基于大模型的学生写作水平智能交互评估系统

王天瑞 王谦益 盛鹏

摘要

学生写作能力的培养是教育的重要目标之一，然而传统的写作评估方式存在诸多问题，如教师评估任务繁重、评估标准模糊、评估过程主观性强等。近年来，大语言模型（LLM）在自然语言处理领域的突破为智能写作评估提供了新的可能性。本文介绍了一种基于大模型的学生写作水平智能交互评估系统，该系统通过多维度评估写作内容并生成评语和建议，帮助学生提升写作能力，并减轻教师的评估任务。

关键词：大模型；辅助写作；多维度评估；文本总结；少样本训练

1 引言

学生写作评估是教学过程中不可或缺的环节，它可以帮助学生了解自身写作水平，发现写作中的问题，并针对性地进行改进。然而，传统的写作评估方式存在诸多问题。在传统写作中，教师的评估任务量大。随着学生人数的增加，教师需要花费大量时间进行写作批改，这增加了教师的工作负担，也影响了评估效率。[1] 其次评估标准模糊: 传统评估方式往往缺乏明确的评估标准，导致评估结果主观性强，难以保证评估的公平性和准确性。[2] 另外评估过程反馈有限: 传统评估方式往往只提供简单的分数或评语，缺乏对学生写作的深入分析和指导，难以帮助学生找到改进的方向。

近年来，大语言模型（LLM）在自然语言处理领域的突破为智能写作评估提供了新的可能性。[2]LLM 具有强大的语言理解和生成能力，可以自动分析学生写作内容，并提供个性化的反馈和建议。因此，基于 LLM 的学生写作水平智能交互评估系统有望解决传统评估方式的不足，提升评估效率和准确性，并为学生提供更有效的写作指导。[3, 4]

2 方法

2.1 模型结构

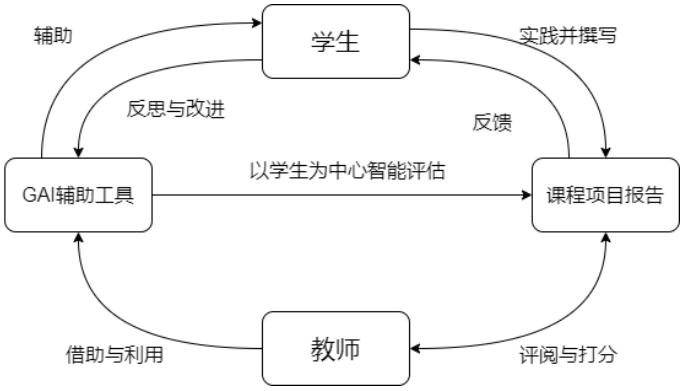

图 1: GAI 学生辅助系统流程

我们推出了一种基于大模型的学生写作辅助系统架构。该架构由学生、教师和由大模型组成的 GAI 辅助工具三个主要部分组成。在这套学习体系中，学生首先进行实践活动并撰写课程项目报告，然后接受教师的评审与打分，同时教师也利用 GAI 辅助工具为学生报告进行打分和评价，最后教师会综合自身和 GAI 工具的得分给出最终评估结果。学生会从最终结果中收到的反馈进行反思和改进，以提高自己的写作能力和水平。其中 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写作任务和提高写作质量。这种基于大模型的写作辅助系统能够有效地整合各方资源和技术优势，为学生们提供一个全面且高效的学习平台，帮助他们不断提升自身的综合素质和能力水平。

2.2 多维度评估方法

在评估学生报告时，我们不能直接让大模型为我们输出理想的结果，而需要为大模型提供一定的评估方法。我们希望评估方法能够细化且全面的评估学生报告，以给出综合且客观的评价。多维度的评估方法是一种综合性的评价方式，它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。我们提出的多维度评估方法包括以下几个方面：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。此外我们还会对参考文献规范性进行检查，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；以及我们希望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。

我们总共设计了 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，而参考文献和课程内容理解我们将其作为次要维度，在评分中占比较小。

2.3 提示词

好的大模型反馈离不开好的提示词的设计。[5] 为了让大模型反馈出更加符合我们需求的内容，我们需要设计提示词对大模型提出需求。在我们为学生报告评估的模型中，我们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；其次我们还会进行词汇解释，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，我们在提示词中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。[6] 经过设计后的提示词如下图：

你是一位教授人工智能导论课程的大学教师，需要对学生提交的课程项目报告进行评估。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：<>""
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：<>"

图 2: 提示词设计

2.4 系统交互模块

图 3: 系统交互界面

为了方便可视化操作，我们设计了系统交互界面。从图 3 中可以看到我们提供了一个直观的用户界面，用于连接大型语言模型（LLM）与学生写作评估功能。界面的左侧区域专门用于输入待评估的报告原文，同时也可以

直接传入文本文件。传入后大模型会根据预设好的提示词对文本内容进行分析。右侧则显示了综合评分以及多维度的详细评估结果。

3 实验

3.1 实验准备

3.1.1 数据集

我们使用的数据集来自课题组收集的计算机系创新实践报告，2022、2023 年人工智能导论课程报告和 2023 年医学人工智能导论课程报告。数据集包含不同年级、不同类型的写作内容。每篇报告都有教师人工打分，以及作者的年纪、专业等信息，较有利于我们后续对比模型结果和对写作者相关属性进行分析。

3.1.2 数据清洗

课程报告会包含很多不需要的冗余文件，例如代码文件、演示文件等。我们对数据集进行清洗后筛选出报告内容完整、文件类型较为统一的报告文件，并使用工具包（例如 PyPDF2）对报告进行文字提取。我们在人工审核校对的时候纠正了对于 PDF 文件页面分栏识别的问题，并且人工提取了每篇报告的参考文献部分避免影响模型对文本分析的效果。但是我们尚未对提取出的纯文本数据进行格式上的规整，还保留脚本提取后的分段、换行等格式。

3.2 多维度评估 + 提示词

我们使用设计好的提示词对大模型进行测试。在模型选择上我们使用文心一言 3.5 版本和智普清言 GLM-4，数据集选择的是 2022 年人工智能导论的报告文本。所有维度和总分的设计是 0 之 10 分，4 个主要维度各占比系数 0.2，次要维度各占比系数 0.1。我们取所有维度和总分的平均得分进行分析。

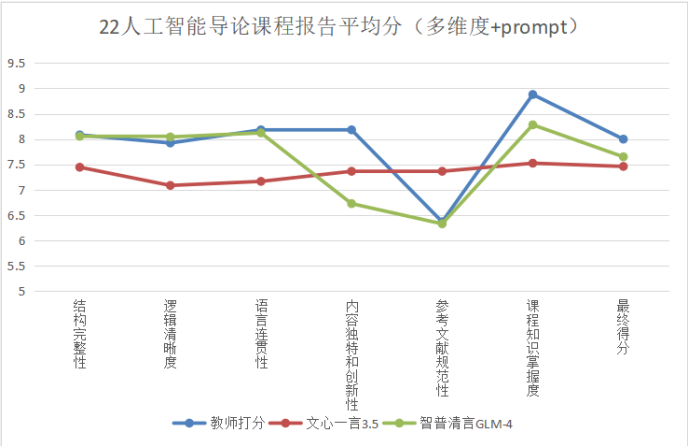

图 4: Enter Caption

对比教师的人工打分，我们发现两种大模型的打分都和人工打分接近。其中文心一言大模型的得分均值偏低，智普清言的得分均值更趋近教师人工打分。文心一言模型的得分波动较小，每个维度基本都维持在相似的范围内。相比较下智普清言的主要维度在内容创新和独创性维度上变化较大。文心一言模型的打分特性较为平均，各维度的得分波动不会太大，而且得分普遍偏低。智普清言的打分波动相对大，但是在趋势上更接近教师人工打分特性，并且得分更接近教师给分。

3.3 文本总结

从上述实验结果来看，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，我们推测是因为大语言模型对长文本的理解能力有限，可能无法全面把握长文本中的关键信息和上下文逻辑。因此，我们尝试使用文本摘要总结来优化模型对文本的理解和打分效果。我们共尝试了两类方法：一类是使用预训练的 Bart 模型生成摘要，通过其强大的生成能力对文本进行压缩和重组；另一类是结合向量数据库与 LangChain 框架，将长文本分片后进行嵌入生成和检索，通过语义匹配提取出最相关的信息进行总结。通过这两种方法，我们希望能够更好地提升模型打分与教师评分之间的一致性。

3.3.1 使用 Bart 模型生成文本摘要

对于英文和中文报告，我们分别采用了适合对应语言的预训练模型：

- 英文报告使用 bart-large-cnn 模型进行文本总结，利用其强大的生成式摘要能力。
- 中文报告使用 mbart-large-50-many-to-many-mmt 模型进行多语言摘要生成。

大致思路如下：
1. 文本分割：按照空行或换行符将报告文本分割成若干较小的段落，每段保持独立语义。
2. 段落摘要生成：对每个段落分别输入模型，生成对应的摘要，并提取其中的关键信息。
3. 摘要拼接：将所有段落的摘要拼接在一起，形成对整篇报告的完整总结文本。

但是实验结果并不理想，对于英文报告，模型生成的摘要会过分关注报告中的某几个部分，而忽略整体的总结性，并且缺乏对全局信息的均衡概括，导致生成摘要不够全面；而对于中文报告，生成的摘要中可能出现中英文混杂的情况，表现为部分内容以英文单词形式输出，此外，还会出现重复的异常符号，如多余的句号、标点符号错误等问题，影响阅读体验。

3.3.2 使用向量数据库 + LangChain 框架生成文本摘要

我们使用向量数据库 + LangChain 框架生成文本摘要，具体思路如下：
1. 文本分割
使用 $RecursiveCharacterTextSplitter$ 对文本进行切分，设置合适的 $chunk\_size$ 和 $chunk\_overlap$，确保每个分块既包含充分上下文信息，又能避免过多冗余。
2. 生成摘要
采用 $ZhipuAIGLM-4$ 模型对每个文本分块生成摘要，提取关键内容，使得每段摘要能够独立概括对应分块的信息。
3. 构建向量数据库
利用嵌入模型将文本分块转换为向量表示，并存储在向量数据库中。通过索引实现高效语义检索，方便后续的问答及验证操作。

4. 问答测试 & 结果保存

基于构建的向量数据库，进行问答测试，以验证模型生成摘要的准确性与全面性。将生成的摘要和问答结果保存到指定路径，便于后续分析。

通过以上操作，我们可以得到如下文本总结过程：

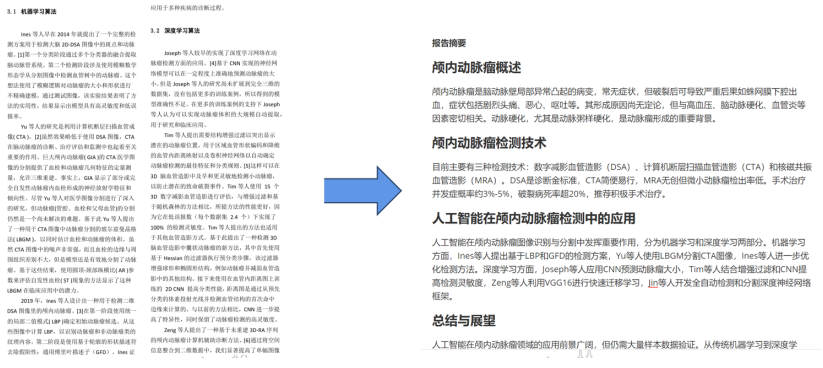

图 5: 文本总结过程

我们使用 22 年人工智能导论的 25 份报告进行了定量实验，进一步验证文本总结方法的可行性，结果如下：

人工打分：

结构完整性平均分	逻辑清晰度平均分	语言连贯性平均分	内容独特性与创新性平均分	参考文献规范性平均分	课程知识掌握度平均分	总分平均分	总分最高分	总分最低分
8.08	7.92	8.18	8.18	6.36	8.88	8.00	9.10	6.30

LLM打分：

	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	课程知识掌握度	总分
原文	7.44	7.08	7.16	7.36	7.36	7.52	7.456
总结	8	7.28	7.36	7.4	6.8	7.72	7.72

图 6: 文本总结实验结果

实验结果表明，通过文本总结后的报告，大语言模型能够更准确地理解报告的核心内容，其评分结果在除参考文献规范性外的所有维度上与人工评分更为一致。

3.4 少样本模型微调

在深入对比了包括百度千帆平台、智普平台等多个主流开源模型平台之后，我们发现，尽管这些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。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，我们决定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。

在市场上，百度千帆平台提供了丰富的模型种类和可调参数，训练过程中具有较高的灵活性和高效性。相较其他平台，百度千帆在模型精调时提供了更多可配置的超参数，以及更强的计算资源支持，这使得我们可以根据实际需求对模型进行个性化调整。因此，我们选择利用百度千帆大模型平台，基于其现有的强大模型进行精调训练，以期实现更好的结果。

3.4.1 数据处理

数据是模型训练的基础。在本次研究中，我们通过精心整合已有的教师评分数据，形成了一套符合训练要求的数据集。首先，我们对所有相关课程的数据进行整理，提取了每个学生的文本作业内容和教师评分，并将其转化为适合模型输入的格式。我们采用了百度千帆平台所要求的标准格式，即 JSON 文件格式，具体包括 ‘messages‘ 字段，每个字段包含 ‘role‘ 和 ‘content‘ 两个部分，分别表示系统消息、用户输入和模型输出。

```
{
    "messages": [
        {
            "role": "system",
            "content": "This is a system message."
        },
        {
            "role": "user",
            "content": "Hello, how are you?"
        },
        {
            "role": "assistant",
            "content": "I'm doing well, thank you!"
```

```
        }
    ]
}
```

在数据选择方面，除了 2022 年春季学期的“人工智能导论”课程数据外，我们还选用了其他相关课程的批改数据进行训练。具体来说，选用的训练数据如下：

- 2022 年春季学期创新实践数据 13 份
- 2023 年秋季学期医学人工智能导论数据 16 份
- 2023 年秋季学期人工智能导论数据 20 份

以上三类数据合计 49 份样本，虽然样本数相对较少，但这些数据能够代表不同课程的评分标准与作业内容。由于平台对数据量的要求为至少 100 条样本，因此我们将这些数据进行了三次上传，以达到最低数据要求，确保模型能够进行有效训练。

数据处理的过程中，我们特别注意到一些细节问题，尤其是如何准确地将评分和评语与文本作业内容进行有效配对。在这个过程中，我们对所有数据进行了严格的筛选，确保每个样本都包含教师的最终评分和详细评语。最终，我们将这些数据整合成符合训练要求的 JSON 格式文件，并上传至百度千帆平台进行训练。

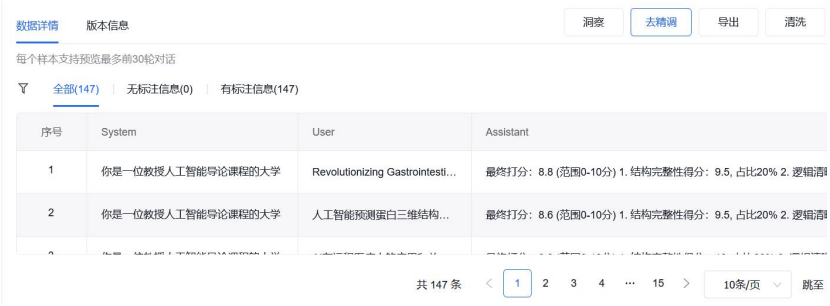

图 7: 模型精调数据集

3.4.2 模型训练

完成数据准备后，我们开始了模型的精调训练。此次训练基于百度千帆平台的 `ernie-3.5-128k` 模型进行。选择该模型是因为其在文本理解与生成

方面表现出了优秀的性能，且能够处理较大规模的语料数据。我们进行了监督微调（Supervised Fine-Tuning，SFT），即通过对已标注的样本数据进行学习，使模型能够生成与人工评分接近的结果。

在训练过程中，我们设定了三个不同的训练轮次，分别为 1 轮、3 轮和 5 轮。通过这样的设置，我们希望能够比较不同训练轮次对模型精度的影响。每一轮训练都会更新模型的参数，理论上，训练轮次越多，模型的性能会随着数据的迭代而不断优化。训练的关键参数包括学习率、batch size（每次更新的样本数量）以及最大训练轮次等。为了在保证高效性的同时避免过拟合，我们特别关注了这些参数的选择。

具体来说，我们使用了以下训练配置：

- 学习率：0.0003，设置较低的学习率，以确保模型在训练过程中能够稳定收敛。
- 序列长度：8192，序列长度为 4096 时发现因输入文本过长出现报错，改为最大限度 8192。
- 训练轮次：分别进行了 1 轮、3 轮和 5 轮训练，旨在评估不同轮次训练的效果差异。

训练完成后，模型的输出结果进行了详细的评估。我们通过平台提供的评估工具，对模型生成的文本进行了分类和统计分析，确保训练结果的质量和有效性。

图 8: 模型精调训练参数配置

图 9: 模型精调训练过程

3.4.3 精调结果与分析

经过训练后，我们将精调后的模型部署，并使用 2022 年春季学期的“人工智能导论”课程数据进行验证测试。我们将模型生成的评分结果分类存储，并对比了不同训练轮次（1 轮、3 轮、5 轮）模型与未经精调的原始模型之间的差异。通过对生成的文本结果进行分析，我们获得了如下几项结论：

- 模型能够较好地理解评分标准，并生成与人工评分相近的结果。尤其是经过 3 轮和 5 轮训练的模型，相较于原始模型，生成的评语更加贴近教师评分的实际标准。
- 在某些情况下，模型生成的评分结果仍缺乏足够的细节，尤其是评分理由和评语方面。这表明，在少样本训练的情况下，模型在生成具体细节上仍然存在一定的不足。
- 测试数据中有 3 份文本因过长，导致模型无法生成结果。经过多次尝试，模型依然未能成功处理这些过长文本，说明现有模型在处理长文本时存在一定的限制。
- 另外，2 份数据未能符合平台要求的系统消息格式，其中 1 份数据仅包含最终得分和评语，另一份数据缺少得分内容，仅有评价信息。这表明，在数据输入格式不规范时，模型的输出结果可能无法正确生成。

尽管存在上述问题，但我们依然从整体上看到了精调后模型的显著提升。特别是在评分一致性和准确性上，精调后的模型相比原始模型显示出了

更强的表现能力。在未来的工作中，我们计划进一步优化模型，尤其是在处理长文本和不规范数据格式方面。我们将尝试引入更多的训练数据，并调整训练参数，以进一步提高模型的泛化能力。

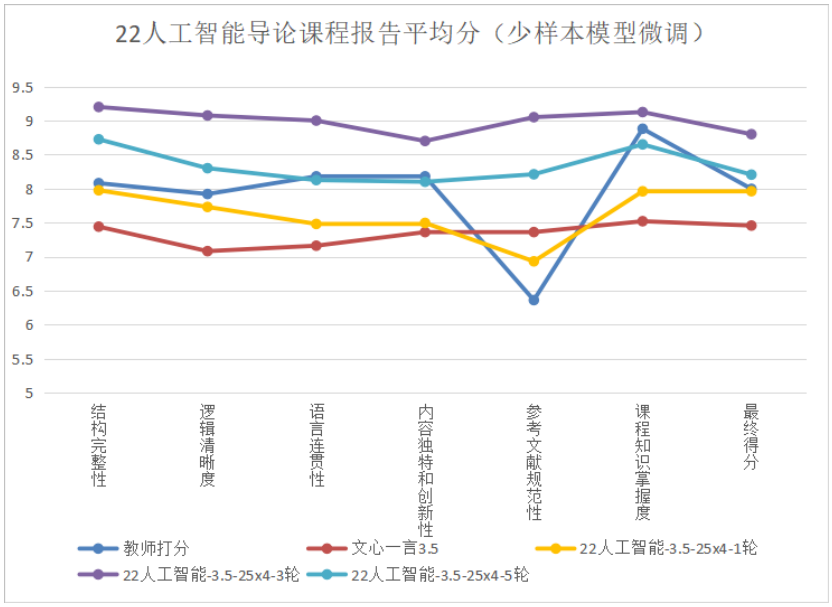

图 10: 精调模型的评分结果与原始模型对比

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，我们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。

3.5 综合对比

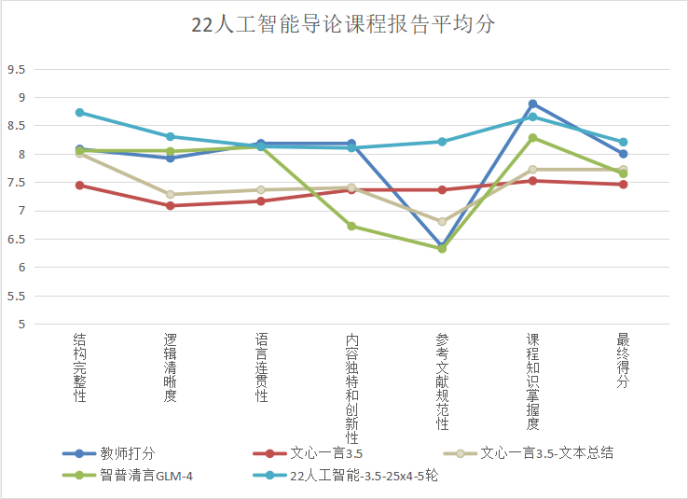

图 11: 三种方法综合对比

通过三种方法对比，我们发现经过在所有方法中少样本训练后的模型在评分上更加接近教师人工打分。但是少样本评分还是出现了同只用多维度 + 提示词方法中打分波动幅度小的问题，这点无法更好的拟合人工打分趋势。而文本总结后的得分虽然总体偏离人工打分多，但是在参考文献和课程知识掌握两个维度出现了更贴合教师打分的明显拐点。

综上所述，经过优化后的模型打分得到了不同程度的提升。文本总结后的大模型打分更加符合教师打分的变化趋势，而少样本微调后的模型在得分数值上更接近人工打分。

4 总结

本报告介绍了一种基于大模型的学生写作水平智能交互评估系统，该系统可以有效地评估学生写作水平，并提供个性化的评语和建议，帮助学生提升写作能力，并减轻教师的评估任务。未来，我们将继续改进系统，提升评估效果，并分析具体得分和学生属性的相关性，进一步完善网页系统，使其能够支持多文件并发请求。

本文介绍了一种基于大模型的学生写作水平智能交互评估系统。我们

提出了利用 GAI 辅助工具提升教师评价和学生写作水平的系统架构，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，我们尝试使用文本总结和少样本模型微调的方式。文本总结的实验结果表明，通过文本总结后的报告，大语言模型能够更准确地理解报告的核心内容，其评分结果在除参考文献规范性外的所有维度上与人工评分更为一致。而在模型微调的实验中，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，我们将继续改进系统，提升评估效果。我们会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。此外我们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。我们也会进一步完善网页系统，使其能够支持多文件并发请求，同时还有自定义的提示词设计。

参考文献

[1] 翟洁, 李艳豪, 李彬彬, 等. 基于大语言模型的个性化实验报告评语自动生成与应用 [J/OL]. 计算机工程,1-10[2024-11-09].https://doi.org/10.19678/j.issn.1000-3428.00EC0069593.

[2] 薛嗣媛, 周建设. 大语言模型在汉语写作智能评估中的应用研究 [J]. 昆明学院学报,2024,46(2):10-22. DOI:10.14091/j.cnki.kmxyxb.2024.02.002.

[3] Link, S., Mehrzad, M., Rahimi, M. (2022). Impact of automated writing evaluation on teacher feedback, student revision, and writing improvement. Computer Assisted Language Learning, 35(4), 605-634. doi:https://doi.org/10.1080/09588221.2020.1743323

[4] Pankiewicz, M., Baker, R. S. (2023). Large language models (GPT) for automating feedback on programming assignments. Ithaca: Retrieved from https://www.proquest.com/working-papers/large-language-models-gpt-automating-feedback-on/docview/2832896659/se-2

[5] A. Ramprasad and P. Sivakumar, "Context-Aware Summarization for PDF Documents using Large Language Models," 2024 International Conference on Expert Clouds and Applications (ICOECA), Bengaluru, India, 2024, pp. 186-191, doi: 10.1109/ICOECA62351.2024.00044.

[6] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., . . . Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. Ithaca: Retrieved from https://www.proquest.com/working-papers/prompt-pattern-catalog-enhance-engineering-with/docview/2779271809/se-2