■ 时论与争鸣

DOI:10.15998/j.cnki.issn1673-8012.2025.05.010

走向人机协同:大语言模型赋能学术论文评审的效能与边界

焦丽珍[1,2],刘　选[3],刘俊杰[4],钟洪蕊[5]

(1.清华大学　信息化技术中心,北京　100084;2.数字化学习技术集成与应用教育部工程研究中心,北京　100081;
3.四川开放大学,成都　610073;4.中央民族大学　教育学院,北京　100081;
5.广州星河汇投资有限公司,广州　511356)

摘　要:大语言模型在学术论文评审领域展现出巨大潜力,但其评审的效能及边界仍有待深入探索。以某教育技术类期刊为例,探索 DeepSeek 大语言模型协助学术论文评审的效能。研究发现,DeepSeek 有助于提升稿件初审的效率和准确性,在评审论文的内在逻辑方面具有较高的可信度,并能为论文的修改和完善提供一定的借鉴意见;与人类专家相比,DeepSeek 存在明显缺陷,如对学术论文评审"一板一眼"、创新性分析深度不足、提供的反馈建设性和适切性不足等。为更好地利用大语言模型赋能学术论文评审,切实为评审实践、学科发展和学术繁荣贡献力量,学界应恪守学术伦理,达成行业共识;加强数据安全,强化高质量数据集,构建垂直领域大模型;开展多元化人机协同评审,构筑机器赋能智能评审的边界。

关键词:大语言模型;DeepSeek;学术论文评审;智能评审;人机协同;评价标准

[**中图分类号**]G649.21;G473.8　[**文献标志码**]A　[**文章编号**]16738012(2025)05011909

感谢清华大学钟晓流、浙江大学翟雪松,江苏师范大学杨现民、李新,中央民族大学魏顺平、卢雨婷,曲阜师范大学徐振国,浙江师范大学逯行等专家对本文提供的帮助。

修回日期:20250603

基金项目:数字化学习技术集成与应用教育部工程研究中心创新基金项目"人工智能赋能教育研究成果评价的实现路径研究"(1421009)

作者简介:焦丽珍,女,山西应县人,清华大学信息化技术中心期刊编辑高级专员,主要从事教育技术和数字出版研究;
刘俊杰,男,河南南阳人,中央民族大学教育学院博士生,主要从事人工智能教育研究;
钟洪蕊,男,山东济宁人,广州星河汇投资有限公司战略执行经理,从事教育数字化转型研究。

通信作者:刘选,男,山西右玉人,四川开放大学研究员,博士,主要从事人工智能、数据治理和数字出版研究。

引用格式:焦丽珍,刘选,刘俊杰,等.走向人机协同:大语言模型赋能学术论文评审的效能与边界[J].重庆高教研究,2025,13(5):119-127.

Citation format:JIAO Lizhen,LIU Xuan, LIU Junjie,et al. Towards human-machine collaboration: efficacy and boundaries of large language models in empowering academic paper review[J]. Chongqing higher education research,2025,13(5):119-127.

一、问题提出

学术论文评审体系的发展与完善是学术界与出版界的核心关切。传统的同行评议方式虽然受到学界的高度认可,但也存在审稿周期长、学术出版滞后等问题,难以适应数智时代知识的更迭速度[1]。ChatGPT、DeepSeek 等大语言模型因其具备的强大文本语义理解能力,为学术论文评审带来新机遇[2]。研究表明,大语言模型在学术论文的语法格式检查、内容全面评审及创新性评估等方面均表现出较大潜力,可减少编辑与审稿专家在基础性内容审查与格式方面花费的时间,大幅提升论文评审效能。同时,大语言模型经过“细粒度评审推理链”数据集的微调与结构化的评审流程训练后,能够完成对学术论文结构框架、文献引用全面性等方面的评审任务,甚至能给出与人类审稿专家具有高度一致性的评审意见[3]。此外,这种基于大语言模型的自动化学术论文评审方式还进一步降低了评审的主观性,评审更加公平、高效[4]。目前,许多期刊已经尝试将大语言模型融入抄袭检测[5]、审稿人选择与分配[6]以及统计数据核查[7]等环节,推动学术论文评审质量与效能的双重提升。然而,大语言模型裹挟而来的算法偏见、算法黑箱、生成幻觉等现实问题,也易带来智能评审质量与结果存在偏差,导致其在实际应用中仍存在诸多挑战。例如,刘俏亮等使用多个大语言模型对学术论文格式、研究主题匹配度、论文重复率进行评价,结果发现,部分大语言模型的评审准确率不足 60%[8]。究其根本,提示词作为大语言模型的核心输入与运行基点,不仅是其融入学术论文评审过程的核心,也是决定评审结果可靠性与有效性的关键。实践中,许多编辑或审稿专家设计的提示词过于简单或模糊,难以切实表达实际的审稿要求,导致评审结果的准确性存在偏差。此外,系统化、全面性的论文评审框架的“缺位”,也导致大语言模型始终以“黑箱”方式评审论文,存在评审结果逻辑性欠缺、评审维度不全面以及评审结果不可重复等问题。

本研究尝试探索大语言模型在学术论文评审应用中的明确路径,通过分析其评审学术论文的效能和适用边界,解决大语言模型融入论文评审过程中面临的提示词设计缺陷与评审框架缺位的双重困境。基于此,本研究通过编制《论文评审框架指标体系》、设计详细的提示词框架及提示词,使大语言模型遵循与人类专家一致的学术评审逻辑进行论文审稿,同时将大语言模型融入学术论文的初审和提供评审意见环节,明确大语言模型赋能学术论文评审的效能与边界。

二、研究设计

(一)研究对象

本研究以某教育技术类期刊为研究对象。该期刊作为一本学科交叉性比较强的学术刊物,审稿环节包括初审、专家外审、复审、终审等。鉴于研究可行性和学术伦理的综合考量,本研究主要进行两个方面的实证检验:一是以 2023 年该期刊远程采编平台中的全部稿件为样本,验证大语言模型在协助编辑初审稿件方面的优势及其有效性;二是随机选取该刊 2023 年已经录用的 30 篇学术论文作为样本,验证大语言模型进行同行评议的评审效果。每篇学术论文的字数均在 10 000 字以上,且均为作者投稿时的原稿,未做任何修改和调整。需要说明的是,基于伦理考量,本研究未将大语言模型直接应用于外审和终审等评审环节,而是将经过人类专家外审、复审与终审的“录用稿件”提供给大语言模型进行再次评审,并将人类专家的评审意见作为基准,对比分析大语言模型赋能学术论文评审的有效性。

(二)工具选择

为比较大语言模型在学术论文智能评审中的实际效果,本研究选择 DeepSeek 作为智能评审工具,并通过以下 3 个步骤对研究样本进行智能评审(如图 1):(1)通过“摘要+关键词”验证 DeepSeek

在初审环节的有效性;(2)系统设计并验证《论文评审框架指标体系》,用于支持设计有针对性的提示词,使 DeepSeek 对学术论文评审更具针对性;(3)验证 DeepSeek 在评审意见方面的优势与有效性。

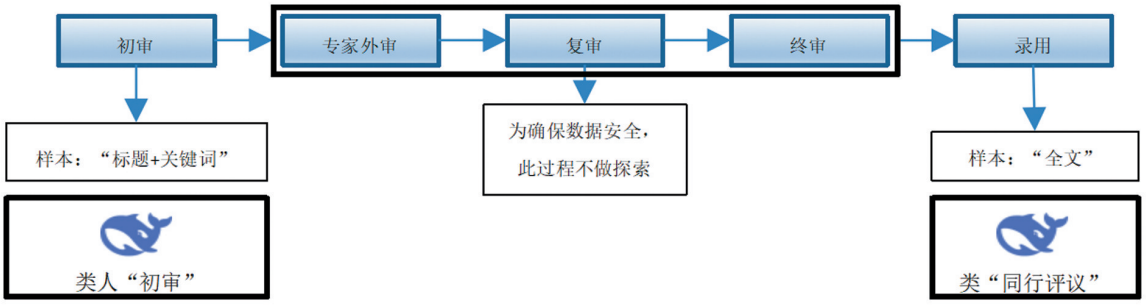

图 1　DeepSeek 赋能论文评审流程图

(三)论文评审框架和标准制定

本研究团队结合案例期刊论文的评审标准,编制《论文评审框架指标体系》,包括选题价值与前沿性、理论深度与逻辑严谨性、研究方法科学性、研究结论、学术规范与表述质量 5 个一级指标。为验证《论文评审框架指标体系》的有效性,本研究采用德尔菲法与层次分析法相结合的方式开展了两轮专业的意见征询。基于对专家规模与征询效率的考量,本研究最终遴选 12 位专家组成意见征询小组,其中包括 11 位期刊审稿专家和 1 位期刊编辑,这些专家均具有副高级及以上职称,在期刊论文评审方面有丰富的工作经验,能够确保意见征询结果的可靠性与全面性。

表 1　各层次指标的权重分析结果

一级指标	权重/%	二级指标	权重/%
选题价值与研究内容(28 分)	28	问题的明确性和现实指向性(13 分)	45.68
		研究内容的创新性(9 分)	31.35
		教育实践/政策的现实意义(6 分)	22.97
研究深度与研究逻辑(23 分)	23	理论基础的适切性(8 分)	36.62
		研究假设/问题的逻辑自洽性(10 分)	43.12
		文献综述的全面性与批判性(5 分)	20.26
研究过程与研究方法(20 分)	20	研究方法适切性(7 分)	36.45
		数据来源的可靠性与样本代表性(9 分)	42.50
		技术工具与分析方法的合理性(4 分)	21.05
研究结论(14 分)	14	研究结论的合理性与可推广性(7 分)	51.28
		结论的凝练性与分析的深刻性(7 分)	48.72
学术伦理与规范(15 分)	15	术语使用的规范性(5 分)	33.22
		学术伦理的规范性(5 分)	32.96
		语言表达的准确性(5 分)	33.82

确定专家小组后,本研究编制《论文评审框架指标体系》专家征询问卷,在问卷中主要介绍研究的背景与目的,确保各位专家了解本研究的实施情境。同时,收集每位专家的职称、研究领域等基本信息,最后要求每位专家对指标体系的熟悉程度和打分依据进行自我评价。在第一轮意见征询中,通过微信向所有专家发送初步的《论文评审框架指标体系》问卷,共回收有效问卷 12 份,回收率为 100%。参照德尔菲法的基本原则,结合专家的评审意见对《论文评审框架指标体系》中的指标进行

增加、删除或修改。在第二轮意见征询中,采用层次分析法对修改后的《论文评审框架指标体系》进行验证,并构建论文评审框架指标体系的层次结构模型,采用“1-9 标度法”进一步设计指标重要性程度比较的专家意见征询问卷。专家需对各一级指标下的所有二级指标进行两两比较,用于表示不同指标间的重要性程度。在第二轮意见征询中,我们通过问卷星再次咨询 12 位专家的意见并收集了重要性程度评分结果。通过构建专家打分表的判断矩阵进行评分一致性检验,结果显示一级指标与各二级指标的 CR 值介于 0.001-0.024,均小于 0.1,表明专家意见具有高度内在一致性。最后,通过特征向量法计算出《论文评审框架指标体系》各指标权重系数(见表 1)。

(四)提示词设计

提示词是指向大语言模型提出的具体问题和指令,它们作为初始输入,向大语言模型传递需要完成的任务,并引导大语言模型生成相应的输出。有效的提示词能够帮助大语言模型更准确地理解用户需求,进而生成符合预期的高质量内容。一般而言,提示词要清晰具体、重点明确、尽量详尽,并且要避免歧义,主要包括背景、角色、任务、示例、格式、语气等要素。为使大语言模型更为有效地评审学术论文,本研究将 COAST、ROSES、BROKE、RTF 等提示框架加以融合,构建 BROTF 提示框架。其中,B(Background)为背景,R(Role)为角色,O(Objectives)为目标,T(Task)为任务,F(Format)为格式。

三、研究结果与讨论

(一)DeepSeek 赋能学术论文智能评审的优势

1. DeepSeek 有助于提升稿件初审的准确性和效率

为证明 DeepSeek 在论文初审环节的有效性,本研究在确保不泄露论文信息的前提下对某教育技术类期刊 2023 年的数据以“摘要+关键词”的方式进行检测,并将评审结果划分为 A、B、C 三个等级。纳入分析的全部稿件信息 2 200 余份,剔除数据信息不完整的样本,DeepSeek 共计生成 1 729 份完整样本,其中录用数据 127 条(含 A 级 106 条、B 级 12 条、C 级 9 条)。其中,DeepSeek 判定等级为 C 的稿件共 689 条,其中 9 条为实际录用的稿件,其他均为退稿。深入挖掘判断等级为 C 级却实际上又被“录用”的稿件发现,其中有 7 篇为跨学科领域的研究,未显示与“技术”相关的关键词,因而被 DeepSeek 视为不符合该领域的用稿标准。由此可见,DeepSeek 在剔除劣质稿件方面具有一定优势,可以为期刊编辑部减负增效,同时把优质稿件误判为退稿稿件的概率较低。

2. DeepSeek 评审论文的内在逻辑具有较高的可信度

基于前述设计并经检验的论文评审框架和提示词,研究对 DeepSeek 评审的 30 篇论文的总体打分情况进行量化统计与分析。结果发现,从每篇论文的得分总分看,优秀等级(≥85 分)的论文有 3 篇,良好等级(80-85 分)的论文有 22 篇,一般等级(<80 分)的论文有 5 篇且均大于 75 分,全部论文总分的平均分为 81.5 分,除得分最低(76 分)的一篇给出“修改后复审”的建议外,其他论文给出的总评建议均是“论文经修改完善后能达到期刊刊发的要求”。由此可以推断,DeepSeek 对论文质量的判断和稿件的最终录用情况具有高度的一致性,其评审结果具有较高的可信度。

具体来看,从论文评审框架的 5 个维度看,选题价值与研究内容维度(28 分),DeepSeek 给出的最低分是 22 分,最高分是 26 分,平均分 24.4 分,占该维度总分的 87.1%;研究深度与研究逻辑维度(23 分),DeepSeek 给出的最低分为 17 分,最高分为 20 分,平均分 18.5 分,占该维度总分的 80.4%;研究过程与研究方法维度(20 分),DeepSeek 给出的最低分为 12 分,最高分为 17 分,平均分 15.4 分,占该维度总分的 77.0%;研究结论维度(14 分),DeepSeek 给出的最低分为 10 分,最高分为 12 分,平均分 10.7 分,占该维度总分的 76.4%;学术伦理与规范维度(15 分),DeepSeek 给出的分数均为 12 分或 13 分,平均分 12.4 分,占该维度总分的 82.0%。由此可见,DeepSeek 在评审论文过程中对 5 个维

度的关注程度依次为“选题价值与研究内容>学术伦理与规范>研究深度与研究逻辑>研究过程与研究方法>研究结论”。这与编辑及专家审稿对论文的评审侧重点基本一致,即先看稿件的选题价值与研究内容,再看该文的研究深度及逻辑是否自洽,并重点判定研究过程是否合情合理,研究方法的应用是否适切,以及研究结论的价值和意义。不同的是,在学术伦理与规范方面,人类专家不如 DeepSeek 的重视程度高,这也从侧面表明 DeepSeek 在评审论文过程中的内在逻辑具有较高的可信度。

表 2　DeepSeek 评审 30 篇论文打分情况

论文评审框架	满分	评审最低分	评审最高分	评审平均分	分数占比/%
选题价值与研究内容	28	22	26	24.4	87.1
研究深度与研究逻辑	23	17	20	18.5	80.4
研究过程与研究方法	20	12	17	15.4	77.0
研究结论	14	10	12	10.7	76.4
学术伦理与规范	15	12	13	12.4	82.0

从对 30 篇论文的总体评价看,DeepSeek 既肯定了论文选题的前沿性和高价值性,方法的科学、严谨、规范或创新,又在理论深度、结论凝练或推广、实证过程(如样本代表性、数据多样性、实验完备性、技术细节等)等方面提出完善建议。这一方面说明 DeepSeek 对论文做出的总评意见是依据其对各评审指标的打分综合得出的,具有内在的逻辑一致性;另一方面也说明 DeepSeek 对论文的总评意见与专家审稿给出的总评意见基本一致,即在首肯论文选题价值与研究内容的前提下,通过学术伦理与规范判断作者的治学态度和学术功底,多从研究深度与研究逻辑、研究过程与研究方法、研究结论等方面对论文提出改进或完善建议。

3. DeepSeek 能够为论文的修改和完善提供一定的借鉴意见

将 DeepSeek 生成的 30 份审稿意见与人类审稿专家的意见从选题价值与研究内容、研究深度与研究逻辑、研究过程与研究方法、研究结论、学术伦理与规范 5 个维度进行对比分析后发现:(1)在选题价值与研究内容维度,人机评审结果一致程度最高,再次验证了选题价值与内容判读在论文评审中的重要性。(2)在学术伦理与规范维度,评审结果的一致程度较低,可能的原因是人类专家在评审过程中鲜有涉及学术伦理与规范方面的考量。(3)在研究深度与研究逻辑、研究过程与研究方法、研究结论等维度,尽管这些均可以归为研究设计范畴,也是人和机器都关注的范畴,但二者却表现出一定的差异:人类专家更像是共同体视野下的伙伴,会根据自己的经验,为稿件的修改提供有边界、可操作的建议,有时哪怕一句点拨的话就能让文章质量再上一个台阶,但也因此常常忽略一些在他们看来并不是很重要的方面(如格式、规范等);大模型是“照章办事”,逐条对标并回答问题,但有时从文章本身的逻辑看,其所提供的修改建议的参考性并不是很强。例如,同样是对“样本量不足的批判”,人类专家可能提出“数据收集需补充说明”,而 DeepSeek 则会强调“N = 12 代表性不足,需要分层抽样控制变量”。可见,DeepSeek 不仅可以在初审环节协助编辑高效地完成对劣质稿件的筛选,协助人类在规范层面减负增效,而且 DeepSeek 生成的审稿建议也可以为作者稿件的修改与完善提供新的视角。

(二)DeepSeek 赋能学术论文智能评审的短板

1. DeepSeek 对学术论文评审“一板一眼”

DeepSeek 提供的审稿意见较为“规矩”:在形式上,每份审稿建议都按照“论文内容概括—论文质量评价—修改建议”的架构呈现;在论文质量评价方面,DeepSeek 提供的审稿意见涵盖选题价值与研究内容、研究深度与逻辑、研究过程与方法、研究结论、学术规范等方面的问题,而且会将相关意见及建议归纳起来,按照“优势—问题”的格式输出,全面详尽,便于作者和编辑参考。相比较而言,人类

专家的评审意见整体上并不全面,更倾向按照“重要性—紧迫性”的方式呈现,且容易受个人学术专长及审稿倾向性的影响。以样本 2 为例,外审专家 1 仅提到统计方法存在缺陷、文献综述缺少国外研究、技术应用和研究流程信息不足等问题,并未指出选题质量、研究过程、研究结论等方面存在的问题。从评审内容看,一方面,DeepSeek 对论文内容概括得较好,展现其具有更好的文本分析能力,能够基于自身的算法逻辑从论文中快速抽取关键信息,并进行总结与概括;另一方面,DeepSeek 对论文质量的评价虽然较为客观,但显得过于普适。如果单纯看一篇文章的审稿建议,可能会觉得 DeepSeek 是一个知识渊博的“大先生”,但当系统阅读了 30 份评审建议后,就会觉得不少内容具有相似性和重复性。例如,在多篇论文评审意见中,DeepSeek 在评审选题价值与研究内容时,均是按照“紧扣教育部相关政策需求,直击实践痛点”这样的套路来说明“研究的现实指向性”,而且几乎每篇文章都会指出文献综述批判性不足的问题。此外,DeepSeek 缺乏对论文本身思想性的评价,所提的修改意见与建议虽然面面俱到,但缺乏对文章类型定位和实际情境的考量。例如,有些论文本身属于理论性研究,但硬要套用固定的评审框架,就显得评审建议虽全面却浮于表面,不切实际。最后,DeepSeek 生成的评审建议缺乏批判性与对话性。DeepSeek 的“生成”不同于人类的“创造”,这在一定程度上削弱了论文独特的语言风格,让作者和评审专家难以实现生成性的对话。

2. DeepSeek 对学术论文评审过程中的创新性分析深度不足

DeepSeek 在评审论文过程中确实能够提供一些准确的信息,尤其是在语法检查、术语规范和文献引用等方面。例如,DeepSeek 针对论文提出一些诸如“统一公式与表格排版(如修正表 2 中的公式重复问题),核对参考文献编号与正文引用的一致性”“将‘数智化’明确定义为数据智能(Data Intelligence)与教学智慧的深度融合过程”“统一术语拼写(如 ChatGPT)”“部分量表信效度报告不完整(如 SAM 量表的 Cronbach's α 未分维度报告)”“引用格式错误,如文献未按顺序标注;外文文献(如 Mark Granovetter, 1985)未补充中文译名”等。这些细节性差错有助于作者和编辑更好地完善和规范论文。然而,DeepSeek 在对论文的创新性和研究深度进行评价时,往往比较含混且缺乏深度。这可能是因为 DeepSeek 主要依赖已有的数据和知识库,一些专业文献库和最新研究文献无法获得,且其缺乏专家对问题的直觉认知和经验积淀。尽管 DeepSeek 的性能已经超越基准水平,但在理解复杂概念和执行实际任务方面,其准确性还有待提高。对学术研究的评价是一项具有高度创造性的工作,而不是是非问答。理论上,当有明确的评审框架时,DeepSeek 有可能通过评估学术论文的质量来取代人工同行评审者,但也可能产生误导性的错误答案或不完整的答案,所以对 DeepSeek 生成的评审意见的准确性进行严格测试是必要的。

3. DeepSeek 提供反馈意见的建设性和适切性不足

DeepSeek 虽能指出学术论文中存在的格式错误、数据缺失等显性缺陷或基础性问题,并针对该部分问题提出具有建设性的修改意见,但这些建议往往浮于表面,缺乏对论文深层次结构和内容的洞察,导致反馈意见的建设性和适切性,与人类审稿专家相比,仍存在一定的差距。因此,在提供建设性论文反馈意见方面,DeepSeek 尚不能完全满足学术论文评审的需求。实践中,同行评议专家在审稿过程中可以互换角色,先立足作者的角度进行理解,再转换为专家角度进行审视,与作者进行超时空对话。例如,同行评议专家根据“学习动机水平高低”提出层层递进的追问:(1)为何学习动机平均值低反而学习动机水平高?如果是因为问卷得分越高,动机水平越低,则需要在研究工具部分说明;(2)课程实施周期也会影响学习动机水平,因此除了需要说明一节课 45 分钟的信息外,还要说明课程持续了几周,这些信息有助于我们评估动机的差异是否是短期的新鲜感或是真的有影响;(3)动机量表是何时施测的?施测了几次?这些信息也需要在研究流程中体现。从这些追问中,作者可以感受专家评审的专业性及其细致认真的态度。但是,由于审稿专家个人学术专长和风格的不同,可能仅

会关注他们认为更关键的内容,并不会按照“指出问题—分析不足—提出建议”的逻辑闭环提供较为完善的审稿意见。从编辑审稿的实践经验看,有时候审稿专家提出的意见具备一定的“启发性”,可以为研究的进一步深入提供重要参考,而 DeepSeek 针对一篇理论性文章所提的诸如“在‘知识传播策略’部分补充具体案例(如Ⅵ型组的教案设计流程),并设计实施路线图(如教师干预的阶段性步骤)”的建议显然已经超越该研究的边界和重心,即 DeepSeek 提出的评审意见虽较为全面但也较为泛化,有时并不适切。

四、研究建议

(一)恪守学术伦理,达成行业共识

高校在科学研究范式转型中应紧密依托新一轮科技变革,植根中国科学研究实践土壤,立足已有范式寻求创变,探索一条符合我国国情且具有鲜明中国特色的高校科学研究道路。具体来讲,应加强跨学科合作与协同创新、规范数据使用标准、推动可解释性 AI 的发展,建立健全科研伦理审查机制[9]。有研究指出,在论文评审过程中,审稿人应遵循人工智能使用披露规范,声明人工智能使用情况[10]。为保证大语言模型在学术论文评审中的有效性和可靠性,建议相关部门联合期刊界制定相应的指导方针和评审标准,包括大模型的使用范围、评审流程、结果解释、责任归属等方面,以确保评审过程的公正透明。尽管学术研究质量的定义多种多样,但原创性、方法论的严谨性以及对社会的影响力评估,通常被认为是学术论文评价的 3 个核心要素。因此,围绕创新性、方法论和影响力构建更科学、详尽的评审标准,是引导大语言模型进行学术论文评审的更为基础性的研究工作。我们认为,应在确保遵循学术伦理的前提下,对学术论文进行分类评审,即理论思辨类、实证研究类、混合研究类、技术开发类等不同范式的研究论文应该在基本的评审框架和指标体系中有所侧重,不能一概而论。同时,鉴于不同学科领域专业期刊的特色和偏好差异,通用的大语言模型及学术论文评审框架可能并不适用于所有期刊的论文评审。不同期刊应基于自身特点部署本地化大语言模型,并结合学术论文的评审框架,最大化发挥大语言模型的智能评审效能。

(二)加强数据安全,强化高质量数据集建设,构建垂直领域大模型

没有高质量的数据集,大模型就是“无本之木”。如何保证数据的高质量和安全,成为构建大模型的重要维度。国家高度重视高质量数据集的建设,相继出台数据要素领域政策文件。例如,2024 年 1 月,国家数据局等十七部门联合印发《“数据要素×”三年行动计划(2024—2026 年)》指出,努力建设好高质量数据集,以发挥数据要素价值乘数效应[11];也有媒体提到高质量数据是大模型竞争的下一代[12]。刘烈宏提出,高质量数据与人工智能的结合,将进一步发挥数据和人工智能的倍增效应[13]。学术研究是一项创造性工作,如果将还未曾发表的研究直接输入通用大模型,一是会产生重要的安全隐患,二是在论文评审的专业性层面略显逊色。为确保数据安全,建议学术出版机构合力构建垂直领域大模型,在把握数据安全的同时确保训练数据的多样性,因为大语言模型在学术评审中能否提供精准性和建设性的反馈意见,取决于训练数据的多样性和算法的优劣。当前 DeepSeek 的深度思考主要基于可获取的互联网资源和深度学习算法,而诸多专业的数据库(如中国知网、Web of Science 等)并未纳入其训练数据范畴。因此,可以通过增加训练数据的多样性和专业性来构建专门用于学术论文评审的大模型,通过优化算法提升其计算性能和数据质量,进而提高大语言模型对特定领域学术论文评审的广度和深度。

(三)开展多元化人机协同评审是学术评审的新方向

人机协同已成为教育、出版等领域共同关注的话题,目前已有研究在人机协同的理念层面开展了诸多探索,如郭蕾蕾提出需打造“师—机—生”协同共进的教育新生态,但现阶段的相关研究大多停

留在框架构建与模式设想层面,缺乏具有可操作性的人机协同实践方式与系统性的证据验证[14]。当前使用大语言模型进行学术评审大多聚焦伦理问题,鲜有研究将其应用于审稿流程并进行效果评估[15]。这种重理念、轻实践的发展样态导致“人机如何协同”“在什么阶段协同”等问题凸显,制约了大语言模型在学术评审领域的应用深度。在审稿方面,人类专家具有批判性思维,大模型是数据驱动型思维,如果将审稿全盘交给大模型,容易出现方向偏差,继续沿用传统的“期刊编辑—审稿专家”审稿模式,则存在审稿周期较长、研究成果“失效”等问题。鉴于 DeepSeek 等大语言模型在某些评审任务上的优势,建议探索人机协同的新型评审模式。例如,可以让 DeepSeek 等大语言模型负责稿件的初步审查,如格式和语法检查,让外审专家专注对论文深度的专业评价、建设性修改意见反馈等方面。这种模式既可以提高评审效率,又能保证评审质量。另外,编辑同样可以借鉴 DeepSeek 等大语言模型提供全面性分析与意见反馈的特点,挖掘有价值且可供参考的评审意见,以供作者完善论文。由此,充分利用大语言模型、专家和编辑的各自优势,重构多元化人机协同评审模式,从而为“学术论文评审是为了更好地改进和提升论文质量”的终极目标保驾护航。

五、结　　语

以 DeepSeek 为代表的大语言模型已经对学术研究全流程产生了重要影响,甚至改变了知识的生产方式,探寻大语言模型应用的伦理和边界不仅必要而且急迫。期刊论文评审作为一项复杂的工作,通过“边界”的探寻可以引导期刊从业者及同行评议专家更加有效地使用大语言模型。事实上,论文的评审和发表并不完全受所谓的“标准”左右,加之当下学术界“稿件生产和稿件发表”的供需严重失衡,导致并不是所有“品质较佳”的稿件都有发表的机会。本研究仅以某教育技术类期刊为例,探索了 DeepSeek 赋能学术论文评审的有效性,由于学术研究的类别多样,后续本研究团队将会依据学术研究类别——理论类研究、实证类研究、技术类研究等,探索 DeepSeek 在哪些环节以及如何协同人类专家审稿,从而实现审稿质量和效率的双面提升。当然,任何一项研究探索的落脚点必然要面向实践,后续本研究团队也将尝试部署本地大模型,让“人机协同”评审真正落地。

参考文献:

[1] 王丽丽,王银宏,杨永强,等. 国内外英文科技期刊同行评议的方法与质量控制研究[J]. 编辑学报,2024,36(S2):37-43.

[2] WU T, HE S, LIU J, et al. A brief overview of ChatGPT: the history, status quo and potential future development [J]. IEEE/CAA journal of automatica sinica, 2023,10(5):1122-1136.

[3] ZHOU H, HOU Y, LI Z, et al. How well do large language models understand syntax? an evaluation by asking natural language questions[J]. arXiv preprint arXiv:2311. 08287,2023:1-20.

[4] ZHU M, WENG Y, YANG L, et al. DeepReview: improving LLM-based paper review with human-like deep thinking process[J]. arXiv preprint arXiv:2503. 08569,2025:1-25.

[5] MEMON A R. Similarity and plagiarism in scholarly journal submissions: bringing clarity to the concept for authors, reviewers and editors[J]. Journal of Korean medical science, 2020,35(27):e217.

[6] ZHAO X, ZHANG Y. Reviewer assignment algorithms for peer review automation: a survey[J]. Information processing & management, 2022,59(5):103028.

[7] 喻国明,林昱彤,李昀玥. 作为新型内容生产力的生成式 AI:发展局限与未来进路[J]. 出版广角,2024(14):22-30.

[8] 刘俏亮,刘东亮,张洁. 国产 AIGC 大模型辅助稿件初审的研究:以信息科学类论文为例[J]. 编辑学报,2024,36(5):548-552.

[9] 赵晓伟,王小雨,王艺蓉,等. 人工智能赋能高校科学研究范式创新:价值、风险与进路[J]. 重庆高教研究,2025,13(1):9-20.

[10] 解贺嘉,张家烁,张久珍.人工智能环境下学术期刊出版伦理规范内容建设研究[J].现代出版,2024(12):55-66.
[11] 十七部门关于印发《"数据要素×"三年行动计划(2024—2026 年)》的通知[EB/OL].(2024-01-05)[2025-05-26].https://www.cac.gov.cn/2024-01/05/c_1706119078060945.htm.
[12] 魏蔚.高质量数据:大模型竞争的下一站[N].北京商报,2024-12-03(03).
[13] 李宏伟.以高质量数据促进人工智能发展[N].中国改革报,2025-03-26(01).
[14] 郭蕾蕾.生成式人工智能驱动教育变革:机制、风险及应对——以 DeepSeek 为例[J].重庆高教研究,2025,13(3):38-47.
[15] SCHINTLER L A, MCNEELY C L, WITTE J. A critical examination of the ethics of AI-mediated peer review[J]. arXiv preprint arXiv:2309.12356,2023:1-21.

(责任编辑:张海生　杨慷慨　校对:杨慷慨)

Towards Human-Machine Collaboration: Efficacy and Boundaries of Large Language Models in Empowering Academic Paper Review

JIAO Lizhen [1,2], LIU Xuan[3], LIU Junjie[4], ZHONG Hongrui[5]
(1. *Information Technology Center*, *Tsinghua University*, *Beijing* 100084, *China*;
2. *Engineering Research Center of Integration and Application of Digital Learning Technology*, *Ministry of Education*, *Beijing* 100081, *China*;
3. *The Open University of Sichuan*, *Chengdu* 610073, *China*; 4. *School of Education*, *Minzu University of China*, *Beijing* 100081, *China*;
5. *Guangzhou Galaxy Technology Co.*, *Ltd.*, *Guangzhou* 511356, *China*)

Abstract: Large Language Models (LLM) have shown great potential in the field of academic paper review, but their effectiveness and boundaries still need to be further explored. Focusing on the DeepSeek LLM, taking an educational journal as an example, an exploration was made on the ability of the LLM to assist in the review of academic papers. The results show that DeepSeek can help improve the efficiency and accuracy of the initial review of manuscripts, have a certain degree of credibility in the internal logic of the review papers, and provide some reference for the revision and improvement of the papers. Compared with human experts, DeepSeek also has obvious shortcomings, such as rigidly rule-based for academic paper review, insufficient depth of innovative analysis, and insufficient constructiveness and appropriateness of feedback provided. In order to make better use of LLM to empower academic paper review, and effectively contribute to review practice, discipline suggestions, and academic prosperity, the academic community should abide by academic ethics and reach an industry consensus; strengthen data security, strengthen high-quality datasets, and build large models in vertical domains; carry out multi-human-machine collaborative review to build the boundary of machine-empowered intelligent review.

Key words: large language models; DeepSeek; academic paper review; intelligent review; human-machine collaboration; evaluation criteria