

基于E5-SHAP算法的可解释英语作文自动评分语言模型

王兵¹, 单瑞雪¹, 邢海燕², 李盼池¹

(1. 东北石油大学计算机与信息技术学院, 黑龙江 大庆 163318;

2. 东北石油大学机械科学与工程学院, 黑龙江 大庆 163318)

摘要: 针对英语作文自动评分系统因依赖复杂深度学习模型而缺乏可解释性的问题, 提出了一种基于E5-SHAP算法的可解释英语作文自动评分模型。该模型基于E5-Base模型编码器提取文本特征, 结合均值计算和回归层实现评分输出, 并引入自适应加权机制, 从语法、句法、词汇多样性等6个维度综合评估作文质量。模型采用LoRA微调技术优化特定层参数, 提高对作文特征的适应性。通过SHAP算法计算各特征对最终评分的影响, 从而提供清晰的评分依据和解释路径, 提升评分过程的透明性和可信度。实验结果表明, 与现有模型相比, 该模型在ELLIPSE数据集和自建数据集上的表现均有所提升, 二次加权卡帕值(QWK)达0.84, 在准确性和可解释性上优于现有模型。

关键词: 英语作文; 自动评分模型; E5-SHAP算法; 可解释性

中图分类号: TP391.1, G434

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202530

An interpretable automated essay scoring model for English compositions based on SHAP algorithm

WANG Bing¹, SHAN Ruixue¹, XING Haiyan², LI Panchi¹

1. School of Computer and Information Technology, Northeast Petroleum University, Daqing 163318, China

2. School of Mechanical Science and Engineering, Northeast Petroleum University, Daqing 163318, China

Abstract: In response to the lack of interpretability in English composition automatic scoring systems due to their reliance on complex deep learning models, an interpretable English composition automatic scoring model was proposed based on the E5-SHAP algorithm. This model was based on the E5 base model encoder to extract text features, combined with a mean calculation and a regression layer to achieve scoring output. It introduced an adaptive weighting mechanism to comprehensively evaluate the quality of compositions across six dimensions, including grammar, syntax, and vocabulary diversity. The model utilized LoRA fine-tuning technology to optimize specific layer parameters and enhance adaptability to compositional features. By using the SHAP algorithm to calculate the impact of each feature on the final score, a clear scoring basis and explanation path was provided to enhance the transparency and credibility of the scoring process. The experimental results show that compared with existing models, the performance of this model has been improved on both the ELLIPSE dataset and self-built dataset, with a quadratic weighted kappa value (QWK) of 0.84, which is superior to existing models in accuracy and interpretability.

Key words: English composition, automated essay scoring, E5-SHAP algorithm, interpretability

收稿日期: 2025-03-05; 修回日期: 2025-06-13

通信作者: 王兵, wangbing0812@sina.com

基金项目: 黑龙江省自然科学基金联合引导项目 (No. LH2024E012); 黑龙江省高等教育教学改革研究项目 (No. SJGYB2024479)

Foundation Items: Heilongjiang Natural Science Foundation Joint Guidance Project (No. LH2024E012), Heilongjiang Province Higher Education Teaching Reform Research Project (No. SJGYB2024479)

0 引言

英语作为最多国家使用的官方语言,也是世界上使用最广泛的第二语言^[1],在国际的政治、经济、文化方面被广泛应用,对于人们的工作、学习和交流都具有重要意义。其中,英语写作作为一种重要的输出形式和基本技能,受到越来越多语言研究者和教育学家的重视。我国非常重视英语教育,目前英语的教学工作主要集中在听、说、读、写、译5个方面,为了增强和培养学生的英语综合能力,英语写作在教学体系中更是占了很大的比重。英语作文是我国英语学习者的重要考试题型与重点考查方式,因此,培养学生的写作能力成为学生英语教学改革的热点。

随着现代化教育技术的发展,作文自动评分(automated essay scoring, AES)系统在英语写作教学中被广泛应用。教育的提升与变革必须依靠智能科学与技术,特别是当下火热的人工智能技术^[2]。作文自动评分利用计算机技术对作文进行评估与计分。与传统的教师批改学生英语作文教学模式相比,英语作文自动批改系统能极大地缓解学生在英语写作方面对教师的依赖,有效激发学生的自主学习动力和积极性,并且消除教师的主观因素,使批改结果更加客观公正。国内目前已经研发出许多英语作文自动批改系统,例如批改网、冰果英语等。但是大量反馈指出“批改网”给出的分数偏高^[3],在词汇和语法方面反馈较多,而在内容、逻辑、一致性和篇章结构方面反馈较少,并且目前大多数自动评分系统存在着“黑箱”问题,使教师和学生难以理解评分过程,质疑其公平性和可靠性。因此,本文提出一种基于E5-SHAP算法的英语作文自动评分模型,其采用LoRA(low-rank adaptation)技术微调后的E5-Base模型进行文本特征提取,再利用回归层生成多维度评分值,并引入自适应权重机制反映作文整体质量。为提高评分过程的可解释性和透明性,本文结合SHAP(Shapley additive explanations)可解释算法,利用合作博弈论来解释模型的预测,通过计算SHAP值评估作文质量的不同特征对模型分数预测的影响,提供对英语作文自动评分的解释。实验结果表明,该模型在多个维度上具有较高的评分准确率,二次加权卡帕值(QWK)达到0.84,验证了该模型在AES任务中的有效性。

1 相关工作

国内外学者对英语作文自动评分模型已经进行了大量研究,根据使用方法的的不同,本节将从基于特征工程模型、基于深度学习模型和基于预训练语言模型3个方面对前人的工作进行梳理和讨论。

早期的AES模型主要依赖于机器学习技术,特别是基于特征工程模型。这些系统通常需要人工设计大量可以从文本中提取出来并用于模型训练的特征。梁茂成等^[4]最早提出了AES的建模方法,并从英语文章中提取了语法错误、句法信息等特征,采用线性回归方法进行评分,取得了初步的成果。李晨亮等^[5]使用基于梯度相似性的多主题联合预训练方法和特征工程技术手工设计了3类特征,提升了自动作文评分模型的准确性和泛化能力。林美乐^[6]采用隐含狄利克雷分配、机器学习和特征工程技术,对英语作文进行多个特征提取和评分研究,构建了准确高效的评分模型。PEG(project essay grade)^[7]是国外最早的自动评分系统之一,它通过写作结构特征、词汇等浅层语义特征对作文进行评分。Chen等^[8]从文章内容、行文组织和整体结构3个方面构建了特征,并采用排序信息的方法对作文评分。

基于神经网络的深度学习模型在AES领域取得了许多研究成果。黄凯^[9]利用卷积神经网络(convolutional neural networks, CNN)获取语句的优美性特征,并融合主题特征对作文评分。栾笛^[10]使用长短时记忆网络(long short-term memory, LSTM)和CNN技术,完成了中考英文作文自动评分系统的设计与实现,提高了阅卷效率。周险兵等^[11]利用Doc2Vec模型提取主题信息,并结合句法特征和CNN、LSTM提取的语义信息,对英语作文进行评分。黎艳秋等^[12]结合GloVe词向量与CNN,构建了英语在线考试评分模型,实验结果显示,其评分一致性和准确性更优于传统方法。Alikaniotis等^[13]采用双层双向LSTM对作文进行自动评分。Jin等^[14]提出了两阶段的神经网络模型,在处理与主题无关的作文自动评分任务时取得了较好效果。Masaki等^[15]提出将手工设计的篇章级特征集成到基于神经网络的作文自动评分模型DNN-AES中,准确率有一定提高。

近年来,预训练模型被广泛应用于AES领域。陈宇航等^[16]提出的多维度特征增强的作文自动评

分方法,使用CNN、LSTM、基于变换器的双向编码器表示(bidirectional encoder representations from transformers, BERT)技术提取作文特征,实现自动评分,在公开数据集上性能显著提升。于明诚等^[17]使用XLNet预训练模型及多尺度上下文网络,提升作文自动评分性能,在公共数据集ASAP数据集上的平均QWK值达0.81。马钰等^[18]使用图卷积网络(graph convolutional network, GCN)和微调BERT技术完成作文自动评分任务,提出GFTB模型。实验表明,该方法在公共数据集ASAP上取得了较好性能,有效提升了作文自动评分的准确性。Song等^[19]提出了用于作文自动评分的预训练策略,分为3个阶段:弱监督预训练、跨主题微调,以及针对特定主题的微调。Abdulkarem等^[20]使用BERT技术改进科学领域自动作文评分任务,通过解决词汇外(out-of-vocabulary, OOV)和句子级嵌入问题,提高了评分的准确性和效率。

以上研究表明,基于特征工程和深度学习的作文评分模型能够有效提高作文自动评分的性能,基于预训练语言模型的特征分析能更准确地理解作文内容。有研究指出,大模型的“幻觉”问题实质反映出其推理逻辑尚未成熟,仅靠性能评估难以发现其潜在缺陷^[21]。因此本文引入SHAP可解释算法揭示模型在各评分维度中可能存在的偏差来源,从而避免“高精度-低可靠”的技术陷阱。

SHAP作为一种强大的可解释性算法,在国内外的机械、金融等领域已经取得了显著成果。Hamilton等^[22]使用SHAP框架和机器学习算法解决电力系统中影响瞬态稳定极限因素的研究问题,得出SHAP值可提供关于特征效应的解释,有助于理解整个稳定性边界的结果。Antonini等^[23]使用随机森林算法和SHAP解释方法解决地质岩石的分类问题,发现不仅能有效解释模型预测、提高模型透明度,而且有较高的分类准确率。Machlev等^[24]提出了一种结合SHAP、LIME等多种可解释人工智能(explainable artificial intelligence, XAI)技术的电能质量扰动分类器解释方法,通过评分和优化组合,提升分类器输出的透明度,帮助专家做出可靠决策。

以上研究证实SHAP框架中基于SHAP值的特征选择算法具有可行性,能够捕获复杂特征,并评估特征的重要性。通常AES模型基于复杂的神经网络结构,其内部决策过程对人们来说是不透明的,导致教师和学生质疑分数,很难根据评分结果

来改进教学和学习,SHAP算法的可解释性正好可以弥补此缺陷。因此,本文结合基于预训练语言模型E5-Base和可解释性算法SHAP,从多维语义的角度出发,提出了基于E5-SHAP的可解释英文作文自动评分模型。实验结果表明,该模型能够有效提高作文自动评分性能,并具有良好的可解释性。

2 E5-SHAP评分模型

2.1 模型整体框架

本文提出的E5-SHAP模型是一种创新的英语作文自动评分模型,该模型主要融合了E5-Base模型的文本表示能力、多维度自适应评分和SHAP可解释性分析。整体框架如图1所示,具体分为以下3个部分。

(1) 文本特征提取

E5-Base模型是一种先进的文本嵌入模型,通过高质量数据对比学习和强大的BERT基底,能够生成文本的深层次文本表示。因此本文基于E5-Base模型,通过LoRA技术进行高效微调,降低参数量,并提升模型适配能力。微调后的E5-Base模型能够适应作文评分任务,捕捉文本的多维度特征。

(2) 多维度自适应评分

在多维度评分过程中,E5-SHAP模型加载6个基于E5-Base预训练模型微调的子模型,6个子模型分别对应6个评分维度。每个子模型提取输入文本在对应维度的特征,再添加一个线性回归层,将各个维度的特征向量映射到标准化的评分范围内。在此基础上,引入自适应权重机制,通过动态权重调整策略,为不同评分维度分配合理的权重。最后将6个评分维度的得分整合为最终的综合得分。

(3) 评分可解释性分析

为了解决英语作文自动评分模型缺乏可解释性的问题,本文引入SHAP可解释性算法,对6个评分维度分别构建解释模型,利用SHAP算法量化各输入特征对模型输出分数的正负影响,通过SHAP力图实现解释可视化,从而直观展示特征与分数的因果关系。

2.2 模型评分过程

2.2.1 数据预处理

数据预处理是构建英语作文自动评分模型的第一步,首先要对输入的英语作文进行文本清洗和分词,去除无关信息,处理可能的乱码、特殊字符;再将作文文本截断,转换为适合模型输入的格式;

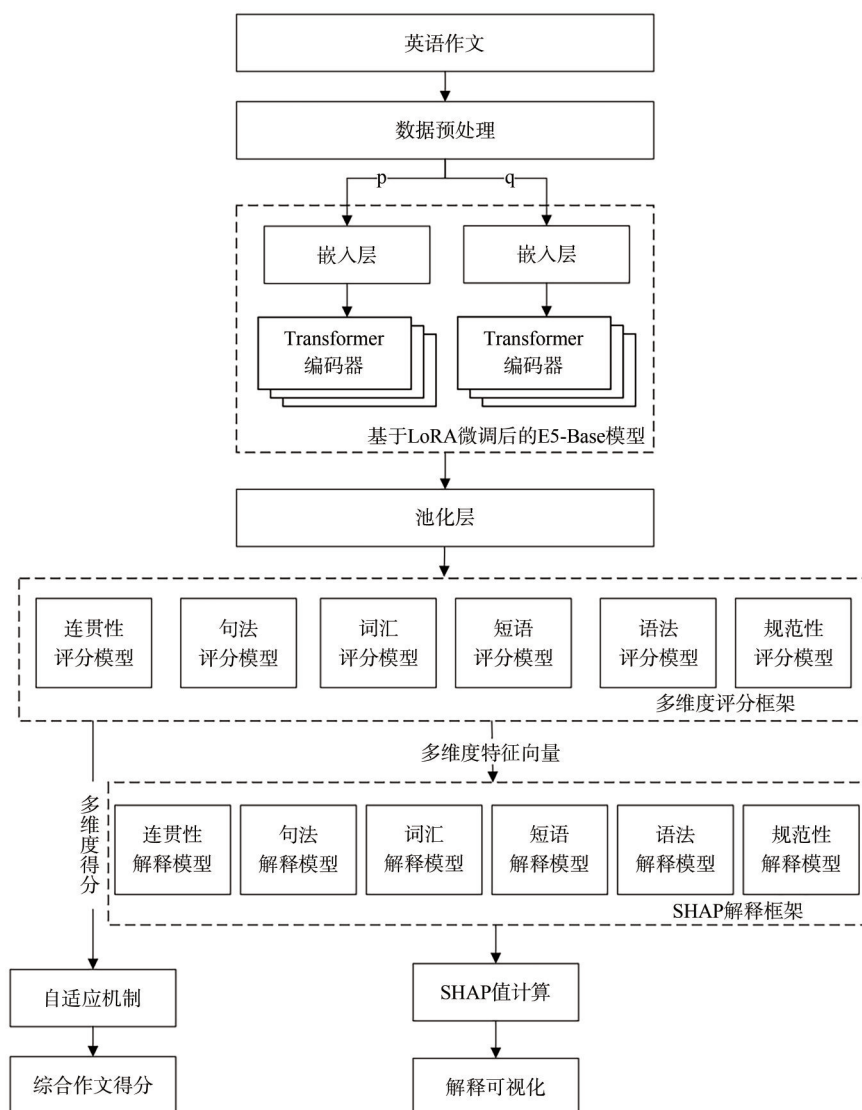


图1 E5-SHAP模型框架

最后将每篇作文与其评分进行配对。每篇作文与其在6个维度上的评分对应，形成一个 (q, p) 对，其中 p 表示作文文本， q 表示各个评分维度的值。每篇作文描述为分数-作文文本对 (q, p) ，能够使E5-SHAP模型的相似性度量更具针对性。

2.2.2 文本特征提取

在构建英语作文自动评分模型时，文本特征提取是核心步骤之一。本文采用LoRA方法对E5-Base模型进行高效微调。LoRA通过微调E5-Base模型的部分参数，能够保留E5-Base模型原有的强大语言能力，适应作文评分任务，提升模型的表现。同时，相比传统的全参数微调方法，冻结预训练模型的大部分权重，仅通过插入低秩矩阵进行参数更新，显著降低了计算开销，并减少了模型的训练时

间，能够达到更加精确和高效的文本特征提取。

E5-Base模型采用多层自注意力机制和前馈神经网络来处理输入文本。在自注意力层中，模型通过计算Query和Key之间的相似度来捕捉输入文本中不同部分之间的相关性。为了使模型更好地适应作文评分任务，本文在E5-Base模型的自注意力层中对Query和Key的投影矩阵进行微调。在微调过程中，对于输入向量 X ，调整后的Query向量和Key向量分别计算如下：

$$Q = XW_Q^{\text{LoRA}} = X(W_Q + A_Q B_Q) \quad (1)$$

$$K = XW_K^{\text{LoRA}} = X(W_K + A_K B_K) \quad (2)$$

其中， A_Q 和 B_Q 为低秩矩阵，用于Query投影， A_K 和 B_K 用于Key投影， W 是预训练的E5-Base自注意

层中对应投影（Query或Key）的权重矩阵。

通过这一操作，LoRA在保持原始参数的基础上对Query和Key向量进行了微调，以适应作文评分任务的特定需求。在微调过程中还使用了Adam优化器，结合权重衰减来防止模型过拟合。然后将预处理好的数据输入微调后的E5-Base模型中，对作文的Token序列 $T=(t_1, t_2, \dots, t_n)$ 进行特征提取，每个Token经过嵌入层和编码层处理后会生成一个表示向量 h_i ，经过E5-Base编码后得到的Token的表示矩阵 H 为：

$$H = (h_1, h_2, \dots, h_n), h_i \in \mathbb{R}^d \quad (3)$$

其中， h_i 为第 i 个Token的表示， d 是模型的嵌入维度， $\mathbb{R}^d$ 表示 d 维实数向量空间。

由于每个 h_i 仅包含该Token的局部信息，因此本文在E5-Base模型输出层之后使用均值池化方法计算所有Token的表示向量的平均值，得到全局特征表示 h_{pooled} ：

$$h_{\text{pooled}} = \frac{1}{n} \sum_{i=1}^n h_i \quad (4)$$

其中， n 为Token数量。

池化层的操作可以将不同长度的Token序列映射为一个固定维度的全局表示，从而为后续的评分阶段提供紧凑而全面的特征表示。

2.2.3 多维度自适应评分

E5-SHAP模型采用多维度自适应评分方法，通过构建多维度评分框架和自适应加权机制，实现对文本连贯性、句法结构、词汇复杂性、短语用法、语法正确性和作文规范性这6个维度的独立分析，并生成综合得分。以下将详细阐述评分原理和实现步骤。

(1) 多维度评分框架

在文本特征提取模块中，E5-Base模型经过池化层操作，生成了固定维度的全局特征表示 h_{pooled} 。对于每个评分维度 $i \in \{1, 2, \dots, 6\}$ ，定义一个独立的线性回归层，将 h_{pooled} 映射到目标评分范围：

$$y_i = \sum_{j=1}^d w_i^{(j)} h_{\text{pooled}}^{(j)} + b_i \quad (5)$$

其中， $w_i^{(j)}$ 表示权重向量 w_i 中第 j 个元素， $h_{\text{pooled}}^{(j)}$ 表示全局特征向量 h_{pooled} 中第 j 个元素， b_i 是第 i 个维度的偏置项， y_i 是第 i 个维度的预测评分。

通过回归层将输入作文的全局特征表示映射到评分空间，生成6个独立的初始评分：

$$y = [y_1, y_2, \dots, y_6] \quad (6)$$

模型采用均方误差（MSE）作为损失函数：

$$L_{\text{MSE}} = (y_i - y_i^{\text{true}})^2 \quad (7)$$

其中， y_i^{true} 表示人工标注的真实分数。

通过最小化 L_{MSE} ，模型可以找到合适的参数。通过优化上述目标函数，模型能够在每个评分维度上生成可靠的初始评分。这些评分直接反映了作文在不同维度上的表现，并且每个评分维度的结果都是独立的，为自适应加权机制提供输入。

(2) 自适应加权机制

自适应加权机制是基于多维度评分的初始结果，通过动态优化各维度权重来生成最终的综合得分。此过程采用梯度优化方法（Adam优化算法），不需要手动指定每个维度的权重，模型可以根据训练数据优化各个维度的权重。通过动态优化权重，模型能够在不同的评分维度之间找到最佳平衡，避免某些维度的评分影响过大或过小。本文通过定义目标函数和约束条件，逐步更新权重以最小化误差，具体步骤如下。

首先输入作文在6个评分维度上的初始评分 $y_1, y_2, y_3, y_4, y_5, y_6$ ，计算评分的均值并定义损失函数 L 为：

$$\bar{y} = \frac{1}{6} \sum_{i=1}^6 y_i \quad (8)$$

$$L = \sum_{i=1}^6 ((y_i \cdot w_i) - \bar{y})^2 \quad (9)$$

其中， $\bar{y}$ 是所有模型评分的平均值，作为每次评分的参考基准。

为使每个评分维度的权重能够动态调整，本文为每个评分维度 $i \in \{1, 2, \dots, 6\}$ 定义一个可学习的权重参数 θ_i ，为确保权重总和为1并满足非负性值，权重通过Softmax函数进行归一化处理，计算式如下：

$$w_i = \frac{\exp(\theta_i)}{\sum_{j=1}^6 \exp(\theta_j)} \quad (10)$$

其中， θ_i 是待学习的模型参数。

权重生成过程由Softmax函数实现，使模型能够动态调整各维度的重要性。具体优化步骤如下：

- 初始化权重参数 $\theta = \{\theta_1, \theta_2, \dots, \theta_6\}$ 为小随机值或均匀分布；
- 使用Adam优化算法，通过反向传播逐步

优化损失函数 L ;

- 在每次迭代中, 根据当前的 θ 计算权重 w_i , 并据此更新模型参数。

经过训练后, 可获得6个评分维度的最终权重 $w = \{w_1, w_2, \dots, w_6\}$, 最终的综合评分 G_{final} 计算式如下:

$$G_{\text{final}} = 6 \times \sum_{i=1}^6 w_i y_i \quad (11)$$

以上评分结果有效结合了多维度评分与自适应加权机制, 生成的综合评分更加准确且具有实际意义。

2.3 模型可解释性分析

为了解决英语作文自动评分模型缺乏可解释性的问题, 本文在6个LoRA微调后的E5-Base子模型基础上引入了SHAP可解释性方法。与其他可解释性方法相比, SHAP能够提供一致性和局部精确性保证, 适用于复杂神经网络结构, 且支持对不同输入维度的重要性进行统一量化。LIME (local interpretable model-agnostic explanations) 等方法也能提供局部解释, 但其解释结果受邻域采样策略影响较大, 稳定性和一致性相对较弱, 且不适用于多维统一解释的场景。因此, SHAP更适合本评分系统在高维度、深层模型架构下的解释需求。具体地讲, 将每篇作文经过6个子模型后所得的6个维度特征向量, 分别作为SHAP模块的输入特征, 其输出为分析这些维度得分对最终综合评分的影响贡献, 即SHAP值, 从而明确指出特定词汇、句子结构或段落逻辑等对分数的正面或负面影响, 并提供可视化结果, 使评分过程透明化, 增强评分的可解释性。

SHAP算法通过一个线性函数来近似原始模型 f 的输出。这个线性函数 g 可以表示为:

$$g(z) = \varphi_0 + \sum_{i=1}^M \varphi_i z_i \quad (12)$$

其中, φ_0 是基线值, 表示无特征输入时的预测值; M 是特征的总数; φ_i 是特征 i 的SHAP值, 表示该特征对模型预测结果的贡献; z_i 是二进制变量, 表示特征 i 是否存在。

由此可见SHAP值可以解释为特征对预测结果的影响, 在E5-SHAP英语作文评分模型中, 可以理解为某个特征影响作文的最终得分。SHAP值可以通过求解以下方程得到:

$$\varphi_i = \sum_{S \subseteq \{1, \dots, M\} \setminus i} \frac{|S|!(M-|S|-1)!}{(M-1)!} [f_x(S \cup \{i\}) - f_x(S)] \quad (13)$$

其中, $f_x(S)$ 表示输入作文文本 x 和特征集合 S 时模型的预测值, $S \subseteq \{1, \dots, M\} \setminus i$ 表示不包含特征 i 的所有特征子集。这个方程计算了特征 i 在所有可能的特征子集中的边际贡献。

在E5-SHAP英语作文自动评分模型中, 具体解释过程如下: 将英语作文中的各个词汇作为输入特征 i , 作文的得分作为预测目标; 使用SHAP框架, 得到每个特征 i 对作文得分的贡献即SHAP值, 并据此进行解释和分析。

- 计算基准预测值: 首先计算所有特征的平均预测值, 作为基准预测值 $E[f(x)]$ 。
- 计算特征贡献: 对于每个特征 i , 计算它在所有可能的特征子集 S 中的边际贡献, 并使用上述计算式计算其SHAP值 φ_i 。
- 多维度解释评分: SHAP值 φ_i 直接表明特征 i 对作文得分的贡献。 $\varphi_i > 0$, 表示特征 i 对得分有正向贡献; $\varphi_i < 0$, 表示特征 i 对得分有负向影响。分析各个特征的SHAP值, 可以了解哪些词汇、短语或其他特征对最终评分的影响最大, 从而解释评分的依据。

通过计算本文6个维度特征和各个词汇特征的SHAP值, 可以了解每个特征对英语作文评分的具体贡献, 同时SHAP库提供了多种可视化工具, 本文使用SHAP力图展示不同特征在预测评分时的贡献和作用, 从而有助于直观地理解模型如何做出评分决策。

3 实验与结果分析

本节将介绍实验数据集和评估方法, 对实验配置进行说明, 以及对本文提出的模型与E5-Base模型和其他AES模型及闭源模型的性能进行对比分析, 最后通过SHAP力图解释E5-SHAP模型的评分依据。

3.1 数据来源和评价指标

为验证基于E5-SHAP算法的英文作文自动评分模型的性能表现, 本节使用3个数据集。数据集1来自ELLIPSE语料库。ELLIPSE语料库是一个免费的资源, 包含来自8~12年级英语语言学习者的大约9 000个写作样本, 对这些样本在整体语言能力

和分析能力的各个方面进行评分,包括连贯性、句法、词汇、短语用法、语法和规范性6个维度。数据集2来自国际英语考试(IELTS/TOEFL)历史作文题库,覆盖2018—2024年考生匿名数据,其中包含10 200篇作文(IELTS 6 000篇,TOEFL 4 200篇),评分维度与上述一致。数据集3自于大学英语作文随堂考试,由大学教师对上述6个维度进行评分,大约12 000个写作样本,数据集3基本信息见表1。

表1 大学英语作文数据集

子集	作文题目	平均长度/词	作文数目/篇
1	How to succeed in university	300	3 003
2	How to manage time in university	300	2 197
3	How to deal with stress in college	300	2 020
4	Happiness	200	2 923
5	Teenager smoking	200	1 982

为了与多数英语作文评分方法保持一致,本文采用精确度、F1值、QWK等作为评价指标。其中F1值是衡量分类模型性能的一个重要指标,它综合了精确率和召回率(Recall)两个方面的性能。F1值如式(14)所示:

$$F1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

其中, Precision表示模型预测为正类别的样本中,真正为正类别的比例; Recall则表示真正为正类别的样本中,被模型预测为正类别的比例。F1值通过计算Precision和Recall的调和平均数,能够更全面地反映模型的效果。

QWK是一种一致性检验方法,用于评估模型的结果与实际结果是否具有一致性。设作文的评分可以划分为 N 个等级,则QWK如式(15)和式(16)所示:

$$QWK = 1 - \frac{\sum_{i,j} W_{ij} O_{ij}}{\sum_{i,j} W_{ij} E_{ij}} \quad (15)$$

$$W_{ij} = \frac{(i-j)^2}{(N-i)^2} \quad (16)$$

其中, O 为一个 n 阶直方图矩阵, O_{ij} 表示专家评分为 i 且模型评分为 j 的作文数量, W 表示基于专家评分和模型评分的二次加权矩阵, E_{ij} 表示专家评分为 i 且模型评分为 j 的概率的乘积,其中需要对 E_{ij} 和 O_{ij} 进行归一化处理。

3.2 实验配置

本文的实验环境配置如下。

- 硬件环境: NVIDIA GeForce RTX 4090 GPU。
- 软件环境: 包括Anaconda用于管理Python环境,核心框架为PyTorch 2.4.1, Transformers库版本为4.40.1, CUDA版本为12.6。

为验证E5-SHAP模型的性能表现,本研究将数据集按8:1:1划分为训练集、测试集和验证集,以保证模型具备良好的泛化能力。参数设置方面,采用较小的学习率(0.000 4)和较长的训练轮次(37个epoch),充分训练模型并降低过拟合风险。设置LoRA参数rank为8, alpha为2,训练过程中启用bf16精度以加快速度且不影响精度。优化器选用AdamW,设置betas=(0.9, 0.999)和epsilon= 1×10^{-8} ,以提升训练稳定性。其他参数见表2。

表2 E5-SHAP参数设置

参数名	参数值
per_device_train_batch_size	8
per_device_eval_batch_size	8
dropout	0.1
logging_steps	1 000
evaluation_strategy	epoch
warmup_steps	500
save_strategy	epoch

3.3 实验结果分析

3.3.1 总体性能评估

本文对比了E5-SHAP模型与E5-Base模型在数据集1、数据集2和数据集3上的6个作文评分维度的性能,具体结果见表3~表5。

表3 在数据集1上使用多个评估标准的模型性能

维度	精确度		F1值		QWK	
	E5-Base	E5-SHAP	E5-Base	E5-SHAP	E5-Base	E5-SHAP
连贯性	0.77	0.83	0.65	0.86	0.63	0.84
句法	0.73	0.92	0.71	0.90	0.69	0.86
词汇	0.66	0.94	0.64	0.94	0.61	0.80
短语用法	0.78	0.74	0.76	0.72	0.74	0.72
语法	0.62	0.91	0.60	0.88	0.58	0.88
规范性	0.72	0.80	0.70	0.84	0.68	0.85

实验结果表明, E5-SHAP模型在3个数据集的6个评分维度上均优于E5-Base,尤其在语法、规范性与连贯性等语言结构与篇章层面要求较高的维度上,精确度、F1值及QWK均提升显著(如数据

表4 在数据集2上使用多个评估标准的模型性能

维度	精确度		F1值		QWK	
	E5-Base	E5-SHAP	E5-Base	E5-SHAP	E5-Base	E5-SHAP
连贯性	0.68	0.88	0.66	0.80	0.64	0.82
句法	0.65	0.86	0.63	0.84	0.61	0.83
词汇	0.60	0.85	0.60	0.73	0.56	0.80
短语用法	0.70	0.75	0.58	0.72	0.66	0.86
语法	0.58	0.86	0.56	0.73	0.61	0.80
规范性	0.65	0.85	0.63	0.83	0.65	0.81

表5 在数据集3上使用多个评估标准的模型性能

维度	精确度		F1值		QWK	
	E5-Base	E5-SHAP	E5-Base	E5-SHAP	E5-Base	E5-SHAP
连贯性	0.62	0.91	0.61	0.89	0.60	0.90
句法	0.63	0.92	0.62	0.90	0.61	0.89
词汇	0.61	0.90	0.60	0.88	0.58	0.90
短语用法	0.76	0.78	0.74	0.75	0.72	0.78
语法	0.65	0.93	0.63	0.92	0.61	0.85
规范性	0.68	0.95	0.67	0.93	0.65	0.87

集3中语法维度QWK由0.61升至0.85)。相较之下,短语用法维度提升不显著,这与该维度的表达模式多样性及语义边界模糊性相关,限制了模型判别能力的提升。E5-SHAP在整体表现的提升,核心在于引入了LoRA微调策略,使基础模型能够在有限训练数据下高效适应多维评分任务。相比原始模型,LoRA微调增强了对文本语义细节与上下文依赖的捕捉能力,从而提高了模型在各维度评分中的稳定性与一致性。

3.3.2 消融实验结果

本文设计了消融实验,以验证各组成模块对整体评分效果的影响,其中包括原始的E5-Base模型、LoRA微调后的E5-Base模型、基于固定权重的6个LoRA微调后的E5-Base子模型、基于自适应加权的6个LoRA微调后的E5-Base子模型,即本文提出的E5-SHAP模型,具体实验结果见表6。

实验结果表明,不同模型结构在英语作文评分任务中的性能差异显著。基础模型E5-Base在数据集2上的QWK表现较弱,表明其在捕捉写作特征

表6 E5-SHAP模型消融实验QWK性能对比

模型	数据集1	数据集2	数据集3
E5-Base	0.73	0.65	0.80
LoRA微调后模型	0.76	0.78	0.83
LoRA+固定权重模型	0.79	0.80	0.86
LoRA+自适应加权模型	0.83	0.81	0.89

方面存在局限。引入LoRA微调后,模型的特征表示能力明显增强,QWK指标普遍提升。进一步通过6个子模型的固定权重融合,模型整体性能得到提升,QWK最高达0.86,说明多维协同评分有助于提升判别能力。然而固定权重机制难以动态适应不同作文维度的重要性差异。因此,本文提出自适应加权机制。实验显示,该机制进一步提升了模型性能,QWK最高达到0.89,验证了其在处理维度间语义耦合与不平衡性方面的有效性。

E5-SHAP模型引入自适应加权机制以集成6个评分子模块,并结合SHAP可解释性框架。相较于基础的单一编码器模型E5-Base,该模型在计算复杂度上有所增加。在本文实验配置下,单篇作文的平均处理时长为1.8~2.3 s,整体时间处于可接受范围。实验结果表明,E5-SHAP在提升评分精度的同时兼顾了解释性与运行效率,验证了其结构设计合理性。

3.3.3 与其他AES模型性能对比

为了更全面地评估模型的性能,本文还将E5-SHAP模型与其他AES模型及闭源模型进行了对比,结果见表7。

表7 传统AES模型及闭源模型的QWK性能指标

模型	QWK
CNN ^[25]	0.31
LSTM ^[25]	0.40
CNN+LSTM ^[11]	0.79
DNN-AES ^[15]	0.80
BERT ^[16]	0.80
XLNET ^[17]	0.81
GFTB ^[18]	0.79
MotionGPT ^[26]	0.72
GPT-4V ^[27]	0.63
Llama ^[28]	0.83
E5-SHAP	0.84

表7汇总了传统AES方法及闭源大模型在QWK指标下的整体评分性能。传统深度学习方法(如CNN、LSTM)由于语义建模能力受限,整体QWK较低,难以有效捕捉写作中的长距离依赖关系。BERT、XLNet等通用预训练模型虽具备一定的语义理解能力,QWK可达0.80左右,但未针对作文评分任务进行结构适配,维度感知不足。闭源大模型(如GPT-4V、LLaMA)表现提升但资源成本高,维度可解释性弱,难以推广。本文提出的E5-SHAP模型在结构上融合了LoRA微调、多维度子模型和自适应加

权机制, QWK指标达到了0.84, 而且资源消耗低。

3.3.4 SHAP 力图解释分数

为了进一步验证模型的可靠性与可解释性, 本文基于 SHAP 力图对 E5-SHAP 模型的评分结果进行解释与可视化分析。

图2展示了 SHAP 值在语法维度的特征贡献。其中, youa、allowsthem 等拼写或结构错误被标蓝表示负贡献; 语法结构清晰的表达 (如 phones are not allowed in the school) 则以红色高亮显示, 代表正向提升评分。图3句子连贯性特征中, 多次出现关键词衔接 (如 students、phones) 和逻辑连接词 (如 However、Also) 保持主题连贯, 并标红表示正贡献, 显著提升了整体得分。

为了评估 SHAP 方法在英语作文自动评分任务中的适用性与解释质量, 本文引入两项可解释性评估指标——Faithfulness (忠实性) 与 Stability (稳定性), 分别用于衡量模型解释结果对关键特征的依赖程度和其在轻度语义扰动下的一致性。

图4中 Faithfulness 曲线展示了在实验中, 通过逐步移除 SHAP 值最高的词语并记录模型评分的变

化情况, 结果表明得分下降趋势明显, 6个维度的平均 AUC (area under the curve) 为0.83, 说明模型的预测依赖 SHAP 识别的关键特征。图4中 SHAP 稳定性热图为在 Stability 实验中, 通过对作文进行轻度语义扰动 (如近义词替换、语序调整), 评估扰动前后 SHAP 分布的相似性, 6个维度的平均皮尔逊相关系数达到0.84, 验证了解释结果在语义等价条件下具备较强的稳定性。

由以上实验结果可得, E5-SHAP 模型不仅在评分方面准确性较高, 其解释结果也具备忠实性与稳定性。

4 结束语

本文致力于解决英语作文自动评分系统中的准确性与可解释性问题, 通过引入基于 E5-SHAP 算法的英语作文自动评分模型, 实现了对英语作文更准确且易于理解的评分。该模型利用 E5-Base 预训练语言模型的深度语义理解和文本表示能力, 有效提取了作文中的关键特征。本文引入 LoRA 技术实现高效微调, 在降低参数冗余的同时, 提升模型在

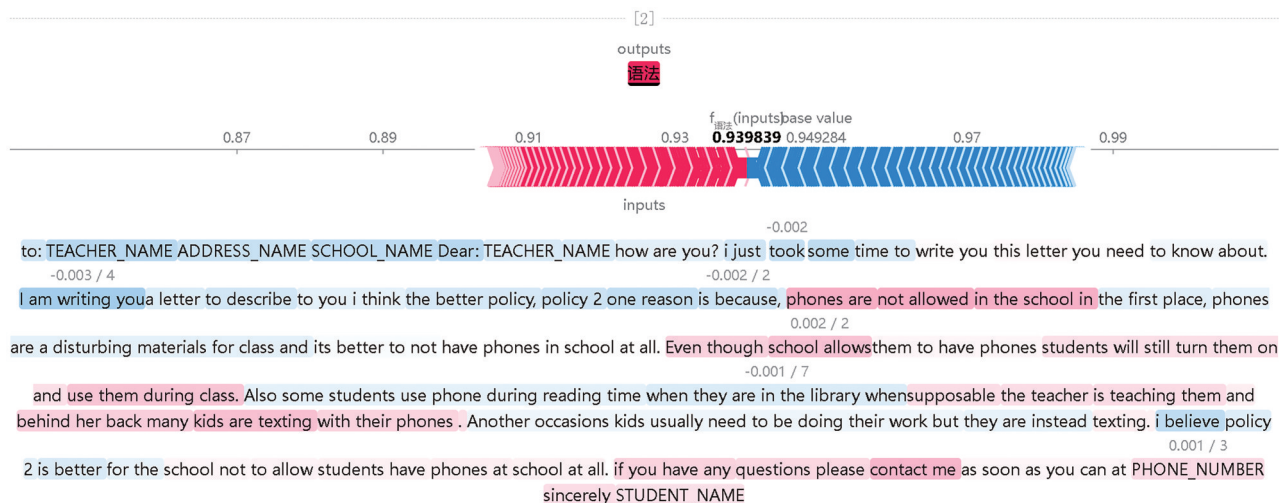


图2 SHAP 力图解释语法评分

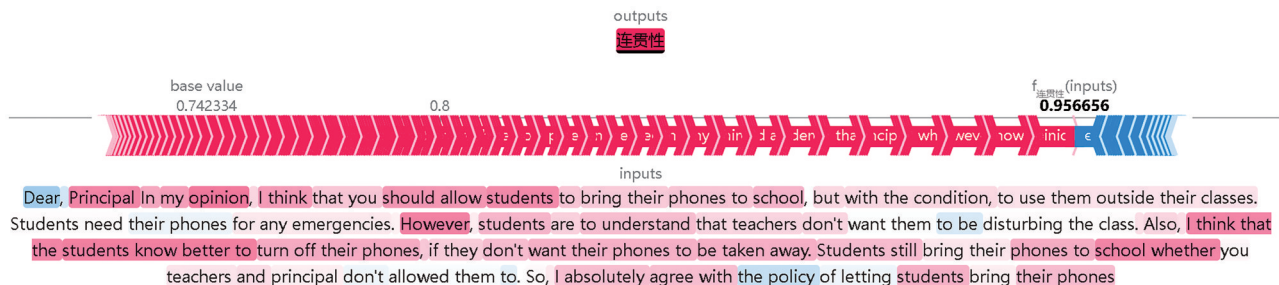
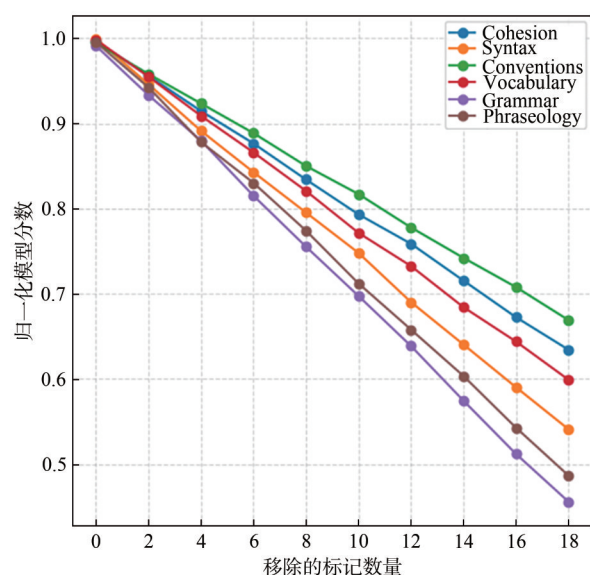
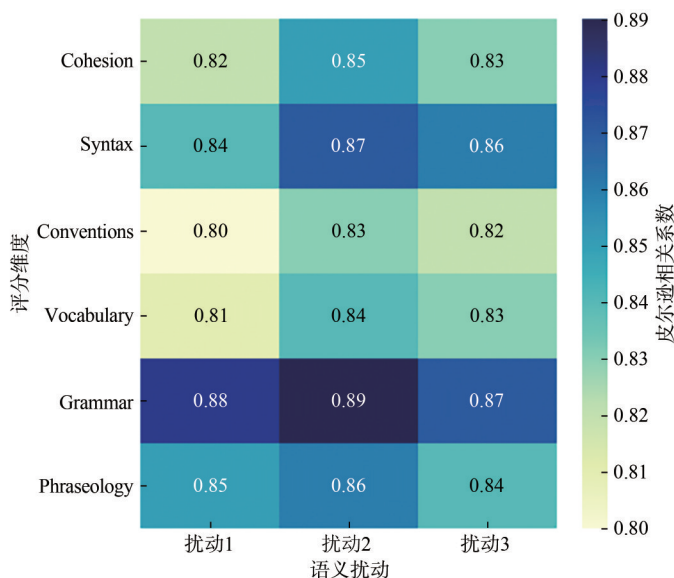


图3 SHAP 力图解释连贯性评分



(a) 各评分维度的Faithfulness曲线



(b) 语义扰动下的SHAP稳定性

图4 Faithfulness曲线与SHAP稳定性热图

评分任务中的适应能力。为了提取作文整体语义信息，添加一个均值池化层将各个词的表示压缩为一个固定大小的整体表示，从而有效地捕捉作文的整体特征。最后通过一个线性回归层来实现并从多维度对作文进行打分，并通过SHAP值分析所得分数，从而提供了清晰的评分依据和决策路径。实验结果表明，与传统的英语作文自动评分系统相比，本文提出的E5-SHAP模型在整体评分上表现更优，特别是在QWK这一指标上展现出更高的准确性。同时通过SHAP算法的应用，还可得到每个特征对模型分数预测的影响。除此之外，此次研究仍存在进一步优化的空间。E5-SHAP模型主要针对结构规范、语言表达较标准的英语作文设计，训练数据以标准考试和课堂作文为主，因而在处理结构自由、表达个性化较强的创意写作时，模型表现可能存在一定偏差，评分和解释的稳定性有所下降。这主要源于训练数据的分布限制以及现有模型对非典型语言风格的建模能力不足。未来研究中，计划结合风格感知机制及多模态信息，增强模型在多样化写作风格中的适应性和鲁棒性。相信在不断的研究和努力下，英语作文自动评分系统将会在教育领域发挥更加重要的作用，为学生的英语写作能力的提升和教学质量的改善贡献更大的力量。

参考文献：

[1] 陶雯馨, 俞万杰. 浅析全球化背景下的空间站语言及中国空间站语言传播策略[J]. 现代语言学, 2023, 11(6): 2510-2515.

TAO W X, YU W J. Analysis of space station language in the context of globalization and Chinese space station language communication strategy[J]. Modern Linguistics, 2023, 11(6): 2510-2515.

[2] 王飞跃. 智能科技与K21教育：未来社会的未来学校与未来师生[J]. 智能科学与技术学报, 2024, 6(3): 281-283.

WANG F Y. K21 education and intelligent sciences: future schools and teachers/students in future society[J]. Chinese Journal of Intelligent Science and Technology, 2024, 6(3): 281-283.

[3] 于洲, 时真妹. 二语写作中iWrite机评反馈和教师反馈的对比研究[J]. 创新教育研究, 2022, 10(2): 214-223.

YU Z, SHI Z M. A comparative study of iWrite machine feedback and teacher feedback in second language writing[J]. Creative Education Studies, 2022, 10(2): 214-223.

[4] 梁茂成, 文秋芳. 国外作文自动评分系统评述及启示[J]. 外语电化教学, 2007(5): 18-24.

LIANG M C, WEN Q F. A critical review and implications of some automated essay scoring systems[J]. Computer-Assisted Foreign Language Education, 2007(5): 18-24.

[5] 李晨亮, 吴鸿涛. 基于梯度相似性的自动作文评分多主题联合预训练方法[J]. 电子科技大学学报, 2022, 51(4): 558-564.

LI C L, WU H T. A gradient-similarity based multi-topic jointly pre-training method for automated essay scoring[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(4): 558-564.

[6] 林美乐. 基于隐含狄利克雷分配和机器学习的英语作文自动评分研究[D]. 上海: 上海师范大学, 2024: 1-57.

LIN M L. Research on Automated English Essay Scoring Based on Latent Dirichlet Allocation and Machine Learning[D]. Shanghai: Shanghai Normal University, 2024: 1-57.

[7] PAGE E B. Grading essays by computer: progress report[C]//Proceedings of the 1967 Invitational Conference on Testing Problems. Princeton: Educational Testing Service, 1967: 87-100.

[8] CHEN H B, XU J G, HE B. Automated essay scoring by capturing relative writing Quality¹[J]. The Computer Journal, 2014, 57(9): 1318-1330.

[9] 黄凯. 英语作文自动评分关键技术的研究与实现[D]. 武汉: 华中师范大学, 2019: 1-53.

- HUANG K. Research and implementation of the key technology of English automatic essay scoring[D]. Wuhan: Central China Normal University, 2019: 1-59.
- [10] 栾笛. 中考英文作文自动评分系统设计与实现[D]. 武汉: 华中科技大学, 2021: 1-65.
- LUAN D. Design and implementation of an automated English essay scoring system for the high school entrance examination[D]. Wuhan: Huazhong University of Science and Technology, 2021: 1-65.
- [11] 周险兵, 樊小超, 任鸽, 等. 基于多层次语义特征的英文作文自动评分方法[J]. 计算机应用, 2021, 41(8): 2205-2211.
- ZHOU X B, FAN X C, REN G, et al. Automated English essay scoring method based on multi-level semantic features[J]. Journal of Computer Applications, 2021, 41(8): 2205-2211.
- [12] 黎秋艳, 刘佳伟, 王鹏, 等. 基于 GloVe-CNN 算法的英语在线考试主观题自动评分模型[J]. 桂林理工大学学报, 2023, 43(1): 155-160.
- LI Q Y, LIU J Y, WANG P, et al. Automatic scoring model of English online examination subjective questions based on GloVe-CNN algorithm[J]. Journal of Guilin University of Technology, 2023, 43(1): 155-160.
- [13] ALIKANIOTIS D, YANNAKOUDAKIS H, REI M. Automatic text scoring using neural networks[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2016: 715-725.
- [14] JIN C C, HE B, HUI K, et al. TDNN: a two-stage deep neural network for prompt-independent automated essay scoring[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stroudsburg: ACL, 2018: 1088-1097.
- [15] UTO M, XIE Y K, UENO M. Neural automated essay scoring incorporating handcrafted features[C]//Proceedings of the 28th International Conference on Computational Linguistics. International Committee on Computational Linguistics, 2020: 6077-6088.
- [16] 陈宇航, 杨勇, 帕力旦·吐尔逊. 多维度特征增强的作文自动评分[J]. 新疆师范大学学报(自然科学版), 2023, 42(3): 43-49, 58.
- CHEN Y H, YANG Y, PALIDAN T E X. Enhance multi-dimensional features for automatic essay scoring[J]. Journal of Xinjiang Normal University (Natural Sciences Edition), 2023, 42(3): 43-49, 58.
- [17] 于明诚, 党亚固, 吴奇林, 等. 基于多尺度上下文的英文作文自动评分研究[J]. 计算机工程, 2024, 50(3): 259-266.
- YU M C, DANG Y G, WU Q L, et al. Research on automatic scoring for English essay based on multi-scale context[J]. Computer Engineering, 2024, 50(3): 259-266.
- [18] 马钰, 杨勇, 任鸽, 等. 基于 GCN 和微调 BERT 的作文自动评分方法[J]. 计算机与现代化, 2024(9): 33-37, 44.
- MA Y, YANG Y, REN G, et al. Automated essay scoring method based on GCN and fine tuned BERT[J]. Computer and Modernization, 2024(9): 33-37, 44.
- [19] SONG W, ZHANG K, FU R J, et al. Multi-stage pre-training for automated Chinese essay scoring[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2020: 6723-6733.
- [20] ABDULKAREM A, KRIVTSUN A. Bert embedding and scoring for scientific automatic essay grading[C]//Proceedings of the 12th International Conference on Applied Innovations in IT. [S.l.:s.n.], 2024: 115-121.
- [21] 王飞跃. DeepSeek 呼唤 DeepThink: 重视 AI 治理与社会范式变革[J]. 智能科学与技术学报, 2025, 7(1): 1-3.
- WANG F Y. DeepSeek calls DeepThink: rethinking AI governance and societal paradigm shift[J]. Chinese Journal of Intelligent Science and Technology, 2025, 7(1): 1-3.
- [22] HAMILTON R I, PAPADOPOULOS P N. Using SHAP values and machine learning to understand trends in the transient stability limit[J]. IEEE Transactions on Power Systems, 2024, 39(1): 1384-1397.
- [23] ANTONINI A S, TANZOLA J, ASIAIN L, et al. Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task[J]. Applied Computing and Geosciences, 2024, 23: 100178.
- [24] MACHLEV R, PERL M, BELIKOV J, et al. Measuring explainability and trustworthiness of power quality disturbances classifiers using XAI: explainable artificial intelligence[J]. IEEE Transactions on Industrial Informatics, 2022, 18(8): 5127-5137.
- [25] TAGHIPOUR K, NG H T. A neural approach to automated essay scoring[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2016: 1882-1891.
- [26] JIANG B, CHEN X, LIU W, et al. Motiongpt: human motion as a foreign language[J]. Advances in Neural Information Processing Systems, 2023, 36: 20067-20079.
- [27] LEE G G, ZHAI X M. NERIF: GPT-4V for automatic scoring of drawn models[J]. arXiv preprint, 2023, arXiv: 2311.12990.
- [28] MANSOUR W, ALBATARNI S, ELTANBOULY S, et al. Can large language models automatically score proficiency of written essays?[J]. arXiv preprint, 2024, arXiv: 2403.06149.

[作者简介]



王兵 (1982-), 女, 博士, 东北石油大学计算机与信息技术学院副教授, 主要研究方向为智慧教育、量子图像处理、深度学习等。



单瑞雪 (2000-), 女, 东北石油大学计算机与信息技术学院硕士生, 主要研究方向为智慧教育、深度学习等。



邢海燕 (1971-), 女, 博士, 东北石油大学机械科学与工程学院教授、博士生导师, 主要研究方向为智慧教育、陆地及海洋钻井平台与油气管道检测机器人设计等。



李盼池 (1969-), 男, 博士, 东北石油大学计算机与信息技术学院教授、博士生导师, 主要研究方向为智慧教育、量子衍生计算、深度学习等。