

融合虚拟现实和大语言模型的师范生讲课训练评价框架

杨 怡^{1,2,3} 张洪源^{1,2,3} 焦方旭^{1,2,3} 赵 珊^{1,2,3} 高丽金^{1,2,3*} 杨 扬^{1,2,3}
YANG Yi ZHANG Hongyuan JIAO Fangxu ZHAO Shan GAO Lijin YANG Yang

摘 要

师范生的教学实践对其教学能力的提升至关重要,因此,对师范生的教学实践表现进行科学合理的评价尤为关键,特别是对其讲课训练的评价。常用的教学实践评价如微格教学评价、档案袋评价、同伴和专家评价存在主观性强、反馈不及时等问题。尤其在边疆民族地区,由于教育资源分布不均衡、师资力量相对薄弱,这些方法存在评价标准具有差异性、评价反馈不及时、评价方法普适性低等不足。文章提出一种教学实践评价新范式:结合虚拟现实技术和大语言模型技术的讲课训练评价框架,依据教育学的基本原理,通过设计可计算模型、附加专业知识库、优化提示语三大策略,为师范生讲课训练提供科学、及时、针对性强的教学反馈。通过对 60 名师范生的实证测试,该框架与专家评价结果高度一致,并能提供更加个性化的教学建议。

关键词

师范生;讲课训练评价;虚拟现实;大语言模型

doi: 10.3969/j.issn.1672-9528.2026.05.040

0 引言

在师范生的培养过程中,对师范生的教学实践评价具有至关重要的意义,尤其是对师范生讲课训练的评价。讲课训练是师范生在教学实践活动中不可或缺的关键环节,不仅直接关系到师范生教学能力的提升和职业素养的养成,更是检验师范生培养成果的关键环节^[1]。通过科学、合理的讲课训练评价,能够深入了解师范生的专业素养、教学技巧及其对教育理论知识的运用能力,为其专业发展提供针对性的指导和建议,从而促进其成长为具备专业教学技能和高尚教育情怀的优秀教师,为我国教育事业的长远发展奠定坚实基础。

在边疆民族地区师范生的培养过程中,讲课训练评价是比较薄弱的一环。边疆民族地区教育资源有限且分布不均衡,导致师范生的讲课训练评价工作难以开展。并且由于

师资力量的相对薄弱、指导教师多为非师范专业的博士生,他们在为师范生的讲课训练提供专业的实践技能指导方面存在一定的局限性^[2],这样的教师差异化^[3]促使其探索评价和智能的深度融合。另外,在讲课训练评价工作中没有一套相对完整的评价体系和评价标准,这一环节的工作显得尤为不足,不仅影响师范生专业素养的全面提升,而且不利于推动教育扶贫工作的深入展开和智慧教育事业在边疆民族地区的发展。

鉴于此,本研究致力于探索一种教学实践评价新范式:基于虚拟现实(virtual reality, VR)和大语言模型(large language model, LLM)的师范生讲课训练评价框架,弥补边疆民族地区师范生讲课训练评价的不足,该框架还可以被系统性地扩展至师范生参与的教学实践活动中,包括但不限于教育见习、教育实习、模拟教学以及课堂教学竞赛等各个环节,以实现师范生教学能力全面而深入的评估。师范生教学实践评价的方法经历了从传统的、依赖于人工的同伴互评法、教师评价法、微格教学评价法^[4]和档案袋法^[5],到基于人脸识别技术的 S-T(Student-Teacher)分析法^[6]、基于弗兰德斯互动分析系统(flanders interaction analysis system, FIAS)的课堂分析法^[7]、基于运动追踪技术的评价方法和课堂智能视频分析技术的转变。这一演进过程标志着教学实践评价的持续进步和向智能化方向的显著发展。然而,这些方法都存在评价反馈不及时、评价标准不统一、评价结果过度量化、无法提供个性化教学建议等问题,对边疆民族地区师范生的教

1. 云南师范大学信息学院 云南昆明 650500

2. 云南师范大学模式识别与人工智能实验室 云南昆明 650500

3. 云南省智慧教育重点实验室(云南师范大学)

云南昆明 650500

[基金项目] 2024 年度民族教育信息化教育部重点实验室、云南省智慧教育重点实验室开放基金课题“基于生成式人工智能的智慧教学助手关键技术研究”(YNSE2024B002); 2024 年度民族教育信息化教育部重点实验室、云南省智慧教育重点实验室开放基金课题“基于生成式大模型的师范生基本技能自动评价研究”(YNSE2024C002)

学实践工作推进产生较大的阻力。

2024年3月28日,教育部发布4项行动助推人工智能赋能教育,提出将人工智能融入数字教育对外开放,推动大模型在课堂中的应用,并决定在西藏、青海、新疆等地开展国家中小学智慧教育平台的全域应用试点^[8]。随着生成式人工智能浪潮的兴起,大语言模型也在教学实践评价领域逐渐崭露头角。虽然大语言模型在自然语言处理中表现优异,但是在对专业知识的理解上不够深入,导致了模型幻觉的产生,使评价结果不准确。另外大语言模型是基于大规模语料库进行训练的,在真实的教学实践评价中如何构建一套大语言模型能够理解的提示语方案也是影响评价效果的重要因素。

为了解决目前在师范生教学实践评价工作中存在的问题,本研究提出一种融合大语言模型与虚拟现实技术的师范生讲课训练评价框架,并将该框架扩展应用于师范生教学实践评价中,以提升评价质量以及评价结果的客观性。该框架使用虚拟现实技术构建虚拟教学环境,以大语言模型作为评价主体。依据“教师应具备扎实的专业知识、掌握专业的教学技能、为人师表、以身作则”的教育学原理^[9]和《中学教育专业师范生教师职业能力标准(试行)》^[10],提出适配于此框架的讲课训练评价指标体系;实时分析师范生的讲课训练录音,并且生成个性化的教学评价和建议。

1 相关工作

1.1 基于人工的师范生教学实践评价

基于人工的师范生教学实践评价,是一种传统的评价手段,其通过人的主观判断对师范生课堂教学进行评估。主要方法有:微格教学评价、档案袋评价法、同伴互评法、教师评价法、自我评估法、访谈法及课堂观察法^[11]。基于人工的师范生教学实践评价方式结合了主观判断和能力考察,可以较好地把握实际教学过程的复杂性。但同时也有一定的局限,此类方法主观性强、反馈滞后、评价标准差异化,这些局限降低了评价的全面性和准确性。其有一定的优点,也存在一定的不足,将多种不同人工评价方式结合使用能够解决现实中的问题,但这也会导致主观性强、反馈延迟,而且师资水平的参差不齐会导致评价的效果不佳,无法实现常态化评价^[12]。

1.2 基于智能技术的师范生教学评价

为了克服传统的人工评价主观性强、反馈延迟、评价结果参差不齐的缺点,师范生的教学评价正逐渐引入人工智能、机器学习以及自然语言处理等技术,为师范生提供更加客观、高效和全面的评价结果。在当前的人工智能时代背景下,将促进师范生教学实践评价在评价主体、内容、方式、结果等方面变革,促进智慧评价的高效开展^[13]。陈增照等^[14]设计

了利用头部姿态估计和师生互动分析的教师注意力智能分析系统,通过分析教师注意力分配来进行教学评估。但是教师对不同学生、不同位置的注意力是不同的,仅依靠注意力的差异不能完整反映整个课堂教学过程的评价,存在一定的片面性。雨课堂^[15]是由清华大学研究团队开发的面向混合式教学场景的课堂系统,能为师生在课前、课中、课后三个阶段的表现提供数据支持和相应的帮助,通过记录和整理教学数据,能帮助量化教学成果。蔡慧英等^[16]利用课堂视频数据进行研究,深入理解人机协同下教师教学反思能力的发展,并提出依靠人工智能提升教师职业技能的建议。郑金存等^[17]提出基于KINECT运动跟踪技术的师范生授课技能评价系统,通过骨骼追踪和人脸识别,实时分析并绘制师范生的头部轨迹,使其快速了解自己的课堂状态,以便提高教学质量。这些技术的发展和运用,都为课堂评价的智能化和精准化奠定良好基础。

但也存在一定的局限,这类方法在处理复杂的教学场景和个性化需求时仍有限制,无法提供个性化的教学改进建议,高度量化的评价结果也无法提供实质性的帮助。所以,寻求新的师范生教学实践活动评价方式势在必行,提升评价的准确性和有效性至关重要。

1.3 基于大语言模型的师范生教学实践评价

大语言模型是人工智能和自然语言处理技术发展的最新成果,它体现了自然语言处理的优秀性能,而自然语言处理是教育互动的关键组成部分^[18],大语言模型也顺理成章成为推动智慧教育快速发展的催化剂。Long等^[19]首次使用大语言模型(generative pre-trained transformer-4, GPT-4)对课堂对话进行分析编码,并将其与人工标注结果进行对比研究,从而提高教学质量。龙禹辰等^[20]提出了一种基于多任务生成辅助的大语言模型情绪识别框架,目的是识别对话中的复杂情绪。方海光等^[21]利用课堂智能分析大模型对课堂教学行为进行智能分析判别,并构建了一个基于课堂智能分析大模型的教师教学能力分析框架,为教师教学能力的个性化培训与智能评估提供借鉴。Bewersdorff等^[22]探讨了大语言模型在自动识别学生实验协议中的逻辑错误方面的潜力,提出基于语言模型的教师教学反馈机制,展现了大语言模型在简化教师评估过程中的作用。褚乐阳等^[23]构建了大模型驱动的循证实践行动框架,开发了(generative pre-trained transformer 4 evidence-based practice, GPT-4EB)平台,为教师提供基于自然语言交互的个性化指导、反馈和评估。卢宇等^[24]构建的基于大语言模型的教学智能体提到智能体对项目作品生成相应的过程性与结果性评价。余胜泉等^[25]提出通过增强大模型构建通用人工智能教师架构,将通用人工智能教师应用于个性化指导与反馈场景,促进大模型在智慧教育领域的发展。

由此可见,大语言模型在教师课堂教学评价方面具有巨大的潜力,能够处理大量的课堂对话数据,提取有价值的教学特征,从而为教师提供反馈和建议,促进教学质量的提升。尽管将大语言模型应用于教育评价的研究取得了一定的进展,但仍存在一些不足之处。一是未根据大语言模型的文本处理能力和教育学原理,构建出一套适用于师范生讲课训练的内容和表达层面的评价指标体系;二是提示语设计有待优化,如文献[21]的评价标准下的提示语策略设计尚不完善,无法充分发挥大语言模型的作用。文献[23]在理论层面提出了通用人工智能教师架构,提供了一定的理论基础,但是并未付诸实践。而且由于训练数据的缺失或错误,导致大语言模型知识滞后现象的产生,大语言模型在专业知识领域的理解深度不高,阻碍了其在教学实践评价领域的广泛应用。与以往研究不同,本研究进入实践阶段,成功构建了一套师范生讲课训练评价框架,深入探究优化了师范生讲课训练评价方法,为师范生教学实践评价提供新途径。

2 师范生讲课训练评价框架的构建

本研究针对大语言模型在师范生讲课训练评价中存在的不足,包括缺乏针对内容和表达层面的评价标准、提示语设计策略的不完善,以及大模型对专业知识理解深度不足等问题,同时考虑到传统方法的主观性强、反馈延迟、评价主体能力参差不齐等局限性,提出并构建了一种基于虚拟现实技术和大语言模型的师范生讲课训练评价框架。通过三维元宇宙教学场景的搭建,不仅丰富了教学的内容和形式,还为讲课场景多元化提供了可能。同时,为了更好地实现课堂证据的采集,采用了流式语音识别技术,将讲课语音数据实时转录为文本,以提高转录的质量和速度。在使用大语言模型进行评价的过程中,针对教学过程中的教学内容以及表达逻辑两方面进行了评价标准设计,通过结合教育学原理的讲课训练评价标准的设计,使大语言模型提出更合理的教学建议。除上述工作外,为提高大语言模型的评价能力,运用设计可计算模型、附加专业知识库、优化提示语三大核心手段来提高评价性能,以解决其对专业知识评价不准确,评价标准和提示语策略不完善等问题,确保其在师范生讲课训练评价中的准确性和有效性。基于虚拟现实和大语言模型的师范生讲课训练评价框架的整体架构如图1所示。

2.1 测试数据

为了系统地验证本架构在讲课训练评价方面的性能,并为后续模型微调训练提供数据基础,构建了一个专用的数据集。该数据集包含了来自边疆民族地区某师范大学已经取得教师资格证的60名师范生的课堂教学实践录音。根据教学内容的不同,本研究将参与者分为三个平行组,每组20人。

教学内容包括信息技术课程的三个基础:二进制的概念、计算机网络按地理范围分类,以及计算机的类型和应用。针对每个主题的知识点准备了3份PPT资料,每份包含3页内容。每位学生根据所提供的PPT进行自由讲课训练,实时记录其教学语音,以此形成了本研究的测试基准数据集。因为每位学生只需围绕3页PPT内容讲解一个知识点,所以每段录音时长大约为5 min,确保了内容的简洁与重点突出。为了验证本研究提出的师范生讲课训练评价框架的实际效用,邀请了该师范大学在师范生教学实践评价领域的专家,对这60名师范生产生的讲课训练数据进行专业评价,并将其作为可对照的专家评价结果数据集,这不仅确保了该框架评价结果的客观性和权威性,也为评价框架提供实证检验,如图1所示。

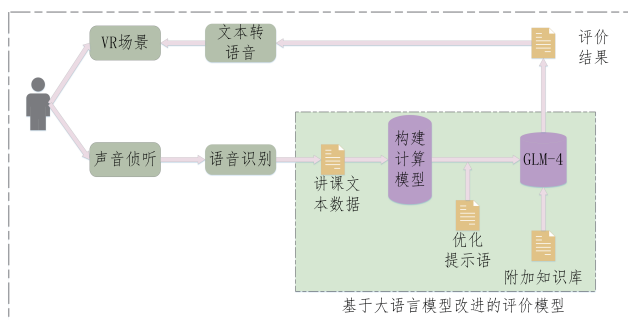


图1 讲课训练评价框架

2.2 搭建基于虚拟现实技术的交互式讲课训练评价场景

为了打破传统二维在线教学的技术发展瓶颈,构建视觉沉浸、听觉沉浸、交互沉浸的三维元宇宙教学实践场景,本研究将通过虚拟现实技术构建VR交互式讲课训练场景,实现讲课训练场景多元化,并将讲课训练评价框架应用于智慧教育领域,以提升智慧教育领域的科技创新水平,为边疆民族地区的师范生多元化教学实践提供可能,促进智慧教育高质量发展和教育公平。

该框架的三维场景导入和通信数据的处理在Unity3D引擎开发平台中编写脚本实现。在构建虚拟三维场景时,本研究选择HTC Vive系列头戴式显示器作为虚拟三维场景的展示平台。三维教室场景的部署采用了网络资源,并利用VIVE串流功能,通过有线或无线连接到PC端,实现了PC端场景的无缝传输至头戴式设备。这一技术允许用户在VR眼镜中体验到低延迟、高保真的虚拟三维场景。

由于本研究使用的评价主体是大语言模型,在使用此模型进行教学评价时只专注于对课堂上能够转录为文本的教学证据进行评价。通过语音识别技术将师范生讲课内容转录为文本,将文本输入模型进行整体分析评教。为了实现高效且实时的语音识别效果,借助(fundamental audio and speech recognition, FunASR)实时语音听写服务包处理流式音频数

据。采用此方法后，原本需耗时 1 min 进行语音识别的任务如今能够几乎实时完成。对于大约十分钟的课堂音频数据，转录时间从使用前的约 6 min 显著减少至几乎与录制同时完成，实现了真正的实时转录。实时语音转录技术的应用显著提高了本评价框架的运作效率，大幅缩短了师范生等待讲课训练评价结果的时间，从而有助于优化教学反馈的即时性。

在构建了交互式讲课训练评价的虚拟现实场景如图 2 所示，师范生通过头戴式显示器进入这一虚拟空间，从而获得一种沉浸式的教学体验。师范生能够利用语音指令激活讲课训练评价功能，进而开始讲课。在此过程中，Unity 平台将启动录音功能，同步进行语音的录制与转录，将师范生的讲授内容转换成文本形式。转录任务完成后，系统将通过调用大语言模型的 API（application programming interface）接口，将讲课内容传递至本研究提出的针对讲课范围和重难点的可量化计算模型，再基于语义进行相似度的计算。随后，计算模型的计算结果连同特定的提示语一同被输入至大语言模型中。待模型处理完毕，将给出相应的师范生讲课训练评价并提出相应的教学反馈建议。以上描述便是基于虚拟现实技术和大语言模型技术的交互式讲课训练评价框架的全流程操作。



图 2 虚拟教室场景

2.3 师范生讲课训练评价框架的基座模型

大语言模型和以往预训练模型的最显著区别之一就是模型所具备的涌现能力，主要表现在复杂推理、知识理解、交互和创造等方面^[26]。本研究是基于 GLM-4（Generative Language Model -4）模型构建的师范生讲课训练评价框架，GLM-4 是智谱 AI 开发的一种大型语言模型^[27]，是清华大学（knowledge engineering group, KEG）实验室提出的系列模型的一部分^[28]。其实施了多阶段的后训练过程，包括监督微调（supervised fine-tuning, SFT）和基于人类反馈的强化学习（reinforcement learning from human feedback, RLHF）。GLM-4 在多个通用指标上的表现与 OpenAI 公司的 GPT-4 非常接近，甚至在某些方面超过了 GPT-4。例如，

在指令遵循方面，GLM-4 的表现接近 GPT-4-Turbo；在处理长上下文任务方面，能够与 GPT-4 Turbo（128K）相媲美；在中文对齐方面，GLM-4 优于 GPT-4^[29]。GLM-4 的性能表现较为优秀，在此基础上构建师范生讲课训练评价框架能优化教学实践评价的效果，提高师范生教学实践能力。鉴于该语言模型具备处理大规模文本数据的能力，并具备强大的文本推理与生成技能，聚焦于师范生讲课的文本数据，针对这些数据设计了一套科学的评价指标，为模型的师范生讲课训练评价提供依据，从而提升其在师范生教学实践评价任务中的效能。

2.4 制定适用于大语言模型的讲课训练评价指标

科学合理的讲课训练评价指标是决定评价效果的关键因素，这些指标能够全面反映师范生的教学能力，并为师范生讲课训练提供明确的指导方向。边疆民族地区因教育资源分布不均、师资力量相对薄弱且水平不一，致使师范生的讲课训练评价工作发展存在阻碍，因而制定一套科学合理的讲课训练评价指标对于边疆民族地区师范生培养具有重要意义。因此，本研究基于大语言模型，依据教育学原理，设计了一套师范生讲课训练评价指标体系，弥补当前在利用语言模型进行师范生讲课训练评价时所面临的评价指标缺失的空白。

由于本研究使用的测试模型是 GLM-4，是一个大规模的语言训练模型，因此基于此模型研究合适的评价标准极其重要。致力于将教学过程中的语音数据转录为文本数据，传入 GLM-4 展开课程评价，所以评价指标并不需要完全覆盖师范生技能的方方面面。主要针对师范生培养过程中的专业能力和教学能力，从内容层面和表达层面进行评价。而对于一些无法转换成文本的教学证据，如讲课情感、讲课仪态姿态等方面的评价将是下一步的研究内容。

刘燕等^[30]将师范生教学实践评价分为两个模块，包括实践认识模块和理论素养模块，实践认识模块包括教师职业理解与认识、沟通与合作、反思与发展；理论素养模块包括一般教育学知识、学科教育学知识。它们是教师对教育学、教育心理学、学科知识、学生特征和学习背景的综合理解，是师范生能否成为一名合格的教师必须具备的理论基础。结合教育部印发的《中学教育专业师范生教师职业能力标准（试行）》^[10]中提到的四大能力：师德践行能力、教学实践能力、综合育人能力、自主发展能力，本研究提出一套适用于大语言模型的讲课训练评价指标，评价指标及其说明如表 1 所示。其中，针对讲课范围和教学重难点两个难以评估的评价指标，设计了可量化计算模型以辅助大语言模型在这两个层面提出科学合理的评价。

表 1 讲课训练评价指标

评价指标		说明
内容层面	讲课范围	确保师范生能够根据学科特点和学生需求,运用适宜的教学方法达成课程目标
	教学重难点	确保师范生教学聚焦关键知识,适应学生的个体差异,提升教学效率
	教学准确性	确保师范生教学内容的准确性,避免错误观念误导学生,影响学生成长
	例子使用	确保师范生讲课使用举例说明,辅助将抽象的理论具体化,激发学生的学习兴趣,提高教学质量
表达层面	语言逻辑性	确保师范生讲课语言具有逻辑性,帮助学生清晰掌握知识点,形成系统化知识结构
	语言连贯性	确保师范生讲课语言具有连贯性,能够清晰、有逻辑地表达思想,有助于高效利用课堂时间
	违规词使用	确保师范生讲课不使用敏感词汇,引导学生树立正确的人生观、价值观
	语言层次性	确保师范生讲课语言具有层次性,能够由浅入深,逐步引导学生理解和掌握知识

3 提高模型评价性能的策略

3.1 构建讲课范围和重难点的可量化计算模型

由于该框架的评价证据是师范生讲课语音转录的文本数据,因此本框架的讲课训练评价只针对文本数据可达的内容层面和表达层面。在设计此框架时,一个核心挑战是将评价任务委托给一个缺乏人类主观意识的主体——大语言模型。关键在于如何设计模型使其能够准确地识别教学内容的相关范围和核心概念,并对师范生的讲课进行客观评价。为了实现这一目标,研究了基于大语言模型的内容判断的可计算模型,该模型通过计算文本间的语义相似度,作为辅助信息输入评价模型,以提高判断准确性。

鉴于中文语言的独特性,仅仅依赖基于句子的相似度计算往往不够精确,因为常会遇到句子结构不同但含义相同

的情况。因此,结合基于语义的相似度计算方法显得更为合理。在计算模型中,采用了两个层次的相似度计算策略:第一层专注于语义层面的相似度分析,而第二层则专注于文本层面。通过这种双层结构的计算,能够显著提升相似度计算的准确度,从而更准确地捕捉到文本之间的内在联系。计算模型的工作流程图如图3所示。通过语音转录技术,将师范生的语音内容转化为文本,并将这些文本内容输入 GLM-4 中。模型随后执行关键词和主旨句的提取任务,将提取到的核心信息整理并放入一个集合中。然后进行第一层语义相似度的计算。

阿里巴巴达摩院开发的自然语言处理预训练模型 StructBERT^[31]通过在预训练过程中引入结构化知识,具备了强大的上下文感知能力,这使得其在语义相似度计算方面表现出色。针对教学重难点,先让模型提取教学重难点的关键词,组成集合 A , 专家评判的关键词组成集合 B , 调用 StructBERT 的 API 接口,对集合 A 和 B 进行第一层语义相似度的计算。如果语义相似度超过预设的阈值 0.5,便认定均在语义上相近,并使用集合 B 中的对应元素替换集合 A 的元素,从而简化后续第二层句子相似度的计算。在计算第二层文本相似度时,采用 Jaccard 系数和 Dice 系数,其基于两个集合的交集和并集来度量集合之间的相似性。计算公式如式(1)~(2)所示。针对教学范围,集合 A 和集合 B 分别代表模型和专家判定的主旨句,其余计算方法相似。一旦计算出两个集合间的相似度,将这一结果以及相应的提示语一同发送给 GLM-4 模型,为模型提供教学范围和教学重难点的量化参考,辅助模型在师范生讲课训练评价过程中对教学范围和教学重难点做出更为精准的评估。

其中,集合 A 是经过第一层语义相似度计算更新后的关键词集合(主旨句集合),集合 B 是由专家评判的关键词(主旨句)构成。 $A \cup B$ 表示两个集合的并集数量, $A \cap B$ 表示两个集合交集的数量, $|A|$ 和 $|B|$ 分别表示集合元素数量。

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

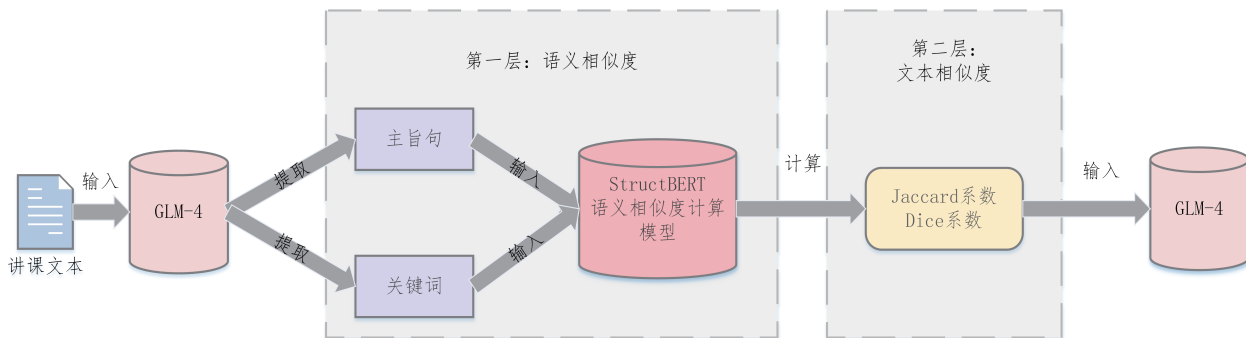


图 3 计算模型工作流程

$$\text{Dice}(A,B)=\frac{2|A\cap B|}{|A|+|B|}$$

(2)

3.2 优化基于大语言模型的讲课训练评价提示语

在使用大语言模型进行师范生讲课训练评价的实验过程中，研究发现存在提示语策略不完善的问题，而精准打磨合适的提示语，使模型在得出推理结果前先形成思维逻辑^[32]，对确保模型给出合理的教学改进意见具有重要作用。本研究经过多轮测试优化，设计了一套模型容易理解、回复格式整齐的提示语策略。提示语优化过程如图4所示。本研究最先采用后缀式提示语进行教学评价，融入讲课训练评价指标，目的是引导模型根据教学内容和评价指标提出改进建议。实验结果显示，该方法生成的评价效果不佳。所以本研究调整了提示语策略，使用后缀提示语为模型提供课程重点知识背景，引导模型在进行师范生讲课训练评价时参考所提供的背景知识。测试结果表明，模型对一些专业错误依然无法识别。

为解决在专业知识领域适配性不足的问题，本研究在GLM-4的基础模型上引入专业知识库，并减少了提示语中背景知识的部分。除此之外，给大语言模型设定一个“资深的教学评课专家”的身份，增强模型的角色认同感和代入感，提高评价结果的准确性和专业性。在多次试验中发现，在该版提示语上的师范生讲课训练评价，有时模型不能充分关注所有评价指标，有遗漏某一指标的情况发生。为了解决此问题，本研究设定了一套标准化的固定回复格式，提升评价结果的规范性和可读性。经过多轮反复提示语优化后，模型回复已基本符合研究需求。同时随着优化的深入，提示语的长度不断增加，导致输入文本过长产生不良评价效果。经过探究测试发现两种策略可以减少长文本影响：一种是采用成对的标点符号来分隔提示语和讲课内容，确保模型能够清晰地区分二者的分界；另一种策略是引入多轮对话机制，通过分步骤地提供提示语，逐渐引导模型生成更精准的回复。但为了实现评价反馈及时的目标，减少评价过程的时间消耗，本研究选择第一种策略使用成对标点符号来分隔提示语和讲课内容。通过评价指标的嵌入、为模型设定特定角色、设定固定的回复格式、使用成对标点符号分隔等一系列优化方法，最终GLM-4对师范生讲课训练评价取得了更优的效果，为师范生讲课训练提供了科学合理的教学反馈建议。

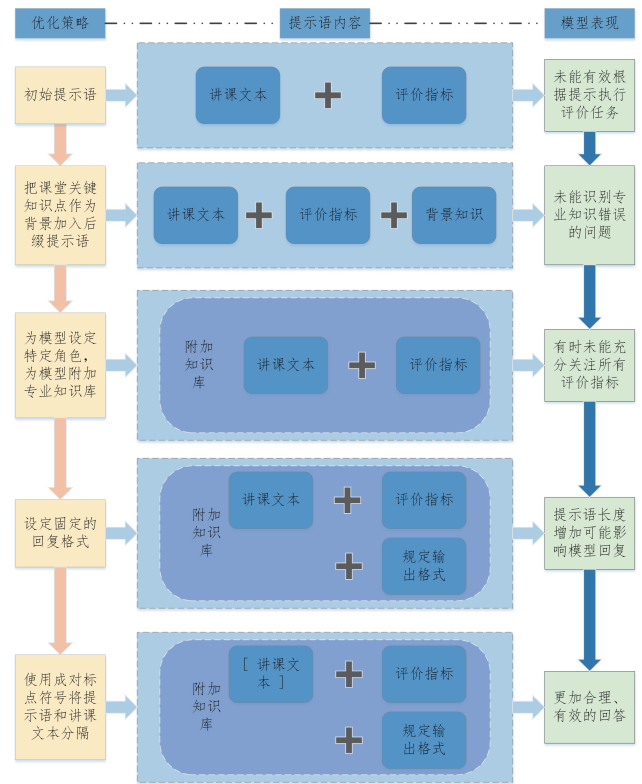


图4 提示语优化路线

3.3 外接具有专业课程知识的知识库

在测试过程中，观察到模型无法有效识别出一些专业性较强的讲解错误。尽管此前在提示语中集成了背景知识模块，以期通过引入相关专业知识来增强模型的判断能力，但是模型表现不佳。鉴于此，推测有两个原因导致此问题的产生：一是由于背景知识的加入导致提示语长度增加，从而影响了模型对长文本的处理能力；二是由于教学领域训练数据不足，导致模型对教学领域专业知识理解深度不够。为了解决这一问题，采用基于LangChain构建知识库的方法，构建知识库技术如图5所示，通过整合外部知识库，为大语言模型注入了丰富的教学领域专业知识，实现了大语言模型

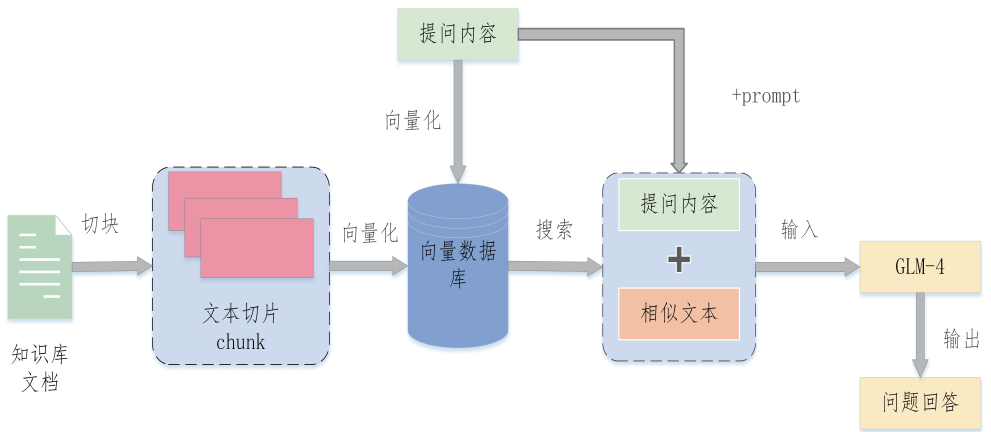


图5 外接知识库技术路线

与特定专业领域知识的深度融合，以此提升模型的专业知识储备，解决模型在识别专业错误方面的不足。

经过几轮实验测试，发现通过构建知识库并将提示语中的背景知识部分移除，不仅有效缩短了提示语的长度，减轻了长文本对模型性能的影响，而且显著提高了模型在识别师范生讲课过程中专业错误的能力。

4 师范生讲课训练评价框架的效果

4.1 外接专业知识库的前后效果对比分析

表 2 列出了加知识库前后模型评价与专家评价的结果，由表可以看出，知识库的构建显著提高了模型识别专业性较强的教学学科知识的能力，成功解决了之前大语言模型无法识别专业性错误的问题。例如，在讲课文本中，存在一个明显的错误：“2 的 8 次方等于 265 种状态”。在未添加知识库之前，模型无法识别此类专业性错误。然而，在添加了知识库之后，模型能够准确识别并指出此类错误。更多类似的例子在表 2 中详细列出。

4.2 添加可计算模型的前后效果分析

为直观评估在该框架中引入可计算模型前后的效果，本文采用 BERTScore^[33] 对模型评价结果的质量进行量化分析。BERTScore 的核心思想是利用（bidirectional encoder representations from transformers, BERT）模型的预训练上下文嵌入来计算候选句子和参考句子之间的相似度。由于可计算模型的引入是为了提升模型在讲课范围与教学重难点两个评价指标上的表现，因此实验的 BERTScore 得分聚焦于这两个维度进行评估。实验结果如表 3 所示，其中，精确率（Precision, P ）

表 3 加计算模型前后模型评价质量量化分析

课程主题	计算模型	讲课范围			教学重难点		
		P	R	F_1	P	R	F_1
二进制的概念	未加	0.671	0.669	0.670	0.660	0.659	0.659
	已加	0.707	0.698	0.701	0.697	0.727	0.711
计算机的类型和应用	未加	0.680	0.729	0.702	0.682	0.695	0.687
	已加	0.720	0.730	0.723	0.709	0.728	0.718
计算机网络按地理范围分类	未加	0.684	0.702	0.692	0.671	0.693	0.682
	已加	0.740	0.739	0.739	0.696	0.690	0.692

表示模型生成的评价结果中，与专家评价结果相似的部分所占的比例；召回率（Recall, R ）表示专家评价结果中，被模型生成的评价结果所覆盖的部分所占的比例； F_1 分数（ F_1 -score）是精确率和召回率的调和平均数，表示模型生成的评价结果与专家评价结果之间的综合相似度。结果显示，在引入可计算模型后， P 、 R 和 F_1 分数均有不同程度的提升，表明模型在讲课范围和教学重难点两个指标上的评价结果与专家评价的相似性得到了显著增强。

4.3 模型评价和专家评价效果对比分析

为了评估该框架下大语言模型的评价效能，本研究将大模型的评价结果与专家评价结果进行了对比，通过人工统计使用频次来统计模型评价和专家评价之间的一致性。表 4 列出了模型评价结果与专家评价结果在 60 个测试样本下的一致频次统计。由表 4 一致频次统计表可以看出，在各样本的测试结果中，对于不同的评价指标模型与专家一致的频次较高，表明模型评价结果与专家评价结果相似性较高。

表 2 加知识库前后的效果对比

未加知识库的模型评价	加知识库的模型评价	专家评价
内容准确性较高：教师对二进制的定义、规则和应用进行了详细的解释，未发现明显错误或误导信息。	内容准确性方面，大部分内容正确，但存在一些口误和专业错误，例如“二进制的数是 110.01”在转换成十进制时的解释有误，位权计算不正确。	讲课内容少部分不准确、有错误，二进制转换成十进制时，很明显地把二进制数“110.01”读成十进制数。
内容准确性：本节课的内容表述准确无误，没有发现错误的信息或不准确的描述。	讲课中存在一些口误和专业错误，如“比如现在有一个 8 位的二进制数，那它就可以表示二的 8 次方，然后等于 265 种不同的状态。”应为“二的 8 次方等于 256 种不同的状态”。	讲课内容少部分不准确、有错误，2 的 8 次方应该是 256，而不是 265。
内容准确性：讲课内容对二进制的定义、原理以及优缺点的描述准确无误，没有发现错误或不准确的表述。	教师在讲解短除法时的确存在错误，主要是关于商的计算和描述不够准确。正确的短除法步骤应该是：将十进制数除以 2，得到商和余数，然后将商继续除以 2，再次得到商和余数，重复这个过程，直到商为 0。	讲课内容少部分不准确、有错误，十进制转换成二进制时，很明显地把余数与商的位置混淆了。

表 4 模型评价和专家评价一致频次统计

评价指标	一致频次
讲课范围	56
教学重难点	56
教学准确性	57
违规词使用	60
语言逻辑性	58
语言连贯性	59
例子使用	58
语言层次性	58

在测试阶段，本研究以初高中信息技术课程为基准，选取了“二进制的概念”“计算机的类型和应用”以及“计算机网络按地理范围分类”三个主题课程进行效果评估。本研究使用 SnowNLP^[34] 库对上文提出的八条评价指标进行量化处理，通过计算每条评价指标的情感得分，判断模型评价与专家评价的一致性。如图 6 中 A、B、C 图所示，1~8 分别为上文提出的 8 个评价指标。由图 6 可以看出模型评价结果和专家评价结果情感得分差值较小，最大差值仅为 0.166 2，表明二者在评价结果的正负情感倾向上基本相似。

由于师范生讲课训练评价结果以大量文本形式呈现，针对每个主题再基于情感得分计算模型评价和专家评价的（quadratic weighted kappa, QWK）系数，以揭示模型评价和专家评价之间的一致性水平，QWK 系数如图 6 中 D 图所示，结果显示，3 个主题课程的 QWK 系数均超过 0.7，模型和专家对各主题课程的评分达到整体一致。

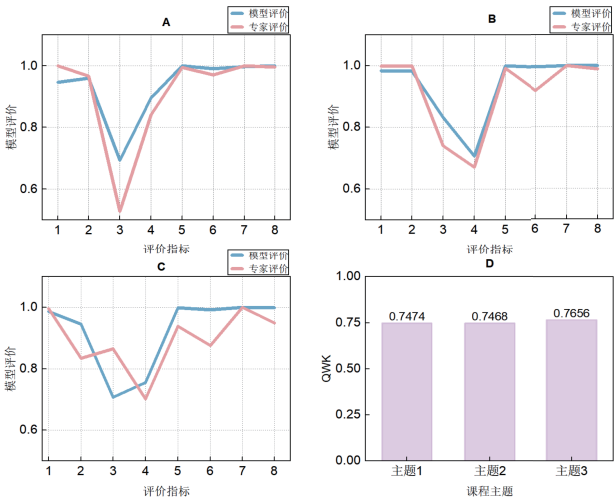


图 6 模型评价与专家评价对比

4.4 消融实验结果对比

为了验证本研究的方法有效，设计了一组消融实验，逐步添加方法以探讨不同方法对师范生讲课训练评价的影响。实验结果如表 5 所示，其中 A、B、C、D、E 方法分别是直接给讲课文本、加入评价指标和提示语、加入针对讲课范围和重难点的可量化计算模型、附加具有专业知识的外部知识库、为模型设定特定角色并进一步优化提示语。由于师范生讲课训练评价结果文本过长，不利于展示，于是在进行消融实验结果对比时，在大语言模型给出评价结果的基础上让其再进一步总结评价结果，仅展示评价结果中的关键部分。结果显示，通过逐步添加这类优化方法，该框架下的大语言模型对师范生讲课训练的评价逐步趋向专业化、合理化，提升

表 5 不同优化方法的消融实验结果

方法					结果
A	B	C	D	E	
✓	—	—	—	—	这段讲课文本讲解了二进制概念并阐述了其重要性，但是在语言流畅性、术语使用严谨性以及具体示例的详细解释上存在一定不足。
✓	✓	—	—	—	该讲课文本在教学内容和逻辑结构上表现较好，围绕二进制主题进行讲解，并适当使用了例子来帮助学生理解。但在语言表达上存在一些口头禅和不流畅之处，部分细节的准确性需要加强。总之文本的授课范围明确，重难点处理得当，但在语言连贯性、层次性和教学准确性方面有待提升。
✓	✓	✓	—	—	这位教师的讲课在内容上基本覆盖了二进制的基础知识，但在细节覆盖和重难点深入解释上有所不足，整体内容准确。在表达层面，教师语言逻辑性强、连贯，举例说明且由浅入深，表现良好。建议教师在重难点的讲解和操作步骤的详细解释上加以改进，以提升教学质量
✓	✓	✓	✓	—	本次讲课内容范围基本合适，涉及二进制的基本概念、与十进制的转换方法及计算机采用二进制的优点。在内容准确性方面，大部分描述正确，但存在一些口误和专业错误。讲课过程中未出现违规词汇，保持了语言规范性。表达层面整体流畅，逻辑清晰，但部分语句过长，可能导致学生理解困难。举例说明和由浅入深的讲解策略有助于学生理解
✓	✓	✓	✓	✓	从教育学专家的评价角度：教师对二进制概念的讲解全面，包括定义、特性和与十进制的转换，内容相关性高。在突出重点和难点方面做得不错，特别是二进制的重要性及其在计算机中的应用。教学准确性总体达标，但有表述错误需纠正。语言使用规范，表达条理清晰，举例恰当

了大语言模型的评价效果,使其能够根据既定的评价指标为师范生的讲课训练提供更准确、更专业的评价和教学改进建议。

5 结论

本研究使用大语言模型作为讲课训练评价的核心,不仅对师范生的教学内容进行精准评价,还借助其文本处理能力,为师范生量身定制个性化的教学改进建议。与传统的评价方法相比,采用大语言模型进行评价不仅展现了更高的智能化水平,而且在专业性和全面性上均有所提升。通过建立“评价-建议”闭环反馈模型,深入探究师范生教学能力评价背后的教育挑战和教学规律,从而向师范生提供智能化且精确的教学优化建议。该架构不仅提升了评价的质量和效率,为边疆民族地区师范生的技能提升和能力培养提供了新的技术支撑,而且对边疆民族地区教育事业的进步产生了积极影响,开创了师范生讲课训练评价的新范式。

尽管大语言模型强大的文本处理能力在此架构中能很好地对师范生的讲课文本进行个性化的评价和建议,但是,目前该评价框架能够处理的数据维度有限,难以满足全景化多角度建模分析的需求,仅仅从师范生的讲课文本来对其整体教学效果进行评价具有局限性,与现实教学可能存在差距,并不能从多个维度对其教学能力进行评估。随着视觉大模型的不断训练和优化,视觉大模型能够更加准确地识别和理解图像中的内容,具有强大的图像识别和处理能力。在此背景下,如何提升智能评价的多维性和科学性,建立全景化多角度智能评价模型,提升模型可评价的讲课训练维度,将成为未来研究基于大模型的智能评价的核心问题。

参考文献:

- [1] 高海荣,李永宇,王雅苹,等.“三习贯通”提升师范生教育教学实践能力路径探索[J]. 郑州师范教育,2024,13(2):24-27.
- [2] 周贵寅,唐静,徐欣,等.基于OBE教育理念的化学专业师范生教学实践能力培养探索与实践[J]. 当代教育理论与实践,2024,16(3):109-114.
- [3] 孙立会,周亮.生成式人工智能融入国家中小学智慧教育平台的实践逻辑[J]. 中国电化教育,2024(8):71-79.
- [4] 王斌华.教师评价模式:微格教学评价法[J]. 全球教育展望,2004,33(9):43-47.
- [5] 王斌华.教师评价模式:教学档案袋[J]. 教育理论与实践,2004(13):24-28.
- [6] 韦雁仙,李霖,段宁贵.基于人脸识别的课堂教学行为S-T分析系统设计[J]. 梧州学院学报,2024,34(3):36-45.
- [7] 漆帝志.基于弗兰德斯互动分析系统的课堂教学行为分析:以研究高中信息技术课《广度优先搜索方法》为例[J]. 教育信息技术,2023(11):50-53.
- [8] 教育部.教育部发布4项行动助推人工智能赋能教育[EB/OL].(2024-03-28)[2026-05-06].http://www.moe.gov.cn/jyb_xwfb/xw_zt/moe_357/2024/2024_zt05/mtbd/202403/t20240329_1123025.html.
- [9] 方娇,梁敏.我国教师专业发展研究综述[J]. 教育教学论坛,2012(15):106-109.
- [10] 教育部办公厅.教育部办公厅关于印发《中学教育专业师范生教师职业能力标准(试行)》等五个文件的通知[EB/OL].(2021-04-02)[2026-05-07].http://www.moe.gov.cn/srcsite/A10/s6991/202104/t20210412_525943.html.
- [11] 高桢婕.国内教师教学技能评价研究综述[J]. 科技视界,2021(13):48-50.
- [12] 鲁圆圆,陈增照,陈荣,等.智能技术助推教师课堂教学行为评价的应用框架研究[J]. 现代教育技术,2022,32(12):76-84.
- [13] 吴立宝,曹雅楠,曹一鸣.人工智能赋能课堂教学评价改革与技术实现的框架构建[J]. 中国电化教育,2021(5):94-101.
- [14] 陈增照,石雅文,王梦珂,等.智能技术支持的教师注意力分析与评价研究[J]. 现代教育技术,2024,34(4):100-111.
- [15] 张立山,冯硕,李亭亭.面向课堂教学评价的形式化建模与智能计算[J]. 现代远程教育研究,2021,33(1):13-25.
- [16] 蔡慧英,卢琳萌,顾小清.人机协同教研会促进教师教学反思能力的发展吗?:基于课堂视频智能分析技术的实证研究[J]. 现代远距离教育,2023(1):40-49.
- [17] 郑金存,陈演桐.利用KINECT运动跟踪技术评价师范生授课技能的实践研究[J]. 电子技术,2018,47(12):54-56.
- [18] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners[C]// NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems. New York: ACM, 2020: 1877-1901.
- [19] Long Yun, Luo Haifeng, Zhang Yu. Evaluating large language models in analysing classroom dialogue[J/OL]. npj Science of Learning,2024[2026-05-07].<https://www.nature.com/articles/s41539-024-00273-3>.
- [20] 龙禹辰,勾智楠,陈宇欣,等.基于大语言模型的多任务生成式重构对话情绪识别[J]. 计算机应用研究,2025,42(7):1964-1971.
- [21] 方海光,洪心,舒丽丽,等.基于课堂智能分析大模型的教师教学能力分析框架及其应用研究[J]. 现代教育技术,2024,34(2):43-52.
- [22] Bewersdorff A, Seßler K, Baur A, et al. Assessing student errors in experimentation using artificial intelligence and large language models: a comparative study with human raters[J]. Computers

and Education: Artificial Intelligence, 2023, 5: 100177.

[23] 褚乐阳, 刘泽民, 王浩, 等. 大模型支持的教师循证实践: 行动框架与案例应用 [J]. 开放教育研究, 2024, 30(4): 91-103.

[24] 卢宇, 余京蕾, 陈鹏鹤. 基于大模型的教学智能体构建与应用研究 [J]. 中国电化教育, 2024(7): 99-108.

[25] 余胜泉, 熊莎莎. 基于大模型增强的通用人工智能教师架构 [J]. 开放教育研究, 2024, 30(1): 33-43.

[26] 徐磊, 胡亚豪, 潘志松. 针对大语言模型的偏见性研究综述 [J]. 计算机应用研究, 2024, 41(10): 2881-2892.

[27] GLM Team. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools[PP/OL].(2024-07-30) [2026-05-07].<https://doi.org/10.48550/arXiv.2406.12793>

[28] Wang Weihang, Lü Qingsong, Yu Wenmeng, et al. CogVLM: visual expert for pretrained language models[PP/OL]. (2024-02-04)[2026-04-07].<https://doi.org/10.48550/arXiv.2311.03079>.

[29] 智谱 AI. 智谱 AI 推出新一代基座模型 GLM-4[EB/OL]. [2024-09-07]. <https://zhipuai.cn/devday>.

[30] 刘燕, 荆国茹, 李玉玲. 师范生全学期教育实习评价指标体系构建: 以内蒙古师范大学生物科学专业为例 [J]. 教育教学论坛, 2022(25): 164-168.

[31] Chen Caixia, Chen Jun, Shi Chun. Research on credit evaluation model of online store based on SnowNLP[C]//International Conference on Intelligent Automation, Energy and Environment Engineering (ICAEER 2018). Les Ulis: EDP Sciences, 2018: 03039.

[32] 张小艳, 闫壮. 融合大语言模型的三级联合提示隐式情感分析方法 [J]. 计算机应用研究, 2024, 41(10): 2900-2905.

[33] Zhang Tianyi, Kishore V, Wu F, et al. BERTScore: evaluating text generation with BERT[PP/OL].(2020-02-24)[2026-05-07]. <https://doi.org/10.48550/arXiv.1904.09675>.

[34] Lu Sicheng, Liu Qiang, Zhang Zhikui. Sentiment Analysis of Weibo Platform Based on LDA-SnowNLP model[C]//2023 2nd International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM). Piscataway: IEEE, 2023: 505-510.

【作者简介】

杨怡(2000—), 女, 云南大理人, 硕士研究生, 研究方向: 智慧教育、计算机视觉。

张洪源(2000—), 男, 河北秦皇岛人, 硕士研究生, 研究方向: 智慧教育、计算机视觉。

焦方旭(2000—), 男, 山东济南人, 硕士研究生, 研究方向: 智慧教育、计算机视觉。

赵姗(1995—), 女, 云南昭通人, 博士, 讲师, 研究方向: 计算机视觉、图像处理。

高丽金(1981—), 通信作者(email:gaolijin@ynnu.edu.cn), 女, 云南曲靖人, 硕士, 讲师, 研究方向: 模式识别、图像处理。

杨扬(1980—), 男, 云南昆明人, 博士, 研究方向: 计算机视觉、模式识别。

(收稿日期: 2025-10-13 修回日期: 2026-05-11)

(上接第 165 页)

参考文献:

[1] 徐丽君. 智慧校园学生行为数据采集与异常检测平台 [J]. 长江信息通信, 2025, 38(8): 102-104.

[2] 谭暑秋, 汤国放, 涂媛雅, 等. 教室监控下学生异常行为检测系统 [J]. 计算机工程与应用, 2022, 58(7): 176-184.

[3] 廖鹏, 刘宸铭, 苏航, 等. 基于深度学习的学生课堂异常行为检测与分析系统 [J]. 电子世界, 2018(8): 97-98.

[4] 任孟其, 宋汝鑫, 王萌, 等. 面向智慧校园的学生异常行为检测 [J]. 哈尔滨师范大学自然科学学报, 2017, 33(3): 20-24.

[5] 余建, 林志兴, 肖香梅. 采用分形维数聚类算法的云计算用户异常行为检测方法 [J]. 三明学院学报, 2025, 42(3): 27-38.

[6] 李晨琛, 窦敬, 孙霁云, 等. 基于姿态表征的轻量化行人异常行为检测算法 [J]. 西安邮电大学学报, 2025, 30(3): 94-102.

[7] 蔡佳炜, 李建文, 谢胜达. ReMass-iForest+: BDS 广播星历异常检测方法 [J]. 导航定位学报, 2025, 13(3): 85-98.

[8] 刘一宁, 陈柏安, 杜鹏程, 等. 基于 MDLOF-iForest 和 M-KNN-Slope 的公共建筑负荷异常数据识别与修复 [J]. 综合智慧能源, 2025, 47(3): 62-72.

[9] 李雨笑, 胡居荣, 邢延潇, 等. 基于改进孤立森林的海面小目标检测方法 [J]. 雷达科学与技术, 2024, 22(6): 628-636.

【作者简介】

王妍(1993—), 女, 江苏南通人, 硕士研究生, 讲师, 研究方向: 大数据分析。

孙克争(1983—), 男, 江苏赣榆人, 硕士研究生, 副教授, 研究方向: 算法优化。

朱亚丽(1987—), 女, 江苏南通人, 硕士研究生, 副教授, 研究方向: 人工智能。

(收稿日期: 2025-09-22 修回日期: 2026-05-08)