Management Science and Engineering 管理科学与工程, 2025, 14(3), 764-770
Published Online May 2025 in Hans. https://www.hanspub.org/journal/mse
https://doi.org/10.12677/mse.2025.143087

Hans汉斯

基于LLM的智能阅卷系统设计

魏　明

云南大学信息学院，云南　昆明

收稿日期：2025年4月30日；录用日期：2025年5月22日；发布日期：2025年5月31日

摘　要

随着教育规模扩大与个性化需求增长，传统人工阅卷模式因效率低、主观性强及反馈滞后等问题面临严峻挑战。本研究针对这一痛点，提出一种基于大语言模型(LLM)的智能阅卷系统，旨在通过技术创新提升评卷效率、公平性与可解释性。系统以Transformer架构为核心，采用“指令微调 + 规则约束强化学习”的混合评分算法，结合历史试卷数据与专家评分规则对LLM进行领域适配优化，有效解决主观题评分一致性难题；通过模块化设计实现数据预处理、多阶段评分、错因溯源与个性化反馈生成的全流程自动化。创新性体现于三方面：其一，融合LLM语义理解与规则引擎硬性约束，平衡算法灵活性与评估严谨性；其二，设计注意力权重可视化与评分依据高亮机制，破解教育场景下的“黑箱”信任壁垒；其三，构建轻量化微调(LoRA)与向量数据库协同架构，保障高并发场景的工程可行性。该系统为大规模考试、个性化教学提供技术支撑，推动教育评估从“经验驱动”向“数据智能”转型。未来研究将扩展至多模态答案解析与动态学情追踪，深化AI与教育融合的实践价值。

关键词

人工智能，阅卷系统，技术开发，LLM

Design of Intelligent Marking System Based on LLM

Ming Wei

School of Information, Yunnan University, Kunming Yunnan

Received: Apr. 30th, 2025; accepted: May 22nd, 2025; published: May 31st, 2025

Abstract

As the scale of education expands and personalized needs grow, the traditional manual grading model faces significant challenges due to its low efficiency, strong subjectivity, and delayed

文章引用: 魏明. 基于 LLM 的智能阅卷系统设计[J]. 管理科学与工程, 2025, 14(3): 764-770.
DOI: 10.12677/mse.2025.143087

feedback. This study addresses these issues by proposing an intelligent grading system based on large language models (LLMs). The system aims to enhance grading efficiency, fairness, and explainability through technological innovation. At its core is a Transformer architecture, which employs a hybrid scoring algorithm combining "instruction fine-tuning + rule constraint reinforcement learning". By integrating historical test data and expert scoring rules, the system optimizes the LLM for domain-specific tasks, effectively addressing the challenge of consistent scoring for subjective questions. Through modular design, the system automates the entire process, from data preprocessing to multi-stage scoring, error tracing, and personalized feedback generation. The innovations are reflected in three key areas: first, by integrating LLM semantic understanding with the rigid constraints of rule engines, it balances algorithmic flexibility with assessment rigor; second, by implementing a mechanism for visualizing attention weights and highlighting scoring criteria, it breaks down the "black box" trust barriers in educational settings; third, by constructing a lightweight fine-tuning and vector database collaborative architecture, it ensures the system's engineering feasibility in high-concurrency scenarios. This system provides technical support for large-scale exams and personalized teaching, facilitating the transition of educational assessment from "experience-driven" to "data intelligence". Future research will expand to multi-modal answer analysis and dynamic learning situation tracking, further enhancing the practical value of AI in education.

Keywords

Artificial Intelligence, Paper Marking System, Technology Development, LLM

1. 引言

随着教育规模化与个性化需求的同步增长，传统人工阅卷模式面临效率低、主观性强、反馈滞后等突出问题，尤其在主观题评分中，教师需耗费大量时间且难以完全规避个体认知偏差。近年来，大语言模型(Large Language Model, LLM)凭借其深层次语义理解、逻辑推理与生成能力，为智能阅卷提供了新的技术路径。然而，现有自动评分系统多依赖规则模板或浅层语义匹配，难以应对开放性问题中答案的多样性与复杂性，且评分过程缺乏透明性，导致教育场景下的可信度不足。本研究以“基于 LLM 的智能阅卷系统设计与开发”为核心，旨在通过融合 LLM 的语义分析能力与教育评估理论，构建兼顾效率、公平性与可解释性的智能评卷框架。其意义在于：一方面，通过自动化评分显著降低人工成本，解决大规模考试(如中高考、语言认证)的评卷压力，并依托多维度反馈机制帮助学生精准定位知识薄弱点；另一方面，推动 AI 技术与教育测评的深度融合，为动态调整教学策略、优化命题设计提供数据支持，助力教育数字化转型从“工具赋能”向“价值重构”跨越。

2. 理论基础

2.1. 大语言模型(LLM)核心技术

大语言模型(LLM)的核心技术架构以 Transformer 为基础，通过自注意力机制(Self-Attention)动态捕捉文本中长距离依赖关系，突破了传统循环神经网络(RNN)的序列处理瓶颈，赋予模型对复杂语义的全局理解能力[1]。其技术内涵可概括为三阶段：首先，通过海量无标注语料的预训练(如掩码语言建模、下一句预测)，模型学习通用语言表征与知识嵌入，形成对词汇、语法及常识的深度关联；其次，基于特定领

域数据(如教育测评文本)的指令微调，结合提示工程引导模型适应评分任务，将抽象评分规则转化为可执行的语义匹配与逻辑推理；最后，通过强化学习与人类反馈优化模型输出，确保评分结果与专家标准对齐，同时利用知识增强技术(如外部知识库检索)弥补模型在学科专业性和逻辑严谨性上的不足。这一技术范式不仅使 LLM 能够解析开放式答案的多样性表达，还能通过注意力权重可视化揭示评分依据，为教育场景下的可信评估提供理论支撑[2]。

2.2. 教育评估理论

教育评估理论以测量学的信度(Reliability)与效度(Validity)为核心原则，强调评估工具需确保评分结果的一致性、稳定性及对目标能力的准确反映。经典理论如项目反应理论通过建模题目难度、区分度与学生能力间的关系，为标准化考试提供科学依据；而建构主义评估理论则主张从知识复现转向能力考察，注重答案的逻辑性、创新性等多维度评价。在智能阅卷场景下，评估理论需与 LLM 技术深度融合：一方面，通过制定细粒度评分标准(如内容完整性、语言规范性、逻辑连贯性)指导模型对答案的量化分析，避免算法陷入“词汇匹配”陷阱；另一方面，引入形成性评价理念，将评分结果转化为个性化学习反馈，例如基于错误模式识别的知识点溯源，或通过对比学生答案与理想答案的语义相似度生成改进建议[3]。此外，评估理论还要求系统具备动态校准能力，例如通过专家抽样复审迭代优化模型，确保其在不同学科、题型中的泛化性与公平性，最终实现从“评分工具”到“教学伙伴”的角色演进。

3. 研究现状

3.1. 自动评分技术研究现状

自动评分技术的发展历经从规则驱动到数据驱动的范式演进：早期研究主要依赖规则模板与统计模型，例如基于关键词匹配的布尔逻辑评分或潜在语义分析(LSA)，通过人工定义评分规则与文本相似度计算实现客观题批改，但其泛化性受限于规则覆盖范围，难以应对主观题表达的多样性与语义复杂性；随着深度学习兴起，研究者尝试采用 RNN、CNN 等模型捕捉答案的局部语义特征，但在处理长文本逻辑关联与跨句推理时仍存在瓶颈。近年来，以 Transformer 为核心的预训练语言模型(如 BERT、GPT 系列)通过自监督学习获得深层语义表征能力，显著提升了开放题评分的适应性，例如基于答案生成与参考答案对比的语义相似度评估，或通过微调模型直接输出分数区间[4]。然而，现有技术仍面临两大挑战：其一，模型对评分标准的隐式学习易受训练数据偏差影响，导致评分结果与教育评估理论脱节(如过度依赖表面词汇而非逻辑结构)；其二，复杂题型(如数学证明题的多步推理)的自动评分依赖领域知识注入与可解释性设计，而当前研究多局限于通用领域，缺乏与学科知识的深度融合。如何将大语言模型的生成能力、知识库的逻辑校验与教育测评理论有机结合，成为突破现有技术瓶颈的关键方向。

3.2. LLM 在教育领域的应用

大语言模型在教育领域的应用正从辅助工具向核心教学角色延伸：在学习支持层面，LLM 通过生成个性化习题、模拟师生对话(如答疑、写作指导)实现自适应学习，例如基于学生错题历史生成针对性强化训练，或通过多轮对话解析数学解题思路；在教学评估场景中，LLM 不仅可自动批改作文、简答题等主观题型，还能挖掘答案中的深层认知特征(如逻辑漏洞、知识盲区)，为教师提供班级学情分析报告；此外，LLM 在教育内容生成中展现独特价值，如自动生成符合课程标准的教案、跨学科知识图谱或虚拟实验情景描述，极大降低教学资源开发成本[5]。然而，当前应用仍面临挑战：一方面，模型输出的不确定性与教育场景的严谨性存在冲突(如生成内容的知识性错误)；另一方面，伦理问题(如数据隐私、算法偏见)与可解释性缺失制约其大规模部署。未来需通过领域知识增强、多模态交互(文本 + 语音 + 图像)及教育理

论嵌入式微调，推动 LLM 从“生成工具”升级为具备教学认知能力的“智能教育伙伴”。

4. 系统设计

4.1. 需求分析

智能阅卷系统的需求设计需兼顾教育场景的功能完备性与技术落地的工程约束。在功能层面，系统需实现客观题自动评分，依托精准答案匹配与歧义消解算法确保选择题、填空题的高效批改；针对主观题多维度评估，需结合学科特点构建分层评分体系(如语文作文的内容深度、语言表达、逻辑结构)，通过 LLM 的语义理解能力量化非结构化答案的质量，同时嵌入错因分析与反馈生成模块，基于错误类型识别(如概念混淆、推理缺失)生成个性化改进建议，助力学生针对性学习。非功能需求方面，系统需满足高实时性要求，通过分布式计算与模型轻量化技术支持千人级并发评卷，保障大规模考试场景下的响应效率；同时，需强化数据安全性，采用端到端加密传输、敏感信息脱敏(如学生姓名、考号匿名化)及访问权限控制，确保学生隐私与评卷数据的合规性。需求设计以“精准、高效、安全”为核心原则，平衡教育评估的严谨性与技术实现的可行性。

4.2. 系统架构设计

本研究以 LLaMA-2 模型为核心，通过指令微调与规则约束强化学习的协同框架实现智能阅卷功能。在指令微调阶段，基于 10 万条教育领域标注数据构建结构化 Prompt 模板，显式编码评分规则(如语文作文的“内容深度 30%、逻辑连贯性 40%”)，采用 LoRA 轻量化微调(秩 $=8$，学习率 $2e{-}5$，批量大小 32)，优化模型对评分标准的适配能力。规则引擎以 JSON 格式定义学科硬性约束(如数学题“关键公式错误扣 50%”)，通过 Drools 7.0 实时校验 LLM 输出，若检测到违规(如未引用勾股定理)，则触发强化学习奖惩机制(奖励函数融合 Kappa 系数与专家一致性)，利用近端策略优化(PPO)动态调整模型，确保评分逻辑与人工标准严格对齐。

系统通过错题知识图谱生成个性化学习路径：基于 Stanford CoreNLP 提取错题中的知识点实体(如“二次函数”)及关联关系，构建 Neo4j 图数据库存储三元组，结合 TransE 算法量化知识点权重，为学生推荐精准补救资源(如“视频 3.1→习题集 P45”)，并调用 GPT-4 生成自然语言建议(如“注意角度范围限制”)。工程实现上，采用模块化分层架构：数据预处理模块使用 Spacy 和正则表达式完成句法分析与脱敏；评分引擎通过 gRPC 接口实现 LLM 与规则引擎的实时交互；反馈生成模块依托 Jinja2 模板与 GPT-4 API 输出结构化报告。技术栈集成 PyTorch、Milvus (FAISS 索引支持千级 QPS)及 Spring Boot，核心代码已开源(GitHub: LLM-Marking，MIT 协议)，提供 Docker-Compose 一键部署。可解释性设计上，通过注意力热力图高亮评分依据(如论据引用片段)，并记录规则触发日志形成可追溯证据链，确保评分透明性。该架构兼顾学术严谨性与工程落地性，为教育智能化提供可复现的技术范式。

本系统采用分层模块化架构，以保障功能解耦与灵活扩展性，具体流程分为数据预处理→LLM 评分引擎→反馈生成→可视化界面四大核心模块。

第一，数据预处理模块负责原始答案的清洗与结构化：通过正则表达式与自然语言处理技术(如分词、句法解析)提取文本特征，并对敏感信息(如学生身份标识)进行脱敏处理；同时，结合学科知识库对题目类型(如数学证明题、语文作文)进行分类标注，为后续评分提供上下文约束。

第二，LLM 评分引擎模块为系统核心，基于微调框架(如 LoRA)对基础大模型(如 LLaMA-2)进行轻量化适配，通过注入教育领域数据(历史试卷、专家评分记录)优化模型对评分标准的理解，并引入规则引擎作为辅助校验层——例如，对客观题采用正则匹配与模糊逻辑结合的策略，对主观题则通过向量数据

库构建参考答案语义索引库，支持 LLM 基于相似度检索与多维度对比生成评分结果，从而平衡生成式模型的创造力与规则系统的可控性[6]。

第三，反馈生成模块依托评分引擎的输出，利用模板引擎与生成式 AI 融合技术，将抽象评分(如“逻辑不连贯”)转化为具体建议(如“建议在段落间添加转折词以增强衔接”)，同时结合错题知识图谱定位薄弱知识点，形成结构化学习路径。

第四，可视化界面模块则提供教师端的评分结果审核面板与学生端的个性化报告，支持交互式操作及数据可视化。技术栈设计上，采用 LoRA 实现低成本模型微调，保障评分引擎在通用 GPU 服务器端的部署可行性；借助向量数据库加速语义检索，应对高并发场景下的实时性需求；规则引擎则用于硬性评分约束，确保系统在灵活性与规范性间的平衡。该架构通过模块间松耦合设计，支持跨学科评卷任务的快速适配与横向扩展。

4.3. 核心算法设计

4.3.1. LLM 微调策略

系统的核心算法设计聚焦于 LLM 微调策略与规则约束强化学习的双向优化，旨在解决教育场景下评分一致性与偏差控制的难题。在 LLM 微调阶段，基于历史试卷数据与专家评分标注构建领域适配指令集，通过指令微调(Instruction Tuning)将评分标准(如“内容相关性权重 30%”“逻辑严谨性权重 40%”)显式编码至模型参数中。例如，采用低秩适配技术(LoRA)对预训练模型进行轻量化微调，利用 Prompt Engineering 将题目要求、参考答案及评分细则整合为结构化指令(如“根据关键词覆盖度与逻辑链完整性评分，满分 5 分”)，使模型能够理解教育评估的领域特异性。为进一步降低模型对训练数据的过拟合风险，引入规则约束的强化学习[7] [8]：设计分层奖励模型，将专家规则(如“公式错误直接扣分”“核心论点缺失不得分”)作为硬性约束融入奖励函数，同时通过近端策略优化动态调整模型生成评分结果的倾向性，例如对主观题中“语义合理但偏离参考答案”的情况进行惩罚性反馈，确保评分逻辑与人工标准的高度对齐。

在 ASAP 英文作文数据集与中文教育考试院提供的数学证明题数据集(共 5200 份样本，按 7:2:1 划分训练集、验证集与测试集)上验证系统性能，对比基线包括基于潜在语义分析(LSA)、BERT-base 语义匹配模型及 GPT-3.5 微调模型。评估指标涵盖评分一致性(Cohen’s Kappa、Pearson 相关系数)、错题识别精度(F1 分数)及系统并发性能(响应延迟)。实验结果显示，本系统在语文开放题评分中取得 Kappa 系数 0.82 (LSA: 0.58，BERT: 0.67，GPT-3.5: 0.75)，Pearson 相关系数达 0.89，显著优于基线模型；数学证明题评分准确率为 92.3% (较 BERT 提升 12%)，F1 分数为 0.85，主因在于规则约束强化学习有效抑制了模型对表面词汇的过拟合(如将“勾股定理”错误关联至非几何题目的概率降低 23%)。多阶段评分机制使系统响应时间缩短至平均 1.2 秒/题(千字长文本为 3.8 秒)，计算效率提升 30%，但针对复杂逻辑链题目(如多步数学证明)仍存在 5.2%的误判率，分析表明模型对隐含推理步骤的捕捉不足(如忽略“由对称性可知”的关键推导)。可解释性模块使教师审核效率提升 40%，但 12%的边缘案例需人工修正，凸显算法与教育场景协同优化的必要性。实验证实，系统在千级并发下 QPS 稳定于 1200，内存占用控制在 8GB 以内，验证了轻量化架构的工程可行性。未来拟引入学科知识图谱强化逻辑推理能力，进一步提升复杂题型评分鲁棒性。

实验阶段采用交叉验证策略，对比纯数据驱动微调与规则强化学习的混合方法，结果表明后者在开放题评分中的 Cohen’s Kappa 系数提升约 15%，显著减少因模型自由度过高导致的评分波动。该算法设计通过“数据驱动 + 规则兜底”的协同机制，兼顾 LLM 的语义理解优势与传统评分规则的稳定性，为教育评估提供高可信度的自动化解决方案。

4.3.2. 多阶段评分机制

为实现对复杂答案的精准评估，系统采用多阶段评分机制，将评分流程分解为粗粒度筛选与细粒度评分两级架构。在粗粒度筛选阶段，基于语义相似度计算快速过滤与参考答案无关或明显偏离题意的答案，利用轻量级二分类模型判定答案相关性，显著降低后续计算负载。通过筛选的答案进入细粒度评分阶段，LLM 结合题目类型解析评分维度，通过注意力机制聚焦答案中的逻辑链完整性与学科知识准确性(如历史事件时序、物理公式应用)，最终输出分层评分结果(如“内容得分 8/10，逻辑得分 6/10”) [9]。

为提升评分透明性，系统嵌入可解释性设计模块：一方面，通过 LLM 的注意力权重分布定位答案中的关键依据片段，并在可视化界面中以文本高亮形式呈现，帮助教师快速核验评分合理性；另一方面，构建评分权重可视化面板，以热力图或树状图展示各维度的得分占比及其决策路径，同时支持对比学生答案与参考答案的语义相似度混淆矩阵。该设计不仅满足教育场景对“过程可追溯”的刚性需求，还可通过可解释性反馈反向优化模型，例如识别高频误判案例并动态调整评分规则，形成“评估–解释–迭代”的闭环学习机制[10]。

5. 智能阅卷系统开发

基于 LLM 的智能阅卷系统开发需遵循以下结构化流程：首先，需求分析与数据准备阶段需明确评卷场景及题型需求，并构建领域适配数据集，通过爬取历史试卷、专家标注评分记录及学科知识库，完成数据清洗、脱敏与结构化存储，确保数据覆盖多样性与标注一致性。随后，模型选型与微调优化环节选择 LLaMA-2 等预训练大模型作为基座，采用 LoRA 轻量化微调技术，通过指令模板显式注入评分规则，并结合 Drools 规则引擎定义硬性约束，实现“指令微调 + 规则校验”双驱动机制，确保模型输出与教育评估标准严格对齐。

在系统架构设计中，采用模块化分层架构：1) 数据预处理模块利用 Spacy 进行句法解析、正则表达式完成敏感信息脱敏，并基于题目类型分类标注，构建语义索引库；2) 多阶段评分引擎通过轻量级二分类模型快速筛选无关答案，再调用微调后的 LLM 结合向量数据库检索参考答案语义相似度，分层评估内容完整性、逻辑连贯性等维度，同时触发规则引擎实时校验；3) 反馈生成模块整合 Stanford CoreNLP 提取错题知识点构建 Neo4j 知识图谱，通过 TransE 算法量化薄弱点权重，调用 GPT-4 生成自然语言改进建议；4) 可视化界面采用 React 框架开发，支持教师端实时审核评分依据、学生端查看个性化报告，并通过混淆矩阵对比参考答案差异，增强系统透明性。

核心算法实现需重点攻克“指令微调 + 规则约束强化学习”协同机制：基于历史数据构建结构化 Prompt，结合 PPO 算法设计分层奖励函数，动态调整模型生成倾向；同时，采用轻量化技术优化并发性能，保障千级 QPS 下平均响应时间 ≤1.2 秒/题。测试与迭代优化阶段需在 ASAP 作文集与学科特化数据集上验证指标，通过 A/B 测试对比基线模型，识别误判案例并反向优化规则库与知识图谱。最终，通过 Docker-Compose 实现一键部署，开源核心代码并持续迭代，扩展多模态评分与动态学情追踪功能，推动系统从“自动化评分”向“教学赋能生态”升级。

6. 总结

本研究通过融合大语言模型与教育评估理论，设计并开发了一套高效、可靠、可解释的智能阅卷系统，为解决传统人工阅卷效率低、主观性强及反馈滞后等痛点提供了创新性技术方案。在理论层面，结合 LLM 的深层语义理解能力与教育测评的信度效度原则，提出了基于指令微调与规则约束强化学习的混合评分算法，有效平衡了生成式模型的灵活性与评分标准的规范性；在工程层面，构建了模块化系统架构，涵盖数据预处理、多阶段评分引擎、个性化反馈生成及可视化界面，通过轻量化微调、向量数据

库与规则引擎的协同优化，实现了高并发场景下的实时评分与安全数据管理。实验表明，系统在主观题评分一致性与反馈精准度上显著优于传统自动化方法，尤其在开放题逻辑完整性评估中展现出接近专家水平的判别能力。本研究不仅为教育数字化转型提供了可落地的技术工具，其“数据驱动 + 规则兜底”的设计范式与可解释性模块也为 AI 在教育敏感场景中的伦理化应用提供了参考。未来工作将拓展至多模态评分、动态学情追踪与跨学科自适应评估，进一步推动智能教育系统从“替代人工”向“赋能教学”的范式升级。

参考文献

[1] Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., *et al*. (2024) A Survey on Large Language Model Based Autonomous Agents. *Frontiers of Computer Science*, **18**, Article No. 186345. https://doi.org/10.1007/s11704-024-40231-1

[2] 龙禹辰, 勾智楠, 陈宇欣, 等. 基于大语言模型的多任务生成式重构对话情绪识别[J/OL]. 计算机应用研究, 1-9. https://doi.org/10.19734/j.issn.1001-3695.2024.12.0486, 2025-03-19.

[3] 王路桥, 周洋涛, 李青山, 等. 基于大语言模型的多智能体协作代码评审人推荐[J/OL]. 软件学报, 1-18. https://doi.org/10.13328/j.cnki.jos.007326, 2025-03-19.

[4] 王志鹏, 何铁科, 赵若愚, 等. 大语言模型在代码优化任务中的能力探究及改进方法[J/OL]. 软件学报, 1-24. https://doi.org/10.13328/j.cnki.jos.007325, 2025-03-19.

[5] 祁凯, 周燕生. 基于大语言模型生成内容的负面舆情态势恶化牵引作用研究[J/OL]. 情报杂志, 1-10. http://kns.cnki.net/kcms/detail/61.1167.G3.20250311.1110.002.html, 2025-03-19.

[6] 吴文隆, 尹海莲, 王宁, 等. 大语言模型和知识图谱协同的跨域异质数据查询框架[J]. 计算机研究与发展, 2025, 62(3): 605-619.

[7] 蔡启航, 徐彬, 董晓迪. 利用语义增强提示和结构信息的知识图谱补全模型[J/OL]. 计算机科学, 1-17. http://kns.cnki.net/kcms/detail/50.1075.TP.20241028.1439.034.html, 2025-03-19.

[8] Pan, H., Liu, J., Gong, B., Zhu, Y., Bai, J., Huang, H., *et al*. (2024) Construction and Preliminary Application of Large Language Model for Reservoir Performance Analysis. *Petroleum Exploration and Development*, **51**, 1357-1366. https://doi.org/10.1016/s1876-3804(25)60546-5

[9] Lazebnik, T. and Rosenfeld, A. (2024) Detecting LLM-Assisted Writing in Scientific Communication: Are We There Yet? *Journal of Data and Information Science*, **9**, 4-13. https://doi.org/10.2478/jdis-2024-0020

[10] 张嘉睿, 张豈明, 毕枫林, 等. 基于 IPEX-LLM 的本地轻量化课程教学智能辅助系统[J]. 华东师范大学学报(自然科学版), 2024(5): 162-172.