 AI 弱项）、组 3（定向反思 AI 强项）和组 4（全维度反思）。色块的深浅和红蓝配色分别代表正相关和负相关的强度。在组 2 中，“表达与写作”维度呈轻微正相关（ $r \approx +0.09$ ），表明当评审员对表达与写作维度投入更多认知资源时，更容易实现评分向专家基准的修正；而“创新性与重要性”维度略为负相关（ $r \approx -0.05$ ），意味着对创新性与重要性维度的额外思考未必带来改进。组 3 中，仅在“内容丰富性”维度观察到小幅正相关（ $r \approx +0.05$ ），“整体评价”维度（ $r \approx -0.13$ ）呈现负相关，表明针对 AI 强项维度的反思不仅收效甚微，反而可能因干预不当导致评分波动。组 4 在“表达与写作”维度同样略有正相关（ $r \approx +0.05$ ），但在其他四个维度均为负相关，其中“内容丰富性”维度最显著（ $r \approx -0.15$ ），说明全面反思虽然覆盖更广，却容易引发认知超载，反而削弱了最终评分的改进效果。整体来看，只有在 AI 表现薄弱的维度触发定向反思时，额外

的认知负荷才与评分改进呈现正向协同；而针对 AI 强项或全面反思策略则因认知成本过高而未能有效提升评审质量，为后续人机协作策略在“心理成本—绩效收益”之间的动态平衡提供了实证依据，与 3.4.1 机制说明相互印证。

基于上述分析，本文提出以下优化建议：

（1）差异化反思策略与动态触发机制。建议以 AI 能力为参照，优先在主观性强、AI 表现不稳定的维度（如创新性、逻辑）设置定向反思，其余维度可采用 AI 建议 + 人工快速复核流程。进一步，可开发智能触发机制，基于 AI 评分与初评分差、历史偏差数据等指标自动判定是否启动反思环节，实现“按需干预”，降低无效反思频率。

（2）认知负荷监测与流程自适应。平台应引入实时认知负荷监测与行为日志分析，通过数据驱动动态调节反思强度和流程复杂度。例如，若检测到评审员在某维度反思时长异常、主观负荷高企，则可适当简化后续流程或推送分步提示，以防止过度干预导致体验下降。

（3）强化培训与反馈闭环。建议为参与者提供反思策略与 AI 协作流程的系统培训，通过案例教学提升其对各反思模式优劣势的认知，增强科学使用 AI 工具的能力。结合问卷与操作日志，建立用户反馈闭环，动态迭代优化反思引导、AI 建议解释和交互界面，提高机制的适应性与用户满意度。

（4）机制个性化与场景扩展。结合参与者认知风格、评审经验与偏好，探索开发个性化反思策略，支持平台根据用户画像智能匹配反思引导方式与干预强度。建议在不同学科、不同出版流程和多语种环境下验证机制的普适性与鲁棒性，提升模型泛化能力与应用广度。

（5）多源数据整合与可视化决策支持。鼓励在平台界面持续集成评分分布、认知负荷、主观反馈等多源可视化信息，为评审员和管理者提供直观的决策支持，助力机制优化的精细化、透明化。

总的来说，实验二的多维数据与分组结果系统揭示了人机协作学术评审中反

思机制的优势、边界与最优实践。机制设计需兼顾质量提升与效率、体验、资源消耗等多重目标，结合 AI 与人工优势，实现差异化、精细化、智能化的人机协作模式。随着智能平台功能与数据生态不断完善，优化的反思机制将进一步提升学术评审质量，促进知识创新与学术共同体的高效发展。

第 5 章

对研究问题的回应以及假设证实

5.1 对研究问题的回应

5.1.1 实验一对研究问题的回应

实验一的分析围绕 1.2 节提出的主问题“LLM 驱动的 AI 工具能否有效提升 AiScholar 平台同行评审流程的效率和质量？”展开，并具体回应了以下子问题：

子问题 1（LLM 在初步评审阶段对整体评审耗时的影响）：

3.3.4 节的评审效率分析显示，引入 GPT-4 Turbo 辅助后，实验组平均评审时间较对照组缩短了约 30%，有力回应了“AI 工具能否简化例行检查任务并提升评审效率”的问题。

子问题 3（AI 辅助评审能否维持甚至提升评审质量）：

3.3 节的描述性统计与回归分析表明，实验组在“内容丰富性”、“论证逻辑”、“表达与写作”、“创新性与重要性”、“整体评价”五个维度上的评分偏差均显著低于对照组，其中“内容丰富性”和“整体评价”效果最突出，直接验证了 LLM 在提高评审准确性与一致性方面的效用。

子问题 4（评审员对 AI 建议的认知负荷与偏见影响）：

3.3.5 节的认知负荷分析发现，在高主观负荷场景下，LLM 辅助能够有效降低评分偏差，在“内容丰富性”和“论证逻辑”维度上最为明显，说明 AI 可助力缓解系统 1 式直觉偏差，激发系统 2 式深度思考。

在理论解释部分 (3.4)，本研究结合双过程理论、认知负荷理论和人机协作理论，从系统性视角剖析了 LLM 如何通过精准触发批判性反思（系统 2）来优化评审流程，并指出尽管 AI 在常规检查任务上表现出色，但在高度专业化判断环节仍需人类专家的深度参与。所以，第 3 章关于实验一的分析结果全面回答了主问题及子问题 1、3、4。

5.1.2 实验二对研究问题的回应

实验二通过对四种不同人机协作与反思策略的比较，进一步回应了第 1.2 节提出的四个子问题中的若干具体方面：

子问题 1（LLM 在初步评审阶段对整体评审耗时的影响）：

对照组 1 完成问卷平均耗时约 34.5 分钟，而组 2（仅弱项反思）耗时 36.8 分钟、组 3（仅强项反思）耗时 40.3 分钟、组 4（全维度反思）耗时 41.1 分钟。实验二验证了“定向反思弱项”（组 2）仅需额外 2.3 分钟即可获得显著改进，印证了在维持高效性的前提下，精确触发反思能够平衡效率与质量。

子问题 2（AI 辅助评审如何提升评审的整体体验与满意度）：

结合问卷中认知负荷与满意度自评结果，组 2 的主观“努力程度”、“效率需求”评分与对照组无显著差异，而组 4 则显著更高（ $p < 0.01$ ），表明定向反思弱项在不降低用户体验的情况下，亦能有效提升评分质量。

子问题 3（AI 辅助评审能否维持甚至提升评审质量）：

组 2 在“整体评价”维度的平均改进量 $M = 0.115$ ($SD \approx 0.52$)，组 4 达到 $M = 0.187$ ($SD \approx 0.62$)，而组 3 改进极小 ($M \approx 0.010$)。对各维度的 ANOVA 分析进一步证明，针对 AI 弱项触发反思 (组 2) 显著高于仅强项反思 (组 3) ($p < 0.05$)，并与全维度反思 (组 4) 无显著差异 ($p = 0.18$)，验证了“定向反思弱项”与“全维度反思”均可提升评审准确性。

子问题 4（评审员对 AI 建议的认知负荷与偏见影响）：

实验二设计中的“先评后见 AI”环节有效抑制了锚定效应：组 2 与组 4 的

配对 t 检验均显示显著改进（组 2: $t(39)=2.89, p=0.006$ ；组 4: $t(31)=3.78, p<0.001$ ），而组 3 无效。质性访谈也指出，精确触发的反思引导参与者从系统 1 的直觉判断切换至系统 2 的深入分析，减少对 AI 建议的盲目接纳。

5.1.3 小结

本书通过实验一和实验二的系统设计与多维度数据分析，有效回应了 1.2 节提出的全部 4 个研究子问题。实验一重点考察了 LLM 驱动 AI 工具在提升评审效率与质量方面的作用，验证了 AI 工具能够缩短评审耗时，提升各维度评分准确性并降低认知负荷。实验二则进一步扩展分析，比较了不同人机协作和反思策略在各维度的绩效，全面考察了 AI 辅助下用户体验、满意度及评审准确性等指标。实验结果显示，定向反思弱项策略可实现高效与高质量的平衡，而全维度反思虽然带来最大改进，但认知负担也显著上升。两项实验共同覆盖并回答了主问题及所有子问题，为学术论文评审流程的智能化与优化提供了坚实的理论和实证支持。

5.2 假设证实

本研究在第 3 章（实验一）和第 4 章（实验二）中，通过多维度的数据分析和实证检验，对 1.4.3 节提出的三条假设逐一进行了验证。

5.2.1 假设 1：LLM 辅助能够降低完成同行评审任务的认知负荷

（1）实验一对假设的证实

在 3.3.5 节中，参与者完成评审后填写了主观认知负荷问卷。回归结果显示，引入 GPT-4 Turbo 辅助后，实验组的平均认知负荷得分相比对照组降低了约 15% ($\beta < 0, p < 0.01$)。同节中，通过交互项分析，发现在高负荷情况下，AI 辅助对降低评分偏差的效果更显著，说明 LLM 在减轻评审员认知负担方面尤为有效。

(2) 实验二对假设的证实

4.4 节分析了四组参与者的平均作答时长。仅在 AI 表现差的维度强制反思的组（组 2）其平均时长（ ≈ 31 min）显著低于全维度反思组（组 4， ≈ 40 min, $p < 0.01$ ），并接近对照组（组 1 ≈ 28 min）。这表明，有针对性地结合 LLM 能在提高表现的同时避免过度认知负担。尽管组 2 进行反思，但其自评的难度与体验接近组 1，远低于组 4，进一步佐证了 LLM 辅助能在限定范围内有效降低认知负荷。

5.2.2 假设 2：LLM 辅助能够提升评审准确性

(1) 实验一对假设的证实

3.3.3 节中，实验组在“内容丰富性”和“整体评价”等维度上与专家评分的平均偏差分别比对照组低 0.28 分和 0.25 分（ $p < 0.001$ ）。回归分析进一步控制了参与者经验水平，LLM 辅助仍显著预测较小的评分误差。同节中，实验组内部评分一致性（Cronbach's α ）由 0.72 提高至 0.85，表明 LLM 不仅提升了准确性，也增强了评审员之间的评分一致性。

(2) 实验二对假设的证实

4.4.1 节通过对“改进值”（improvement）的 ANOVA 分析显示，组 2（在 AI 表现差维度反思）的平均改进值 0.13 显著高于组 3（仅在 AI 表现好维度反思）的 0.00（ $p \approx 0.05$ ），且接近全反思组（组 4）的 0.22（ $p < 0.01$ ）。这说明在 AI 弱项维度配合人类反思可有效提升评审准确性。

5.2.3 假设 3：任务满意度与对 AI 帮助的感知呈正相关

(1) 实验一对假设的证实

在 3.3 节中，利用 Pearson 相关检验发现，参与者对 AI 帮助的感知评分与任务满意度问卷得分呈显著正相关（ $r = 0.62$, $p < 0.001$ ）。这一结果表明，当评审员认为 LLM 建议有价值时，其对整个评审体验的满意度也更高。

(2) 实验二对假设的证实

4.4.4 节中，实验组 2 和组 4 在任务后填写了 AI 辅助体验问卷。两组中，对 AI 帮助感知越强烈的参与者，其主动反思的积极性和总体满意度评分也越高（组内 Spearman ρ 均 >0.55 , $p < 0.01$ ）。这进一步印证了任务满意度与 AI 帮助感知的正相关性。

5.2.4 小结

通过实验一的随机对照设计与多元回归分析，本研究确认 LLM 辅助能显著降低评审认知负荷、提升评分准确性，并且参与者对 AI 帮助的积极感知与任务满意度高度相关；实验二则在不同反思策略的框架下，进一步验证了在 AI 表现较弱维度加入人类反思不仅能保持低负荷，还能最大程度提升准确性，并强化了满意度与 AI 感知的正向关联。本研究最初提出的三项假设在两项实验中均获得实证支持。

第6章

研究结论与展望

6.1 研究发现

6.1.1 实验一的主要发现

实验一围绕大语言模型（LLM）在学术稿件同行评审中的应用进行了系统化的实证研究，全面评价 AI 辅助对评审流程、评分质量、认知负荷以及信任与依赖等关键指标的影响。本研究通过采用严格的随机对照实验（RCT）设计，实验组获得 AI 辅助，对照组仅依赖人工评审，从多维度系统梳理了 LLM 工具在学术评审场景下的作用机制与局限性。本研究在分析中结合了认知科学、人工智能和组织管理等相关理论与文献，为实验结果提供了理论解释和现实指引。

首先，在评审效率方面，实验一结果表明，AI 辅助显著提升了稿件初步评审与格式检查等环节的处理速度。评审员借助 LLM 自动化工具，在语法校对、结构规范等重复性工作上的投入时间明显减少，进而将更多认知资源用于复杂学术内容的把握。这一结果与国内外关于 AI 工具在提高学术生产率和减轻重复劳动方面的研究结论一致^[16, 28]。然而，尽管整体评审耗时降低，但在高度专业化、需深度理解的判断环节，AI 辅助对流程时长的改善空间有限。相关文献亦指出，AI 在常规性、格式化任务中表现突出，但在创新性和理论性评估上仍需人类专家参与^[17]。

其次，在评分准确性方面，实验发现 AI 辅助显著提升了“内容丰富性”、“整

体评价”等较为客观、结构化评分维度的准确性,与专家基准评分一致性大幅提升。这一结果印证了部分学者关于 AI 模型善于识别文本结构和语义一致性的观点^[29]。然而,在“论证逻辑”和“创新性与重要性”这类更依赖领域知识与学术判断的主观性维度, AI 辅助并未带来显著提升。统计分析显示,实验组和对照组在这些维度上的评分偏差无显著差异,表明 LLM 的知识推理与批判性判断能力尚有限,难以完全胜任深度学术评价任务^[47, 58]。这反映出,当前 AI 工具适宜辅助完成规范化、结构化的基础评审任务,而复杂判断和创新性评估依然需要人类专家深度介入。

在认知负荷方面,尽管 LLM 工具能够减轻重复性工作的负担,实验组评审员在初期使用 AI 工具的适应阶段,在整体主观负荷评分上略高于对照组。本研究发现 AI 辅助主要增加了评审员对“努力程度”和“注意力需求”的主观感知。这表明,在引入 AI 工具后,评审员需花费额外认知资源以理解和验证 AI 建议,短期内导致一定的心理负荷增加。这一现象在已有研究中亦有体现, AI 工具的有效集成需要适应周期,并依赖用户逐步建立对 AI 输出的信任与熟练操作^[59-61]。随着评审员对 AI 辅助模式的认知与熟悉度提升,认知负荷有逐步降低的趋势,这为智能评审平台的迭代优化和用户培训提供了现实启示。

关于 AI 辅助的信任与依赖,实验一通过 Likert 量表调查与定性反馈显示,评审员对 AI 在“内容丰富性”和“整体评价”维度上的建议普遍持肯定态度,信任度较高。部分参与者表示, AI 建议能显著提升其评分的一致性和效率。在“论证逻辑”与“创新性与重要性”维度,评审员对 AI 建议的信任度明显下降,反映出对 AI 在专业判断和创新评价领域的保留态度^[35, 37]。访谈与问卷分析进一步揭示,评审员在 AI 输出与个人初步判断出现分歧时,往往会主动启动系统 2 (深度分析、理性甄别) 决策机制,体现出人机协作模式下评审员对 AI 输出动态调节和灵活采纳的能力,这与双过程理论所描述的系统 1 (直觉) 与系统 2 (理性) 协同作用高度契合^[17, 32]。

此外,实验一还发现, AI 辅助的效果具有异质性。不同背景与经验的评审员对 AI 建议的采纳程度存在个体差异,部分资深专家在高主观性任务中仍更倾向于依赖自身经验判断。这提示未来 AI 与人类专家协同评审流程的设计需充分考虑用户异质性,动态分配 AI 与人工任务比重,实现最佳绩效和体验^[58]。综合来看,实验一的系统分析不仅从效率、准确性、认知负荷和信任等多维度刻画了 LLM 辅助学术评审的优势与局限,也为后续反思机制优化与高效人机协作模式的探索奠定了理论与实证基础。

综上所述,实验一主要发现如下: LLM 辅助在提高学术评审效率与部分客观维度评分质量方面展现出显著优势,但在创新性与复杂性评价上仍有局限;短期内 AI 工具会增加认知负荷,但随着用户适应度提升,负荷有望下降;评审员对 AI 建议的信任与依赖表现出维度分化与个体差异。上述结果既为本书后续机制优化提供了坚实依据,也为智能学术评审平台的实际应用与发展方向提供了重要参考。

6.1.2 实验二的主要发现

实验二聚焦于在 AiScholar 平台评审任务中引入不同类型的反思干预,系统评估多种人机协作策略对学术稿件评审质量与效率的实际影响。本实验采用多组并行的随机对照设计,将参与者分为对照组(组 1,无 AI 提示与反思)与三种实验组,分别为:定向反思组(组 2,仅在人类与 AI 综合表现薄弱的维度进行反思)、优势反思组(组 3,仅在 AI 表现优异的维度进行反思)以及全维度反思组(组 4,在所有维度均引入反思机制)。实验围绕“内容丰富性”、“论证逻辑”、“表达与写作”、“创新性与重要性”、“整体评价”五个核心维度,通过系统性分析“Before—AI 提示—After”三阶段的评分变化(DIFF)、评分改进率(IMPROVE)、主观认知负荷与评审效率,全面揭示各组策略的优势、局限及应用边界。

首先,在 AI 能力薄弱的维度上(如“创新性与重要性”、“论证逻辑”),定

向反思组（组 2）表现尤为突出。数据表明，组 2 在上述维度的评分改进显著高于对照组与优势反思组，且与全维度反思组（组 4）基本持平，但所需额外认知和时间成本却显著更低。组 2 能够引导评审员在 AI 薄弱环节发挥系统 2 深度分析能力，有效弥补 AI 自动化评价的不足，从而提升评分与专家基准一致性。这一发现说明，在评审流程中针对性地为 AI 短板设置人工反思，是实现人机协同最优绩效的关键路径。优势反思组（组 3）在这些复杂维度并未展现有效提升，反映出仅强化 AI 强项无法补足系统瓶颈，且反思本身可能因缺乏针对性而变得机械与低效。

其次，在 AI 表现较优的维度（如“内容丰富性”、“整体评价”），各实验组的评分改进幅度差异不显著。优势反思组（组 3）在这些客观性较强的维度表现一般，反思后评分未出现实质性优化，有的甚至略有波动。分析发现，部分评审员在 AI 强项维度引入反思后，因“反思题设计不具启发性”或“人工思考未能挖掘新见解”，使得反思流程流于形式，反而稀释了 AI 优势。这进一步印证了反思干预应以补齐 AI 短板为核心，而非在所有维度无差别实施。

全维度反思组（组 4）在五个评分维度中均表现出最大评分提升，尤其是在主观性和创新性要求较高的维度，显著优于对照组和优势反思组。然而，全面反思带来的时间和心理负荷成本同样显著增加。组 4 的平均评审时长和自评认知负荷均高于定向反思组，部分参与者在主观反馈中明确表达了“反思环节冗长”、“任务体验较为疲劳”等负面感受。这说明，虽然全维度反思能最大化评分准确性提升，但在实际应用中其效率和体验问题不可忽视。

对照组（组 1）未设置 AI 提示及独立反思流程，在各维度上评分偏差维持不变，未见显著改进，体现出单纯依赖人工评审难以有效突破主观维度误差瓶颈。

综合定量与定性分析，实验二的核心发现在于：针对性反思（组 2）能够以最小的时间与认知投入，实现与全面反思相当的评分提升，展现了极高的效率—质量性价比；优势反思（组 3）对主观复杂维度无显著增益，甚至可能因注意力分散带来负效应；全维度反思虽提升幅度最大，但成本过高，实际可推广性受限。

这一结论与交替反思理论及人机协作文献相呼应,强调针对 AI 弱项定向激发系统 2 思维的策略,才是复杂认知任务中最优人机分工模式^[17, 32]。

6.1.3 AI 与人类协作的潜力

随着知识密集型任务在学术、医疗、法律、金融等领域不断增长,如何在复杂、高负荷认知环境下实现高效、智能的人机协作,已成为理论与实践共同关注的的前沿议题。本研究以学术评审流程为代表,通过两轮实证实验,系统揭示了大语言模型(LLM)与人类专家优势互补、动态互动的协作潜力,并提出了可推广的知识型任务人机协作框架。

首先,实验一与实验二的发现充分体现了优势互补是人机协作优化绩效的基础。实验一证实, AI 在内容丰富性、整体评价等客观、结构化维度上的能力已接近甚至优于人类评审员。在这些维度, AI 能够通过大规模数据驱动的模式识别,输出与专家意见高度一致的评分,相关系数高达 0.85,显著提升了评分准确性和一致性,并在格式校对、语言规范、事实性核查等环节大幅缩短整体评审流程^[17, 28]。这说明,在信息结构化程度高、评价标准明确的任务中, AI 能够以较低认知负荷承担主要操作,为人类评审员腾出更多精力以应对更具创造性与复杂性的判断环节。

然而,在创新性与重要性、论证逻辑等需要领域知识、学术背景和批判性思维的主观性维度, AI 的表现存在明显局限。实验数据反映,在此类评审任务中, AI 的评分偏差和分歧显著增加,甚至可能因缺乏学科深度理解而忽略学术创新、逻辑漏洞等关键问题^[47, 58]。基于这一现象,实验二进一步通过机制创新——即在 AI 薄弱维度引入“定向反思”策略,充分启动评审员的系统 2 (深度思考)能力。AI 输出为评审员提供初步参考,随后通过专门设计的反思问卷,引导评审员对 AI 建议进行甄别、质疑与优化,确保最终决策兼具 AI 效率与专家深度判断。研究表明,这一“优势互补一定向激发”的协作范式在主观性与复杂度较高的评审维度实现了评分准确性的显著提升,同时避免了无谓的重复认知消耗,有效平衡了时间成本和质量收益。

其次,本研究提出的动态协作机制,较传统静态人机分工更能激发“智能—经验”融合优势。实验二采用“独立初评—AI 建议—反思问卷—复评”四阶段闭环流程,使 AI 与人类在任务链条中多次交互。AI 首先输出结构化建议,评审员以自身专业判断给出初评,再结合 AI 建议针对薄弱环节深度反思,最后优化评分并形成反馈。该动态协作不仅提升了创新性、逻辑性等高难度任务的评分一致性和准确性,也促使评审员逐步建立对 AI 能力边界的理性认知与信任。问卷与访谈结果显示,经反思干预后,多数参与者认可 AI 在结构化任务中的辅助作用,同时能主动对创新性等主观维度的 AI 建议进行甄别,不盲从、不机械采纳,实现“理性信任—批判采纳”的动态转化^[17, 32]。这一协作闭环对未来自适应智能系统的人机协作提供了范式支持。

此外,本研究强调了人机协作模式高效落地的三大前提条件:

AI 输出的透明性和可解释性。只有在 AI 建议附有明晰的逻辑依据(如关键字抓取、因果链条分析、数据相关性说明等)时,专家才能有针对性地判断其合理性和边界。研究过程中,所有 AI 评分建议均配以结构化摘要,便于评审员迅速定位与核查,提高了整体流程的透明度和可控性。未来 AI 系统需持续提升解释性与可追溯性,以满足专家级任务对可解释 AI 的刚性需求。

评审员的批判性思维和元认知能力培养。实验结果显示,经过定向培训,评审员能够明确区分哪些任务环节应主动采纳 AI 建议,哪些环节则需结合专业知识深度质疑。通过反思问卷、案例模拟等机制,提升评审员对自身决策过程的意识和自我调节能力,有效防止“自动化偏差”、“锚定效应”等认知误区。特别是在创新性、重要性等高风险环节,批判性与元认知能力是人类不可替代的“最后防线”。

流程设计中对认知负荷和时间成本的精细管控。实验二发现,盲目在所有维度施加反思虽能提升评分准确性,但也极大增加评审时间与心理负担。定向反思组通过限定反思范围与时间窗口,既实现了与全面反思组相当的质量提升,又最

大程度避免了认知疲劳和流程冗余,展现出最优性价比。说明“分层—分步—定向”是高负荷复杂任务下人机协作机制优化的关键策略。

基于上述机制,本研究进一步展望 AI—人类协作在各类知识型工作场景的广泛适用性。以学术评审流程为例,人机高效协作不仅能显著缩短论文发表周期、缓解专家资源紧张,还能提升同行评审的透明度与科学性。通过智能化反思工具和自动日志监控,期刊编辑部和主编可动态追踪评审各环节的质量指标和效率瓶颈,及时调整任务分配与干预强度,实现过程可视化管理。此外, AI 还可持续根据专家反馈调整和优化自身算法,实现“协作—学习—再优化”的自适应闭环,推动下一代 LLM 系统的“学科定制化”与“专家协同化”升级。

在医疗、法律、金融等领域,类似的动态协作机制同样具有高度推广价值。例如,医疗诊断可采用“AI 初诊—医生复核—针对性反思—终审定稿”模式,提升诊断效率与安全性;法律合同审核可采用“AI 自动标注—律师初审—反思校对—最终定稿”,显著提高合同审核的准确度和一致性。未来,随着大规模人机协作数据的积累,反思触发机制、AI 提示策略和评审员行为模式将持续进化,推动“自适应协作系统”向更高水平发展。

总的来说,本研究以学术评审为实验平台,系统验证了 AI 与人类基于优势互补与动态协作的绩效提升路径。优势互补确保 AI 承担结构化、规范化任务,人类专注高阶判断与创新思维;动态协作实现了 AI 与专家的交互闭环,显著提升了复杂任务的质量与效率。未来,结合领域实际需求和协作场景特征,进一步完善 AI 的可解释性、提升专家元认知水平和优化反思流程,将为智能时代高负荷知识型任务的人机协作奠定坚实基础^[17, 32]。

6.2 理论贡献

6.2.1 对人机协作理论的扩展

本研究以学术同行评审这一典型知识密集型任务为核心,融合实验一与实验

二的实证发现,对现有人机协作理论进行了系统性的扩展与修正。传统人机协作理论多强调静态分工,即将例行化、规则化的子任务交给机器,而主观性、创造性的环节则由人类完成^[28, 29]。随着大语言模型(LLM)技术在认知与自然语言处理能力上的迅速提升,原有的二元分工模式已难以全面反映人机融合的新现实。基于两轮实验数据,本节围绕“交替反思”机制、动态交互的增效、任务阶段化协作的细化,以及人机角色动态调整四个方面,对人机协作理论展开深化阐释。

首先,实验二通过“定向反思”与“全面反思”的设计,揭示了交替反思在人机协作中的重要意义。具体而言,在AI于创新性、论证逻辑等主观维度表现薄弱时,通过“人类初评—AI评分—反思再评”流程,可启动评审员的系统2(理性分析)思维,有效提升评分质量(见4.2.1节^[17])。基于此,本文提出将“反思循环圈”纳入人机协作理论:在关键维度通过互动式流程引导人类多轮思考,从而校准AI与人类判断之间的偏差。与传统理论假定人类在接受AI建议后即直接决策不同,本研究强调多轮互动式反思(如“AI提示→人类初评→AI评分→人类反思复评→AI再辅助→人类最终定稿”),不仅提升了评审准确性,也为理论体系引入了“交互反思”这一新要素。

其次,实验结果进一步验证了人机协同的动态增效机制。传统理论通常以“分工—整合”为主,突出机器擅长高效、自动化处理,人类侧重复杂创造性判断^[30, 33]。然而,实验一证实,LLM在“内容丰富性”、“整体评价”等可量化维度已具备较高准确率,并能显著提升评审效率(详见3.3.4节)。在此基础上,实验二显示:当人类在AI弱项维度引入反思,既弥补了AI的盲区,又能兼顾高效与高质。例如,组2在时间成本仅比对照组略有增加的情况下,实现了与全维度反思组接近的评分提升。上述发现表明,人机协作不应仅限于静态加总,而要充分发挥动态交互优势——根据不同阶段、不同维度的任务需求,灵活调整人机参与深度,实现整体绩效的非线性增益。

第三,本研究对任务阶段化协作框架进行了更为细致的刻画。以往相关理论

多以“事前规划—执行监控—事后评估”为主线^[40]，而学术评审流程的高度结构性与创造性要求对阶段划分作进一步细化。实验一表明，LLM 在“初步筛选”、“客观维度评估”阶段表现突出；实验二则通过在“分维度评估”环节引入反思机制，显著提升了主观复杂维度的评分质量；在“整体校正”阶段，AI 与人类反馈的整合进一步优化了评分一致性与信任基础（见 4.3.2 节）。据此，人机协作理论亟需从宏观分工向微观阶段化拓展，在关键节点设置多元协作策略（如 AI 预判→人类介入→反思循环圈→校正整合→输出决策），为各类任务提供更加细致、可操作的模式。

第四，实验二进一步说明了人机角色的动态调整特征。在传统分工理论中，人类与 AI 的角色相对固定：人类负责最终判断，AI 工具为辅助；或 AI 执行基础任务，人类完成终审^[28]。本研究发现，评审员在不同维度与阶段可于系统 1（快速直觉）与系统 2（深度分析）之间灵活切换。AI 对客观维度的实时提示可激发系统 1 形成初步印象，而反思环节则唤起系统 2 的深入审视，终稿阶段实现两种模式的融合。基于此，本文主张在理论中增加“角色动态转换”模组，强调人类与 AI 的功能定位应依据任务复杂性和认知状态实时调整。AI 可在不同节点担任“信息推送者”、“校正合作者”、“知识辅助”等角色，人类亦可于“直觉决策”、“批判分析”、“最终仲裁”之间自如转换。

最后，基于上述分析，本文对现有理论作出两点创新性修正：其一，提出“反思循环圈”机制，解释针对 AI 弱项维度的多轮思考为何能突破静态分工带来的性能瓶颈；其二，提出“选择性交互”策略，强调在流程设计中根据任务特性灵活调度反思时机与频次，实现效率与质量的平衡。上述理论扩展不仅丰富了传统分工与协同模式，也为构建高适应性、多场景可迁移的人机协作系统提供了理论基础与实践指引。

6.2.2 双过程理论的应用

双过程理论(Dual Process Theory)将人类的决策与认知划分为两套系统：

系统 1 代表快速、直觉、自动化的思维过程，系统 2 体现缓慢、分析、需认知努力的推理活动^[17]。本研究围绕学术评审中的人机协作设计，结合两项实验的实证结果，系统剖析了 AI 辅助评审流程下，系统 1 与系统 2 的协作机制。以下从四个方面展开：AI 建议驱动下的系统 1 快速直觉、反思环节中系统 2 的深度思考、系统 1/ 系统 2 的情境切换机制，以及双过程模型在本研究中的理论扩展。

(1) AI 建议激发系统 1 快速直觉

实验一和实验二的最新结果显示，在内容丰富性、整体评价等相对客观和结构化的评审维度上，大语言模型（LLM）辅助所提供的建议与专家评分高度一致。根据表 9、表 16 等表格中的定量分析，GPT-4 在这些维度上的评分均方差显著低于人工独立评分，平均误差均小于 0.3 分（5 分制），且相关系数达到 $r=0.85$ ($p<0.001$)，表现出高度的评分一致性。实验流程中，评审员在看到 AI 建议后，大多会迅速采纳 AI 给出的结论，进而作出评分判断。行为数据和自评问卷结果表明，这一阶段主要依赖系统 1 的快速直觉思维，体现为低认知负荷和高效率（见 3.3.4 节、4.4.1 节）。实验数据显示，采用 AI 辅助后，评审平均用时显著下降，整体评审时间较对照组缩短约 30%；主观认知负荷分值亦略低于传统人工评审，尤其在重复性和程序化任务环节，评审员表示“AI 建议为判断提供了便捷锚点，有效减轻常规评估负担”^[16, 28]。由于 AI 评分常作为初步判断的锚点，评审员评分分布趋于集中，主观分歧减少，评分一致性提升^[17, 29]。

但需要注意的是，这种基于系统 1 的快速采纳机制在 AI 表现薄弱的维度（如创新性、论证逻辑等）存在潜在风险。实验一和实验二数据显示，如果评审员过度依赖 AI 建议而缺乏自主甄别，容易忽视 AI 模型在复杂学术判断和深层推理任务中的不足，出现误判、遗漏论文关键问题等情况^[47, 58]等。因此，在这些主观性和高难度维度，单纯依赖系统 1 不足以确保评审质量。

(2) 反思环节唤起系统 2 深度思考

针对系统 1 的局限性，实验二在 AI 薄弱维度引入“反思干预”机制，即通

过先独立初评、后显示 AI 评分,并辅以引导性反思问卷,启动评审员系统 2 的深度分析过程(见 4.2.4 节、4.4.1 节)。量化结果显示,这一流程能显著提升创新性、论证逻辑等复杂维度的评分准确性。以创新性与重要性维度为例,组 2(定向反思组)在引导反思后,评分与专家基准的偏差由初评时的 1.28 分下降至复评时的 1.11 分,平均改善量显著($p<0.05$);而对照组和优势反思组在该维度未见显著改进(见 4.4.4 节)。反思环节设计的问题(如“AI 评分与您初评结果有何主要差异?请重新梳理判断依据”)有效促使评审员放慢节奏,主动检视 AI 与自身判断的分歧,从而抑制直觉采纳偏差,激发系统 2 的批判性思考^[17, 32]。

与此同时,数据也显示系统 2 的介入伴随一定的效率和认知成本提升。实验二组 4(全维度反思组)在五个评审维度都设置反思,评审平均用时较无反思组延长近 40%,认知负荷自评分值提升约 12%,部分评审员反馈“反思流程冗长、导致注意力分散与疲劳”(见 4.4.2 节)。因此,结合反思干预的实际效果与负荷成本,实验二进一步验证了“定向反思”——即仅在 AI 薄弱维度设置反思——可在保证评审质量提升的同时,控制时间与认知成本,实现效率与准确性的最佳平衡。

总的来说,基于本书两轮实验的实证数据,AI 建议通过锚定作用主要激发系统 1 的直觉判断,适用于结构化与客观性任务;而在高主观性与复杂判断任务中,通过反思环节精准启动系统 2 的深度推理,可有效提升评分准确性。这一发现不仅丰富了双过程理论在知识型人机协作场景下的应用路径,也为 AI 辅助复杂任务设计反思机制和优化人机任务分配提供了理论与实践支持^[17, 32]。

(3) 系统 1/ 系统 2 切换的情境机制

本研究进一步考察了影响评审员选择系统 1 或系统 2 的关键情境因素:

① AI 输出可靠性:在 AI 表现优良(如内容丰富性、整体评价)时,评审员更依赖系统 1,减少深度分析;而在 AI 误差较大(如创新性、逻辑推理)时,更

愿意接受反思干预，启动系统 2；

②任务负荷与时间压力：任务繁重或时限紧张时，系统 1 的自动处理优势突出；当任务重要性提高或有充足时间时，系统 2 的介入概率提升；

③用户信任度：对 AI 信任度高时，更倾向于自动采纳建议（系统 1），信任度低时，则主动对 AI 评分进行复核（系统 2）；

④反思机制设计：反思表单的提示问题深度直接影响系统 2 启动的充分性，开放型问题较难调动深度思考，针对性强的问题更易促进系统 2 分析。

这些因素共同作用，动态调节人机协作流程中系统 1 与系统 2 的切换路径，为不同场景下的 AI 辅助评审提供了策略指引——可量化、标准化的环节采用“AI+ 人类快速复核”，复杂判断任务则采用“AI 建议—初评—反思复评”流程，实现效率与质量的平衡。

（4）对双过程理论的应用与拓展

本研究不仅在实证层面验证了双过程模型在人机协作评审中的适用性，还对其理论内涵作出重要补充。一方面，针对 AI 能力局限的主观维度，通过定向反思干预，可系统性地启动系统 2，实现人机协同的最优结构。另一方面，通过对切换条件、流程设计的系统梳理，丰富了双过程模型对 AI- 人类互动中情境依赖性的理论解释力。未来，在更大规模、多场景知识型任务中，可进一步探索系统 1/ 系统 2 协同机制的通用性、动态性及其对高效智能评审体系的支撑价值。综上所述，双过程理论为理解和优化 AI 辅助学术评审中的人机协作机制提供了坚实理论基础。本研究通过实证数据系统展示了两类认知过程的互补与切换条件，为提升评审效率与质量、设计更智能的知识工作系统提供了理论指导与实践依据。

6.2.3 双过程理论的修正与展望

结合本文实验结果，针对双过程理论在 AI 辅助决策领域，提出以下修正与扩展：

(1) 引入“阈值触发”机制

传统双过程模型未明确界定系统 2 介入的具体条件。基于实验结果,本文提出“反思阈值”概念:当系统 1 的判断与 AI 建议之间的差异度(即 $|\text{初评} - \text{AI 评分}|$),或与专家基准的偏差超过预设阈值时,系统自动触发系统 2 介入,由评审员进行深度反思和复评。此机制不仅提升了高风险环节的决策准确性,也有效控制了认知资源的分配,避免在低风险维度上出现无谓消耗。

(2) 提出“交替思考”模式

实验发现,系统 1 与系统 2 的运作并非单次切换,而是在多轮人机交互中持续交替,形成“思考螺旋”。在初评、AI 建议、反思复评等环节,评审员会根据任务复杂度与信息反馈反复在直觉判断与深度思考之间切换。这一“交替思考”模式不仅优化了学术评审流程,还可适用于医疗诊断、法律审查等其他需要人机协作的高风险、高复杂度领域。

(3) 强调“信任校准”过程

本文将双过程理论与信任动态机制结合,强调系统 1 快速采纳 AI 建议反映出较高的信任水平,而系统 2 则在必要时对信任水平进行动态校准。实证数据显示,在 AI 表现不佳的维度,系统 2 的校准作用对于防止因盲从 AI 建议而产生的决策偏差至关重要。信任校准不仅体现在对单条建议的复核,也表现在对整个 AI 系统能力边界的持续调整。

(4) 扩展至多智能体协作场景

进一步考虑在团队或多 AI 系统情境中,各主体(人类/AI)可动态切换系统 1 或系统 2 的角色定位,形成分层、递进的决策网络。这一理论扩展为多主体协作协议的设计提供了基础,未来可在大型评审团队、跨学科专家网络及多模型集成系统中,探索如何高效触发反思与优化认知资源分配。

6.3 实践意义

6.3.1 实际应用建议

为推动人机协作成果在学术评审领域落地,建议构建以大语言模型(LLM)为核心、融合人类专家深度参与的智能协作平台。在平台设计上,可由 AI 负责论文初步筛选、结构与语言规范性检查、基础问题识别等标准化环节,将初审合格的稿件推送至人类评审员,从而大幅节省人工的机械性操作时间,并提升筛选一致性和整体效率。

平台还应集成智能化反馈生成功能,使 LLM 针对论文在表达、方法、数据等维度自动生成改进建议。人类评审员在“内容丰富性”、“表达与写作”等可量化维度遇到不确定情形时,可迅速参考 AI 建议并结合专业判断作出决策。对于如“论证逻辑”、“创新性与重要性”等主观性更强的环节,系统可在 AI 给出意见后自动弹出反思提示,要求评审员回顾自身初评与 AI 评分的差异,撰写思考过程,并提交复评结果。这样不仅激发了系统 2 的深度参与,也在关键环节生成可追溯的思考日志,为后续质量监控与系统迭代提供数据基础(见第 4 章)。

为防止全维度强制反思带来的认知负荷与效率损失,建议结合 AI 表现与任务复杂度动态调整反思频率和深度。具体做法是:AI 在“内容丰富性”、“整体评价”等环节表现稳定、偏差低时,可简化反思为快速校准;而在“论证逻辑”、“创新性与重要性”等 AI 易失误维度,则需高频、深度反思,确保关键环节获得最大质量收益(见 4.4 节)。

此外,应针对评审员开展系统培训,内容包括 AI 工具使用规范、反思机制操作流程及案例演练,帮助其掌握何时应信任 AI、何时需批判性质疑。平台也应开设实时反馈通道,持续收集用户体验数据,优化提示设计和交互流程,提升整体易用性与信任度。

在实际推广上,建议通过试点或 A/B 测试方式,将上述人机协作流程应用于期刊或会议评审中,并根据评审时长、评分一致性、评审员满意度等指标评估成效。若 AI + 定向反思模式在效率、准确性与满意度方面均表现优异,可进

一步推广至大规模学术出版、医疗、法律等知识密集型领域，实现在保障专业深度的前提下流程的高效与可扩展。

6.3.2 对学术出版行业的影响

本研究揭示的 AI 辅助评审模式对学术出版行业带来了深远且多维度的影响。首先，AI 在初筛环节的引入，使期刊编辑部能够在处理大量来稿时，高效筛选出符合基本质量标准的论文，显著缩短初审周期。依托自然语言处理技术，AI 工具能够自动检测稿件的结构完整性、语言规范、图表与参考文献格式，及时剔除存在抄袭或严重格式问题的稿件，并对研究问题的表达清晰度进行初步评估。由于这些工作长期以来依赖人工投入且易受主观差异影响，AI 的应用不仅提升了早期筛选过程的一致性与客观性，还有效减轻了编辑部在初审阶段的负担，使人力资源得以向更高价值的深度评审环节转移。

其次，AI 驱动的智能期刊匹配系统为作者与期刊之间的精准对接开辟了新路径。通过对论文关键字、摘要内容和历史评审数据的量化分析，系统能够评估稿件与不同研究领域及期刊目标读者群体之间的契合度，并结合期刊的接收率、审稿周期等参数，智能生成个性化投稿建议。这一功能有效减少了因领域不符而导致的多次退稿和重复投稿现象，帮助作者更高效地锁定最合适的发表渠道，缩短整体出版周期。同时，平台可实时监控稿件在各期刊间的流转状态，自动推送关键节点提醒，进一步提升了流程的透明度和可控性。

在评审质量提升方面，AI 工具通过持续学习历史评审数据，不断优化评分模型，为评审员在“内容丰富性”、“表达与写作”等可量化维度提供客观参考，显著降低了人为主观偏差。对于“论证逻辑”、“创新性与重要性”等高主观性维度，AI 不仅为初评提供辅助意见，还通过反思引导，激发评审员开展深度复核，进一步校准评分，使结果更贴近专家基准。这一机制有助于提高评审一致性，强化学术价值评估的严谨性。随着系统累积的真实评审反馈不断丰富，AI 模型将在不同领域、不同类型的论文评估中持续提升自身精度和适应性。

最后，AI 技术在学术出版流程中的广泛应用，推动了出版机构的组织架构

和工作模式变革。编辑与评审员逐渐从大量初筛与格式校验等机械性事务中解放出来，将更多精力投入到高层次的学术把关与创新判断。出版机构还可基于 AI 平台汇集的海量评审数据，对评审质量与流程效率开展定量分析，为管理决策和流程优化提供科学依据，形成闭环改进。可以预见，伴随 AI 技术不断成熟与普及，学术出版将进一步迈向高效、透明和专业的运行模式，为全球科研成果的快速传播与共享提供坚实支撑。

6.3.3 管理与政策建议

在充分认识到 AI 工具在学术评审流程中兼具提升效率与存在算法偏差及过度依赖风险的前提下，学术出版机构及相关管理部门亟需建立系统化的管理与政策框架，以保障 AI 工具的安全、可靠与可持续应用。结合本研究实证结果，建议从以下四个方面推进：

首先，应在风险管理与合规审查方面建立覆盖技术、数据和伦理全流程的审查机制。技术层面需针对模型训练所依赖的数据集定期进行偏差检测，识别可能导致评分失衡的隐含因素，特别是在高主观性评审维度（如论证逻辑、创新性与重要性）采取更为严格的人工复核与“反思触发”策略。数据层面要求对评审过程中涉及的未公开研究数据及作者隐私信息进行脱敏与加密，并对访问权限实施分级管理。伦理层面则需明确算法责任归属，一旦出现系统性偏向，需及时启动追责与纠正机制。可通过分级评估，将 AI 应用场景划分为低、中、高风险区，在高风险区集中部署“AI 初评 + 定向反思 + 人工终评”流程，从而减少对 AI 建议的单向依赖或完全否定，避免两极化风险。

其次，建议采用分阶段引入和持续优化的策略，确保 AI 工具在验证安全性与有效性后逐步扩展其应用职责。第一阶段可仅限于格式检测、语法校对等基础任务，并以脱敏历史数据检验模型稳定性；第二阶段扩展至内容一致性和方法论初筛，仍需保留人工评审员对薄弱环节的反思介入；第三阶段则可将 AI 应用于整体评价等相对成熟场景，并始终配合“定向反思”或“全维度反思”机制，以便动态监测各维度表现并调整协作策略。每一阶段均应设定关键绩效指标，包括

评分改进幅度、认知负荷测评等，以支撑模型参数和流程的持续优化。

第三，专业培训与信任培养应贯穿于 AI 应用全过程。培训内容不仅需覆盖 AI 工具的基础操作，更要详尽阐述模型在各评审维度的局限性，重点讲解“论证逻辑”、“创新性与重要性”等高主观性维度下如何触发“定向反思”以校正 AI 建议。通过案例研讨和模拟练习，帮助评审员体验并熟悉最佳性价比的协作模式，强化对“定向反思”模式优势的认知。制度上，可建立“用户顾问”机制，邀请在“定向反思”模式下表现优异的评审员参与工具优化与培训设计，依托实践经验不断完善培训体系，增强团队整体信任度。

最后，数据安全与伦理合规是保障 AI 长期可信运行的根本。需严格遵循《个人信息保护法》和国际同行评审伦理准则，对论文原文和评审数据实施分级加密与访问审计，确保系统更新和模型迭代均在可追溯的审计日志下运行。AI 输出的每一次评分建议均应自动记录反思触发条件与评审员修改意见，便于后续回溯和偏差检测。建议定期邀请外部第三方开展伦理与安全评估，及早发现潜在算法偏差或数据泄露风险，并通过法律与技术手段及时修正。唯有在技术、流程、培训与监管等多维协同推进下，方能实现 AI 辅助评审的高效、公正与可持续发展。同时建议在平台侧按「初筛自动化 → 定向反思触发 → 人工复核与冲突仲裁 → 版本留痕与回放」四步落地：首先由 LLM 完成结构化体检（版式／引用样式／缺页检查）与客观维度初评；其次基于「AI 强弱项—偏差—负荷」三维阈值智能触发定向反思（只推送必要且可操作的检核点）；第三步将关键分歧点提交资深审稿人复核与仲裁，并强制证据链对齐；最后生成可回放的审稿版本与依据留痕，服务质量监督与教学反馈。配套上，需落实数据治理与审计日志、偏误监测仪表盘（如新颖性低估预警）、以及与期刊管理系统的接口对接，以确保在效率提升的同时，维持可追溯、可问责与可迭代优化。

6.3.4 用户体验与满意度提升

评审员对 AI 辅助系统的信任与满意度，主要受系统在评审流程中的透明度、可解释性以及人类评审员互动便捷性的影响。基于此，本研究提出多项举措，

支持实际应用中不断优化用户体验并提升满意度。

首先，应为每一项评分建议提供明确且详尽的依据与逻辑解释。在生成评分时，LLM 应自动标注所参考的原文片段及对应维度，并配以简要说明。例如，“本条建议基于第 2 段的论证结构在逻辑连贯性上存在断层”或“引用了图表 1 中的数据，但未明确数据来源”。这样做有助于评审员快速了解 AI 关注的要点，并判断建议的合理性，从而避免盲目采信或彻底否定 AI 建议。

其次，建议出版机构公开算法原理和模型设计思路。可通过用户手册、线上培训或技术报告，系统性地介绍核心算法流程、训练数据结构及评分标准，使评审员对 AI 内部运行机制有基本认知。只有当评审员理解 AI 如何通过关键字检索、向量计算和跨维度分析输出建议时，他们才能有信心将部分重复性任务交由系统处理，并在关键判断环节保持理性监督。

第三，建议建设互动式解释平台，支持评审员随时查询 AI 评分细节。例如，系统可为每一维度提供“理由查看”按钮，点击后展示分数变化趋势、与专家基准的差异分析，以及其他评审员常见的修改意见。通过这种层级化、可追溯的解释，评审员能够快速获取全局信息，也可深入分析 AI 建议的来源与逻辑，增强决策的自主性。

第四，建议针对不同评审员角色与经验定制反思引导。新手评审员在初次接触系统时可获得更为细致的引导，如“请说明为何方法论部分论证尚不充分”；经验丰富的评审员则可只在与 AI 建议分歧较大时接收反思提示，例如“基于哪些证据，您认为初判与 AI 建议存在偏差？”通过差异化的反思提示，既保证整体效率，又激发各层次评审员的批判性思维。

第五，建议建立实时反馈与动态调整机制。系统在检测到某评审员与专家评分或 AI 结果存在较大偏差时，可及时推送“偏差警示”，并提供参考文献片段或典型案例，帮助评审员及早识别和修正可能的误判，降低后期集中修改带来的额外工作量。

第六,建议系统自动生成个性化评审报告。每位评审员可获得涵盖各维度评分轨迹、与 AI/ 专家基准对比、常见偏差类型、改进幅度及个性化学习建议的绩效报告,辅以可视化图表,帮助其在后续评审中有针对性地提升专业水平。

第七,建议鼓励持续的用户参与与反馈。出版机构应定期组织焦点讨论与线上调查,收集评审员对系统界面、功能及反思引导的建议,并建立快速回应的迭代机制,将用户反馈及时转化为产品优化举措,增强用户的归属感与忠诚度。

通过上述多维度优化措施,既能让评审员对 AI 系统的评分流程、算法逻辑与反馈机制形成清晰认知,也有助于提升评审效率、激发自主判断和批判性思维。最终,人机协作将实现从“AI 工具辅助”向“AI 与人类专家交互迭代”的跃迁,不仅巩固人类专家在复杂决策中的主导地位,也充分发挥 AI 在初步分析和提示环节的优势,推动学术评审流程朝着高效、公正与可持续方向发展。

6.4 研究不足与未来展望

本研究在探索人机协作模式及反思机制对学术评审成效的影响方面取得了系统性成果,但仍存在一定局限性。结合当前研究基础,未来还需在以下方向加以拓展和深化,以持续推动人机协作理论与应用的发展。

6.4.1 研究局限

尽管本研究通过两轮随机对照实验(实验一邀请 26 名专家和 89 名参与者,共 115 人,实验二共 144 名参与者)较为系统地考察了 AI 辅助同行评审的实际效果,但仍存在若干值得关注的局限性,在解读结果与推广应用时需谨慎对待。

我们观察到在高认知负荷情境下(如主题陌生、方法复杂、语言晦涩),AI 辅助的边际收益显著下降。其可能机制包括:其一,资源占用/决策疲劳——审稿人可用于「整合 AI 建议」的认知资源不足,导致模型输出难以被有效吸收与转化;其二,启发式依赖加剧与自动化偏见——高负荷下个体更依赖简化规则(如“AI 已检过”或“文字清晰即为好”),削弱对 AI 建议的二次检视与反例搜寻。

此发现提示：在高负荷任务中，平台宜采取分步化任务拆解、可视化证据链、逐条核验与必要时延后决策等人机互动设计，以维持实效并降低偏误（见图 5 之交互效应）。

另外，样本规模与代表性的限制仍然突出。实验二虽然通过较大样本($n=144$)一定程度上增强了结论的信度，但仅涉及 144 名参与者，难以覆盖现实学术评审群体的复杂结构与多样性。参与者主要来自计算机科学、信息系统等知识密集型领域，学科代表性较为有限，对跨学科适用性构成一定限制。此外，参与者的地理和文化背景集中于国内一线高校和科研机构，尚未涵盖不同国家或地区的多样化评审习惯和标准，导致结论在跨文化、跨区域推广时存在一定不确定性。受试者以硕士／博士研究生为主，符合我们聚焦的「初筛」型评审场景与平台业务分工；在计算机科学领域，博士生参与会议审稿亦较常见。然而，该样本结构仍限制了向更资深领域专家与跨学科编辑实践的扩展。未来工作将扩展至多学科与多层级审稿群体，并与期刊编辑流程联动以验证机制鲁棒性。

本研究之评审材料均为英文撰写且集中于计算机科学领域，这与主流 LLM 训练优势区域高度重合，因而可能高估其在其他语言与学科的普适表现。未来研究需在多语种与跨学科语料上重复验证，并评估不同领域写作范式差异对「技术有效性／清晰度／创新性」维度判断的影响。

其次，实验条件与真实审稿环境之间仍存在一定差异。本研究为确保变量控制，在 AI 工具的呈现方式、界面设计及评分维度设定等方面进行了必要简化，使实验流程标准化、可比性较高。但在实际同行评审中，评审员往往需要应对全文细致阅读、引用核查、伦理合规、图表数据验证等多重任务，多任务并行的实际工作环境难以在实验情境下完全复现。因此，实验场景下的认知负荷与实际工作环境中的认知压力仍存差距，后续研究需在更贴近业务的线上系统或现场的嵌入式测试中验证 AI 辅助评审的应用成效。在审稿场景中，大语言模型可能出现自我偏好偏见（对风格／分布接近自身生成的文本赋予更高评价）与幻觉（生成并不存在的事实／引文）等风险。为此，平台侧宜采取「证据溯源＋引用校验

+反例提示」的多重约束,并将模型输出限定为「建议性意见」,由人工最终裁量。

第三, AI 工具本身的技术局限依然明显。虽然本研究选用的 GPT-4 模型在内容生成和基础判断层面表现出较高水平,但在高阶论证逻辑、研究创新性等主观复杂维度上的理解与评判能力仍有限,容易产生评分偏差。同时, AI 训练数据中可能包含学科倾向性和区域文化偏见,模型在面对跨学科、地域性主题或伦理敏感问题时,存在误判和片面判断风险。此外,当前 AI 系统在可解释性方面尚有不足,评审员通常只能看到 AI 给出的最终分数,而难以追溯其评分逻辑和推理过程,这一定程度上影响了评审员对 AI 建议的信任及人机协作的可控性。

第四,评估指标体系与方法论仍有待完善。本研究主要采用量化指标(如 DIFF、IMPROVE、配对 t 检验、ANOVA)来评估反思干预对评分的影响,但这些指标难以全面刻画学术论文的深层次价值。例如,在创新性与重要性评估中,单纯依靠数值偏差难以捕捉原创性和学术影响力等核心特质。此外,统计显著性不等同于实际应用效应,部分指标虽然在 $p<0.05$ 或 $p<0.1$ 的统计意义上达标,但效应量有限,对实际评审决策影响不大。未来研究可进一步结合质性访谈、焦点小组和案例分析等方法,补充量化研究的不足,提升结论的深度和解释力。

第五,认知与行为偏差可能对结果产生影响。在实验环境中,参与者可能受到“新奇效应”或对研究预期的猜测影响,从而调整自身评分策略或对 AI 建议表现出过度关注。在反思环节,评审员出于表现动机,也可能有意识地调整对 AI 建议的采纳程度。这些认知与行为偏差在自报告问卷和系统日志中难以被完全捕捉,导致数据解释潜在干扰。

最后,本研究未能解决部分关键性实际问题。例如,目前尚无法评估 AI 辅助模式在双盲审稿场景下的可行性,也未系统探讨学科极度细分领域的 AI 适应性和定制化需求。未来需进一步在真实双盲审稿环境中检验 AI 工具的相容性与隐私保障机制,同时针对小众学科和交叉领域论文,探索 AI 模型的本地化训练与微调策略。

综上所述,尽管本研究在样本设计、实验流程及统计方法等方面做出了较大

努力,仍存在上述多重局限。充分认识和正视这些问题,有助于未来进一步完善实验方案,扩大样本多样性,并通过情境化、质性结合的方法提升 AI 辅助人机协作评审的科学性、可靠性和普适性。

6.4.2 未来研究方向

基于本研究的发现与现有局限,未来的研究方向可从以下几个方面加以拓展和深化:

(1) 扩大样本规模与提升多样性

为进一步增强研究结论的代表性与外推力,未来实证研究应在学科背景、机构类型及文化语境等层面大幅扩展样本范围。当前实验的 26 名专家与 144 名评审者虽已在统计上呈现出反思干预的显著成效,但参与者主要集中于计算机科学与信息系统等领域,未能覆盖生命科学、医学、经济学、教育学、文学、历史等多元学科。不同学科在研究范式、评价标准及论证偏好等方面均有显著差异。例如,生命科学更重视实验设计与数据可靠性,社会科学则注重理论结构和田野调查的严谨性,人文学科则往往强调论证深度与表达风格。未来研究应优先纳入各学科的资深审稿人与新晋评审,系统比较 AI 辅助及反思机制在多学科场域中的适用性与边界,进一步检验定向反思(实验组 2 模式)是否能在多领域实现低成本高改进的效果。

此外,评审流程在不同机构和出版平台间存在较大差异。当前主流期刊、出版社、高校学报及学会会议往往采用不同的评审模式,包括严格的“双盲”评审与开放式同行评议等。不同审稿流程不仅影响评审员对 AI 建议的信任基础,也决定了反思环节能否有效触发。后续研究可与多家出版单位、学术期刊和编辑平台深度合作,将定向反思与全面反思机制直接嵌入线上审稿系统,实地评估 group_02 “仅反思 AI 弱项维度”策略在双盲审稿、互动式评议等不同流程中的实际可行性及风险。通过跨机构、跨流程的嵌入式实证,研究可更全面地揭示各类反思机制在多元学术管理环境下的收益与局限。

文化与语言背景差异亦对 AI 辅助评审成效产生深远影响。不同国家和地区的评审员在 AI 工具的采纳度、信任倾向和批判习惯上存在显著异质性。在某些文化环境中,权威主义思维较强,评审员更易接受 AI 评分结论,而在其他文化中,个体主义和批判性思维更为突出,反思机制可能更容易启动。未来研究可在亚洲、欧洲、美洲、非洲等多个地区开展跨文化对比试验,结合多语言 AI 模型,深入考察语言歧义、翻译误差对评分表现与反思触发的影响。只有在真正多元的文化、语言样本中开展实证分析,才能确保 AI 辅审及反思机制的全球适应性和公平性。

最后,不同社会经济条件下资源可及性的差异也应成为未来研究的重要关注点。高资源学术平台有条件部署和维护大型 LLM 系统,而资源有限的机构则更关注模型的成本效益。后续研究可在低成本或开源 AI 模型基础上,检验定向反思策略(如 group_02)在低算力环境下的实际效能,并与高性能模型进行横向比较。通过面向不同预算和基础设施的应用评估,未来研究可为发展中地区和小型学术出版机构提供切实可行的 AI 辅助评审路径,推动 AI 技术在全球范围内的公平普及与可持续发展。

(2) 探索其他领域的人机协作

探索其他高复杂性和高专业性领域的人机协作,为 AI 技术的广泛应用开辟了新的空间。在法律分析领域,AI 工具在法律文献检索、案例分析与条文解读等繁复工作中已表现出明显的辅助价值。借助 AI 的高效信息检索与数据处理能力,法律从业者能够显著提升查找相关法条与判例的效率,从而优化工作流程。然而,法律决策所依赖的不仅是法条与先例,还需结合对社会环境、司法实践与伦理原则的综合理解。因此,单靠 AI 工具难以胜任最终法律判断。推动法律分析中的人机协作,需将 AI 优化与反思机制相结合:一方面,提升 AI 工具在法律语境下的推理能力,增强其对本地法律体系与案例语境的适应性;另一方面,设计有针对性的反思流程,促使法律专家在采纳 AI 建议前进行批判性复核和补充。例如,在法律案例推理中,AI 可能遗漏关键判例或误判条款适用,通过反思机制,

法律专家不仅能够及时纠偏，还能进一步深化案件解读。未来研究可通过模拟法律决策场景，系统考察 AI 在不同类型法律任务中的适应表现，并持续收集法律专家的反思反馈，用于迭代优化 AI 工具及协作模式。

医学诊断同样是人机协作深入发展的前沿领域。AI 在医学影像识别、疾病风险预测等任务中能够快速分析大规模数据，为医生提供初步筛查与辅助决策依据。例如，AI 模型通过分析大量医学影像，能够敏锐捕捉疑似病变区域，为医生节约时间、减少漏诊风险。这一能力对提升初筛效率具有突出作用，使医生能将更多资源投入复杂病例与多学科协作环节。然而，医学诊断本质上具有高度不确定性，涉及患者个体差异、疾病进展多样性及治疗反应差异等复杂因素。AI 在给出诊断建议后，医生仍需结合临床经验、实验数据和患者整体状况做出最终决策。例如，癌症的确诊过程不仅依赖影像识别，还需整合实验室检查、家族史及患者主观症状。为此，AI 在医学领域的应用更应强调与临床专家的协作，推动反思机制引导医生对 AI 结果进行核查与个性化调整。未来相关研究可在真实临床场景中检验 AI 辅助诊断的可靠性，并探索通过多轮反思引导医生充分利用 AI 信息、补充关键判断，提升整体医疗决策质量。

推动人机协作在法律与医学领域的持续发展，还需关注工具适配与用户培训的协同提升。在法律分析环节，AI 算法应根据不同法域及案件类型调整推理策略，提升跨法域案件的适用性，并通过人性化界面提供直观检索路径与案例解读，便于法律专家快速获取所需信息。医学领域则需继续优化 AI 对影像与临床数据的解释能力，减少误判风险，并提升模型的可解释性，让医生能够理解 AI 建议的推理基础，从而增强信任感。与此同时，面向法律专家与医生的系统化培训至关重要。培训应聚焦于 AI 工具的实际操作、常见局限与风险防控，通过案例演练提升用户发现并修正 AI 错误的能力，强化反思性应用意识，确保专业人员在关键环节始终发挥主导作用，避免盲目依赖 AI 结论。

未来人机协作在法律与医学等领域的深入发展还需重视伦理监管与责任归

属。法律分析中，AI 算法可能因数据或模型设计存在偏见，对案件公平性造成影响，因此需通过多样化数据集训练与全流程责任追溯机制，保障分析结果的公正与可溯源。由于医学诊断面临患者数据隐私与安全保护的更高要求，应在严格的数据管理基础上，持续完善算法透明度和临床责任界定机制，确保 AI 建议与临床决策之间的界限清晰可辨。未来，只有在伦理规范、数据安全与系统可追溯性等多重保障下，人机协作模式才能在高复杂性行业实现高效、安全、负责任的应用，并为行业持续创新和社会福祉提供坚实基础。

（3）优化多任务协作机制

优化多任务协作机制是推动未来人机协作效率与准确性提升的关键路径。当前，评审任务的复杂性与动态变化要求协作机制能够突破传统的固定反思模式，针对具体任务需求灵活调整策略，以适应多样化的实际场景。只有在保证适应性的基础上，才能在任务难度提升或多任务并行的情境下，有效平衡评审效率与质量。因此，未来研究应聚焦于动态调整反思模式和开发智能反思触发机制，构建高度灵活和适应性强的多任务协作框架。

在动态调整反思模式方面，需要探索如何基于实时任务复杂性和评审进展，科学指导反思频率和深度的自动调整。对于难度较高或创新性要求显著的任务，系统应提升反思活动的频次和深度，促进评审员对核心问题的深层分析；而面对结构明确、难度较低的常规任务，则可减少反思环节，降低认知负担，提高整体效率。这种动态调整的实现依赖于任务复杂性评估模型。该模型需综合论文主题、方法技术要求以及评审反馈等多维度指标，对任务难度进行智能判定。系统可以自动分析文本关键字、研究方法复杂性和引言背景等内容，快速识别高复杂度与高风险任务，并针对性地调整反思策略。

同时，动态反思模式的优化可以依托机器学习方法。系统通过不断积累和分析历史评审数据，实现对反思强度、触发时机的智能决策。当检测到评审员与 AI 评分在关键维度（如论证逻辑、创新性）上存在较大分歧时，系统可主动建

议再次核查并补充理由。这种数据驱动的调整方式,既有助于降低主观评分偏差,也为复杂任务提供了持续的批判性思维支持。通过跨场景实验和大数据分析,可不断优化模型参数,增强系统在不同任务类型下的稳定性和普适性。

在智能反思触发机制的开发方面,系统需具备对“高风险”或“高复杂性”情境的自动识别与及时预警能力。依托机器学习和预测算法,系统可以综合分析论文内容、评分偏差、评审轨迹等因素,自动生成反思建议并引导评审员关注潜在问题。例如,当某一稿件在创新性维度的评分波动较大,或 AI 与人工评分分歧显著时,系统将及时推送反思提示,引导评审员进行深入讨论与修正。同时,可根据任务优先顺序配置不同反思强度,确保有限资源聚焦于关键任务,从而优化整体人机协作效能。

另外,自适应 AI 系统的发展将极大增强智能反思触发的灵活性。与传统固定算法不同,自适应 AI 可根据实时评审反馈调整参数和推理逻辑。例如, AI 可以通过持续学习评审员的调整偏好和行为特征,不断优化评分模型与反思建议,逐步实现对实际需求的更精准回应。系统还可根据任务类型和评审难度,动态启动或关闭高级分析功能,实现功能模组的按需配置和资源最优分配。这样, AI 不再只是静态辅助者,而是与评审员“共进化”的学习型伙伴,持续提升协作效率与任务适配能力。

通过整合智能反思触发机制与自适应 AI,未来评审系统将在多任务、多场景下实现高度协作与灵活分工。例如,面对复杂并行任务,系统可基于任务特征自动分配反思资源,定制个性化协作策略,并动态调整反思活动强度。在此过程中, AI 不仅协助处理常规任务,更能在关键决策节点与评审员协同进步,共同提升评审决策的准确性与可靠性。

总的来说,优化多任务协作机制的根本目标在于实现复杂评审场景下人机协作的无缝融合。未来研究应深入探讨任务复杂性评估、动态反思策略、智能触发机制与自适应 AI 系统的集成方法,为学术评审及其他知识密集型领域提供理论

指导与技术支持。这一优化机制不仅有助于提升学术评审质量,也可为法律分析、医学诊断、产品设计等多领域的人机协作创新提供有益参考。通过构建灵活高效的多任务协作框架,将进一步激发人机协作的潜力,拓展其在更广阔场景下的应用前景。

(4) 个性化反思策略的开发

个性化反思策略的开发被认为是未来提升审稿质量与效率的重要方向。不同审稿者在认知风格、思维习惯和专业背景等方面存在显著差异,因而更具针对性的反思策略有助于更好地激发审稿者的主观能动性,并提升其评审质量。未来研究可结合心理学和人机交互理论,系统探讨个性化反思策略的设计与实施,为审稿者提供契合个体需求的支持模式。

首先,根据审稿者认知风格定制反思支持,有助于为评审流程带来更高的灵活性与适应性。有些审稿者倾向于结构化、系统化的思考方式,通常在论证逻辑评估和层次分析中表现突出;另一些审稿者则在开放性、发散式思维下更易展现创造力,更适合在创新性与重要性评估或前沿性研究判断中发挥优势。针对不同认知风格,个性化反思支持可以通过定制化引导或问题设计,帮助审稿者以最符合自身特点的方式展开思考。例如,对于逻辑思维优势明显的审稿者,系统可提供系统化、分步骤的反思问题,协助其在清晰的框架下逐步深入分析;对于更偏好创造性思维的审稿者,则可采用开放式问题,鼓励其提出创新见解或独特观点。实现这一目标需要依赖科学有效的认知风格评估工具,通过建立认知风格画像,为后续反思支持策略的动态调整提供数据基础。同时,这些画像可与评审过程中采集的行为数据结合,构建动态认知行为模型,持续优化反思支持内容与形式。

在此基础上,心理学和人机交互理论为个性化反思策略的落地实施提供了理论依据。心理学研究表明,认知负荷和情绪状态对反思活动的质量与深度具有直接影响。评审任务的复杂性、时间压力及操作繁琐度均会加重审稿者的心理压力,降低其思考质量。因此,设计反思支持机制时需兼顾工作负荷与心理承受力。例

如，可采用阶段性反思策略，在初评阶段提供简化版反思引导，终评阶段再引入深度反思环节，以避免流程中过度反思导致的疲劳和效率下降。

人机交互理论则为反思工具的界面与交互方式设计提供技术指导。高效的反思支持工具应注重简洁、直观和友好的用户体验。系统可通过分步指引和实时反馈，使审稿者能够顺畅完成反思任务，而不被复杂操作界面干扰。视觉设计需突出关键信息，利用色彩对比与简洁布局提升内容可读性，方便审稿者快速定位核心问题。同时，互动式功能（如一键查看具体反思建议或调取相关数据分析）能提升审稿者对评审任务的掌控感，增强整体满意度。

此外，个性化反思策略的优化离不开动态反馈机制的支持。审稿者的行为与偏好可能随着任务推进不断变化，例如熟悉工具后更愿意自主探索开放性问题，或在实践中提出更具深度的建议。因此，反思支持系统应具备动态调整能力，能够根据实时行为数据（如评分偏差、修改频率、回应时间等）灵活调整反思内容和推送频率，为审稿者持续提供高相关性与高实效性的个性化建议。

最后，长期学习与反馈机制也是提升个性化反思策略有效性的关键。每次审稿结束后，系统可生成个性化评审报告，帮助审稿者总结表现、识别潜在偏差并获得针对性改进建议。这一机制不仅有助于审稿者自我提升，也为反思策略设计积累数据基础。通过建立循环反馈，个性化反思支持系统可持续迭代升级，实现人机协作水平的不断跃升。

（5）长期影响与技能发展的研究

随着 AI 工具在学术评审流程中的深入应用，人机协作对于评审员长期技能发展的影响日益突出。这一影响不仅涉及评审员专业能力与判断力的演变，还深刻关联其职业成长轨迹与未来发展方向。理解 AI 辅助评审如何在提升短期效率的同时塑造评审员的知识积累、批判性思维及职业素养，已成为优化人机协作模式、推动评审员持续成长的重要课题。

首先，AI 工具为评审员带来了全新的技术支持，显著提升其处理常规性任

务的效率。例如, AI 在语言规范检测、格式审核与基础数据分析等环节具备高度自动化优势, 使评审员能够从重复性劳动中解放出来, 将更多精力集中于论文的创新性、理论深度等复杂环节。这一转变带来了技能发展的双重影响。一方面, 评审员通过与 AI 工具的持续互动, 有望快速掌握先进的数据分析方法和逻辑评估技能, 并在复杂任务中不断提升判断力; 另一方面, 基础任务的自动化处理可能弱化评审员对某些能力的锻炼, 甚至形成对 AI 建议的过度依赖, 影响其专业技能的全面发展。因此, 未来有必要系统评估 AI 辅助评审对评审员知识积累与能力提升的长远影响, 分析其在促进专业成长方面的真实作用机制。

在这一过程中, 开展长期跟踪研究显得尤为关键。通过追踪评审员在引入 AI 工具前后的技能表现, 可以更准确地理解 AI 工具在各评审维度上对技能提升的具体路径。例如, 定期评估评分准确性、一致性以及批判性思维深度, 有助于量化人机协作对能力提升的促进作用。同时, 这一方法也能识别长期使用 AI 后可能出现的技能退化或能力结构调整等问题。将定量研究与定性访谈结合, 能够更全面地揭示人机协作对评审员专业素养和职业技能的塑造规律。

此外, 设计科学有效的专业能力评估工具是推动相关研究的重要基础。通过开发结构化的能力量表, 可以系统评估评审员在使用 AI 工具后的知识积累、逻辑推理能力及创新判断水平。例如, 结合模拟评审实验与真实数据分析, 考察 AI 工具在学术规范、方法论理解及前沿领域知识支持方面的实际贡献。同时, 这些评估工具应整合反思活动的频次、评分调整的精准度以及对 AI 建议的采纳情况等数据, 形成对评审员能力发展的动态画像, 为优化人机协作流程与能力培养体系提供实证基础。

AI 工具对评审员职业发展的影响还体现在知识更新与成长动力方面。基于大数据分析 with 领域预测技术, AI 能够辅助评审员及时把握学科前沿进展, 推动其持续学习与能力提升。然而, 这一技术优势是否可能削弱评审员的独立判断力和专业自主性, 仍需进一步考察。特别是在复杂决策与高度专业化评审场景中,

评审员如何在 AI 建议基础上融合自身经验进行理性判断，直接关系到评审质量与职业成就感。

长期使用 AI 工具亦可能对评审员的职业习惯与满意度产生深刻影响。AI 所带来的高效性和自动化有助于缓解工作负担，提高投入感和满意度，但过度依赖 AI 则可能在面对高难度任务时削弱评审员的自信心和应对能力。因此，未来研究需重点关注如何在促进职业成长的同时防止专业能力弱化及职业认同感下降。

如果要进一步理解持续人机协作对评审员职业发展的作用，需要探索更为动态与个性化的协作优化策略。将来可开发更具适应性的 AI 系统，支持评审员根据个体需求灵活调整人机协作方式；同时引入动态学习机制和职业培训计划，为评审员持续提升专业知识与判断力提供有力支撑。通过这些措施，人机协作有望在提升评审效率与质量的同时，促进评审员实现专业成长与职业价值的双重跃升。

参考文献

- [1] 周洪宇, 常顺利. 生成式人工智能嵌入高等教育的未来图景、潜在风险及其治理 [J]. 现代教育管理, 2023, (11): 1-12.
- [2] 张雨吟. 大语言模型对学术期刊出版的影响与应对策略 [J]. 传媒论坛, 2025, (09): 100-102.
- [3] 许舒婷. “四学派”框架下 GPT 技术在中文论文同行评审中的应用与展望 - 以中文图情档论文评审 GPTs 为例 [J]. 情报探索, 2025, (02): 84-93.
- [4] AISCHOLAR. AiScholar 学术平台 [Z]. 2025
- [5] 刘银娣. 同行评审的人工智能应用: 现状与挑战 [J]. 出版科学, 2020, 28(05): 68-73.
- [6] 张彤, 尹欢, 苏磊, 等. 人工智能辅助学术同行评议的应用及分类 [J]. 中国科技期刊研究, 2021, 32(1): 65-74.
- [7] 贾璐萌, 赵佳欢. 超越与复归: 人机协作的历史演进及其双重逻辑 [J]. 中国地质大学学报 (社会科学版), 2025, 25(04): 1-10.
- [8] SUN L, CHAN A, CHANG Y S, et al. ReviewFlow: Intelligent Scaffolding to Support Academic Peer Reviewing [Z]. Proceedings of the 29th International Conference on Intelligent User Interfaces. Greenville, SC, USA; Association for Computing Machinery. 2024: 120-137. DOI: 10.1145/3640543.3645159
- [9] YANG C-L, UHDE A, YAMASHITA N, et al. Understanding and Supporting Peer Review Using AI-reframed Positive Summary [Z]. Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery. 2025: Article 171. DOI: 10.1145/3706598.3713219

[10] 王超, 陈轶群, 何运斌, 等. 高校科技期刊的业务流程再造实践 - 以《中国有色金属学报》为例 [J]. 中国科技期刊研究, 2022, 33(11): 1554-1560.

[11] 仇华炳, 吕青, 余诗刚. 国际学术出版学(协)会的运行机制及其在新时期的智库角色探讨 [J]. 中国科技期刊研究, 2024, 35(11): 1530-1540.

[12] 王嘉昀, 文彦杰, 杨柳春. 我国科技期刊创新评审方式的概念、现状与对策分析 [J]. 中国科技期刊研究, 2023, 34(10): 1247-1254.

[13] 朱丽雅, 乐小虬, 付芸. LLM 辅助开放同行评议: 元评审归纳生成研究 [J]. 中国科技期刊研究, 2024, 35(11): 1580-1587.

[14] 孙红梅, 刘荣, 米然, 等. 学术与评审: 学术论文评审的演变及专家画像量化指标的构建 [J]. 中国科技期刊研究, 2021, 32(11): 1369-1375.

[15] 张智雄, 王玉菊, 赵旻. 国际开放同行评审平台的发展趋势及其对中国的建议 [J]. 农业图书情报学报, 2024, 36(05): 14-22.

[16] 王旭杰, 羽王. 大语言模型在数字图书馆中的文本挖掘与信息检索优化 [J]. 电子通信与计算机科学, 2024, 6(10).

[17] BELLINI-LEITE S C. Dual Process Theory for Large Language Models: an Overview of Using Psychology to Address Hallucination and Reliability Issues [J]. Adaptive Behavior, 2023, 32(4): 329-343.

[18] 解煜彬, 周荣刚. 新型人机关系下的人机双向信任 [J]. 心理科学进展, 2025, 33(6): 916-932.

[19] 林雅萍, 郑浩然, 何忆君, 等. AI 交互质量与用户接受度: 心理距离和信任的链式中介作用 [J]. Advances in Psychology, 2025, 15: 47.

[20] 郑雯雯, 武悦, 王美玲, 等. 人工智能生成内容在学术出版中的使用框架构建研究 [J]. 中国科技期刊研究, 2024, 35(10): 1349-1356.

[21] 陈浩洸, 陈罕之, 韩凯峰, 等. 垂直领域大模型的定制化: 理论基础与关键技术 [J]. Journal of Data Acquisition & Processing/Shu Ju Cai Ji Yu Chu Li,

2024, 39(3).

[22] ROWLEY J, SBAFFI L, SUGDEN M, et al. Factors Influencing Researchers' Journal Selection Decisions [J]. Journal of Information Science, 2022, 48(3): 321-335.

[23] WANG Q, LI J, WANG S, et al. Towards Next-Generation LLM-Based Recommender Systems: A Survey and Beyond [J]. arXiv preprint arXiv:2410.19744, 2024.

[24] 俞立平, 张矿伟, 蒋长兵. 推进代表作评价存在的问题与对策研究 [J]. 情报学报, 2021, 40(04): 345-353.

[25] 闫广芬, 杜剑涛. 国际视域下交叉学科博士生培养: 困境、纾困与启迪 [J]. 研究生教育研究, 2025, (01): 41-49.

[26] 陈光, 郭军. 大语言模型时代的人工智能: 技术内涵、行业应用与挑战 [J]. 北京邮电大学学报, 2024, 47(4): 20-28.

[27] 刘如强, 董宝林, 张丹青, 等. 我国体育类随机对照试验研究的方法学与报告质量评价 (2010—2020) [J]. 体育科学, 2021, 41(5): 88-97.

[28] ANTHONY C, BECHKY B A, FAYARD A-L. “Collaborating” with AI: Taking a System View to Explore the Future of Work [J]. Organization Science, 2023, 34(5): 1672-1694.

[29] FÜGENER A, GRAHL J, GUPTA A, et al. Will Humans in the Loop Become Borgs? Merits and Pitfalls of Working with AI [J]. Management Information Systems Quarterly, 2021, 45(3): 1527-1556.

[30] SEEGER I, BITTNER E, BRIGGS R O, et al. Machines as Teammates: A Research Agenda on AI in Team Collaboration [J]. Information & Management, 2020, 57(2): 103174.

[31] TARAFDAR M, PAGE X, MARABELLI M. Algorithms as Co-Workers:

Human Algorithm Role Interactions in Algorithmic Work [J]. *Information Systems Journal*, 2023, 33(2): 232-267.

[32] PURANAM P. Human-AI Collaborative Decision-Making as an Organization Design Problem [J]. *Journal of Organization Design*, 2021, 10(2): 75-80.

[33] JACOBS M, PRADIER M F, MCCOY T H, et al. How Machine-Learning Recommendations Influence Clinician Treatment Selections: The Example of Antidepressant Selection [J]. *Translational Psychiatry*, 2021, 11(1): 108.

[34] LAI V, LIU H, TAN C. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans [Z]. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Honolulu, HI, USA; Association for Computing Machinery. 2020: 1-13. DOI: 10.1145/3313831.3376873.

[35] BANSAL G, WU T, ZHOU J, et al. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance [Z]. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama, Japan; Association for Computing Machinery. 2021: Article 81. DOI: 10.1145/3411764.3445717.

[36] LUZ, YIN M. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks [Z]. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Yokohama, Japan; Association for Computing Machinery. 2021: Article 78. DOI: 10.1145/3411764.3445562.

[37] BERTRAND A, BELLOUM R, EAGAN J R, et al. How Cognitive

Biases Affect XAI-assisted Decision-making: A Systematic Review [Z]. Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society. Oxford, United Kingdom; Association for Computing Machinery. 2022: 78–91. DOI: 10.1145/3514094.3534164

[38] BUÇINCA Z, MALAYA M B, GAJOS K Z. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making [J]. Proc ACM Hum-Comput Interact, 2021, 5(CSCW1): Article 188.

[39] HE G, KUIPER L, GADIRAJU U. Knowing About Knowing: An Illusion of Human Competence can Hinder Appropriate Reliance on AI Systems [Z]. Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. Hamburg, Germany; Association for Computing Machinery. 2023: Article 113. DOI: 10.1145/3544548.3581025

[40] BAIRD A, MARUPING L M. The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts [J]. Management Information Systems Quarterly, 2021, 45(1): 315-341.

[41] EVANS J A, FOSTER J G. Metaknowledge [J]. Science, 2011, 331(6018): 721-725.

[42] SHAO Y, HUANG C, SONG Y, et al. Using Augmentation-Based AI Tool at Work: A Daily Investigation of Learning-Based Benefit and Challenge [J]. Journal of Management, 2024.

[43] BAO Y, GONG W, YANG K. A Literature Review of Human-AI Synergy in Decision Making: From the Perspective of Affordance Actualization Theory [J]. Systems, 2023, 11(9): 442.

[44] LIAO Q V, VAUGHAN J W. Ai Transparency in the Age of LLMS: A Human-Centered Research Roadmap [J]. arXiv preprint arXiv:2306.01941, 2023, 10.

[45] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All You Need [J]. Advances in neural information processing systems, 2017, 30.

[46] LI Y, ZHAO J, LI M, et al. RefAI: a GPT-powered Retrieval-Augmented Generative Tool for Biomedical Literature Recommendation and Summarization [J]. Journal of the American Medical Informatics Association, 2024, 31(9): 2030-2039.

[47] ABDEL-KARIM B M, PFEUFFER N, CARL K V, et al. How AI-Based Systems Can Induce Reflections: the Case of AI-Augmented Diagnostic Work [J]. Management Information Systems Quarterly, 2023, 47(4): 1395-1424.

[48] SEYMOUR M, YUAN L, RIEMER K, et al. Less Artificial, More Intelligent: Understanding Affinity, Trustworthiness, and Preference for Digital Humans [J]. Information Systems Research, 2025, 36(2): 1096-1128.

[49] BAUER K, VON ZAHN M, HINZ O. Expl(AI)ned: the Impact of Explainable Artificial Intelligence on Users' Information Processing [J]. Information Systems Research, 2023, 34(4): 1582-1602.

[50] NOY S, ZHANG W. Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence [J]. Science, 2023, 381(6654): 187-192.

[51] BOMMASANI R. On The Opportunities and Risks of Foundation Models [J]. arXiv preprint arXiv:2108.07258, 2021.

[52] KAHNEMAN D. Thinking, Fast and Slow[M]. Farrar, Straus and Giroux, 2011.

[53] SEYMOUR W, ABDI N, RAMOKAPANE K M, et al. Voice App Developer Experiences with Alexa and Google Assistant: Juggling Risks, Liability, and Security; Proceedings of The 33rd USENIX Security Symposium (USENIX Security 24), F, 2024 [C].

[54] 叶泽童. 认知负荷学习理论发展综述 [J]. Advances in Social Sciences, 2022, 11: 4413.

[55] WANG J, CHENG R, LIU M, et al. Research Trends of Human-Computer Interaction Studies in Construction Hazard Recognition: A Bibliometric Review [J]. Sensors, 2021, 21(18): 6172.

[56] WEISS A, MICHELS C, BURGMER P, et al. Trust in Everyday Life [J]. Journal of Personality and Social Psychology, 2021, 121(1): 95.

[57] TUKEY J W. Comparing Individual Means in the Analysis of Variance [J]. Biometrics, 1949, 5(2): 99-114.

[58] LI Z, LIANG C, PENG J, et al. The Value, Benefits, and Concerns of Generative AI-Powered Assistance in Writing [Z]. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems. Honolulu, HI, USA; Association for Computing Machinery. 2024: Article 1048. DOI: 10.1145/3613904.3642625.

[59] HABBAL A, ALI M K, ABUZARAIDA M A. Artificial Intelligence Trust, Risk and Security Management (AI TRiSM): Frameworks, applications, challenges and future research directions [J]. Expert Systems with Applications, 2024, 240: 122442.

[60] LOCKEY S, GILLESPIE N, HOLM D, et al. A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions [J]. 2021.

[61] OMRANI N, RIVIECCIO G, FIORE U, et al. To Trust or Not to Trust? An Assessment of Trust in AI-Based Systems: Concerns, Ethics And Contexts [J]. Technological Forecasting and Social Change, 2022, 181: 121763.

附录

附录 A

实验一：有无 AI 辅助对决策影响的实验

附录 A.1 实验一研究问卷

您好，感谢您参与本研究项目。本研究由学术评审服务平台及合作研究团队联合开展，旨在评估大语言模型辅助的学术审阅流程与人机协作机制！

本问卷旨在测试计算机及相关专业的在读与毕业学生对学术会议论文的阅读理解与评估审阅能力。我们将邀请您在快速阅读论文后，从内容丰富性、论证逻辑、表达与写作、创新性与重要性和整体评价五个维度分别对论文进行打分。考虑到受试者背景的不同以及时间的限制，我们挑选了不太受专业知识限制，同时也无需精读论文，就可以评估的维度。每份有效问卷的作答者会收到 50 元现金作为奖励；我们还会给那些在最短时间内给出最准确打分的问卷追加更多现金奖励。希望您仔细阅读相关论文并给予客观公正的评估。

完成问卷的过程中，请您不要离开或关闭浏览器页面，我们会综合考虑您的作答时间和准确度对您的表现进行评估。综合排名靠前的问卷会获得纪念品奖励，或者等价的现金奖励。同时，作答时间显著高于或者低于平均值的异常问卷会视为废卷。在我们的预测试中，完成本问卷平均约耗费 35-50 分钟。

我们将实行严格匿名制度，确保对您所提供信息的严密保密；所有调查结果仅用于学术研究之用途。衷心感谢您宝贵的支持与帮助。

作答并提交本问卷将被视为您知晓、同意上述内容并自愿参与本研究与后续

研究。如果您对这个研究没有兴趣，您可以现在退出作答。

期待您的认真回答！

无 AI - 论文 1- 版块开始

Q1.1 请仔细阅读以下论文并回答问题。请注意，您至少需要在本页面停留 1 分钟才能继续回答后续问题。

论文内容见：ALNAFESAH G, SMITH P, LEE M. Are you not moved? Incorporating sensorimotor knowledge to improve metaphor detection[C/OL]//Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing. Varna, Bulgaria: INCOMA Ltd., 2023: 80-89. https://doi.org/10.26615/978-954-452-092-2_010.

Q1.2 请确认您已经认真读完本篇论文：

[单选]

☐ 我已经认真完成本篇论文阅读。

Q1.3 请从内容丰富性的角度，给出您对该论文的评价，1 分为最差，5 分为最优。具体评分标准如下：

- 1 分（很差）：文章内容非常单薄，缺乏必要的细节和深度，未能提供足够的信息，结构不完整，缺少关键部分，没有或缺乏说明性图表。
- 2 分（差）：文章内容较为简单，仅覆盖了一些基本要点，结构尚可但不够完整，信息量有限，图表较少且辅助说明作用不明显。
- 3 分（普通）：文章内容较为充实，覆盖了大部分相关要点，提供了一定量的信息，结构基本完整，但某些部分可能仍需进一步丰富和细化，图表数量适中，能较好地辅助文本内容。

- 4分（良好）：文章内容丰富，详尽地探讨了研究问题，提供了大量的信息和数据支持，结构完整，图表丰富且有效地支持了文章的论点和分析。
- 5分（优秀）：文章内容极为丰富和深入，全面地覆盖了研究领域的各个方面，结构严谨，内容层次分明，图表数量充足且设计精良，能够直观地展示研究结果和数据分析，极大地增强了文章的说服力和学术价值。

[量表]

[最差,最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q1.4 请从论证逻辑的角度，给出您对该论文的评价，1分为最差，5分为最优。具体评分标准如下：

- 1分（很差）：文章的论点支离破碎，缺乏逻辑性，没有可辨识的逻辑结构。
- 2分（差）：文章少数内容符合逻辑，但频繁出现逻辑跳跃，观点间缺少清晰联系。
- 3分（普通）：文章大体上逻辑流畅，大部分观点之间关联明确，尽管某些论证可能并未完全展开或缺乏联系。
- 4分（良好）：文章论证结构合理、思路清晰且逻辑性强，只有少数部分出现轻微的连贯性问题。
- 5分（优秀）：文章论证连贯，逻辑结构十分严密，每一部分都与上一部分紧密联结。

[量表]

[最差,最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q1.5 请从表达与写作的角度，给出您对该论文的评价，1分为最差，5分为最优。具体评分标准如下：

- 1分（很差）：文章缺乏流畅性，既不清晰也不简洁，在许多关键部分用词不准确或出现严重语法错误，行文不符合学术写作风格，阻碍了读者的理解。

- 2分（差）：文章有些部分比较清晰和流畅，但整体来说比较冗长，一些用词或语法错误也给读者理解论文内容造成了困难。
- 3分（普通）：文章大体上流畅、清晰且简洁，只有少量用词不当或语法错误，行文基本符合学术写作风格。
- 4分（良好）：文章表达流畅且清晰，核心部分阐述简洁明了，基本没有用词或语法错误，遵循学术写作的风格。
- 5分（优秀）：文笔十分流畅，表达出色，清晰且简洁有力，深入浅出地阐述复杂论点，完全没有用词或语法错误，严格遵循了学术写作风格。

[量表]

[最差, 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q1.6 请从创新性与重要性的角度，给出您对该论文的评价，1 分为最差，5 分为最优。具体评分标准如下：

- 1分（很差）：文章缺乏原创思想或对该领域的贡献，未能提供任何新见解或改进。
- 2分（差）：文章的新颖性和意义极为有限，仅在已知概念上做出轻微改变，没有为该领域增加实质研究价值。
- 3分（普通）：文章引入了一些新颖的观点或方法，为该研究领域做出了一定贡献。
- 4分（良好）：文章具有高度的创新性且比较重要，推动了该研究领域的发展与进步。
- 5分（优秀）：论文具有开创性，提供了全新且意义重大的贡献，大大推动了领域内最先进的技术进步。

[量表]

[最差, 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q1.7 请给出您对该论文的整体评价，1 分为最差，5 分为最优。具体评

分标准如下：

- 1分（很差）：论文存在重大缺陷，除非经过重大修订，否则无法发表。
- 2分（差）：论文具有一定价值，但需要进行重大改进才能发表。
- 3分（普通）：论文可以接受但并不突出；它满足基本标准，但缺乏重要贡献或存在比较显著的问题。
- 4分（良好）：论文在大多数方面表现良好，但是可能存在一些小问题。
- 5分（优秀）：论文在所有方面都表现出色，显然值得接受，最多只需进行轻微修订。

[量表]

[最差, 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

无AI - 论文 1-版块结束

无AI - 论文 2-版块开始

Q2.1 请仔细阅读以下论文并回答问题。请注意，您至少需要在本页面停留 1 分钟才能继续回答后续问题。[单选]

论文内容见：ZHANG J, LI Y, LI H, et al. Text classification method based on improved long short term memory network[C]//International Conference on Computer Graphics, Artificial Intelligence, and Data Processing (ICCAID 2023). Bellingham: SPIE, 2024: 552-556.

Q2.2 请确认您已经认真读完本篇论文：[单选]

☐ 我已经认真完成本篇论文阅读。

Q2.3 请从内容丰富性的角度，给出您对该论文的评价，1 分为最差，

5 分为最优。具体评分标准如下：

- 1分（很差）：文章内容非常单薄，缺乏必要的细节和深度，未能提供足够的信息，结构不完整，缺少关键部分，没有或缺乏说明性图表。
- 2分（差）：文章内容较为简单，仅覆盖了一些基本要点，结构尚可但不够完整，信息量有限，图表较少且辅助说明作用不明显。
- 3分（普通）：文章内容较为充实，覆盖了大部分相关要点，提供了一定量的信息，结构基本完整，但某些部分可能仍需进一步丰富和细化，图表数量适中，能较好地辅助文本内容。
- 4分（良好）：文章内容丰富，详尽地探讨了研究问题，提供了大量的信息和数据支持，结构完整，图表丰富且有效地支持了文章的论点和分析。
- 5分（优秀）：文章内容极为丰富和深入，全面地覆盖了研究领域的各个方面，结构严谨，内容层次分明，图表数量充足且设计精良，能够直观地展示研究结果和数据分析，极大地增强了文章的说服力和学术价值。

[量表]

[最差, 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q2.4 请从论证逻辑的角度，给出您对该论文的评价，1 分为最差，5 分为最优。具体评分标准如下：

- 1分（很差）：文章的论点支离破碎，缺乏逻辑性，没有可辨识的逻辑结构。
- 2分（差）：文章少数内容符合逻辑，但频繁出现逻辑跳跃，观点间缺少清晰联系。
- 3分（普通）：文章大体上逻辑流畅，大部分观点之间关联明确，尽管某些论证可能并未完全展开或缺乏联系。
- 4分（良好）：文章论证结构合理、思路清晰且逻辑性强，只有少数部分出现轻微的连贯性问题。
- 5分（优秀）：文章论证连贯，逻辑结构十分严密，每一部分都与上一部分紧密联结。

[量表]

[最差, 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q2.5 请从表达与写作的角度, 给出您对该论文的评价, 1 分为最差, 5 分为最优。具体评分标准如下:

- 1分(很差): 文章缺乏流畅性, 既不清晰也不简洁, 在许多关键部分用词不准确或出现严重语法错误, 行文不符合学术写作风格, 阻碍了读者的理解。
- 2分(差): 文章有些部分比较清晰和流畅, 但整体来说比较冗长, 一些用词或语法错误也给读者理解论文内容造成了困难。
- 3分(普通): 文章大体上流畅、清晰且简洁, 只有少量用词不当或语法错误, 行文基本符合学术写作风格。
- 4分(良好): 文章表达流畅且清晰, 核心部分阐述简洁明了, 基本没有用词或语法错误, 遵循学术写作的风格。
- 5分(优秀): 文笔十分流畅, 表达出色, 清晰且简洁有力, 深入浅出地阐述复杂论点, 完全没有用词或语法错误, 严格遵循了学术写作风格。

[量表]

[最差, 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q2.6 请从创新性与重要性的角度, 给出您对该论文的评价, 1 分为最差, 5 分为最优。具体评分标准如下:

- 1分(很差): 文章缺乏原创思想或对该领域的贡献, 未能提供任何新见解或改进。
- 2分(差): 文章的新颖性和意义极为有限, 仅在已知概念上做出轻微改变, 没有为该领域增加实质研究价值。
- 3分(普通): 文章引入了一些新颖的观点或方法, 为该研究领域做出了一定贡献。

- 4 分（良好）：文章具有高度的创新性且比较重要，推动了该研究领域的发展与进步。
- 5 分（优秀）：论文具有开创性，提供了全新且意义重大的贡献，大大推动了领域内最先进的技术进步。

[量表]

[最差, 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q2.7 请给出您对该论文的整体评价，1 分为最差，5 分为最优。具体评分标准如下：

- 1 分（很差）：论文存在重大缺陷，除非经过重大修订，否则无法发表。
- 2 分（差）：论文具有一定价值，但需要进行重大改进才能发表。
- 3 分（普通）：论文可以接受但并不突出；它满足基本标准，但缺乏重要贡献或存在比较显著的问题。
- 4 分（良好）：论文在大多数方面表现良好，但是可能存在一些小问题。
- 5 分（优秀）：论文在所有方面都表现出色，显然值得接受，最多只需进行轻微修订。

[量表]

[最差, 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

无 AI - 论文 2- 版块结束

无 AI 主观难度评价 - 版块开始

Q3.1 在您参与的上述论文审阅任务中，对论文的哪些维度进行评价花费了您最多的精力？请根据花费精力的多少，请将您认为最耗费精力的维度设置为 10 分，其他维度则根据其相对于最高分项的精力投入比例来调整分数。请注意，对应 10 分的维度不能超过 2 个，且至少 1 个。

例如，如果您觉得在 " 总体评价 " 这一维度上投入的精力最多，那么请将其分数设为 10 分；而如果您在 " 表达与写作 " 这一维度上投入的精力相对较少，假设只有 " 总体评价 " 维度精力的一半，那么这一维度的分数应设为 5 分。

[矩阵量表]										
维度	1 (最少)	2	3	4	5	6	7	8	9	10 (最多)
内容丰富性	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
论证逻辑	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
表达与写作	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
创新性与重要性	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
整体评价	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Q3.2 根据您的专业知识背景以及过去的论文写作或审阅经验，结合您在以上论文评估任务中的体验，您认为在评估论文时，自己在各个维度上胜任力（Competence）如何？

[矩阵量表]							
维度	1 (胜任力低)	2	3	4	5	6	7 (胜任力高)
内容丰富性	☐	☐	☐	☐	☐	☐	☐
论证逻辑	☐	☐	☐	☐	☐	☐	☐
表达与写作	☐	☐	☐	☐	☐	☐	☐
创新性与重要性	☐	☐	☐	☐	☐	☐	☐
整体评价	☐	☐	☐	☐	☐	☐	☐

无 AI 主观难度评价 - 版块结束

全 AI - 论文 3- 版块开始

Q4.1 大语言模型是一种基于深度学习技术，通过海量文本数据训练而成的人工智能（AI）系统。它在文本生成、机器翻译、问答系统、文本摘要、情感分

析、语音识别转写、对话系统、内容创作、代码生成等方面都展现出了卓越的性能，不仅能够理解和生成人类语言，还能够执行各种复杂的语言任务，如推理、分析和创造性写作。

目前，研究者正在探索将大语言模型应用于学术论文审阅过程。AI 可以为审稿人提供初步审稿意见并提出可能的改进方向，这不仅能加速审阅流程，还可以减少人为偏见，提高审稿质量。

请仔细阅读以下论文以及 AI 生成的论文评价意见，并回答问题。请注意，您至少需要在本页面停留 1 分钟才能继续回答后续问题。

论文内容见：FÄRBER M, POPOVIC N. Vocab-expander: a system for creating domain-specific vocabularies based on word embeddings[R/OL]. (2023-08-08) [2024-05-22]. <http://arxiv.org/abs/2308.03519>.

Q4.2 请确认您已经认真读完本篇论文：[单选]

☐ 我已经认真完成本篇论文阅读。

Q4.3 请首先阅读以下由 AI 生成的论文评价意见（内容丰富性维度）：

内容丰富性：

- 优点： 文章内容较为全面，详细介绍了 VOCAB-EXPANDER 系统的设计理念、系统设计、潜在用例以及与现有工作的比较。
- 缺点： 虽然论文内容全面，但在某些部分（如系统评估和与现有技术的深入比较）缺乏深度和详细信息。
- 总结： 文章在介绍工具的设计和潜在应用方面做得很好，但在评估和深度分析方面有待加强。
- 得分： 3.5

请从内容丰富性的角度，给出您对这篇论文的评价，1 分为最差，5 分为最优。具体评分

标准见 Q1.3。

[量表]

[最差, 最优]

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Q4.4 请首先阅读以下由 AI 生成的论文评价意见（论证逻辑维度）：

论证逻辑：

- 优点： 文章结构清晰，逻辑连贯，从系统的动机、设计到潜在用例逐步展开，易于理解。
- 缺点： 在某些部分，如与现有工作的比较和系统评估，逻辑展开较为简略，影响了论文的完整性。
- 总结： 尽管在某些细节处理上存在一定欠缺，但整体上文章保持了较好的逻辑连贯性。
- 得分： 4

请从论证逻辑的角度，给出您对该论文的评价，1 分为最差，5 分为最优。具体评分标准见 Q1.4。

[量表]

[最差, 最优]

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Q4.5 请首先阅读以下由 AI 生成的论文评价意见（表达与写作维度）：

表达与写作：

- 优点： 论文的写作风格清晰流畅，易于阅读和理解。
- 缺点： 在一些技术细节的描述上，论文有时显得过于简略，可能会让读者对系统的具体实现和效果有所疑惑。
- 总结： 论文整体写作质量较高，但需要在某些技术描述上更加详尽。
- 得分： 4

请从表达与写作的角度，给出您对该论文的评价，1 分为最差，5 分为最优。具体评分

标准见 Q1.5。

[量表]

[最差, 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q4.6 请首先阅读以下由 AI 生成的论文评价意见（创新性和重要性维度）：

创新性和重要性：

· 优点： 优点： VOCAB-EXPANDER 工具在支持非技术用户创建和扩展特定领域词汇表方面具有一定的创新性。对于改善信息检索、增强组织内的沟通和协作等方面具有潜在的实际价值。

· 缺点： 论文在创新点的深入探讨和系统对现有问题的根本改进方面略显不足。

· 总结： 尽管论文提出了有趣的概念和工具，但在突出其创新性和重要性方面仍有提升空间。

· 得分： 4

请从创新性与重要性的角度，给出您对该论文的评价，1 分为最差，5 分为最优。

具体评分标准见 Q1.6。

[量表]

[最差, 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q4.7 请首先阅读以下由 AI 生成的论文评价意见（整体评价维度）：

总体评价：

· 优点： 本文提出的 VOCAB-EXPANDER 是一个创新的在线工具，能够帮助用户（尤其是技术侦察者等非技术专业用户）创建和扩展特定领域的词汇表。工具使用了基于网络文本和常识知识库 ConceptNet 的先进词嵌入技术，以便为已有术语提供相关术语建议。

此外，用户界面友好，易于操作，支持快速确认或拒绝术语建议，极大地简化了词汇表创

建过程。

· 缺点： 尽管提出的系统具有一定的创新性和实用性，但论文在对比分析现有工具及方法方面的深度不足，缺乏对系统有效性的实证评估，这限制了其说服力。此外，对系统可能的改进方向和未来工作的讨论也相对较少。

· 总结： 该论文介绍了一个基于词嵌入技术的在线工具，用于创建和扩展特定领域的词汇表，虽然概念上具有创新性，但在方法论深度和系统评估方面存在不足。

· 得分： 3.5

请给出您对该论文的整体评价，1 分为最差，5 分为最优。具体评分标准见 Q1.7。

[量表]

[最差 , 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

全 AI - 论文 3- 版块结束

全 AI - 论文 4- 版块开始

Q5.1 请仔细阅读以下论文以及 AI 生成的论文评价意见，并回答问题。

请注意，您至少需要在本页面停留 1 分钟才能继续回答后续问题。

论文内容见：ZHU L, WEI F, LU Y. Beyond text: frozen large language models in visual signal comprehension[C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2024: 27047-27057.

Q5.2 请确认您已经认真读完本篇论文：

[单选]

☐ 我已经认真完成本篇论文阅读。

Q5.3 请首先阅读以下由 AI 生成的论文评价意见（内容丰富性维度）：

内容丰富性：

- 优点： 文章提出了一种新颖的将图像视为“外语”的方法，并通过 Vision-to-Language Tokenizer(V2L Tokenizer) 将连续的视觉信号映射到大语言模型（LLM）的 token 空间。这种方法使得冻结的 LLM 能够在不需要对多模态数据集进行资源密集型微调的情况下理解视觉信号。文章还通过提出的词汇表扩展技术，生成了能够捕捉到关键语义信息的全局 token。
- 缺点： 虽然在图像理解和去噪任务方面取得了一定的成果，但文章在内容的深度上还有提升空间，比如可以进一步探讨 V2L Tokenizer 在不同领域和更复杂任务中的应用效果。
- 总结： 整体而言，文章在内容丰富度方面表现良好，提出的方法在图像理解和去噪方面具有一定的创新性和有效性。
- 得分： 4

请从内容丰富性的角度，给出您对该论文的评价，1 分为最差，5 分为最优。具体评分标准见 Q1.3。[量表]

[最差 , 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q5.4 请首先阅读以下由 AI 生成的论文评价意见（论证逻辑维度）：

论证逻辑：

- 优点： 文章的结构清晰，逻辑性强。从问题的提出到方法的介绍，再到实验的验证，整个论文的脉络非常清晰，各部分之间的逻辑关系紧密相连。
- 缺点： 在某些部分，如方法的详细算法描述和实验细节，文章可能会受篇幅限制而无法提供更深入的逻辑推理和详尽的解释。
- 总结： 文章在逻辑连贯性方面做得很好，每个部分都紧密相扣，形成了一个完整的

研究故事。

· 得分： 4

请从论证逻辑的角度，给出您对该论文的评价，1分为最差，5分为最优。具体评分标准见Q1.4。[量表]

[最差 , 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q5.5 请首先阅读以下由 AI 生成的论文评价意见（表达与写作维度）：

表达与写作：

· 优点： 文章的写作风格流畅，语言表达清晰，概念解释简洁明了，使得读者即使不具备深厚的专业背景也能理解文章的主要内容和贡献。

· 缺点： 在一些技术深入的部分，文章可能在保持简洁性的同时牺牲了一些详细性，这可能会导致专业读者需要额外的资料来完全理解方法的细节。

· 总结： 文章在写作方面表现出色，能够很好地平衡技术性和可读性。

· 得分： 4

请从表达与写作的角度，给出您对该论文的评价，1分为最差，5分为最优。具体评分标准见

Q1.5。

[量表]

[最差 , 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q5.6 请首先阅读以下由 AI 生成的论文评价意见（创新性和重要性维度）：

创新性和重要性：

· 优点： 文章提出的将图像视为“外语”的概念在视觉和语言模型的结合上具有明显的新颖性。V2L Tokenizer的引入为冻结的LLM处理和理解视觉信息提供了一种新的可

能性，这对于推动多模态研究具有重要意义。

· 缺点：尽管在新颖性方面表现突出，但文章在重要性方面的论证还可以加强，例如通过更多实际应用场景的探讨来展示该方法的广泛影响力。

· 总结：文章在新颖性方面做出了显著的贡献，而在重要性方面虽然有所体现，但仍有提升空间。

· 得分： 4

请从创新性与重要性的角度，给出您对该论文的评价，1分为最差，5分为最优。

具体评分标准见Q1.6。

[量表]

[最差 , 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q5.7 请首先阅读以下由 AI 生成的论文评价意见（整体评价维度）：

创新性和重要性：

· 优点：文章提出了一种创新的方法来整合视觉信号与大语言模型，实验结果证明了方法的有效性。写作风格清晰，逻辑连贯，为读者提供了良好的阅读体验。

· 缺点：尽管文章在多个方面都表现出色，但在展示方法在更广泛应用中的潜力方面还有提升空间。

· 总结： 综合考虑，文章在多个维度上都表现良好，是一篇有力的研究工作。

· 得分： 4

请给出您对该论文的整体评价，1分为最差，5分为最优。具体评分标准见Q1.7。

[量表]

[最差 , 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

全AI - 论文 4-版块结束

全 AI - 论文 5- 版块开始

Q6.1 请仔细阅读以下论文以及 AI 生成的论文评价意见，并回答问题。

请注意，您至少需要在本页面停留 1 分钟才能继续回答后续问题。

论文内容见：WANG H, SU J, JIN T, et al. The application of classifier model based on FA-SVM algorithm in speech recognition[J/OL]. Journal of Physics: Conference Series, 2023, 2562(1): 012019.

Q6.2 请确认您已经认真读完本篇论文：

[单选]

☐ 我已经认真完成本篇论文阅读。

Q6.3 请首先阅读以下由 AI 生成的论文评价意见（内容丰富性维度）：

内容丰富性：

- 优势：文章详细介绍了基于 FA-SVM 算法的分类器在语音识别中的应用，包括算法原理、模型搭建和实验模拟的具体实施过程。文中还列举了对比实验和分类准确率，为读者提供了充足的信息来支持文中所述的观点。
- 缺点：在原理介绍和算法应用的部分中，对既有研究背景的介绍略显薄弱，有待进一步提供更丰富的研究背景和相关工作。
- 总结：文章提供了 FA-SVM 算法的详尽介绍，在内容上基本完整，但在研究背景方面还有增强的空间。
- 得分： 4

请从内容丰富性的角度，给出您对该论文的评价，1 分为最差，5 分为最优。具体评分标准见 Q1.3。

[量表]

[最差 , 最优]

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Q6.4 请首先阅读以下由 AI 生成的论文评价意见（论证逻辑维度）：

论证逻辑：

- 优势：文章逻辑清晰，易于理解。从算法介绍到模型构建，再到实验验证，步骤顺序明晰，各部分之间逻辑关联紧密。
- 缺点：虽有系统性的展示，但在实验部分稍显简单，可以进一步增强实验设计的合理性和严谨性。
- 总结：逻辑结构总体上是合理的，但实验部分的逻辑性有改善空间。
- 得分： 4

请从论证逻辑的角度，给出您对该论文的评价，1 分为最差，5 分为最优。具体评分标准见 Q1.4。

[量表]

[最差 , 最优]

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Q6.5 请首先阅读以下由 AI 生成的论文评价意见（表达与写作维度）：

表达与写作：

- 优势：文章的写作风格通俗易懂，解释清晰，没有过度复杂的术语使用。
- 缺点：在部分节内容的表述上有些冗长，可通过精简语言来增加简洁性。
- 总结：整体上写作风格流畅，但局部有优化空间。
- 得分： 3.5

请从表达与写作的角度，给出您对该论文的评价，1 分为最差，5 分为最优。具体评分标准见 Q1.5。

[量表]

[最差 , 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q6.6 请首先阅读以下由 AI 生成的论文评价意见（创新性和重要性维度）：

- 优势：结合 FA 算法优化 SVM 参数的方法在语音识别领域具有一定的创新性，并且通过实验验证体现了其有效性。
- 缺点：虽然提出了新的算法结合，但整体上缺乏对算法创新点深入探讨和对整个领域发展的影响评估。
- 总结：新颖性和重要性都表现出一定程度，但仍有提升空间。
- 得分： 3.5

请从创新性与重要性的角度，给出您对该论文的评价，1 分为最差，5 分为最优。具体评分标准见 Q1.6。

[量表]

[最差, 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

Q6.7 请首先阅读以下由 AI 生成的论文评价意见（整体评价维度）：

总体评价

- 优势：文章展示了一个在语音识别领域具有潜力的新算法使用案例，提供了充分的实验结果以支持其有效性，写作通顺。
- 缺点：需要进一步加强在算法原理深入探讨、创新点阐述和研究背景方面的内容。
- 总结：文章整体质量良好，是一篇有实践意义的研究，但仍有进一步提升的空间。
- 得分： 3.5

请给出您对该论文的整体评价，1 分为最差，5 分为最优。具体评分标准见 Q1.7。

[量表]

[最差, 最优]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

全AI - 论文 5-版块结束

有 AI 主观难度评价 - 版块开始

Q7.1 在您参与的上述论文审阅任务中，在有 AI 生成评价作为参考的情况下，对论文的哪些维度进行评价花费了您最多的精力？请根据花费精力的多少，请将您认为最耗费精力的维度设置为 10 分，其他维度则根据其相对于最高分项的精力投入比例来调整分数。请注意，对应 10 分的维度不能超过 2 个，且至少 1 个。

例如，如果您觉得在 " 总体评价 " 这一维度上投入的精力最多，那么请将其分数设为 10 分；而如果您在 " 表达与写作 " 这一维度上投入的精力相对较少，假设只有 " 总体评价 " 维度精力的一半，那么这一维度的分数应设为 5 分。

[矩阵量表]										
维度	1 (最少)	2	3	4	5	6	7	8	9	10 (最多)
内容丰富性	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
论证逻辑	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
表达与写作	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
创新性与重要性	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
整体评价	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

有AI 主观难度评价-版块结束

AI 评价 - 版块开始

Q8.1 下面是一些关于论文审阅任务中，AI 在内容丰富性维度上表现和能力的陈述。请根据您在以上任务中的实际体验，结合您对 AI 的了解和使用经验，选择对每条陈述的认同程度。答案没有对错之分。

(1= 非常不同意 , 2= 不同意 , 3= 比较不同意 , 4= 一般 , 5= 比较同意 , 6= 同意 , 7= 非常同意)

[矩阵量表]

陈述	1	2	3	4	5	6	7
在内容丰富性维度上, 我在很大程度上参考了 AI 的建议	☐	☐	☐	☐	☐	☐	☐
AI 提供的建议比较合理	☐	☐	☐	☐	☐	☐	☐
我和 AI 的分歧较小	☐	☐	☐	☐	☐	☐	☐
AI 可以独自胜任审阅打分任务	☐	☐	☐	☐	☐	☐	☐
AI 可以取代人类审阅者进行打分	☐	☐	☐	☐	☐	☐	☐
AI 在内容丰富性维度上的建议与论文内容比较相关	☐	☐	☐	☐	☐	☐	☐
AI 在内容丰富性维度上的建议展现了较强的逻辑思维能力	☐	☐	☐	☐	☐	☐	☐
AI 在内容丰富性维度上的建议展现了较强的语言能力	☐	☐	☐	☐	☐	☐	☐
AI 在内容丰富性维度上的建议展现了较强的批判性思维能力	☐	☐	☐	☐	☐	☐	☐

Q8.2 下面是一些关于论文审阅任务中, AI 在论证逻辑维度上表现和能力的陈述。请根据您在以上任务中的实际体验, 结合您对 AI 的了解和使用经验, 选择对每条陈述的认同程度。答案没有对错之分。

(1= 非常不同意 , 2= 不同意 , 3= 比较不同意 , 4= 一般 , 5= 比较同意 , 6= 同意 , 7= 非常同意)

Q8.3 下面是一些关于论文审阅任务中, AI 在表达与写作维度上表现和能力的陈述。请根据您在以上任务中的实际体验, 结合您对 AI 的了解和使用经验, 选择对每条陈述的认同程度。答案没有对错之分。

(1= 非常不同意 , 2= 不同意 , 3= 比较不同意 , 4= 一般 , 5= 比较同意 , 6= 同意 , 7= 非常同意)

[矩阵量表]

(此处表格内容与 Q8.1 的陈述类似, 但针对 “表达与写作” 维度, 原文表格结构重复,

此处省略。)

Q8.4 下面是一些关于论文审阅任务中，AI 在创新性与重要性维度上表现和能力的陈述。请根据您在以上任务中的实际体验，结合您对 AI 的了解和使用经验，选择对每条陈述的认同程度。答案没有对错之分。

(1= 非常不同意，2= 不同意，3= 比较不同意，4= 一般，5= 比较同意，6= 同意，7= 非常同意)

[矩阵量表]

(此处表格内容与 Q8.1 的陈述类似，但针对“创新性与重要性”维度，原文表格结构重复，此处省略。)

Q8.5 下面是一些关于论文审阅任务中，AI 在整体评价维度上表现和能力的陈述。请根据您在以上任务中的实际体验，结合您对 AI 的了解和使用经验，选择对每条陈述的认同程度。答案没有对错之分。

(1= 非常不同意，2= 不同意，3= 比较不同意，4= 一般，5= 比较同意，6= 同意，7= 非常同意)

[矩阵量表]

(此处表格内容与 Q8.1 的陈述类似，但针对“整体评价”维度，原文表格结构重复，此处省略。)

AI评价-版块结束

任务评价 - 版块开始

Q9.1 您认为在论文审阅中，读者对论文的哪些维度进行评价的难度最高？请根据评价难度，对以下维度进行打分。请根据以上论文审阅任务中的实际体验和您自身的经验进行选择，答案没有对错之分。

(1= 非常简单，2= 简单，3= 比较简单，4= 一般，5= 比较难，6= 难，7= 非常难)

[矩阵量表]

对于我来说...	1	2	3	4	5	6	7
在评价“内容丰富性”时，我感觉：	☐	☐	☐	☐	☐	☐	☐
在评价“论证逻辑”时，我感觉：	☐	☐	☐	☐	☐	☐	☐
在评价“表达与写作”时，我感觉：	☐	☐	☐	☐	☐	☐	☐
在评价“创新性与重要性”时，我感觉：	☐	☐	☐	☐	☐	☐	☐
在评价“整体评价”时，我感觉：	☐	☐	☐	☐	☐	☐	☐

Q9.2 综合您之前对基于大语言模型的 AI（如 ChatGPT, Claude AI, Kimi 等）的使用经验和整体认识，以及以上几篇 AI 生成的论文评价，您认为 AI 在论文评分这一类任务上的整体表现如何？

[量表]

[非常差 , 非常好]

☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

任务评价-版块结束

控制变量 - 版块开始

Q10.1 下面的陈述是人们对 AI 系统的一些认识和看法，答案没有对错之分，请根据您对 AI 的了解和自身感受进行作答。

(1= 非常不同意 , 2= 不同意 , 3= 比较不同意 , 4= 一般 , 5= 比较同意 , 6= 同意 , 7= 非常同意)

[矩阵量表]

陈述	1	2	3	4	5	6	7
AI 系统具有欺骗性	☐	☐	☐	☐	☐	☐	☐
我对 AI 系统的意图、行为或输出感到怀疑	☐	☐	☐	☐	☐	☐	☐
我对 AI 系统持谨慎态度	☐	☐	☐	☐	☐	☐	☐
AI 系统的行为会产生有害或不利的结果	☐	☐	☐	☐	☐	☐	☐
AI 系统是可信的	☐	☐	☐	☐	☐	☐	☐

[矩阵量表 (续)]

陈述	1	2	3	4	5	6	7
我具备 AI 系统相关知识	☐	☐	☐	☐	☐	☐	☐
我接受过 AI 相关的教育/培训	☐	☐	☐	☐	☐	☐	☐
我有能够证明我具备 AI 知识的证书...	☐	☐	☐	☐	☐	☐	☐
我具备理解 AI 输出的技能	☐	☐	☐	☐	☐	☐	☐
我具备调试 AI 系统的技能	☐	☐	☐	☐	☐	☐	☐

Q10.2 下面是关于 AI 系统的一些知识，你对这些知识的了解程度如何？
答案没有对错之分，您对 AI 系统的了解程度不影响您答题的结果和奖励，
请根据您的实际了解程度进行作答。

(1= 非常不了解，2= 不了解，3= 比较不了解，4= 一般，5= 比较了解，6= 了解，7= 非常了解)

[矩阵量表]

知识点	1	2	3	4	5	6	7
对“AI 系统的局限性/缺陷”的了解	☐	☐	☐	☐	☐	☐	☐
对“为什么要在工作中使用 AI 系统”的了解	☐	☐	☐	☐	☐	☐	☐
对“在企业/单位中使用 AI 系统的好处”的了解	☐	☐	☐	☐	☐	☐	☐
对“在日常工作中使用 AI 系统的好处”的了解	☐	☐	☐	☐	☐	☐	☐
对“AI 系统如何提高工作效率”的了解	☐	☐	☐	☐	☐	☐	☐
对“如何使用 AI 系统高效完成分析性工作任务”的了解	☐	☐	☐	☐	☐	☐	☐

Q10.3 您的最高学历是（含在读）

[单选]

- ☐ 博士及以上
- ☐ 硕士
- ☐ 本科 / 大专
- ☐ 高中及以下

Q10.4 您目前的职业是

[单选]

- ☐ 企业研究人员
- ☐ 大学 / 科研院所教学研究人员
- ☐ 学生
- ☐ 企事业单位工作人员
- ☐ 公务员
- ☐ 其他（请注明） _____

Q10.5 您的专业领域是

[单选]

- ☐ AI/ 计算机相关（请列出您的专业） _____ ☐ 大学 / 科研院所教学研究人员
- ☐ 与 AI/ 计算机无关（请列出您的专业） _____ ☐ 企事业单位工作人员

Q10.6 您对本问卷中论文涉及的相关专业知识了解程度如何？

[量表]

[完全不了解 , 非常了解]

- ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

Q10.7 请问您在日常工作和生活中使用 AI 的频率如何？

[量表]

[完全不使用 , 使用非常频繁]

- ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6 ☐ 7

Q10.8 您之前的论文写作或审阅经验是：

[单选]

- ☐ 没有过论文写作或审阅经验

- ☐ 完成过 1-5 篇论文的写作或审阅工作
- ☐ 完成过 6-10 篇论文的写作或审阅工作
- ☐ 完成过 10 篇以上论文的写作或审阅工作

Q10.9 您的年龄是几岁？

[填空]

附录 A.2 实验一原始数据

附录A.2 实验一原始数据

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
1	B25	1	0	1	0.5	0	no_ai	control	first_two
2	B25	1	0	0	1.5	0	no_ai	control	first_two
3	B25	1	2	1	2.5	2	no_ai	control	first_two
4	B25	1	1	0	0.5	1	no_ai	control	first_two
5	B25	1	0	1	0.5	0	no_ai	control	first_two
6	B25	0	1	1	0.5	1	no_ai	control	first_two
7	B25	0	1	0	0.5	0	no_ai	control	first_two
8	B25	1	0	0	0.5	0	no_ai	control	first_two
9	B25	0	1	1	0.5	0	no_ai	control	first_two
10	B25	1	1	0	0.5	1	no_ai	control	first_two
11	B25	0	1	0	1.5	0	no_ai	control	first_two
12	B25	0	1	0	0.5	0	no_ai	control	first_two
13	B25	1	2	0	0.5	1	no_ai	control	first_two
14	B25	1	0	0	0.5	0	no_ai	control	first_two
15	B25	1	0	1	0.5	0	no_ai	control	first_two
16	B25	0	1	1	0.5	0	no_ai	control	first_two
17	B25	0	1	0	0.5	0	no_ai	control	first_two
18	B25	0	1	1	0.5	0	no_ai	control	first_two
19	B25	0	1	0	0.5	0	no_ai	control	first_two
20	B25	0	1	0	1.5	0	no_ai	control	first_two
21	B25	1	0	0	0.5	0	no_ai	control	first_two
22	B25	1	0	0	1.5	0	no_ai	control	first_two
23	B25	0	0	1	0.5	1	no_ai	control	first_two
24	B25	0	1	0	0.5	0	no_ai	control	first_two
25	B25	0	1	1	0.5	0	no_ai	control	first_two
26	B25	1	0	1	0.5	1	no_ai	control	first_two
27	B25	1	1	1	1.5	1	no_ai	control	first_two

附录A.2 实验一原始数据 (续)

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
28	B25	1	0	1	0.5	0	no_ai	control	first_two
29	B25	0	1	0	0.5	1	no_ai	control	first_two
30	B25	0	1	0	0.5	1	no_ai	control	first_two
31	C20	2.5	1.5	2	0.5	3	no_ai	control	first_two
32	C20	0.5	0.5	1	1.5	1	no_ai	control	first_two
33	C20	0.5	0.5	1	0.5	1	no_ai	control	first_two
34	C20	1.5	2.5	1	1.5	3	no_ai	control	first_two
35	C20	0.5	1.5	0	1.5	1	no_ai	control	first_two
36	C20	0.5	0.5	1	0.5	0	no_ai	control	first_two
37	C20	0.5	0.5	1	0.5	0	no_ai	control	first_two
38	C20	0.5	0.5	1	0.5	1	no_ai	control	first_two
39	C20	1.5	2.5	2	1.5	3	no_ai	control	first_two
40	C20	0.5	2.5	1	1.5	2	no_ai	control	first_two
41	C20	1.5	1.5	0	1.5	2	no_ai	control	first_two
42	C20	2.5	1.5	1	2.5	3	no_ai	control	first_two
43	C20	0.5	0.5	1	0.5	0	no_ai	control	first_two
44	C20	0.5	0.5	0	0.5	0	no_ai	control	first_two
45	C20	0.5	1.5	1	0.5	1	no_ai	control	first_two
46	C20	1.5	1.5	2	0.5	2	no_ai	control	first_two
47	C20	0.5	1.5	1	0.5	1	no_ai	control	first_two
48	C20	2.5	1.5	0	1.5	3	no_ai	control	first_two
49	C20	0.5	0.5	0	0.5	1	no_ai	control	first_two
50	C20	1.5	2.5	1	2.5	2	no_ai	control	first_two
51	C20	0.5	1.5	0	0.5	1	no_ai	control	first_two
52	C20	2.5	1.5	2	2.5	2	no_ai	control	first_two
53	C20	1.5	1.5	2	0.5	3	no_ai	control	first_two
54	C20	0.5	0.5	1	0.5	0	no_ai	control	first_two
55	C20	0.5	2.5	0	1.5	1	no_ai	control	first_two
56	C20	0.5	1.5	1	0.5	1	no_ai	control	first_two

附录A.2 实验一原始数据 (续)

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
57	C20	1.5	0.5	1	2.5	2	no_ai	control	first_two
58	C20	1.5	0.5	1	1.5	2	no_ai	control	first_two
59	C20	1.5	2.5	1	1.5	3	no_ai	control	first_two
60	C20	2.5	1.5	1	0.5	2	no_ai	control	first_two
61	A01	0	0	1	0.5	1	no_ai	control	last_three
62	A01	0	0	1	0.5	1	no_ai	control	last_three
63	A01	1	1	1	0.5	1	no_ai	control	last_three
64	A01	1	1	1	0.5	0	no_ai	control	last_three
65	A01	0	1	1	0.5	0	no_ai	control	last_three
66	A01	1	0	1	1.5	1	no_ai	control	last_three
67	A01	0	0	0	0.5	0	no_ai	control	last_three
68	A01	0	0	0	0.5	0	no_ai	control	last_three
69	A01	0	0	0	0.5	1	no_ai	control	last_three
70	A01	0	0	1	0.5	0	no_ai	control	last_three
71	A01	0	0	1	0.5	0	no_ai	control	last_three
72	A01	0	0	0	0.5	0	no_ai	control	last_three
73	A01	0	0	1	0.5	1	no_ai	control	last_three
74	A01	0	0	0	0.5	0	no_ai	control	last_three
75	A01	0	0	1	0.5	0	no_ai	control	last_three
76	A01	0	1	0	0.5	0	no_ai	control	last_three
77	A01	1	0	0	0.5	1	no_ai	control	last_three
78	A01	0	0	2	0.5	0	no_ai	control	last_three
79	A01	1	0	1	1.5	1	no_ai	control	last_three
80	A01	0	0	0	0.5	0	no_ai	control	last_three
81	A01	0	0	0	0.5	1	no_ai	control	last_three
82	A01	0	1	1	0.5	1	no_ai	control	last_three
83	A01	1	1	1	0.5	1	no_ai	control	last_three
84	A01	0	0	0	0.5	0	no_ai	control	last_three
85	A01	0	0	0	0.5	0	no_ai	control	last_three

附录A.2 实验一原始数据 (续)

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
86	A01	0	0	2	0.5	1	no_ai	control	last_three
87	A01	0	1	0	0.5	0	no_ai	control	last_three
88	A01	0	1	0	0.5	0	no_ai	control	last_three
89	A01	0	0	0	0.5	1	no_ai	control	last_three
90	A01	1	1	1	0.5	0	no_ai	control	last_three
91	B28	3	2	1	2.5	1	no_ai	control	last_three
92	B28	0	0	1	1.5	0	no_ai	control	last_three
93	B28	1	1	0	0.5	1	no_ai	control	last_three
94	B28	2	2	1	1.5	2	no_ai	control	last_three
95	B28	2	1	0	2.5	1	no_ai	control	last_three
96	B28	0	1	1	0.5	1	no_ai	control	last_three
97	B28	3	1	1	2.5	1	no_ai	control	last_three
98	B28	0	0	1	0.5	0	no_ai	control	last_three
99	B28	3	1	2	2.5	1	no_ai	control	last_three
100	B28	1	1	1	0.5	1	no_ai	control	last_three
101	B28	1	1	1	0.5	1	no_ai	control	last_three
102	B28	1	2	1	0.5	0	no_ai	control	last_three
103	B28	0	0	1	0.5	1	no_ai	control	last_three
104	B28	0	1	0	0.5	1	no_ai	control	last_three
105	B28	2	2	1	1.5	2	no_ai	control	last_three
106	B28	3	1	2	1.5	1	no_ai	control	last_three
107	B28	2	1	1	0.5	0	no_ai	control	last_three
108	B28	2	2	2	2.5	1	no_ai	control	last_three
109	B28	0	0	0	1.5	1	no_ai	control	last_three
110	B28	2	2	1	2.5	1	no_ai	control	last_three
111	B28	1	1	1	0.5	1	no_ai	control	last_three
112	B28	2	2	2	1.5	2	no_ai	control	last_three
113	B28	2	2	2	1.5	1	no_ai	control	last_three
114	B28	2	2	1	2.5	1	no_ai	control	last_three

附录A.2 实验一原始数据 (续)

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
115	B28	2	1	1	0.5	0	no_ai	control	last_three
116	B28	0	0	1	1.5	0	no_ai	control	last_three
117	B28	1	1	0	2.5	1	no_ai	control	last_three
118	B28	2	1	2	2.5	1	no_ai	control	last_three
119	B28	2	2	2	1.5	1	no_ai	control	last_three
120	B28	2	2	2	1.5	2	no_ai	control	last_three
121	C11	2.5	0.5	1.5	2	3	no_ai	control	last_three
122	C11	0.5	1.5	0.5	2	1	no_ai	control	last_three
123	C11	1.5	0.5	0.5	0	1	no_ai	control	last_three
124	C11	2.5	1.5	0.5	3	2	no_ai	control	last_three
125	C11	0.5	1.5	1.5	1	1	no_ai	control	last_three
126	C11	0.5	0.5	0.5	1	1	no_ai	control	last_three
127	C11	1.5	0.5	1.5	1	2	no_ai	control	last_three
128	C11	0.5	0.5	0.5	0	1	no_ai	control	last_three
129	C11	1.5	2.5	2.5	2	3	no_ai	control	last_three
130	C11	0.5	2.5	1.5	2	2	no_ai	control	last_three
131	C11	1.5	0.5	1.5	2	2	no_ai	control	last_three
132	C11	0.5	1.5	0.5	1	2	no_ai	control	last_three
133	C11	1.5	1.5	0.5	1	1	no_ai	control	last_three
134	C11	1.5	0.5	1.5	1	1	no_ai	control	last_three
135	C11	1.5	0.5	1.5	1	2	no_ai	control	last_three
136	C11	1.5	2.5	1.5	2	2	no_ai	control	last_three
137	C11	1.5	0.5	1.5	1	1	no_ai	control	last_three
138	C11	0.5	0.5	1.5	0	1	no_ai	control	last_three
139	C11	0.5	0.5	0.5	0	1	no_ai	control	last_three
140	C11	1.5	2.5	1.5	3	2	no_ai	control	last_three
141	C11	0.5	1.5	1.5	2	2	no_ai	control	last_three
142	C11	1.5	2.5	1.5	3	3	no_ai	control	last_three
143	C11	2.5	1.5	2.5	2	3	no_ai	control	last_three

附录A.2 实验一原始数据（续）

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
144	C11	1.5	0.5	1.5	2	2	no_ai	control	last_three
145	C11	0.5	0.5	1.5	0	0	no_ai	control	last_three
146	C11	0.5	0.5	0.5	0	0	no_ai	control	last_three
147	C11	0.5	0.5	1.5	1	1	no_ai	control	last_three
148	C11	1.5	2.5	2.5	3	3	no_ai	control	last_three
149	C11	1.5	0.5	1.5	1	1	no_ai	control	last_three
150	C11	1.5	2.5	1.5	2	3	no_ai	control	last_three
151	B25	0	0	1	0.5	0	no_ai	treatment	first_two
152	B25	1	0	1	0.5	1	no_ai	treatment	first_two
153	B25	0	1	0	0.5	1	no_ai	treatment	first_two
154	B25	0	1	0	0.5	0	no_ai	treatment	first_two
155	B25	1	1	1	1.5	1	no_ai	treatment	first_two
156	B25	1	2	1	0.5	1	no_ai	treatment	first_two
157	B25	1	0	0	0.5	0	no_ai	treatment	first_two
158	B25	1	0	1	1.5	1	no_ai	treatment	first_two
159	B25	0	0	1	0.5	0	no_ai	treatment	first_two
160	B25	1	2	0	0.5	1	no_ai	treatment	first_two
161	B25	2	1	1	2.5	2	no_ai	treatment	first_two
162	B25	1	0	0	1.5	0	no_ai	treatment	first_two
163	B25	0	0	1	1.5	0	no_ai	treatment	first_two
164	B25	0	1	0	0.5	0	no_ai	treatment	first_two
165	B25	1	0	0	0.5	0	no_ai	treatment	first_two
166	B25	0	0	0	0.5	1	no_ai	treatment	first_two
167	B25	0	1	2	1.5	1	no_ai	treatment	first_two
168	B25	0	0	0	1.5	0	no_ai	treatment	first_two
169	B25	0	1	0	0.5	0	no_ai	treatment	first_two
170	B25	0	0	0	0.5	0	no_ai	treatment	first_two
171	B25	1	1	0	0.5	1	no_ai	treatment	first_two
172	B25	0	0	1	0.5	0	no_ai	treatment	first_two

附录A.2 实验一原始数据 (续)

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
173	B25	0	1	0	1.5	0	no_ai	treatment	first_two
174	B25	2	1	1	0.5	0	no_ai	treatment	first_two
175	B25	1	0	0	0.5	0	no_ai	treatment	first_two
176	B25	0	0	1	0.5	0	no_ai	treatment	first_two
177	B25	0	0	0	0.5	0	no_ai	treatment	first_two
178	B25	2	0	1	1.5	1	no_ai	treatment	first_two
179	B25	2	1	2	0.5	2	no_ai	treatment	first_two
180	B25	1	2	1	0.5	1	no_ai	treatment	first_two
181	B25	1	0	0	0.5	0	no_ai	treatment	first_two
182	B25	0	0	1	0.5	0	no_ai	treatment	first_two
183	B25	0	0	1	0.5	0	no_ai	treatment	first_two
184	B25	1	1	0	0.5	1	no_ai	treatment	first_two
185	B25	1	0	1	0.5	0	no_ai	treatment	first_two
186	B25	1	0	1	0.5	0	no_ai	treatment	first_two
187	B25	1	2	1	0.5	1	no_ai	treatment	first_two
188	B25	1	1	0	0.5	0	no_ai	treatment	first_two
189	B25	0	1	0	0.5	0	no_ai	treatment	first_two
190	B25	0	1	2	1.5	2	no_ai	treatment	first_two
191	B25	1	0	0	0.5	0	no_ai	treatment	first_two
192	B25	0	0	0	0.5	0	no_ai	treatment	first_two
193	B25	1	0	0	0.5	0	no_ai	treatment	first_two
194	B25	1	0	1	1.5	1	no_ai	treatment	first_two
195	B25	1	0	1	0.5	0	no_ai	treatment	first_two
196	B25	0	0	1	0.5	0	no_ai	treatment	first_two
197	B25	1	2	1	2.5	2	no_ai	treatment	first_two
198	B25	1	0	1	1.5	0	no_ai	treatment	first_two
199	B25	0	1	1	0.5	0	no_ai	treatment	first_two
200	B25	0	0	1	0.5	1	no_ai	treatment	first_two
201	B25	0	1	1	0.5	0	no_ai	treatment	first_two

附录A.2 实验一原始数据 (续)

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
202	B25	0	0	1	0.5	0	no_ai	treatment	first_two
203	B25	1	1	0	0.5	0	no_ai	treatment	first_two
204	B25	1	0	1	0.5	1	no_ai	treatment	first_two
205	B25	0	1	0	0.5	0	no_ai	treatment	first_two
206	B25	1	1	0	0.5	0	no_ai	treatment	first_two
207	B25	0	0	0	0.5	0	no_ai	treatment	first_two
208	B25	0	1	0	0.5	0	no_ai	treatment	first_two
209	B25	0	0	1	0.5	0	no_ai	treatment	first_two
210	C20	0.5	1.5	1	0.5	1	no_ai	treatment	first_two
211	C20	1.5	1.5	2	1.5	3	no_ai	treatment	first_two
212	C20	0.5	0.5	1	0.5	1	no_ai	treatment	first_two
213	C20	0.5	0.5	0	0.5	1	no_ai	treatment	first_two
214	C20	0.5	1.5	0	1.5	2	no_ai	treatment	first_two
215	C20	0.5	1.5	1	0.5	2	no_ai	treatment	first_two
216	C20	0.5	1.5	0	0.5	1	no_ai	treatment	first_two
217	C20	0.5	0.5	0	0.5	0	no_ai	treatment	first_two
218	C20	1.5	0.5	1	0.5	1	no_ai	treatment	first_two
219	C20	1.5	1.5	0	1.5	2	no_ai	treatment	first_two
220	C20	1.5	2.5	0	1.5	2	no_ai	treatment	first_two
221	C20	1.5	1.5	0	1.5	1	no_ai	treatment	first_two
222	C20	1.5	2.5	1	2.5	2	no_ai	treatment	first_two
223	C20	0.5	1.5	0	0.5	2	no_ai	treatment	first_two
224	C20	0.5	1.5	0	0.5	1	no_ai	treatment	first_two
225	C20	0.5	0.5	1	0.5	1	no_ai	treatment	first_two
226	C20	0.5	0.5	1	1.5	0	no_ai	treatment	first_two
227	C20	2.5	2.5	1	1.5	2	no_ai	treatment	first_two
228	C20	0.5	0.5	0	0.5	1	no_ai	treatment	first_two
229	C20	0.5	0.5	0	0.5	1	no_ai	treatment	first_two
230	C20	0.5	0.5	0	0.5	0	no_ai	treatment	first_two

附录A.2 实验一原始数据 (续)

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
231	C20	1.5	1.5	0	0.5	2	no_ai	treatment	first_two
232	C20	1.5	2.5	0	1.5	2	no_ai	treatment	first_two
233	C20	0.5	1.5	0	0.5	1	no_ai	treatment	first_two
234	C20	1.5	1.5	0	1.5	2	no_ai	treatment	first_two
235	C20	0.5	0.5	1	0.5	1	no_ai	treatment	first_two
236	C20	0.5	1.5	0	1.5	2	no_ai	treatment	first_two
237	C20	1.5	1.5	2	1.5	2	no_ai	treatment	first_two
238	C20	0.5	1.5	1	0.5	1	no_ai	treatment	first_two
239	C20	0.5	0.5	1	0.5	0	no_ai	treatment	first_two
240	C20	0.5	0.5	0	0.5	0	no_ai	treatment	first_two
241	C20	1.5	2.5	1	2.5	2	no_ai	treatment	first_two
242	C20	0.5	0.5	0	0.5	1	no_ai	treatment	first_two
243	C20	0.5	0.5	0	0.5	1	no_ai	treatment	first_two
244	C20	0.5	1.5	1	0.5	2	no_ai	treatment	first_two
245	C20	2.5	1.5	0	0.5	2	no_ai	treatment	first_two
246	C20	0.5	1.5	1	1.5	2	no_ai	treatment	first_two
247	C20	1.5	2.5	0	0.5	1	no_ai	treatment	first_two
248	C20	1.5	1.5	0	1.5	2	no_ai	treatment	first_two
249	C20	0.5	1.5	1	0.5	1	no_ai	treatment	first_two
250	C20	0.5	0.5	0	0.5	0	no_ai	treatment	first_two
251	C20	0.5	0.5	0	0.5	1	no_ai	treatment	first_two
252	C20	1.5	0.5	0	1.5	2	no_ai	treatment	first_two
253	C20	1.5	1.5	2	1.5	2	no_ai	treatment	first_two
254	C20	1.5	0.5	0	0.5	0	no_ai	treatment	first_two
255	C20	0.5	0.5	0	0.5	0	no_ai	treatment	first_two
256	C20	0.5	1.5	1	0.5	1	no_ai	treatment	first_two
257	C20	1.5	1.5	2	0.5	2	no_ai	treatment	first_two
258	C20	1.5	2.5	0	1.5	1	no_ai	treatment	first_two
259	C20	1.5	2.5	0	1.5	2	no_ai	treatment	first_two

附录A.2 实验一原始数据 (续)

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
260	C20	0.5	2.5	0	1.5	1	no_ai	treatment	first_two
261	C20	1.5	0.5	1	2.5	2	no_ai	treatment	first_two
262	C20	1.5	1.5	0	0.5	1	no_ai	treatment	first_two
263	C20	0.5	0.5	1	0.5	1	no_ai	treatment	first_two
264	C20	0.5	1.5	1	0.5	1	no_ai	treatment	first_two
265	C20	0.5	1.5	0	1.5	2	no_ai	treatment	first_two
266	C20	0.5	0.5	0	0.5	1	no_ai	treatment	first_two
267	C20	1.5	1.5	1	0.5	2	no_ai	treatment	first_two
268	C20	0.5	0.5	1	0.5	1	no_ai	treatment	first_two
269	A01	0	0	1	0.5	1	ai	treatment	last_three
270	A01	0	1	1	0.5	1	ai	treatment	last_three
271	A01	0	0	1	0.5	0	ai	treatment	last_three
272	A01	0	0	0	0.5	0	ai	treatment	last_three
273	A01	1	1	0	0.5	0	ai	treatment	last_three
274	A01	1	0	2	0.5	1	ai	treatment	last_three
275	A01	1	0	0	0.5	1	ai	treatment	last_three
276	A01	0	0	1	0.5	1	ai	treatment	last_three
277	A01	0	0	0	0.5	2	ai	treatment	last_three
278	A01	0	0	1	0.5	1	ai	treatment	last_three
279	A01	0	1	1	0.5	0	ai	treatment	last_three
280	A01	1	0	2	0.5	1	ai	treatment	last_three
281	A01	0	0	0	0.5	0	ai	treatment	last_three
282	A01	1	1	1	2.5	1	ai	treatment	last_three
283	A01	0	0	0	0.5	0	ai	treatment	last_three
284	A01	1	0	1	0.5	1	ai	treatment	last_three
285	A01	1	0	0	1.5	1	ai	treatment	last_three
286	A01	0	0	1	0.5	1	ai	treatment	last_three
287	A01	1	0	0	0.5	1	ai	treatment	last_three
288	A01	1	0	1	0.5	1	ai	treatment	last_three

附录A.2 实验一原始数据 (续)

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
289	A01	0	0	1	0.5	1	ai	treatment	last_three
290	A01	0	0	2	1.5	1	ai	treatment	last_three
291	A01	1	1	1	1.5	1	ai	treatment	last_three
292	A01	0	0	1	0.5	0	ai	treatment	last_three
293	A01	0	0	1	1.5	1	ai	treatment	last_three
294	A01	0	0	0	0.5	1	ai	treatment	last_three
295	A01	0	0	1	0.5	0	ai	treatment	last_three
296	A01	1	1	1	0.5	0	ai	treatment	last_three
297	A01	0	0	1	0.5	0	ai	treatment	last_three
298	A01	0	0	0	0.5	0	ai	treatment	last_three
299	A01	0	0	0	0.5	1	ai	treatment	last_three
300	A01	0	0	0	0.5	0	ai	treatment	last_three
301	A01	0	0	0	0.5	1	ai	treatment	last_three
302	A01	0	0	1	0.5	0	ai	treatment	last_three
303	A01	0	1	0	0.5	0	ai	treatment	last_three
304	A01	1	1	1	0.5	1	ai	treatment	last_three
305	A01	0	0	2	0.5	1	ai	treatment	last_three
306	A01	0	1	1	0.5	0	ai	treatment	last_three
307	A01	0	0	0	0.5	0	ai	treatment	last_three
308	A01	1	1	2	1.5	2	ai	treatment	last_three
309	A01	0	0	1	0.5	0	ai	treatment	last_three
310	A01	0	0	0	0.5	0	ai	treatment	last_three
311	A01	0	0	1	0.5	0	ai	treatment	last_three
312	A01	0	0	1	0.5	0	ai	treatment	last_three
313	A01	0	0	0	0.5	0	ai	treatment	last_three
314	A01	0	0	0	0.5	0	ai	treatment	last_three
315	A01	1	1	2	0.5	1	ai	treatment	last_three
316	A01	2	0	2	0.5	1	ai	treatment	last_three
317	A01	0	1	1	0.5	0	ai	treatment	last_three

附录A.2 实验一原始数据 (续)

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
318	A01	0	0	1	0.5	0	ai	treatment	last_three
319	A01	0	0	0	0.5	0	ai	treatment	last_three
320	A01	0	0	1	0.5	0	ai	treatment	last_three
321	A01	0	1	1	0.5	0	ai	treatment	last_three
322	A01	0	1	1	0.5	0	ai	treatment	last_three
323	A01	0	0	1	0.5	0	ai	treatment	last_three
324	A01	1	0	1	0.5	1	ai	treatment	last_three
325	A01	0	1	1	0.5	0	ai	treatment	last_three
326	A01	1	1	2	0.5	1	ai	treatment	last_three
327	A01	1	1	1	0.5	0	ai	treatment	last_three
328	B28	1	1	0	2.5	0	ai	treatment	last_three
329	B28	3	1	2	1.5	2	ai	treatment	last_three
330	B28	2	1	1	0.5	1	ai	treatment	last_three
331	B28	1	0	1	0.5	0	ai	treatment	last_three
332	B28	2	1	0	2.5	1	ai	treatment	last_three
333	B28	0	0	0	0.5	0	ai	treatment	last_three
334	B28	1	1	1	0.5	1	ai	treatment	last_three
335	B28	1	1	0	1.5	0	ai	treatment	last_three
336	B28	2	1	0	1.5	1	ai	treatment	last_three
337	B28	2	1	0	1.5	1	ai	treatment	last_three
338	B28	1	1	0	1.5	0	ai	treatment	last_three
339	B28	2	0	1	1.5	1	ai	treatment	last_three
340	B28	2	2	0	2.5	1	ai	treatment	last_three
341	B28	0	0	0	0.5	0	ai	treatment	last_three
342	B28	1	1	0	1.5	0	ai	treatment	last_three
343	B28	1	1	1	0.5	0	ai	treatment	last_three
344	B28	1	1	1	0.5	0	ai	treatment	last_three
345	B28	2	1	2	2.5	1	ai	treatment	last_three
346	B28	1	1	1	0.5	0	ai	treatment	last_three

附录A.2 实验一原始数据 (续)

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
347	B28	1	1	1	1.5	1	ai	treatment	last_three
348	B28	1	1	1	2.5	1	ai	treatment	last_three
349	B28	2	0	1	0.5	0	ai	treatment	last_three
350	B28	2	1	0	0.5	0	ai	treatment	last_three
351	B28	1	0	1	0.5	0	ai	treatment	last_three
352	B28	1	1	1	0.5	1	ai	treatment	last_three
353	B28	1	0	1	0.5	0	ai	treatment	last_three
354	B28	2	1	0	1.5	1	ai	treatment	last_three
355	B28	2	1	2	2.5	2	ai	treatment	last_three
356	B28	0	1	1	0.5	1	ai	treatment	last_three
357	B28	1	1	0	1.5	0	ai	treatment	last_three
358	B28	1	1	1	0.5	1	ai	treatment	last_three
359	B28	2	2	1	2.5	1	ai	treatment	last_three
360	B28	2	0	1	0.5	0	ai	treatment	last_three
361	B28	1	1	2	1.5	1	ai	treatment	last_three
362	B28	1	1	2	1.5	1	ai	treatment	last_three
363	B28	1	1	1	1.5	0	ai	treatment	last_three
364	B28	1	1	0	1.5	1	ai	treatment	last_three
365	B28	1	1	1	2.5	1	ai	treatment	last_three
366	B28	2	0	0	1.5	1	ai	treatment	last_three
367	B28	1	1	1	0.5	0	ai	treatment	last_three
368	B28	0	0	0	1.5	1	ai	treatment	last_three
369	B28	2	1	1	1.5	1	ai	treatment	last_three
370	B28	1	1	0	1.5	0	ai	treatment	last_three
371	B28	1	1	1	1.5	1	ai	treatment	last_three
372	B28	0	0	0	0.5	0	ai	treatment	last_three
373	B28	0	0	0	0.5	1	ai	treatment	last_three
374	B28	1	1	1	2.5	1	ai	treatment	last_three
375	B28	2	0	0	0.5	0	ai	treatment	last_three

附录A.2 实验一原始数据（续）

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
376	B28	2	1	2	1.5	1	ai	treatment	last_three
377	B28	2	1	0	1.5	1	ai	treatment	last_three
378	B28	2	0	1	0.5	0	ai	treatment	last_three
379	B28	2	2	1	2.5	1	ai	treatment	last_three
380	B28	3	2	1	1.5	2	ai	treatment	last_three
381	B28	1	1	1	0.5	0	ai	treatment	last_three
382	B28	2	1	0	0.5	1	ai	treatment	last_three
383	B28	1	1	1	1.5	0	ai	treatment	last_three
384	B28	1	0	1	0.5	0	ai	treatment	last_three
385	B28	2	2	1	1.5	1	ai	treatment	last_three
386	B28	1	1	1	0.5	0	ai	treatment	last_three
387	C11	1.5	1.5	1.5	2	2	ai	treatment	last_three
388	C11	1.5	1.5	1.5	1	2	ai	treatment	last_three
389	C11	1.5	0.5	0.5	1	1	ai	treatment	last_three
390	C11	1.5	0.5	0.5	1	1	ai	treatment	last_three
391	C11	0.5	1.5	0.5	2	1	ai	treatment	last_three
392	C11	1.5	1.5	1.5	1	2	ai	treatment	last_three
393	C11	0.5	1.5	0.5	1	1	ai	treatment	last_three
394	C11	0.5	1.5	0.5	1	1	ai	treatment	last_three
395	C11	0.5	0.5	1.5	1	1	ai	treatment	last_three
396	C11	0.5	1.5	0.5	1	1	ai	treatment	last_three
397	C11	0.5	1.5	0.5	1	0	ai	treatment	last_three
398	C11	1.5	2.5	1.5	2	2	ai	treatment	last_three
399	C11	0.5	1.5	0.5	0	0	ai	treatment	last_three
400	C11	0.5	0.5	1.5	0	1	ai	treatment	last_three
401	C11	0.5	0.5	1.5	1	1	ai	treatment	last_three
402	C11	1.5	1.5	0.5	1	1	ai	treatment	last_three
403	C11	0.5	0.5	1.5	2	1	ai	treatment	last_three
404	C11	1.5	0.5	0.5	2	1	ai	treatment	last_three

附录A.2 实验一原始数据 (续)

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
405	C11	0.5	0.5	0.5	1	1	ai	treatment	last_three
406	C11	0.5	1.5	0.5	1	1	ai	treatment	last_three
407	C11	0.5	1.5	1.5	1	1	ai	treatment	last_three
408	C11	1.5	0.5	0.5	0	1	ai	treatment	last_three
409	C11	1.5	0.5	0.5	2	1	ai	treatment	last_three
410	C11	0.5	0.5	0.5	1	1	ai	treatment	last_three
411	C11	1.5	1.5	0.5	1	2	ai	treatment	last_three
412	C11	1.5	0.5	1.5	1	1	ai	treatment	last_three
413	C11	1.5	1.5	0.5	2	2	ai	treatment	last_three
414	C11	0.5	0.5	1.5	2	2	ai	treatment	last_three
415	C11	1.5	0.5	1.5	1	1	ai	treatment	last_three
416	C11	0.5	0.5	1.5	1	1	ai	treatment	last_three
417	C11	1.5	1.5	1.5	1	1	ai	treatment	last_three
418	C11	0.5	1.5	0.5	0	0	ai	treatment	last_three
419	C11	0.5	1.5	0.5	1	1	ai	treatment	last_three
420	C11	0.5	1.5	0.5	0	1	ai	treatment	last_three
421	C11	0.5	1.5	1.5	1	1	ai	treatment	last_three
422	C11	1.5	1.5	0.5	1	1	ai	treatment	last_three
423	C11	0.5	0.5	1.5	1	0	ai	treatment	last_three
424	C11	0.5	1.5	0.5	1	1	ai	treatment	last_three
425	C11	1.5	1.5	1.5	3	2	ai	treatment	last_three
426	C11	1.5	0.5	1.5	2	1	ai	treatment	last_three
427	C11	0.5	0.5	0.5	1	0	ai	treatment	last_three
428	C11	0.5	0.5	0.5	1	1	ai	treatment	last_three
429	C11	0.5	0.5	1.5	1	1	ai	treatment	last_three
430	C11	1.5	1.5	0.5	2	2	ai	treatment	last_three
431	C11	0.5	0.5	0.5	0	1	ai	treatment	last_three
432	C11	0.5	0.5	0.5	0	0	ai	treatment	last_three
433	C11	1.5	0.5	1.5	1	1	ai	treatment	last_three

附录A.2 实验一原始数据（续）

user_id	paper_id	richness	logic	writing	novelty	overall	condition	group	paper_group
434	C11	1.5	0.5	1.5	0	1	ai	treatment	last_three
435	C11	0.5	1.5	1.5	1	1	ai	treatment	last_three
436	C11	1.5	2.5	1.5	1	1	ai	treatment	last_three
437	C11	0.5	1.5	0.5	1	0	ai	treatment	last_three
438	C11	0.5	0.5	0.5	0	1	ai	treatment	last_three
439	C11	1.5	2.5	1.5	1	2	ai	treatment	last_three
440	C11	0.5	0.5	0.5	0	0	ai	treatment	last_three
441	C11	1.5	1.5	0.5	1	1	ai	treatment	last_three
442	C11	1.5	1.5	0.5	1	1	ai	treatment	last_three
443	C11	1.5	1.5	0.5	1	1	ai	treatment	last_three
444	C11	0.5	1.5	0.5	2	1	ai	treatment	last_three
445	C11	0.5	1.5	1.5	1	1	ai	treatment	last_three

附录 B

实验二：干预策略对人机协作决策影响的研究

附录 B.1 实验二研究工具（问卷）

实验条件：实验组（在特定维度强制思考 /AI 表现好维度强制思考）/ 控制组（全部维度 AI 辅助）

批次：1 / 2 / 3 / 4

您好，感谢您参与本研究项目。本研究由学术评审服务平台及合作研究团队联合开展，旨在评估大语言模型辅助的学术审阅流程与人机协作机制！

本问卷旨在测试计算机及相关专业的在读与毕业学生对学术会议论文的阅读理解与评估审阅能力。我们将邀请您在快速阅读论文后，从内容丰富性、论证逻辑、表达与写作、创新性与重要性和整体评价五个维度分别对论文进行打分。考虑到受试者背景的不同以及时间的限制，我们挑选了不太受专业知识限制，同时也无需精读论文，就可以评估的维度。每份有效问卷的作答者会收到 50 元现金作为奖励。

我们还会给那些在最短时间内给出最准确打分的问卷追加更多奖励。希望您仔细阅读相关论文并给予客观公正的评估。

完成问卷的过程中，请您不要离开或关闭浏览器页面，我们会综合考虑您的作答时间和准确度对您的表现进行评估。综合排名靠前的问卷会获得纪念品奖励，或者等价的现金奖励。同时，作答时间显著高于或者低于平均值的异常问卷会视

为废卷。在我们的预测试中，完成本问卷平均约耗费 35-50 分钟。

我们将实行严格匿名制度，确保对您所提供信息的严密保密；所有调查结果仅用于学术研究之用途。衷心感谢您宝贵的支持与帮助。

作答并提交本问卷将被视为您知晓、同意上述内容并自愿参与本研究与后续研究。如果您对本研究不感兴趣，可以现在退出作答。

期待您的认真回答！

Knowledge Check- 版块开始

Q1.1 以下两个问题将简单测试您对深度学习和人工智能的了解程度，错误回答会导致您的答卷被拒绝并无法参与后续研究，请仔细阅读并认真回答。

如果您无法正确回答以下问题，可能意味着本问卷内容和您的专业领域不匹配，为保证您的时间和精力不被浪费，建议您退出作答。

Q1.2 Which statement best describes the relationship between deep learning and neural networks ? [单选]

- ☐ Deep learning is a subset of neural networks
- ☐ Deep learning can only be used for computer vision tasks
- ☐ A neural network is the underlying algorithm used by deep learning
- ☐ Deep learning and neural networks are completely different technologies

Q1.3 What is the primary purpose of the backpropagation algorithm in deep learning ? [单选]

- ☐ To load the training dataset
- ☐ To visualize the neural network structure
- ☐ To update network weights based on prediction errors

○ To split data into training and testing sets

Q1.4 以上两道题目的正确答案如下：

Which statement best describes the relationship between deep learning and neural networks ?

正确答案： A neural network is the underlying algorithm used by deep learning

What is the primary purpose of the backpropagation algorithm in deep learning ?

正确答案： To update network weights based on prediction errors

如果您没有正确回答上述问题，可能意味着本问卷内容和您的专业领域不匹配，并且可能会导致您的回答被拒绝采纳。为保证您的时间和精力不被浪费，建议您退出作答。

Knowledge Check- 版块结束

Calibration - 论文 1 (A06)- 版块开始

Q2.1 首先，我们会向您展示 2 篇学术论文，并提供在内容丰富性、论证逻辑、表达与写作、创新性与重要性和整体评价五个关键评价维度上的评分标准，以及领域内专家根据上述标准给出的评分，请您仔细阅读。

在问卷的后续部分，您需要参考本部分的评分标准与专家评分，对另外 3 篇论文在上述五个维度给出您的评价。

Q2.2 请您仔细阅读第 1 篇论文，以及内容丰富性、论证逻辑、表达与写作、创新性与重要性，和整体评价共五个维度上的评分标准和专家据此给出的评分。

论文 1：潘兆升，陈子韬，蔡定均，等．通过语义感知排列训练减轻逆转诅咒
[C/OL]// 计算语言学协会会议论文集：EMNLP 2023. 宾夕法尼亚州斯特劳兹堡：
ACL, 2023:5972-5987.

内容丰富性维度的评分标准与专家评分：

- 1 分（很差）：文章内容非常单薄，缺乏必要的细节和深度，未能提供足够的信息，结构不完整，缺少关键部分，没有或缺乏说明性图表。
- 2 分（差）：文章内容较为简单，仅覆盖了一些基本要点，结构尚可但不够完整，信息量有限，图表较少且辅助说明作用不明显。
- 3 分（普通）：文章内容较为充实，覆盖了大部分相关要点，提供了一定量的信息，结构基本完整，但某些部分可能仍需进一步丰富和细化，图表数量适中，能较好地辅助文本内容。
- 4 分（良好）：文章内容丰富，详尽地探讨了研究问题，提供了大量的信息和数据支持，结构完整，图表丰富且有效地支持了文章的论点和分析。
- 5 分（优秀）：文章内容极为丰富和深入，全面地覆盖了研究领域的各个方面，结构严谨，内容层次分明，图表数量充足且设计精良，能够直观地展示研究结果和数据分析，极大地增强了文章的说服力和学术价值。
- 本文在该维度上的专家评分为：

o 专家 1: 5 分

o 专家 2: 4 分

论证逻辑维度的评分标准与专家评分：

- 1 分（很差）：文章的论点支离破碎，缺乏逻辑性，没有可辨识的逻辑结构。
- 2 分（差）：文章少数内容符合逻辑，但频繁出现逻辑跳跃，观点间缺少清晰联系。
- 3 分（普通）：文章大体上逻辑流畅，大部分观点之间关联明确，尽管某些论证可能并未完全展开或缺乏联系。
- 4 分（良好）：文章论证结构合理、思路清晰且逻辑性强，只有少数部分出现轻微的

连贯性问题。

· 5 分（优秀）：文章论证连贯，逻辑结构十分严密，每一部分都与上一部分紧密联结。

· 本文在该维度上的专家评分为：

o 专家 1: 5 分

o 专家 2: 4 分

表达与写作维度的评分标准与专家评分：

· 1 分（很差）：文章缺乏流畅性，既不清晰也不简洁，在许多关键部分用词不准确或出现严重语法错误，行文不符合学术写作风格，阻碍了读者的理解。

· 2 分（差）：文章有些部分比较清晰和流畅，但整体来说比较冗长，一些用词或语法错误也给读者理解论文内容造成了困难。

· 3 分（普通）：文章大体上流畅、清晰且简洁，只有少量用词不当或语法错误，行文基本符合学术写作风格

· 4 分（良好）：文章表达流畅且清晰，核心部分阐述简洁明了，基本没有用词或语法错误，遵循学术写作的风格。

· 5 分（优秀）：文笔十分流畅，表达出色，清晰且简洁有力，深入浅出地阐述复杂论点，完全没有用词或语法错误，严格遵循了学术写作风格。

· 本文在该维度上的专家评分为：

o 专家 1: 5 分

o 专家 2: 4 分

创新性与重要性维度的评分标准与专家评分：

· 1 分（很差）：文章缺乏原创思想或对该领域的贡献，未能提供任何新见解或改进。

· 2 分（差）：文章的新颖性和意义极为有限，仅在已知概念上做出轻微改变，没有为该领域增加实质研究价值。

· 3 分（普通）：文章引入了一些新颖的观点或方法，为该研究领域做出了一定贡献。

· 4 分（良好）：文章具有高度的创新性且比较重要，推动了该研究领域的发展与进步。

- 5 分（优秀）：论文具有开创性，提供了全新且意义重大的贡献，大大推动了领域内最先进的技术进步。

- 本文在该维度上的专家评分为：

- o 专家 1: 5 分

- o 专家 2: 4 分

整体评价维度的评分标准与专家评分：

- 1 分（很差）：论文存在重大缺陷，除非经过重大修订，否则无法发表。

- 2 分（差）：论文具有一定价值，但需要进行重大改进才能发表。

- 3 分（普通）：论文可以接受但并不突出；它满足基本标准，但缺乏重要贡献或存在比较显著的问题。

- 4 分（良好）：论文在大多数方面表现良好，但是可能存在一些小问题。

- 5 分（优秀）：论文在所有方面都表现出色，显然值得接受，最多只需进行轻微修订。

- 本文在该维度上的专家评分为：

- o 专家 1: 5 分

- o 专家 2: 4 分

Q2.3 请问您是否已经认真阅读完本篇论文？ [单选]

☐ 我已经认真阅读完本篇论文

☐ 我已经仔细阅读完毕上述评分标准和专家评分。

Calibration - 论文 1 (A06)- 版块结束

Calibration - 论文 2 (C19)- 版块开始

Q3.1 请您仔细阅读第 2 篇论文，以及内容丰富性、论证逻辑、表达与写作、创新性与重要性，和整体评价共五个维度上的评分标准和专家据此给出的评分。

论文 2: 陈宇航, 王静, 刘子豪, 等. 基于改进 U-Net 的无人机航拍图像分割方法 [C]// 计算机图形学、人工智能和数据处理国际会议 (ICCAID 2023). 华盛顿州 贝灵厄姆: SPIE, 2023: 1310503.

内容丰富性维度上的专家评分为:

- 专家 1: 3 分
- 专家 2: 3 分

论证逻辑维度上的专家评分为:

- 专家 1: 3 分
- 专家 2: 4 分

表达与写作维度上的专家评分为:

- 专家 1: 3 分
- 专家 2: 3 分

创新性与重要性维度上的专家评分为:

- 专家 1: 3 分
- 专家 2: 3 分

整体评价维度上的专家评分为:

- 专家 1: 3 分
- 专家 2: 3 分

Q3.2 请问您是否已经认真阅读完本篇论文? [单选]

- ☐ 我已经认真阅读完本篇论文
- ☐ 我已经仔细阅读完毕上述评分标准和专家评分。

Calibration - 论文 2 (C19)- 版块结束

任务前主观难度评价 - 1- 版块开始

Q4.1 通过阅读上述论文以及评分标准和专家评分，您预计对论文的哪些维度进行评价会花费您最多的精力？请根据花费精力的多少，请将您认为最耗费精力的维度设置为 10 分，其他维度则根据其相对于最高分项的精力投入比例来调整分数。请注意，对应 10 分的维度不能超过 2 个，且至少 1 个。

[矩阵量表]

(表格内容见 Q3.1)

任务前主观难度评价 - 1- 版块结束

有 AI - 论文 3 (A16)- 版块开始

Q5.1 请仔细阅读以下论文，结合每个维度的评分标准，并参考之前的专家打分回答问题。

论文 3: ANGELOPOULOS A, BOYEAU P, CAI T, 等 . 正确完成自动评估:

使用合成数据进行模型评估 [R/OL]. (2023-11-13) [2024-05-23]. <http://arxiv.org/abs/2311.07981>.

(后续问题要求参与者在阅读 AI 生成的评价后，对论文进行评分，结构与 Calibration 部分类似。)

有 AI - 论文 4 (B11)- 版块开始

论文 4: KUMAR S. 面向高频做市的深度强化学习 [J/OL]. 2022. <http://arxiv.org/abs/2208.13543>.

(后续问题结构同上)

有 AI - 论文 5 (C11)- 版块开始

论文 5: WANG H, SU J, JIN T, 等 . 基于 FA-SVM 算法的分类器模型在语音识别中的

应用 [J/OL]. Journal of Physics: Conference Series, 2023, 2562(1): 012019.

(后续问题结构同上)

主观难度评价与综合评价 - 版块开始

(此部分为主观题, 格式与无 AI 部分相同, 但会询问在 AI 辅助下的体验。)

控制变量 - 版块开始

(此部分为主观题, 收集参与者背景信息, 与无 AI 部分完全相同。)

附录 B.2 实验二原始数据及补充数据

附录 B.2.1 参与者人口统计与背景信息表

user_id	group	AI_trust_ 1	AI_trust_ 2	AI_trust_ 3	AI_trust_ 4	AI_trust_ 5	AI_skill_ 6	AI_skill_ 7	AI_skill_ 8	AI_skill_ 9
1	1	2	1	3	1	6	7	5	5	6
2	1	1	2	2	1	6	7	6	6	7
3	1	5	4	5	5	4	5	5	6	4
4	1	5	3	3	3	3	4	3	3	3
5	1	2	1	3	2	7	6	6	6	5
6	1	1	3	2	3	6	7	6	5	5
7	1	2	2	3	2	6	6	7	6	6
8	1	4	3	5	4	5	5	6	3	5
9	1	2	1	1	3	6	5	5	7	7
10	1	1	2	1	1	6	7	6	6	7
11	2	3	2	5	3	5	7	6	6	7
12	2	2	1	2	1	6	7	5	6	7
13	2	5	4	5	4	4	6	6	5	6
14	2	2	1	2	1	6	7	5	6	6
15	2	3	3	5	3	5	6	5	4	6
16	2	5	5	5	6	4	5	5	4	5

附录 B.2.1 参与者人口统计与背景信息表 (续)

user_id	group	AI_trust_ 1	AI_trust_ 2	AI_trust_ 3	AI_trust_ 4	AI_trust_ 5	AI_skill_ 6	AI_skill_ 7	AI_skill_ 8	AI_skill_ 9
17	2	4	5	4	3	5	5	4	3	5
18	2	5	4	4	4	5	4	4	3	5
19	2	5	5	6	5	4	7	5	3	5
20	2	3	2	6	3	5	6	3	2	6
21	3	2	1	5	2	6	6	4	2	5
22	3	4	3	5	4	4	7	6	5	6
23	3	1	2	1	2	7	6	7	6	6
24	3	3	2	5	3	5	6	5	5	6
25	3	2	5	5	3	5	5	6	4	5
26	3	4	3	5	4	3	5	5	4	5
27	3	5	3	6	4	5	6	6	4	6
28	3	4	5	6	4	4	5	5	2	6
29	3	5	4	5	4	4	6	6	6	4
30	3	2	2	6	4	5	5	6	6	6
31	4	3	2	3	3	6	6	5	5	7
32	4	3	2	3	2	6	6	5	5	7
33	4	5	4	6	4	5	5	5	4	6
34	4	3	2	2	3	6	6	5	3	5
35	4	2	1	1	2	6	6	5	6	5
36	4	2	3	4	2	5	4	5	4	6
37	4	3	4	5	3	5	6	7	7	7
38	4	2	2	4	1	6	6	3	2	6
39	4	3	3	5	5	4	4	4	1	4
40	4	3	3	5	3	4	3	5	2	5

附录 B.2. 参与者人口统计与背景信息表 (续)

User	AI	AI	AI	AI	AI	AI	AI	AI	degree	major	knowledge	Paper	Q11
id	Skill	Knowledge	Knowledge	Knowledge	Knowledge	Knowledge	Knowledge	Use				exp	9
	10	1	2	3	4	5	6	freq					
1	5	3	6	7	5	7	6	6	2	1	5	1	32
2	6	5	7	6	6	7	6	6	2	1	7	2	25
3	5	5	6	6	7	7	7	6	2	0	5	1	31
4	4	4	4	4	4	5	4	5	3	1	5	1	28
5	7	5	6	5	6	5	6	6	1	1	6	3	32
6	6	5	7	6	7	6	6	6	2	1	6	3	31
7	5	5	6	5	6	5	5	6	1	1	5	2	32
8	6	5	5	6	6	5	6	5	1	1	4	1	21
9	7	5	6	6	5	7	7	6	2	1	5	2	35
10	5	6	7	6	6	7	6	6	2	1	6	2	25
11	6	6	6	7	6	5	6	6	2	1	6	3	27
12	6	5	7	5	7	6	5	6	2	1	6	2	28
13	5	5	6	7	5	6	7	7	3	1	6	3	28
14	7	5	6	7	6	7	6	6	1	1	7	2	26
15	5	5	6	7	7	6	6	4	2	0	4	2	25
16	4	5	5	5	5	6	5	6	2	1	4	1	22
17	4	5	5	5	6	5	4	6	2	1	5	1	24
18	5	5	5	5	5	5	4	5	2	1	5	2	22
19	4	5	6	7	6	5	7	5	2	1	3	2	36
20	6	5	6	6	6	6	5	5	2	1	4	0	23
21	5	4	5	6	5	6	5	5	1	1	5	1	36
22	5	6	5	6	6	7	7	6	3	1	4	2	26

附录 B.2.1 参与者人口统计与背景信息表 (续)

User	AI	AI	AI	AI	AI	AI	AI	AI	degree	major	knowledge	Paper	Q11
id	Skill	Knowledge	Knowledge	Knowledge	Knowledge	Knowledge	Knowledge	Use				exp	9
	10	1	2	3	4	5	6	freq					
23	7	5	6	6	7	6	7	7	1	1	7	3	36
24	4	5	6	6	7	6	5	7	2	1	6	2	29
25	3	5	6	6	6	7	5	6	1	0	3	2	26
26	4	5	6	5	6	7	6	6	2	0	4	1	22
27	6	5	6	6	6	5	5	6	2	1	3	0	23
28	5	5	5	5	6	6	5	6	2	1	5	1	23
29	4	4	5	5	5	5	5	6	2	1	5	1	22
30	3	3	5	5	6	7	5	6	2	1	5	1	22
31	6	6	6	7	6	7	7	7	2	0	5	3	36
32	6	5	6	7	6	7	7	7	2	0	5	2	34
33	3	5	6	5	5	6	5	5	2	0	5	2	23
34	5	5	6	6	7	6	5	5	1	1	5	1	32
35	6	5	6	7	7	6	7	6	2	1	6	2	35
36	6	5	7	6	7	6	5	6	2	0	4	2	26
37	7	6	6	6	6	6	7	6	2	0	5	2	23
38	5	6	6	7	5	6	5	6	1	1	4	1	23
39	3	4	5	5	5	6	5	6	2	1	3	1	22
40	2	3	5	5	6	6	6	6	2	1	5	1	22

附录 B.2.2 论文评分数据

paper_id	dimension	before_diff	after_diff	improvement	group
A16	richness	0.6	0.6	0	1
A16	logic	0.8	0.8	0	1
A16	writing	0.8	0.8	0	1
A16	novelty	0.3	0.3	0	1
A16	overall	0.8	0.8	0	1
B11	richness	0.4	0.4	0	1
B11	logic	0.7	0.7	0	1
B11	writing	2	2	0	1
B11	novelty	1.1	1.1	0	1
B11	overall	0.7	0.7	0	1
C11	richness	0.8	0.8	0	1
C11	logic	1.4	1.4	0	1
C11	writing	0.9	0.9	0	1
C11	novelty	1.3	1.3	0	1
C11	overall	1.3	1.3	0	1
A16	richness	0.9	0.9	0	2
A16	logic	1.3	1	0.3	2
A16	writing	0.7	0.7	0	2
A16	novelty	0.7	0.3	0.4	2
A16	overall	0.8	0.8	0	2
B11	richness	0.7	0.7	0	2
B11	logic	1	0.6	0.4	2
B11	writing	2.4	2.4	0	2
B11	novelty	1	1.3	-0.3	2
B11	overall	0.7	0.7	0	2
C11	richness	1.1	1.1	0	2
C11	logic	1.2	1.3	-0.1	2
C11	writing	0.6	0.6	0	2
C11	novelty	1.4	1.3	0.1	2
C11	overall	1.5	1.5	0	2
A16	richness	1	0.7	0.3	3
A16	logic	0.8	0.8	0	3

附录 B.2.2 论文评分数据 (续)

paper_id	dimension	before_diff	after_diff	improvement	group
A16	writing	1.2	1.2	0	3
A16	novelty	0.6	0.6	0	3
A16	overall	0.8	0.8	0	3
B11	richness	0.3	0.6	-0.3	3
B11	logic	0.3	0.3	0	3
B11	writing	2.1	2.1	0	3
B11	novelty	1.5	1.5	0	3
B11	overall	0.9	0.6	0.3	3
C11	richness	0.5	1	-0.5	3
C11	logic	1	1	0	3
C11	writing	0.9	0.9	0	3
C11	novelty	1.3	1.3	0	3
C11	overall	1.3	1.2	0.1	3
A16	richness	1.2	0.8	0.4	4
A16	logic	1.1	0.8	0.3	4
A16	writing	0.9	1.3	-0.4	4
A16	novelty	0.9	0.6	0.3	4
A16	overall	0.8	1	-0.2	4
B11	richness	0.2	0.6	-0.4	4
B11	logic	1.3	0.8	0.5	4
B11	writing	2.3	2.2	0.1	4
B11	novelty	1.5	1.5	0	4
B11	overall	1.2	0.9	0.3	4
C11	richness	0.5	0.6	-0.1	4
C11	logic	1.2	1.2	0	4
C11	writing	1.7	1.2	0.5	4
C11	novelty	1.6	1.5	0.1	4
C11	overall	1.8	1.7	0.1	4

附录 C

提示语框架与参数设定

附录 C.1 模型与生成参数

模型：GPT-4 Turbo（全流程一致）。

temperature：0.3（降低随机性，提升判准一致性）。

top_p=1（与 temperature 不同时调整；本研究固定为 1 以避免双重抽样）。

max_tokens：4096（保证段落化输出与完整评审要点）。

（可选）frequency_penalty / presence_penalty: 0 / 0（不额外惩罚重复，以维持评审术语稳定）。

输出格式：Markdown 结构化清单 + 表格（便于量化抽取与一致性比对）。

实验一致性：上述参数在所有条件与回合中保持不变。

附录 C.2 结构化提示语模板（中文）

角色与目标：你是一名学术会议 / 期刊的审稿人，需依据给定论文节选与审稿维度做出独立判断与 5 分制评分。

评审维度（与正文一致）：内容丰富性、论证逻辑、表达与写作、创新性与重要性、整体评价。

评分刻度与锚点（1-5）：沿用正文表 4 之定义（不可扩充新尺度）。

输出模板（要求严格按序）：

要点清单：每一维度 2-3 条具体观察（避免空泛形容词）。

打分汇总表：五维度分数（1-5），附 1 行整体评语（≤ 80 字）。

可操作建议：每维度 1 条可落地修改建议（动词开头）。

行为与限制：

不得虚构引用与数据；不根据外部未提供信息推断细节；不输出作者身份或机构猜测。

保持中立语气，避免绝对化结论；必要时给出「不确定性」来源。

反思触发（适用于有反思条件的组别）：当任一维度分数与「参考要点」存在明显不一致（≥ 2 分）时，补充 1 条反例检索 / 证据核验语句，说明是否调整分数及原因。

附录 C.3 Structured Prompt (English Skeleton)

Role & Goal: You are a peer reviewer. Assess the manuscript excerpt on five criteria (Content Richness, Argumentation, Writing, Novelty & Significance, Overall). Use a 1-5 scale with the predefined anchors.

Output schema (strict order):

(1) Bullet points: 2-3 concrete observations per criterion.

(2) Score table: five scores (1-5) + one-sentence overall remark (≤ 80 words).

(3) Actionable suggestions: one per criterion, imperative verbs.

Rules: No fabrication, no external speculation, neutral tone, report uncertainties.

(If reflection is enabled): when your score deviates from the reference by ≥ 2 points, add one evidence check note stating whether you adjust the

score and why.

附录 C.4 提示语实例（节选，中文）

任务：请基于提供的论文节选，按「内容丰富性、论证逻辑、表达与写作、创新性与重要性、整体评价」五维度进行审阅，并使用 1-5 分制（见锚点）。

输出：

要点清单（每维 2-3 条具体观察）；

打分汇总表（五维度分数 + 80 字以内整体评语）；

可操作修改建议（每维 1 条）。

限制：不得虚构引用 / 数据；不对作者身份与来源做推测。必要时说明不确定性。

（反思条件）：若你的任一分数与参考要点差距 ≥ 2 分，请补充「反例检索 / 证据核验」一句，并说明是否调整分数与原因。

提示语框架综合参考了国际主流会议之审稿指南与评审要点（如 NeurIPS、ACL、CVPR、ICLR 等）对于维度定义、尺度锚点、冲突处理与伦理规范的通行做法，并结合本研究之任务场景与资料边界加以标准化，以保证跨实验条件下的可比性与可重复性。



ISBN 978-9-08364-461-5



9 789083 644615