文章编号:2096-1472(2023)10-0027-05
DOI:10.19644/j.cnki.issn2096-1472.2023.010.006

基于生成式人工智能的学业评价应用研究——以 ChatGPT 为例

陈思文，孔亚琪，刘　宇

(南京邮电大学教育科学与技术学院，江苏　南京　210023)

✉ 1021163315@njupt.edu.cn; 1021163408@njupt.edu.cn; yliu@njupt.edu.cn

摘　要:目前,传统的学业评价方法在反映学生的各项技能与知识掌握情况方面尚存一定不足,评价过程需要较多的时间与资源且难以实现个性化评价。文章首先探讨了 ChatGPT(Chat Generative Pre-trained Transformer)在学业评价中的生成与应用,对学生学习数据进行诊断、激励、指导、干预。其次使用 Bi-LSTM(Bi-directional Long Short-Term Memory)模型对评价文本进行情感分析,并使用 BERT(Bidirectional Encoder Representation from Transformers)模型进行文本相似度检测,对 ChatGPT 评价内容与教师评价内容进行对比,结果显示:ChatGPT 的评价内容情感在客观上更为积极,其评价内容文本相似度达到教师评价的 75.21%以上,已具备实际应用价值与潜力。最后探讨了生成式 AI 在学业评价应用中的风险与启示。

关键词:生成式人工智能;AIGC;学业评价;ChatGPT

中图分类号:TP311.5　　**文献标志码**:A

Research on the Application of Generative Artificial Intelligence in Academic Assessment —A Case Study of ChatGPT

CHEN Siwen, KONG Yaqi, LIU Yu

(*College of Education Science and Technology, Nanjing University of Posts and Telecommunications, Nanjing* 210023, *China*)

✉ 1021163315@njupt.edu.cn; 1021163408@njupt.edu.cn; yliu@njupt.edu.cn

Abstract: Currently, traditional methods of academic assessment often fall short in accurately reflecting students' mastery of various skills and knowledge. Its evaluation process requires considerable time and resources while struggling to achieve personalized evaluations. This research first explores the generation and application of ChatGPT (Chat Generative Pre-trained Transformer) in educational assessment, aiming to diagnose, motivate, guide, and intervene in students' learning progress using their academic data. Then, the study employs a Bi-LSTM (Bi-directional Long Short-Term Memory) model for sentiment analysis of evaluation text and utilizes a BERT (Bidirectional Encoder Representation from Transformers) model for text similarity detection. ChatGPT generated evaluations are compared with those provided by teachers. The results reveal that ChatGPT's evaluations exhibit more positive sentiment, with a text similarity of over 75.21% of the teacher's evaluation, demonstrating its practical applicability and potential. Lastly, the study examines the risks and insights associated with the implementation of generative AI in educational assessment.

Key words: generative AI; AIGC(Artificial Intelligence Generated Content); academic evaluation; ChatGPT

0　引言(Introduction)

学业评价也被称为教育评价或学术评价。通过学业评价,教师、学校或教育系统能够评估学生的学习进度程度,了解学生的知识和技能水平。学业评价通常涉及各种形式的测试和评估,包括标准化测试、项目评价、口头评价、书面作业、课堂参与等[1]。尽管学业评价在教育过程中发挥了重要作用,但传统的学业评价方法存在一些短板,一是在反映学生各项技能和知识水平方面存在一定的局限性,二是需耗费大量时间和资源进

收稿日期:2023-02-07
基金项目:南京邮电大学教育科学"十三五"规划课题重点课题项目(GJS-XKT1901)

行评分，而且难以进行个性化评价。近年来，生成式人工智能（Artificial Intelligence Generated Content，AIGC）的出现为这些问题提供了可行的解决方案。通过利用深度学习和自然语言处理技术，生成式人工智能能够对学生的作业和考试进行高效、公正且全面地评价，从而提供更详尽的反馈，并更好地满足个性化教学的需求[2]。因此，生成式人工智能在学业评价中的应用具有巨大的潜力及价值。然而，在使用 AIGC 技术时，需综合考虑多方面的因素以保证其作用最大化，为学习评价提供了创新路径。本文研究以 ChatGPT 为例，结合学业评价的生成与应用，验证 AIGC 在学业评价中的应用效果及其风险应对策略。

1 基于生成式人工智能的学业评价生成研究（Research on academic evaluation generation based on Artificial Intelligence Generated Content）

近年来，生成式人工智能已经在教育领域得到了广泛的应用。学业评价是教育评价的一个重要领域，其作用是帮助学生了解自身的学习表现与亟待改进之处，也能帮助教师更好地指导学生学习，提高教学质量。然而，传统的学业评价方法易出现评价趋于主观性，难以快速评价一定数量学生等问题。因此，在学业评价中使用 AIGC 可提高评价效率和准确性，在一定程度上也降低了评价中人的主观性的影响[3]。针对生成式人工智能在学业评价中的应用，本文采用真实的学生学习数据作为输入信息，并将这些数据分别输入 ChatGPT（一个大型的语言模型）和交给两位数学专业教师用于评价，ChatGPT 和两位数学专业教师将针对学生的学习行为表现，从诊断、激励、指导、干预 4 个方面进行评价。

为了更好地使 ChatGPT 作为评价者对学生的学习和行为数据进行诊断、激励、指导、干预，需要先编写合适的 Prompt（提示词）引导 ChatGPT 成为一个评价者。其中，Prompt 是一种文本片段，其目的是指导 ChatGPT 根据给定的条件生成特定类型的文本输出，可理解为在给定的上下文中，使用某一主题或话题引导模型生成使用者所需的相关文本[4]。若要使用 ChatGPT 对学生进行学习评价并评估学生的课程表现，需要先使用合适的 Prompt 指导 ChatGPT 生成正确的文本输出，在此过程中应考虑以下几个方面。

（1）输入的信息：需要收集学生课堂内外表现的信息，如学生的出勤率、课堂表现和潜在的课堂问题等。

（2）评估的要素：确定用于评估学生表现的要素。可以根据学生的课堂表现，分析学生对某些概念的掌握程度，回答问题的能力，主动提出问题的频率，以及对课程的积极参与度等。

（3）Prompt 的生成：编写合适的 Prompt，并使用它引导 ChatGPT 为每名学生评估他们的表现。例如，给定一名学生表现评估的 Prompt，ChatGPT 会基于其内部的模型，生成一个文本输出，这个输出会包含一名学生的综合表现，对学生的学习表现进行概括，给出学习改进建议或其他此类有关文本。

在这个过程中，需要尝试不同的 Prompt，确保 ChatGPT 生成符合预期的文本输出。同时，需要利用反馈告知 ChatGPT 其生成的文本是否合适，以便 ChatGPT 进一步优化生成模型。基于 ChatGPT 的学业评价生成流程如图 1 所示。

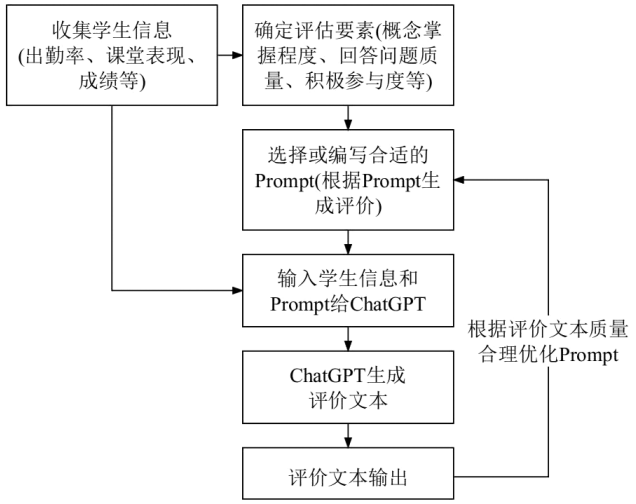

图 1 基于 ChatGPT 的学业评价生成流程图

Fig. 1 Flow chart of academic evaluation generation based on ChatGPT

本文研究选择了 UCI Machine Learning Repository（加州大学欧文分校机器学习数据库）中的 Student Performance 数据集作为学生信息数据来源，此数据集包含学习两个课程（数学和葡萄牙语）的学生表现数据，涵盖了学生的个人信息、家庭背景、学习习惯、课堂表现和成绩等维度，为本文研究提供了一个全面且深入的视角理解学生的学习情况与生活情况，数据集中的部分学生学习数学课程数据如表 1 所示。

在数据处理阶段首先进行数据清洗，检查并处理数据集中的缺失值和异常值以确保数据的准确性；其次对数据进行了预处理，将二元变量转换为文字变量，例如将 1 和 0 转换为“是”和“否”，这样处理的目的是使教师可以更直观、更轻松地了解学生的数据，从而更好地对学生的学习数据进行评价决策。

表 1 Student Performance 数据集中三名学生学习数学课程的表现数据

Tab.1 Data of three students' performance in mathematics courses in the Student Performance dataset

类别	学生 1	学生 2	学生 3
学校	Gabriel Pereira	Mousinho da Silveira	Mousinho da Silveira
性别	女	男	女
年龄/岁	15	18	17
选择此学校的原因	其他	其他	课程偏好
课程挂科次数/次	3	1	0
是否参加付费补课	是	否	是
逃学次数/次	10	10	0
是否参与课外活动	否	否	是
是否期望获得高等教育	否	是	是
考试 1 分数/分（总分为 20 分）	7	11	16
考试 2 分数/分（总分为 20 分）	8	11	15
期末考试分数/分（总分为 20 分）	10	13	15
每周数学学习时间/h	2～5	2～5	2～5

Prompt 构建完成后，以数据集中给出的数学课堂中随机选择三名学生的学习数据为例，将每名学生学习数学课程的表

现数据以及 Prompt 输入给 ChatGPT,即可得到三名学生基于 ChatGPT 的学业评价内容文本。

2 基于 ChatGPT 的学业评价内容验证与分析(Validation and analysis of academic evaluation content based on ChatGPT)

2.1 评价内容情感分析

情感分析是一种自然语言处理技术,它的目标是识别和提取文本中的主观信息,如情绪、观点、情感等[5]。在学业评价中进行情感分析的主要原因是更深入地理解评价者的态度和情绪倾向[6]。这种理解有利于判断评价的积极性或消极性,以及评价的强度和情感色彩。首先,情感分析可以帮助量化评价内容的情感倾向,通过这种方法可以将主观的、定性的评价转化为可以量化和比较的数据。通过文本情感分析技术可以更公正、客观地比较 ChatGPT 生成的评价内容和教师的评价内容。其次,情感分析可以揭示评价者的情绪状态和态度,这对于理解评价者的观点与意图至关重要。再次,情感分析有利于发现潜在的问题和改进点。例如,如果情感分析结果显示某位教师的评价总是倾向于消极,即需要进一步研究其评价方法和内容,检查是否有需要改进的地方。同样,如果 ChatGPT 生成的评价过于消极,那么需要调整 Prompt 生成策略,使其更好地反映学生真实的学习情况。

在进行情感分析的过程中,本文研究使用中文自然语言处理开源数据集 weibo_senti_100k 作为数据源,此数据集包含 10 万多条附带情感标注的新浪微博评论,其中正向评论和负向评论各约 5 万条。首先,对输入的中文文本数据进行预处理,预处理包括分词和构建词汇表的过程。其次,使用 jieba 分词库将文本切分为单个词语,并构建词汇表(vocab)存储词语和对应的索引。

在预处理数据之后,将数据集分为训练集和测试集,使用 Bi-LSTM 模型进行情感分析,这是因为 Bi-LSTM 在处理序列数据方面具有优秀的性能。情感分析涉及对文本进行时序建模,以捕捉文本中的上下文信息和语义结构。Bi-LSTM 作为一种循环神经网络(RNN)的变体,能够有效地处理序列数据,并具有一定的记忆能力。构建 Bi-LSTM 模型,该模型包括一个嵌入层(Embedding)、一个双向 LSTM 层(Bi-LSTM)、一个全连接层(Fully Connected)和一个 Dropout 层(Dropout),其模型网络结构图如图 2 所示。首先,将词汇索引序列作为输入,通过嵌入层将每个词语转换为固定维度的词嵌入向量。其次,将嵌入向量输入双向 LSTM 层中得到隐藏状态。最后,将隐藏状态经过拼接和全连接层操作后,通过 Dropout 层得到模型的输出结果。在训练过程中,使用二元交叉熵损失函数(BCEWithLogitsLoss)作为优化目标,并使用 Adam 优化器进行参数更新。

二元交叉熵损失函数是用于二分类问题的一种常用损失函数。假设有一个二分类问题,其真实标签为 y(取值为 0 或 1),模型预测的概率为 p。那么,二元交叉熵损失函数可以定义如下:

$$loss=-[y\cdot\ln(sigmoid(logits))+(1-y)\cdot\ln(1-sigmoid(logits))] \tag{1}$$

其中,$logits$ 表示模型输出的未经 $sigmoid$ 函数处理的结果,y 表示真实标签(取值为 0 或 1),$sigmoid(\cdot)$ 表示 $sigmoid$ 函数,$\ln(\cdot)$ 表示以 e 为底的自然对数。

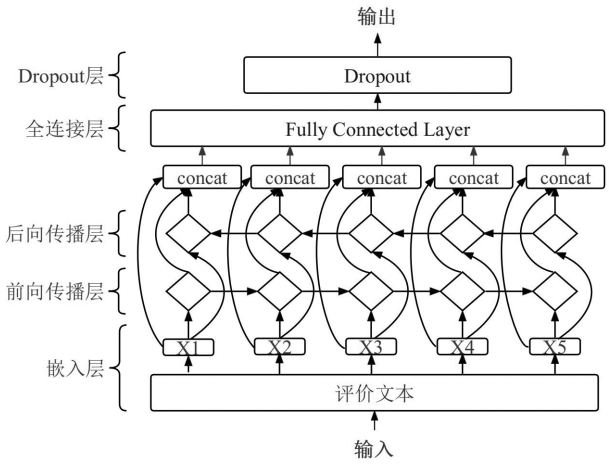

图 2　Bi-LSTM 模型网络结构图

Fig. 2 Bi-LSTM model architecture diagram

Adam(Adaptive Moment Estimation)优化器是一种用于深度学习模型的优化算法,它结合了 Momentum 和 RMSprop 两种优化算法的优点。Adam 优化器的更新规则如下。

(1)计算梯度的一阶矩估计和二阶矩估计:

$$m_t=\beta_1 m_{t-1}+(1-\beta_1)g_t \tag{2}$$

$$v_t=\beta_2 v_{t-1}+(1-\beta_2)g_t^2 \tag{3}$$

其中,m_t 和 v_t 分别是梯度的一阶矩估计和二阶矩估计,β_1 和 β_2 超参数(通常设为 0.9 和 0.999),g_t 为在时间步 t 的梯度。

(2)对一阶矩估计和二阶矩估计进行偏差修正:

$$\hat{m}_t=\frac{m_t}{1-\beta_1^t} \tag{4}$$

$$\hat{v}_t=\frac{v_t}{1-\beta_2^t} \tag{5}$$

其中,$\hat{m}_t$ 和 $\hat{v}_t$ 是偏差修正后的一阶矩估计和二阶矩估计,t 是当前的时间步。

(3)使用修正后的一阶矩估计和二阶矩估计更新参数:

$$\theta_t=\theta_{t-1}-\alpha\frac{\hat{m}_t}{\sqrt{\hat{v}_t}+\varepsilon} \tag{6}$$

其中,θ_t 是在时间步 t 的参数,α 是学习率,ε 是一个很小的数(通常设为 10^{-8}),以防止除以 0。在训练和验证函数中,模型的预测结果和真实标签被输入损失函数中计算损失,然后通过反向传播和优化器更新模型参数。

此外,本文研究使用朴素贝叶斯模型和逻辑回归模型两种机器学习模型进行模型训练,使用 Bi-LSTM 模型进行性能对比,各模型在测试集上的 *ACC*(准确率)和 *LOSS*(损失)指标如表 2 所示,从表 2 中的数据来看,Bi-LSTM 模型在情感分析任务上明显优于朴素贝叶斯模型和逻辑回归模型的机器学习模型。

表 2　Bi-LSTM 模型与朴素贝叶斯模型和逻辑回归模型的性能对比

Tab.2 Performance comparison of Bi-LSTM model with naive Bayes and Logistic Regression models

模型	*ACC*	*LOSS*
朴素贝叶斯模型	0.784 9	0.242 1
逻辑回归模型	0.796 6	0.251 2
Bi-LSTM 模型	0.980 7	0.052 5

Bi-LSTM(双向长短期记忆)模型是一种循环神经网络

(RNN)模型,它能够处理序列数据,并且能够捕捉序列中的长期依赖关系,这使得它在处理文本数据时,能够理解文本的上下文信息,从而提高模型的预测性能。①双向信息流:Bi-LSTM模型不仅能像传统的LSTM模型那样从前往后处理序列,还能从后往前处理序列,这使得Bi-LSTM模型在预测某个位置的输出时,能够同时考虑到该位置前后的所有信息,从而提高模型的预测准确性。并且,LSTM模型通过引入门控机制,能够有效地避免在训练深层网络时常见的梯度消失和梯度爆炸问题。这使得模型能够学习到更深层次的特征,从而提高模型的预测性能。②模型的泛化能力:从表2中的数据来看,Bi-LSTM模型的损失明显低于朴素贝叶斯模型和逻辑回归模型,说明Bi-LSTM模型在文本情感分析任务上的泛化能力更强。

使用训练完成的Bi-LSTM模型分别对ChatGPT和教师的评价文本内容进行预测。首先对评价内容文本进行相同的停用词预处理,其次将其转化为向量,并使用Bi-LSTM模型进行预测。输出模型对新文本的预测概率,可以帮助使用人了解模型对新文本的情感倾向的预测情况。以上过程可了解到Bi-LSTM模型对不同文本的情感倾向的预测情况,从而进行后续的分析与决策。ChatGPT的评价内容与教师的评价内容的文本情感分析结果统计图如图3所示,图3中ChatGPT+学生1表示ChatGPT对学生1的评价内容的情感分析结果,分数越接近1,说明评价内容越积极。同样,教师1+学生1表示教师1对学生1评价内容的情感分析结果,通过对比显示,ChatGPT对学生的学业评价内容在情感表现上更为积极。

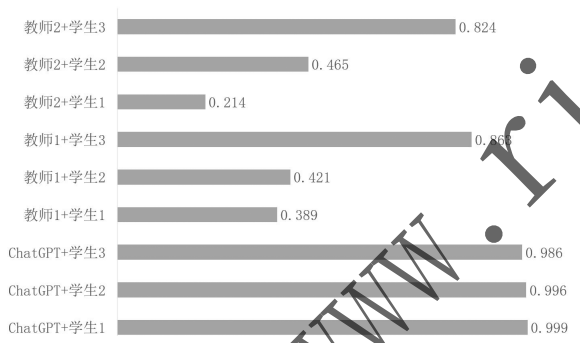

图3 评价文本情感分析结果统计图

Fig. 3 Statistical chart of sentiment analysis results of evaluation text

2.2 评价内容文本相似度检测(Text similarity detection in evaluation content)

基于AIGC的学业评价生成与应用效果验证过程中,对ChatGPT的评价内容与教师的评价内容进行中文文本相似度检测是至关重要的。量化ChatGPT生成的评价内容与教师的评价内容在语义上的相似度,以此评估ChatGPT的评价质量与教师的评价质量。通过这种方式可以了解ChatGPT是否能够生成与教师相似的、高质量的评价内容,从而评估AIGC在学业评价场景中的应用价值。

本文研究使用BERT模型进行中文文本相似度检测。BERT(Bidirectional Encoder Representations from Transformers)是一种基于Transformer的预训练模型,其模型网络结构如图4所示。Transformer模型的核心是自注意力机制(Self-Attention Mechanism),它能够捕捉文本中的长距离依赖关系。BERT模型通过双向的Transformer编码器,能够捕捉到文本中的上下文信息。

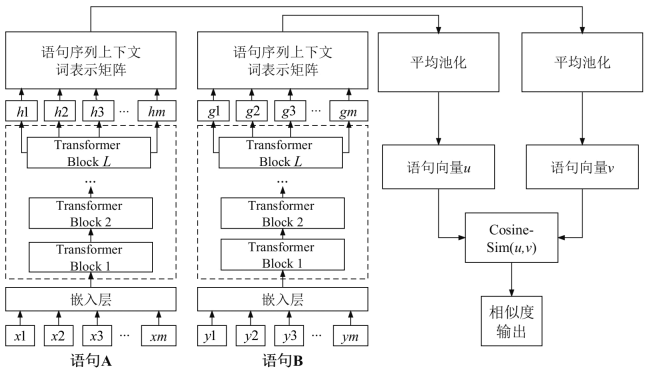

图4 BERT模型网络结构图

Fig. 4 BERT model architecture diagram

在数据处理过程中,首先对评价内容文本进行停用词的去除,其次使用BERT的分词器对文本进行分词,并转化为模型需要的输入格式。因为BERT模型在预训练阶段已经学习到了丰富的语言表示,所以研究在训练过程中使用预训练的BERT模型,不需要进行额外的训练。

本文研究使用余弦相似度公式进行评价文本的相似度计算。余弦相似度是一种基于向量空间的度量,它可以衡量两个向量夹角的余弦值,表示两个向量的相似度。余弦相似度的计算公式如下:

$$\text{Cosine-Sim}(\boldsymbol{u},\boldsymbol{v})=\frac{\boldsymbol{u}\cdot\boldsymbol{v}}{\|\boldsymbol{u}\|\cdot\|\boldsymbol{v}\|} \tag{7}$$

其中,$\boldsymbol{u}$和$\boldsymbol{v}$是2个语句的向量,$\boldsymbol{u}\cdot\boldsymbol{v}$表示$\boldsymbol{u}$和$\boldsymbol{v}$的点积,$\|\boldsymbol{u}\|$和$\|\boldsymbol{v}\|$分别表示$\boldsymbol{u}$和$\boldsymbol{v}$的模。余弦相似度的取值范围为$-1\sim1$,值越大,则表示文本越相似。

训练好模型后,输入ChatGPT的评价内容和教师的评价内容,即可计算ChatGPT和教师之间的文本相似度,文本相似度结果如表3所示,表3中的"ChatGPT+教师1"表示ChatGPT对某名学生的评价内容与教师1对某名学生评价的内容之间的文本相似度,以此类推。结果显示:ChatGPT的学习评价内容的相似度达到了教师评价的75.21%以上,已具备实际的应用价值及具有一定的市场潜力。

表3 ChatGPT与教师评价内容相似度检测结果

Tab.3 Similarity detection results between ChatGPT's generated content and teacher evaluations

类别	学生编号	评价相似度/%
ChatGPT+教师1	学生1	80.55
	学生2	84.37
	学生3	81.88
ChatGPT+教师2	学生1	77.98
	学生2	81.23
	学生3	75.21

3 ChatGPT在学业评价中的挑战与启示(Challenges and insights of ChatGPT in academic evaluation)

3.1 ChatGPT在学业评价中的局限和挑战

ChatGPT在学习评价中不可避免存在一定的主观性和不

确定性[7]。由于模型的训练过程中使用的数据和评价标准在很大程度上决定了所生成的评价结果的质量,因此需要充分认识其局限性,努力寻找更准确和可靠的评价方法。ChatGPT 在处理基本的语法和语义问题上表现良好,但其较难处理复杂的问题或特殊领域的评价,例如针对具体学科、职业或社会背景的评价。因此,必须寻找其他不同类型的模型和算法,在多样化的评价场景和问题中获得更好的评估效果。并且 ChatGPT 模型需要使用大量的用户数据进行训练和调整,这些数据可能涉及用户隐私保护等问题,同时数据量必须达到一定规模才能对模型进行有效训练和优化[8]。所以,采取切实有效的数据采集和审核策略,确保数据的质量和隐私安全,是目前研究者面临的最大挑战。ChatGPT 模型的不透明性以及算法的复杂性产生的评价结果较难被人们所理解,这种不透明性会影响对评价结果的准确性和可靠性的判断,并且随着各种学习场景的不断变化,ChatGPT 模型的适应能力不可避免地会受到一定的影响[9]。如果模型无法处理新的场景或问题,那么需要准备更新和改进模型,确保它能够适应不断变化的学习需求,为学习者提供更准确、可靠的评价和反馈服务。

3.2 ChatGPT 在学业评价中的启示和促进

ChatGPT 模型基于深度学习算法,可以对大量的自然语言数据进行有效的训练和处理。这使得模型可以对不同学生的学习表现进行个性化的评价,提供更加精准的反馈,这对于提高学生的学习动力和效果具有很大的促进作用。学业评价通常需要大量的人力和时间投入,而 ChatGPT 模型可以自动化评价和反馈过程,有助于教育机构和教育工作者提高评价的效率和精确度,使教育资源得到更高效的利用[10]。ChatGPT 模型的应用和研究,为教育的评价、反馈和个性化服务提供了新的思路和方法,推动了教育行业的发展和创新[11]。ChatGPT 模型的应用也提供了新的思路和方法,促进了教育评价方式的创新。教育领域可以对人工智能和自然语言生成等先进技术进行更深入的研究,探索更加准确、全面和高效的评价方式,满足不断变化的学习需求并进一步实现个性化学习[12]。ChatGPT 模型在学习评价中具有广泛的应用前景,需要进一步探索和开发更加高效和准确的评价方法,为社会提供更加优质、个性化和便捷的教育服务。

4　结论(Conclusion)

本文研究主要关注于生成式人工智能系统 ChatGPT 在教育领域的应用潜力,并基于其启发性内容生成、对话情境理解、序列任务执行和程序语言解析 4 项核心能力,探讨该系统在学业评价中的应用效果。本文研究使用真实的学习数据,对学生学习数据进行评价,并对其在评价过程中的诊断、激励、指导和干预进行了相应的评估和比较。结果表明,与两位教师的评价内容相比,ChatGPT 生成的评价内容情感更积极,评价文本相似度达到了教师评价的 75.21%以上。此研究证明了基于 AIGC 的学习评价潜力,证明了其优良的自然语言理解和生成能力在提供学业评价的诊断和指导方面的可应用性。AIGC 可在教育领域中为学生提供更加精准和个性化的学习支持服务,提高学生的学习效果和成果,有望在未来得到更为广泛的应用。

尽管 ChatGPT 在自然语言理解、生成和多模态数据处理方面具有优势,但它仍存在一些技术局限性,例如系统可能会缺乏深入的语义理解或知识表示能力,导致其对某些学术领域(如数学或物理学)的特定领域知识进行推理或解释方面存在困难。因此,在将其应用于学习评价的过程中,需要谨慎考虑其适应性与总体准确性。

参考文献(References)

[1] 喻平. 基于核心素养的高中数学课程目标与学业评价[J]. 课程·教材·教法,2018,38(1):80-85.

[2] QADIR J. Engineering education in the era of ChatGPT: promise and pitfalls of generative AI for education[C]// IEEE. 2023 IEEE Global Engineering Education Conference(EDUCON). Piscataway:IEEE,2023:1-9.

[3] 夏琪,程妙婷,薛翔钟,等. 从国际视野透视如何将 ChatGPT 有效纳入教育:基于对 72 篇文献的系统综述[J]. 现代教育技术,2023,33(6):26-33.

[4] COOPER G. Examining science education in ChatGPT: an exploratory study of generative artificial intelligence[J]. Journal of Science Education and Technology,2023,32(3):444-452.

[5] 胡慧君,杨雨烟,易洋,等. 基于细粒度信息感知 BERT-LEP 的情绪分类方法[J]. 计算机工程与科学,2023,45(4):751-760.

[6] 罗玉萍,潘庆先,刘丽娜,等. 基于情感挖掘的学生评教系统设计及其应用[J]. 中国电化教育,2018(4):91-95.

[7] 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.

[8] 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/OL]. 现代教育技术,2023,33(8):1-10[2023-07-31]. http://kns.cnki.net/kcms/detail/11.4525.N.20230713.2009.002.html.

[9] 郑永和,周丹华,张永和,等. 计算教育学视域下的 ChatGPT:内涵、主题、反思与挑战[J]. 华东师范大学学报(教育科学版),2023,41(7):91-102.

[10] 李海峰,王炜. 生成式人工智能时代的学生作业设计与评价[J]. 开放教育研究,2023,29(3):31-39.

[11] 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.

[12] FERGUS S,BOTHA M,OSTOVAR M. Evaluating academic answers generated using ChatGPT[J]. Journal of Chemical Education,2023,100(4):1672-1675.

作者简介:

陈思文(2000-),女,硕士生。研究领域:学习评价,学习分析。

孔亚琪(1998-),男,硕士生。研究领域:深度学习,教育人工智能,学习分析。本文通信作者。

刘　宇(1971-),女,博士,副教授。研究领域:软件工程,数字化学习,学习分析,课程与教学理论。