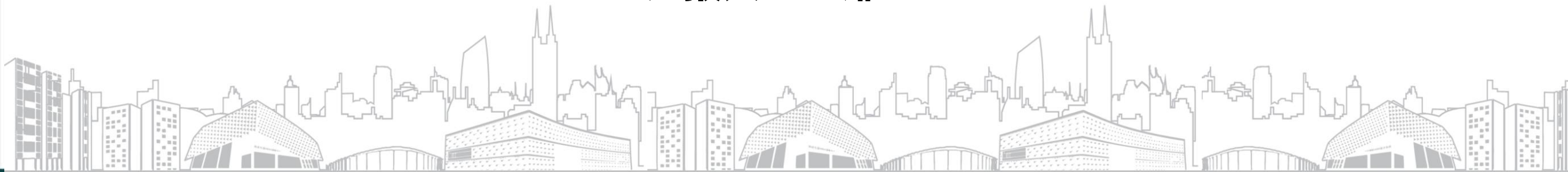


基于大语言模型的毕业设计评分框架研发

院系：计算机科学与工程系

指导老师：刘江教授

汇报人：王谦益



1

研究背景

2

研究内容

3

实验结果

4

总结展望



1

研究背景



1.1 研究背景

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

检验学生知识整合能力、科研素养与实践技能

衔接校园学习与职业发展

直接反映高校人才培养水平，对学生学术能力认证、职业素养塑造及学习能力具有深远影响



1.1 研究背景

现有人工评估的挑战：

1. 人工评阅模式受限于教师精力，评分时间较为紧张
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
3. 传统评语以总体性建议为主，缺乏精准反馈，学生获得的评价针对性不强



1.1 研究背景

- 《深化新时代教育评价改革总体方案》

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

- 《教育信息化2.0行动计划》

推动人工智能在教学、管理等方面的全流程应用

LLMs的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。



1.2 相关工作 — 自动化作文评分

Are Large Language Models Good Essay Graders?

Kundu, Anindita

kundu2@ualberta.ca

Barbosa, Denilson

denilson@ualberta.ca

September 23, 2024

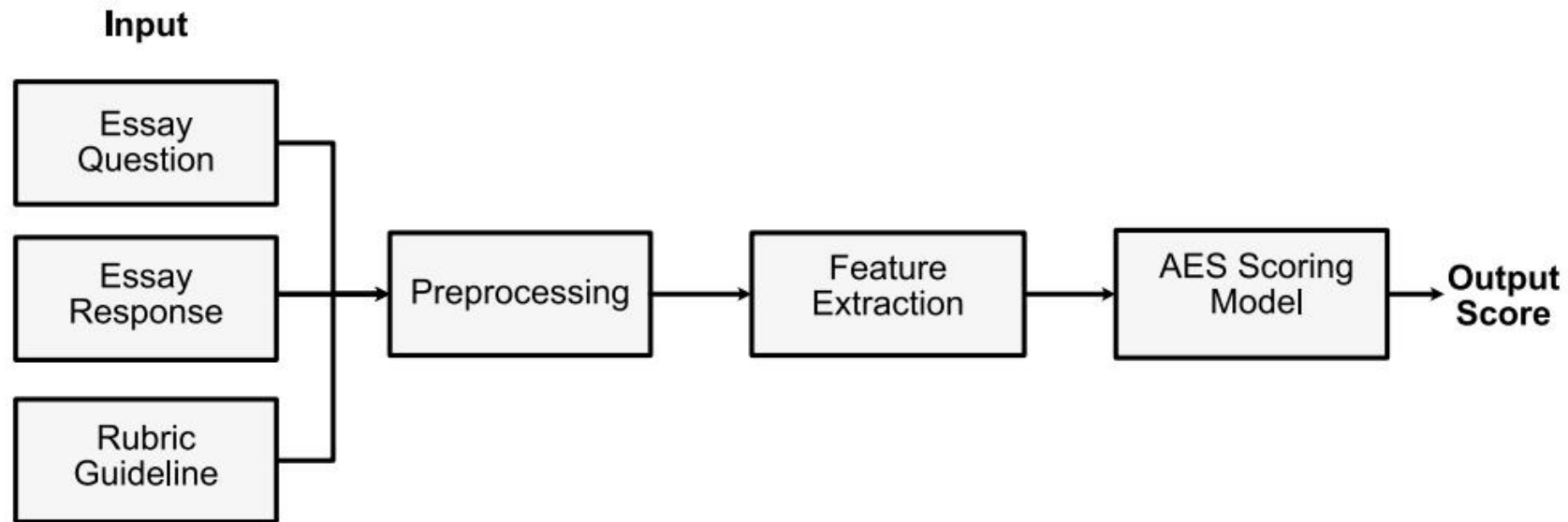


Figure 1: Automated Essay Scoring Pipeline

1.2 相关工作 — 自动化作文评分

Automatic Essay Multi-dimensional Scoring with Fine-tuning and Multiple Regression

Kun Sun^{*1} and Rong Wang^{†2}

¹Department of Linguistics, University of Tübingen, Germany

²Institute of Natural Language Processing, Stuttgart University, Stuttgart, Germany

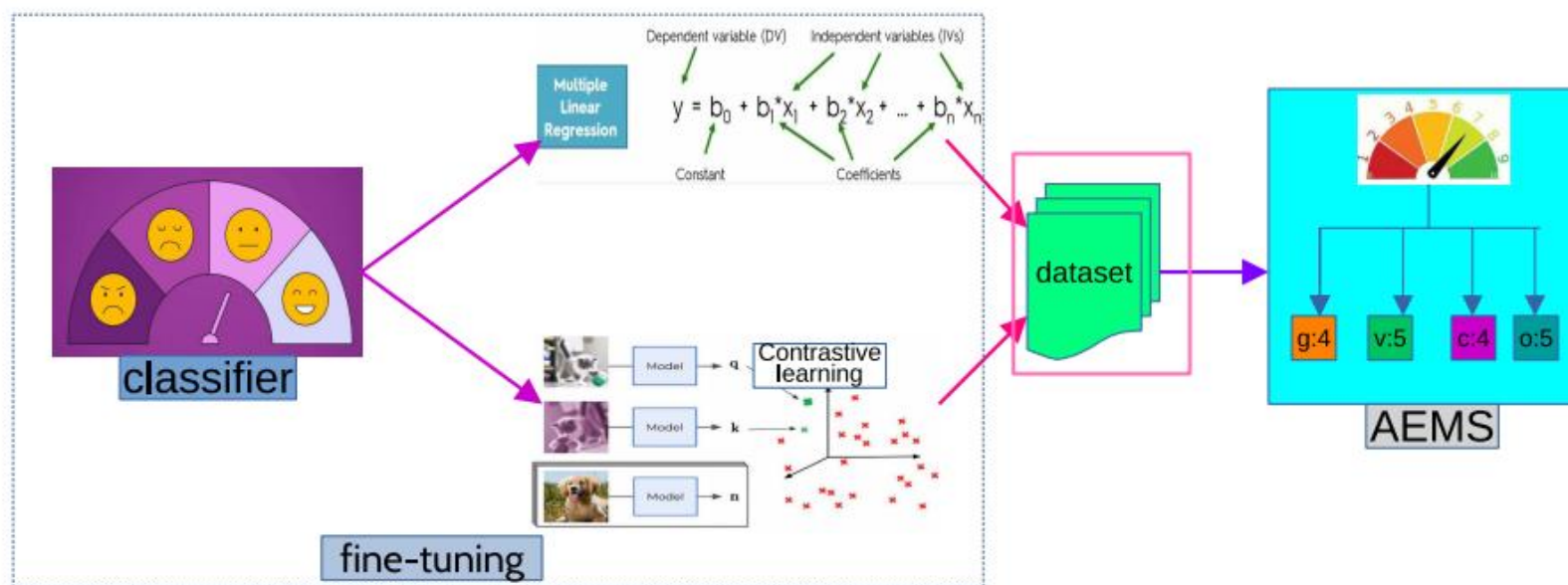


Figure 1: The roadmap of developing AEMU in the present study

1.2 相关工作 — 自动化作文评分

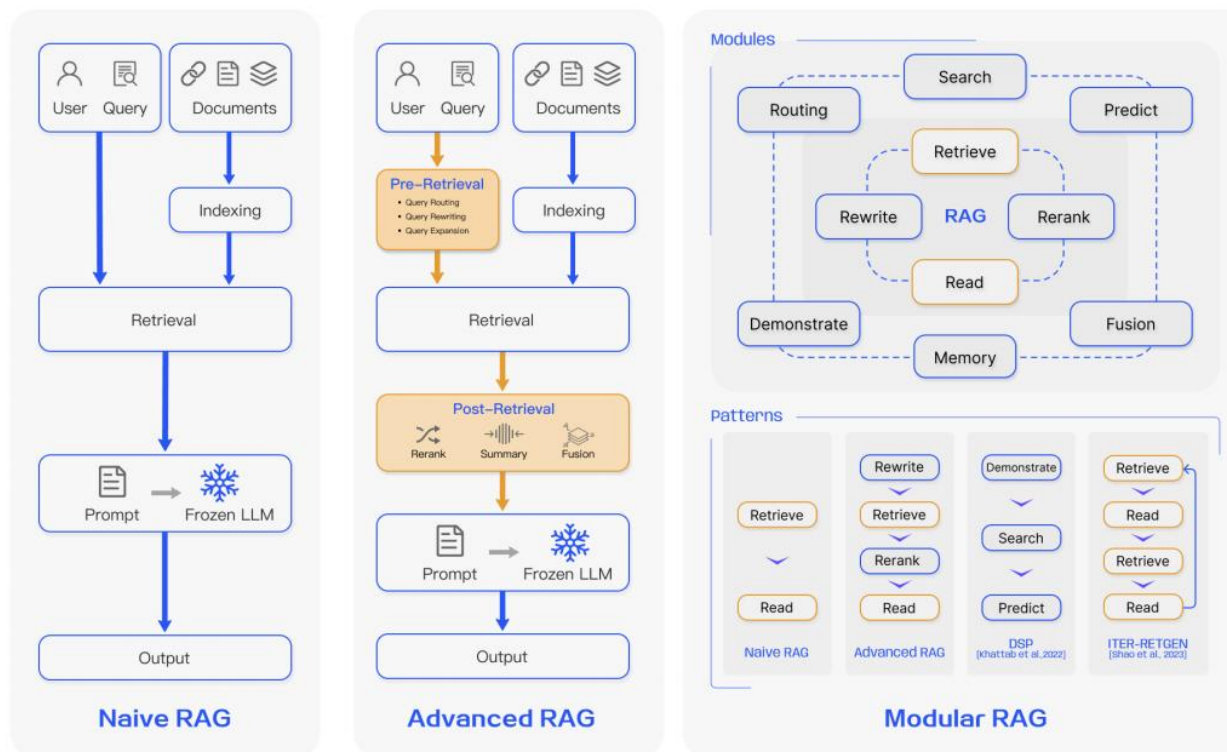
Retrieval-Augmented Generation for Large Language Models: A Survey

Yunfan Gao^a, Yun Xiong^b, Xinyu Gao^b, Kangxiang Jia^b, Jinliu Pan^b, Yuxi Bi^c, Yi Dai^a, Jiawei Sun^a, Meng Wang^c, and Haofen Wang^{a,c}

^aShanghai Research Institute for Intelligent Autonomous Systems, Tongji University

^bShanghai Key Laboratory of Data Science, School of Computer Science, Fudan University

^cCollege of Design and Innovation, Tongji University





1.3 相关工作 — 大语言模型

| 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 |

1.4 研究内容

自动评分不足：

评分与人工打分之间仍存在显著差异，且相关性较弱。

研究内容：

搭建基于大语言模型的毕业设计评分框架，**多维度**（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）评估毕设内容并生成评语和建议，帮助学生改善论文内容， 辅助教师快速打分。

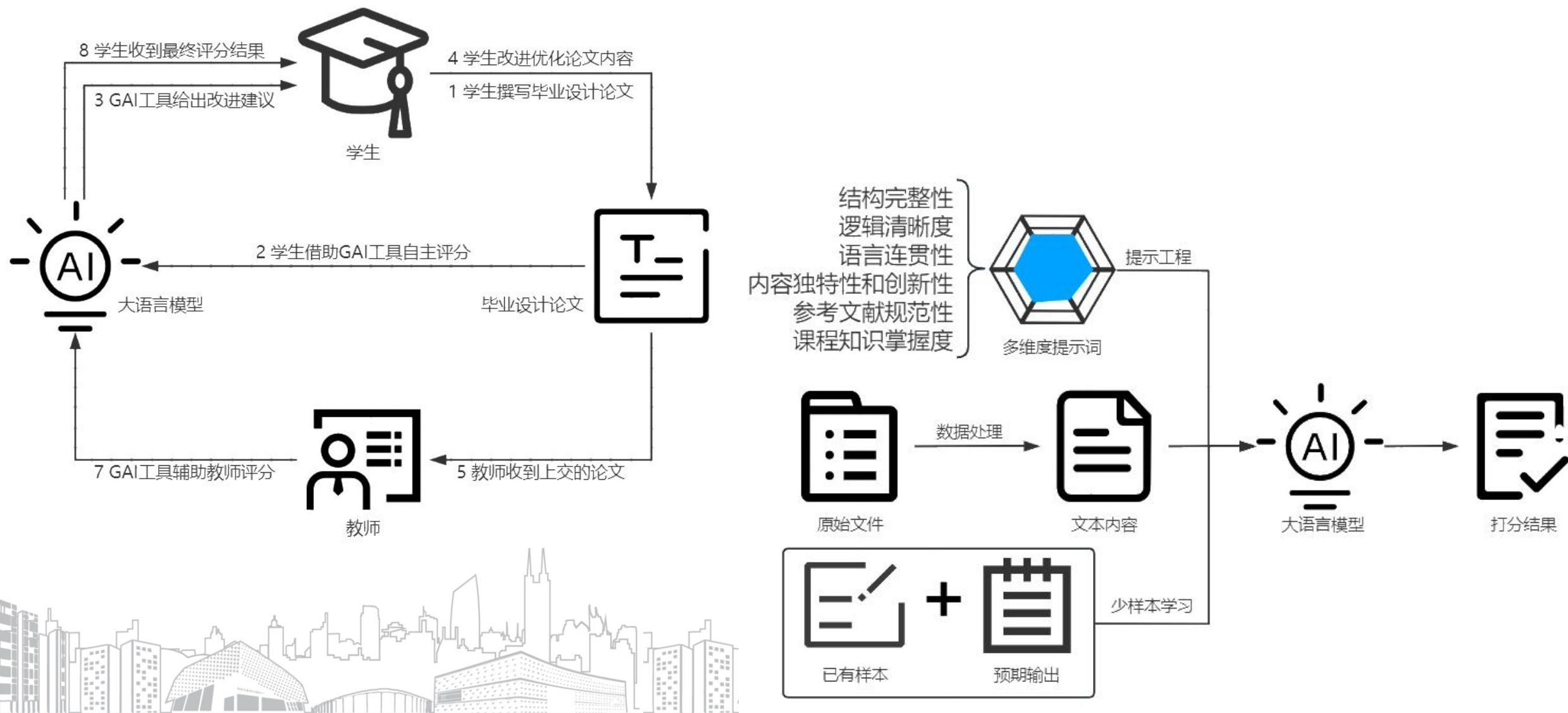


2

研究内容



2.1 研究框架



2.2 多维度评分

- 1. 结构完整性得分:占比20%: 评估对象的总体框架是否严谨, 内部布局是否合理
- 2. 逻辑清晰度得分:占比20%: 考察论证过程的条理性和连贯性
- 3. 语言连贯性得分:占比20%: 语言的运用是否恰当得体, 句子结构是否符合语法规范
- 4. 内容独特性和创新性得分:占比20%: 内容是否有新意或有独到见解
- 5. 参考文献规范性得分:占比10%: 检查引用文献的格式是否符合学术界的通用标准
- 6. 课程知识掌握度得分:占比10%: 了解学生对于相关课程的知识的理解和应用程度

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



2.3 提示词设计

1. 设计角色背景

2. 阐述任务目标

3. 解释评分指标

4. 控制输出模版

"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：<>"

2.4 模型微调

通过prompt进行微调

利用文本及教师打分，进行少样本学习

少样本学习 (Few-Shot Learning) 是针对标注数据稀缺场景设计的机器学习范式，其核心目标是让模型仅需少量样本（如每个类别仅1-5个样本）即可完成高效学习。

```
[{
  '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）\\n1. 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：<>\\n3. 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：<>\\n5. 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）\\n1. 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：<>\\n3. 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：<>\\n5. 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\\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 （省略后续内容）\\n\\n最终打分：8.2（范围0-10分）\\n\\n1. 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%\\n4. 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 （省略后续内容）'
}]
```

3

实验结果



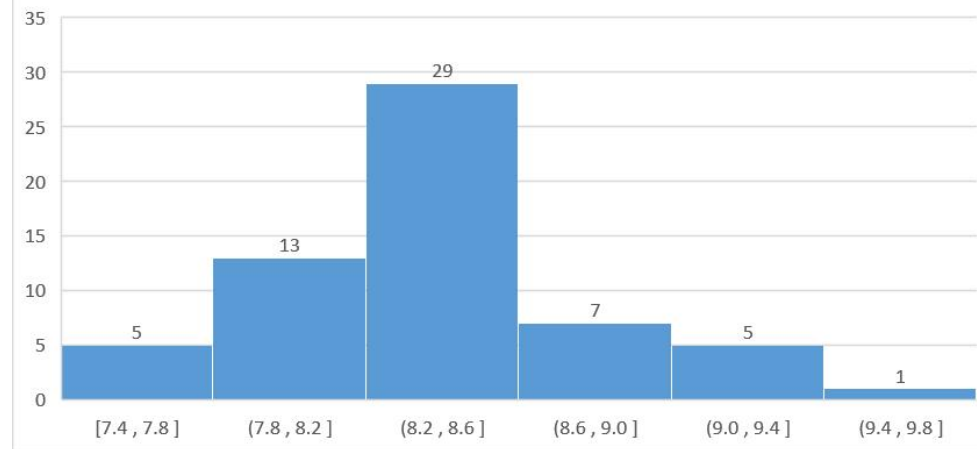
3.1 数据集构造

数据处理:

1. 对筛选出的PDF和Word报告进行文字提取, 去除页面布局和格式的影响
2. 根据定义好的评分标准, 统计教师对每一部分内容的量化打分

```
def get_txt(in_path):  
    """  
    提取PDF文字内容  
    """  
    reports = {}  
    full_text = []  
    with pdfplumber.open(in_path) as pdf:  
        for page_number, page in enumerate(pdf.pages, start=1):  
            # 页数编码标签  
            page_tip = "\n# page " + str(page_number) + ":" if page_number != 1 else "# page 1:"  
            # 优先提取文本  
            text = page.extract_text()  
            if text and text.strip():  
                full_text.append(text)  
            else:  
                # 处理扫描件图片OCR  
                # print("need use OCR to scan")  
                # return "failed"  
                img = page.to_image(resolution=300).original  
                img_bytes = io.BytesIO()  
                img.save(img_bytes, format='PNG')  
                img = Image.open(img_bytes)  
                ocr_text = pytesseract.image_to_string(img, lang='chi_sim') # 支持中文OCR  
                full_text.append(ocr_text)
```

总分分布直方图



```
def get_img(in_path, img_path):  
    """  
    提取PDF所有图片  
    """  
    # 创建图片文件夹  
    if not os.path.exists(img_path):  
        os.makedirs(img_path)  
  
    # 读取PDF文件  
    pdf_document = fitz.open(in_path)  
  
    # 遍历每一页  
    for page_number in range(len(pdf_document)):  
        page = pdf_document.load_page(page_number)  
        image_list = page.get_images(full=True)  
  
        # 提取图片  
        for img_index, img in enumerate(image_list):  
            xref = img[0] # 图片的交叉引用编号  
            base_image = pdf_document.extract_image(xref)  
            image_data = base_image["image"]  
            image_form = base_image["ext"]  
            image_name = ""  
            for i, info in enumerate(base_image):
```

3.2 实验设置

实验模型 —— 阿里云百联平台

通义千问-Plus、通义千问2.5-14B-1M、通义千问-Turbo、DeepSeek-V3、DeepSeek-R1

实验内容

1. 无提示词
2. 有提示词
3. 模型精调



3.3 评价标准

1. 平均分 $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i; \bar{t} = \frac{1}{n} \sum_{i=1}^n t_i$

2. 平均绝对误差(Mean Absolute Error, MAE)

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

3. 均方误差(Mean Squared Error, MSE)

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

4. 皮尔逊相关系数(Pearson Correlation Coefficient, PCC)

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

其中y为模型打分结果, t为教师评分结果



3.4.1 多维度提示词的性能验证

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。

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

| 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 |

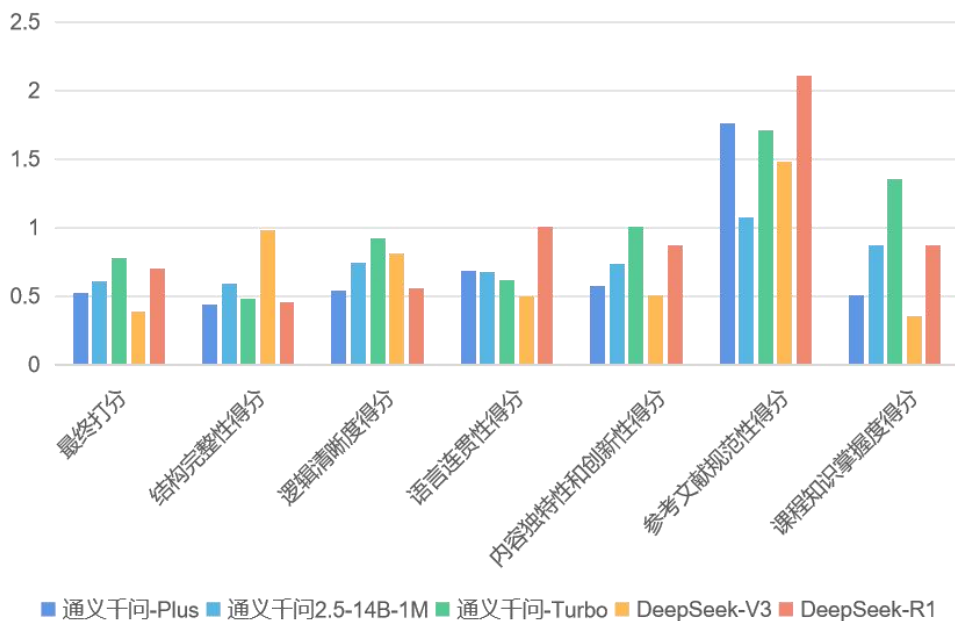


3.4.1 多维度提示词的性能验证

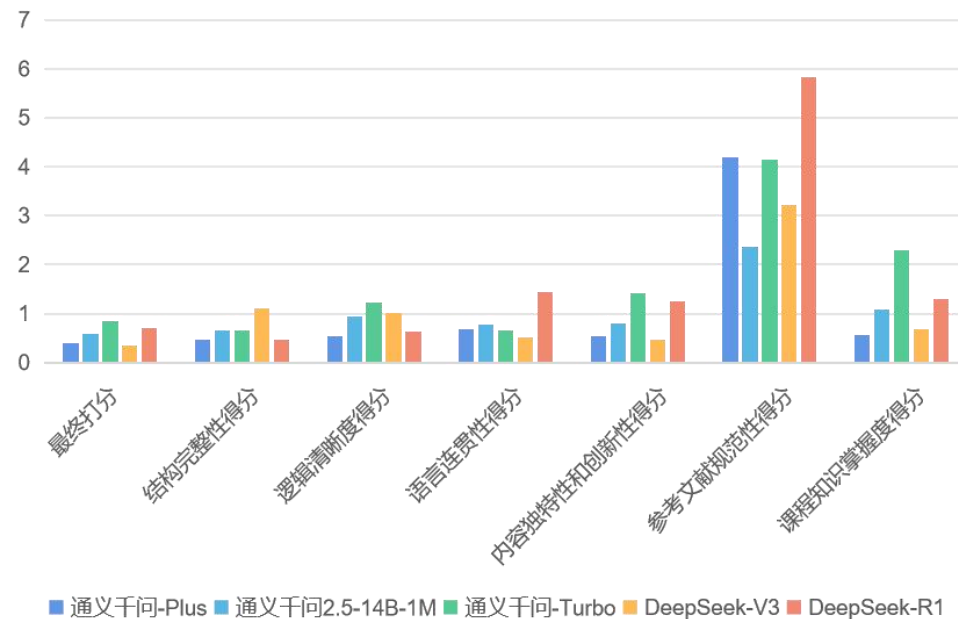
| 模型 | 最终 打分 | 结构 完整性 | 逻辑 清晰度 | 语言 连贯性 | 内容独特性 和创新性 | 参考文献 规范性 | 课程知识 掌握度 |
|--------------------|----------|-----------|-----------|-----------|---------------|-------------|-------------|
| 教师评分 结果 | 8.39 | 8.20 | 8.31 | 8.15 | 8.36 | 8.83 | 9.05 |
| 通义千问 Plus | 7.96 | 8.34 | 7.87 | 7.62 | 8.10 | 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 |

3.4.1 多维度提示词的性能验证

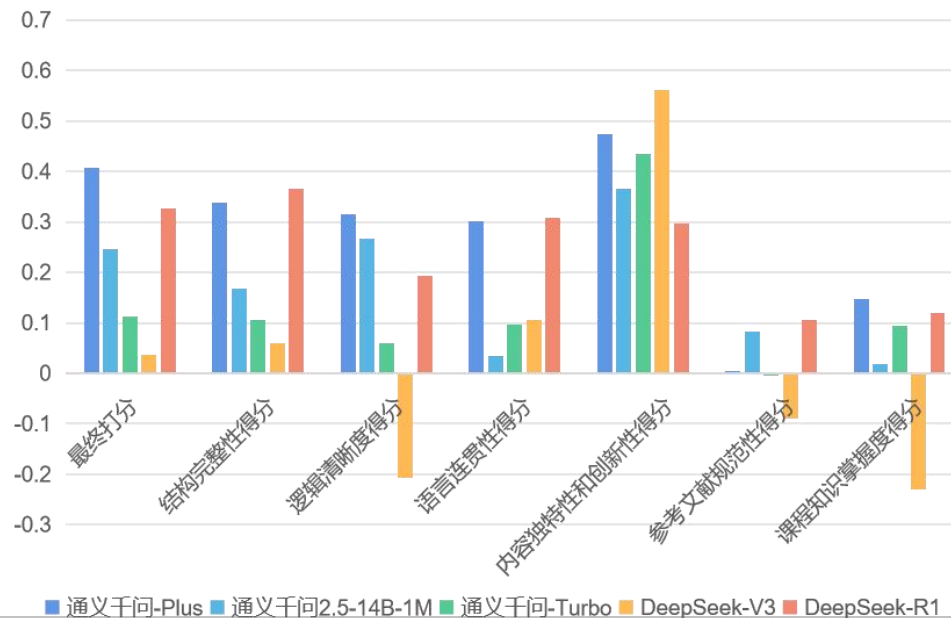
提示语调控下各模型打分平均绝对误差(MAE)



提示语调控下各模型打分均方差(MSE)



提示语调控下各模型打分皮尔逊相关系数(PCC)





3.4.2 少样本提示语

| 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 |

模型打分结果得到提升

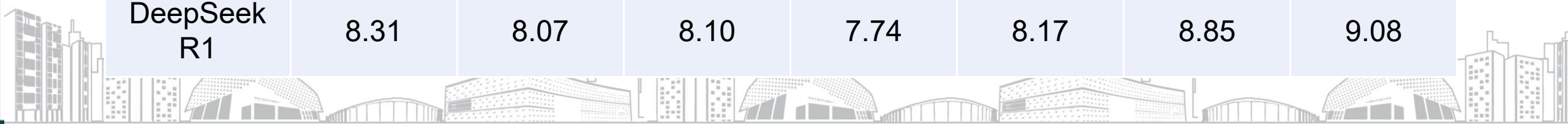
其中DeepSeek-V3和通义千问2.5-14B-1M在贴合教师评分标准方面表现突出

本次精调训练成功提升了模型的评分效果，为后续的模式优化和实际应用打下了坚实的基础



3.4.1 多维度提示词的性能验证

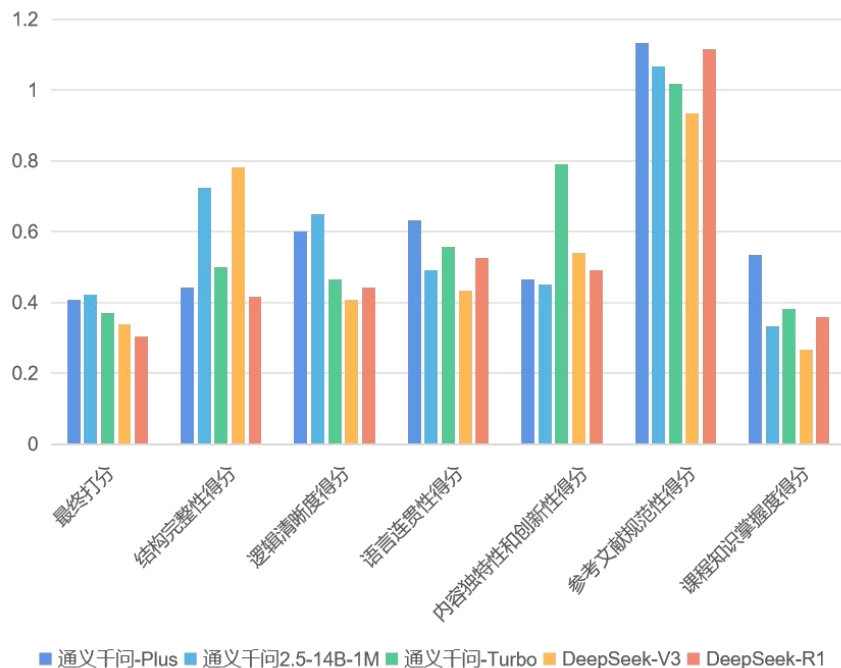
| 模型 | 最终打分 | 结构完整性 | 逻辑清晰度 | 语言连贯性 | 内容独特性和创新性 | 参考文献规范性 | 课程知识掌握度 |
|-----------------|------|-------|-------|-------|-----------|---------|---------|
| 教师评分结果 | 8.39 | 8.20 | 8.31 | 8.15 | 8.36 | 8.83 | 9.05 |
| 通义千问 Plus | 8.66 | 8.34 | 8.81 | 8.48 | 8.51 | 9.90 | 9.35 |
| 通义千问 2.5-14B-1M | 8.73 | 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 |



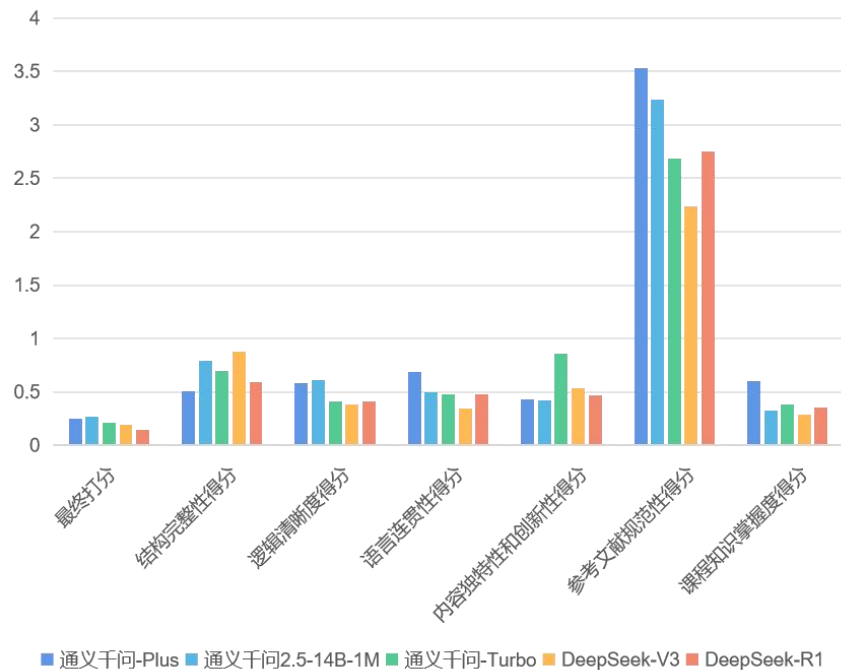


3.4.2 少样本提示语

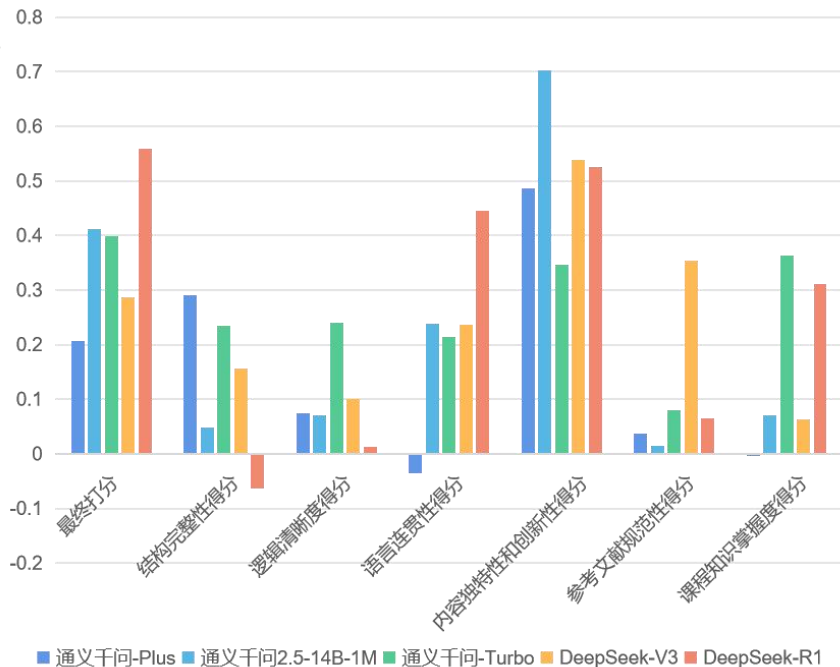
模型微调后各模型打分平均绝对误差(MAE)



模型微调后各模型打分均方误差(MSE)



模型微调后各模型打分皮尔逊相关系数(PCC)



3.5 智能评估系统

论文原文

分类号
U D C

编号
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

提交

综合得分
8.2



login

评语：

1. 结构完整性得分：9, 占比20%，原因如下：论文结构完整，包含引言、方法、实验、系统设计等标准章节，章节逻辑清晰且层次分明

2. 逻辑清晰度得分：8, 占比20%，原因如下：研究脉络清晰，问题-方法-验证的论证链条完整，但实验对比维度可进一步深化

3. 语言连贯性得分：8, 占比20%，原因如下：专业术语使用规范，学术表达准确，但部分长句存在冗余现象

4. 内容独特性和创新性得分：7.5, 占比20%，原因如下：多维度提示词设计与少样本微调方案具有创新性，但技术突破性需强化

修改意见：

建议核查参考文献时间线准确性，优化实验对照组设计，补充技术方案创新性论述，精简冗余表述提升可读性





3.5 智能评估系统

```
# other pre-defined variables
contents = None

class Report(BaseModel):...

# build API
app = FastAPI()

# set CORS
origins = [...]

app.add_middleware(
    CORSMiddleware,
    allow_origins=origins, # the list of origins that are allowed to make requests
    allow_credentials=True,
    allow_methods=["*"], # the list of HTTP methods that are allowed
    allow_headers=["*"], # the list of HTTP headers that are allowed
)
```

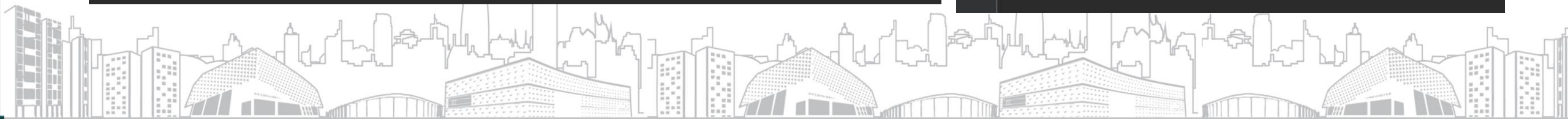
```
def prepare_prompt(count, text):...

def get_score(model_id, text, max_few_shot):...

@app.post("/Committing/")
async def score_report(report: Report):...

# Uploading Part
@app.post("/Uploading/")
async def update_report(file: UploadFile):...

if __name__ == '__main__':
    uvicorn.run(app, host="127.0.0.1", port=8000)
```



4

总结与展望



总结

1. 根据多维度评价指标设计提示词，使用不同大模型测试
2. 通过提示语微调对评估效果进行提升
3. 设计了可交互的网页系统

未来改进

1. 尝试不同模型、微调方式改善结果
2. 利用RAG技术，进一步完善流程
3. 完善网页系统，多文件并发请求



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.



参考文献

- [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.



参考文献

- [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.





感谢聆听

2025.5

