DOI:10.13811/j.cnki.eer.2024.05.008

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

罗 恒，廖小芳，茹琦琦，王志锋

（华中师范大学 人工智能教育学部，湖北 武汉 430070）

[摘 要] 教师评语是形成性评价分析的重要数据来源，但现有教师评语生成的质量在社会性、客观性和个性化方面存在一定的不足，生成式人工智能支持的教师评语为解决该问题提供了可能。生成式教师评语是指教师通过认知诊断技术得到的智能诊断数据与生成式人工智能平台进行交互，从而让生成式人工智能平台模拟教师社会性语言，生成数据驱动的个性化教师评语。在初中数学课堂中的实证研究发现：生成式教师评语有效地提高了学生的自我调节学习能力（Cohen's d=1.08，p=0.000）和学习动机（Cohen's d=0.59，p=0.003），对学生的深度访谈结果表明，生成式评语能作为考试的一种总结反思工具，体现了教师对学生个性化的关注和鼓励。研究结果为人工智能技术助力个性化、精准化教学提供了研究思路，为教育环境中持续评估和改进生成式人工智能技术的使用提供了建议。

[关键词] 生成式人工智能；教师评语；智能评测；个性化教学；初中数学

[中图分类号] G434 [文献标志码] A

[作者简介] 罗恒（1983—），男，湖北武汉人。副教授，博士，主要从事教学设计、数字化学习研究。E-mail：luoheng@mail.ccnu.edu.cn。

一、引 言

作为教育生态变革的重要驱动力，人工智能技术在教育评价改革，实现规模化教育与个性化培养的有机结合方面的实践应用广受关注[1]。其中，由机器学习模型提供支持的人工智能生成内容（Artificial Intelligence Generated Content，AIGC）能够自动化生成文本、图像、视频、音频等多模态数据，具备很强的对话情境理解能力和启发性内容生成能力。自 2022 年以来，生成式人工智能在文本生成方向的应用（例如，CTRL 模型、ERNIE-GEN 模型、GPT 模型）纷纷涌现，引发了教育内容生产方式的变革[2]。生成式人工智能采用预训练—提示（Pre-train-Prompt）的学习模式，借助概率对用户输入的文本上下文进行模式识别，根据句法规则生成文本内容，并结合用户输入的文本进行反馈强化学习，从而提供多轮次、流畅、自然的内容生成。生成式人工智能基于自然语言的对话能力，使其在教学评价与反馈中具备良好的应用潜力[3]。

教师评语作为教育评价的重要环节已经引发了越来越多的关注，评语通常是教师以描述性语言来概述学生学习情况，并以文本形式呈现给学生本人或其家长和教育行政部门[4]。中共中央、国务院于 2020 年 10 月印发了《深化新时代教育评价改革总体方案》，提出要“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”[5]。然而现有的教师评语反馈在社会性、客观性和个性化方面存在一定的局限性。首先，评语反馈的社会性不足，现有的评语生成系统生成的文本内容不能实现和学生的交互，不具备多轮对话的能力。其次，教师评价的来源多为教师的主观经验，缺乏客观性。最后，教师评语的撰写容易陷入抽象化、模式化、空洞化、教条化的误区，现有

基金项目：2021 年度国家自然科学基金面上项目“细粒度的智能学习诊断及其可解释性机制研究”（项目编号：62177022）

的教师评语生成的途径大多是通过评语语料库或者关键字生成的一些重复的短语,缺乏对学生学习过程和认知状态的准确描述。

生成式教师评语为这三个局限性提供了解决方案。生成式教师评语是指教师通过认知诊断技术得到的学习分析数据与生成式人工智能平台进行交互,从而让生成式人工智能平台模拟教师社会性语言,生成数据驱动的个性化教师评语。在生成式教师评语中,数据是核心指标,教师是设计者和监督者,学生是内容使用者。首先,在社会性方面,与其他自然语言处理模型相比,AIGC 具有良好的语言生成能力,能够更好地理解用户的问题,并且 AIGC 具有连续多轮对话的能力,能够在教师的监督下生成更加具有交互性的教师评语。其次,在客观性方面,AIGC 可以利用知识诊断数据实现对学习过程的科学描述,为学生提供数据驱动下更加准确客观的学习状态点评。最后,在个性化方面,AIGC 的内容生成能力可以根据学生的学习数据提出针对性的改进建议和错题练习,让学生得到更加精准化和个性化的评价反馈。

2022 年 11 月 30 日,OpenAI 发布了 ChatGPT,仅仅用了两个月的时间,成为史上用户增长速度最快的消费级应用程序。ChatGPT、文心一言和 Bard 等 AIGC 产品对教育领域的影响及应用引起了国内外学者的探讨与关注,但目前的研究大多是从 AIGC 的基本原理、智能涌现、能力边界以及工具价值等维度进行概念性论述[6-8],缺乏对 AIGC 在真实课堂中应用的实证研究。另外,现有的实证研究涉及的实践应用主要是为学生个性化自适应学习赋能,包括编程代码生成[9]、语言翻译[10]、课程知识问答[11]和适应性学习[12]等,需要进一步探讨 AIGC 在面对面课堂教学情境中的应用模式和路径。本研究使用的 AIGC 平台为基于飞桨深度学习平台和文心知识增强大模型而研发的文心大模型 4.0。

本研究在真实的初中数学课堂环境中进行,聚焦“如何通过 AIGC 支持的生成式教师评语促进数学学习”这一核心主题,探索人工智能技术支持下教师评语生成的新型方式,探讨数智化时代提供个性化、精准化的教学评价的有效途径,探究生成式教师评语对数学学习成绩、自我调节能力以及学习动机的影响。基于此,本研究提出以下四个研究问题:

1. 如何依托认知诊断模型和 AIGC 平台生成教师评语?

2. 生成式教师评语能否提高学习者的学习成绩,为什么?

3. 生成式教师评语能否发展学习者的自我调节学习能力,为什么?

4. 生成式教师评语能否促进学习者学习动机,为什么?

二、文献综述

(一)教师评语的数据基础

教师评语是教师对学生某一阶段发展状态的较为全面且富有个性的质性评价。Matsumura 等提出,从形式上评语可以分为认知特征类评语和情感特征类评语[13]。认知特征类是一些可采纳评语,如指出问题、提出建议、定位问题、给出解决办法;情感特征类主要指称赞、批评两类评语。

反映学生学习过程的数据是生成更加准确、客观的生成式教师评语的重要基础。传统的评语数据是教师根据主观经验生成的,随着计算机技术的发展,教师只要在评语生成程序内输入关键词,就可以利用评语生成系统得到模板化的教师评语数据。而人工智能技术的广泛应用使评语的数据生成更加智能和个性化[14]。本研究主要使用了认知诊断技术实现智能学习诊断,为评语的内容生成提供了数据基础。认知诊断模型在项目反应理论的基础上,基于学生的交互行为(如答题数据、测试数据)来挖掘学习者的潜在认知状态(知识点掌握程度和熟练程度),进而预测学习者在特定学习任务中的表现[15]。常见的认知诊断模型有 IRT 模型、DINA 模型和神经认知诊断模型。

(二)教师评语的内容生成

评语的内容生成是数据和社会性的语言整合的结果。评语是自然语言处理(Natural Language Processing)的子问题,自然语言理解和自然语言生成是评语生成过程中的重要组成部分。现有计算机生成的评语在语义表达的准确性、社会情感性等方面存在一定的局限性。让计算机更好地理解教师意图并根据学生数据信息来生成更加个性化和精准化的评语文本,是评语生成的技术难题之一。文心一言等 AIGC 产品的面世为个性化教师评语的生成提供了新的解决方案。

文心一言是对大语言模型(Large Language Model,简称 LLM)训练的结果,因此,对文本具有更强的理解能力,在教师评语生成过程中表现出良好的潜力。相比较于传统的聊天机器人,基于大规模语料库训练的文心一言能够结合少量的提示词,实现个性化数字资源高效创建、对话式人机协同学习、素质导向的教育评价[16]。基于文心一言的语言理解、对话交互、文本生成等方面的优势,文心一言在本研究中扮演教师评语内容生产者的角色。

三、研究方法

(一)参与者与研究情境

本研究在湖北省 W 市某中学采用准实验研究法开展了为期 6 周的教学实验,在真实的初中数学课堂中检验生成式教师评语的效果。本研究选择的两个班级的学生在学习成绩和师资配备上基本一致,其中一个班级为实验班级(七年级 18 班),另一个班级为对照班级(七年级 5 班)。剔除无效被试后(无效被试为没有填写问卷或不认真作答的学生),共有 117 名实验被试,其中,男生 62 人,女生 55 人,平均年龄 12.9 岁。

(二)教学干预:生成式教师评语

生成式教师评语的建构包括评语的数据生成和内容生成两个阶段。整体的生成路径如图 1 所示。首先,教师采集学生的个人信息、试题信息以及作答情况等测评数据,这些数据能够反映学习者的学习过程和认知状态。其次,利用本文作者自主搭建的智能学习诊断实验平台[17]对测评数据进行信息挖掘与建模。

在数据生成阶段,智能学习诊断平台使用神经认知诊断模型(Neural-cognitive Diagnostics),该模型综合考虑学生因素、题目因素以及它们之间的相互作用。智能学习诊断平台通过分析学生的答题记录,提取学生学习过程中的各类特征,诊断学生对知识点和六种初中数学核心素养(数学抽象、逻辑推理、数学建模、直观想象、数学运算和数据分析)的掌握情况,从而得到学习者错题和知识点的对应定位。具体的挖掘和分析过程如下:首先,用 $S=\{s_1,s_2,\cdots,s_N\}$ 表示学生集合,$E=\{e_1,e_2,\cdots,e_M\}$ 表示试题集合,并以人工标注的方式将试题所考察的知识点情况存入一个矩阵 Q,$Q\in\{0,1\}^{(M\times K)}$。其次,对于每个学生,将其编码成一个维度为学生总数 N 的学生 one-hot 向量 x^s,通过与一个可训练的学生知识掌握矩阵 A 进行乘积,得到该学生的知识掌握嵌入向量 h^s,其中 $h^s\in(0,1)^{(1,K)}$;而对于每道试题,先将试题 one-hot 向量 x^e 与 Q 相乘,得到每道题所对应的知识相关度向量 Q_e,接着构造可训练的矩阵 B 和 D,以同样的方式对试题知识点难度向量 h^{diff} 和试题区分度向量 h^{disc} 加以嵌入表征。得到学生和试题的向量表示后,构建交互函数并通过多层神经网络:

$$x=Q_e\circ(h^s-h^{diff})\times h^{disc}$$

$$f_1=\phi(W_1\times x^T+b_1)$$

$$f_2=\phi(W_2\times f_1^T+b_2)$$

$$y=\phi(W_3\times f_2^T+b_3)$$

其中,ϕ 是激活函数,此处使用 sigmoid 函数,训练以预测作答和实际作答结果的交叉熵作为神经网络的损失函数,训练结束后 h^s 即为学生对某个知识点的掌握情况诊断。最终,通过对测评数据的挖掘,本研究得到了学生的个人信息、知识点掌握情况、核心素养情况和错题定位等,这是教师与文心一言平台进行交互的数据基础。

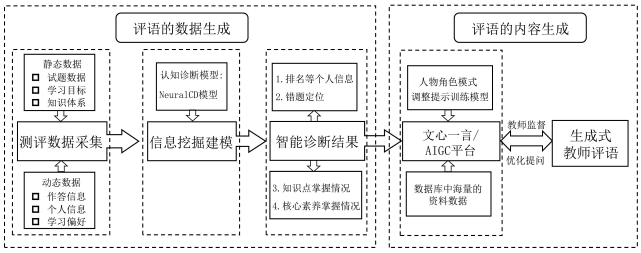

图 1　生成式教师评语的生成路径

在内容生成阶段,教师根据智能诊断结果对文心一言进行提问。教师与文心一言互动的质量取决于提问设计,如何向其提问成为获得有效反馈的关键[18]。本研究采用优化提问设计原则[19],优化提问指的是研究者在获得“初始提问”的反馈信息后,基于初始提问进行优化后的提问。整体提示编写框架参照 White[20] 提出的人物角色模式,具体提示编写框架见表 1。在教师的监督和多轮对话迭代下,最终生成的教师评语如图 2 所示。

表 1　提示编写框架

提示框架	解　释	示　例
背景和目的	阐述问题情境和目标	湖北省武汉市某中学进行了阶段测试
动机	说明解决此问题的重要性	教师评语将作为学生阶段测试的形成性评价结果之一
结构和关键点	指明模型充当的角色并呈现角色的输出内容	假设你是一名七年级数学教师
示例实施方式	编写清晰而具体的指令	根据“赵 *,本次阶段考试数学班级排名等级 C,和上次月考相比分数提高了 2 分,校排名提高了一个等级,同旁内角知识点掌握薄弱,逻辑推理的核心素养较弱”等数据信息,评语中需要包括个性化学习策略建议和社会性点评。请给她提供个性化的具有社会情感的教师评语
影响	审查结果的优势、适用性以及可能的改进。如果运行结果不理想,重新定义你的观点和提示	在进行评语生成时,如果成绩和排名等级发生变化时要进行推理可能的原因,并对于不同成绩的学生采用不同的学习策略建议

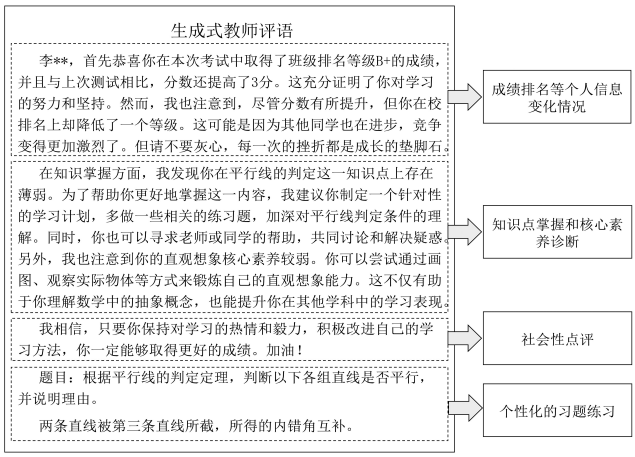

图 2　生成式教师评语示例

(三)研究过程

本研究采用准实验设计来检验生成式教师评语对学生数学学习成绩、自我调节学习和学习动机的影响。整体研究过程如图 3 所示。参与者在第一阶段和第三阶段分别进行了自我调节学习、学习动机的前后测。第二阶段,实验组和对照组同步学习七年级数学第五章,并进行了四次测试,以监测成绩变化。区别在于,对照组在每次测试后仅获得含有排名和分数变化情况的成绩单,教师会对班级整体考试情况进行口头反馈,而实验组除此之外,还会收到个性化的生成式教师评语。

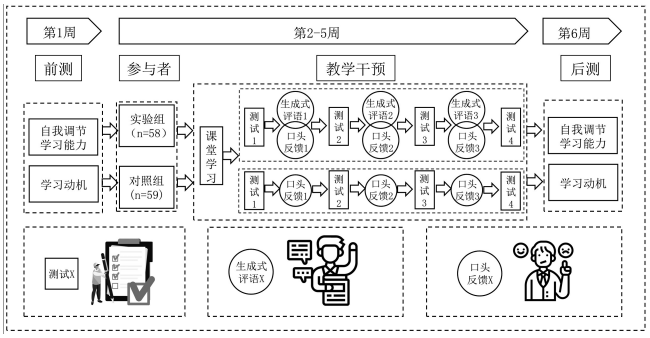

图 3　总体实验流程

(四)数据收集与分析

1. 数据收集

本研究收集了知识测验成绩、问卷数据和访谈数据。知识测验试卷由 10 年以上授课经验的教师开发,评估学生对数学知识的理解,涵盖选择题、填空题和解答题等 24 个题目,总分 100 分。学习动机问卷改编自 McAuley[21]的 IMI 动机量表,包括兴趣、能力、价值和压力四个维度的 11 个题项。自我调节学习问卷改编自 Barnard[22],包括目标设定、任务策略、时间管理、寻求帮助和自我评价等 13 个题项。问卷均以 5 分李克特量表来测量。共回收有效问卷 117 份,自我调节学习和学习动机问卷的 Cronbach's α 值分别为 0.898 和 0.885,内部一致性较高。此外,本研究还在实验组随机选择 6 名学生进行半结构化访谈,访谈围绕生成式教师评语的使用感受以及对学生可能产生的影响等七个方面展开,通过诸如“你能简要描述一下你对教师评语的使用感受吗”“教师评语中哪些设计和点评给你留下了深刻印象”以及“阶段测试后收到教师点评,你的学习有发生什么变化吗”等问题展开。每个人访谈时长为 5~8 分钟，录音后对访谈进行部分转录,生成 8131 个中文单词的文本内容进行定性分析。

2. 数据分析

在知识测验与问卷调查的定量数据分析中,本研究首先使用描述性统计分析来了解参与者的学习成绩、自我调节学习能力以及学习动机在均值上的差异。其次,在差异性分析方面,学习成绩和问卷数据的 K-S 检验结果均满足正态分布(p>0.05)和方差齐性检验。因此,本研究采用重复测量方差分析来分析知识测验数据,采用配对样本 T 检验来探究学习者在前测和后测时自我调节学习能力和学习动机的差异,采用独立样本 T 检验来探究两组学生在自我调节学习和学习动机的后测差异。本研究采用 IBM SPSS 21 进行统计分析。

在访谈数据方面,改编自 Chen 等[23]的编码方案,本研究从认知、元认知和情感三个方面对访谈记录进行定性分析。本研究遵循了 Braun 等[24]提出的主题分析程序,对文本内容进行了定位、识别和分类,以便进一步分析和主题生成。最后识别出 85 个代码,11 个节点,其中认知反馈 19 个节点、元认知反馈 45 个节点和情感反馈 21 个节点。编码主要由本研究的第一、第二作者使用 NVivo 12 进行分析，编码过程中出现的任何有争议的问题通过所有作者参加的每周会议解决。质性数据的分析在本文中主要用来支持对量化结果的三角互证和解释解读。

四、研究结果

(一)前测数据分析

由于本研究是在真实课堂中的准实验研究,为减少无关变量的影响,本研究采用独立样本 T 检验分析对照组与实验组的数学学习成绩、自我调节学习能力和学习动机的前测是否存在差异。结果显示,两个组初始成绩(测试 1)(p=0.898≥0.05)、自我调节学习能力前测（p=0.146≥0.05)和学习动机前测（p=0.056≥0.05)均无统计学差异。

(二)学习成绩差异

学生整体学习成绩变化如图 4 所示。实验组在收

到第一次和第二次生成式评语反馈后，学习成绩有一定的提升且与对照组有显著性的差异，但是实验组在收到第三次评语反馈后，学习成绩呈现下降的趋势，两组在第四次测试成绩上不存在显著性差异。

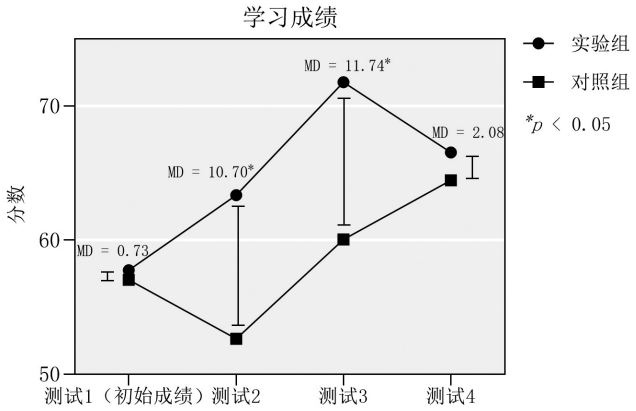

图 4　实验组和对照组学习成绩比较

为了进一步探究学习者的成绩变化与组别以及时间的关系，本研究采用重复测量方差分析评估实验组和对照组之间四次阶段测试学习成绩变化情况。经 Shapiro-Wilk 检验，各组数据服从正态分布且符合球形检验（p=0.144>0.05）。重复测量方差分析结果显示，组别的主效应不显著（F=1.45，p=0.23，ηp^2=0.012）；测量次数的主效应显著（F=29.17，p<0.001，ηp^2=0.202）；测量次数与组别的交互效应显著（F=11.10，p<0.001，ηp^2=0.088）。

（三）自我调节学习和学习动机差异

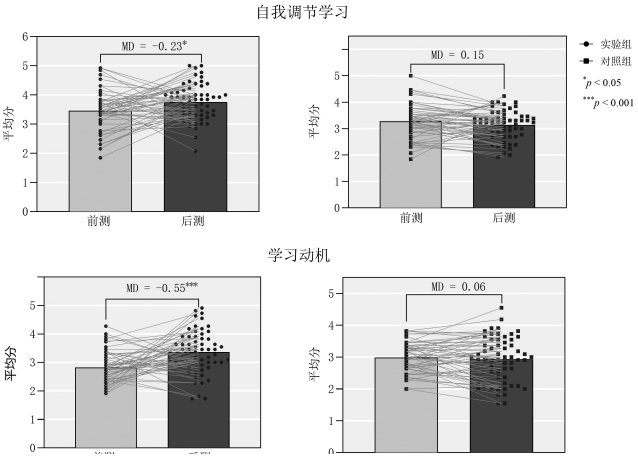

图 5　实验组和对照组自我调节学习和学习动机前后测比较

如图 5 所示，对照组的自我调节学习和学习动机在前后测中没有显著性差异。实验组在收到三次生成式评语反馈后，实验组的自我调节学习能力均值提高 0.23，标准差减少 0.06，说明实验组在收到三次生成式教师评语后，自我调节学习能力显著提高（p=0.014），整体数据更加集中。实验组的学习动机均值提高 0.55，标准差增加 0.21，结果表明实验组学习动机在实验期间得到了提高（p=0.000），但数据的波动性更大。

对实验组和对照组的后测成绩进行独立样本 T 检验，结果见表 2，实验组自我调节学习后测得分显著高于对照组（MD=0.63，Cohen's d=1.08，p=0.000），证明生成式教师评语能显著提升学生的自我调节学习能力。实验组学习动机后测得分高于对照组（MD=0.43，Cohen's d=0.59，p=0.003），表明生成式教师评语同样能显著增强学生的学习动机。

表 2　两组后测的自我调节学习和学习动机评估得分比较

类别	独立样本 T 检验					
	控制组（n=59）M ± SD	实验组（n=58）M ± SD	df	t	Cohen's d	p
自我调节学习	3.13±0.54	3.76±0.62	115	−5.83	1.08	0.000
学习动机	2.93±0.70	3.36±0.75	115	−3.19	0.59	0.003

五、研究讨论

（一）对研究问题的讨论

针对问题 1，生成式教师评语的生成包括数据生成和内容生成两个阶段，分别体现了个性化教师评语的数据性和社会交互性的原则。生成式教师评语可以通过测评数据采集、信息挖掘建模、诊断结果输入和智能评语生成四个步骤生成。首先，随着人工智能技术在教师评语中的应用，教师评语的生成来源应不仅仅局限于教师的主观经验，而应更多地反映学习者的学习过程和认知状态，使他们对自己的学业成就水平有一个客观的认识，实现数据驱动的精准化形成性评价。另外，具有社会情感的交互是教师评语的重要评价指标[25]，本研究要充分利用文心一言等 AIGC 应用在语义表达和语义理解上的优势，生成更便于学生理解并更能体现对学生发展性关注的评价。教师在整个评语过程中起到主导和监督作用，通过多轮对话迭代，让 AIGC 应用成为教师的智力支持和工作助手[26]，大幅减轻教师工作负担，助力精准化、个性化形成性评价。

针对问题 2，生成式教师评语对学习成绩的整体效果影响是一个有趣的发现。研究结果表明，生成式评语对学习成绩有一定的促进作用，访谈数据编码结果表明生成式教师评语中提供改进线索或试题练习对于提升学习者学习成绩至关重要。有受访者表示，相比较于知识点掌握情况诊断，他们对评语中提供的题目练习更感兴趣，练习题目的数量、质量以及答案

反馈也是影响生成式教师评语评价质量的关键因素。这个发现和 Hattie 等[27]的观点一致，他对 74 项教师评语研究进行元分析发现,最有效的评语形式是那些提供改进线索或强化策略给学习者的评语。在第一、二次评语反馈发放后,实验组的学习成绩提高,但是第三次评语反馈并未对实验组的学习成绩产生显著的影响。一种可能的解释是由于新奇效应,Huang[28]将新奇效应定义为对用户来说新的、不熟悉的、意想不到的体验。学生第一次和第二次收到生成式教师评语时,相比较于以前收到的笼统的、模板化的点评,生成式教师评语中含有学习者学习情况诊断、针对性错题练习,这会引发学生的好奇心和感知可用性。然而,新奇效应是短期的，随着生成式教师评语发放次数的增加,这种新奇效应会逐渐消失[29]。学生对评语内容的学习不仅仅靠自身的兴趣驱动,教师更要引导学生关注评语的内容。

针对问题 3,生成式教师评语发展了学习者的自我调节学习能力。生成式教师评语提供了学习者学习过程数据,反映了学习者的认知状态,使学生从被动接受评价结果转变为主动建构自我评估,这是自我调节的关键环节[30]。教师评语的反馈涉及自我调节学习过程的三个阶段的循环,在计划阶段,教师评语会引导学生制订学习计划，受访者提到,“我会根据老师的建议来计划我下一步的学习方向，以前都没注意到自己这个平行线的判定这个知识点掌握得不太好，我准备在课后多刷一些类似的题”。在表现阶段,学习者会用已有的知识经验和动机信念对评语内容进行解读,评语会帮助学习者对以上的认知参与过程进行自我监控,从而对学习进行针对性的调整。在自我反思阶段,学生会根据评语的内容评估自己的学习状况，并进行反思，决定如何改变自己的行为,“评语让我反思到自己在学习上面的不足，也能及时让我知道自己的进步和退步”。同时,这也体现了技术赋能教育的深层目标,即通过新兴信息技术，提升学习者的自我调节学习能力,实现学习者自主性学习、个性化学习的现实需求。

针对问题 4,生成式教师评语能有效促进学习者的学习动机。教师评语中的反馈内容会提供与学生自己的考试结果相关的信息,从而吸引学生的注意。根据动机设计的注意、关联、信心和满意度(ARCS)理论[31],注意和关联是诱导和维持学习动机的两个基本设计特征。除此之外,生成式教师评语中含有很多鼓励性的、表扬性的话语,有学习者提到“以前从来没有老师给我们发过这种,感觉很震撼,老师把你考得好的和考得不好的都写出来了,还给了我很多鼓励,感觉很惊喜吧”。教师积极的评价语言,使被评价者感受到了教师的肯定与认可,这种积极的情感反馈对学生学习体验和认知建构至关重要[32]。以 AIGC 为标志的生成式人工智能正在快速渗透到教与学的研究与实践中,人类教育所呼唤的个性化知识问答与人性化情感陪伴已成为现实。在数字技术赋能教育评价的过程中,教师需要在彰显教育教学过程中技术固有优势的前提下,更好地发挥主导作用,通过多轮对话交互,激发技术的情感、动机、态度、审美等育人属性,提升师生交互的亲密性和归属感,从而提升学习者的学习动机。

(二)教学启示

本研究结果对于生成式教师评语在教学实践中的应用有以下几点启示：

首先,本研究提供了一个生成式人工智能辅助教师教学的应用实例,教师应充分认识到教师评语的重要价值并利用文心一言等 AIGC 的工具属性来辅助教育评价的内容产出,使教师由主观经验式教学转向数智驱动式教学。教师在日常教学过程中,可以更加关注学习者的学习过程和认知状态,在评语中体现对学生发展性的关注,从而启发学生去思考自己的学习表现,更好地改进自己的学习。其次,由于文心一言等 AIGC 平台是基于对提示词及逻辑关系的匹配来生成答案,具有不理解语义和真实世界的技术局限,同时也存在输出信息不实、隐私和安全等问题,因此,教师提示词的编写和提问方式非常重要,建议教师在对文心一言进行提问时,编写清晰而具体的指令,同时给模型思考的时间,不断调整提示来训练模型,充分发挥教师的监督作用,以求达到更优的生成效果。最后,生成式教师评语的使用对象是学生,因此,教师需要合理地引导学生来使用生成式评语,带领学生认真阅读评语的内容,并且及时讲解评语中的个性化练习题目,实现迅速且高效的学习反馈。学生可以利用评语中的知识点和核心素养能力诊断情况来改进和调整自己的学习。

(三)研究局限性和未来展望

在未来，本研究的三个主要局限性应予以解决。首先,在研究设计上,目前评语数据生成的第一阶段仍然需要教师对试题信息、作答数据进行标注,这是一个相当烦琐的过程,希望后期能有更好的自动化采集学习者测评数据的方法。而且目前生成式评语的文本生成还是需要教师对文心一言等 AIGC 平台进行提问,后期建议实现评语生成的集成自动化,减少教师的工作量。其次,在生成式教师评语对成绩的影响上,本研究是针对初中数学课堂的为期六周的干预研

究。本研究发现个性化教师评语未能持续有效地提升学习者的学习成绩。因此,在后续的设计中,应该考虑研究延长实验周期,增加教师评语反馈次数,以观察教师评语对学习成绩的影响。后续的研究还可以深入探究对成绩的影响效果在学段、班级规模、学科等调节变量上的差异。最后,对于学习者学习动机和自我调节能力的测量,本研究仅采用问卷的方式来获得学习者的相关数据,问卷测量固有的局限性可能会损害统计结果的可信度。因此,建议未来的研究可以使用不同的研究工具收集学习者多模态学习数据,从而更加准确地测量学习者的学习动机和自我调节学习能力。

六、结　语

本研究为文心一言等 AIGC 在教育教学中的实践应用提供了实证证据,将人工智能技术和教师评语有机整合,创设了一种新型的数据驱动的、具有社会交互性的教师评语生成模式,构建了一种基于证据取向的评价模式。该评语生成模式兼具数据性和社会性的两大特点,实现了人工智能技术赋能精准化、个性化的教育评价。研究结果表明,利用人工智能技术得到的个性化教师评语能够在初中数学课堂中有效地进行评估和反馈,并提高学习者的数学学习动机和自我调节学习能力。

人工智能技术的赋能为智慧教育实践提供了强大支撑,当人工智能越来越多地参与到教育中,教师作为教育活动的主导者和技术赋能教育高质量发展的关键变量[33],应该在个性化学习、教师负担、教师自身成长等方面抓住生成式人工智能带来的机遇,在学习目的、教学过程和设计、评价方式等方面积极应对生成式人工智能带来的挑战[34],推动教学模式从“师—生”二元结构转变为“师—机—生”三元结构,在优化教学服务供给与学习需求匹配度、促进“师—机—生”协同与合作等方面提供支持,不断适应双脑协同、智力共生的学习评价新思维。

[参考文献]

[1] 吴砥,郭庆,吴龙凯,等.智能技术赋能教育评价改革[J].开放教育研究,2023,29(4):4-10.

[2] ZHAO W X, ZHOU K, LI J, et al. A survey of large language models[EB/OL]. arXiv preprint, (2023-11-24)[2024-04-29]. https://doi.org/10.48550/arXiv.2303.18223.

[3] 刘邦奇,聂小林,王士进,等.生成式人工智能与未来教育形态重塑:技术框架、能力特征及应用趋势[J].电化教育研究,2024,45(1):13-20.

[4] 郁晓华,战晓瑜.教师评语知多少?——探析文本后面的情感价值[J].电化教育研究,2022,43(7):97-105.

[5] 中共中央国务院. 深化新时代教育评价改革总体方案 [EB/OL].(2020-10-13)[2024-04-29]. https://www.gov.cn/zhengce/2020-10/13/content_5551032.htm.

[6] 刘明,吴忠明,廖剑,等.大语言模型的教育应用:原理、现状与挑战——从轻量级 BERT 到对话式 ChatGPT[J].现代教育技术,2023,33(8):19-28.

[7] 彭绍东.AIGC 时代基于双向赋能的人工智能教育创新框架[J].教育文化论坛,2023,15(4):12-26.

[8] 杨宗凯,王俊,吴砥,等.ChatGPT/生成式人工智能对教育的影响探析及应对策略[J].华东师范大学学报(教育科学版),2023,41(7):26-35.

[9] 孙丹,朱城聪,许作栋,等.基于生成式人工智能的大学生编程学习行为分析研究[J].电化教育研究,2024,45(3):113-120.

[10] LYU Q, TAN J, ZAPADKA M E, et al. Translating radiology reports into plain language using ChatGPT and GPT-4 with prompt learning: results, limitations, and potential[J]. Visual computing for industry, biomedicine, and art, 2023,6(1):1-9.

[11] 徐康,蒋凌云,黄海平,等.基于 ChatGPT 的留学生编译原理课程实践方法[J].软件导刊,2023,22(9):227-231.

[12] RAMAZAN Y, GIZEM F Y K. Augmented intelligence in programming learning: examining student views on the use of ChatGPT for programming learning[J]. Computers in human behavior: artificial humans,2023,1(2):100005.

[13] MATSUMURA L C, PATTHEY-CHAVEZ G G, VALDÉS R, et al. Teacher feedback, writing assignment quality, and third-grade students´ revision in lower-and higher-achieving urban schools[J]. The elementary school journal, 2002,103(1):3-25.

[14] 李宁.基于数据挖掘的自动评语生成方法的研究[D].天津:天津财经大学,2012.

[15] LEIGHTON J P, GIERL M J. Defining and evaluating models of cognition used in educational measurement to make inferences about examinees´ thinking processes[J]. Educational measurement: issues and practice, 2007,26(2):3-16.

[16] 卢宇,余京蕾,陈鹏鹤,等.生成式人工智能的教育应用与展望——以 ChatGPT 系统为例[J].中国远程教育,2023,43(4):24-31.

[17] MA L, ZHANG X, WANG Z, et al. Designing effective instructional feedback using a diagnostic and visualization system: evidence from a high school biology class[J]. Systems, 2023,11:364-377.
[18] LIU P, YUAN W, FU J, et al. Pre-train, prompt, and predict: a systematic survey of prompting methods in natural language processing[J]. ACM computing surveys, 2023,55(9):1-35.
[19] 王丽,李艳,陈新亚,等.ChatGPT 支持的学生论证内容评价与反馈——基于两种提问设计的实证比较[J].现代远程教育研究,2023,35(4):83-91.
[20] WHITE J, FU Q, HAYS S, et al. A prompt pattern catalog to enhance prompt engineering with chatGPT [EB/OL]. arXiv preprint, (2023-2-21)[2024-04-29]. https://doi.org/10.48550/arXiv.2302.11382
[21] MCAULEY E, DUNCAN T, TAMMEN V. Psychometric properties of the intrinsic motivation inventory in a competitive sport setting: a confirmatory factor analysis[J]. Research quarterly for exercise and sport, 1987,60:48-58.
[22] BARNARD L, PATON V, LAN W. Online self-regulatory learning behaviors as a mediator in the relationship between online course perceptions with achievement[J]. International review of research in open and distance learning,2008,9(2):1-11.
[23] CHENG K H, LIANG J C, TSAI C C. Examining the role of feedback messages in undergraduate students´ writing performance during an online peer assessment activity[J]. The internet and higher education, 2015,25:78-84.
[24] BRAUN V, CLARKE V. Using thematic analysis in psychology[J]. Qualitative research in psychology, 2006,3(2):77-101.
[25] 陈惠英.个性化评语的作用及撰写方法——以北京市普通高中学生综合素质评价中的教师评语为例[J].教育测量与评价(理论版),2011(10):32-35.
[26] 沈书生,祝智庭.ChatGPT 类产品:内在机制及其对学习评价的影响[J].中国远程教育,2023,43(4):8-15.
[27] HATTIE J, TIMPERLEY H. The power of feedback[J]. Review of educational research, 2007,77(1):81-112.
[28] HUANG M H. Designing website attributes to induce experiential encounters[J]. Computers in human behavior, 2003,19(4):425-442.
[29] SHIN G, FENG Y, JARRAHI M H, et al. Beyond novelty effect: a mixed-methods exploration into the motivation for long-term activity tracker use[J]. JAMIA open, 2019, 2(1):62-72.
[30] ZIMMERMAN B J, SCHUNK D H. Self-regulated Learning and performance: an introduction and an overview. Handbook of self-regulation of learning and performance [C]. New York: Routledge, 2011:1-12.
[31] LI K, KELLER J M. Use of the ARCS model in education: a literature review[J]. Computers & education, 2018,122:54-62.
[32] DUIJNHOUWER H. Feedback effects on students´ writing motivation, process, and performance[M]. Utrecht: Utrecht University,2010.
[33] 田小红,季益龙,周跃良.教师能力结构再造:教育数字化转型的关键支撑[J].华东师范大学学报(教育科学版),2023,41(3):91-100.
[34] 宋萑,林敏.ChatGPT/生成式人工智能时代下教师的工作变革:机遇、挑战与应对[J].华东师范大学学报(教育科学版),2023,41(7):78-90.

Generative AI-supported Teacher Comments: An Empirical Study Based on Junior High School Mathematics Classrooms

LUO Heng, LIAO Xiaofang, RU Qiqi, WANG Zhifeng
(Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan Hubei 430070)

[Abstract] Teacher comments are an important source of data for formative assessment. However, the quality of existing teacher comments has certain deficiencies in terms of sociality, objectivity and personalization, and generative AI-supported teacher comments provide the possibility of solving these problems. Since intelligent diagnostic data obtained by teachers through cognitive diagnostic technology interacts with the generative AI platform, the generative AI platform can simulate teachers´ social language and generate data-driven personalized teacher comments. An empirical study in a junior high school mathematics classrooms finds that generative teacher comments effectively improve students´ self-regulated

learning ability (Cohen´s d=1.08, p=0.000) and learning motivation (Cohen´s d=0.59, p=0.003). Through in-depth interviews with students, the results suggest that generative comments can be used as a summative and reflective tool for exams, reflecting the teacher´s personalized attention and encouragement to students. The results provide research ideas for AI technology to facilitate personalized and precise teaching and learning, and provide recommendations for continuous evaluation and improvement of the use of generative AI technology in educational settings.

[Keywords] Generative Artificial Intelligence; Teacher Comments; Intelligent Assessment; Personalized Teaching; Junior High School Mathematics

(上接第 50 页)

group awareness tools to promote self-regulated, peer-regulated and group-regulated learning, which are driven by individual, pedagogical and environmental factors. In this process, learners need to experience group awareness reception and understanding, including information viewing, information selection, value judgment, self-comparison, and meaning construction. In view of the above conclusions, the study gives the corresponding pedagogical enlightenments, including guiding learners to understand the content of group awareness tools and assisting learners to apply the group awareness tools for regulation.

[Keywords] Socially Regulated Learning; Group Awareness Tools; Mechanism; Grounded Theory; Qualitative Analysis

(上接第 57 页)

that each teaching unit has its own focus on the elements of computational thinking. Then, based on the nine-level classification system of knowledge structure, a multi-level progressive target system of computational thinking is formed, which is vertically refined layer by layer and is horizontally upgraded in the level of computational thinking (knowledge) structure. This design scheme can end the embarrassing situation between content teaching and computational thinking in information technology curriculum, so that computational thinking can be truly realized in the teaching approach that is integrated and unified with the content of the course.

[Keywords] Information Technology Curriculum; Computational Thinking; Target System; Multi-level Progressive System; Nine-level Classification System of Knowledge Structure