可解释自动批阅模型构建与应用

卢 宇[1,2] 章 志[1] 马安瑶[1] 陈鹏鹤[1,2]

（1. 北京师范大学 教育学部教育技术学院，北京 100875；2. 北京师范大学未来教育高精尖创新中心，北京 102206）

[摘要] 自动批阅是数字化教学平台与智能化教育评价的重要实现形式和基本功能。基于深度学习的自动批阅模型逐步成熟但其内部结构复杂且决策过程不透明，导致用户难以信任其批阅结果并影响大规模部署。本研究提出了可解释自动批阅模型的基本框架，包含自动批阅基础模块、自动批阅解释模块与自动批阅交互模块。在此基础上，本研究构建了可解释自动批阅模型的实例并嵌入智能导学系统开展准实验研究。实验结果表明，嵌入可解释自动批阅模型的智能导学系统，有效提升了学习者对自动批阅功能和系统整体的信任度，也有助于提高技术接受度，交互模块的解释性信息也不会增加学习者的认知负荷。最后，本研究提出了可解释人工智能在教育领域开展自动批阅的研究建议和展望。

[关键词] 自动批阅；深度神经网络；可解释人工智能；人机交互；智能导学系统

[中图分类号] G434 [文献标识码] A [文章编号] 1007-2179(2023)05-0098-08

一、引言

自动批阅旨在实现对学习者开放性或半开放性作答的自动评分，是构建数字化教学平台与智能化教育评价的重要实现形式和基本功能（祝智庭等，2022）。然而，基于深度学习的自动批阅模型决策过程复杂，可能产生不易察觉且不可预测的错误输出。即使模型可以输出正确批阅结果，其决策过程宛如黑箱（卢宇等，2022），无法针对批阅结果提供明确评分依据。自动批阅模型的决策复杂性和决策依据缺失，会导致教师和学习者等教育用户难以信任批阅结果，从而影响其在日常教学和高利害考试中的大规模部署和应用。

对复杂人工智能模型的运行过程与输出结果进行适当与合理的阐释，是当前人工智能领域的重要研究方向，也被称为可解释人工智能（Explainable Artificial Intelligence, xAI）。可解释人工智能（Arrieta et al., 2020）旨在设计和应用技术方法和手段，直接或间接地解释复杂人工智能模型的决策过程和结果，帮助用户理解模型和系统，从而建立人机间的信任关系，推动相关人工智能应用的规模化普及。例如，可解释人工智能可以揭示复杂人工智能模型，向用户说明系统隐性的决策规则与底层机制，或者告知用户对当前模型的决策结构起到主

[收稿日期] 2023-09-03 [修回日期] 2023-09-05 [DOI 编码] 10.13966/j.cnki.kfjyyj.2023.05.010

[基金项目] 北京市教育科学“十四五”规划 2021 年度重点课题“人工智能驱动的新一代智能导学系统构建研究”（CHAA21036）。

[作者简介] 卢宇，副教授，博士生导师，北京师范大学教育学部，北京师范大学未来教育高精尖创新中心，研究方向：人工智能及其教育应用（luyu@bnu.edu.cn）；章志，硕士研究生，北京师范大学教育学部教育技术学院，研究方向：可解释人工智能教育应用；马安瑶，硕士研究生，北京师范大学教育学部教育技术学院，研究方向：学习者建模及其教育应用；陈鹏鹤，讲师，北京师范大学教育学部，北京师范大学未来教育高精尖创新中心，研究方向：教育知识图谱、自然语言处理及其教育应用（chenpenghe@bnu.edu.cn）。

[引用信息] 卢宇，章志，马安瑶，陈鹏鹤（2023）. 可解释自动批阅模型构建与应用[J]. 开放教育研究,29(5)：98-105.

导或重要作用的输入特征等解释性信息。可解释人工智能已在金融、交通、医疗等诸多垂直领域有较为广泛的应用,但在教育领域的研究和应用尚处于起步阶段(刘桐等,2022)。

二、文献综述

(一)自动批阅模型

从技术维度划分,自动批阅模型可简单分为基于规则、基于传统自然语言处理与基于深度学习三类。基于规则的自动批阅模型主要基于概念图、学科知识等,通过专家构建规则库,分析用户答案的组成、语法等进行评分批阅。该类模型的规则通常评分依据明确,可解释性好。但此类模型的准确度通常不高,难以处理复杂多变的学习者作答,且学习者可以通过使用大量关键短语、复杂句型等欺骗模型获取高分(袁莉等,2021)。基于传统自然语言处理的自动批阅模型,可以基于预先定义的自然语言文本特征,在高维向量空间计算用户答案与参考答案的文本相似度,进而构建简单的分类或回归模型,实现对用户答案的自动批阅(谭红叶等,2019)。此类模型的准确性高,但难以直观解释并提供自动批阅依据。基于深度学习的自动批阅模型,是当前学术界和工业界研究和部署的热点和主要方向(Ramesh & Sanampudi, 2022)。此类模型利用深度神经网络及海量训练数据,挖掘学习者作答中的深层语义信息,实现个性化精准评分与反馈。此类模型常利用监督式或自监督式机器学习算法,内部参数量大且计算过程抽象复杂,无论是专业人员和教育用户都难以理解和信任。典型模型有基于长短期记忆神经网络与卷积神经网络的 EMD 模型(Kumar et al, 2017)、基于注意力机制的 Att-Grader 模型(谭红叶等,2022)等。

因此,如何针对解释性较差或无法解释的自动批阅模型,构建更加科学合理的自动批阅模型,使其所支撑的系统服务功能对于用户可解释且可信任,是亟待解决的重要问题。

(二)可解释人工智能

在可解释人工智能领域,人工智能模型可简单分为白盒模型和黑盒模型(曾春艳等,2021)。白盒模型指该模型内部结构直观清晰,决策逻辑易于理解,如线性回归与决策树等。黑盒模型的内部结构和决策过程较为复杂,以涵盖循环神经网络、卷积神经网络、图神经网络等深度学习模型为代表,需要利用可解释人工智能技术加以阐释。

可解释人工智能技术分全局解释方法与局部解释方法(Arrieta et al, 2020):全局解释方法通过设计算法揭示模型的运行机制与决策逻辑等全局性关键信息。例如,给定数据样本集合 D 与深度学习模型 M,解释模块可以构建一个在性能表现上逼近 M 的可解释全局模型 m_g,然后通过 m_g 的解释逻辑形成对 M 的全局解释。常见的全局解释方法包括知识提取(Adadi & Berrada, 2018)与激活最大化(Erhan et al., 2009)等。

局部近似方法不直接解释模型本身,更多聚焦实例个体,揭示模型对个体输入作出决策的依据(Lundberg & Lee, 2017)。具体而言,针对深度学习模型 M 及其多维向量输入 x,局部解释方法通过计算 x 的不同维度对 M 输出结果的影响程度,解析深度学习决策结果的主要依据,形成对个体实例输入的科学解释。常见的解释方法有局部近似、反向传播和特征反演等(Guidotti et al., 2018)。反向传播方法借助深度神经网络的反向传播机制,将模型的决策信息逐层向输入方向传播,得到每个模型输入的关联值,从而计算哪些输入特征对模型决策产生了重要影响(Simonyan et al., 2013)。特征反演法利用给定模型某一层的激活,尝试找到一个输入,使其通过模型时产生相同或相似的激活,从而形成对模型决策过程的解释 (Du et al., 2018)。局部解释方法的适用性通常更加广泛。

综上,本研究的核心是:如何选取适当方法,构建与合理解释教育领域的自动批阅模型。

三、可解释自动批阅模型构建

针对日趋复杂的自动批阅模型难以解释和取得用户信任问题,本研究提出可解释自动批阅模型(见图 1)。该模型包括自动批阅基础模块、自动批阅解释模块和自动批阅交互模块三个部分。

(一)自动批阅基础模块

自动批阅基础模块的核心是基于深度学习的自动批阅模型,通常由卷积神经网络(CNN)、循环神经网络(RNN)、变换器(transformer)等结构单元构建。具体而言,基础模块构建包含设计、训练与

评价三个阶段。1)设计阶段：确定自动批阅的目标和需求，例如，面向语文学科的简答题文本类作答或面向信息技术学科的编程类作答等，然后对学生作答数据进行预处理，并合理选择结构单元与模型架构。2)训练阶段：基于大规模学生作答与学科专业知识，使用相关框架和库（如TensorFlow与PyTorch等）开发模型，并利用优化算法训练模型。3)评价阶段：基于准确率、召回率等指标，并结合学科专家经验，评价和验证自动批阅基础模型。

（二）自动批阅解释模块

自动批阅解释模块是整个模型的关键，旨在利用不同的可解释人工智能方法，解释基础模块的自动批阅过程和结果，增强基础模型的透明度，提升学习者、教师和家长等用户的理解与信任度。解释模块的输入数据主要源于基础模块，包含基础模块的批阅结果信息，还包括深度学习模型的内部结构、权重参数及其输入数据信息。在此基础上，解释模块选取不同的方法，多维度解释基础模块。一方面，它可以解释每一批阅结果的判定依据，说明依据何种信息给出学习者的分数和反馈；另一方面，它可以解释批阅过程的决策逻辑，解释批阅模型输出结果的过程中进行了怎样的判断。无论解释信息是什么，本模块都需要将其输出到自动批阅交互模块，与批阅结果融合并反馈给用户。

（三）自动批阅交互模块

自动批阅交互模块旨在通过科学合理的方式，将自动批阅结果及其解释信息展示给用户，提升其对批阅结果的理解与信任。交互模块的设计需要符合自适应性、准确性、完整性及可理解性的用户接口设计原则（Rai, 2020），也需要考虑教育用户的特殊性和认知特点。自动批阅交互模块包含可理解的结果显示、可解释的批阅反馈、可信任的交互机制三部分。可理解的结果显示旨在利用文本、图片甚至虚拟代理等形式向用户显示批阅结果，尤其是相对负面的批阅结果，需尽量采用轻松活泼的可视化形式，还可以提供各级批阅结果的案例、参

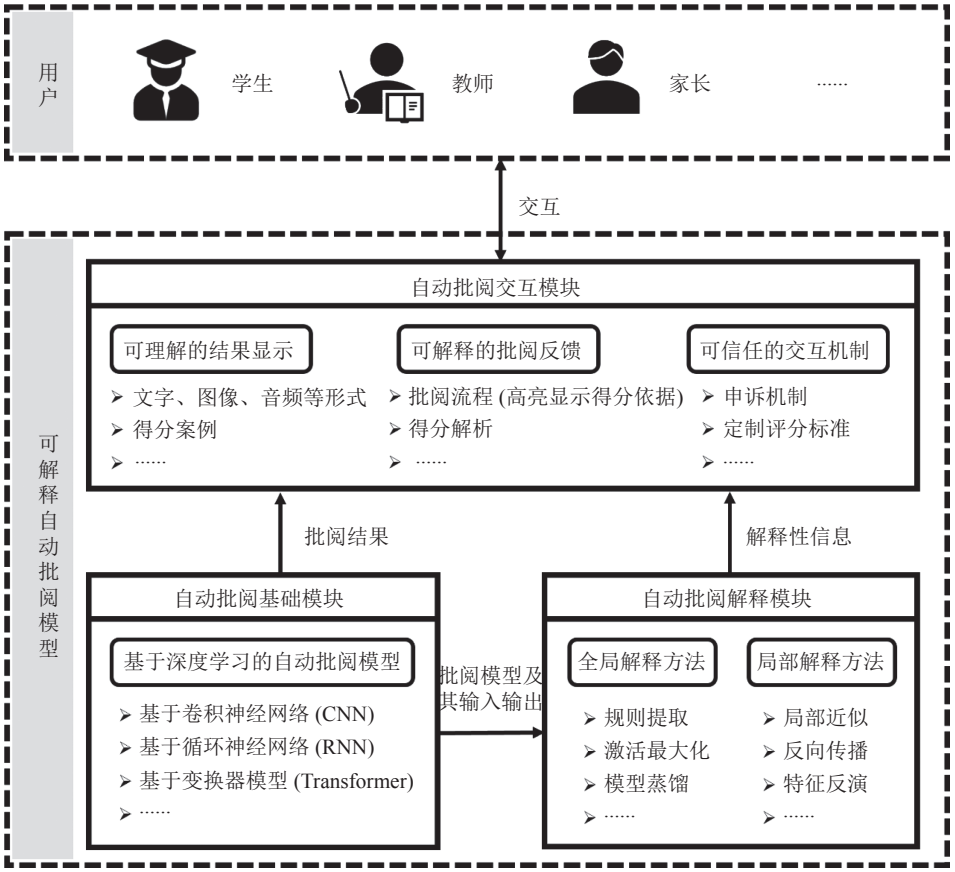

图1　可解释自动批阅模型基本框架

考答案和评分标准等辅助信息。可解释的批阅反馈旨在充分利用解释模块提供的信息，为用户呈现客观的解释性批阅反馈。系统可以对用户答案的各组成部分，分别进行反馈并提供细颗粒度得分，描述评分逻辑与得分依据。重要的解释性信息，可以通过颜色、字体等设置突出显示。可信任的交互机制提供申诉机制、评分标准定制等。申诉机制允许用户询问存疑的评分结果，系统可以再次验证或转为人工批阅。评分标准定制允许教师或家长自定义题目得分点及分数权重，使系统更加灵活且满足特定需求。此外，系统还可以增加样本批阅演示、评分方法详解等功能，提高用户信任度和使用体验。

四、可解释自动批阅模型实现

可解释自动批阅模型可以用多种形式实现，下文以具体案例说明。

（一）基础模块实现

基础模块采用基于注意力机制的深度学习模型 Att-Grader（谭红叶等，2019）。该模型的输入包括参考答案和用户答案，输出为用户答案不同得分的概率值。模型主要由编码层、注意力层、输出层构成。编码层先对用户答案与参考答案进行切分，并通过双向长短期记忆网络对两者分别编码，得到包含上下文信息的用户答案向量U^s与参考答案向量U^r。注意力层利用双向注意力机制计算两者的相似度矩阵及注意力向量$\overline{U}^s$与$\overline{U}^r$。输出层将两个注意力向量进行组合得到拼接矩阵 A，并利用卷积神经网络用户答案与参考答案的相似匹配程度，计算用户答案的最终得分概率。经过在语文、数学等学科数据集的验证，该基础模块的自动批阅准确率以及与教师批阅的一致率表现良好。

（二）解释模块实现

基础模块基于较为复杂的深度学习模型，解释模块可以选择局部解释方法的 LIME（Local Interpretable Model-agnostic Explanations）方法。LIME 方法的核心思想是构建可解释的简单模型来近似复杂模型的局部边界，并基于该简单模型得到原复杂模型输入与输出的关联值，从而解释原复杂模型的决策（Ribeiro et al., 2016）。LIME 模型的优势之一是其与被解释模型的结构无关，适用性较广。

具体来说，给定实例 x 与复杂模型 f，LIME 方法对复杂模型 f 关于实例 x 决策的解释可表示为：

$$\xi(x) = argmin_{g\in G} L(f,g,\pi_x) + \Omega(g) \qquad (1)$$

其中，G 代表简单模型的集合，g 代表某个简单模型，π_x定义实例 x 的邻域的大小，$\Omega(g)$代表简单模型 g 的复杂度。简单模型 g 和复杂模型 f 的预测差距通过函数 L 测量。函数 L 如公式（2）所示：

$$L(f,g,\pi_x(z)) = \sum_{z,z'\in Z} \pi_x(z)(f(z) - g(z'))^2 \qquad (2)$$

其中，z 为扰动实例 x 所生成的实例，z'为实例 z 中非零特征的一部分，Z 代表扰动生成的数据集，$\pi_x(z)$代表实例 x 与扰动实例 z 的相似度。在计算 $f(z)$时，z 中的特征会映射到其原本在实例 x 中的特征值。换言之，LIME 方法利用新生成的数据集 Z 以及复杂模型给出的预测结果 $f(z)$，对简单模型 g 展开训练，直至找到与复杂模型 f 的局部预测差距最小的简单模型 g；然后基于简单模型 g 的权重参数组件，得到实例 x 各个特征与模型输出 $f(x)$的关联值，从而得到对模型决策的解释性信息。

具体而言，假设某数学主观题的参考答案为“两点之间线段最短；且点到直线的距离，垂线段最短”，满分为 2 分。学生 A 的作答为“小丽的依据是两点之间线段最短”，且自动批阅基础模块 Att-Grader 给出学生 A 的自动批阅分数为 1 分。针对该自动批阅结果，LIME 方法可以通过六个基本步骤生成解释性信息（见图 2）。

1）步骤一：依据学生 A 作答的“小丽的依据是两点之间线段最短”答案，生成扰动数据集 Z，通过余弦相似度计算得到扰动数据集中所有实例与学生 A 作答的相似度，从而得到扰动实例的相似度$\pi_x(z)$。

2）步骤二：将扰动数据集 Z 的所有实例输入 Att-Grader 模型，得到每个实例的预测值$f(z)$。

3）步骤三：基于扰动数据集 Z 以及 Att-Grader 模型的预测值$f(z)$，训练得到简单模型 g。

4）步骤四：将扰动数据集 Z 的所有实例输入简单模型 g，得到简单模型的预测值$g(z)$。

5）步骤五：基于扰动实例与用户答案 x 的相似度$\pi_x(z)$、Att-Grader 模型预测值$f(z)$、简单模型预测值$g(z)$，计算出简单模型 g 与 Att-Grader 模型的局部预测差距 L。

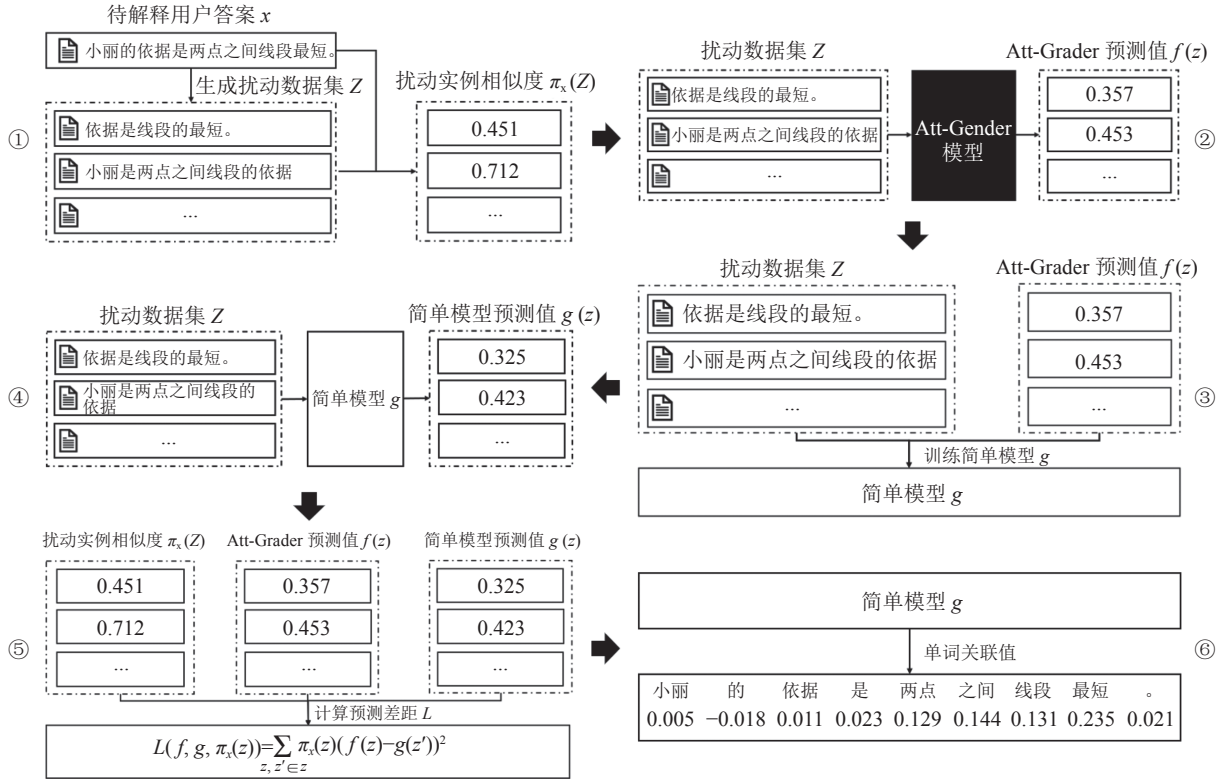

图 2　LIME 方法解释 Att-Grader 模型预测得分步骤

6)步骤六：循环执行步骤三至五，保留与模型 f 的局部预测差距最小的简单模型 g；基于简单模型 g 的权重参数，得到学生 A 作答中各个单词与模型评分 $f(x)$ 的关联值。

通过以上六个基本步骤，研究者可以得到学生 A 作答的各个单词与模型预测评分的关联值，且可以发现“两点”“之间”“线段”“最短”的关联值较大。因此，当前自动批阅得分的依据是答案中有“两点之间线段最短”这一关键信息。

(三)交互模块实现

基础模块的自动批阅结果与解释模块的解释性信息，经过与参考答案进行对比和可视化设计，可以共同在交互模块中进行呈现，实现对自动批阅结果的用户端解释。交互模块由批阅流程、得分解析、题目解析三部分构成。其中，批阅流程呈现评分的具体过程，得分解析呈现自动批阅结果的依据，题目解析呈现题目的参考答案。批阅流程与得分解析的部分截图见图 3。在批阅流程部分，针对学生 A 的作答，交互模块会自动计算其答案中所有单词关联值绝对值的平均值，并将大于平均值的单词作为得分判断依据进行高亮显示。在得分解析部分，交互模块会将高亮单词与题目的得分点进行余弦相似度计算，从而生成自动批阅得分的判断依据信息，例如学习者答案包含哪些得分点以及缺失哪些得分点等。

通过该交互模块，学习者可以了解自己答案中哪些部分得到分数且对最终得分的影响较大，从而理解自动批阅功能的评分结果。此外，如果发现有不合理的内容，交互模块也为学习者提供了申诉和人工批阅途径。

五、可解释自动批阅模型应用成效

(一)实验设计与实施

本研究采用准实验研究方法，将所设计和实现的可解释自动批阅模型嵌入智能导学系统，作为实验组系统。原有智能导学系统具有相同的自动批阅功能，但不具备解释模块及相应的交互能力，作为对照组系统。实验流程见图 4。所有被试被随机分配到对照组与实验组，两组被试均使用无解释能力的自动批阅服务功能，完成练习并查看评分反馈，以初步体验系统。系统体验完成后，研究人员针对信任度、技术接受度、认知负荷三个维度对两

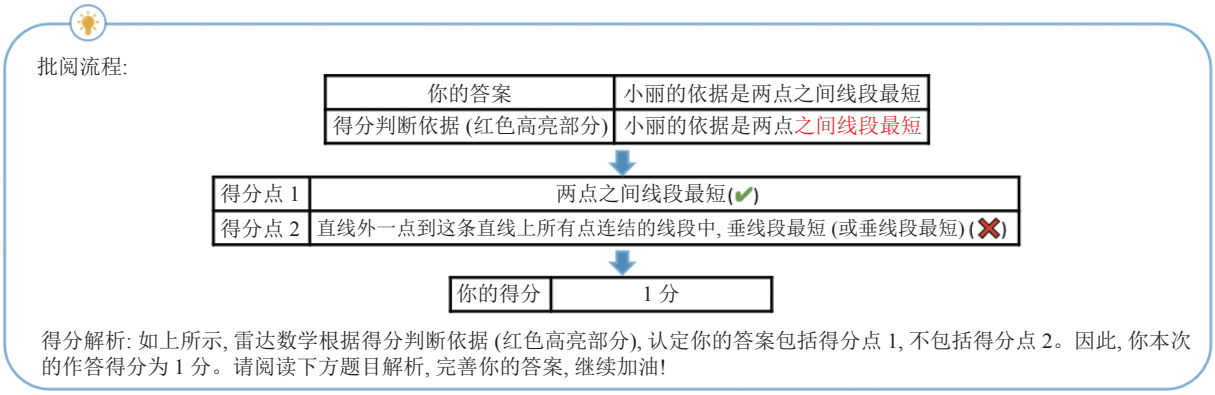

图 3　交互模块的批阅流程与得分解析部分截图

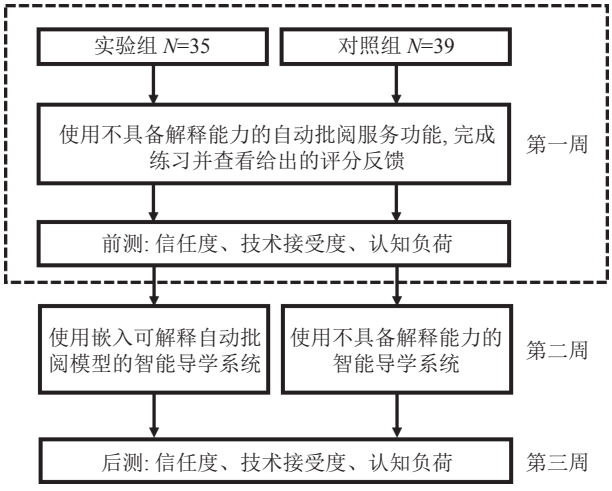

图 4　可解释自动批阅模型教育应用实验设计

组被试开展前测。前测完成后，实验组使用嵌入可解释自动批阅模型的智能导学系统，对照组使用不具备解释能力的智能导学系统。最后，研究人员对所有被试开展相同维度的后测。本研究选取北京市某初中七年级 80 名学生为实验对象，剔除问卷结果缺失的被试，实验组与对照组分别为 35 人和 39 人。实验采用北京师范大学自主研发的“雷达数学”智能导学系统(卢宇等，2023)。

实验采用的问卷均为里克特五点量表。信任度问卷改编自 Hoffman 等(Hoffman et al., 2019)的智能系统用户信任度问卷，共包含 8 道题，涵盖被试对服务功能本身以及评分结果的信任度测量。技术接受度问卷和认知负荷问卷改编自 Hwang 等(Hwang et al., 2013)的技术接受度问卷和认知负荷问卷，分别有 13 道和 8 道题，涵盖被试对服务功能认知有用性和认知易用性的测量，以及其服务功能使用的心智负荷和心智努力的测量。

(二)实验结果

本研究首先对实验组与对照组的前测信任度、技术接受度、认知负荷得分数据进行独立样本 t 检验(见表 1)。结果显示，实验组和对照组的前测信任度(p=0.780>0.05)、技术接受度(p=0.899>0.05)、认知负荷(p=0.835>0.05)的得分均不存在显著差异，表明两组对自动批阅服务功能的信任程度、技术接受程度和使用过程的认知负荷水平均较为接近。

表 1　前测独立样本 t 检验结果

维度	组别	样本数	均值	标准差	t	p
信任度	实验组	35	2.411	0.160	0.280	0.780
	对照组	39	2.353	0.135		
技术接受度	实验组	35	2.202	0.130	0.128	0.899
	对照组	39	2.180	0.122		
认知负荷	实验组	35	3.507	1.004	0.209	0.835
	对照组	39	3.452	1.262		

注：*p<0.05，**p<0.01，***p<0.001。

在此基础上，本研究对实验组和对照组的后测信任度、技术接受度、认知负荷进行独立样本 t 检验(见表 2)。结果显示，实验组与对照组的后测信任度(p=0.000<0.001)与技术接受度(p=0.022<0.05)的得分存在显著差异，且实验组的信任度(均值=3.229)与技术接受度(均值=2.385)得分均高于对照组的信任度(均值=2.385)与技术接受度(均值=2.189)。同时，实验组与对照组的认知负荷(p=0.434>0.05)得分不存在显著差异。

最后，本研究对两组的前测与后测在信任度、技术接受度、认知负荷维度分别进行配对样本 t 检验，其结果见表 3。实验组的前后测认知负荷(p=0.127>0.05)得分不存在显著差异，但信任度(p=0.000<0.001)与技术接受度(p=0.016<0.05)的得分

表 2　后测独立样本 t 检验结果

维度	组别	样本数	均值	标准差	t	p
信任度	实验组	35	3.229	0.112	5.162	**0.000*****
	对照组	39	2.385	0.118		
技术接受度	实验组	35	2.591	0.093	2.352	**0.022***
	对照组	39	2.189	0.143		
认知负荷	实验组	35	3.199	0.914	−0.772	0.434
	对照组	39	3.404	1.306		

注：*p<0.05，**p<0.01，***p<0.001。

表 3　两个组的前、后测实验数据配对样本 t 检验结果

维度	组别	样本数	前测均值	后测均值	t	p
信任度	实验组	35	2.411	3.229	−4.226	**0.000*****
	对照组	39	2.353	2.385	−1.108	0.275
技术接受度	实验组	35	2.202	2.591	−2.532	**0.016***
	对照组	39	2.180	2.189	−0.049	0.961
认知负荷	对照组	35	3.507	3.432	1.565	0.127
	实验组	39	3.452	3.404	0.181	0.857

注：*p<0.05，**p<0.01，***p<0.001。

存在显著差异，且后测中信任度（均值=3.229）与技术接受度（均值=2.591）得分均高于前测信任度（均值=2.411）与技术接受度（均值=2.202）。对照组测信任度（p=0.275>0.05）、技术接受度（p=0.961>0.05）、认知负荷（p=0.857>0.05）的前后测得分均不存在显著差异。

（三）结果讨论

数据分析表明，嵌入可解释自动批阅模型的智能导学系统，可以向学习者提供自动批阅得分和失分的依据等信息，有效提升了学习者对自动批阅功能和系统整体的信任，也有助于提高技术接受度。交互模块的解释性信息也不会增加学习者的认知负荷。学习者对不具备解释能力的自动批阅功能和智能导学系统的信任度普遍较低。缺乏解释性信息，使学习者不会随着交互次数的增加而逐步信任和接受系统。这些都可能影响学习者再次使用自动批阅服务甚至整个智能导学系统。

六、总结与展望

自动批阅作为数字化教学平台的重要服务功能，是当前和未来智能化教育评价的重要手段和方法。本研究提出的可解释自动批阅模型，可以有效启发和帮助各类数字化教学平台的设计和实践应用，提升用户对机器批阅结果的信任度，增加系统的用户友好性。本文提出以下研究展望：

首先，随着以 ChatGPT 为代表的生成式人工智能快速演进，以超大规模神经网络为基础的教育领域大模型正在迅速发展。因此，设计和研发可解释的教育大模型亟需受到重视，以确保教育领域可以充分利用相关人工智能技术，快速提高各类数字教育平台的通用性与智能性。

其次，教育系统与模型通常有自身的独特性，可解释人工智能领域的通用方法难以对其充分合理解释。因此，鼓励设计和研发针对教育系统和模型的专有解释方法与技术，建立检验系统决策合理的规范流程，及时提示和预警违背教育原则的不合理决策是一项重要工作。

最后，自动批阅、开放学习者模型等关键性功能与模块，需要考虑学习者与教师等不同用户的心理与认知特点，建立相应的用户心理需求模型，分别设计用户交互方式与用户接口，开展多轮用户测试与迭代优化，确保所设计的解释性服务功能满足不同角色的实际心理需求。

［参考文献］

[1] Adadi, A., & Berrada, M.(2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)[J]. IEEE Access, 6: 52138-52160.

[2] Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F.(2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI[J]. Information Fusion, 58: 82-115.

[3] Du, M., Liu, N., Song, Q., & Hu, X. (2018). Towards explanation of DNN-based prediction with guided feature inversion[C]. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 1358-1367.

[4] Erhan, D., Bengio, Y., Courville, A., & Vincent, P.(2009). Visualizing higher-layer features of a deep network[J]. University of Montreal, 1341(3): 1.

[5] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D.(2018). A survey of methods for explaining black box models[J]. ACM Computing Surveys (CSUR), 51(5): 1-42.

[6] Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2019). Metrics for explainable AI: Challenges and prospects[J]. arXiv preprint arXiv: 1812.04608.

[7] Hwang, G. J., Yang, L. H., & Wang, S. Y.(2013). A concept

map-embedded educational computer game for improving students' learning performance in natural science courses[J]. Computers & Education, 69: 121-130.

[8] Kumar, S., Chakrabarti, S., & Roy, S. (2017). Earth mover's distance pooling over Siamese LSTMs for automatic short answer grading[C]. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2046-2052.

[9] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions[C]. Proceedings of the 31st International Conference on Neural Information Processing Systems, 30: 4768-4777.

[10] 刘桐,顾小清(2022). 走向可解释性:打开教育中人工智能的“黑盒”[J]. 中国电化教育,(5):82-90.

[11] 卢宇,骈扬,陈鹏鹤(2023). 新型智能导学系统构建及其关键技术 [J]. 中国远程教育,43 (7):30-38+46.

[12] 卢宇,章志,王德亮,陈鹏鹤,余胜泉(2022). 可解释人工智能在教育中的应用模式研究 [J]. 中国电化教育,(8):9-15+23.

[13] Rai, A.(2020). Explainable AI: From black box to glass box[J]. Journal of the Academy of Marketing Science, 48(1): 137-141.

[14] Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring systems: A systematic literature review. Artificial Intelligence Review, 55(3): 2495-2527.

[15] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier[C]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.

[16] Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps[J]. arXiv preprint arXiv: 1312.6034.

[17] 谭红叶,郭亚鑫,李茹(2022). 主观题自动批阅研究进展、应用与挑战 [J]. 人工智能,(2):14-20.

[18] 谭红叶,午泽鹏,卢宇,段庆龙,李茹,张虎(2019). 基于代表性答案选择与注意力机制的短答案自动评分 [J]. 中文信息学报,33 (11):134-142.

[19] 袁莉,曹梦莹,约翰·加德纳,迈克尔·奥利里(2021). 人工智能教育评估应用的潜力和局限 [J]. 开放教育研究,27 (5):4-14.

[20] 曾春艳,严康,王志锋,余琰,纪纯妹(2021). 深度学习模型可解释性研究综述 [J]. 计算机工程与应用,57 (8):1-9.

[21] 祝智庭,胡姣(2022). 教育数字化转型:面向未来的教育“转基因”工程 [J]. 开放教育研究,28 (5):12-19.

(编辑:赵晓丽)

Research on the Construction and Application of Explainable Automatic Scoring Model

LU Yu[1,2], ZHANG Zhi[1], MA Anyao[1] & CHEN Penghe[1,2]

(1. *School of Educational Technology, Faculty of Education, Beijing Normal University, Beijing* 100875, *China*; 2. *Advanced Innovation Center for Future Education, Beijing Normal University, Beijing* 102206, *China*)

Abstract: *Automatic scoring is an important realization form and basic function of intelligent educational evaluation. Deep learning-based automatic scoring models are gradually maturing but their internal structure is becoming increasingly complex and the decision-making process is not transparent, which makes it difficult for users to trust their score results and affects large-scale deployment. Therefore, this study proposes a basic framework for an interpretable automatic score model, which includes an automatic score base module, an automatic score interpretation module, and an automatic score interaction module. On this basis, this study constructs an example of the interpretable auto-score model and embeds it into an intelligent tutoring system, and further carries out quasi-experiments to verify its effectiveness. The experimental results show that the intelligent tutoring learning system embedded with the interpretable auto-score model effectively enhances learners' trust, and also contributes to their technology acceptance. Meanwhile, the explanatory information in the interaction module does not increase the cognitive load of learners. Finally, this study provides suggestions and outlooks for the research of explainable AI in educational areas.*

Key words: *automatic scoring*; *deep neural network*; *explainable artificial intelligence*; *human-computer interaction*; *intelligent tutoring system*