分类号 编 号

U D C 密 级



**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**

**院 系：** **计算机科学与工程系**

**专 业：** **计算机科学与技术**

“姓名、学号、指导教师、年级与专业、年月日”均用四号宋体打印，不得手写，各栏目下划线需统一长度

**指导教师：** **刘江教授**

2025 年 月 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：

2025 年 月 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存在教师工作负荷过重、评估标准主观性强、反馈机制滞后等挑战。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的设计。其次，本研究还通过少样本提示对模型进行了微调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；提示设计

[ABSTRACT] :Graduation projects 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 projects, we designed a multi-dimensional scoring standard covering six dimensions: structural integrity, logical clarity, language fluency, content uniqueness, and innovation. Based on this, we designed the prompt. 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 projects 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; Graduation projects; Intelligent assessment; Prompt design

目录

**[1. 引言 1](#_Toc23482)**

[1.1. 研究背景与意义 1](#_Toc14693)

[1.2. 相关工作 2](#_Toc27072)

[1.2.1. 自动化作文评分 2](#_Toc29776)

[1.2.2. 大语言模型 2](#_Toc23211)

[1.3. 研究内容 3](#_Toc19111)

[1.4. 章节安排 4](#_Toc12796)

**[2. 基于大语言模型的毕业设计评分框架 4](#_Toc13524)**

[2.1. 框架结构 4](#_Toc9652)

[2.2. 多维度评分提示词设计 6](#_Toc15118)

[2.3. 模型微调 8](#_Toc1666)

**[3. 实验结果及分析 9](#_Toc30544)**

[3.1. 数据集构建 10](#_Toc2646)

[3.2. 实验设置 11](#_Toc17446)

[3.3. 评价标准 11](#_Toc5245)

[3.4. 实验结果 12](#_Toc26321)

[3.4.1. 多维度提示词的性能验证 12](#_Toc12216)

[3.4.2. 少样本提示语 17](#_Toc14220)

[3.4.3. 实验结果总结 20](#_Toc20968)

**[4. 智能评估系统设计 21](#_Toc8441)**

**[5. 总结与展望 24](#_Toc32329)**

**[参考文献 25](#_Toc22529)**

**[致谢 27](#_Toc16977)**

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。传统评分模式面临教师工作负荷过载、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(Large Language Models, LLMs)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. 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：一是教学规模扩张与教师评价负荷之间的矛盾，高校扩招导致毕业设计数量激增，而教师评价精力有限；二是统一标准要求与主观评价偏差之间的矛盾，人工评分易受评价者专业背景、经验认知等因素影响；三是过程性评价需求与结果导向反馈之间的矛盾，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](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD/9180?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/_blank)、[大数据](https://baike.baidu.com/item/%E5%A4%A7%E6%95%B0%E6%8D%AE/1356941?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/_blank)等[现代信息技术](https://baike.baidu.com/item/%E7%8E%B0%E4%BB%A3%E4%BF%A1%E6%81%AF%E6%8A%80%E6%9C%AF/2426261?fromModule=lemma_inlink" \t "https://baike.baidu.com/item/_blank)，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. 相关工作
     1. 自动化作文评分

当前，自动化作文评分(Automated Essay Scoring, AES)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(Fine-tuning)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(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]。然而，现有研究多聚焦于通用作文场景，针对毕业设计论文的学术性、创新性等差异化评估需求仍缺乏系统性解决方案。

* + 1. 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1.1）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。

当前主流大语言模型呈现三大技术趋势：其一，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；其二，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；其三，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。

**表1.1 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 | 高阶功能收费较高，免费版存在调用次数限制 |
| 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 | 实时数据更新滞后，部分垂直领域专业性需加强 |
| DeepSeek | 支持长上下文，中文优化出色，适合学术和研究场景 | 生态工具链较新，企业级支持文档和案例较少 |
| 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 | 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 | 使用成本高昂，企业合规审查严格 |
| 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 | 企业级服务支持不足，社区贡献质量参差 |
| Gemini Pro  (Google) | 多模态融合深度学习，在科研文献分析、代码生成场景效率高 | 移动端部署优化不足，企业级API调用限制较多 |

* 1. 研究内容

传统毕业设计评分存在三重困境：其一，人工评阅模式受限于教师精力，面对大量长篇幅的毕业设计论文，教师评分时间较为紧张，难以应对大规模学生群体的差异化评价需求；其二，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；其三，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。

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，基于多维度评分标准进行提示词设计，并对模型微调，使评分结果不断接近教师打分。此外，还研发了一个毕业设计智能评估系统。帮助学生不断优化毕业设计、提高自身写作能力；辅助教师进行评估工作，在一定程度上减少评估所需时间，尽可能减轻教师压力。

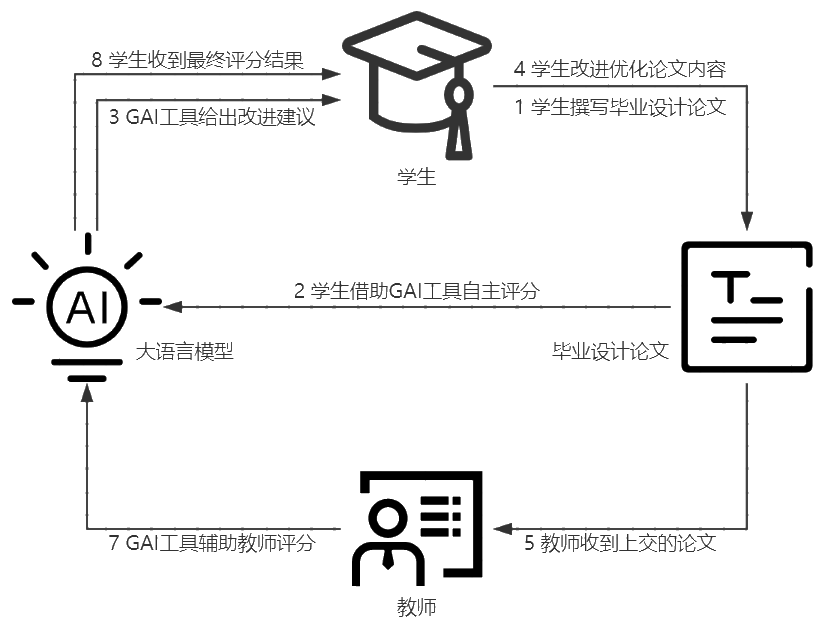
* 1. 章节安排

本研究分为以下几个部分：

1. 引言部分，介绍AI+教育的大环境背景，智能作文评估的国内外研究现状，毕业设计论文智能评估的难点，引出本研究的框架设计：基于大语言模型进行提示词设计、模型微调，通过多维度打分贴合教师评分结果，帮助学生自我提升、辅助教师评分。
2. 基于大语言模型的毕业设计评分框架部分，描述具体的评分框架结构，展示多维度评价指标以及提示词设计，展示模型微调过程。
3. 实验结果及分析部分，首先展示数据集构造，阐明实验过程中使用的数据以及数据集构造过程。而后表明实验设置、实验结果并对数据进行分析，得出实验结论。
4. 智能评估系统设计部分，展示针对用户设计的智能评估系统详情以及设计过程，突出本研究框架与用户的交互性。
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。
6. 基于大语言模型的毕业设计评分框架
   1. 框架结构

为了实现辅助教师评分，帮助学生改进论文内容的目的，设计了如图2.1所示的基础框架。

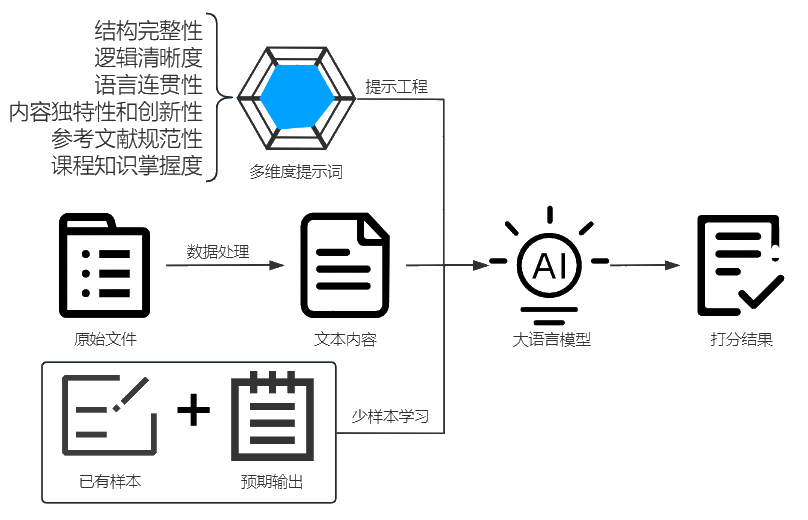
此架构由学生、教师和基于大语言模型的GAI辅助工具三个主要部分组成。在这套框架结构当中，学生首先进行毕业设计实验以及论文撰写工作；然后通过GAI辅助工具给自己的论文打分，根据评分以及修改建议不断优化改进自己的毕业设计论文；最后学生提交论文。同时，教师收到学生提交的毕业设计论文之后，可以利用GAI辅助工具减轻时间等方面压力，辅助自己完成毕业设计论文评分工作；然后学生会收到教师最终评分结果以及GAI辅助工具给出的修改建议，学生将在此基础上进一步完善自己的毕业设计论文。



**图2.1：框架结构**

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。

如图2.2所示，在大语言模型打分之前，会先对原始文件进行数据处理，只保留文字部分。而后通过多维度评分实现通过提示词调控模型的目的。同时，会结合已有样本以及预期输出，通过少样本学习对模型进行微调，以获得更好的打分结果。



**图2.2：实验方法**

* 1. 多维度评分提示词设计

在进行毕业设计论文评分的时候，大语言模型本身返回的结果往往不是十分理想，评分时没有固定的标准、输出结果也没有遵循一定的模版，问题较为严重。为了优化模型效果，需要为大语言模型提供一套合理的评估方案以及详细的打分标准。本实验的目的是让大语言模型基于评分方案，准对毕业设计论文给出详细且全面的评估结果，得到综合而又客观的反馈。而多维度评分方法就是一种可采取的综合性评价方案，可以通过对多个维度的考量来全面地衡量某一对象或现象的特征和价值。

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[8]。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”[9]。因此，综合各个高校的评分原则不难发现，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。

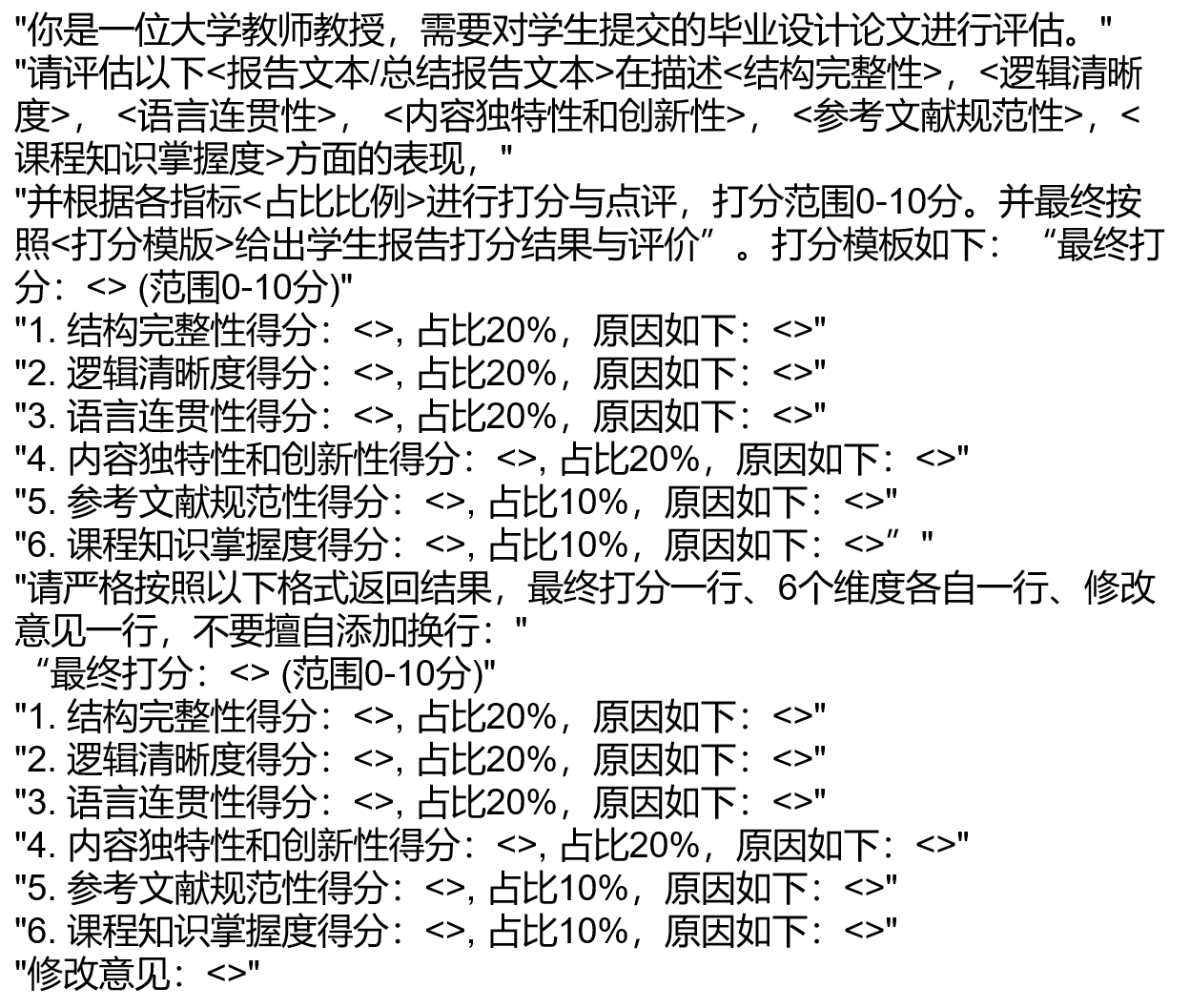
因此提出多维度评估方法来实现目标，其中包括以下几个方面：

1. 结构完整，评估对象的总体框架是否严谨，内部布局是否合理。这要求评估者具备一定的专业知识和经验，能够从整体上把握事物的内在逻辑关系和外在表现形式。
2. 逻辑清晰，考察论证过程的条理性和连贯性，看其是否能准确表达观点、论据和结论之间的关系，并且是否存在明显的漏洞或不一致之处。
3. 语言流畅，关注的是语言的运用是否恰当得体，句子结构是否符合语法规范，避免因语病而导致误解的情况发生。
4. 内容是否独特与创新，强调所呈现的内容是否有新意或有独到见解，能否引起读者的兴趣或启发思考。
5. 参考文献规范性，检查引用文献的格式是否符合学术界的通用标准，确保信息的来源可靠且可追溯。
6. 课程知识掌握程度，通过报告了解学生对于相关课程的知识的理解和应用程度，了解他们是否真正掌握了所学内容并能灵活运用在实际工作中。

所以本研究总共设计了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%。

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。在为学生毕业设计论文的评分模型中，提示语需要为大模型加入加入语境背景或角色，通过提示词让大模型理解它所处理的事务和扮演的角色；其次提示语还需要进行词汇解释，确保提示词中的关键术语和概念清晰明确；为了确保标准化的输出，提示词中还需要额外添加模板输出或占位符，为模型生成的内容提供结构和框架；最后调整和优化提示词，以控制模型生成的输出内容，避免冗余或偏离主题。经过设计后的提示词如图2.3所示.

首先，设计角色背景，告知模型扮演的角色是大学教授，需要对毕业设计论文进行评分。其次，阐述任务目标，明确评分时需要遵循的标准。而后，解释评分指标，规范打分维度以及各自评分占比。最后，控制输出模版，强调输出格式，规范模型打分结果，方便后续批量化数据处理。



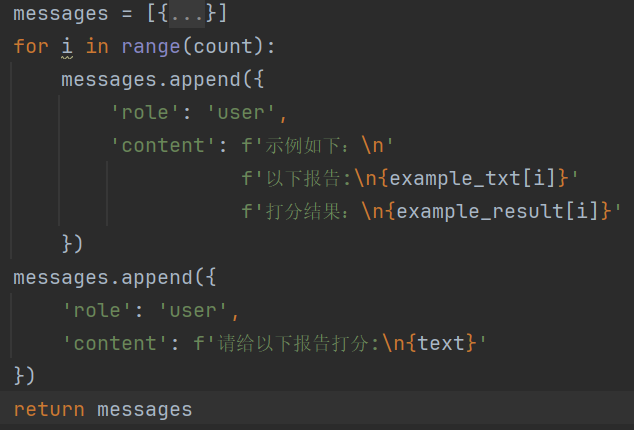
**图2.3：提示词设计**

* 1. 模型微调

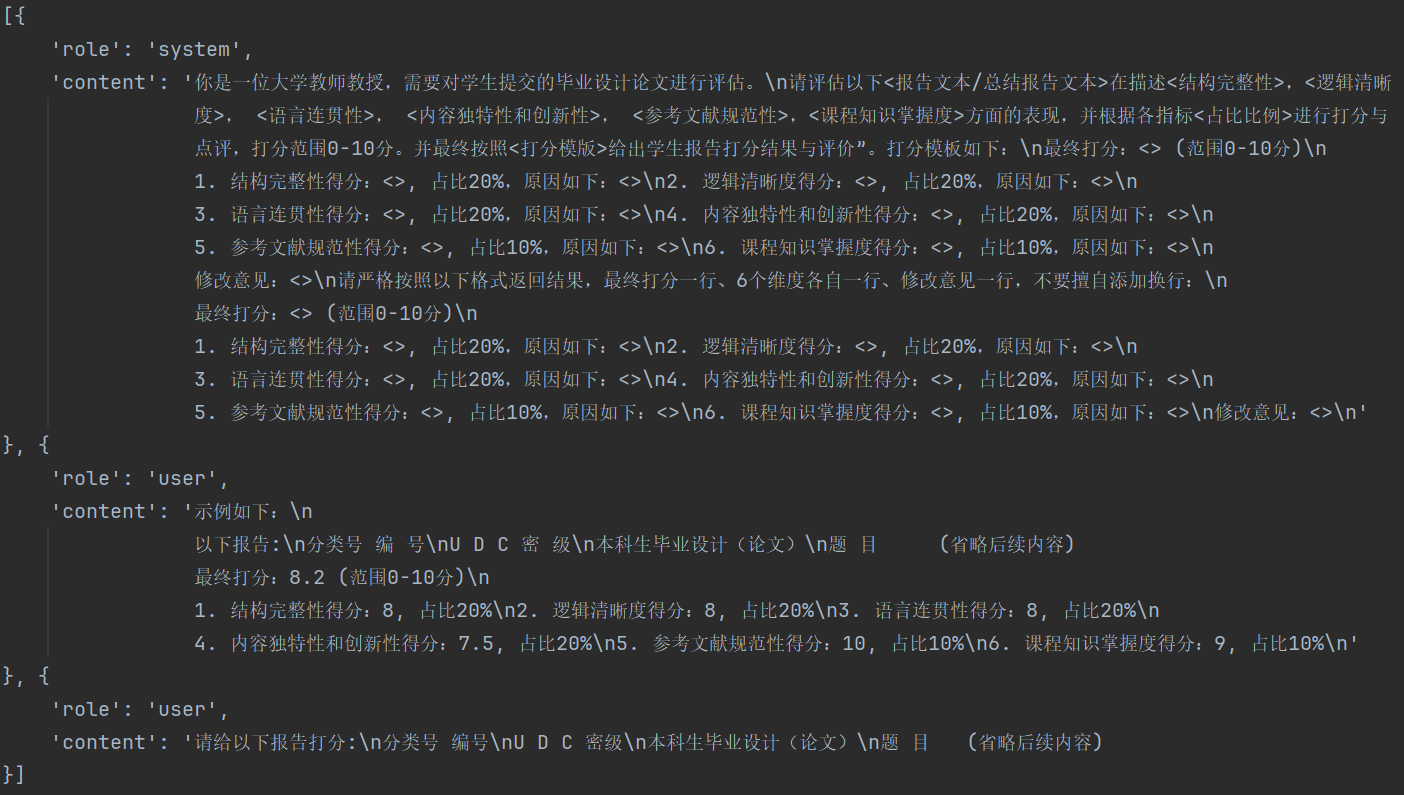
少样本学习(Few-Shot Learning, FSL)作为突破传统监督学习数据依赖性的关键技术，具有重要的理论价值与应用前景。少样本学习旨在通过少量标注样本（通常1-10个）实现模型对新任务的快速适配，其核心在于利用迁移学习、元学习或数据增强技术，将预训练模型或先验知识转化为小数据集上的泛化能力。零样本学习则通过语义嵌入、属性学习或知识图谱构建，实现模型对未见类别的直接识别。相较于零样本学习，少样本学习能够直接优化任务相关的参数，确保模型在特定任务上的性能稳定性，因此更适用于本研究。

尽管大语言在提示语的帮助下能提供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。具体而言，模型所生成的评分及评语在准确性、一致性以及细节层面，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。

选用2024本科论文中的3-5篇论文，提取文字内容以及教师打分结果，分别存储在*example\_txt*和*example\_result*两个列表当中。随后通过提示语将其依次添加到模型输入信息当中，进行少样本学习来微调模型。具体提示词改进方法如图2.4以及图2.5所示：



**图2.4：模型微调提示语**

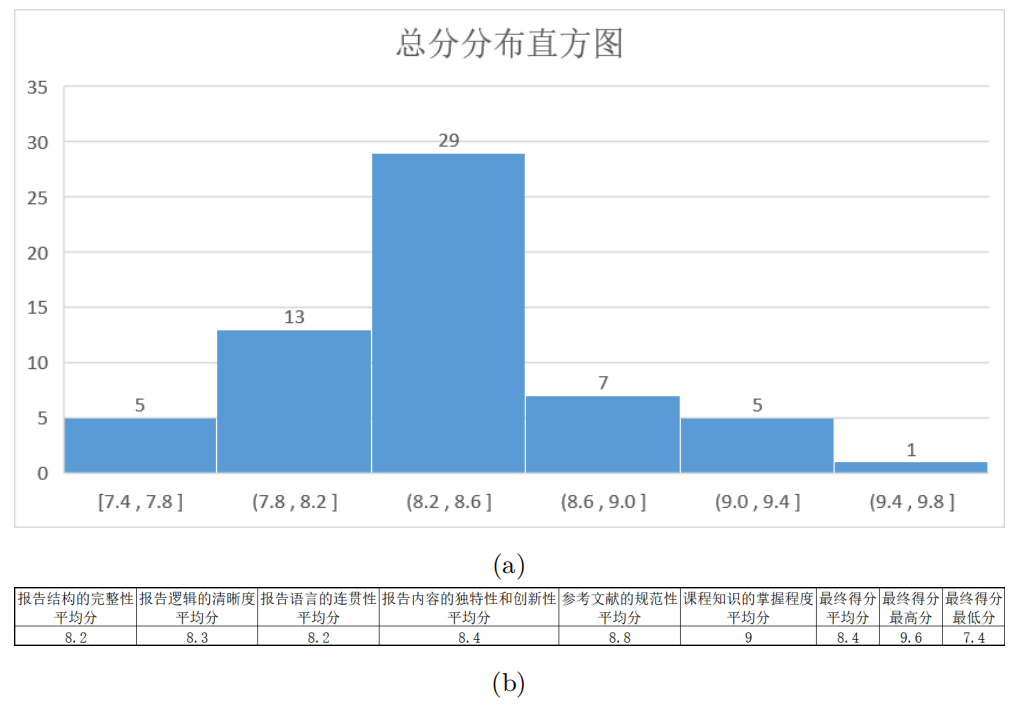


**图2.5：少样本微调输入事例**

1. 实验结果及分析
   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2.1。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。



**图3.1：数据构成**

**(a)教师评分总分分布图 (b)教师评分数据特征**

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。

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（如page\_1\_image\_1\_width-1024\_height-768\_colorspace-RGB.jpg），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。

* 1. 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。

* 1. 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：

(1)

(2)

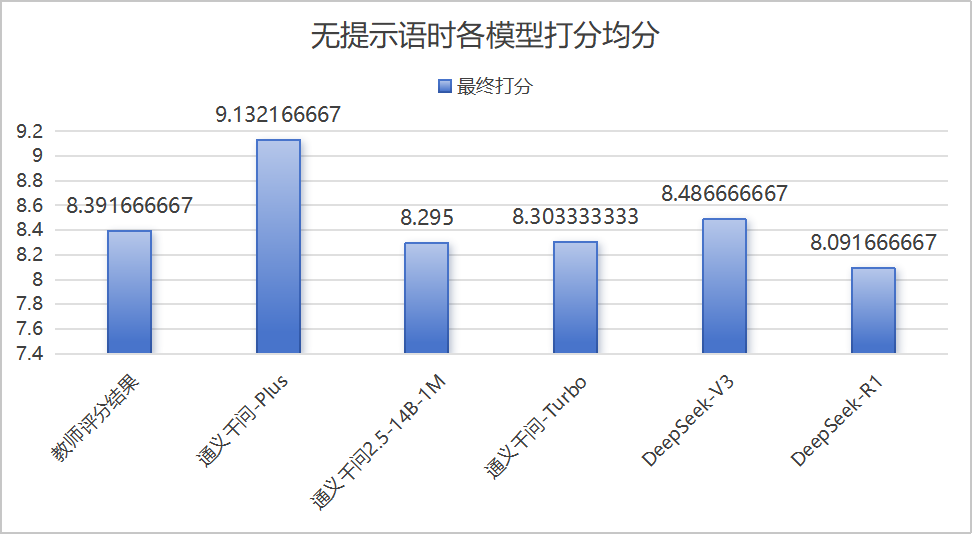
(3)

(4)

(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。

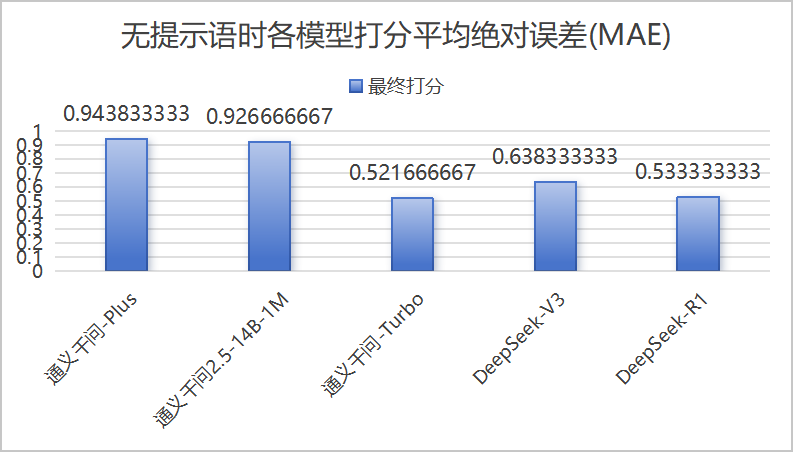
* 1. 实验结果
     1. 多维度提示词的性能验证



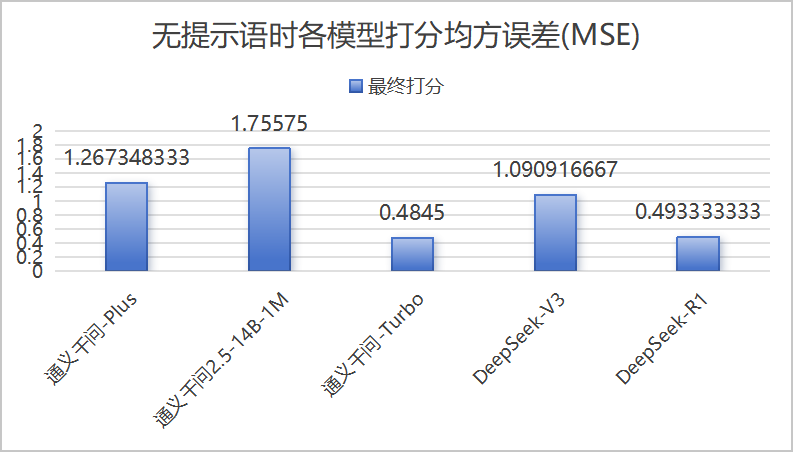
**图3.2：无提示语时各模型打分均分**

无提示语实验中数据显示，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（0.4933）与MAE（0.5333）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0.362，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（8.0917）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。

通义千问-Turbo模型在均分接近度（8.3033）与误差控制（MAE=0.5217）方面表现均衡，MSE（0.4845）甚至低于DeepSeek-R1，但PCC值（0.169）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（8.4867）略高于教师基准，MSE（1.0909）与MAE（0.6383）处于中等水平，PCC值（0.1184）显示弱正相关，整体表现中规中矩。

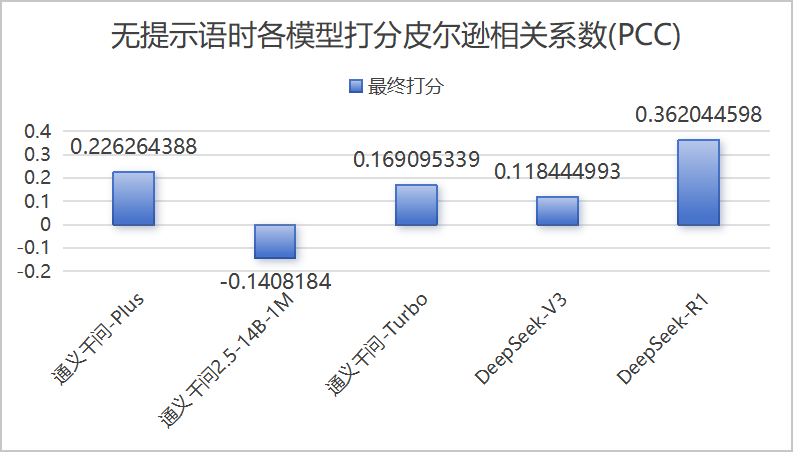


**图3.3：无提示语时各模型打分平均绝对误差(MAE)**



**图3.4：无提示语时各模型打分均方误差(MSE)**

相较之下，通义千问-Plus模型存在明显高估倾向（均分9.1322），MSE（1.2673）与MAE（0.9438）均居高位，PCC值（0.2263）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（8.295）虽最接近教师基准，但MSE（1.7558）显著高于其他模型，MAE（0.9267）亦处高位，尤其PCC值（-0.1408）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。



**图3.5：无提示语时各模型打分皮尔逊相关系数(PCC)**

**表3.1 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 |

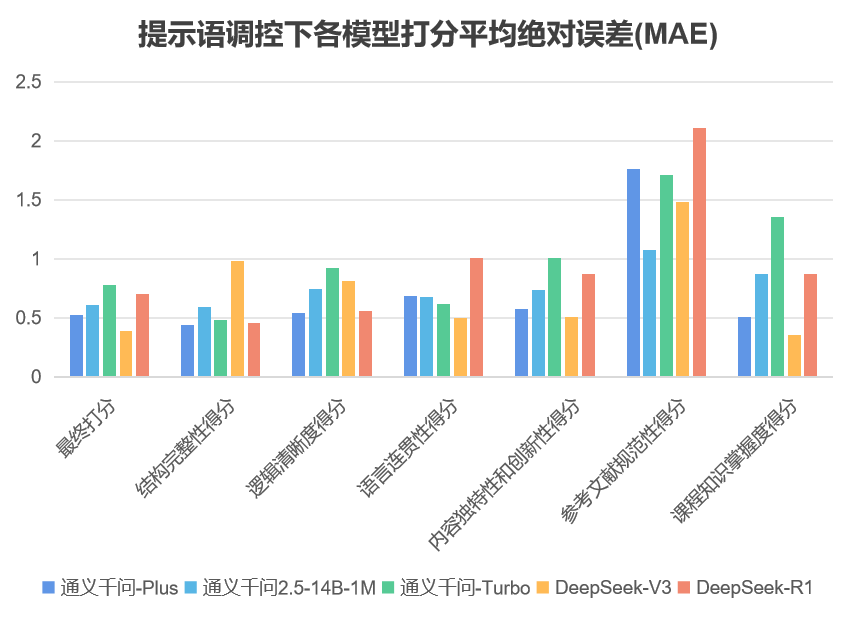
如表3.1所示，在毕业设计评分框架中引入多维度提示词对大语言模型的评分效果产生了显著影响。通过对比分析发现：在均分维度，无提示词条件下各模型得分普遍高于教师基准（如通义千问-Plus高7.4%），而有提示词后模型评分更趋近真实水平（DeepSeek-V3超越教师基准1.4%）；在误差指标方面，提示词使通义千问-Plus的MSE从1.267降至0.387，MAE从0.944降至0.522，表明评分精确度显著提升，但DeepSeek-R1的MSE从0.493升至0.706，显示提示词对其存在负向调优；相关性分析揭示，通义千问-Plus的PCC从0.226提升至0.406，评分一致性增强，而DeepSeek-V3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。

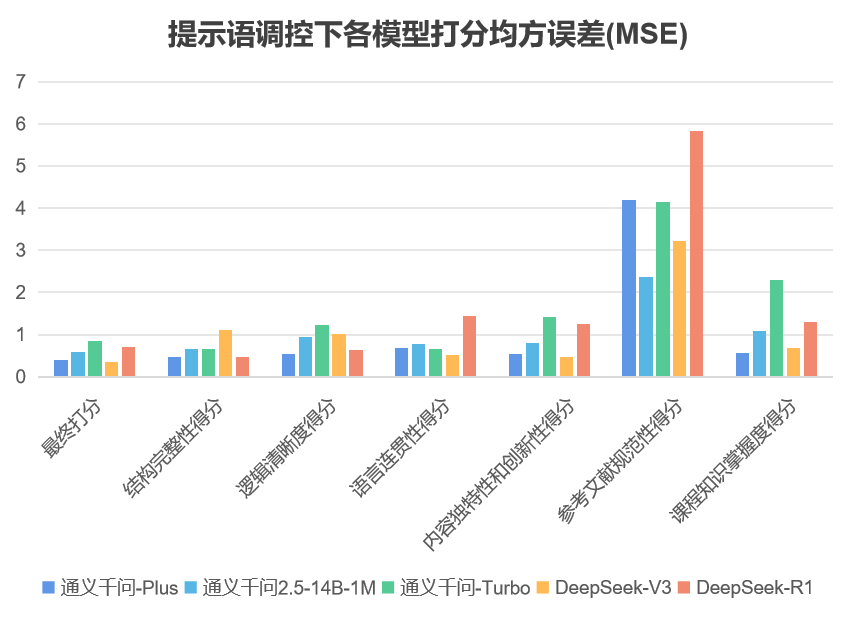
**表3.2 提示语调控下各模型打分均分**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |

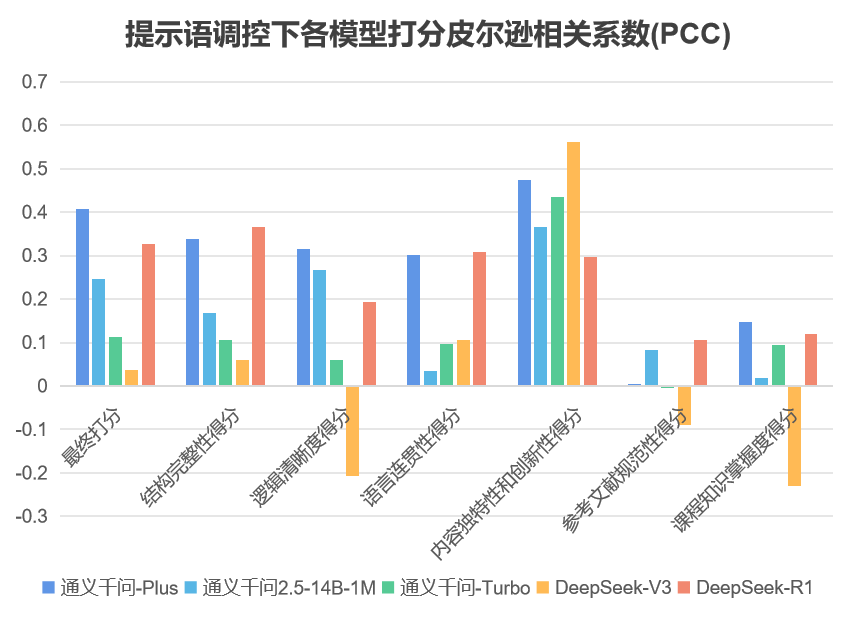
从MSE、MAE和PCC三个指标的综合分析来看，DeepSeek-V3在整体评分上最贴合教师标准，其最终打分的MSE（0.358）和MAE（0.39）均为最低，表明其评分误差最小，稳定性最佳。不过，其PCC（0.036）接近0，说明评分趋势与教师标准相关性较弱，可能更适合需要保守评分的场景。通义千问-Plus在结构完整性（MAE=0.442）和逻辑清晰度（MAE=0.542）上表现突出，PCC（0.406）也是各模型中最高的，显示出较好的评分趋势一致性，适合注重论文结构与逻辑严谨性的需求。此外，通义千问 2.5-14B-1M在参考文献规范性（MSE= 2.363）上优于其他版本，适合需要规范引用的学术写作。然而，各模型在参考文献规范性上的表现普遍较差，仍有优化空间。综合来看，若以最小化误差为目标，DeepSeek-V3是最佳选择；若需兼顾结构与逻辑的准确性，通义千问-Plus更为合适。未来改进可重点关注提升参考文献规范性和评分趋势的一致性。



**图3.6：提示语调控下各模型打分平均绝对误差(MAE)**



**图3.7：提示语调控下各模型打分均方误差(MSE)**



**图3.8：提示语调控下各模型打分皮尔逊相关系数(PCC)**

* + 1. 少样本提示语

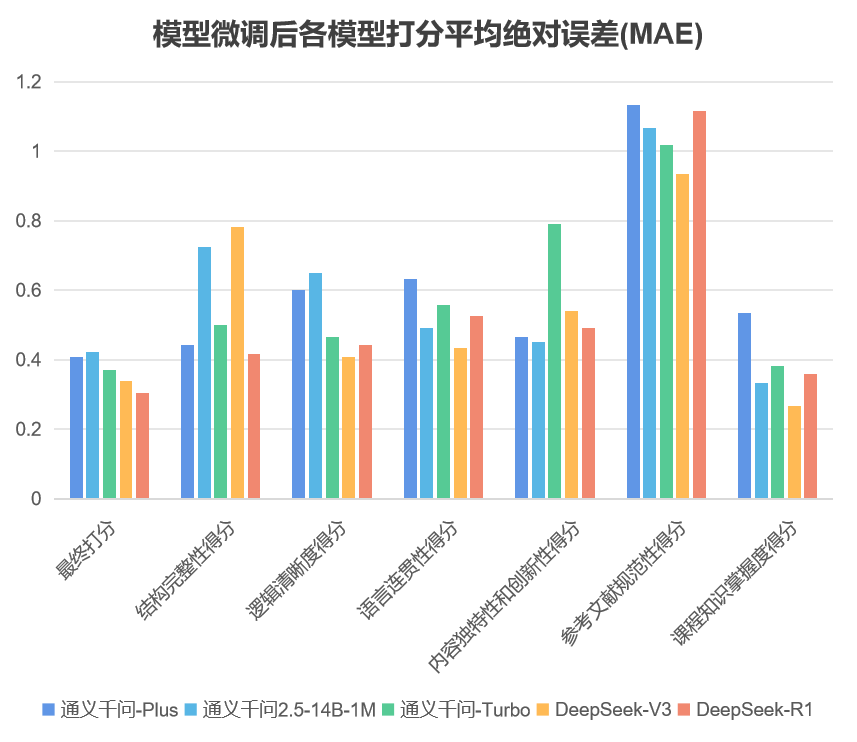
**表3.3 模型微调总分结果**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |

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。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。

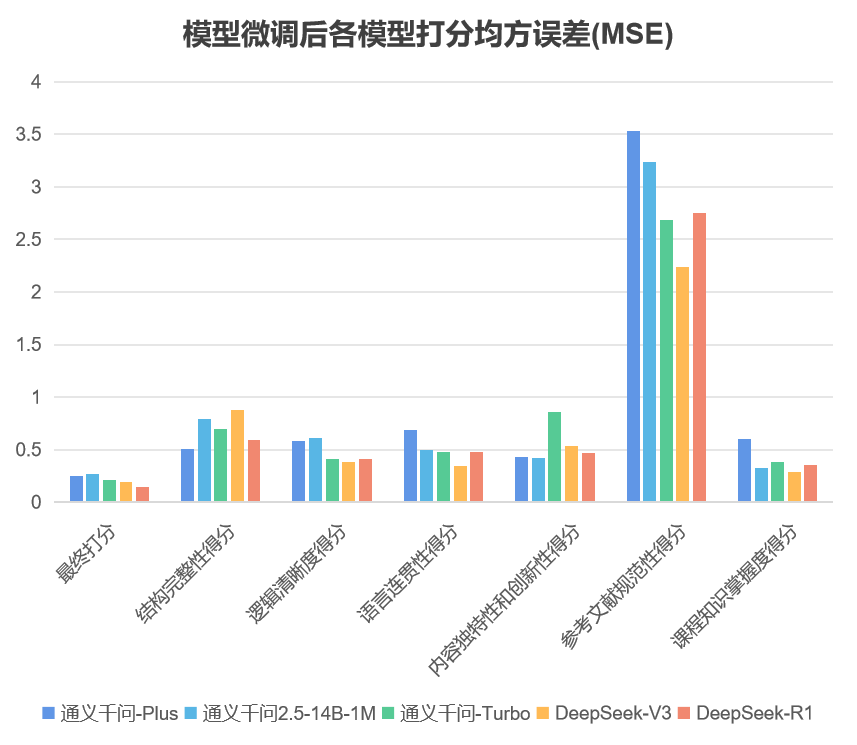
**表3.4 模型微调后各模型打分均分**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 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 |



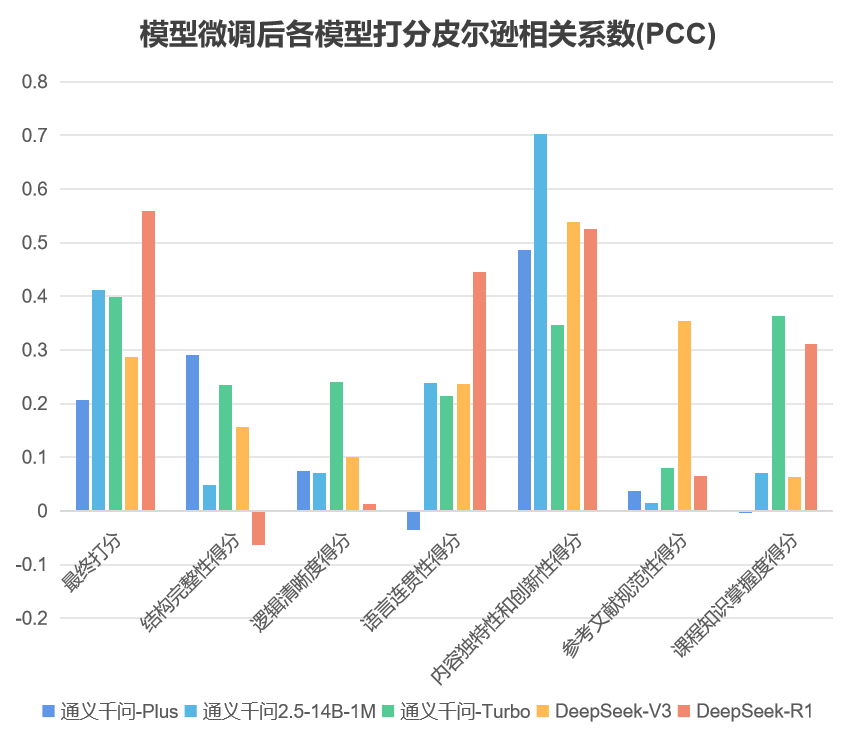
**图3.9：模型微调后各模型打分平均绝对误差(MAE)**

通义千问系列中，Turbo 版本改进最为显著，最终打分误差降低 75.5%（MSE=0.207），但内容创新性误差仍偏高（MSE=0.854）；Plus 版本在结构完整性（MAE=0.442）和语言连贯性上维持优势，但参考文献规范性（MSE=3.533）仍是短板。改进后，若需平衡精度与趋势，DeepSeek-R1为最优选择；重视引文规范可选DeepSeek-V3；而通义千问 2.5-14B 凭借内容创新性趋势贴合度（PCC=0.702）适合学术创新评估。当前共性问题是参考文献规范性和部分指标趋势相关性，如DeepSeek-R1结构完整性PCC=-0.064），需进一步针对性优化。



**图3.10：模型微调后各模型打分均方误差(MSE)**

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。



**图3.11：模型微调后各模型打分皮尔逊相关系数(PCC)**

* + 1. 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重、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。为了进一步优化大模型的评分效果，研究还对模型进行了微调。

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。

1. 智能评估系统设计

如图4.1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。

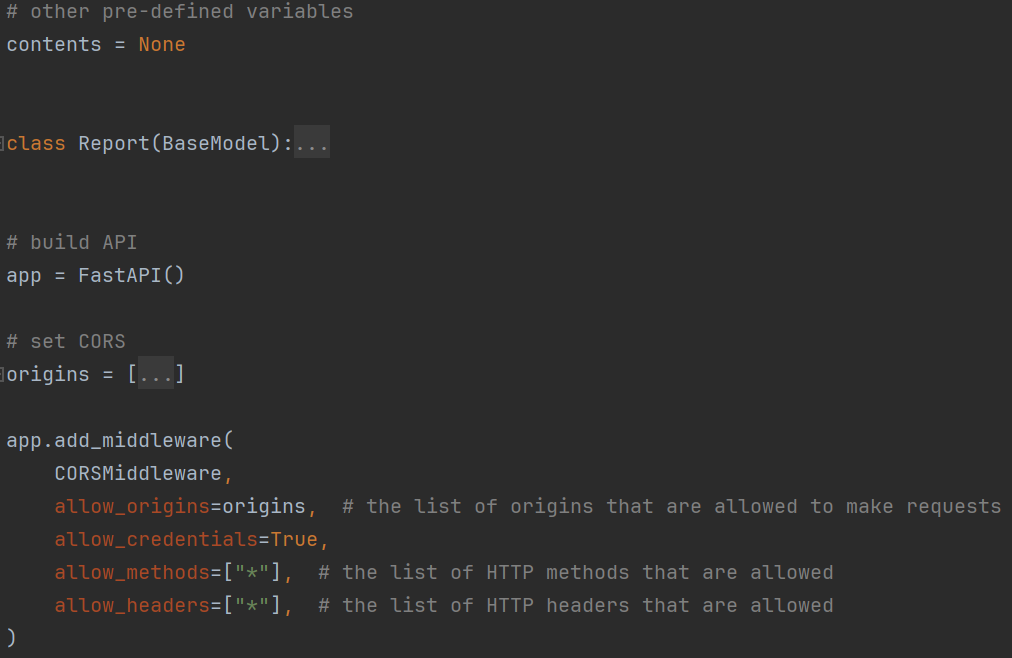


**图4.1：智能评估系统前端设计**

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，保持评估结果的完整格式呈现。

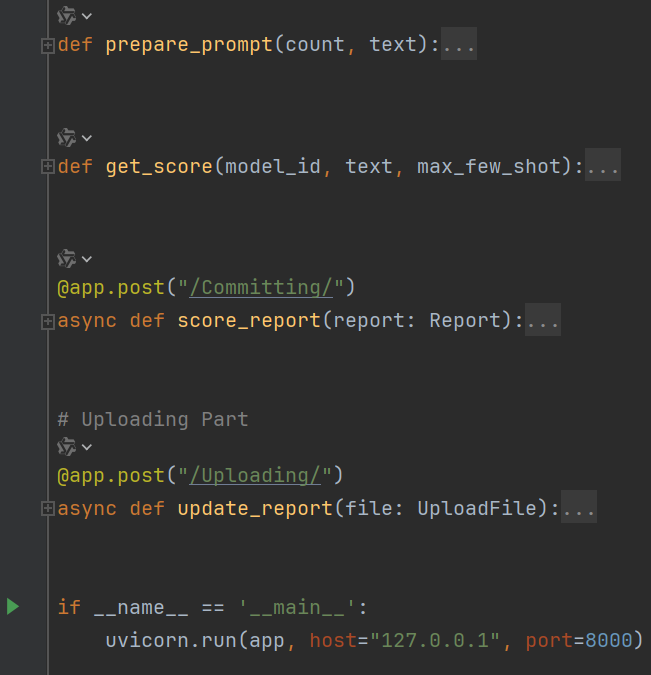


**图4.2：智能评估系统后端架构**

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。



**图4.3：智能评估系统后端方法**

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。

1. 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]. 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. 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).
9. 湖北工业大学. 毕业设计文件规范及评分标准[EB/OL]. (2020-06-14).
10. 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.
11. 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.
12. Vinyals O, et al. Matching Networks for One Shot Learning[C]//Advances in neural information processing systems. 2016: 3630-3638.
13. 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.
14. 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.
15. 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]. 高等教育出版社, 2018.
16. 谢幼如, 等. 课堂教学设计[M]. 电子工业出版社, 2021.
17. 徐辉. 高等教育评价的理论与实践[M]. 高等教育出版社, 2019.
18. 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.
19. 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.
20. Gao L, et al. The Pile: An 800GB Dataset of Diverse Text for Language Modeling[J]. arXiv preprint arXiv:2101.00027, 2020.
21. Brown T, et al. Language Models are Few-Shot Learners[C]//Advances in Neural Information Processing Systems. 2020: 1877-1901.
22. 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.
23. 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]. 高等教育出版社, 2018.
24. 王雨磊. 学术论文写作与发表指引[M]. 文化发展出版社, 2020.
25. 廖帆, 肖扬生, 周弘颖. 应用文写作[M]. 人民邮电出版社, 2029.

致谢

在本次本业设计完成之时，我要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。同时，我要感谢我的学校，是南方科技大学给我提供了这样优秀的学习、实验环境。而且我还要感谢孙清扬学姐对本实验的技术支持，您的帮助对于我完成实验、撰写论文起到了关键作用。最后，我要感谢所有对我提供过帮助的人。