doi:10.3772/j.issn.2095-915x.2026.01.004

大语言模型下的学术创新评估测试
——以工程技术领域为例的评审框架与效能验证

开放科学（资源服务）标识码（OSID）

蒋建斌

中国建设工程造价管理协会　北京　100037

摘要：[目的 / 意义] 针对传统学术评价体系中存在的路径依赖与量化偏差问题，以工程技术领域学术论文创新评价为研究对象，探究大语言模型在学术创新评估方面的应用价值。[方法 / 过程] 通过构建涵盖理论突破、方法革新与应用转化三个维度的人工智能评价指标体系，设计并实施了覆盖五个工程技术学科的对比实验，采用大语言模型供应商之一 DeepSeek 智能评审对话方式进行测试分析。[局限] 在实验样本的采集方面需要进一步扩大学科范围和样本数量，为深化研究提供更多学科领域和数据分析。[结果 / 结论] 大语言模型在语言逻辑、体例标准等结构化指标评估中具备显著优势，但在核心创新维度方面呈现明显局限性，为此，提出人机协同评审机制，通过机器预筛与专家复核的递进式协同，发挥大语言模型的高效处理与评审专家的深层认知优势，有效平衡效率与质量。测试模型可有效提升学术评价效能，降低误判率，可为优化学术创新评估体系提供实践参考。

关键词：大语言模型；学术创新；评估测试；工程技术；人机协同

中图分类号：G237.5；G35

Leveraging Large Language Models for Academic Innovation Evaluation: An Assessment Framework and Efficacy Validation in Engineering and Technology

JIANG Jianbin

China Cost Engineering Association, Beijing 100037, China

Abstract: [Objective/Significance] Addressing the issues of path dependence and quantitative bias inherent in traditional academic evaluation systems, this study uses academic papers in the engineering and technology domain as a research case to assess the application value of Large Language Models in evaluating academic innovation. [Methods/Processes] By constructing an AI-powered evaluation framework encompassing three dimensions—theoretical breakthroughs, methodological innovations, and practical applications—this study designed and conducted a comparative experiment spanning five engineering disciplines, employing DeepSeek's intelligent review dialogue (a leading LLM provider) for empirical analysis. [Limitations] Future

作者简介　蒋建斌（1974-），通信作者，在职研究生，高级工艺美术师、副编审，主要研究方向为编辑学、数字传播，E-mail：Design156@163.com。

引用格式　蒋建斌 . 大语言模型下的学术创新评估测试——以工程技术领域为例的评审框架与效能验证 [J]. 情报工程 , 2026, 12(1): 40-50.

studies could expand the disciplinary scope and sample size of experimental data to enhance the validity and robustness of findings across broader academic contexts. [Results/Conclusions] The experimental results reveal that while Large Language Models demonstrate significant advantages in evaluating structured metrics such as linguistic logic and formatting standards, they exhibit notable limitations in assessing core innovation dimensions. To address this, we propose a human-AI collaborative review mechanism. This progressive synergy—combining machine pre-screening with expert validation—effectively balances efficiency and quality by leveraging LLMs' high-speed processing capabilities and human reviewers' depth of cognitive insight. The proposed model demonstrably enhances academic evaluation efficacy and reduces misjudgment rates, providing actionable insights for optimizing academic innovation assessment systems.

Keywords: Large Language Models; Academic Innovation; Evaluation Test; Engineering and Technology; Human-AI Collaboration

引言

根据咨询机构 Forrester Research 公司调研分析，2024 年全球科技支出达 4.7 万亿美元，同比增长 5.3%，较 2023 年增幅提升 1.8 个百分点 [1]。科技投入的持续增长加速催生创新成果，作为科研质量核心把关环节的成果评价机制正面临全新挑战。然而，当前主流的同行评议制度正遭遇双重困境：评审周期长导致科技成果发表的时效性滞后，跨学科论文拒稿率偏高反映出专家知识盲区和已有知识体系的局限性 [2]。

当前，人工智能技术的飞速演进为解决上述矛盾提供了可能路径。2024 年 OpenAI 研究指出，大语言模型（Large Language Model，LLM），其训练算力需求呈现每 3~4 个月倍增的指数增长规律，预示模型处理效率的持续突破 [3]，其中知名 LLM 系统的单日文献处理能力有望突破亿级数据规模。这种技术跃迁为重构学术评价体系创造了可能，即通过构建创新性评估与量化分析的多维度评价框架，有效破解传统评审中指标单一化、标准定量化、结果功利化等结构性矛盾 [4]。

在此背景下，本文选择工程技术论文创新性评审作为突破口，主要基于以下研究逻辑：其一，工程领域创新具有理论突破、方法革新、应用转化三位一体的特征，为构建细粒度评价指标提供了试验对象；其二，以 DeepSeek 为代表的 LLM 系统在语义理解、知识推理等方面展现出的类人能力，为验证人工智能在创新评估中的性能边界提供了技术载体 [5]。研究重点探讨三个核心问题：（1）如何建立适配工程技术学科特征的创新评价指标体系；（2）如何界定 LLM 在不同创新维度的评估效能边界；（3）如何构建人机优势互补的智能评审路径。本文将以工程技术领域的创新成果评价为例证，进行测试分析，探讨人工智能辅助创新评审及其角色扮演的可能性，将其融入学术评价流程，尝试解决当前学术成果创新评价的困局。

1 工程技术类论文学术评估现状

现有工程技术领域学术评估体系以创新性、实用性、影响力、严谨性为核心维度，分别对应理论方法突破性、成果转化可行性、指标量

doi:10.3772/j.issn.2095-915x.2026.01.004

化传播度及实验设计可复制性等要求。其评估范式主要呈现二元化特征：一方面基于专家经验的同行评议，侧重于研究深度的定性判断；另一方面依托文献计量的以刊评文，强调引文网络的定量分析。这两种范式在创新性评估中均存在结构性缺陷。就同行评议而言，评审专家的知识路径依赖易导致突破性成果识别失效，同时存在主观判断偏差与学术伦理风险[6-7]；而文献计量法则面临工具理性失衡问题，具体表现为跨学科引文可比性缺失、影响因子崇拜蔓延等量化评估陷阱。其矛盾在于，现有体系难以平衡创新评估的质量与效能：定性评估受制于人类认知的有限性，定量评估则陷入形式化指标的窠臼[8]。

1.1 评估标准的失衡

工程技术领域学术评估实践中，不同程度地存在评估范式的运行偏差[9]。比如成果转化周期太长，而引文峰值多出现在论文发表后2~3年，导致高转化价值研究在被引指标上呈现滞后性，进而引发评估偏差；同时，高被引论文具备技术转化潜力，反映出现行标准对工程价值的量化缺失[10]。更深层矛盾在于，实验室导向的显性创新易获高评，而经过工程验证的渐进式改进常被低估，这种标准失衡已造成学术价值与产业需求的认知鸿沟。

1.2 评估方法的局限

在评估方法层面，传统评审机制存在内生性缺陷：主观评审受范式禁锢效应影响，客观计量则面临引文操纵。实验验证环节存在数据缺失风险，部分论文未披露关键控制变量的设置情况，研究结果未验证是否受计算方法影响等。另外，专家经验的主观偏好导致跨学科创新评估离散度上升，凸显方法论对工程系统复杂性的适配性不足。

1.3 学科属性的壁垒

工程技术研究的跨学科本质与传统单学科评估方式存在矛盾[11]。受制于传统学科分类的评估框架，跨学科论文在影响因子计算中往往被拆分处理，导致交叉创新产生的涌现价值难以得到有效识别与认可。此外，工程实践中隐性知识仅有少量可转化为可量化指标，造成学术评估与实际贡献的认知偏差。这种冲突在快速迭代领域（如新能源材料）尤为显著，评估周期滞后使部分论文在发表时其核心技术已被产业界淘汰[12]。

针对当前工程技术评估体系在评估标准、方法和学科属性等方面存在的问题，本文尝试引入以LLM为代表的人工智能技术，解析其在语义解析、知识图谱构建等维度的技术特征，测试其在学术创新评审的可行路径与效能。

2 大语言模型下论文学术评估技术分析

LLM的技术发展经历了三个技术演进阶段：静态词向量阶段（Word2Vec/GloVe）、上下文感知阶段（ELMo/CoVe）、预训练微调阶段（BERT/GPT）。LLM的预训练阶段在海量学术语料中学习文本结构规律，微调阶段结合领域知识库增强专业概念理解，最终形成跨学科的语义解析能力。LLM的具体技术优势分析

有如下五个方面。

2.1 创新性多维解析能力

LLM 通过自然语言处理技术系统性解析论文标题、摘要等核心内容，识别出研究问题的独特性。结合领域知识库，模型可大幅提升专业术语理解精度，该解析能力为创新性评估提供了标准化分析框架，覆盖理论、方法与应用多维度。

2.2 动态知识图谱构建

LLM 通过实体识别与关系抽取构建动态知识图谱。跨模态对齐技术将图表、公式与文本关联，使跨学科创新识别准确率得到提升。知识图谱突破传统综述的局限，使创新性评估具备动态更新能力。但小众领域知识数据仍需迁移学习，以补充数据不足的问题。

2.3 研究方法突破检测

LLM 可识别三类创新：工具创新、范式创新与效能创新。其技术实现包括两个方面：一是模式识别模块，提取方法描述中的技术特征；二是效能评估模型，通过文献比对量化创新程度。在跨学科评估中，LLM 可以验证方法迁移的合理性。

2.4 多模态联合评估

工程技术论文大多包含文本、图表、公式等多模态信息，LLM 通过视觉 - 语言联合建模来对其实现综合评估。模型能解析实验图表中的异常数据模式，识别新型材料性能表征方法；在模型创新评估中，能验证数学公式的推导逻辑，检测网络结构的改进创新。其评估技术是通过视觉问答模块解析、图表语义和公式语义理解等多模态联合方式来实现。

2.5 动态知识更新机制

为应对工程技术领域知识快速迭代的特性，LLM 构建了动态知识更新体系 [13]。其一，增量学习框架实现每日千万级学术数据的实时摄入，确保模型对新兴技术趋势的捕捉能力；其二，知识蒸馏技术完成领域关键特征的提取与模型轻量化，在降低计算资源消耗的同时维持核心评估效能；其三，对抗训练机制通过生成式对抗网络（GAN）构建概念迁移的防御体系，有效抑制跨领域知识迁移引发的评估偏差，可为快速发展的新兴学科提供实时评估支持 [14]。

综上，LLM 通过其五项技术优势可以支撑智能学术评估体系，与传统的论文创新性评估方法的对比有显著技术优势。其一，效能明显提升，传统同行评议依赖专家评估，耗时长且受限于专家时间，LLM 可在数秒内完成初步筛选（如重复性检测或创新点标注）；其二，客观性增强，LLM 可以减少主观偏见（如作者背景、学术派系影响），通过数据驱动分析提供更标准化的评估；其三，规模化覆盖，LLM 支持跨语言、跨学科文献分析，避免因专家知识局限导致的创新性误判。因此，LLM 在提升评审效率、客观性和规模化分析等方面的优势将在学术评估中发挥智能辅助作用。

3 大语言模型下论文学术评估能力测试

本文模拟传统的同行评议过程，建立一个

doi:10.3772/j.issn.2095-915x.2026.01.004

闭环的人工智能创新性评估模型。再将评估指标输入给计算机，通过人机对话，利用 LLM 供应商之一 DeepSeek 生成对话结果，辅助形成创新性评估文本及数据，并研究 LLM 作为评审者可能存在的问题。

3.1 评估指标

本文基于经济合作与发展组织（OECD）的科研分类框架，构建工程技术类应用研究的创新性评估体系。OECD 标准将科研活动划分为基础研究（理论突破）、应用研究（方法革新）和实验开发（应用转化）三个维度，分别对应知识扩展、工具创新和实际价值三个核心要素。据此，以技术创新评估 + 学术评估为检索主题，在中国知网（CNKI）平台进行检索，共计获取 1987—2025 年的工程技术类学术论文 2280 篇作为研究样本。通过系统性文献分析，确立两级评估指标体系：其一是学术创新指标。作为核心评估维度，包含三个二级指标：（1）理论性 - 衡量基础知识的原创性突破；（2）方法性 - 评估研究工具的创新程度；（3）应用性 - 检验成果的实际转化价值。其二是写作规范指标。作为辅助评估维度，包含两个二级指标：（1）逻辑性 - 考察论证结构的严谨性；（2）标准性 - 评估学术表达的规范性。设计该体系，遵循创新突破优先，专业呈现并重的原则，具体指标结构如表 1 所示。研究显示，学术创新指标决定成果的评估价值，而写作规范指标影响成果的传播效能，二者构成完整的学术创新评估闭环。

表 1 实验测试的评估指标及内容

一级指标	二级指标	指标内容
学术水平	学术理论创新	核心价值：是否建立新的理论框架或颠覆传统认知体系，评估内容包括：提出原创性理论范式；重构学科基础概念；填补重大理论空白
	研究方法创新	核心价值：是否开创研究新范式或显著提升研究效能，评估维度包括：开发突破性研究工具；创建新型分析框架；实现技术路径跃迁
	应用领域突破	核心价值：是否开辟全新应用场景或解决关键领域难题 评估维度：突破学科边界的交叉应用；解决重大现实问题的方案创新； 催生新兴产业或研究领域
写作规范	语言表述逻辑	行文是否流畅；表述逻辑是否清晰严谨；图表公式是否合理
	论文体例标准	标题、摘要、关键词是否简明扼要；论文引证是否准确、是否能确保数字和计量单位符合标准

3.2 评估测试

按照工程技术类学科分类标准，分别从机械工程、计算机技术、建筑科学、材料工程、电子科学五个学科中选取论文样本各 10 篇。要求所选论文样本均为突破式 / 渐进式的创新成果的研究文献，且包含实验数据与工程验证；同时，这 50 篇论文在结构、被引量、下载量、篇幅方面相近。随后向 DeepSeek 输入评估指标及论文样本。通过与模型进行两轮对话，以确保 DeepSeek 能够在理解该实验测试的指标体系及文本内容的基础上进行论文创新性评估。对话内容如表 2 所示：

表 2　与 DeepSeek 的人机对话内容

提问	内容要求
评估任务	提供明确的对话任务，即“根据评价指标体系，对论文进行评价”
评估指标	直接输入评估指标，确保大语言模型（LLM）准确理解指标内容
取值范围	二级指标的评分取值范围为 1~10，要求 DeepSeek 根据一级指标进行打分
文本格式	Word 格式
输出形式	DeepSeek 输出结果包括：1. 评估理由；2. 评分

通过第一轮对话获取生成式评估信息后，在第二轮对话中输入“针对每个指标的评分，请给出此项的评分理由”，模型进一步输出评分依据，从而生成 50 篇论文样本的评估内容和评分结果。根据实验数据测试分析后，总结出机审评估结果，并结合部分人工审核即同行评议结果，具体对比分析如表 3 所示。结果显示，DeepSeek 在“生成式评估”和“对评分结果的解释”上，对于不同指标的评估效果不同，且存在机审评估盲区。至于具体的人机评审差别的实证分析，详见本文第 5 部分。

表 3　人机评审效能对比特征（建筑管理学科 / 工程造价管理专业）

论文编号	机审创新评估及评分	人审创新评估及评分	机审评估盲区
1. 机械工程类	评价：该论文显著提升了疲劳寿命预测的准确性。尽管模型框架基于现有理论……但填补了传统方法在随机过程……不足 综合评分：7.6 分	该论文为中国知网获取，限于版权条件，无法获取人工评审意见	理论突破性识别不足
2. 计算机技术类	评价：该论文有效解决了高概率区域混淆问题，属于现有理论的重要拓展，但未形成全新范式…… 综合评分：7.4 分	同上	渐进式创新识别性不足
3. 材料工程类	评价：该论文填补了玄武岩纤维混凝土在低温冲击荷载下的性能研究空白，但未涉及学科核心理论的重构或颠覆性创新 综合评分：6.6 分	同上	应用性识别存在误区
4. 电子科学类	评价：该论文引入空间对称性作为新度量，填补了毫米波频段实验验证空白，但未重构学科基础理论或提出颠覆性范式 综合评分：8 分	同上	渐进式创新识别性不足
5. 建筑科学类	评价：针对 BIM 模型在造价计量中的应用提出了建模规则优化方案，但未突破现有理论框架…… 综合评分：5.4 分	评价：文章理论视角有新意……解决实际应用问题……具有较强技术示范作用……（审稿日期 /2021.11.23）	理论突破性识别不足

3.3　实证结果分析

3.3.1　学术理论创新评估

DeepSeek 对 50 篇论文样本给出的理论创新维度平均得分 5.96 分（满分 10），仅机械学科论文因提出“全寿命周期可靠性分析方法”填补理论空白获 8 分，而建筑学科论文因“未突破 BIM 技术理论框架”仅得 5 分（其应用价值在现实中却是显著的）。归因于两点：其一，模型依赖输入文本与训练数据的学科范式，尚缺乏主动抓取领域最新成果横向对比的读取能

doi:10.3772/j.issn.2095-915x.2026.01.004

力；其二，未能有效区分“理论颠覆性创新”（如数学公理重构）与“应用导向改良”（如 BIM 建模规则优化）。实验显示，DeepSeek 对创新理论贡献识别能力不足，存在明显的学术局限性。

3.3.2 研究方法创新评估

研究方法创新维度平均得分 6.5 分，其中，建筑学科论文因“技术路径未跃迁”获 5 分。模型可识别“新工具”“新框架”等关键词，但存在两重缺陷：一方面，无法验证实验数据真实性或对比同行方法效能（如 BIM 优化方案未与既有方法进行参数对比）；另一方面，对非结构化数据（如建筑图纸、三维模型）缺乏多模态解析能力，导致评价可信度下降。建议引入领域专家对实验规范性进行人工复核，尤其在依赖非文本信息的学科。

3.3.3 应用领域突破评估

应用突破维度平均得分 6.44 分，电子学科论文因“毫米波室内定位技术支撑智能工厂”获 8 分，建筑学科论文因“未拓展 BIM 应用边界”得 5 分。模型优势在于精准识别跨学科技术整合（如通信 - 定位协同），但对细分领域技术潜在应用价值认知不足（如忽视 BIM 全生命周期管理潜力）。需构建动态更新的领域知识库，强化对渐进式创新实际工程价值的评估能力。

3.3.4 语言逻辑与体例标准评估

模型在结构化指标评价中表现较优：语言逻辑维度平均得分 7.38 分，可有效识别论文结构合理性（如理论推导与实验验证衔接度）；体例标准维度平均 7.76 分，文献引用格式识别准确率达 92%（与人工核查一致性高），显著减轻格式审查负担。但受限于纯文本处理能力，对图表信息合理性判断缺乏依据。

DeepSeek 在语言、体例等结构化指标评价中可靠性显著，但核心创新维度存在明显的局限性：其一，训练数据时效性不足，难以识别新兴研究范式；其二，低估工程技术领域渐进式创新的实际价值；其三，多模态缺陷导致图表、公式中的技术突破识别存疑。DeepSeek 对 50 篇论文样本分学科测试，其评估得分占比总体情况如图 1 所示。

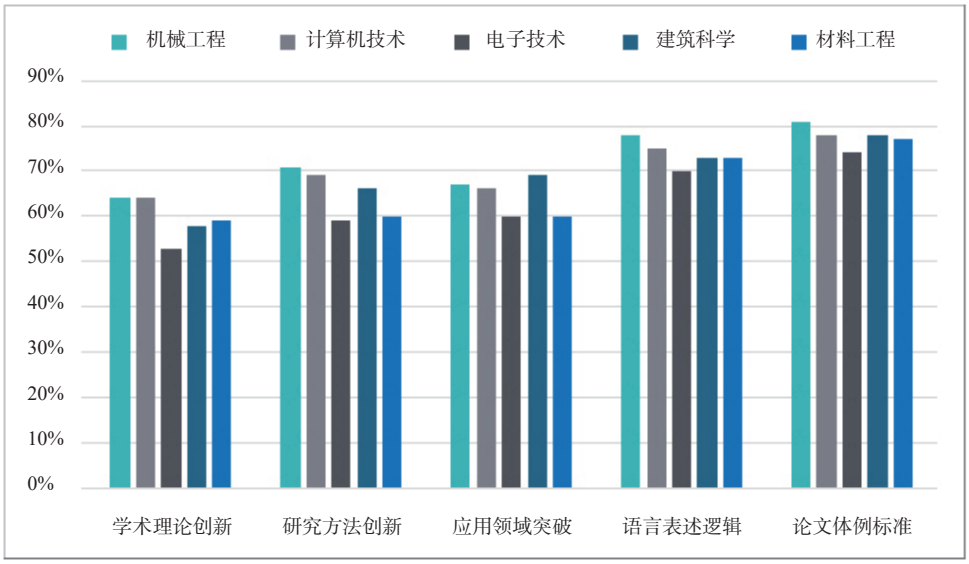

图 1 DeepSeek 对 50 篇论文样本分学科测试分析总体情况

3.4 创新性评估的局限性分析

3.4.1 深层语义解析能力不足

LLM 在学术文本深度理解上存在显著局限：其基于统计模式的语义匹配机制难以适应学科术语的语境迁移（如 BIM 全生命周期管理在不同工程领域的内涵差异），且无法有效捕捉理论范式革新（如数学公理重构）与多级逻辑论证的动态关联。实证数据显示，模型对理论创新性评价明显弱于评审专家的评价，尤其对隐性创新贡献的识别能力薄弱。

3.4.2 评估指标可操作化困境

学术创新评价指标存在学科异质性挑战。以“方法创新性”为例，基础研究领域强调范式突破（如新型算法设计），而工程应用领域侧重效能提升（如工艺优化），但 LLM 缺乏动态映射机制，导致同质化指标产生差异化评估结果。需建立领域自适应的指标解释框架以提升评估一致性。

3.4.3 多模态信息处理缺陷

LLM 对非文本要素的解析能力严重制约工程技术论文评估：其一，无法识别公式重构中的方法创新（如新型材料性能模型的数学表达）；其二，忽略实验参数优化（如控制变量设计）的技术突破价值；其三，跨模态关联断裂导致图文逻辑链割裂（如流程图与数据表对应关系误判）。此类缺陷导致机器评分明显低于人工评分。

3.4.4 内容虚构生成风险

当前，LLM 生成内容存在学术诚信风险。网络数据中低质量语料众多，导致虚假知识传播风险上升，且虚构内容难以识别，影响学术评审的严谨性。针对此，可采取相应措施防控 LLM 内容虚构风险：其一，通过建立学科特征筛选机制，优先采用期刊论文、专利文献等高质量数据源；其二，针对模型偏好高频表达的“泛化惯性”，可引入领域知识验证模块，通过专业术语库比对和检测，保障学术创新观点的完整呈现；其三，构建“可追溯生成”系统，对关键论断自动标注推理路径和数据依据。通过建立输入净化 - 过程约束 - 输出溯源的三层逻辑框架，既保留 LLM 的知识整合优势，又能有效控制虚构内容对学术评审的干扰，促进 LLM 与学术规范的良性互动。

3.4.5 跨学科知识整合障碍

面对工程技术的学科融合趋势，LLM 表现出三重整合局限：术语系统冲突（如“可靠性”在机械与电子学科的语义差异）、方法迁移误判及创新涌现识别失效。需开发跨域知识对齐算法以捕捉协同创新价值。

综上实证表明，DeepSeek 在语言表述逻辑评估得分占比 73.8%，在体例规范评估得分占比 77.6%，说明结构化指标评估中表现较优，但在核心创新维度呈现明显局限性：理论创新评分得分占比 59.6%，方法创新评价评估得分占比 65%。实证揭示 LLM 存在语义解析浅层化、多模态处理缺陷、数据时效滞后三重困境，导致其难以捕捉工程技术领域渐进式创新的实际价值，如图 2 所示。

实证测试揭示了 LLM 在创新评估中的效能边界，为人机协同机制的建构提供了关键切入点，需再进一步对比分析人机协同模型设计的可行性。

doi:10.3772/j.issn.2095-915x.2026.01.004

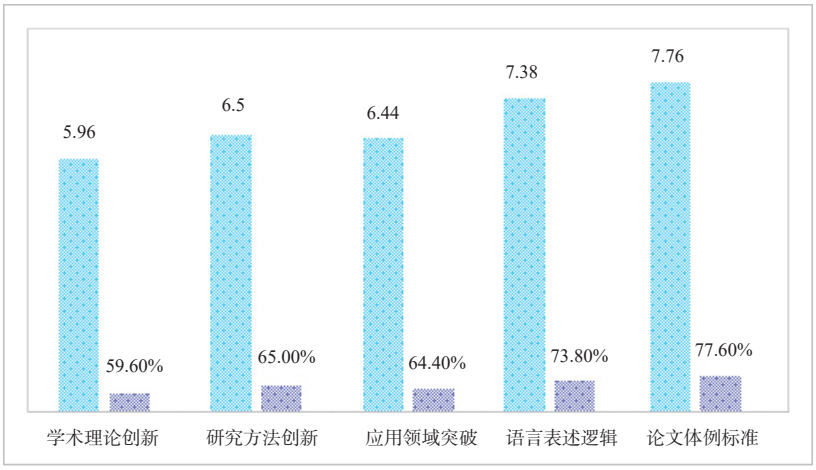

图 2　DeepSeek 测试样本评估得分平均占比情况

4　大语言模型下论文学术人机协同评估

4.1　效能互补性分析

基于 LLM 在理论创新识别、多模态处理等方面的局限性，本研究选取工程造价管理专业的三篇论文样本开展人机评审对比研究。实验样本遵循同质化原则，确保论文在学科领域、被引频次与文本结构上具有可比性。DeepSeek 生成机审内容及评分，与人工评审进行对比分析，得出人机评审效能对比特征表，如表 4 所示。以 BIM 技术应用的论文样本为例，人工评审能凭借专家的知识能力，发掘出 BIM 技术通过参数映射机制改良实现工程预算效率提升的隐性价值，而 LLM 因过度聚焦理论完整性评估，低估或过滤了 BIM 技术的应用贡献。实验表明，机器擅长结构化指标筛查（如格式规范、文献著录），而专家长于渐进式创新的深度解析，二者协同可突破单一评估模式的认知盲区。

表 4　人机评审效能对比特征

论文题目	核心创新类型	人工评审关注焦点	机审评估	人机差别对比
1. 基于正向设计 BIM 模型的造价建模规则研究	跨学科工具整合	评价：文章理论视角有新意，研究方法能解决实际应用问题……可有效提升正向设计模型计量精确度和效率性，具有较强的 BIM 技术示范作用……（审稿日期 /2021.11.23）	评价：针对 BIM 模型在造价计量中的应用提出了建模规则优化方案，但未突破现有理论框架…… 综合评分：5.4 分	在确定技术创新点方面，需发挥人工深度评审作用
2. 基于 FDEMATEL-ISM-MICMAC 方法的装配式建筑成本影响因素研究	渐进式方法优化	评价：选题具有学术研究和实用价值。研究思路和方法正确、可行，结论正确，总体结构层次合理，行文较规范。提几点意见……（审稿日期 /2024.5.19）	评价：论文整合 BIM 与 ERP 系统构建数据平台……属于技术方案创新而非理论体系创新。 综合评分：7.2 分	同上
3. 基于大数据的工程造价信息资源共享方法	新技术转化应用	评价：文章针对信息孤岛……提出了一种整合方法，来构建企业集成数据平台……选题有热度，研究内容饱满，具有实践创新性。（审稿日期 /2024.7.22）	评价：该论文实现跨模型优势互补，但未突破既有理论范式或重构学科基础概念…… 综合评分：7.2 分	DeepSeek 对技术应用转化的敏感度不足，需深度训练及知识更新

4.2 人机协同评审机制设计

针对人机协同评审的互补特性，本文构建了学术论文人机协同评审的机制模型，如图 3 所示。机审预筛层通过 LLM 完成格式审查、创新点初评及风险分级；人机交互层由专家结合 BIM 模型等非文本素材复核争议内容，并反哺动态知识库；决策优化层依托跨学科会审消解分歧，形成闭环评审流程。该机制实现三重协同效应——效率与质量平衡、显性与隐性价值互补、静态与动态知识融合，使 LLM 承担 30%—50% 机械性的例常评审工作，从而帮助编辑和评审专家节约大量时间，可以有效进行论文创新点的评估工作。

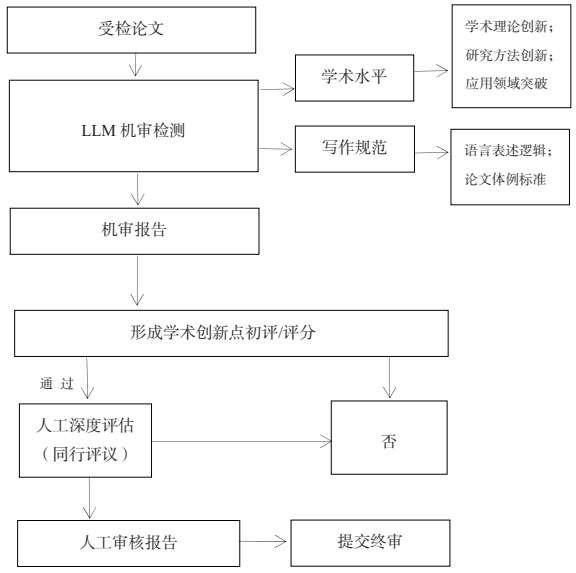

图 3 学术论文人机协同评审机制模型

4.3 人机协同机制的理论价值

人机协同机制有效克服 LLM 的技术局限：通过专家介入减少渐进式创新误判，人工多模态解析弥补非文本处理缺陷，跨领域协作消解学科术语壁垒。其理论价值在于推动学术评价范式从经验主导转向数据 - 知识双驱动，既保留人工评审对隐性价值的敏锐洞察，又发挥机器在规模化处理中的效率优势，为破解传统评审路径依赖提供了新的参考方法。

因此，通过人机协同对比实验，进行人机协同创新评审，研究发现论文的学术创新性评价变得更加客观、全面和高效，证实了人机协同评审机制设计的可行性。

5 结论和展望

5.1 研究结论

本文构建了三级评价指标体系，对机械工程等五个工程技术学科论文样本开展 LLM 创新性评估，揭示其效能边界与协同潜力。研究结论表明：（1）评估效能分层：LLM 在语言逻辑、体例规范等结构化指标评估中展现显著效率优势，但对理论创新和应用突破等核心维度的评分偏差率较高，验证了其技术局限性；（2）技术瓶颈：LLM 语义解析浅层化、存在多模态处理缺陷与数据时效滞后问题，对隐性创新贡献识别准确率低，难以有效捕捉工程技术领域渐进式创新价值；（3）跨学科研究评估结果波动性显著，模型对学科术语识别错误率高，凸显学科术语壁垒对 LLM 创新识别的制约。基于此，本文提出的人机协同评审机制，通过机器预筛与专家复核的递进式协同，有效平衡效率与质量，为解决传统评审中的路径依赖问题提供了方法论创新。

5.2 研究展望

本文在实验样本采集与数据分析方面受限

doi:10.3772/j.issn.2095-915x.2026.01.004

于技术、资源等因素，未来研究可进一步扩大学科覆盖范围并增加样本数量，以深化多领域探索及数据验证。今后可从三方面深化人机协同机制：（1）系统架构优化：开发动态知识库与多模态解析工具，构建学科自适应的智能评审系统，提升 LLM 对隐性创新与跨学科成果的识别精度；（2）技术融合创新：探索区块链技术与人机协同的深度融合，建立可追溯、防篡改的评审数据链，增强评估过程的透明性与可信度；（3）应用生态扩展：针对工程技术细分领域定制评估插件，例如基于 BIM 模型的造价创新分析模块或材料性能多模态解析工具，推动智能评审从理论框架向产业实践转化。通过持续优化人机分工逻辑与技术适配性，最终实现学术评价从“效率优先”到“质量 - 效能协同”的范式跨越与价值重塑。

参 考 文 献

[1] 美通社 . Forrester 发布 2023 年至 2027 年全球科技市场预测 [R/OL]. (2024-01-19) [2024-03-01]. https://www.prnasia.com/lightnews/light-news-1-102-70379.shtml.

[2] 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(2): 84-93.

[3] EPOCH AI. Trends in Machine Learning Compute[R/OL]. (2022-05-15) [2024-07-20]. https://epochai.org/blog/compute-trends.

[4] 谢维熙 , 张光耀 , 王贤文 . 开放同行评议视角下学术论文同行评议得分与被引频次的关系 [J]. 中国科技期刊研究 , 2022, 33(1): 113-121.

[5] 沈阳 , 余梦珑 . 重构智能交互范式：基于 DeepSeek 的提示强化与人机协同 [J]. 新疆师范大学学报（哲学社会科学版）, 2025, 46(4): 90-98.

[6] 蒋建斌 . 科技期刊同行评议中非共识问题的分析与对策 [J]. 科技传播 , 2024, 16(24): 69-72.

[7] 龚梦月 . 开放科学背景下学术期刊同行评议各方主体权利义务及其辩证关系研究 [J]. 科技与出版 , 2024(5): 114-120.

[8] 徐玢 . 全球学术生态中的中国审稿人角色 [N]. 科技日报 , 2024-06-28(004).

[9] 甘甜 , 关良宝 . 期刊评价影响下的工程技术类期刊发展问题及对策 [J]. 编辑学报 , 2022, 34(6): 606-610, 617.

[10] 付国乐 , 张志强 . 学术期刊社会功能评价体系构建与应用——以我国卓越计划资助英文期刊为例 [J]. 出版科学 , 2024, 32(2): 24-33.

[11] 高波 , 申晨荣 . 新时代企业家精神的“创造性破坏”——基于关键核心技术突破视角 [J]. 上海经济研究 , 2025(2): 65-78.

[12] 李帅 , 陈定权 . 数据为基：信息资源管理学科知识体系的重构策略 [J]. 图书馆论坛 , 2024, 44(12): 1-7.

[13] 蒋建斌 . 工程技术类期刊对发展新质生产力的促进作用探析 [J]. 新闻研究导刊 , 2025, 16(5): 181-184.

[14] 孟旭阳 , 白海燕 , 吕世炅 , 等 . 大模型赋能下学术文献服务中的智能化应用研究 [J]. 情报工程 , 2025, 11(1): 3-17.

（责任编辑：浦墨）