分类号_____ 编 号_____

U D C_____ 密 级_____

南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

本科生毕业设计（论文）

题　　目：基于大语言模型的毕业设计评分框架研发

姓　　名：王谦益

学　　号：12111003

系　　别：计算机科学与工程系

专　　业：计算机科学与技术

指导教师：刘　江

2025 年 6 月 6 日

诚信承诺书

1.本人郑重承诺所呈交的毕业设计（论文），是在导师的指导下，独立进行研究工作所取得的成果，所有数据、图片资料均真实可靠。

2.除文中已经注明引用的内容外，本论文不包含任何其他人或集体已经发表或撰写过的作品或成果。对本论文的研究作出重要贡献的个人和集体，均已在文中以明确的方式标明。

3.本人承诺在毕业论文（设计）选题和研究内容过程中没有抄袭他人研究成果和伪造相关数据等行为。

4.在毕业论文（设计）中对侵犯任何方面知识产权的行为，由本人承担相应的法律责任。

作者签名：王谦益

2025 年 6 月 6 日

基于大语言模型的毕业设计评分框架研发

王谦益

（计算机科学与工程系　指导教师：刘江 章晓庆）

［摘要］：毕业设计是高校学生综合运用专业知识与技能的学术研究实践，其完成质量直接关联学生毕业资格的认定与学术能力的认证。然而传统的毕业设计评估中，教师评分效率仍有一定提升空间、评估标准主观性强且各有差异、反馈机制较为滞后等问题。近来，大语言模型(Large Language Models, LLMs)凭借强大的知识推理与内容创造能力，已被广泛应用于多种领域，包括自动文本评分(Automated Essay Scoring, AES)。本研究提出了一个基于大语言模型的毕业设计评分框架。首先，基于毕业设计特点与评分需求，设计了涵盖结构完整性、逻辑清晰度、语言流畅性、内容独特与创新性等六个维度的多维度评分标准，并基于此进行提示词(prompt)的设计。其次，本研究还通过少样本提示对模型进行了微调(Fine-tuning)以进一步优化大模型的评分效果。为了验证方法有效性，本研究构建了一个包含 60 篇毕业设计的数据集，并在此数据集上基于通义千问-Plus、通义千问 2.5-14B-1M、通义千问-Turbo、DeepSeek-V3、DeepSeek-R1 进行了实验。实验结果表明，本研究提出的基于大语言模型的毕业设计评分框架有效提高了评分结果的准确性与可靠性。此外，本研究还基于此方法开发了一个毕业设计智能评分系统，可用于辅助教师提升评估效率、促进教学个性化，从而降低人力成本并推动教育公平与技术融合。

［关键词］：大语言模型；毕业设计智能评估；提示词设计

[ABSTRACT] : Graduation designs are academic research practices where college students comprehensively apply their professional knowledge and skills. The quality of its completion directly affects the recognition of graduation qualifications and the certification of academic abilities. However, traditional graduation project evaluations face challenges such as excessive workload for teachers, strong subjectivity in evaluation criteria, and lagging feedback mechanisms. Recently, large language models (LLMs) have been widely applied in various fields, including Automated Essay Scoring(AES), due to their powerful knowledge reasoning and content creation capabilities. This study proposes a graduation project scoring framework based on large language models. First, considering the characteristics and scoring needs of graduation designs, we designed a multi-dimensional scoring standard covering six dimensions: structural integrity, logical clarity, language fluency, content uniqueness, and innovation, and the prompt is designed based on this. Second, this study also fine-tuned the model with few-shot prompts to further optimize the scoring effect of large models. To verify the effectiveness of the method, this study constructed a dataset containing 60 graduation designs and conducted experiments using models such as Tongyi Qwen-Plus, Tongyi Qwen 2.5-14B-1M, Tongyi Qwen Turbo, DeepSeek-V3, and DeepSeek-R1. The experimental results validate that our graduation project scoring framework effectively improves the accuracy and reliability of the scoring results. In addition, this study also developed an intelligent graduation design scoring system based on this method, which can be used to assist teachers to improve the evaluation efficiency and promote teaching personalization, so as to reduce labor costs and promote education equity and technology integration.

[Keywords]: Large language model; Intelligent evaluation of graduation

design; Prompt design

目录

1. 引言

毕业设计是高等教育本科人才培养体系中的核心实践环节，是学生综合运用专业理论知识、技术方法与创新思维解决复杂工程或学术问题的系统性训练过程。作为本科教育的“最后一公里”，毕业设计不仅承载着检验学生知识整合能力、科研素养与实践技能的重要职能，更是衔接校园学习与职业发展的关键桥梁。其质量直接反映高校人才培养水平，对学生学术能力认证、职业素养塑造及终身学习能力发展具有深远影响。

毕业设计论文评分是高等教育教学质量保障体系的核心环节，其结果直接影响学生毕业资格认定及学术能力评价。传统评分模式存在教师评分时间存在缩减余地、评估标准主观性强、反馈机制滞后等瓶颈。随着教育部《教育信息化 2.0 行动计划》的深入实施[1]，人工智能技术与教育场景的深度融合已成为教育现代化改革的重要方向。本研究基于大语言模型技术，构建毕业设计智能评分框架，旨在通过自动化多维评估提升评分效率与公平性，为教育评价改革提供技术支撑。

1.1 研究背景与意义

在高等教育质量保障体系中，毕业设计论文评分作为核心评价环节，其重要性日益凸显。该环节不仅直接关联学生毕业资格的认定与学术能力的认证，更承载着检验学生知识整合能力、科研素养与实践技能的关键职能。然而，随着高等教育普及化进程的加速，传统人工评分模式逐渐暴露出三大核心矛盾：

1. 教学规模扩张毕业设计数量不断增加，教师评分效率仍有一定的提升空间，可以通过使用工具为教师提供辅助，提高教师工作。

2. 统一标准要求与主观评价偏差之间存在矛盾，人工评分易受评价者专业背景、经验认知等因素影响。

3. 传统评分往往滞后于设计过程，难以实现动态指导。

这些矛盾直接制约着评价结果的信效度，甚至引发社会对高等教育质量公平性的关注。

近年来，国家层面促进着 AI+教育的理念的推广。教育部《教育信息化 2.0 行动计划》明确提出“推动人工智能在教学、管理等方面的全流程应用[1]”，《新一代人工智能发展规划》强调“围绕教育、医疗、养老等迫切民生需求，加快人工智能创新应用”，而《深化新时代教育评价改革总体方案》则要求“创新评价工具，利用人工智能、

大数据等现代信息技术，探索开展学生各年级学习情况全过程纵向评价、德智体美劳全要素横向评价”[2]。

LLMs 的突破为破解上述需求提供新路径。其强大的语义理解能力可实现文本内容深度解析，而生成式反馈机制则能提供个性化改进建议。在此条件下，基于 LLM 的评分系统可通过语义分析、多维度指标建模及动态反馈生成，实现从“经验驱动”向“数据-模型双驱动”的评价模式转型，其价值不仅体现为评估效率的量化提升，更在于推动教育评价从“结果判定”向“能力诊断”的范式跃迁。

1.2 研究内容

传统毕业设计评分存在三重困境：

1. 人工评阅模式会消耗教师精力，面对大量的长篇幅毕业设计论文，教师评分所花费的时间依旧能有缩短的余地，可以通过一些途径实现一定程度上提供帮助，从而辅助教师更好的工作。

2. 不同学科对论文的学术规范、方法论要求差异显著，不同教师的评价标准也不尽相同，评分标准多依赖主观经验，导致跨教师、跨学科评估结果可比性不足。

3. 传统评语以总体性建议为主，缺乏对论文逻辑、方法论等深层次问题的精准反馈，学生获得的评价针对性不强，对毕设的改进较小。

因此针对上述需求，本研究提出基于大语言模型的毕业设计论文评估框架，如图 1.1 所示。此架构由学生、教师和基于大语言模型的 GAI 辅助工具三个主要部分组成。在这套框架结构当中，学生首先进行毕业设计实验以及论文撰写工作；然后通过 GAI 辅助工具给自己的论文打分，根据评分以及修改建议不断优化改进自己的毕业设计论文；最后学生提交论文。同时，教师收到学生提交的毕业设计论文之后，可以利用 GAI 辅助工具减轻时间等方面压力，辅助自己完成毕业设计论文评分工作；然后学生会收到教师最终评分结果以及 GAI 辅助工具给出的修改建议，学生将在此基础上进一步完善自己的毕业设计论文。

在此框架中，学生可以从 GAI 辅助工具给出的结果当中收到较为细致的反馈，从而更好地修正自己毕业设计中的缺点与不足；教师可以用过利用 GAI 工具，在一定程度上减少评分所需要的时间，减轻自己的压力。而 GAI 辅助工具不仅能辅助教师完成打分任务，还可以为学生提供各种形式的帮助和服务，如文本分析、语法检查、词汇推荐等，从而帮助学生更好地完成论文撰写工作、提高论文质量。

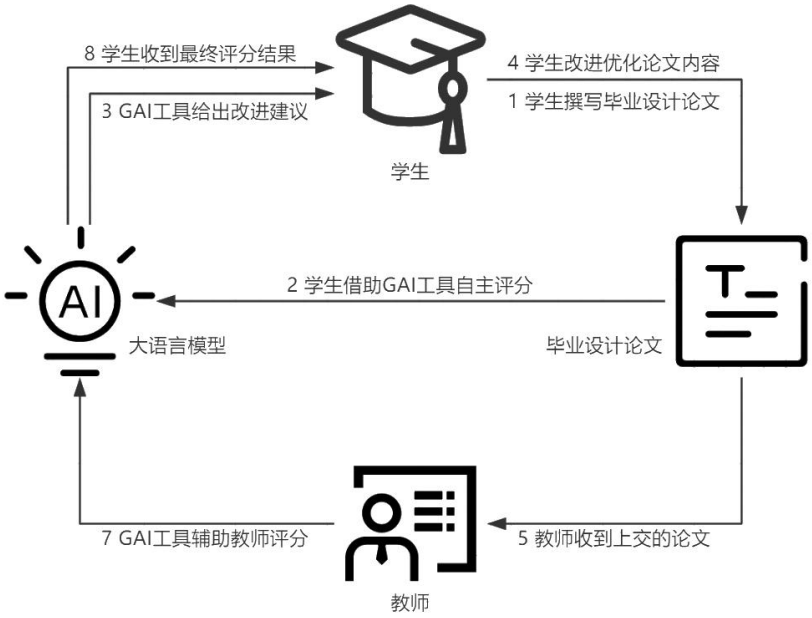


图 1.1 框架结构

此外，本研究还设计了一个毕业设计智能评估系统。通过与系统交互，使用者可以通过粘贴文本或上传文件的方式提交需要评分的毕业设计论文。在用户点击提交按钮之后，前端网页将会把输入内容传输到后端服务器。在将输入处理为合适的格式之后，会通过大语言模型的 API 调用接口访问模型，进行智能评分。最后，后端将模型输出结果处理后按照指定格式返回给前端网页，详细展示给使用者。

1.3 研究难点

本研究的主要目标是研发一个基于大语言模型的毕业设计评分框架，因此使用大语言模型进行论文打分这一环节至关重要，然而却面临诸多难点。

首先是论文输入部分。不同学校、专业的论文格式不尽相同，同时文档的文件格式也会出现差异，因此需要对数据进行预处理。首先引入多模态输入模式，支持 PDF/Word 等格式解析。在数据预处理过程中不仅提取文字部分，还会将图片一并提取、转存，以备不时之需。同时，集成 OCR 模块处理扫描版论文，实现全文内容抽取。

其次问题在于大语言模型评分结果上。智能打分结果不仅需要实现不同指标多维度评分、细致展示原因以及建议，还需要尽可能贴合教师打分结果。然而普通智能作文评估结果评分单一，与人工结果有较明显差异，且并非针对毕业设计论文设计、学术专业性较差。因此本研究设计多维度评分机制，针对性评估毕业设计论文，同时通过少样本学习对模型进行针对性微调，以改进最终的大语言模型打分结果。

最后需要将模型打分结果展示给使用者，需要保障结果简单易读、清晰明了，支持学生、框架、教师之间多方互动。因此本研究设计了智能评估系统，开发可视化评分报告系统，提供雷达图展示多维得分、分点罗列各个维度的评分原因并提供修改建议。

1.4 章节安排

本研究分为以下几个部分：

1. 引言部分。首先阐述毕业设计论文人工评分的困境，国家推动 AI 与教育融合的大背景趋势。而后阐明本研究的具体内容、框架设计和相关应用场景。最后表明本研究的难点以及相应解决方案。

2. 相关工作部分。首先介绍智能作文评估的国内外研究现状，针对典型案例进行分析。而后针对现有大语言模型进行调研、分析，列出各个模型优缺点。

3. 基于大语言模型的毕业设计评分框架部分。首先详细讲解本研究的试验方法和相关研究重点。其次讲解本研究各个部分的实验原理以及设计过程，展示多维度评价指标、提示词设计、模型微调以及智能评估系统。

4. 实验结果及分析部分。首先阐明实验过程中使用的数据来源以及数据集构造过程。其次表明实验设置，即实验中使用的大语言模型类型、进行的具体实验内容。而后，阐述实验效果的评价指标和相关计算过程。最后，展示实验结果并对数据进行分析，得出实验结论。

5. 总结与展望部分。整理本研究的研究内容，表明本研究设计框架的有效性，并表明现有工作的不足与未来展望。

2. 相关工作

2.1 自动化作文评分

当前，自动化作文评分领域已形成两类技术路线：

1. 传统统计模型：以 E-rater、IntelliMetric 为代表，基于语法特征、词汇复杂度等浅层指标建模，但难以捕捉逻辑连贯性等深层维度[3]；

2. 深度学习模型：BERT、GPT 等预训练模型逐步成为主流。例如，ChatGPT 在 AES 任务中实现拼写错误检测准确率 92.3%，但与人类评分者的一致性（皮尔逊相关系数）仅 0.68[4]；Llama 模型通过微调可将一致性提升至 0.75[5]。

多维评分技术是近年研究热点。文献[6]提出基于微调与多元回归的 AEMS 系统，在词汇、语法、连贯性三维度上的评分误差率较单维模型降低 41%。此外，检索增强生成(Retrieval-Augmented Generation, RAG)技术通过整合外部知识库，使 LLMs 在学术规范检测任务中的 F1 值提升 23%[7]。

自动化作文评分（AES）作为自然语言处理和教育技术交叉领域的重要方向，近年来在模型能力、应用场景和评价维度等方面取得了显著进展。总体而言，国外研究在模型精度、多维评分及系统部署方面走在前列，而国内研究正逐步突破技术瓶颈，聚焦中文语境下的模型适配与资源建设。

当前 AES 系统的技术发展主要呈现以下几个趋势：

1. 从特征工程到深度语义理解：早期模型依赖人工设计的词汇、句法等浅层语言特征，而近年来的大多数系统已转向基于 Transformer 架构的预训练语言模型（如 BERT、GPT、RoBERTa 等），使作文评分具备了对上下文语义、逻辑一致性和写作风格的综合理解能力。这种技术跃迁显著提升了评分的准确性与一致性[4]。

2. 从单一总分到多维评分结构：现代写作教学强调写作能力的多维度构成，例如词汇运用、语法掌握、结构完整性、语义连贯性、内容创新等。因此，多维评分模型成为研究热点。相关系统如 AEMS[6]已能在多个维度分别预测作文得分，增强了评分的细致性与诊断价值。

3. 从封闭语料到知识增强生成：传统模型多训练于固定语料，而检索增强生成（RAG）等新型方法支持在生成评分/评语过程中动态引入外部知识库，从而提升作文中知识引用、事实准确性等高阶能力的评估表现[7]。

4. 模型开放化与低资源适应性增强：随着 LLaMA、Vicuna 等开源大模型的发布，评分系统的开发门槛被大大降低。此外，参数高效微调技术如 LoRA、Prompt-tuning 也被引入 AES 任务，使得小样本训练成为可能，特别适合应用于作文题目变换频繁的实际教学场景[5]。

尽管技术不断演进，AES 系统在理论研究与实际部署中仍面临以下主要挑战：

1. 评分公平性与偏差控制问题：大模型虽能学习语言模式，但容易在评分中产生偏差，特别是对文风不常见或内容涉及敏感话题的作文，存在打分不一致或误判现象。此外，评分系统对不同性别、地区、语言水平的学生是否公正，仍是评估系统可信度的关键难题。

2. 跨题迁移能力较弱：目前大多数 AES 系统在特定题目下表现良好，但当作文

题材发生变化时，模型评分准确性显著下降，说明系统对“写作能力”的本质理解尚不充分。这也是 AES 系统大规模部署至学校教学中的关键障碍之一。

3. 评分维度定义模糊与数据稀缺：如何科学划分写作能力维度，并为每一维度建立可靠的人工标注体系，是构建高质量多维评分系统的前提。然而当前尤其在中文环境下，缺乏规模化、多样化、跨题目的标准化评分语料库，导致训练样本不足，模型泛化能力受限。

4. 系统可解释性与教学融合问题：教师与学生对 AES 系统的信任依赖于其评分逻辑的可解释性。但目前大多数深度模型为“黑箱”结构，难以输出评分理由或提供有效反馈。此外，AES 系统如何与课堂教学场景、教师批改流程深度融合，仍需进一步研究与实践探索。

综上所述，自动化作文评分正从语言特征分析迈向认知能力建模的新时代。未来系统将不仅是一个评分器，更应成为支持个性化学习与智能反馈的写作指导平台。为此，需要从模型能力、数据资源、评价标准、教学实践等多个层面开展系统性创新，推动 AES 技术真正走向智能教育的核心应用。

2.1.1 国外研究现状

早期 AES 系统多采用人工特征与回归建模技术。典型代表如 ETS 开发的 e-rater 系统，利用句法、词汇多样性、语法错误数量等浅层语言特征构建评分回归模型，在 TOEFL 等标准化考试中得到广泛应用[3]。此外，Vantage 公司推出的 IntelliMetric 系统，采用贝叶斯模型融合专家评分标准，对写作逻辑与结构进行有限建模，但在捕捉语义一致性和篇章连贯性方面仍存在显著不足[3]。

这类系统的优势在于解释性强、部署成本低，但严重依赖人工特征设计，难以应对开放性写作任务，也难以适应跨领域评分需求。

近年来，随着深度学习特别是预训练语言模型(Pre-trained Language Model, PLMs)的快速发展，AES 系统逐步由“特征驱动”转向“数据驱动”。如 BERT、GPT、RoBERTa 等模型具备强大的语义建模与上下文理解能力，成为构建 AES 系统的新主流。

例如，Liu 等评估了 ChatGPT 在作文评分任务中的表现，指出其在拼写错误检测任务中的准确率达 92.3%，但评分结果与人类评分者之间的一致性（皮尔逊相关系数）仅为 0.68[4]。进一步研究发现，通过在 LLaMA 模型上进行微调，一致性可提升至 0.75，显示出开源大模型在教育任务中的可塑性[5]。

多维评分技术近年来成为国际研究热点。Zhang 等提出自动化作文多维评分系统

(Automatic Essay Multi-dimensional Scoring, AEMS)系统，结合 Fine-tuning 与多元回归，在词汇、语法、连贯性三大维度上分别进行评分，有效克服了传统评分系统仅关注总分的问题。实验表明，该方法的多维评分误差率比单一评分模型降低了 41%[6]。

此外，检索增强生成技术也被引入作文评分任务，用于检测引用准确性、学术规范等评价维度。RAG 通过结合外部知识库对模型生成进行引导，使大模型在知识密集型任务中的 F1 值提升 23%，展现出对逻辑合理性与内容可信度的强化能力[7]。

2.1.2 国内研究现状

我国自动化作文评分研究起步较晚，早期研究以基于规则的系统为主。刘庆双等提出使用词频、句长、词性多样性等语言特征构建中文作文评分系统，采用 SVM 算法进行回归分析，验证了中文作文自动评分的可行性[8]。

受限于中文资源与工具的发展，早期 AES 系统多关注语法错误检测与词汇覆盖率评价，难以建模文本的篇章结构与语义连贯性。

近年来，国内研究者开始将深度学习方法引入 AES 任务。徐鹏等基于 BERT 和层次注意力机制提出了一种结构化评分模型，在中学生作文评分数据集上显著提升了评分相关性和稳定性[9]。王韬等则采用多任务学习框架，同时优化评分预测与作文纠错任务，提高了模型的泛化能力[10]。

此外，中文预训练语言模型（如 ERNIE、MacBERT）在该任务中展现出良好效果，结合微调与个性化标注策略，逐渐形成适应中文写作习惯的评分体系。

目前国内 AES 研究仍面临作文数据资源不足、评分标准不统一等问题。大多数研究基于自建小规模语料库，限制了模型的推广应用。为此，学界已启动多个中文写作能力评估项目，尝试构建多维度、多年级、多题材的作文评分数据集，如“汉语写作能力等级评估语料库”等[11]。

同时，部分研究探索引入专家评语、多轮人工标注机制，构建具有可解释性的 AES 训练与评估体系，提升评分系统在实际教学中的信任度与适用性。

2.2 大语言模型

在人工智能技术飞速发展的当下，语言大模型已成为推动产业智能化转型的核心驱动力。国内外科技企业与研究机构持续投入研发，形成了各具特色的技术生态。为系统梳理当前主流语言模型的技术特性与应用价值，本研究针对国内外代表性产品展开多维对比研究，重点考察其中文处理能力、行业适配性及商业化路径等关键指标。

以下调研结果（见表 2.1）从技术优势、功能局限两个维度展开分析，旨在为不同场景下的模型选型提供参考依据。

表 2.1 国内外语言大模型调研表

分类	模型	优点	缺点
中文大模型	百度文心一言	中文语义理解精准，支持多轮对话；企业版提供私有化部署能力；结合知识图谱进行知识增强，跨模态对话能力强	高阶功能收费较高，免费版调用次数受限；生成文本缺乏情感色彩和人情味
	阿里通义千问	跨领域知识覆盖广，支持复杂逻辑推理；提供行业定制化解决方案；基于大规模参数和高效计算平台，性能稳定	实时数据更新滞后，部分垂直领域专业性不足；处理具体数字和动态事件时易出错
	DeepSeek	支持长上下文，中文优化出色，适合学术研究；使用成本低，兼容性强	生态工具链较新，企业级支持文档和案例较少；网络要求严格，服务器易繁忙
	智谱 AI GLM	开源生态完善，支持中英双语；学术文献分析表现突出；基于 Android 系统的自动化任务执行能力强	企业级服务商业化较晚，社区支持待提升；UI 依赖性强，隐私风险需关注
	腾讯混元大模型	依托腾讯数据资源，中文社交媒体、游戏、娱乐领域覆盖广；多模态理解能力出色，结合图片、视频信息深度生成	学术文献分析专业性不足；企业级服务商业化较晚，社区支持待提升
英文大模型	GPT-4 (OpenAI)	创造力与逻辑能力领先，支持多模态输入；API 生态最完善，应用场景广泛	使用成本高昂，企业合规审查严格；实时数据更新依赖 API 调用
	Llama 3 (Meta)	全系列开源，支持商用；在推理优化和硬件适配方面表现优异，适合定制化需求	企业级服务支持不足，社区贡献质量参差；多语言支持弱于中文优化模型
	Gemini Pro (Google)	多模态融合深度学习，科研文献分析、代码生成效率高；支持 38 种语言，服务全球	移动端部署优化不足，企业级 API 调用限制较多；免费版功能受限
	Claude 3 (Anthropic)	复杂逻辑推理和长文本处理能力优异，支持 200k tokens 超长上下文；安全性设计突出，减少有害内容生成	使用成本较高，企业级 API 调用限制多；多模态融合表现一般
	Mistral Large (Mistral AI)	高效能低成本，适合大规模部署；开源策略灵活，支持企业定制优化	企业级服务支持不足，社区贡献质量参差；多模态融合能力有限

当前主流大语言模型呈现三大技术趋势：

1. 多模态交互能力深度融合，支持文本、图像、音频的联合处理与生成，例如 GPT-4V、Gemini 等模型已实现跨模态内容理解。

2. 长文本处理能力突破百万字级分析阈值，Claude 3 模型凭借 200K token 处理能力显著提升长文档解析效率。

3. 推理优化与硬件适配持续强化，Llama 3 系列通过结构化剪枝技术实现低成本高效部署。

中文大模型方面，百度文心一言依托百度搜索数据资源，在中文语义理解与多轮对话场景中表现优异，支持私有化部署，但高阶功能收费较高且免费版存在调用限制；阿里通义千问凭借跨领域知识覆盖与复杂逻辑推理能力，提供行业定制化解决方案，然而实时数据更新存在滞后性，部分垂直领域专业性需加强；DeepSeek 以长文本处理与学术优化为特色，支持 200k tokens 超长上下文，但生态工具链较新，企业级支持文档不足；智谱 AI GLM 基于开源生态与中英双语支持，在学术文献分析场景表现突出，但商业化进程较晚，社区支持体系尚待完善；腾讯混元大模型结合腾讯生态数据，在社交媒体、游戏领域覆盖广泛，但学术垂直领域专业性存在短板。

英文大模型方面，OpenAI 的 GPT-4 以多模态输入支持与 API 生态完善性占据技术领先地位，但企业合规审查严格且使用成本高昂；Meta 的 Llama 3 通过全系列开源策略实现商用场景覆盖，硬件适配优化显著，但企业级服务支持不足，社区贡献质量参差；Google 的 Gemini Pro 在科研文献分析与代码生成场景效率突出，支持 38 种语言，但移动端部署优化不足，企业 API 调用限制较多；Anthropic 的 Claude 3 以复杂逻辑推理与长文本处理能力见长，安全性设计减少有害内容生成，但企业级服务成本较高；Mistral AI 的 Mistral Large 通过高效能低成本策略适合大规模部署，但多模态融合能力有限。

选型建议方面，中文场景可以优先选择文心一言或通义千问，前者适合对数据安全性要求高的企业，后者适合需要行业定制化解决方案的场景；学术研究场景可选用 DeepSeek，开源生态需求可考虑 GLM。英文场景则技术领先性需求优先选择 GPT-4，成本控制需求可选用 Llama 3 或 Mistral Large，科研场景推荐 Gemini Pro。而面临多语言的情况，需要结合模型的多语言支持能力与垂直领域专业性综合评估，例如 Claude 3 在安全性要求高的场景中更具优势。因此，本研究将选用百度文心一言、阿里通义千问以及 DeepSeek 中的大语言模型进行实验。

3. 基于大语言模型的毕业设计评分框架

3.1 实验方法

如图 3.1 所示，本研究提出了一种基于大语言模型的多维度自动化文本评分系统。该系统融合了自然语言处理、提示工程(Prompt Engineering)、少样本学习(Few-shot Learning)等多项技术，构建了一套从文本预处理到自动评分的闭环系统，具备良好的可扩展性和高度的自动化能力。本系统的核心模块包括：原始文本预处理模块、多维提示构建模块、少样本学习支持模块、大语言模型调用与评分生成模块以及评分结果输出模块。

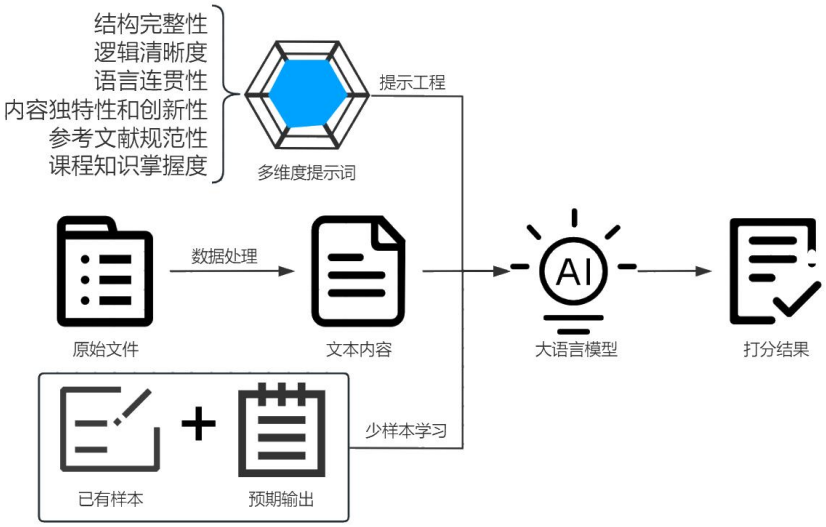


图 3.1 实验方法

在评分流程的最初阶段，系统接收各种格式的原始文本数据（如 Word 文档、PDF、纯文本等），并通过自研的数据解析模块对其进行统一的数据抽取与清洗处理。通过通用文档解析工具（如 pdfplumber, python-docx 等）提取正文内容，将输入文本全部转换为 UTF-8 编码，去除特殊字符与乱码。该模块确保了进入评分流程的文本内容具备高度规范性与一致性，避免由于输入差异而导致模型评分失准的问题，是整个系统的输入保障核心。

为了更精确地引导大语言模型对文本内容进行评分判断，系统首先引入了“提示工程”机制，对评分维度进行细粒度提示设计。图中“多维提示词”即为本模块的产出，具体包含以下维度：结构完整性、逻辑清晰度、语言连贯性、内容独特性与创新性、参考文献规范性以及课程知识掌握程度。通过自然语言构建提示模板，系统能够向模型清晰传达评分任务的具体要求。此模块将传统评分标准结构化地编码进提示词中，以语言指令形式引导模型完成复杂认知性任务，提升了评分的可控性与一致性。

针对大语言模型需要学习评分标准的问题，而后系统引入了少样本学习机制(Few-Shot Learning, FSL)，通过构造“评分示例”作为上下文支持，引导模型进行类比

判断。该模块主要由两部分构成，即已有数据作为学习样本以及依据教师评分结果生成的预期输出，具备较高参考价值。

完成文本预处理与提示构建后，系统将包含原始文本、多维提示词、评分样本的复合提示一并输入至大语言模型中进行推理评估。目前系统可适配主流语言模型如Deepseek、百度文心一言等。模型按提示要求分别给出各个评分维度的评分分值；而后对每一项评分给出简要解释，说明打分依据；最后给出修改意见。该模块是整个系统的决策核心，其性能决定了评分的质量与鲁棒性。

最终，系统将模型返回的打分结果呈现为标准化评分报告，内容包括：最终评分、每个维度的分数、评分所对应的打分理由以及改进建议。综上所述，本研究提出的自动评分系统融合提示工程与少样本机制，构建高维度可控评分框架；设计并实现了模块化、多流程协同的评分系统架构，具备良好的工程可实现性；支持从原始文本解析到最终报告输出的全流程闭环自动评分，提升评分效率；通过多维度指标与解释性输出增强评分透明性与可解释性。

3.2 多维度评分提示词设计

在进行毕业设计论文评分时，直接调用大语言模型往往无法取得理想的评估效果。主要体现在以下几点：（1）评分缺乏统一的标准；（2）评分语言松散无结构，难以提取和对比；（3）缺乏角色意识，模型不理解其职责；（4）难以准确把握评分维度的含义和权重。这些问题表明，想要将大语言模型用于高质量、标准化的论文评分任务，仅凭模型原生能力远远不够，必须辅以精心设计的提示词来引导其行为和输出。

为此，本研究提出一种基于多维度指标体系的提示词设计方案，从评分维度设定、提示语结构设计、角色背景嵌入、术语定义、输出格式规范等多个方面入手，旨在为大语言模型构建一个明确、统一、可执行的评分场景，使其输出更具结构性、可比性与学术性。

3.2.1 多维度评分体系的构建与依据

各个高等院校对于毕业设计论文的评分标准纵然不尽相同，但仍有共通之处，均遵循以下核心原则，以确保评价的科学性与公正性。首先是确保评分标准明确、可量化，以避免主观臆断。例如，中南林业科技大学继续教育学院（2025）的评分细则将“选题意义”“逻辑构建”“专业能力”“学术规范”等维度细化分值，确保评价有据可依。其次强调学术价值与创新性[12]。如有些高校要求优秀论文需“提出新的观点、

方法或技术，或对现有理论进行改进”，并规定创新性评分占总成绩的 30%。同时，论文需体现解决实际问题的能力，并严格遵守学术规范。例如，湖北工业大学（2020）要求毕业设计报告需包含“问题的提出、设计方案的选择与比较、实用性与经济效益评估”等内容，同时强调“论文格式符合要求，中外文用词准确”[13]。

因此，在综合包括南方科技大学、湖北工业大学等众多高校的评分原则后可以发现：毕业设计论文评分标准的制定不仅需要结合学科特点与人才培养目标，还要涵盖学术价值、创新性、实践性及规范性等维度的综合评价。

所以，本研究提出多维度评估方法来实现目标，其中包括以下几个方面，如表 3.1 所示。

表 3.1 多维度评估方法

维度	内容
结构完整	评估论文的总体框架是否合理，内容是否覆盖完整的研究流程，包括引言、相关研究、研究方法、实验与结果、结论等基本结构。要求章节划分清晰、逻辑层次分明
逻辑清晰	考察论文论述过程是否条理清晰、论点是否自洽，是否存在逻辑跳跃、证据缺失、推理不充分等问题。重点分析观点与论据之间的连接是否严密
语言流畅	关注文本语言是否规范，句式是否通顺，用词是否准确，是否存在语法错误或表达歧义。语言应符合学术论文写作规范，避免口语化或模糊表述
内容是否独特与创新	内容是否独特与创新，评价论文是否具有原创性贡献，包括提出新的研究问题、方法、解决方案或理论创新。也包括对已有研究的新视角解读，强调是否具有知识增量
参考文献规范性	检查参考文献的格式是否符合要求，如 APA、GB/T7714 等，同时评估所引用文献的数量、质量与相关性。是否涵盖核心领域文献，引用是否规范、可追溯
课程知识掌握程度	通过论文内容判断学生对相关课程知识（如算法设计、数据分析、工程实践等）的掌握和应用能力。重点是课程知识能否与实践问题有效结合

本研究总共设计了 6 个维度的评估标准，其中前四个维度（结构完整性、 逻辑清晰度、语言流畅性、内容独特与创新性）作为评估报告的主要维度，各自占比 20%；而参考文献和课程内容理解则将其作为次要维度，在评分中占比较小，分别占比 10%。该六维度评分体系不仅有利于覆盖毕业设计论文的评价重点，还可为模型评分提供清晰的指标指引，确保评分结果兼具全面性与针对性。

3.2.2 提示词设计策略与核心组成

优质的大语言模型反馈结果需要好的提示词设计。为了让大模型反馈出更加符合需求的内容，需要设计提示词对大模型提出需求。为了使大语言模型能够准确理解评分任务并输出标准化结果，本研究从提示词工程(Prompt Engineering)的角度出发，参考相关研究经验，根据王东等人在《大语言模型中提示词工程综述》一文中讲述的创建有效提示词的步骤[14]，构建出一套结构化、角色明确的评分提示词模板。

首先，在设计提示词之前，需要明确模型完成的具体任务并理解模型能力[14]。本研究将选用百度文心一言、阿里通义千问以及 DeepSeek 中的大语言模型进行毕业设计论文评分实验，需要让评分结果尽可能贴合教师打分结果，并尽可能较少模型垂直领域专业性不足等问题带来的影响。同时，模型在缺乏角色语境的情况下，往往以聊天或泛评方式回答问题，因此需要进行角色设定以提高评分质量。

因此，依据上述需求本研究完成了初始提示的设计。该模板主要包括角色设定、任务目标说明以及评分指标解释三部分内容。指定模型的身份为“高校教师/教授”或“毕业论文评审专家”，增强其对所处语境的认知，提升模型对任务严肃性与学术性的理解。明确告知模型任务目标是“根据具体评分指标对毕业论文进行评估与打分”，并指出需要依据多维度评分原则进行评分，避免模型生成泛泛评价。

```
你是一位大学教师教授，需要对学生提交的毕业设计论文进行评估。
请评估以下<报告文本/总结报告文本>
在描述<结构完整性>，<逻辑清晰度>， <语言连贯性>， <内容独特性和创新性>， <参考文献规范性>，
<课程知识掌握度>方面的表现，
并根据各指标<占比比例>进行打分与点评，打分范围0-10分。
并最终按照<打分模版>给出学生报告打分结果与评价。打分模板如下：
最终打分：<> (范围0-10分)
1. 结构完整性得分：<>, 占比20%，原因如下：<>
2. 逻辑清晰度得分：<>, 占比20%，原因如下：<>
3. 语言连贯得分：<>, 占比20%，原因如下：<>
4. 内容独特性和创新性得分：<>, 占比20%，原因如下：<>
5. 参考文献规范性得分：<>, 占比10%，原因如下：<>
6. 课程知识掌握度得分：<>, 占比10%，原因如下：<>
请严格按照以下格式返回结果，最终打分一行、6个维度各自一行、修改意见一行，不要擅自添加换行：
最终打分：<> (范围0-10分)
1. 结构完整性得分：<>, 占比20%，原因如下：<>
2. 逻辑清晰度得分：<>, 占比20%，原因如下：<>
3. 语言连贯性得分：<>, 占比20%，原因如下：<>
4. 内容独特性和创新性得分：<>, 占比20%，原因如下：<>
5. 参考文献规范性得分：<>, 占比10%，原因如下：<>
6. 课程知识掌握度得分：<>, 占比10%，原因如下：<>
修改意见：<>
请给以下论文打分：
<论文文本>
```

图 3.2 提示词设计

初步测试提示词效果后发现，模型输出结果较为符合要求，但出现生成内容不规

范或冗余的问题，并没有按照提示词中的格式生成回答。为此，提示词中规定了结构模板，并在语义上限定了输出内容的格式，提高输出质量和可控性。因此，多维度提示词的最终格式设计完成，如图 3.2 所示。

本研究在提示词设计方面的工作，不仅解决了通用大模型用于论文评分时常见的可控性与一致性问题，还具备以下创新性和实用性：

1. 提示词结构全面，覆盖评分全过程：从语境设定到输出控制，全流程引导模型行为。

2. 评分体系标准化，提升可复用性：评分维度设计合理，适用于不同高校与学科，便于推广。

3. 输出格式规范化，利于数据化处理：便于后续与学生反馈系统、成绩系统或 AI 评分系统集成。

3.3 模型微调

Algorithm 1 少样本学习微调过程提示词构造

Require: Integer $count \in [1, 5]$, String $text$

Ensure: Solution String $messages$

Initialize $example\_path, exampl\_txt, example\_result$ ▷ 初始化列表
read all_result from xlsx ▷ 读取所有教师评分结果
for $path$ in $example\_path$ **do** ▷ 遍历文件，获取样本数据
    open $path$ as $file$
    read txt from $file$
    $exampl\_txt$ append txt ▷ 文本存入 $exampl\_txt$
    for $item$ in all_result **do**
        if $item$ is the result of $file$ **then** ▷ 检测结果是否与文件相匹配
            $exampl\_result$ append $item$ ▷ 结果存入 $example\_result$
        end if
    end for
end for
Initialize $prompt$ ▷ 构建原始提示词
$messages = [\{'role': 'system', 'content': prompt\}]$ ▷ 初始化模型输入
for $i = 0$ to $count$ **do** ▷ 依据 count 数量添加少样本学习样例
    $content \leftarrow example\_txt\,[i]\,\&example\_result\,[i]$
    $message$ append $content$
end for
$message$ append $(\{'role': 'user', 'content': f' 请给以下报告打分:\{text\}'\})$
return $message$

图 3.3 少样本学习微调过程提示词构造

少样本学习作为突破传统监督学习对海量标注数据依赖性的关键技术，在资源受限场景下具有重要的理论价值与广阔的应用前景。其核心目标在于利用极少量标注样

本（通常 1-10 个），结合预训练模型蕴含的丰富先验知识，通过特定的学习机制（如迁移学习、元学习或高效的数据增强策略），快速适配新任务并获得良好的泛化能力。与之相比，零样本学习(Zero-Shot Learning)虽然能处理完全未见过的类别，但其通常依赖于外部知识（如语义嵌入、属性描述或知识图谱）进行推理，在特定任务（如主观性强的论文评分）上的性能稳定性和可控性往往不及少样本学习。鉴于本研究任务（本科毕业论文评分）具有明确但主观性强的评价标准，且获取大规模精细标注数据成本高昂，采用少样本学习策略对预训练大语言模型进行针对性微调(Fine-tuning)成为最优选择。 其核心优势在于能够直接利用少量高质量样本，在模型参数空间中优化与评分任务高度相关的区域，从而提升模型在特定评分标准下的理解力、判断力与输出一致性。

尽管当前的大语言模型(LLMs)在精心设计的提示词(Prompt)引导下，能够完成文本生成乃至初步的评分任务，但初步实验表明，其生成的评分结果与专业教师的人工评分之间仍存在显著差距，主要体现在以下方面。首先，模型评分准确性不足，对评分细则理解不够精准，分数偏离教师评分结果。其次，评分一致性存在波动，模型对同一评分标准在不同论文或不同上下文中的把握存在波动，输出稳定性欠佳。

```
[{
    role: system,
    content: 你是一位大学教师教授，需要对学生提交的毕业设计论文进行评估。\n请评估以下<报告文本/总结报告文本>在描述<结构完整性>，<逻辑清晰度>，
             <语言连贯性>， <内容独特性和创新性>， <参考文献规范性>，<课程知识掌握度>方面的表现，并根据各指标<占比比例>进行打分与点评，
             打分范围0-10分。并最终按照<打分模版>给出学生报告打分结果与评价”。打分模板如下：\n最终打分：<> (范围0-10分)\n
             1. 结构完整性得分：<>，占比20%，原因如下：<>\n2. 逻辑清晰度得分：<>，占比20%，原因如下：<>\n
             3. 语言连贯性得分：<>，占比20%，原因如下：<>\n4. 内容独特性和创新性得分：<>，占比20%，原因如下：<>\n
             5. 参考文献规范性得分：<>，占比10%，原因如下：<>\n6. 课程知识掌握度得分：<>，占比10%，原因如下：<>\n
             修改意见：<>\n请严格按照以下格式返回结果，最终打分一行、6个维度各自一行、修改意见一行，不要擅自添加换行：\n
             最终打分：<> (范围0-10分)\n
             1. 结构完整性得分：<>，占比20%，原因如下：<>\n2. 逻辑清晰度得分：<>，占比20%，原因如下：<>\n
             3. 语言连贯性得分：<>，占比20%，原因如下：<>\n4. 内容独特性和创新性得分：<>，占比20%，原因如下：<>\n
             5. 参考文献规范性得分：<>，占比10%，原因如下：<>\n6. 课程知识掌握度得分：<>，占比10%，原因如下：<>\n修改意见：<>\n
}, {
    role: user,
    content: 示例如下：\n
             以下报告:\n分类号 编 号\nU D C 密 级\n本科生毕业设计（论文）\n题 目 ......
             最终打分：8.2 (范围0-10分)\n
             1. 结构完整性得分：8，占比20%\n2. 逻辑清晰度得分：8，占比20%\n3. 语言连贯性得分：8，占比20%\n
             4. 内容独特性和创新性得分：7.5，占比20%\n5. 参考文献规范性得分：10，占比10%\n6. 课程知识掌握度得分：9，占比10%\n
}, {
    role: user,
    content: 请给以下报告打分:\n分类号 编号\nU D C 密级\n本科生毕业设计（论文）\n题 目 ......
}]
```

图 3.4 少样本微调模型 API 调用样例

为了减少上述差距，使模型评分结果更贴合教师的专业标准并具备实际应用价值，本研究使用少样本学习的方式对模型进行微调。 该方案的核心思路是利用有限的高质量教师评分样本作为学习样本，促使模型学习特定评分任务的内在规律和标准，实

现从通用语言能力向专业化评分能力的定向迁移。

为通过微调优化模型评分结果，需要构建高质量微调数据集。首先，精选 2024 届本科毕业论文中的 3-5 篇具有代表性的论文，尽可能涵盖不同得分段、研究方向和写作风格，让模型能学习到更全面的样本。而后，提取论文的完整文字内容，存储在列表 *example_txt* 中。同时，依据教师评分结果提取最终总分、各维度得分等结果，构建样本预期输出，并将这些结构化信息存储在列表 *example_result* 中。具体构建流程如图 3.3 中的伪代码前半部分所示。

少样本学习需要的的样本数据构建完成后，需要进一步完善模型调用的提示词，以实现模型微调的左右。如图 3.3 中的伪代码后半部分所示，在原本的多维度提示词与需打分论文之间添加学习样例。将学习样本与预期输出一一对应，根据微调所需要的样本数量以循环的形式插入提示词当中，具体的模型 API 调用样例如图 3.4 所示。

最后，将带有少样本学习样本的提示词通过调用 API 输入模型，得到模型微调后的打分结果。微调结果取决于样本数量，需要平衡效率以及泛化问题。样本过少会导致模型无法学习鲁棒特征，泛化能力差；反之样本过多则会超出少样本范畴、增加计算成本，同时也可能会有过拟合的潜在隐患。同时调用模型的 API 接口会对输入 token 进行限制，因此本研究在不断测试后，综合各方面考量选择使用 3 个样本进行微调。

3.4 智能评估系统设计

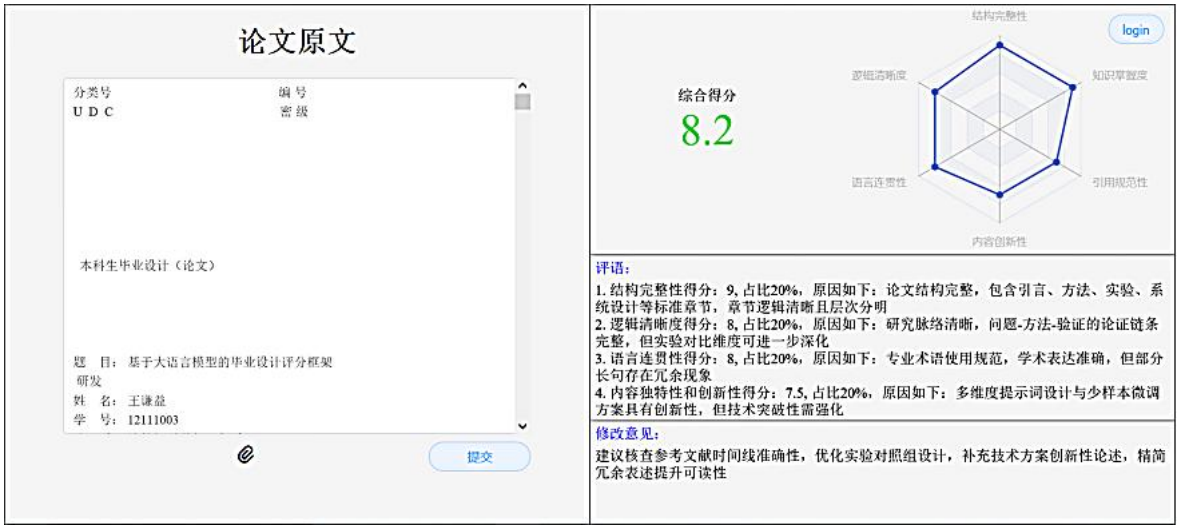


图 3.5 智能评估系统前端设计

如图 3.5 所示，该前端组件采用 Vue 3 框架构建，通过模块化设计实现了智能评估系统的核心交互逻辑，整体架构遵循数据驱动与响应式编程原则。在界面布局上采用双栏对称结构，左侧为输入区域，右侧为结果展示区，通过 flex 布局实现自适应屏

幕尺寸的响应式排版。输入模块集成文本编辑与文件上传双通道，使用 Element Plus 的 el-input 组件构建 18 行高度固定的文本域，配合 v-model 实现报告原文的双向数据绑定；文件上传功能通过 el-upload 组件实现，设置单文件限制与手动上传控制，结合 genFileId 生成唯一标识防止重复提交，上传成功后通过 uploadCondition 状态管理触发后续处理流程。

核心交互逻辑封装在 handleCommit 异步函数中，该函数首先进行空值校验，随后启动 Element Plus 的全屏加载动画，通过 axios 向后端接口发送 POST 请求。在数据可视化层面，采用 ECharts 构建雷达图评分展示，初始化时定义包含结构完整性、逻辑清晰度等 6 个维度的评估指标体系，通过动态配置项实现评分数据的实时更新。为增强用户体验，引入打字机效果逐字渲染评语和修改建议，使用定时器数组管理多个异步写入过程，确保文本动画的流畅执行。

响应式数据管理方面，利用 Vue 的 ref API 创建 reportInput、finalScoreValue 等响应式变量，通过 watch 监听数据变化驱动界面更新。在状态重置逻辑中，提交前会清空历史评分数据，重置雷达图初始状态，并清除所有打字机动画定时器，确保每次提交都是全新的评估流程。组件生命周期管理上，在 onMounted 钩子中完成雷达图实例的初始化，通过 ref 获取 DOM 容器并配置图表基础参数，包括提示框动态定位算法和指标项的最大值设定。

样式设计采用 BEM 命名规范，通过 scoped CSS 实现模块化样式隔离。输入区域使用灰度背景与黑色边框强化视觉层次，按钮组采用 flex 布局实现上传与提交按钮的等比分布，其中提交按钮使用幽灵按钮样式保持界面简洁。评分展示区通过 flex-direction: column 实现垂直分栏，综合评分采用 3em 的鲜绿色字体强化视觉焦点，雷达图容器设置 100%高度确保图表完整显示。评语与建议区域集成 Element Plus 的滚动条组件，通过 v-html 实现富文本内容的动态渲染，保持评估结果格式完整呈现。

后端系统基于 FastAPI 框架构建，采用模块化设计实现论文智能评估的核心业务逻辑，整体架构遵循 RESTful API 设计规范。系统通过 CORS 中间件配置实现了跨域资源共享，允许来自指定前端地址的跨域请求，采用星号通配符开放所有 HTTP 方法和请求头以支持动态交互需求。在数据模型层面，使用 Pydantic 库定义 Report 基类模型，通过类型注解确保接收的论文文本数据符合预期格式，有效保障接口参数校验的严谨性。

核心评估功能由/Committing/端点承载，采用 POST 方法接收前端提交的论文文

本。业务逻辑分为示例准备、模型调用和结果处理三个阶段：首先通过 prepare_prompt 函数构建结构化提示词，该函数动态加载预存的 5 篇示范论文及其评分结果作为少样本学习案例，结合当前待评估文本生成符合大语言模型输入格式的对话历史；随后调用 get_score 函数，通过阿里云百炼平台兼容层访问 OpenAI 接口，使用 deepseek-r1 模型进行推理，过程中实施严格的响应格式化处理，包括去除多余换行符、强制换行符标准化等文本清洗操作，最终返回符合预设模板的评分结果。

文件上传功能通过/Uploading/端点实现，采用异步文件处理机制接收前端上传的论文文档，使用全局变量 contents 暂存文件内容，同时将文件持久化存储至本地 uploaded_files 目录。系统设置完善的错误处理机制，在文件写入异常时返回错误信息，确保服务稳定性。

在模型交互层面，系统构建了精细化的评估指令模板，明确要求模型扮演高校教师角色，从结构完整性、逻辑清晰度等 6 个维度进行专业评估，每个维度包含数值评分、权重占比和文本点评，最终需返回包含综合评分、分项评价和修改建议的结构化结果。通过正则表达式级别的响应格式约束，确保输出结果严格符合前端解析要求。

技术选型方面，采用 FastAPI 框架充分利用其自动数据验证、异步请求处理等特性，结合 uvicorn 服务器实现高性能部署；使用 OpenAI 官方 SDK 确保 API 调用的规范性和安全性，通过配置 base_url 参数实现多云厂商适配；在数据处理流程中，实施示例数据动态加载、响应内容多阶段清洗等策略，有效提升系统的灵活性和鲁棒性。整个后端系统通过清晰的接口定义、严格的数据校验和完善的异常处理，为前端提供稳定可靠的论文评估服务，形成完整的前后端协作闭环。

4. 实验结果及分析

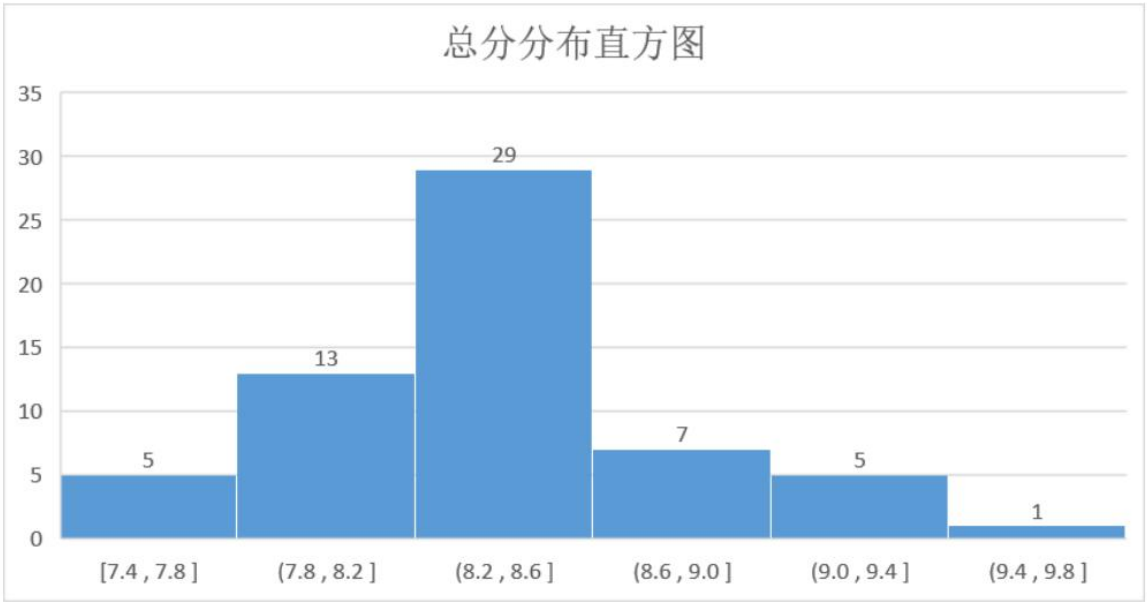
4.1 数据集构建

为了更好地帮助学生改进自己的论文、尽可能为教师评分工作提供帮助，让大模型评分结果更加贴合教师打分结果是必不可缺的。因此需要针对论文以及教师评分结果构造数据集，以方便后续测试以及微调等工作。

为确保数据集质量与评估有效性，本研究选取了南方科技大学 2023、2024 年计算机科学与技术专业本科生毕业论文共 60 篇，其中包括论文 PDF 文件以及教师评分结果，数据构成如图 4.1。其中结构的完整性均分为 8.2 分；逻辑清晰度均分 8.3 分；语言的连贯性均分 8.2 分；内容独特性和创新性均分 8.4 分；参考文献规范性均分 8.8；

课程知识掌握度均分为 9 分，是所有标签中得分最高的；最终总分均分 8.4 分，最高分 9.6 分，最低分 7.4 分。同时，大部分论文处在 7.8 分至 8.6 分，而 8.2 分至 8.6 分有 29 份，是占比最多的区间。

在数据处理时，本研究通过 python 脚本批量化处理样本。代码实现了一个自动化 PDF 内容提取系统，采用模块化设计支持学术论文的批量文本抽取与图像解析，整体架构遵循高内聚低耦合原则。系统核心由文本提取模块、图像处理模块和批处理引擎三部分构成，通过命令行接口触发全流程自动化作业。



(a)

报告结构的完整性 平均分	报告逻辑的清晰度 平均分	报告语言的连贯性 平均分	报告内容的独特性和创新性 平均分	参考文献的规范性 平均分	课程知识的掌握程度 平均分	最终得分 平均分	最终得分 最高分	最终得分 最低分
8.2	8.3	8.2	8.4	8.8	9	8.4	9.6	7.4

(b)

图 4.1 数据构成 (a)教师评分总分分布图 (b)教师评分数据特征

图像提取模块基于 PyMuPDF 库构建，通过 page.get_images(full=True)获取全量图像资源，采用交叉引用编号(xref)机制精确提取每个图像的二进制数据及元信息。系统自动生成包含页码、图像序号和属性信息的标准化文件名，确保图像资源的可追溯性。通过创建独立目录存储图像文件，并在处理完成后自动清空空目录，有效维护了文件系统的整洁性。

批处理引擎采用三层目录遍历机制，支持同时处理多个年度论文集，通过 os.listdir 实现目录遍历，对每个 PDF 文件创建对应的 txt 和 image 子目录。最后通过 pandas 中的 read_excel 方法读取存有教师评分数据的 xlsx 表格文件，构造论文的评分基准，以便用于后续与模型打分结果对比分析。

4.2 实验设置

本次研究进行了多组实验以保证结论的严谨性。首先利用少量数据调研了不同平台不同模型的具体效果，最后综合模型调用、打分结果等多方面因素，选用百炼平台的模型进行后续实验。其中，实验选择的模型包括：通义千问-Plus、通义千问2.5-14B-1M、通义千问-Turbo、DeepSeek-V3、DeepSeek-R1。

在进行大批量实验之前，仍需验证提示词设计的必要性以及合理性。因此，选用了阿里云百联平台上的模型以及全部 60 份毕业设计论文分别进行了无提示词和有提示词的两种不同实验进行测试。

尽管平台提供了高效的文本生成和评分功能，但它们生成的评分结果与教师的人工评分结果之间依然存在明显差距，未能完全符合教师的评分标准。因此，为了更好地契合实际需求，并缩小模型与人工评分之间的差距，需要基于现有的开源大模型，通过少样本训练的方式，针对性地对模型进行精调，以提高其在特定任务中的表现。

4.3 评价标准

实验需要将模型打分结果与教师打分结果进行比对，以衡量实验结果。其中选用平均分，平均绝对误差(Mean Absolute Error, MAE)，均方误差(Mean Squared Error, MSE)以及皮尔逊相关系数(Pearson Correlation Coefficient, PCC)作为评价标准。这些评价指标可根据以下公式来计算：

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{1}$$

$$\bar{t} = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - t_i| \tag{3}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - t_i)^2 \tag{4}$$

$$PCC = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(t_i-\bar{t})}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}\sqrt{\sum_{i=1}^{n}(t_i-\bar{t})^2}} \tag{5}$$

其中 y 为模型打分结果，t 为教师评分结果。平均分是模型对课程项目报告整体质量的直接量化，反映了模型评分倾向。*MAE* 直接反映预测误差的绝对规模，保障对异常值的鲁棒性。*MSE* 强化大误差惩罚，量化了模型评分的波动性和准确性。*PCC* 量化了模型评分与教师评分的线性相关程度，验证模型是否捕捉数据内在模式。

4.4 实验结果

4.4.1 多维度提示词结果分析

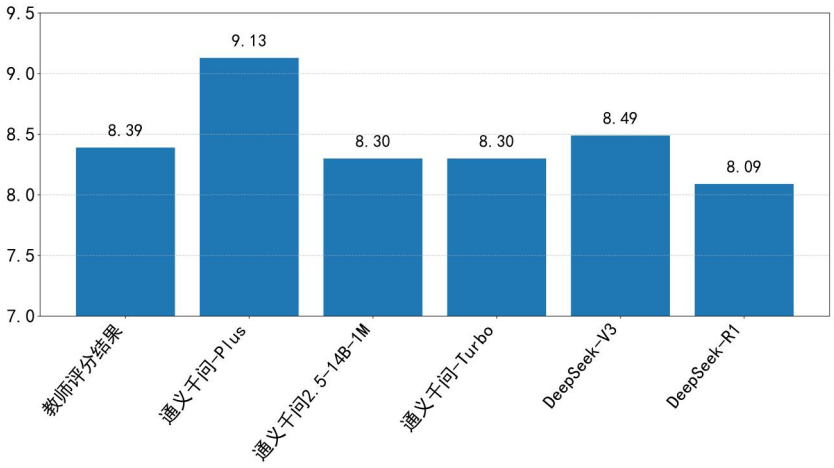


图 4.2 无提示语情况下各个模型最终评分均分

如图 4.2 所示，本研究在无提示词辅助的情况下，使用通义千问-Plus、通义千问2.5-14B-1M、通义千问-Turbo、DeepSeek-V3、DeepSeek-R1 这 5 个模型对 60 份本科毕业设计论文进行评分，并与教师评分结果一起取平均分进行对比。根据图 4.2 可知，通义千问-Plus 与 DeepSeek-V3 的打分结果均分高于教师评分，且通义千问-Plus 的结果（9.13）与教师结果差异显著。同时，通义千问 2.5-14B-1M、通义千问-Turbo 以及 DeepSeek-V3 的模型评估结果均分与教师打分结果极为接近（误差不超过 0.1）。总体而言，无提示语引导下的模型在毕业设计评分中仍有可取之处，但评分结果单一，只能提供整体评分，同时与教师人工评分结果仍有差距。

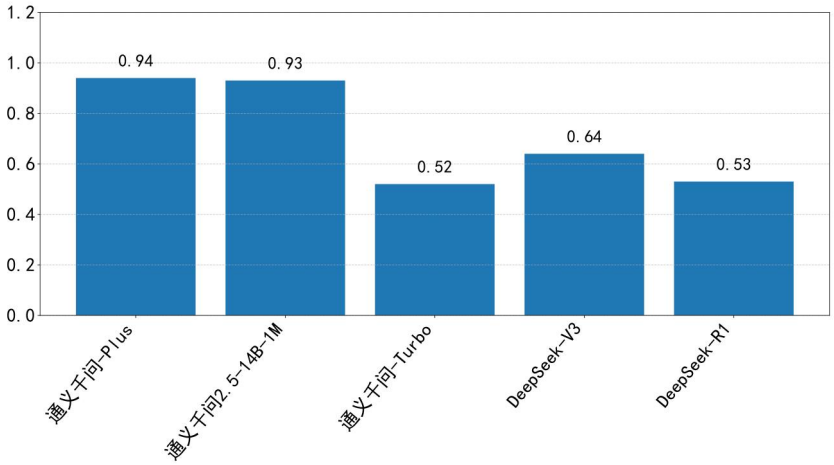


图 4.3 无提示语情况下各个模型最终评分平均绝对误差(MAE)

无提示语引导的情况下，各个模型最终评分与教师评分对比计算出的 MAE、MSE、

PCC 结果分别如图 4.3、图 4.4、图 4.5 所示。依据上述图中数据分析可知，DeepSeek-R1 模型在误差控制与评分趋势一致性方面表现最优，其 MSE（0.49）与 MAE（0.53）均为最低值，表明评分预测值与教师真实值差异最小；同时 PCC 值达 0.36，显著高于其他模型，说明其评分变化趋势与教师评分具有最强正相关性。尽管该模型均分（8.09）略低于教师基准，但其误差控制与趋势跟随能力凸显出稳健的评分一致性。

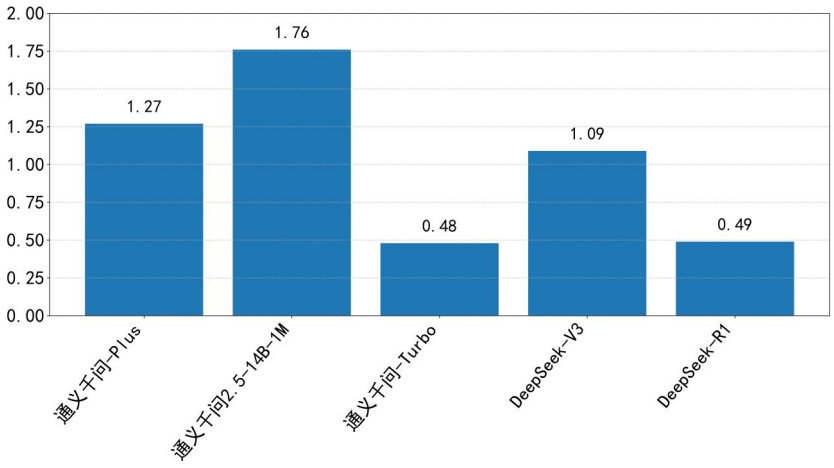


图 4.4 无提示语情况下各个模型最终评分均方误差(MSE)

通义千问-Turbo 模型在均分接近度（8.30）与误差控制（MAE=0.57）方面表现均衡，MSE（0.48）甚至低于 DeepSeek-R1，但 PCC 值（0.17）相对较弱，表明其虽能保持较小评分偏差，但对教师评分趋势的捕捉能力有待提升。DeepSeek-V3 模型均分（8.49）略高于教师基准，MSE（1.09）与 MAE（0.64）处于中等水平，PCC 值（0.12）显示弱正相关，整体表现中规中矩。

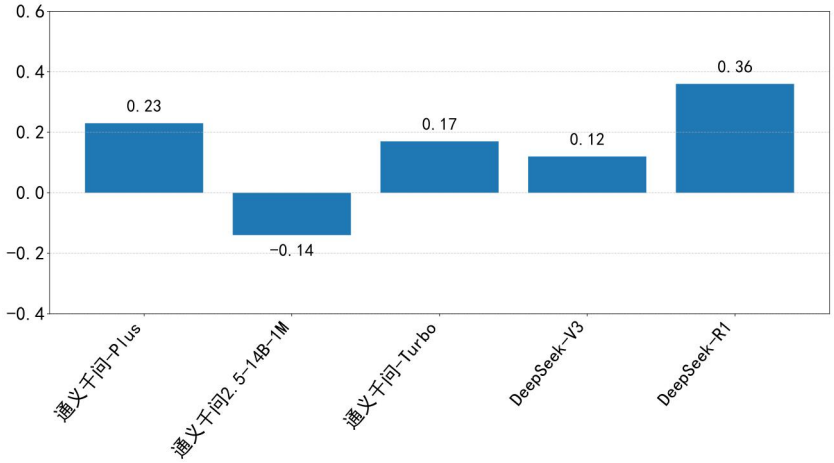


图 4.5 无提示语情况下各个模型最终评分皮尔逊相关系数(PCC)

相较之下，通义千问-Plus 模型存在明显高估倾向（均分 9.13），MSE（1.27）与 MAE（0.94）均居高位，PCC 值（0.23）虽为正值但相关性强度不足，反映出该模型

在评分准确性方面存在系统偏差。最值得关注的是通义千问 2.5-14B-1M 模型，其均分（8.30）虽最接近教师基准，但 MSE（1.76）显著高于其他模型，MAE（0.93）亦处高位，尤其 PCC 值（-0.14）呈现负相关，表明该模型不仅存在较大评分波动，其评分变化趋势甚至与教师基准相悖。

表 4.1 多维度提示词的性能验证总分结果

模型	多维度提示词	平均分	MAE	MSE	PCC
通义千问-Plus	×	9.13	0.94	1.27	0.23
	√	7.96	0.52	0.39	0.41
通义千问 2.5-14B-1M	×	8.30	0.93	1.76	-0.14
	√	7.92	0.61	0.58	0.25
通义千问-Turbo	×	8.30	0.52	0.48	0.17
	√	7.63	0.78	0.85	0.11
DeepSeek-V3	×	8.49	0.64	1.09	0.12
	√	8.43	0.39	0.36	0.04
DeepSeek-R1	×	8.09	0.53	0.49	0.36
	√	7.77	0.70	0.71	0.33

如表 4.1 所示，本研究将无提示词情况下的模型最终打分结果与使用多维度提示词调控后的模型最终打分结果进行对比。参照表 4.1 数据可得，在毕业设计评分框架中引入多维度提示词对大语言模型的评分效果产生了显著影响。通过对比分析发现：在均分维度，无提示词条件下各模型得分普遍高于教师基准（如通义千问-Plus 高 7.4%），而有提示词后模型评分更趋近真实水平（DeepSeek-V3 高于教师基准 1.4%）；在误差指标方面，提示词使通义千问-Plus 的 MSE 从 1.267 降至 0.387，MAE 从 0.944 降至 0.522，表明评分精确度显著提升，但 DeepSeek-R1 的 MSE 从 0.493 升至 0.706，显示提示词对其存在负向调优；相关性分析揭示，通义千问-Plus 的 PCC 从 0.226 提升至 0.406，评分一致性增强，而 DeepSeek-V3 的 PCC 从 0.118 骤降至 0.036，提示词可能改变其评价维度侧重。

综合来看，结构化提示词显著优化了多数模型的评分校准能力，尤其在降低系统误差（MAE 下降 17%-45%）和提升评分效度（PCC 提升 0.08-0.24）方面表现突出，但不同模型架构对提示设计的响应存在差异性，需结合具体模型特性进行提示词优化设计。

表 4.2 提示语调控下各模型打分均分

模型	最终打分	结构完整性	逻辑清晰度	语言连贯性	内容独特性和创新性	参考文献规范性	课程知识掌握度
教师评分结果	8.39	8.20	8.31	8.15	8.36	8.83	9.05
通义千问 Plus	7.96	8.34	7.87	7.62	8.10	7.81	8.64
通义千问 2.5-14B-1M	7.92	8.39	7.69	7.73	8.14	8.99	8.25
通义千问 Turbo	7.63	7.87	7.45	7.72	7.42	7.81	7.70
DeepSeek V3	8.43	8.92	8.87	7.93	8.68	8.22	8.90
DeepSeek R1	7.77	8.26	7.92	7.21	7.68	7.09	8.22

根据表 4.2 数据呈现的多维度评估结果可知，DeepSeek V3 模型展现出了与教师评分高度贴合的评估效果。其最终打分与教师评分差异仅为 0.04，在结构完整性和逻辑清晰度两个核心指标上，该模型仅分别超越了教师评分基准 0.72 和 0.56，显示出其在学术文本组织框架构建和论证逻辑推演方面的突出优势。值得注意的是，尽管在内容创新性维度存在 0.32 的微小差距，但该模型在参考文献规范性和课程知识掌握度等维度仍保持了与教师评分的高度一致性。

相较之下，通义千问系列模型在多个评估维度上呈现出较大的评分偏差。通义千问 Plus 在最终打分、逻辑清晰度和参考文献规范性等指标的评估中，与教师评分存在显著差异。通义千问 2.5-14B-1M 虽然在课程知识掌握度维度表现优于其他版本，但在结构完整性和逻辑清晰度方面的评估仍需优化。通义千问 Turbo 模型在内容创新性和课程知识掌握度两个关键指标上的评分偏差尤为突出。

DeepSeek R1 模型虽然在最终打分和结构完整性等维度接近教师评分，但在内容创新性和课程知识掌握度方面的评估表现相对薄弱。这表明该模型在学术创新点识别和专业知识应用层面的评估能力有待提升。综上所述，DeepSeek V3 模型在毕业设计论文的评估中，其评分结果与教师评分基准具有最高的契合度。

图 4.6 为各提示语调控下各模型打分平均绝对误差。由图 4.6 可知，各模型在最终打分、结构完整性、逻辑清晰度、语言连贯性、内容独特性和创新性、参考文献规范性、课程知识掌握度等维度上展现出不同的评估效果。从平均绝对误差（MAE）值来看，DeepSeek-V3 模型在最终打分（0.39）、结构完整性（0.98）、逻辑清晰度（0.54）和语言连贯性（0.60）等维度上均保持了较低的误差水平，表明其在这些方

面的评估结果与教师评分较为接近。尤其是在最终打分维度上，DeepSeek-V3 的 MAE 值最低，显示出其评分结果与教师评分的高度一致性。

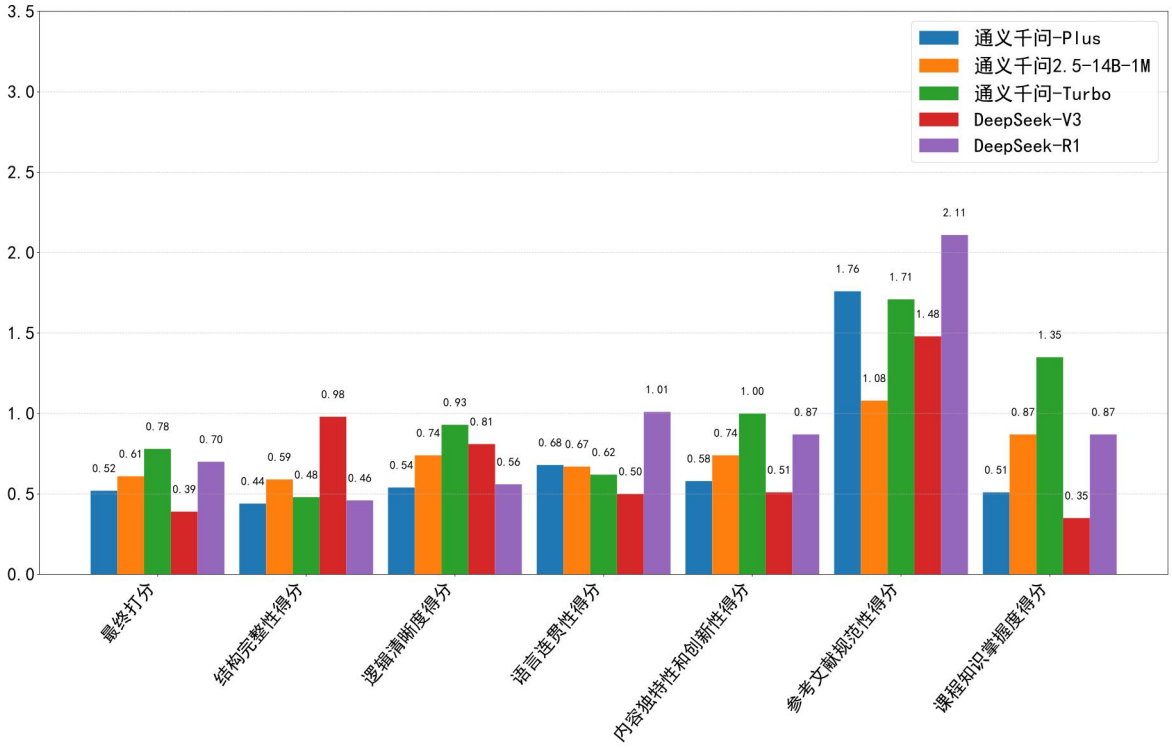


图 4.6 提示语调控下各模型打分平均绝对误差(MAE)

然而，在参考文献规范性（1.76）和课程知识掌握度（0.51）等维度上，DeepSeek-V3 模型的误差相对较大，这表明其在这些方面的评估能力有待提升。相比之下，通义千问系列模型在多个维度上的 MAE 值较高，尤其是在参考文献规范性和课程知识掌握度等维度，显示出较大的评估偏差。例如，通义千问-Turbo 在参考文献规范性维度上的 MAE 值高达 1.00，表明其在此方面的评估结果与教师评分存在显著差异。

DeepSeek-R1 模型在内容独特性和创新性（1.48）以及课程知识掌握度（1.35）等维度上的 MAE 值最高，表明其在这些方面的评估效果相对较差。这可能与该模型在处理创新性和专业知识应用方面的评估任务时存在的局限性有关。

各提示语调控下各模型打分均方误差如图 4.7 所示。由图可知，在最终打分维度上，通义千问-Plus 模型展现了最低的 MSE 值（0.58），表明其在此方面的评估结果与教师评分最为接近。然而，在结构完整性、逻辑清晰度等其他维度上，各模型的 MSE 值存在差异。例如，在逻辑清晰度维度上，通义千问-Turbo 模型的 MSE 值（0.95）低于 DeepSeek-V3（1.24）和 DeepSeek-R1（1.02），表明其在此方面的评估结果与教师评分更为接近。

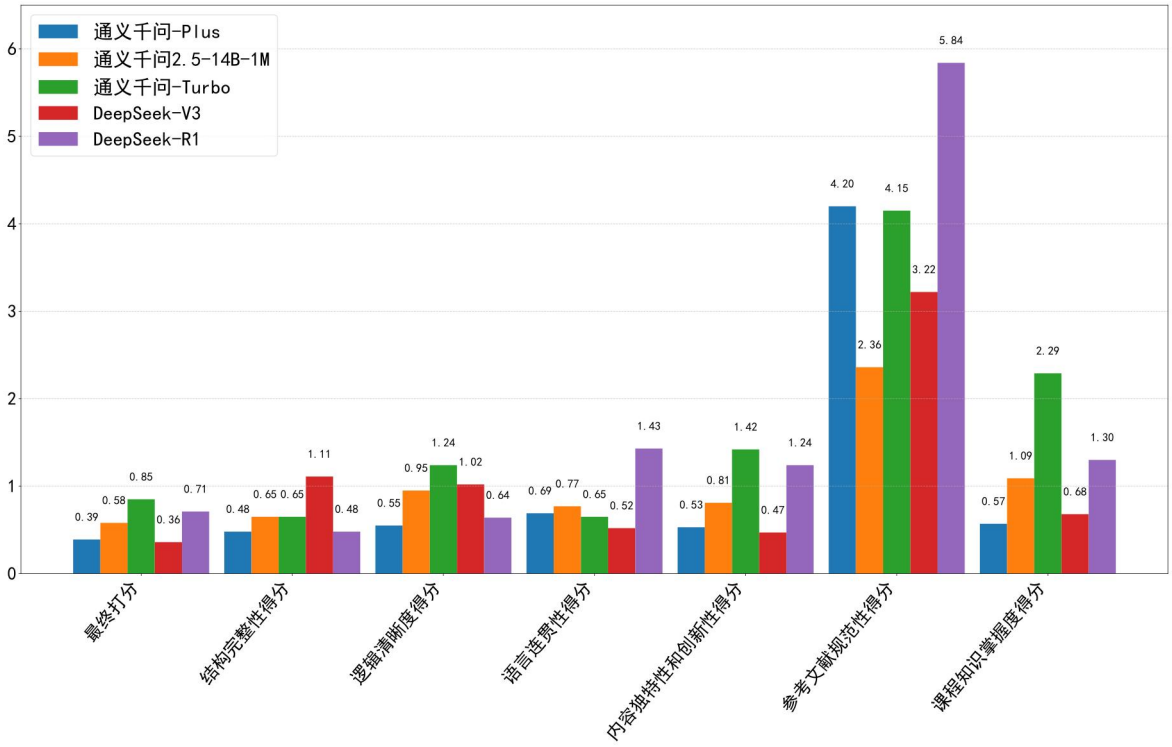


图 4.7 提示语调控下各模型打分均方误差(MSE)

值得注意的是，DeepSeek-V3 模型在多个维度上展现了较高的 MSE 值，尤其是在参考文献规范性（4.20）和课程知识掌握度（5.84）等维度上，这表明其在这些方面的评估结果与教师评分存在显著差异。相比之下，DeepSeek-R1 模型在内容独特性和创新性（1.42）以及课程知识掌握度（2.29）等维度上的 MSE 值也较高，显示出较大的评估偏差。综合各维度上的 MSE 值，通义千问-Plus 模型在整体上展现了较低的评估误差，尤其在最终打分和某些关键维度上表现优异。

图 4.8 为各提示语调控下各模型打分皮尔逊相关系数。从中可以发现通义千问系列模型在多个维度上展现了中等偏上的相关性。其中，通义千问-Plus 在参考文献规范性（0.56）和最终打分（0.41）方面表现尤为突出，显示出其评估结果与教师评分的高度一致性。通义千问-Turbo 则在内容独特性和创新性（0.47）方面展现了最高的相关性，表明其在此方面的评估能力较为接近教师评分。

相比之下，DeepSeek 系列模型在多个维度上的相关性普遍较低。DeepSeek-V3 虽然在内容独特性和创新性（0.44）方面展现了一定的相关性，但在逻辑清晰度（-0.21）和课程知识掌握度（0.02）方面的相关性较低或为负，表明其在此方面的评估结果与教师评分存在分歧。DeepSeek-R1 在多个维度上的相关性均较低，尤其是在课程知识掌握度（-0.23）方面展现了负相关，显示出其评估结果与教师评分的显著差异。

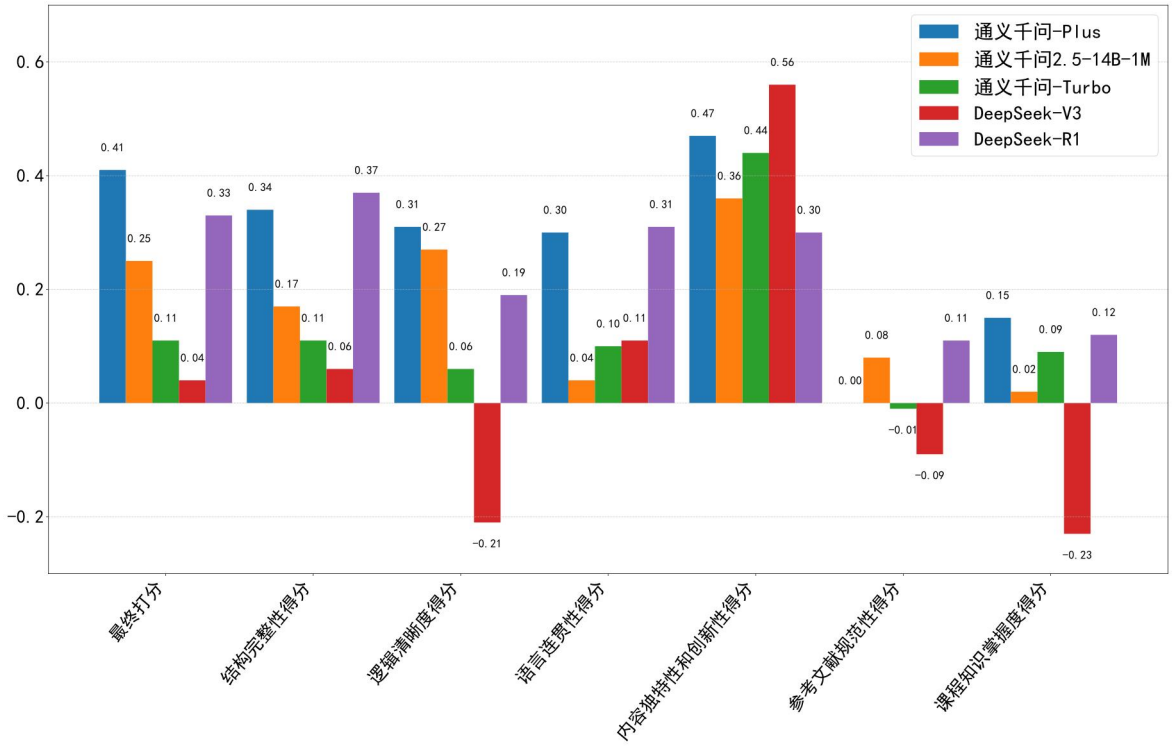


图 4.8 提示语调控下各模型打分皮尔逊相关系数(PCC)

4.4.2 少样本提示语

选用 3 篇 2024 年本科毕业设计论文作为学习样本，以全部 60 篇论文为数据集，针对百炼平台 5 个模型进行试验后，模型打分结果得到提升。表 4.3 为各模型在微调之后给出最终得分与教师结果的比较。

表 4.3 模型微调总分结果

模型	平均分	MAE	MSE	PCC
通义千问-Plus	8.66	0.41	0.25	0.21
通义千问 2.5-14B-1M	8.73	0.42	0.27	0.41
通义千问-Turbo	8.23	0.37	0.21	0.40
DeepSeek-V3	8.51	0.34	0.19	0.29
DeepSeek-R1	8.31	0.30	0.15	0.56

从平均分来看，通义千问 2.5-14B-1M 模型（8.73 分）和 DeepSeek-V3 模型（8.51 分）表现较为突出，显示出较高的整体评估水平。然而，在评估准确性方面，各模型的表现存在差异。DeepSeek-R1 模型在 MAE（0.30）和 MSE（0.15）值上均表现优异，表明其评估结果与教师评分之间的差异较小。同时，DeepSeek-R1 模型在 PCC 值（0.56）上也表现出较高的相关性，进一步验证了其评估结果的可靠性。

相比之下，通义千问-Plus 模型在 PCC 值（0.21）上表现相对较低，表明其评估

结果与教师评分之间的相关性较弱。尽管该模型在平均分（8.66 分）上表现良好，但在评估准确性方面仍有待提升。通义千问-Turbo 模型在 MAE（0.37）和 MSE（0.21）值上表现适中，但在 PCC 值（0.40）上显示出一定的相关性，表明其评估结果与教师评分之间存在一定的关联。

综合各指标来看，DeepSeek-R1 模型在评估准确性和相关性方面均表现出色，显示出其与教师评分的高度贴合度。然而，对于通义千问系列模型而言，尽管在整体评估水平（平均分）上表现良好，但在评估准确性（MAE、MSE）和相关性（PCC）方面仍有提升空间。

表 4.4 模型微调后各模型打分均分

模型	最终打分	结构完整性	逻辑清晰度	语言连贯性	内容独特性和创新性	参考文献规范性	课程知识掌握度
教师评分结果	8.39	8.20	8.31	8.15	8.36	8.83	9.05
通义千问 Plus	8.66	8.34	8.81	8.48	8.51	9.90	9.35
通义千问 2.5-14B-1M	8.73	8.59	8.91	8.54	8.74	9.67	9.12
通义千问 Turbo	8.23	7.75	8.41	7.88	7.82	9.28	8.73
DeepSeek V3	8.51	8.78	8.28	8.03	8.77	9.27	9.02
DeepSeek R1	8.31	8.07	8.10	7.74	8.17	8.85	9.08

表 4.4 为模型微调后各模型打分与教师评分结果的均分对比，其中 DeepSeek-V3 和通义千问 2.5-14B-1M 在贴合教师评分标准方面表现突出。改进后的模型中，DeepSeek-V3 在贴合教师评分标准上表现最为均衡，其核心优势体现在与教师评分高度契合的课程知识掌握度（9.02 vs 教师 9.05）和逻辑清晰度（8.28 vs 教师 8.31），总分差距仅 0.113 分，且通过内容创新性（+0.41）和参考文献规范性（+1.05）的稳健提升实现了均衡优化。通义千问 2.5-14B-1M 虽以总分 8.732 位列第一，但逻辑清晰度（8.91）、结构完整性（8.59）等多项指标显著高于教师标准，存在过度优化风险，更适合强调创新性（8.74）和文献规范（9.67）的场景。通义千问-Plus 和 DeepSeek-R1 分别通过局部强化参考文献规范性（+2.092）和课程知识贴合度（9.08）取得进步，但存在语言连贯性（7.74）或逻辑偏离（8.91）的短板。

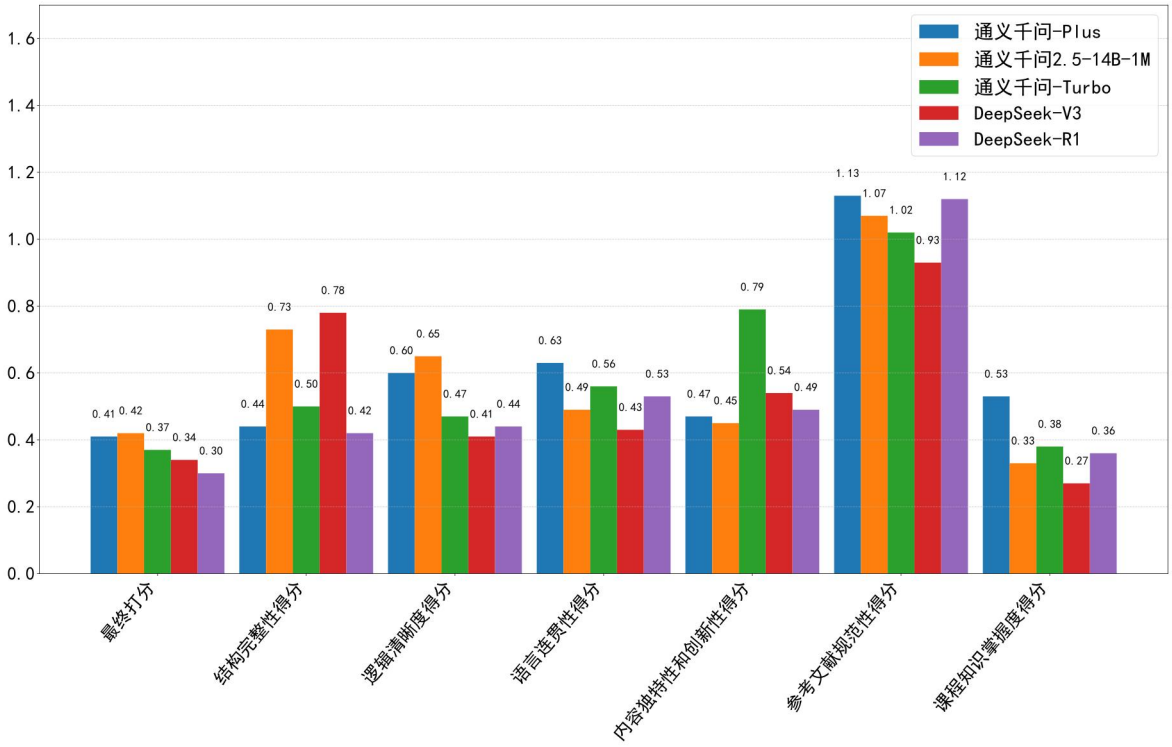


图 4.9 模型微调后各模型打分平均绝对误差(MAE)

使用五个模型进行微调实验后的打分结果与教师结果对比，平均绝对误差如图 4.9 所示。从整体来看，DeepSeek-R1 模型在多个维度上展现了较低的 MAE 值，尤其在课程知识掌握度（0.27）方面，表明其评估结果与教师评分高度接近。此外，DeepSeek-R1 在语言连贯性（0.56）和内容独特性和创新性（0.49）等方面的误差也相对较低，显示出其评估结果的稳定性和准确性。

相比之下，DeepSeek-V3 模型在多个维度上的 MAE 值较高，尤其是在内容独特性和创新性（0.79）和最终打分（0.78）方面，表明其评估结果与教师评分存在一定差异。这可能与该模型在处理创新性和整体打分方面的评估任务时存在的局限性有关。

通义千问系列模型在不同维度上展现了不同的评估效果。通义千问-Plus 在参考文献规范性（1.13）和课程知识掌握度（1.12）方面的误差相对较大，但在其他维度上表现适中。通义千问-2.5-14B-1M 在结构完整性（0.30）和语言连贯性（0.47）方面展现了较低的 MAE 值，显示出较好的评估准确性。通义千问-Turbo 在最终打分（0.34）和课程知识掌握度（0.38）方面展现了较低的 MAE 值，但在结构完整性（0.73）方面的误差相对较大。

图 4.10 为各个模型评分结果与教师结果计算出的均方误差。由图可知，在最终打分维度上，DeepSeek-V3 模型的 MSE 值最低，为 0.25，表明其在整体评估上与教师评分最为接近，评估效果最为贴近。在结构完整性得分维度上，通义千问 2.5-14B-1M

模型的 MSE 值最低，为 0.79，显示出其在评估论文结构完整性方面与教师评分最为一致。在逻辑清晰度得分维度上，通义千问 2.5-14B-1M 模型的 MSE 值最低，为 0.88，表明其在评估论文逻辑清晰度方面与教师评分最为接近。在语言连贯性得分维度上，DeepSeek-R1 模型的 MSE 值最低，为 0.48，显示出其在评估论文语言连贯性方面与教师评分最为一致。在内容独特性和创新性得分维度上，通义千问-Turbo 模型的 MSE 值最低，为 0.34，表明其在评估论文内容独特性和创新性方面与教师评分最为接近。在参考文献规范性得分维度上，DeepSeek-R1 模型的 MSE 值最低，为 0.47，显示出其在评估论文参考文献规范性方面与教师评分最为一致。在课程知识掌握度得分维度上，DeepSeek-R1 模型的 MSE 值最低，为 0.60，表明其在评估学生对课程知识掌握程度方面与教师评分最为接近。

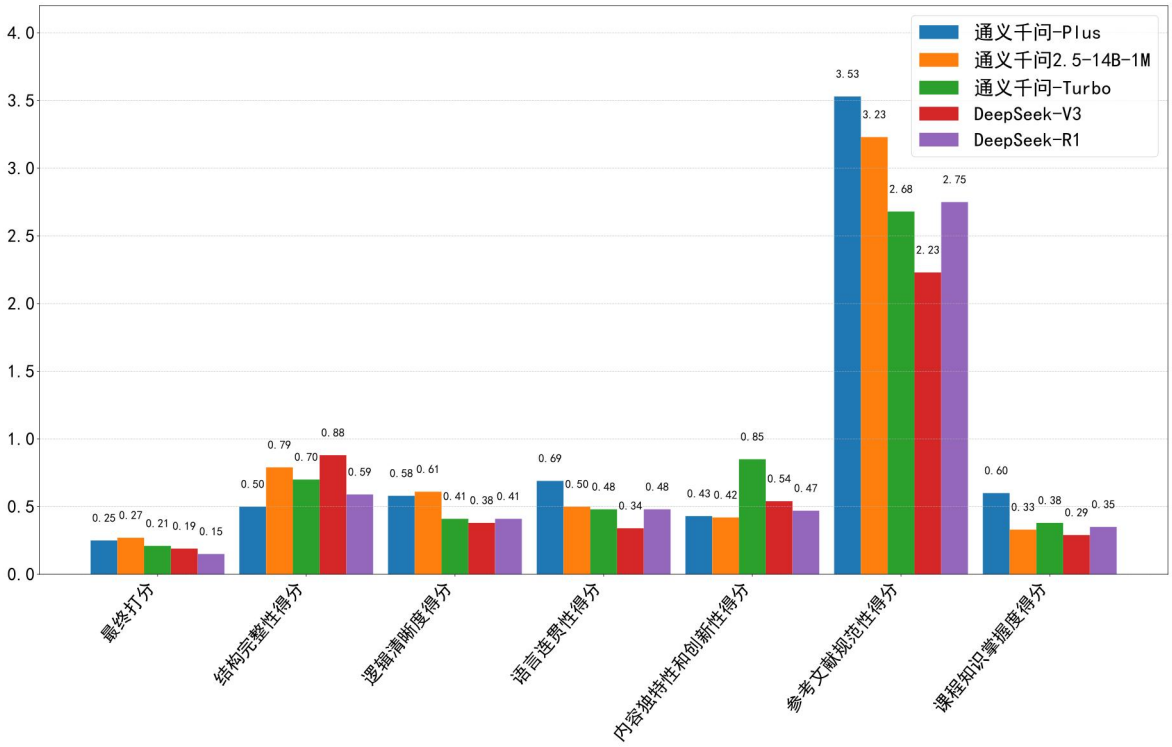


图 4.10 模型微调后各模型打分均方误差(MSE)

综上所述，可以得出以下结论：DeepSeek-R1 模型在语言连贯性得分、参考文献规范性得分和课程知识掌握度得分维度上表现优异，MSE 值最低，表明其在这些方面的评估效果最为贴近教师评分。DeepSeek-V3 模型在最终打分维度上表现最佳，MSE 值最低，显示出其在整体评估上的优势。通义千问 2.5-14B-1M 模型在结构完整性得分和逻辑清晰度得分维度上表现突出，MSE 值最低，表明其在评估论文结构和逻辑方面具有较高的准确性。通义千问-Turbo 模型在内容独特性和创新性得分维度上

表现最佳，MSE 值最低，显示出其在评估论文创新性和独特性方面的优势。

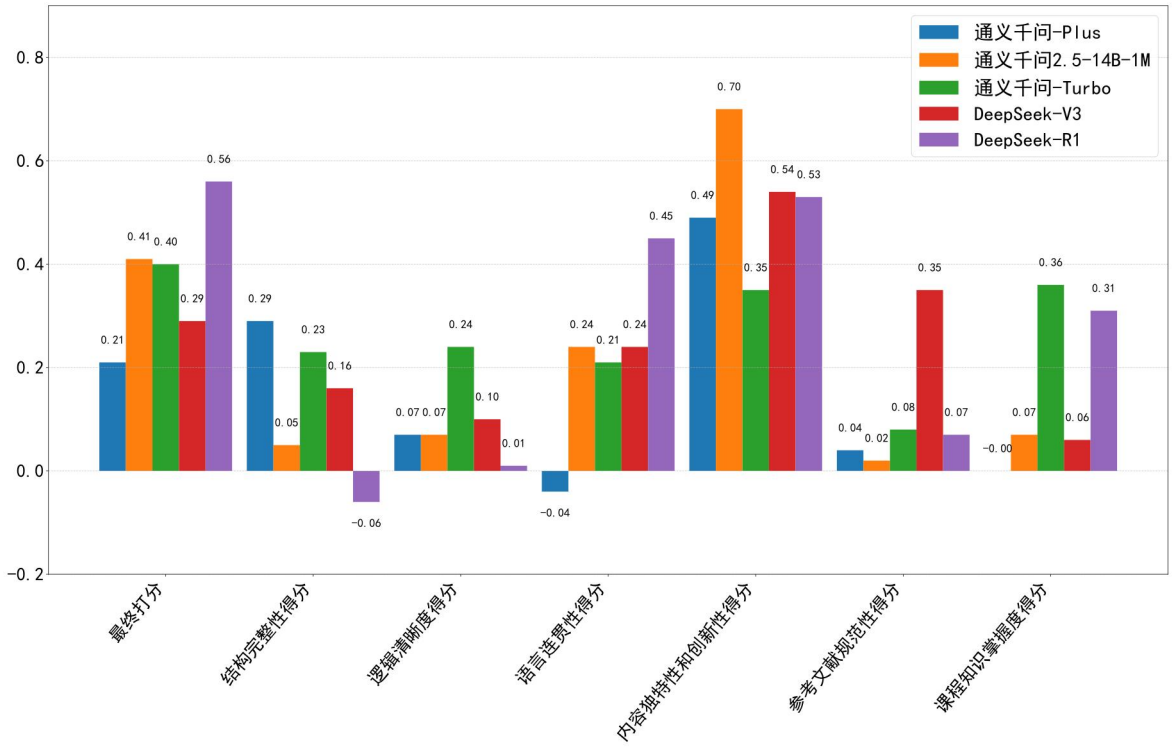


图 4.11 模型微调后各模型打分皮尔逊相关系数(PCC)

微调后各个模型打分结果与教师结果对比，皮尔逊相关系数如图 4.11 所示。从整体来看，DeepSeek-R1 模型在内容独特性和创新性（0.70）方面展现了最高的 PCC 值，表明其在此方面的评估结果与教师评分具有高度相关性。此外，DeepSeek-R1 在课程知识掌握度（0.36）方面的 PCC 值也相对较好，显示出其评估结果的稳定性和准确性。

相比之下，DeepSeek-V3 模型在内容独特性和创新性（0.45）方面也展现了较好的相关性，但在其他维度上的 PCC 值普遍较低，尤其是在逻辑清晰度和课程知识掌握度方面。这可能与该模型在处理创新性和专业知识应用方面的评估任务时存在的局限性有关。

通义千问系列模型在不同维度上展现了不同的评估效果。通义千问-Plus 在参考文献规范性（0.49）方面展现了最高的 PCC 值，表明其在此方面的评估结果与教师评分具有较强的相关性。然而，该模型在其他维度上的 PCC 值普遍较低。通义千问-2.5-14B-1M 在多个维度上的 PCC 值较低，尤其是在逻辑清晰度和课程知识掌握度方面展现了负相关。通义千问-Turbo 在最终打分和课程知识掌握度方面的 PCC 值与教师评分具有一定的相关性，但在其他维度上的表现相对较弱。综合各维度上的 PCC 值，DeepSeek-R1 模型在内容独特性和创新性方面展现了与教师评分最高的相关性，

显示出其评估结果的可靠性和准确性。

总的来说，本次精调训练成功提升了模型的评分效果，为后续的模型优化和实际应用打下了坚实的基础。未来将在更多数据和更强模型的支持下，进一步改善现有结果，并实现更加精准的自动评分系统。

4.4.3 实验结果总结

本研究旨在应对传统毕业设计评分模式中教师工作负荷过、评估标准主观性强、反馈机制滞后等挑战。基于减轻教师压力和为学生提供细致评价的需求，研究提出了使用大语言模型进行毕业设计评分的系统框架。为了让评分多元化、打分结果更加符合需求，本研究提出了多维度评分标准，包括结构完整性、逻辑清晰度、语言流畅性、内容独特与创新性等六个维度。同时为了进一步优化大模型的评分效果，贴近教师评分结果，研究还使用少样本学习的方式对模型进行了微调优化。

实验基于南方科技大学 2023、2024 年计算机科学与技术专业本科生的 60 篇毕业论文构建了数据集。研究使用了多个大语言模型进行实验，包括通义千问-Plus、通义千问 2.5-14B-1M、通义千问-Turbo、DeepSeek-V3、DeepSeek-R1 等，以验证所提出评分框架的有效性。

实验结果表明，本研究提出的基于大语言模型的毕业设计评分框架有效提高了评分结果的准确性与可靠性。提示词辅助与模型微调后，在平均分、MAE、MSE 和 PCC 等评价指标上普遍优于纯 LLM。微调显著提高了模型的评分准确性和可靠性，减少了预测误差，并增强了模型预测分与教师打分之间的线性相关程度。

5. 总结与展望

本研究介绍了一种基于大语言模型的毕业设计论文智能评分系统。其中提出了利用 GAI 辅助工具提升教师评价和学生毕业设计论文质量的框架结构，并通过多维度评估设计大模型的提示词进行实验测试。通过设计合适的提示词，大语言模型能够根据这些维度对写作内容进行分析，并生成综合评分和详细的评估结果，但直接将原文输入给大语言模型打分的结果和教师打分有一定差距。为了优化模型对文本的理解和打分效果，尝试使用少样本模型微调的方式。实验结果表明，精调后的模型在评分一致性和准确性上得到了显著提升。

未来将借鉴 RAG 技术，通过检索增强以及模型微调相结合等方式继续改进系统，提升评估效果。计划进一步完善评估标准，使其更加全面和准确，并能够评估学生的

写作风格、情感表达等方面。此外还可以分析具体得分和学生属性的相关性，提供更详细的评估结果解释，帮助学生理解评估结果，并进行针对性的改进。并且还会进一步完善网页系统，使其能够支持多文件并发请求，同时添加自定义提示词等其他功能设计。

参考文献

[1] 教育部. 教育信息化 2.0 行动计划[Z]. 2018.

[2] 国务院. 深化新时代教育评价改革总体方案[Z]. 2020.

[3] Page E B. Project Essay Grade: PEG[J]. Automated Essay Scoring: A Cross-disciplinary Perspective, 2003: 43-54.

[4] Liu V, et al. Are Large Language Models Good Essay Graders?[J]. arXiv preprint arXiv:2304.01652, 2023.

[5] Chiang W L, et al. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality[J]. arXiv preprint arXiv:2304.11264, 2023.

[6] Zhang Y, et al. Automatic Essay Multi-dimensional Scoring with Fine-tuning and Multiple Regression[C]//Proceedings of the 14th International Conference on Educational Data Mining. 2021: 612–617.

[7] Lewis P, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459–9474.

[8] 刘庆双, 李德毅, 关宏志. 中文作文自动评分的研究与实现[J]. 中文信息学报, 2009, 23(6): 110-115.

[9] 徐鹏, 郝鹏程, 王一凡, 等. 基于 BERT 和层次注意力机制的中文作文自动评分方法[J]. 小型微型计算机系统, 2020, 41(6): 1143-1148.

[10] 王韬, 黄乾, 罗文. 基于多任务学习的中文作文评分研究[J]. 中文信息学报, 2021, 35(2): 129-136.

[11] 陈婷, 刘宇翔, 吕洪波. 中文作文自动评分研究综述[J]. 现代教育技术, 2021, 31(1): 45-52.

[12] 中南林业科技大学继续教育学院. 毕业论文评分标准及细则[EB/OL]. (2025-01-15).

[13] 湖北工业大学. 毕业设计文件规范及评分标准[EB/OL]. (2020-06-14).

[14] 王东清, 芦飞, 张炳会, 等. 大语言模型中提示词工程综述[EB/OL]. (2025-01-01).

[15] 李华, 等. 高校毕业设计管理模式创新研究[J]. 中国高教研究, 2021(5): 45–50.

[16] 王明. 基于评分者信度的论文质量评估研究[J]. 现代教育技术, 2020, 30(8): 76-82.

[17] Vinyals O, et al. Matching Networks for One Shot Learning[C]//Advances in Neural Information Processing Systems. 2016: 3630-3638.

[18] Hong Y, et al. F2GAN: Fusing-and-Filling GAN for Few-shot Image Generation[C]//Proceedings

of the 28th ACM International Conference on Multimedia. 2020: 2842-2850.

[19] Ojha U K, et al. Few-shot Image Generation via Cross-domain Correspondence[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10743-10752.

[20] 教育部高等学校教学指导委员会. 普通高等学校本科专业类教学质量国家标准[M]. 北京: 高等教育出版社, 2018.

[21] 谢幼如, 等. 课堂教学设计[M]. 北京: 电子工业出版社, 2021.

[22] 徐辉. 高等教育评价的理论与实践[M]. 北京: 高等教育出版社, 2019.

[23] 贾积有, 王光迪. 应用大语言模型快速有效分析教育访谈文本[J]. 中国远程教育, 2023(12): 34-42.

[24] 王雅青, 等. Automated Evaluation of Personalized Text Generation using Large Language Models[C]//Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. 2023: 1-12.

[25] Gao L, et al. The Pile: An 800GB Dataset of Diverse Text for Language Modeling[J]. arXiv preprint arXiv:2101.00027, 2020.

[26] Brown T, et al. Language Models are Few-Shot Learners[C]//Advances in Neural Information Processing Systems. 2020: 1877-1901.

[27] 教育部高等教育司. 普通高等学校本科专业设置管理规定[Z]. 2012.

[28] 王雨磊. 学术论文写作与发表指引[M]. 北京: 文化发展出版社, 2020.

[29] 廖帆, 肖扬生, 周弘颖. 应用文写作[M]. 北京: 人民邮电出版社, 2029.

致谢

在本次毕业设计完成之际，我想感谢大学四年间给予我指导、帮助的学校、老师、同学、朋友和家人。

感谢南方科技大学提供的优良环境。南科大开设了多种多样的教学课程，让我能丰富自己的学识。众多图书馆提供了良好的教学环境，让我能不断精进自我。同时，南科大的全英文授课机制为我提供了良好的英语环境，让我受益匪浅。

感谢计算机系对我的栽培。覆盖面极广的专业课程为我提供了扎实的基础，多种多样的选修课更加丰富了我的见识。随着计算机相关知识的不断学习，我逐渐从为了完成作业而编程转变为因为兴趣而编程。每次搭建程序、跑通任务的过程都不断完善着我的知识储备。尽管 debug 的过程算不上幸福，但始终令我收获良多。

感谢老师和学长学姐对我的帮助。特别感谢我的导师刘江老师、章晓庆老师。在实验进行的时候，两位老师都在百忙之中抽时间为我答疑解惑，帮助我解决了很多实验中遇到的问题与难点。不仅如此，学长学姐对我的帮助也不可或缺。他们经常无私的帮我解决在学习过程中遇到问题。尤其要感觉孙清扬学姐，无论我何时有问题，学姐总会及时给我提供帮助。本研究的完成离不开学姐的指导与鼓励。除此之外，也要感谢学校的各位任课教师，是他们上课认真负责，才让能我不断学习到新的知识，不断进步。

感谢我的同学、朋友以及家人。是你们的陪伴让我大学四年变得格外精彩。无论是我的舍友还是其他朋友，每当我遇到困难，总有你们的身影陪伴着我。面对疾病是你们照顾我，面对沮丧是你们宽慰我，面度困境是你们帮助我。感谢大家对我的包容，给我的每一天都带来了快乐和成长。

感谢所有人对我提供的帮助、给予我的陪伴，是你们的付出造就了现在的我，感谢大家。