Service Science and Management 服务科学和管理, 2026, 15(6), 1122-1129
Published Online November 2026 in Hans. https://www.hanspub.org/journal/ssem
https://doi.org/10.12677/ssem.2026.156132

Hans汉斯

基于大语言模型的科技计划项目绩效评价研究

舒　毅，饶兰香*，胡家余，罗怡欣，孙　丹

江西省科技基础条件平台中心软件和大数据研究室，江西　南昌

收稿日期：2026年10月6日；录用日期：2026年10月28日；发布日期：2026年11月10日

摘　要

针对科技计划项目绩效评价维度单一、过度依赖专家主观经验、非结构化材料处理效率低下，成果真实性与创新性难以开展深度分析等问题，提出一种基于大语言模型的智能绩效评价模型。该模型对项目全周期文本开展深度语义理解，从成果真实性、目标相关性、时间有效性、成果创新性四个维度开展辅助定性研判，输出风险提示与参考线索。以江西省2021~2025年1625项省级科技计划项目为样本开展模型推理，模型共触发风险提示97次，涉及74个项目，完成100个项目的材料筛查、信息比对与风险线索输出平均耗时0.5 h，约为人工开展同等审核工作耗时的十分之一，初步表明该方法在减轻评审工作负担、拓展绩效评价分析维度方面的实用性。

关键词

大语言模型，科技计划项目，绩效评价

Research on Performance Evaluation of Science and Technology Plan Projects Based on Large Language Models

Yi Shu, Lanxiang Rao*, Jiayu Hu, Yixin Luo, Dan Sun

Jiangxi Science and Technology Infrastructure Center Software and Big Data Research Lab, Nanchang Jiangxi

Received: October 6, 2026; accepted: October 28, 2026; published: November 10, 2026

Abstract

Addressing issues such as the single dimension of performance evaluation for science and technology plan projects, over-reliance on subjective experience of experts, inefficient processing of unstructured materials, and insufficient in-depth analysis of the authenticity and innovativeness of outcomes, an intelligent performance evaluation model based on a large language model is proposed. This model conducts deep semantic understanding of project text throughout the entire cycle, completes qualitative

*通讯作者。

文章引用: 舒毅, 饶兰香, 胡家余, 罗怡欣, 孙丹. 基于大语言模型的科技计划项目绩效评价研究[J]. 服务科学和管理, 2026, 15(6): 1122-1129. DOI: 10.12677/ssem.2026.156132

evaluation from four dimensions: authenticity of outcomes, relevance to objectives, time effectiveness, and innovativeness of outcomes, and outputs risk warnings and comprehensive research judgments. Using 1625 provincial science and technology plan projects in Jiangxi Province from 2021 to 2025 as test samples, the model triggered risk warnings 97 times, involving 74 projects. Testing showed good consistency between risk warnings and acceptance conclusions; processing 100 projects took an average of 0.5 hours, which is about one-tenth of the traditional review cycle. This preliminarily verifies the practicality of this method in improving evaluation efficiency and deepening analysis dimensions.

Keywords

Large Language Model (LLM), Science and Technology Plan Projects, Performance Evaluation

Copyright © 2026 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

科技计划项目是为解决经济和社会发展中的科学技术问题，所开展的一系列独特的、复杂的并相互关联的科研活动。据统计，2024 年，全国共投入研究与试验发展(R&D)经费 36326.8 亿元，比上年增长 8.9%，研究与试验发展(R&D)经费投入强度为 2.69% [1]。在这一背景下，如何科学、客观、全面地评价科研项目的实施绩效，确保财政资金的配置效率和使用效益，已成为科技管理部门和社会公众共同关注的焦点问题。

为进一步规范科技项目绩效管理，推动绩效评价工作标准化落地，江西省科学技术厅印发《江西省科技计划项目综合绩效评价工作实施细则（试行）》[1]通知，明确对基础研究计划、科技重大专项、重点研发计划等各类项目实施分类评价；针对省自然科学基金项目，配套发布《江西省自然科学基金项目综合绩效评价工作实施细则（试行）》[2]，细化基金类项目的评价组织流程与考核管理要求。政策层面已经建立起分类绩效评价制度框架。

传统科技计划项目绩效评价主要依靠专家人工审阅任务书、结题报告、论文、专利等海量非结构化佐证材料进行判断。面对每年的结题项目，人工评审模式存在评审周期长、人力成本高、主观经验依赖性较强等问题，同时对于成果填报真实性、成果与研究任务匹配程度、成果创新性等内容，很难做到完整的外部交叉核验；同时单纯依靠人工阅读文本，也难以充分挖掘海量材料背后的深层信息，制约绩效评价的客观性与工作效率。

随着大语言模型等人工智能技术快速发展，为科研项目管理带来了更多的实现路径。将大语言模型引入科技计划项目绩效评价，可以承担成果信息抽取、多源数据比对、文本语义解析等重复性工作，弥补传统评审模式的短板。基于此，本文构建一套基于大语言模型的科技计划项目智能绩效评价模型，从成果真实性、目标相关性、时间有效性、成果创新性四个维度实现项目材料自动化研判，输出风险提示与综合参考意见。

2. 研究背景

2.1. 当前绩效评价存在的问题

第一，评价过程缺乏有效核验手段。孙洪义[2]以农业科技计划项目为例，指出在评价过程中，指标

[1]https://kjt.jiangxi.gov.cn/jxskxjst/col/col64258/content/content_2003755757496963072.html
[2]https://kjt.jiangxi.gov.cn/jxskxjst/col/col64258/content/content_2003760772588507136.html

数据收集和整理工作主要由项目承担单位完成，部分承担单位为追求良好的评价结果，可能提供虚假数据或隐瞒信息，数据的真实性、完整性和准确性难以得到保证。王文芳[3]指出，受时间紧、任务重的刚性约束，专家评价异化为“合规性审查”，评价仅聚焦材料齐全、格式规范、指标达标等表层要素，对研究成果的真实性、可靠性、创新性缺乏系统分析。

第二，评价质效不高。刘代玉[4]指出，专业机构组织专家集中验收，时间短、数量多，评价不深入，容易忽略项目的创新性成果。万昊[5]等指出，国家层级的大型研发评价实践“需要消耗巨大的人力、物力和财力”。组织大规模专家评审需要投入大量人力、物力和时间成本。以江西省自然科学基金项目绩效评价为例，每年有几千个项目需要组织专家进行绩效评价，涉及抽取专家、组织专家评审等环节，周期通常较长。一个项目涉及的附件材料有十多项，需要人工逐一筛查。

2.2. 绩效评价研究现状

围绕科技项目绩效评价，国内学者已经开展大量研究，当前的研究主要集中在评价指标体系构建、评价方法优化两个方向。

在评价指标体系构建方面：施筱勇[6]等以绩效评价理论为指导，借鉴国内外公共财政支出项目绩效评价方法，设计了包含相关性、目标实现、创新性、转化与效果、影响和实施管理 6 个准则和 14 个关键指标的科技项目绩效评价指标体系。王忠[7]等基于创新质量和贡献导向，采用德尔菲法识别、赋权和筛选，形成了基础研究类、开发研究类和应用示范研究类三类科研项目的绩效评价指标体系。针对不同类别的科技项目，学者们也开展了针对性的研究，李超[8]等针对辽宁省农业科技项目，运用专家打分法和层次分析法构建了包含经济效益、技术创新、社会效益和企业成长四个维度的绩效评价体系。张玉娇[9]等基于国家自然科学基金 2020 年度绩效评价实践，研究和建立了“三青”项目(青年基金、优青基金、杰青基金)的绩效评价方法与指标体系，并进行了实证分析。石聪聪[10]基于 2022 年验收的贵州省科技计划基础研究项目数据，构建项目绩效评价指标体系，采用层次分析法和熵权法组合计算权重。

在评价方法优化方面，国内学者尝试各种统计学方法引入到项目绩效评价当中。万昊[5]等借助 BP 人工神经网络模型，从“投入–产出”视角选取 36 个指标模拟项目评审专家组的决策打分，并添加贝叶斯正则化修正项提高模型预测精度。这给研究提供了一个独特的视角，即绩效评价可以通过神经网络技术来进行。楼雯[11]等基于要素化大模型从科技成果中提取要素，提高评价效能。马威[12]等、王奕博[13]等分别运用随机层次分析法、最佳最差法等方法对科技项目绩效进行评价。

综合现有文献可以看到：现有研究停留在指标体系构建与优化层面，缺少直接面向项目全周期原始文本，利用大语言模型完成多维度自动研判的落地实践。层次分析法、BP 神经网络等传统统计学与机器学习方法，能够实现指标加权、模拟专家打分，但这类模型高度依赖人工前置完成指标提取与数据整理，仅能接收结构化表格数据作为输入，无法直接解析任务书、论文、专利说明书等原始非结构化结题文档。大语言模型可直接解析项目文档，自动完成成果识别、信息比对与语义解析，以此缓解大批量项目评审中的人工材料筛查压力，这正是本文的研究切入点。

3. 相关概念与技术基础

3.1. 大语言模型

大语言模型(LLM)基于海量文本预训练，具备强大的非结构化文本理解、实体信息抽取、文本解析能力。不同于传统关键词检索，大模型可以理解自然语言背后的语义逻辑，能够对论文摘要、专利权利要求、项目任务书等长文本进行解析，为本研究多维度智能评价提供核心底座。但大模型普遍存在幻觉问题[14]，必须结合外部权威数据源约束校验。

3.2. OCR 识别技术

科技计划项目的归档材料中包含大量扫描件、图片格式的文档——如签字盖章的任务书、专利证书扫描件、软著登记证明等，这些图像格式材料无法被大语言模型直接读取，需要先转换为可编辑文本。OCR技术承担这一前置转换任务：将扫描件或图片中的文字区域识别提取，转换为可编辑、可检索的文本格式。

3.3. 语义向量化

语义向量化将自然语言文本映射到高维向量空间，语义含义相近的文本在向量空间中距离更近，通过余弦相似度可以量化两段文本语义的相似程度。该技术突破传统关键词匹配局限，可以识别“同一科研含义、不同专业术语表述”的文本关联，是相关性分析的技术底座。

3.4. 检索增强生成技术(RAG)

RAG 技术可以通过外部知识检索的方式增强大语言模型的能力，有效减少 LLM 的幻觉现象[15]，通过对接 CrossRef、国家知识产权局等权威数据库，提升绩效评价结果的客观性与可信度。

3.5. 大语言模型用于绩效评价的优势

结合科技计划项目绩效评价的业务场景，对比传统人工评审方式，大语言模型在绩效评价工作中的优势主要体现在三方面。

第一，具备大批量非结构化文档快速处理能力。大语言模型可实现批量完成实体抽取、文本解析工作，能够承担大量机械性筛查任务，有效缓解评审专家和科技管理人员的工作负荷。第二，可实现多源信息自动化交叉核验。借助 RAG 检索增强技术，解决传统评审中很难做到全部成果逐项核验的现实难题，降低人工漏审、错判概率，提升成果真实性核查的客观性。第三，可实现深层语义关联识别。依靠语义向量化技术，模型能够挖掘任务书与成果材料之间的语义关联，不受专业术语表述差异的干扰。

4. 基于大模型的项目绩效评价

本文以大语言模型为核心底座，融合 OCR 文档识别、文本语义向量化、RAG 检索增强生成技术，构建真实性-相关性-时间有效性-成果创新性四维智能评价模型，实现对项目全周期材料自动化解析，整套体系的技术路线如图 1 所示。

4.1. 材料真实性验证

真实性验证核心用于判别项目产出成果是否真实合规。传统评审模式下，专家需人工核验论文、专利、软件著作权等全部佐证材料：通过知网、万方检索论文真伪，核对专利申请号、权属人信息，校验软著登记备案记录。单批次评审往往涉及数百项成果，逐项人工比对工作量巨大，极易出现疏漏。本研究把这一环节交给大语言模型来处理。利用 OCR 识别技术，从非结构化文本中提取标题、作者、期刊名称和 DOI 号等核心标识信息，以这些标识作为索引，构建为检索字段。系统优先比对 DOI 号，如成果材料中填写了 DOI，以 DOI 为唯一标识直接定位记录；对于未填写 DOI 的成果，系统以“标题 + 第一作者 + 期刊名称”为组合条件进行检索。比对结果分为三类：完全一致标记为“真实”，关键字段存在差异标记为“存疑”，检索不到标记为“待人工复核”。

4.2. 相关性分析

项目成果与任务合同书之间是否“对得上”，是判断项目执行质量的另外一个维度。传统专家凭借自身知识储备开展判断，容易出现判断误差。

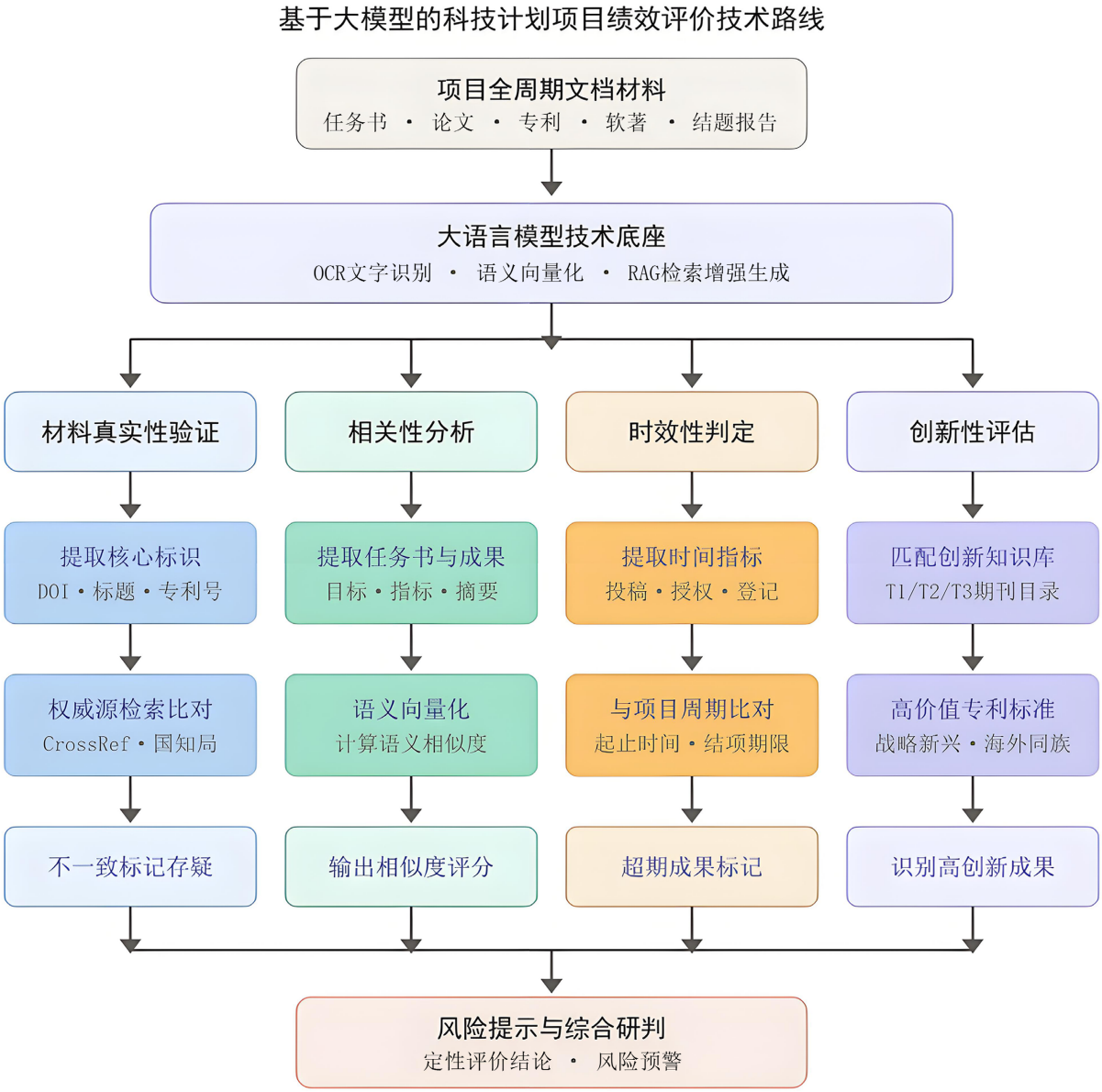


Figure 1. Technical route for performance evaluation of science and technology plan projects based on large models
图 1. 基于大模型的科技计划项目绩效评价技术路线

大模型通过提取任务书中的核心内容(研究目标、关键技术、考核指标)以及成果材料中的关键章节(论文摘要、专利权利要求书、软著说明书)，分别输入到模型里转化成语义向量，然后计算这两组向量在语义空间中的相似度，相似度越高，说明成果与任务书的相关性越强。这一做法突破了传统关键词匹配的局限，能够精准捕捉成果与任务书之间的深层语义关联，有效识别项目成果与任务书的相关性。相关性分析中语义相似度的阈值通过人工标注试验确定：从测试样本中随机抽取 500 组“成果－任务书”，邀请多位具有 5 年以上科技项目管理经验的专家分别标注“相关”与“不相关”，在不同阈值下比较模型判定与人工标注的一致性，0.6 时综合一致率最高，因此选定 0.6 作为相关性判断的最佳阈值。

4.3. 时效性判定

时间有效性即成果产出时效性，核心判断各项产出成果是否落在项目法定实施周期以内。科技计划

项目具备明确的立项、结题时间边界，原则上项目核心产出应当产生于项目执行周期，超周期产出的成果不能直接纳入本项目绩效。不同类别科技项目执行周期差异较大，自然科学基金类项目周期多为 1 至 2 年，重点研发计划项目周期可达 3 年以上。传统评审需要人工逐个查阅论文刊发、专利授权、软著登记时间，和项目起止时间做对照，大批量项目评审场景下人工核对极易发生疏漏。

大模型通过提取各类时间指标(论文的投稿、录用、优先出版、正式刊发时间；专利的申请、授权时间；软著的开发完成、登记时间)，与任务书中明确的“项目起止时间”或“结项期限”自动比对。论文以正式出版时间为准，专利以授权公告时间为准，软件著作权以登记完成时间为准。对于产出时间早于项目立项时间，或是晚于项目结题(含延期)时间的成果，统一标记为超期成果，输出风险提示，提醒评审专家重点甄别该成果是否归属本项目资助产出。

4.4. 创新性评估

创新性评估旨在从批量项目中识别高水平科研产出，对成果创新层级做初步筛查。大语言模型现阶段难以直接完成学术原创价值的深度研判，因此模型依托公开权威标准建立创新知识库，采用规则匹配的方式实现批量初筛。

针对论文成果，引入中国科学技术协会发布的《高质量科技期刊分级目录总汇》[3]，把 T1、T2、T3 等级期刊纳入知识库，若论文发表期刊属于该目录，则标记为高创新等级论文成果。针对专利成果，参照国家知识产权局高价值发明专利判定标准，满足战略性新兴产业专利、海外同族专利、维持年限超 10 年、高质押融资金额、获得国家级专利奖项任一条件，即判定为高价值发明专利。

大模型负责完成成果基础信息抽取，和创新知识库开展检索匹配，完成创新水平初筛。该模块不直接给出最终创新结论，主要用于从海量项目中快速筛选出潜在高价值成果，为专家深度研判提供线索。

4.5. 风险提示与综合研判

真实性验证、相关性分析、时效性判定、创新性评估四个模块独立进行，分别输出各自的标签、量化参考信息以及单项风险标记。模型不对项目直接给出最终验收意见，而是汇总生成结构化风险提示清单。清单将真实性存疑、超期产出、语义相关性偏低等疑点逐条罗列，同时标注识别得到的高创新性成果，形成完整的辅助研判材料，供评审专家查阅。

5. 大模型的绩效评价实践

5.1. 数据预处理

为适配大模型解析要求，需要对项目原始文档开展预处理并构建领域知识库。在数据预处理过程中，先要对项目全周期文档进行系统化整理。对于扫描件调用 OCR 引擎进行文字识别；识别完成后开展文本清洗，去除页眉页脚、无关标记、乱码等噪声信息。进一步区分文档类型(如任务书、进展报告、结题报告)，为每个文本片段添加元数据标签，构建结构化项目文档集，以便大模型能够准确检索和定位关键信息。同时导入《高质量科技期刊分级目录总汇》、高价值发明专利判定规则，构建本地创新知识库，为创新性评估模块提供支撑。

5.2. 大模型测试结果

本研究涉及的模型调用集中在两个环节：一是大语言模型完成信息抽取和语义解析，二是嵌入模型完成文本向量化转换。

[3]https://www.cast.org.cn/xw/gzdt/KJCX/art/2025/art_7affb3ed494c436b9a69aedb8604e400.html

大语言模型方面，基础模型选择 Qwen2.5-72B-Instruct，模型参数量 72 B，采用温度系数 temperature = 0.1 降低输出随机性，用于实体抽取、时间解析、RAG 查询字段组装；嵌入模型采用 text-embedding-3-small，输出向量维度 1536，用于任务书文本与成果摘要的语义向量化转换。

在模型处理过程发现，平均处理(包括材料解析、信息比对与风险线索输出) 100 个项目的耗时约为 0.5 小时。人工完成同等工作量的材料筛查——即逐项核对成果真实性、检查成果时间是否在项目周期内、判断成果内容与任务书的关联度——以一位熟练管理人员的操作速度为基准，单个项目平均耗时约 3 分钟，100 个项目合计需 5 小时，大幅缩短了人工审核的效率。在 1625 个测试样本中，模型共触发各类风险提示 97 次，涉及 74 个项目(部分项目有多项提示)。提示类型包括真实性存疑、超期产出、相关性偏低和创新性突出，分布如表 1 所示。

Table 1. Distribution of risk warning types for large model modules
表 1. 大模型模块风险提示类型分布

提示类型	提示次数	涉及项目数	提示内容说明
真实性	12	7	成果元数据与权威数据库不匹配、成果疑似非本项目产出
有效性	58	43	成果正式发表/授权时间在项目签订任务书前或结题后
相关性	19	16	成果内容与任务书研究方向关联度低于阈值
创新性	8	8	成果指标在知识库清单中检索成功，判定为高创新成果
总计	**97**	**74**	

模型解析结果如下：

1) 真实性存疑：某自然科学基金项目结题材料中填报一篇期刊论文，项目负责人为第一作者。模型检索 CrossRef 返回的记录显示该论文的实际第一作者为另一人，填报信息与数据库记录不一致，模型标记真实性存疑。

2) 有效性存疑：某现代农业领域重点研发项目，项目执行周期 2022-01~2024-12。项目结题报告填报 3 项发明专利，模型提取专利授权公告时间，发现其中 2 项专利授权时间为 2021-08，早于项目立项起始时间。模型标记为超期产出风险，提示专家甄别该专利是否为本项目资助产出。

3) 相关性不足：某生物医药类基金项目，任务书核心任务是开发靶向缓释制剂技术。项目提交一篇论文主题为普通药物体外毒性试验，模型对任务书关键技术文本与论文摘要做语义向量化，通过计算余弦相似度判定为不相关。该论文关键词没有明显冲突，传统关键词匹配无法识别，但语义层面和项目核心任务关联较弱，模型依靠语义相似度识别出该风险并标记存疑。

4) 创新性突出：某重点实验室项目要求产出一项发明专利，模型提取专利 IPC 分类号、专利维持年限信息，与本地知识库高价值专利标准比对，该专利属于战略性新兴产业领域，且海外布局同族专利，模型标记为创新性突出成果。

6. 结论与展望

6.1. 研究结论

针对第一章提出的科技计划项目绩效评价存在评价过程规范性不足、评价质效不高、评价维度单一，非结构化材料处理繁重，过度依赖专家主观经验等现实问题，本文构建融合 OCR 文字识别、语义向量化、RAG 检索增强生成技术的大语言模型智能绩效评价模型，从成果真实性、目标相关性、时间有效性、成果创新性四个维度实现项目全周期材料的自动化辅助研判，并采用江西省 2021~2025 年 1625 项省级结题

科技计划项目开展实证测试。实验过程中，模型完成 100 个项目的材料筛查、信息比对与风险线索输出平均耗时 0.5 小时。以经验工时估算，人工完成同等规模的成果真实性核查、时间核对、相关性判断等基础筛查工作约需 5 小时(每个项目材料平均耗时三分钟，总共 100 个项目需要 5 小时)，模型耗时约为人工的十分之一。依托 RAG 对接权威数据源，可对论文、专利等成果开展真实性校验，降低人工逐条核验带来的漏审风险；借助语义向量化计算能够缓解跨领域评审场景下，专家对细分方向专业术语变体不熟悉造成的研判偏差，拓展传统绩效评价的分析维度，模型输出的风险提示与项目已有的正式验收结论具备较好一致性。

6.2. 研究局限与定位

本研究验证了大语言模型在科技计划项目绩效评价中的可行性，但模型本身仍存在局限，需要在后续研究和实际部署中持续关注。

第一，当前模型仅支持文本类材料的解析，如任务书、结题报告、论文摘要、专利说明书、软著文档等。但针对非文本类材料，如照片、视频等无法对其真实性、技术含量或完成质量做出判断。

第二，模型解析效果高度依赖归档文档质量。当原始扫描件 OCR 识别错乱、关键佐证材料缺失，会造成模型识别准确率下降。

综合上述考虑，本研究定位为：为科技计划项目绩效评价提供辅助判断，帮助评审专家快速完成材料筛查和风险点识别等基础性工作，降低重复性劳动。各维度输出的“存疑”“超期”“相关性偏低”等提供决策依据——最终验收结论由专家基于完整材料综合判断。

基金项目

江西省管理科学项目(项目编号 20252BAA100042)。

参考文献

[1] 2024 年全国科技经费投入统计公报[EB/OL]. 2025-09-29. http://www.gov.cn/lianbo/bumen/202509/content_7042859.htm, 2026-04-03.

[2] 孙洪义, 王之旭, 孙小丁. 农业科技项目绩效评价优化对策探究[J]. 中国农业会计, 2025, 35(19): 36-38.

[3] 王文芳. 江西省科技计划项目综合绩效评价优化路径探析[J]. 科技广场, 2025(6): 90-96.

[4] 刘代玉, 夏润龙, 欧灵, 等. 科技项目绩效评价优化——以湖南省项目验收评价实践为例[J]. 中国科技信息, 2025(20): 136-138.

[5] 万昊, 谭宗颖, 张福俊, 等. 项目验收的同行评议辅助决策评价方法研究——基于贝叶斯正则化修正的 BP 人工神经网络模型[J]. 情报杂志, 2017, 36(11): 192-199.

[6] 施筱勇, 杨云, 迟计, 等. 科技项目绩效评价指标体系研究[J]. 科技管理研究, 2016, 36(10): 39-43.

[7] 王忠, 文宇峰, 孙玉芳, 等. 创新质量和贡献导向下科研项目绩效评价体系研究[J]. 管理科学, 2021, 34(1): 28-37.

[8] 李超, 张玉东, 宋春雨, 等. 辽宁省农业科技项目绩效评价指标体系构建研究[J]. 农业经济, 2024(3): 121-123.

[9] 张玉娇, 苑怡, 冯勇, 等. 基础研究人才计划项目绩效评价方法与实证——以“三青”项目为例[J]. 科学学与科学技术管理, 2023, 44(11): 3-14.

[10] 石聪聪, 宋雪. 贵州省科技计划基础研究项目绩效评价研究[J]. 青海科技, 2024, 31(2): 78-85.

[11] 楼雯, 赵星. 要素化大模型增强的科技成果评价理念前瞻[J]. 中国图书馆学报, 2025, 51(4): 50-65.

[12] 马威, 李屹然. 一种利用 SAHP-云模型的科技投入项目绩效评估方法分析[J]. 西南大学学报: 自然科学版, 2023, 45(10): 160-168.

[13] 王奕博, 陈俊清, 闫文敏, 等. 基于权重偏离和最佳最差法的重大科技工程财政绩效关键要素识别[J]. 哈尔滨工程大学学报, 2025, 46(9): 1910-1920.

[14] 何金骅, 王洪俊. 大语言模型幻觉研究综述[J]. 人工智能与机器人研究, 2026, 15(1): 156-167.

[15] 张超然, 梁素平, 杨杰, 等. 基于大模型 RAG 架构的专利检索 Agent 研究[J]. 中国发明与专利, 2025, 22(2): 14-25.