基于大模型思维链的实验报告评语自动生成研究

Research on Automated Feedback Generation for Experimental Reports based on Large Model Chain of Thought Prompting

王萌[1*] 刘晓燕[1]
[1]江南大学，江苏“互联网+教育”研究基地
[*]wangmengly@163.com

【摘要】 在实验教学中，实验报告是评估学生实践能力与学习效果的重要手段。然而，依赖专业教师对大量的实验报告进行评价和反馈存在效率低、主观性强、评语模版化等问题。为此，文章针对《人工智能应用》课程提出了一种基于 ChatGPT API 和思维链提示的实验报告评语自动生成方法。首先对文本进行预处理与分割，然后通过 API 接口与模型进行交互，利用少样本思维链提示策略自动生成各模块评价结果，并整合生成具有社会情感的评语。结果表明，该方法生成的评语在可读性、相关性、准确性和个性化四个维度上都达到了较高的质量水平，且在除可读性外的其他方面均显著优于简单提示。

【关键词】 ChatGPT API；思维链提示；论文式实验报告；人工智能应用课程

Abstract: *Experimental reports serve as a critical tool for assessing students' practical skills and learning outcomes in laboratory teaching. However, relying on instructors to evaluate large volumes of reports faces challenges such as inefficiency, subjectivity and templated comments. To address these issues, this study proposes an automated comment generation method for experimental reports in the Artificial Intelligence Applications course, leveraging the ChatGPT API and chain-of-thought prompting. First, textual data is preprocessed and segmented. Then, the API interface interacts with the model to automatically generate module-specific evaluations using a few-shot chain-of-thought prompting strategy, which are subsequently synthesized into contextually empathetic comments. Results demonstrate that the generated comments achieve high quality across four dimensions: readability, relevance, accuracy, and personalization. they significantly outperform simple prompting approaches in all dimensions except readability.*

Keywords: ChatGPT API, chain of thought prompting, essay-style experimental report, artificial intelligence application course

1.引言

作为教育生态改革的重要助推器，人工智能技术在教育评价改革，实现规模化评估与个性化反馈的有机结合方面的研究已经引发了广泛关注。其中，实验教学作为教育评价改革的核心场域已在虚拟实验平台、数据采集与分析、实验报告在线生成等方面取得了显著进展。但面对大规模教学班的实验报告批阅过程中，智能化水平相对较低，面临效率瓶颈、反馈迟滞及评语同质化等局限。当前，以 ChatGPT 为代表的大语言模型（Large Language Models, LLMs）凭借出色的自然语言理解与生成能力，使其在教育评价中具备良好的应用潜力。该类模型遵循“预训练-提示”的范式，即先在海量文本数据上自监督预训练构建语言表征（Lee et al., 2023），继而通过动态提示机制实现任务适配。为此，本研究以华东某高校的《人工智能应用》课程为例，提出了基于 ChatGPT API 和思维链提示的实验报告自动评语生成框架。该课程是面向人文社科类专业开设的一门理论与实践相结合的课程，实践教学环节要求学生运用所学的人

工智能算法进行建模和解决实际问题，最终形成具有学术规范性的论文式实验报告，以培养学生的创新、问题解决及学术写作能力。

2.相关工作

2.1. 实验报告自动评价方法相关研究

从技术发展演进的角度，实验报告自动评价方法可以分为基于相似度、基于深度学习和基于 LLMs 的方法。(Chen et al., 2017)针对实验报告中的主观简答题，提出了基于标准答案的多级语义相似度自动评分算法，其评分结果与教师的平均差异仅为 0.46，但该方法无法适用于无参考答案的实验报告。(张迪，2022)通过引入 Real Former 和 Mean-Max-Pooled 方法优化 BERT 模型，实现了大学课程总结性文本的自动评分，准确率达到 79.17%。基于 LLMs 的方法是通过对已经在大规模数据集上预训练的语言模型进行微调，使模型适应评价任务。翟洁等(翟洁等人，2024)首次尝试了将 LLMs 应用于实验报告评语自动生成，通过匹配评估决策树自动生成报告和代码文本的评级结果及依据。但该研究仅从相关性角度对报告内容质量进行评价，且未对 LLMs 进行微调以适应特定的实验评价生成任务。

2.2. ChatGPT 在自然语言生成中的应用现状

评语生成作为自然语言生成（Natural Language Generation, NLG）的一个重要分支，专注于根据输入数据生成逻辑清晰、内容合理的反馈。目前，ChatGPT 在自然语言生成领域已证实具有强大的能力，不仅能从内容、语言、结构维度为作文生成高质量的反馈，还能有效的识别错误并提供具体的纠正策略。同时，ChatGPT 在学术论文的摘要生成、用语规范及创新性评估等方面提供了有力支持。此外在评语生成方面，（罗恒等，2024）将智能诊断数据与模型进行交互，生成数据驱动的个性化教师评语。（薛嗣媛和周建设，2024）从五个方面对模型生成的写作评语进行考量，发现思维链提示能够有效的优化教师撰写评语的过程。

3.基于少样本思维链提示的实验报告评语生成

3.1. 文本预处理与分割

由于模型输入长度受限于 token 数量，而学生的实验报告通常篇幅较长，字符数量远超模型输入上限。因此，本研究先提取实验报告中的文本数据，对其进行简单清洗，如去除多余的空格、换行符等。然后根据教师已规定的框架将报告分割为摘要、引言、实验设计、实验步骤、实验结果、实验总结和参考文献七个模块。其中，参考文献不涉及内容评价，其余六个模块以 JSON 格式存储，以便后续分模块评价。

3.2. 实验报告各模块自动评价

3.2.1. API 接口设置

本研究通过调用 API 接口实现与模型的交互。ChatGPT API 调用方法如图 1 所示。

```
import openai
openai.api_key = "sk-***"
def get_evaluation(user_text):
    response = openai.ChatCompletion.create(
        model="gpt-4",
        temperature=0,
        messages=[
            {"role": "system", "content": system_message()},
            {"role": "assistant", "content": assistant_message()},
            {"role": "user", "content": user_text}
        ]
    )
    return response['choices'][0]['message']['content']
```

图 1 ChatGPT API 调用方法

3.2.2 基于少样本思维链提示的各模块评价生成

提示词工程（Prompt Engineering）是一种用于优化与 LLMs 交互的方法，旨在通过精心设计的提示以引导模型生成所期望的高质量输出。Wei 等（Wei et al., 2022）提出了一种通过引

入思维链提示来激发 LLMs 推理能力的方法，并在不同任务上取得了优异的效果。受此启发，本研究构建了少样本思维链提示策略与 ChatGPT 进行交互，以完成实验报告评语生成任务。具体而言，该策略将评价任务分解为两个阶段：第一阶段为评价指标相关语句识别，即通过少样本示例引导模型提取出描述各评价指标的语句；第二阶段为评价结果生成，即引导模型根据评价标准对相关语句进行评价并输出结果。每个阶段都设置了三种角色，系统消息提供评价的具体规则指令；助手消息提供 3 个语句提取和结果输出示例；用户消息提供各模块具体的待评价内容。

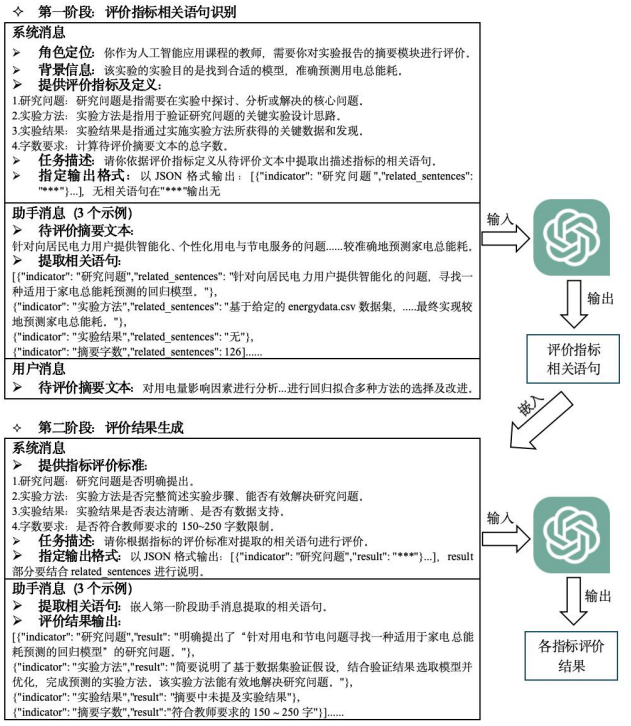

图 2 基于少样本思维链的摘要模块评价生成

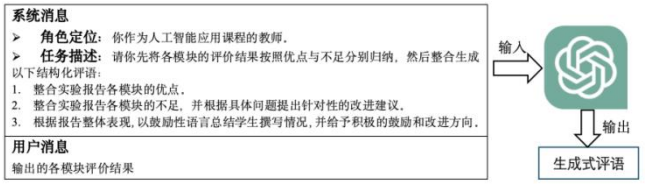

图 3 实验报告评语生成提示

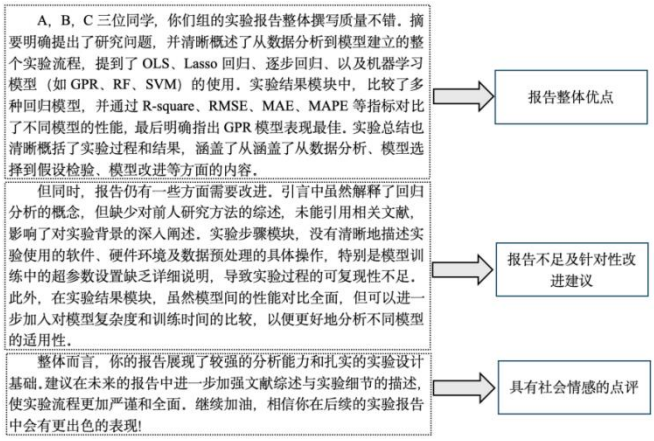

图 4 生成式评语示例

为确保评价过程的全面性和科学性，本研究明确了各模块的评价指标及具体标准。其中，摘要的评价指标包括研究问题、实验方法、实验结果和字数要求；引言包括实验背景和文献综述；实验设计重点评价数据预处理选择、模型选择和模型原理；实验步骤聚焦于实验配置、数据预处理、数据集划分及模型训练；实验结果包括评价指标、模型对比和结果分析；实验总结重点关注是否清晰表达了结果总结、实验局限及改进建议。这些指标的设定不仅为提示词的设计提供了明确的目标导向，也为模型在生成评语时提供了规范的评价依据。

以课程中的线性回归实验为例，给出了摘要模块的提示词，如图 2 所示。其他模块的评价都遵循此提示词设计框架，因此只需针对性的替换指标定义与评价标准、待评价文本、示例即可。

3.3. 实验报告评语生成

本研究的实验评语从各模块的评价结果中归纳总结出实验报告的整体优点和不足，并提出针对性的改进建议。同时，为了更接近人工教师的反馈方式，增加了社会情感的点评部分。为此，本研究设计了评语生成提示，如图 3 所示。在多轮对话优化与迭代下，最终的生成式评语示例如图 4 所示。

4.实验结果分析

为了评估上述生成式评语的质量，本研究选取了另一任务的 55 份实验报告作为实验集，让四名《人工智能应用》课程的任课教师分别从可读性、相关性、准确性、个性化 4 个维度对 ChatGPT 基于两种提示生成的评语进行打分，各个维度的评分规则如表 1 所示。

表 1 可读性、相关性、准确性、个性化的评分规则

指标	10-7 分	7-4 分	4-0 分
可读性	评语无语病、冗长，通俗易懂。表达自然同教师评语相媲美。	评语整体无语病，句子略有冗长。偶有表达不清晰和难以理解的情况。	评语存在多处语病，句子过于冗长导致重点不明，表达生涩影响学生理解。
相关性	评语与实验报告内容完全贴合，能够精准地指出报告的优点与不足，提供了具体且有建设性的反馈。	评语与实验报告内容基本贴合，能够指出报告的一些优点和不足，但可能较为笼统，缺乏具体细节。	评语与实验报告内容关联不大，对学生的表现不具备指导意义。
准确性	评语中不存在事实性错误，准确的反映了学生的报告内容。	评语存在少量表述偏差但未影响整体理解。	评语中存在严重的事实性错误，未正确反正报告内容，对学生产生了误导。
个性化	评语能够识别并评价学生的创新思路或面临的特殊挑战，提出了针对性的建议和鼓励。	评语在某些方面体现了对学生想法的关注，但整体仍有通用性。	评语明显模板化，未体现对学生想法的理解。

表 2 是教师对两种提示类型生成评语的打分情况。从表中可以看出，两种提示类型在可读性方面表现均较好，表明无论基于哪种类型 ChatGPT 生成的评语都清晰流畅。但思维链提示在相关性、准确性以及个性化指标上均显著优于简单提示，说明思维链能通过分步骤的逻辑推理，深入地理解报告内容，生成更贴近原文、针对性及个性化的评语。

表 2 教师对模型生成评语的评分

指标 / 提示类型	实验报告生成式评语的评分平均值			
	可读性	相关性	准确性	个性化
少样本+简单提示	9.18	8.22	7.03	7.71
少样本+思维链提示	**9.23**	**8.67**	**7.99**	**8.40**

5.结语

为了提高教师评阅实验报告的效率和反馈质量，本研究以《人工智能应用》课程为例，提出了基于 ChatGPT API 和思维链提示的实验报告评语自动生成方法。相较于简单提示来说，该方法生成的评语在相关性、准确性及个性化方面均表现出更高的质量和可靠性。

参考文献

罗恒,廖小芳,茹琦琦和王志锋.(2024).生成式人工智能支持的教师评语研究：基于初中数学课堂的实践探索.电化教育研究(05),58-66.doi:10.13811/j.cnki.eer.2024.05.008.

薛嗣媛和周建设.(2024).大语言模型在汉语写作智能评估中的应用研究.昆明学院学报(02),10-22.doi:10.14091/j.cnki.kmxyxb.2024.02.002.

张迪.(2022).基于深度学习的大学课程总结性文本自动评价关键技术研究(重庆理工大学).

翟洁,李艳豪,李彬彬和郭卫斌.基于大语言模型的个性化实验报告评语自动生成与应用.计算机工程 1-10.doi:10.19678/j.issn.1000-3428.00EC0069593.

Lee, G., Hartmann, V., Park, J., Papailiopoulos, D., & Lee, K. (2023). Prompted llms as chatbot modules for long open-domain conversation. arXiv preprint arXiv:2305.04533.

Chen, Y., Liu, X., Huo, P., Li, L., & Li, F. (2017, August). The design and implementation for automatic evaluation system of virtual experiment report. In 2017 12th International Conference on Computer Science and Education (ICCSE) (pp. 717-721).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, 24824-24837.