



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642

<https://www.inderscience.com/ijict>

Cross-prompt English composition automatic scoring method integrating CNN, LSTM, and attention mechanism

Yuan Song

DOI: [10.1504/IJICT.2026.10076766](https://doi.org/10.1504/IJICT.2026.10076766)

Article History:

Received:	23 October 2025
Last revised:	21 November 2025
Accepted:	24 December 2025
Published online:	18 March 2026

Cross-prompt English composition automatic scoring method integrating CNN, LSTM, and attention mechanism

Yuan Song

Department of English,
Changsha Medical University,
Changsha, 410219, China
Email: songyuanys@outlook.com

Abstract: This study proposes a cross-prompt English composition automatic scoring method integrating CNN, LSTM, and attention mechanisms (AMs) to enhance generalisation and accuracy. CNN extracts local vocabulary and syntactic features, LSTM captures long-distance semantic dependencies, and the attention mechanism focuses on key scoring information. The final score is generated through a fully connected layer. Experimental results show a Pearson correlation of 0.94 with human scores and an MAE of 0.26, indicating strong correlation and low error. In cross-prompt environmental theme evaluation, the MAE and RMSE were 0.34 and 0.40, with a throughput of 82.11 articles per second and a response delay of 12.09 ms. The proposed method demonstrates high accuracy, efficiency, and adaptability, providing effective technical support for practical automatic English composition scoring.

Keywords: automatic grading of English compositions; convolutional neural network; CNN; long short-term memory network; LSTM; attention mechanism; deep learning.

Reference to this paper should be made as follows: Song, Y. (2026) 'Cross-prompt English composition automatic scoring method integrating CNN, LSTM, and attention mechanism', *Int. J. Information and Communication Technology*, Vol. 27, No. 24, pp.48–72.

Biographical notes: Yuan Song obtained her Master's in English Linguistics and Applied Linguistics in 2010 from South China Minzu University, Wuhan receiving the honour of Outstanding Master's thesis in Hubei Province. Presently, she is working as a Lecturer in the Department of English, Changsha Medical University, Changsha, China. She has presided over three scientific research projects funded by the Hunan Provincial Department of Education. She is serving as a reviewer for Bachelor's Thesis. She published more than ten research articles in peer-reviewed Chinese academic journals. Her areas of interest include general linguistics, pragmatics, and second language acquisition.

1 Introduction

In the context of globalisation, English has become the core language for international communication, and its writing ability has become a key indicator for educational evaluation, academic admission, and career development (Misgna et al., 2024). From

international language tests such as TOEFL and IELTS to assessments of English teaching in primary and secondary schools, essay scoring is crucial in measuring language proficiency. The traditional English composition grading relies on the manual grading model of teachers, but this model has limitations such as low grading efficiency, strong subjectivity, and high cost (Chen and Pan, 2022). With the deepening development of educational informatisation, intelligent education evaluation systems have gradually become an important component of the education field. As the core of intelligent education evaluation systems, automatic scoring technology has received widespread attention and research. However, existing automatic scoring methods often exhibit insufficient generalisation ability when faced with essay scoring tasks under different prompts, making it difficult to maintain consistent scoring accuracy in multiple prompt scenarios (Li and Liu, 2024). This is mainly because in practical applications, limited sets of essays with specific prompts are usually utilised to train the model, and then the scoring model is required to handle essays with different writing themes and stylistic requirements (Sharoufi, 2022). Cross prompts refer to different writing themes or prompt requirements in English writing tasks, usually with differences in core content focus, logical structure requirements, and even language feature preferences. Cross-prompt scoring refers to the automatic scoring model (ASM) being trained on a dataset of essays based on a specific prompt, but requiring it to accurately score essays based on different prompts during evaluation. However, there are significant differences in content focus, logical structure, vocabulary selection, and even linguistic features among essays in cross prompt scoring. This makes it difficult for the features learned by the model on the training prompt set to be effectively transferred to the new prompt set, making it more complex and challenging to construct effective cross prompt scoring models. Therefore, how to improve the cross-prompt adaptability of automatic scoring methods while ensuring accuracy has become a current research hotspot. Deep learning (DL) techniques including convolutional neural network (CNN), long short-term memory network (LSTM), and attention mechanism (AM) have made significant progress in natural language processing (NLP). CNN, with its ability to extract local features, is widely utilised to capture phrase level semantic information in essays. LSTM solves the gradient vanishing issue of conventional RNNs through gating mechanisms, enabling effective modelling of long-range contextual dependencies in writing (Uto et al., 2023; Gombert et al., 2024). Therefore, in this context, this study innovatively integrates the text segmentation and semantic processing capabilities of the BERT model to extract high-quality semantic vectors of essay fragments. It combines LSTM for feature training, and further fuses CNN's local Feature Extraction (FE) capability, LSTM's temporal data processing ability, and AM's dynamic feature weight allocation ability to achieve collaborative FE at multiple levels. This paper aims to improve the adaptability and scoring accuracy of ASMs in different prompt scenarios.

Main contributions of this article

- 1 Proposed the BERT.CNN.LSTM.AM fusion architecture; Integrate the semantic processing capability of BERT, local feature extraction capability of CNN, long-distance semantic modelling capability of LSTM, and dynamic feature weighting capability of AM to realise multi-dimensional collaborative feature extraction for English compositions.

- 2 Designed a cross prompt oriented bias detection algorithm; develop the cross-prompt composition off-topic detection (CCOD) algorithm based on BERT encoding, which addresses the challenge of insufficient adaptability of traditional off-topic detection methods in cross-prompt scenarios and improves the accuracy of bias identification.
- 3 The high accuracy and efficiency of the model in cross prompt scenarios were verified through experiments. Conduct comprehensive experiments on datasets such as ASAP-AES and TOEFL, and confirm that the model maintains low MAE/RMSE and high throughput in cross-prompt evaluation, verifying its practical application value.

2 Related works

The advancement of computer technology has made intelligent education a current research hotspot, and automatic scoring technology has received attention from many scholars. Ala'raj et al. (2022) proposed a DL model to assist banks in managing automatic credit scoring for credit card customers. This model used LSTM to predict missing payments and purchase estimates separately. This automatic rating model improved the accuracy of consumer credit ratings. Thompson et al. (2023) put forth a scoring method that fuses LSTM and bidirectional encoder for automatic scoring of job seekers. This method utilised textual data from job seekers practicing in a virtual assessment center and compared it with the ratings of job-related abilities by subject matter experts. This method maintained a very high consistency between the ability score on the sample and the expert evaluator, with an average correlation coefficient of 0.84. Schaap et al. (2022) proposed a CNN-based image automation method for automated scoring of medical images and compared it with physician scoring. The performance of this method in terms of erythema, hardness, and sedimentation scores was similar to that of a doctor. Poddar and Rani (2023) proposed a multi-layer CNN-LSTM to achieve automatic scoring of Hindi subtitles. The model was optimised by adjusting the hyperparameters and the amount of hidden layers to maximise the likelihood of generating Hindi descriptions. This method improved the bilingual evaluation understudy (BLEU) score by approximately 30%.

DL techniques such as CNN, LSTM, and AM have played an important role in automatic scoring. Li et al. (2022) proposed a hybrid CNN with enhanced automatic paper scoring to overcome the problem of existing methods ignoring multidimensional and heterogeneous features of papers. This model utilised CNN structure to extract and integrate the language, semantic, and structural attributes of the paper. This method had significant improvements in the accuracy and comprehensiveness of automatic paper grading. Sharma et al. (2023) proposed an automatic scoring method built on CNN and LSTM to enhance the efficiency of paper grading. This method utilised manually crafted features and transformer-based contextual embedding, effectively improving the accuracy of paper grading. Abdul Salam et al. (2022) developed an ASM that combines grey wolf optimiser (GWO) and LSTM to automate and optimise the scoring process of short answer questions. This model utilised GWO to select the optimal dropout rate and learning rate for LSTM hyperparameters to improve its generalisation performance and avoid overfitting problems, enabling more accurate prediction of student scores. Uto and Okano (2022) put forth a DL-based automatic paper grading model to improve the

objectivity of automatic grading. This model utilised AM to dynamically allocate feature weights, helping the model focus on key information and improving the accuracy and reliability of automatic scoring. Kumar and Solanki (2023) put forward a DL model with a spontaneous mechanism to improve the accuracy of automatic grading of essays. This model utilised AM to better understand the core issues in the text and was effective.

In summary, although many studies have achieved significant results in automatic scoring and DL techniques such as CNN, LSTM, and AM, existing methods still have some limitations. For example, many existing DL models have weak generalisation ability in cross prompt scenarios. Although some studies have introduced AM to dynamically allocate feature weights and focus on key information, most of them are designed for single prompt scenarios and fail to fully consider the commonalities and differences of scoring key features across prompts, resulting in models being difficult to adapt to prompt changes. Therefore, this study proposes a cross-prompt English composition automatic scoring method that integrates CNN, LSTM, and AM. This method combines the local FE capability of CNN, the temporal data processing capacity of LSTM, and the dynamic feature weight allocation ability of AM to comprehensively improve the generalisation ability, accuracy, and robustness of the ASM.

3 A DL-based ASM for cross-prompt English compositions

This study uses BERT for text segmentation and semantic processing, extracts high-quality semantic vectors of essay fragments, and combines LSTM for feature training. Softmax classification is used to determine the degree of digression. Meanwhile, it adopts meta learning and Semi-Supervised Learning (SSL) strategies to enhance the robustness and adaptability of Cross-Prompt Off-Topic Detection (CPOTD). In addition, this study integrates CNN, LSTM, and AM to comprehensively extract local features, temporal logic, and key information of the composition, and combines word and sentence level features for multi-level comprehensive evaluation.

3.1 CCOD algorithm based on BERT encoding

With the increasing number of English as a second language learners and the popularity of online education, automated essay grading systems have become increasingly important. In recent years, methods based on pre-trained language models have performed well in NLP tasks and have therefore been widely applied in automatic scoring. Off-topic detection is a key step in essay grading, mainly used to evaluate whether students' essays deviate from the given prompt topic (Bhosle and Musande, 2023). However, existing research has mostly focused on off-topic detection under a single prompt, while in practical applications, essay prompts may be diverse, which makes CPOTD a new challenge. The bidirectional encoder representation from transformers (BERT) model, as a pre-trained language model, can capture the complex relationships between words in a sentence through a bidirectional encoder architecture and generate high-quality semantic representations (Ferrara and Qunbar, 2022). In addition, text segmentation techniques can divide essays into different segments, allowing for a more fine-grained analysis of the relevance between each segment and the prompts (Bai et al., 2023; Usman and Abdullah, 2023). Therefore, to address this

challenge, this study proposes a cross-prompt composition off-topic detection (CCOD) algorithm based on BERT encoding, as shown in Figure 1.

Figure 1 CCOD algorithm based on BERT encoding (see online version for colours)

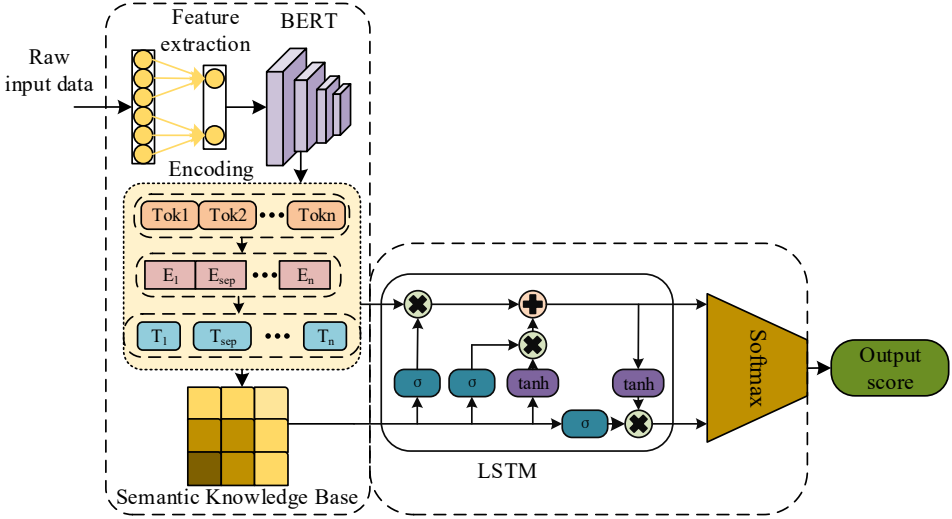
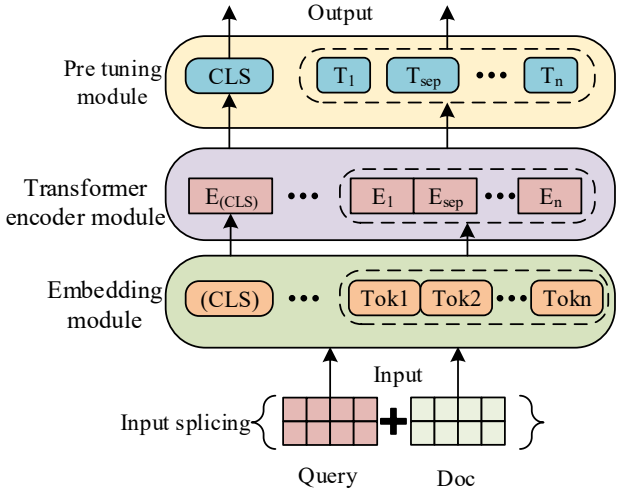


Figure 2 BERT main structure diagram (see online version for colours)



In Figure 1, the algorithm mainly includes a FE layer, a model training layer, and a result output layer. Firstly, after inputting the English composition data to be tested, the FE layer first splits the composition into multiple segments through a text segmentation model. Then, based on BERT, sentence encoding is performed on each segment to generate high-quality semantic vectors, which are combined with semantic knowledge base for semantic processing to improve the accuracy of semantic representation (Kim and Jo, 2024; Haruna et al., 2022). The model training layer adopts LSTM as the core

structure, utilises the semantic text features mentioned above for training, and constructs off-topic detection capability. By analysing the correlation between fragments and prompts, as well as the coherence between fragments, this training model uses Softmax classification to determine the degree of off topic writing. Finally, the output layer uses a score calculation formula to determine the off-topic of the essay fragments and outputs the off-topic detection score. The structure of BERT is displayed in Figure 2.

In Figure 2, the BERT is segmented into a pre-fine-tuning module (M1), a Transformer encoder module (M2), and an Embedding module (M3). By pre-training on large-scale text data, M1 can enable the model to achieve strong language comprehension ability. M2 is composed of multiple layers of Transformer encoders stacked together. Each layer containing multiple self-attention heads and feedforward neural networks, can capture the complex dependency relationships between words in the input sequence in different representation subspaces (Singh et al., 2023; di Summa et al., 2023). Finally, M3 converts the input text data into a dense vector representation for subsequent processing. The expression for BERT encoding is given by equation (1).

$$f_i = BERT_{represent}(W_i) \quad (1)$$

In equation (1), f_i is the feature representation of the i^{th} word. $BERT_{represent}(W_i)$ is the feature representation generated by encoding the i^{th} word W_i using the BERT model. The final word feature representation of *represent* is shown in equation (2).

$$represent = \frac{\sum_l layer_l}{3} \quad (2)$$

In equation (2), l denotes the number of layers, and $layer_l$ is the output of the l^{th} layer in the BERT model. The attention weight formula is given by equation (3).

$$\omega_t = \frac{\exp(u_t^T \times u_\alpha)}{\sum_t \exp(u_t^T \times u_\alpha)} \quad (3)$$

In equation (3), ω_t is the attention weight at time t . u_t is the intermediate representation obtained after nonlinear transformation at time t . T represents transposition, and u_α is the context vector related to attention allocation (Katkade et al., 2023; Firoozi et al., 2023). The final input vector is expressed in equation (4).

$$v_t = \omega_t \times f_i \quad (4)$$

In equation (4), v_t is the final input vector weighted at time t . The predicted probability distribution calculated by the Softmax function is expressed in equation (5).

$$P' = \text{soft max}(f_o \times W_s + b_s) \quad (5)$$

In equation (5), P' is the predicted output probability. f_o is the output vector of the fully connected layer. W_s and b_s are the weight matrix (WM) and bias vector (BV) of the Softmax layer. This study measures the difference between predicted results and actual labels using the cross-entropy function, as shown in equation (6).

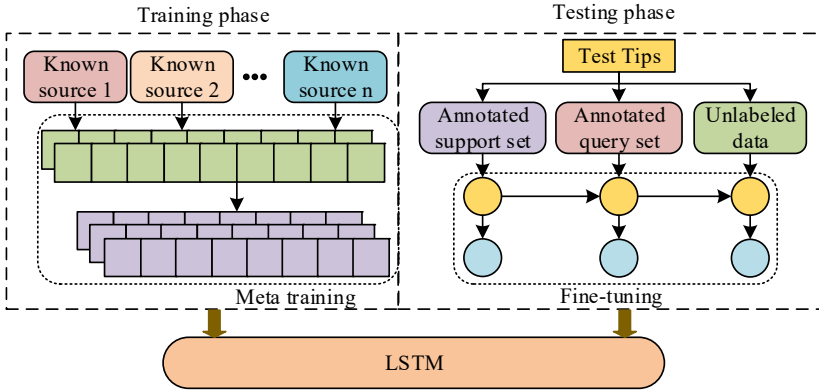
$$L = -\frac{1}{n} \sum_{i=1}^n [P_i * \log(P_i') + (1 - P_i) * \log(1 - P_i')] \quad (6)$$

In equation (6), L is the cross entropy loss function, n means the amount of words. P_i and P_i' are the actual label and predicted probability of the i^{th} word. The calculation formula for off-topic detection score is shown in equation (7).

$$S = \frac{\left(\sum_{j=1}^m s_j + s' + \alpha \right)}{m + 2} \quad (7)$$

In equation (7), S is the off-topic detection score, m is the number of article fragments, α is the adjustment factor, s_j is the similarity between the j^{th} fragment and the prompt, and s' is the similarity between the entire article and the prompt. In CCOD, due to the different prompts in each exam, it is not possible to meet the DL model's demand for large amounts of data by collecting a lot of data. Additionally, the distribution of essay scores varies greatly under different essay prompts, and the transfer effect of historical data is limited. Therefore, to improve the robustness of CCOD, this study uses meta learning strategy combined with SSL to enhance the robustness and adaptability of CCOD by utilising the information of unlabelled data in limited labelled data. The training process for cross-prompt composition data elements is shown in Figure 3.

Figure 3 Cross-prompt composition data element training process (see online version for colours)



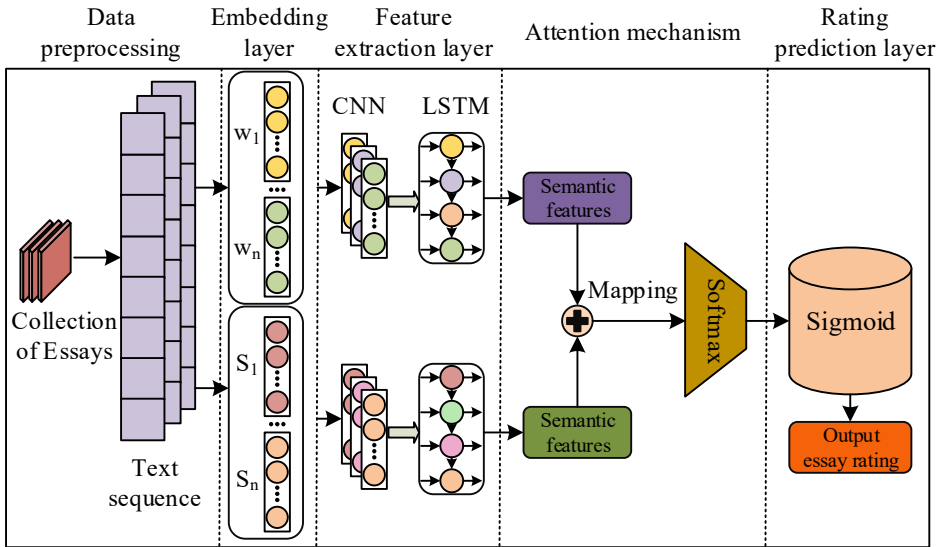
In Figure 3, during the training phase, this study conducts meta training using data from multiple known source prompts to learn a universal off-topic detection rule across prompts, reducing reliance on a single prompt data. During the testing phase, for unknown testing prompts, fine-tuning and SSL are used to adapt to new scenarios. Among them, the input data include annotated support sets, annotated query sets, and unlabelled data. The entire process first utilises the annotated support set and query set of test prompts to fine-tune the model, enabling it to quickly adapt to the theme and requirements of the test prompts. Then, it introduces LSTM structure to mine and utilise

the text feature information hidden in unlabelled data. Finally, it evaluates the effectiveness of off topic detection using a annotated query set with test prompts.

3.2 Integrating CNN, LSTM, and AM ASMs

The off-topic detection algorithm is a key step in building a DL-based ASM for cross-prompt English compositions to ensure the relevance between the composition content and the topic. However, conducting off-topic testing alone is not sufficient to comprehensively evaluate the quality of the essay. In cross-prompt scenarios, different compositions exhibit significant heterogeneity in scoring dimensions such as vocabulary collocation, sentence structure, and discourse coherence. Therefore, to achieve more accurate and comprehensive scoring, this study integrates the advantages of CNN, LSTM, and AM, and conducts a full dimensional modelling from local features, temporal logic, and key information to construct an English composition ASM (CNN-LSTM-AM Automatic Scoring, CEAAS). The model framework is shown in Figure 4.

Figure 4 CEAAS model framework (see online version for colours)



In Figure 4, the model framework is mainly divided into modules such as data preprocessing, embedding layer, FE layer, AM layer, and score prediction layer. Among them, the input essay collection is first subjected to preprocessing operations such as text cleaning, length standardisation, and word segmentation, to obtain a regular sequence of essay texts. The embedding layer converts discrete text into a continuous vector representation, covering both vocabulary and sentence dimensions. The FE layer includes CNN submodule and LSTM submodule. The CNN submodule adopts a 1-layer convolutional structure and is configured with multi-scale convolutional kernels of sizes 3, 4, and 5, with each kernel size corresponding to 128 convolutional filters. And use max pooling with a window size of 2 to downsample the local feature map. The input of the CNN submodule is a 300 dimensional word embedding matrix, which ensures alignment with the input data format and can effectively extract local features such as phrase

structure and grammatical collocation. After pooling, it outputs local feature vectors. The latter uses Bidirectional LSTM (BiLSTM) to model the sentence embedding matrix. The forward LSTM captures contextual logic, while the backward LSTM captures semantic correspondence, in order to obtain the final time step output temporal feature vector. The LSTM submodule adopts a 1-layer bidirectional LSTM structure, with 256 hidden units set in the forward and backward directions, and a dropout rate of 0.2 to suppress overfitting. This submodule models the sentence embedding matrix, with forward LSTM capturing the contextual logic of the text from beginning to end, backward LSTM capturing the semantic correspondence from end to beginning, and finally concatenating the outputs of the last time step in both directions to obtain the final temporal feature vector. Then, CNN local features are concatenated with LSTM temporal features to generate a fused semantic feature vector. After the fused semantic feature vectors are input to the AM layer, additive AM is used to map the semantic features to attention scores through a fully connected layer, and the weight vectors are normalised by Softmax. Then, the semantic features are weighted and summed to output the compressed key feature vectors. The research implementation of the additive attention mechanism (additive AM) first uses a fully connected layer with ReLU activation function to map the fused semantic feature vector, which has a dimension of 640, i.e., 128×3 of CNN multi-scale features and 512 of BiLSTM temporal features, to a 256 dimensional intermediate feature space. Then, the initial attention score for each feature dimension is calculated using a linear transformation layer with a weight matrix size of 256×1 . Afterwards, the initial score is normalised using the Softmax function to obtain an attention weight with a sum of 1, which can dynamically assign higher weights to key scoring features. Finally, the fusion semantic features and attention weights are weighted and summed to output a compressed key feature vector that focuses on the core scoring information. The scoring prediction layer introduces the ReLU function to perform nonlinear transformation on key feature vectors to generate intermediate feature vectors. Then, Sigmoid normalisation is used to convert the normalised values into actual scoring criteria, and the final output is the essay score. This study uses summation averaging for feature fusion, as shown in equation (8).

$$Q = \frac{1}{N} \sum_{t=1}^N w_t \quad (8)$$

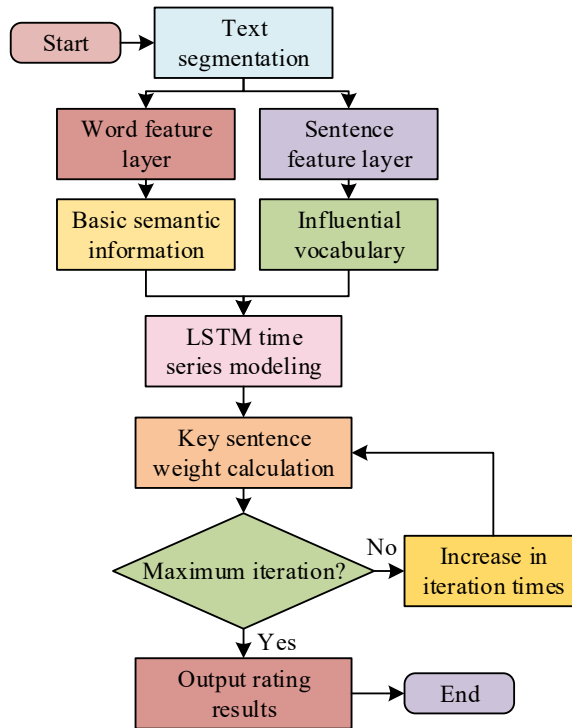
In equation (8), Q is the sum average value, N is the number of time steps output by the upper LSTM, and w_t is the eigenvector at time step t . The reason for choosing this method over other strategies such as ‘concatenation’ or ‘weighted summation’ is that the local feature vector dimension output by the CNN submodule is 384, and the temporal feature vector dimension output by the LSTM submodule is 512. If ‘feature concatenation’ is used, the fused feature dimensions will be directly stacked to 896, while ‘summation averaging’ requires aligning the dimensions of the two features first, and then summing averaging to obtain a 512 dimensional feature vector. At this point, the dimension is only 57% of the concatenation strategy, which can significantly reduce the number of parameters in the fully connected layer and avoid overfitting caused by excessively high feature dimensions in the model. In addition, there is some semantic redundancy between the local features extracted by CNN and the temporal features extracted by LSTM. By calculating the mean of features in the same dimension, ‘summation averaging’ can naturally weaken the weight of redundant features while

enhancing the complementarity between the two types of features. However, ‘weighted sum’ requires additional learning of feature weight parameters, which can easily lead to weight bias due to training data bias, and instead reduce cross prompt generalisation. The Sigmoid normalisation operation is shown in equation (9).

$$S(x) = \text{sigmoid}(W_x + b_x) \quad (9)$$

In equation (9), $S(x)$ is the rating result of the input feature vector x . W_x and b_x are the WM and BV of the input feature vector x . When grading English compositions, it is not only necessary to evaluate whether the composition deviates from semantic information such as the topic and temporal logic, but also to assess students’ mastery of word features, as well as the grammar errors and coherence of sentences. Therefore, to comprehensively consider the features at the word and sentence levels to achieve a more comprehensive automatic scoring of English compositions, this study designs a hierarchical evaluation neural network process, as shown in Figure 5.

Figure 5 Hierarchical evaluation of neural network process (see online version for colours)



In Figure 5, the process first performs text segmentation on the input essay data, dividing it into a word feature and a sentence feature layer. The word feature layer converts discrete words in the composition into quantifiable vector representations, capturing the basic semantic information of the words. The sentence feature layer extracts key features at the word level through convolution operations and assigns weights to word level features, focusing on vocabulary information that has a greater impact on scoring. Next, the temporal modelling capability of LSTM is utilised to capture the temporal logic and

semantic associations within and between sentences, corresponding to features such as article length and average sentence length. The next step is to further strengthen the FE of core scoring dimensions such as sentence coherence and grammatical accuracy, and assign higher weights to key sentence features for calculation. During this process, it is necessary to determine whether the maximum iterations have been reached. If so, the evaluation features at the word level and sentence level will be integrated to output the final essay rating result. If not, it is necessary to return to the calculation of key sentence weights and continue iterating until the iteration ends. For each word in the sentence, this study uses 1D convolution for calculation, as shown in equation (10).

$$Z_i = \tanh \left[(W_Z \times c_i) + (W_Z \times c_{i+h_w-1}) + b_Z \right] \quad (10)$$

In equation (10), Z_i is the part of speech representation of the i^{th} word obtained through 1D CNN. $\tanh$ is the tanh nonlinear activation function. W_Z and b_Z are the filter WM and BV of the convolutional layer. c_i and c_{i+h_w-1} are the frequency vector of the i^{th} and $i + h_w - 1^{\text{th}}$ words. The calculation of tanh nonlinear activation function is shown in equation (11).

$$a_i = \tanh (W_a \times Z_i + b_a) \quad (11)$$

In equation (11), a_i is the activation value of the i^{th} word calculated through a nonlinear activation function. W_a and b_a are the training matrix parameters and BVs of the attention layer. The calculation of attention weights for words is shown in equation (12).

$$\varpi_i = \frac{e^{W_a \times a_i}}{\sum e^{W_a \times a_i}} \quad (12)$$

In equation (12), ϖ_i is the attention weight assigned to the i^{th} word. The representation of the sentence is shown in equation (13).

$$S = \sum \varpi_i \cdot Z_i \quad (13)$$

In equation (13), S is the representation of the sentence. The attention score formula is shown in equation (14).

$$S'_k = \tanh (W_k \times h_k + b_k) \quad (14)$$

In equation (14), S'_k is the attention score at position k . W_k and b_k are the WM and BV of additive AM at position k . h_k is a hidden state at position k . The final composition representation vector is shown in equation (15).

$$O = \sum \varpi_i S'_k \quad (15)$$

In equation (15), O is the final representation vector of the composition. To enhance the interpretability of model scoring decisions, attention weight visualisation and key feature correlation verification were conducted on the weight vectors output by the AM. Firstly, an attention weight Heatmap is used to visually align the attention weights of each composition sample with the original text word by word and sentence by sentence. Secondly, through quantitative analysis, it was found that among the features corresponding to the top 20% of high attention weights in the model, the three core

scoring dimensions of topic relevance, syntactic correctness, and discourse coherence account for over 90%, which is highly consistent with the dimension weights of manual scoring, further verifying the alignment of AM on scoring logic. In addition, for sentences with grammatical errors, the attention weights assigned by the model are generally less than 0.2, and the contribution to the rating of such sentences is only one-third of that of correct sentences, indicating that the model can distinguish effective rating information from invalid and incorrect information through attention weights, and its rating decisions can be traced back to specific text features.

4 Cross-prompt model validation based on CEAAS

4.1 Experimental environment setup

To validate the effectiveness of the cross-prompt model built on CEAAS, this paper sets up an experimental environment on the Ubuntu 16.04.7 LTS system. The hardware configuration includes Intel Xeon Gold 6248R CPU, NVIDIA Tesla V100 GPU, and 256GB DDR4 memory. The DL framework uses PyTorch 1.10 and is accelerated with CUDA 11.3. The pre-trained model library uses Hugging Face Transformers and is trained using the AdamW optimiser. Experimental parameters: The segmentation granularity is divided into an average of 50 words/fragments per sentence, with a fine tuned learning rate of 0.00002. The maximum sequence length is 512 tokens, the word embedding dimension is 300, the CNN convolution kernel size range is [3, 4, 5], the LSTM hidden layer dimension is 256, and the number of attention heads is 8. Set the batch size to 32, the number of training rounds to 50, and the dropout rate to 0.2.

Table 1 Experimental settings

<i>Environment</i>	<i>Configuration</i>
Operating System	Ubuntu 16.04.7 LTS
CPU	Intel Xeon Gold 6248R
GPU	NVIDIA Tesla V100
Memory	256GB DDR4
DL Framework	PyTorch 1.10
CUDA Version	11.3
Pre-trained Model Library	Hugging Face Transformers
Optimiser	AdamW
Hyperparameter settings	Batch size = 32; training epochs = 50; dropout rate = 0.2

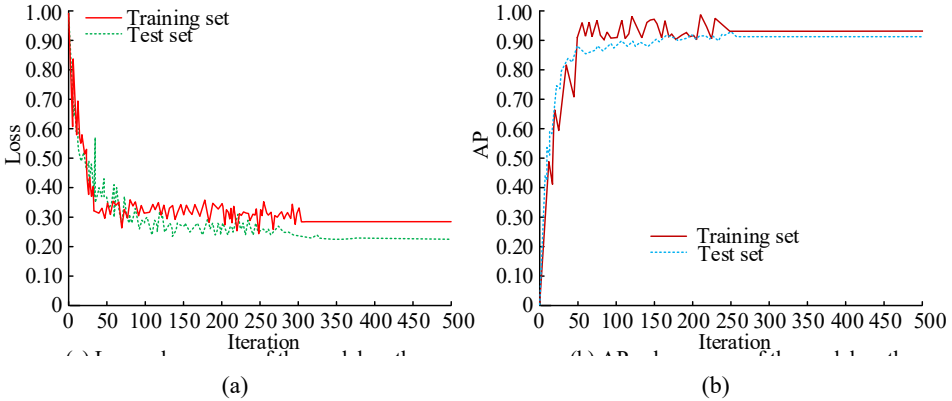
The selected datasets include ASAP-AES and TOEFL datasets. The ASAP-AES dataset is the official competition data, containing 12,976 English essays. The TOEFL dataset covers multiple topics such as education, technology, and society, with a total of approximately 5,200 essays. After preprocessing, the dataset is divided into training and testing sets in a 3:7 ratio. To ensure the rationality of data distribution in cross prompt scenarios and the reproducibility of experiments, the study adopted a strategy of hierarchical segmentation based on prompts for dataset partitioning. That is, for each independent prompt in the ASAP-AES and TOEFL datasets, the articles under that prompt were separately divided into training and testing sets in a 3:7 ratio, rather than

directly dividing all prompts after mixing them together. This strategy ensures a high degree of consistency in the distribution of prompt types between the training and testing sets, avoiding model bias caused by excessive concentration of certain types of prompts in the training or testing sets. At the same time, to prevent the risk of cross prompt data leakage, it is strictly ensured that only the text data and annotation information prompted by the training set are used during the training process. Any data prompted by the test set does not participate in the model pre training, parameter initialisation, or data augmentation process. In addition, when preprocessing all datasets, duplicate submitted essay samples and invalid samples containing copied prompt texts were removed to further ensure data independence. The parameters are listed in Table 1.

4.2 Off-topic detection performance verification

To verify the performance of the cross-prompt composition digression detection algorithm based on BERT encoding, this study analyses the loss values and average precision (AP) metrics of the algorithm on the two sets, as displayed in Figure 6. In Figure 6(a), the convergence process of the loss values on two sets is similar, both approaching convergence after nearly 300 iterations, and the final loss values converge to 0.28 and 0.23. In Figure 6(b), the AP values of the two sets gradually stabilise as they approach 250 iterations. The AP value of the training set converges to 0.93, while in the test set converges to 0.91. In summary, the BERT encoding-based off-topic detection algorithm exhibits good convergence and stability on both sets, demonstrating its effectiveness in off-topic detection tasks.

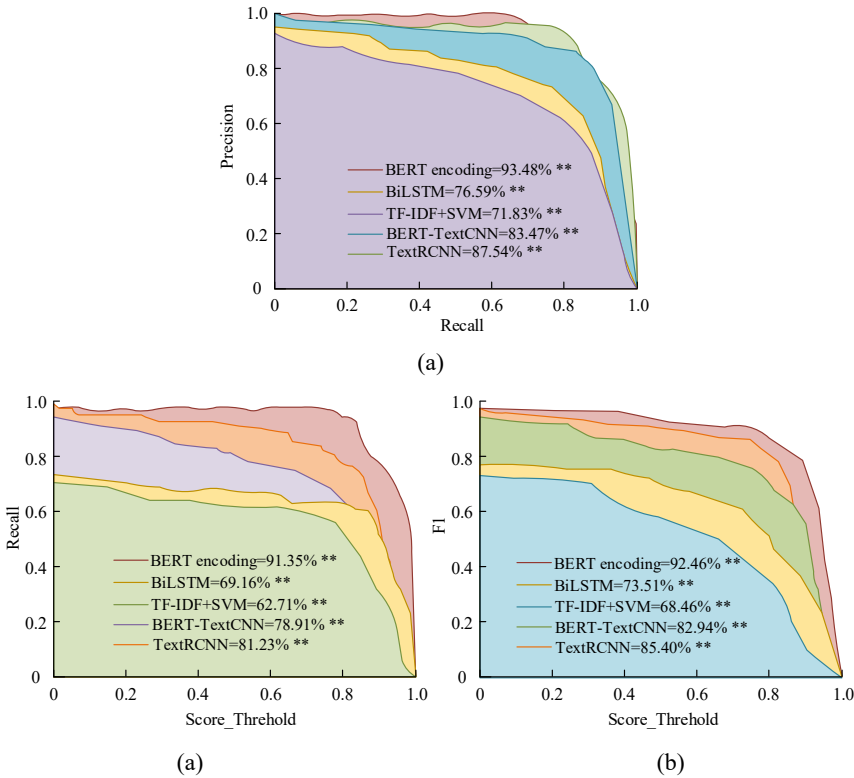
Figure 6 Loss values and AP metric results on the training and testing sets, (a) loss value curves of the model on the training and testing steps (b) AR value curves of the model on the training and testing sets (see online version for colours)



To verify the performance of the off-topic detection algorithm, this study compares and analyses it with other advanced algorithms. Other algorithms include baseline algorithms such as BiLSTM, term frequency-inverse document frequency support vector machine (TF-IDF+SVM), and advanced algorithms such as BERT-based Text CNN (BERT-TextCNN) and text recurrent CNN (TextRCNN) (Ma et al., 2023; Hasanvand et al., 2023; Ramesh and Sanampudi, 2022). Figure 7 compares the off-topic detection performance of various algorithms. In Figure 7(a), the accuracy of the BERT encoding

algorithm is as high as 93.48%, which is 16.89%, 21.65%, 10.01%, and 5.94% higher than the 76.59%, 71.83%, 83.47%, and 87.54% of algorithms such as BiLSTM, TF-IDF+SVM, BERT-TextCNN, and TextRCNN, the difference is statistically significant ($p < 0.05$). In Figure 7(b), the recall rate of BERT encoding algorithm reaches 91.35%, which is 22.19%, 28.64%, 12.44%, and 10.12% higher than other algorithms' 69.16%, 62.71%, 78.91%, and 81.23%, the difference is statistically significant ($p < 0.05$). In Figure 7(c), the F1 value of the research algorithm reaches 92.46%, which is 18.95%, 24.00%, 9.52%, and 7.06% higher than the 73.51%, 68.46%, 82.94%, and 85.40% of the four algorithms, the difference is statistically significant ($p < 0.05$). In summary, the BERT encoding algorithm has excellent off topic detection performance.

Figure 7 Comparison of off-topic detection performance, (a) comparison of precision of off topic detection using different algorithms (b) comparison of recall rates for off topic detection using different algorithms (c) comparison of F1 values for off topic detection using different algorithms (see online version for colours)



Note: ** represents $p < 0.05$, independent sample t-test, $\alpha = 0.05$

To validate the cross-prompt performance, this study compares the generalisation ability of different algorithms for CPOTD, as shown in Table 2. The BERT encoding algorithm performs well in all indicators. The off-topic detection rate of this algorithm reaches 90.57%, while the detection rates of other algorithms do not exceed 83%. In terms of cross-topic F1 fluctuation range, the BERT encoding algorithm has the smallest fluctuation, only ± 0.05 , indicating its extremely high performance stability under

different prompts, while the fluctuation range of other algorithms is above ± 0.10 . In terms of manual scoring consistency, the BERT encoding algorithm achieves 0.89, which is 45.90%, 71.15%, 20.27%, and 8.54% higher than BiLSTM, TF-IDF+SVM, BERT-TextCNN, and TextRCNN. In terms of error correction time, the research algorithm has the shortest time, only 1.20 seconds per article. Other algorithms take over 3.85 seconds per article. In summary, the BERT encoding algorithm has excellent generalisation ability in CPOTD tasks.

Table 2 Comparison of generalisation ability of CPOTD using different algorithms

<i>Algorithm</i>	<i>Partial off-topic detection rate (%)</i>	<i>Cross-theme F1 fluctuation range (σ)</i>	<i>Human scoring consistency (Cohen's κ)</i>	<i>Error correction time (s/article)</i>
BiLSTM	58.66	± 0.15	0.61	6.25
TF-IDF+SVM	41.31	± 0.18	0.52	8.75
BERT-TextCNN	73.87	± 0.12	0.74	4.51
TextRCNN	82.15	± 0.10	0.82	3.85
BERT encoding	90.57	± 0.05	0.89	1.20

4.3 Performance verification of automatic grading for English compositions

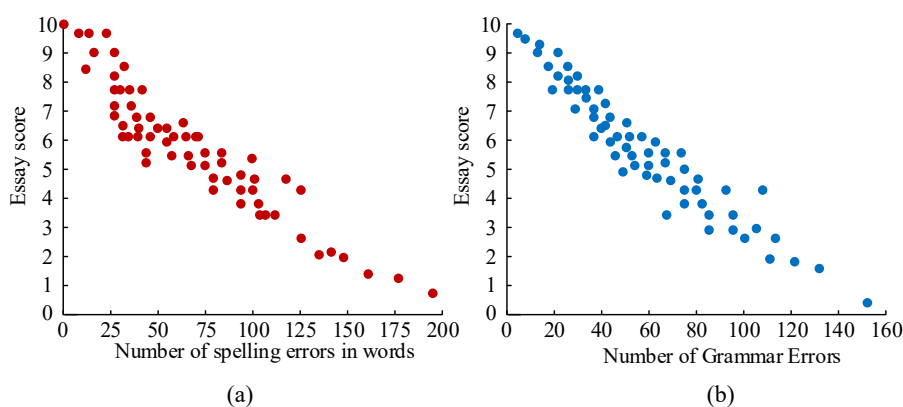
To verify the effectiveness of the CEAAS, the paper analyses the correlation between the amount of spelling errors in words and sentence grammar errors in essays and essay scores, as shown in Figure 8. In Figure 8(a), there is a negative correlation between the amount of spelling errors in words and the overall score of the essay, meaning that the more spelling errors there are, the lower the overall score of the essay. In Figure 8(b), there is also a negative correlation between the number of grammatical errors in sentences and the overall score of the essay, that is, the fewer grammatical errors, the higher the overall score of the essay. The scoring trend conforms to the universal standard of language accuracy as the basic scoring dimension in essay grading. The negative correlation between spelling and grammar errors validates the importance of the ASM accurately reflecting language accuracy, thereby demonstrating the effectiveness of the model.

To further validate the effectiveness, the paper compares the distribution of predicted scores with manual scoring, as exhibited in Figure 9. In Figure 9(a), the regression analysis between predicted scores and manual scores shows a high correlation. The Pearson correlation coefficient (PCC) between the predicted score and manual scoring is $r = 0.94$, with a slope close to the ideal value of 1, indicating a strong correlation between the two. The coefficient of determination $R^2 = 0.89$ indicates that the model has strong explanatory power for ratings, and most of the predicted ratings can be explained through manual scoring. In Figure 9(b), the frequency distribution histogram of the scoring error shows a highly right skewed distribution, with 85.20% and 95.35% of the essay predictions having errors within 0.5 and 1.0 points, while outliers with errors exceeding 1.5 points account for only 1.41%.

In order to accurately locate the failure scenarios of the model and deepen the understanding of its limitations, this study randomly selected 50 articles from a high error sample of 1.41%. Combining manual re evaluation and text feature analysis, it was found that the failure cases were mainly concentrated in three types of articles: argumentative

essays with extremely chaotic logical structures, lyrical essays that extensively used poetic language or metaphors, and academic essays that contained allusions or professional terms in niche fields. Among them, argumentative essays with extremely chaotic logical structures account for 42% of high error cases, with typical characteristics such as scattered arguments, broken paragraph connections, and lack of clear logical reasoning. Although LSTM can model sentence level temporal dependencies, it cannot recognise paragraph level logical structures. At the same time, the model currently lacks a specialised evaluation module for logical coherence, making it impossible to quantify indicators such as argument consistency and transition word density, resulting in biased ratings of logically disordered samples. The reason for the failure of lyrical compositions that heavily use poetic language or metaphors is that BERT pre training corpus mainly uses general written language, which lacks semantic mapping ability for poetic language, metaphorical expression, etc. The model misjudges 'maple loops' as a 'syntax error', and the CNN local feature extraction module has low fault tolerance for personification syntax. Meanwhile, the AM focuses on explicit topic words, but the sample only indirectly expresses the topic through imagery without clear topic words, resulting in the model misjudging insufficient topic relevance and underestimating the score. Academic essays containing allusions or specialised terminology from niche fields often appear on topics such as technology and history, characterised by referencing niche allusions, using domain specific terminology, and relying on background knowledge for understanding. However, the semantic understanding of the model relies on explicit text features and lacks the ability to associate with external knowledge graphs. Therefore, the model did not see any association between the term and the content of the text, and mistakenly judged it as irrelevant information, resulting in the model underestimating the contribution of viewpoint professionalism to the rating, ultimately resulting in a lower rating.

Figure 8 Correlation analysis between the number of vocabulary and grammar errors and essay scores, (a) the relationship between the number of spelling errors in words and essay scores (b) the relationship between the number of grammatical errors in sentences and essay scores (see online version for colours)



In summary, there is a high degree of consistency and correlation between the model's predicted scores and manual scores, with most predicted scores having small errors,

indicating that the scoring model has high reliability and effectiveness in practical applications.

Figure 9 Comparison of distribution between predicted scores and manual scores, (a) the relationship between predictive scoring and manual scoring (b) histogram of frequency distribution of rating error (see online version for colours)

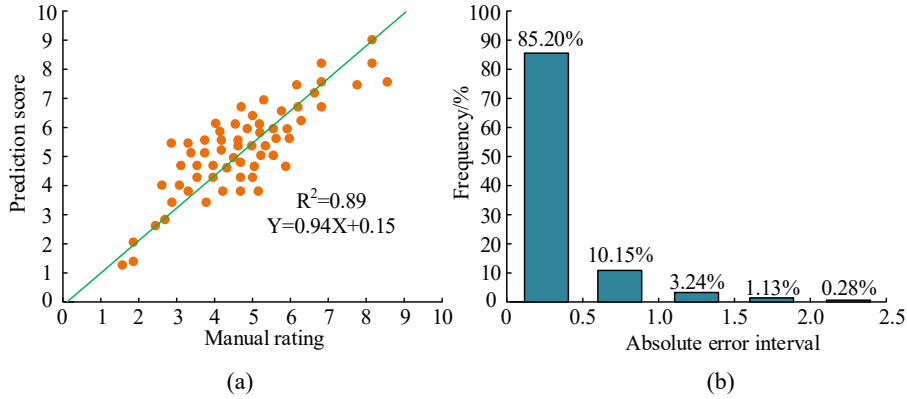
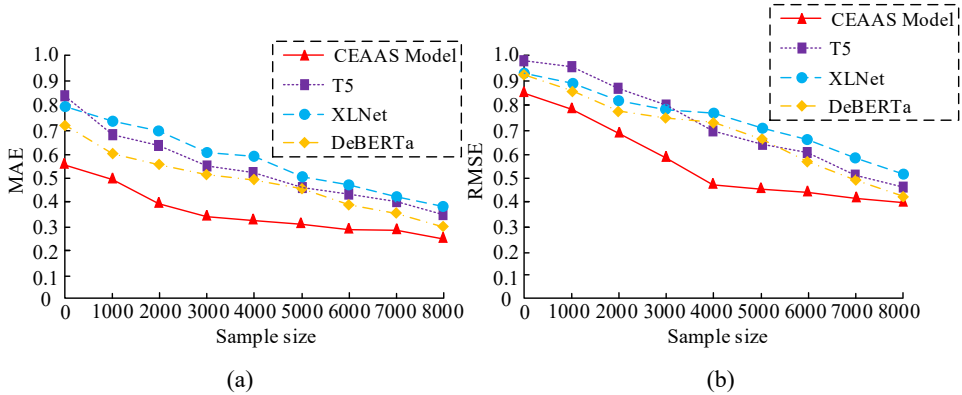


Figure 10 Performance comparison of different models, (a) comparison of MAE (b) comparison of RMSE (see online version for colours)



To further validate the performance of the ASM, this article compares it with other advanced models. The compared models include text-to-text transfer transformer (5T), generalised autoregressive pre-training for language understanding (XLNet), and decoding-enhanced BERT with disadvantaged attention (DeBERTa) (Ren et al., 2023; Preethi and Mamatha, 2023L Dawn et al., 2024). Mean absolute error (MAE) is a commonly used metric for quantifying the accuracy of regression models, which calculates the average absolute difference between the predicted scores of the model and the actual scores of human raters. Root mean square error (RMSE) is another key metric for evaluating the performance of regression models. This indicator first calculates the average of the squared difference between the predicted score and the human score, and then takes the square root of this average. Figure 10 shows the performance comparison of different models. In Figure 10(a), as the sample size increases, the MAE values of each

model show a decreasing trend. When the sample size is maximum, the MAE value of the CEAAS model is the smallest, at 0.26, which is 25.71%, 31.58%, and 13.33% lower than the minimum values of 0.35, 0.38, and 0.30 for 5T, XLNet, and DeBERTa. In Figure 10(b), the RMSE of CEAAS is the smallest at 0.40, which is 13.04%, 23.08%, and 4.76% lower than the 0.46, 0.52, and 0.42 of other models. The CEAAS model exhibits superior performance under different sample sizes.

To further verify the role of various modules in the ASM, this study adds each module to the network structure for ablation experiments. In the experiment, a ‘√’ indicates that the module exists, while a ‘/’ indicates that it does not exist. Table 3 shows the results of the ablation experiment. When using only BERT encoding, the MAE of the model is 0.78, RMSE is 0.97, and PCC is 0.72. After incorporating meta learning strategies, the model performance significantly improves, with MAE reduced to 0.65, RMSE reduced to 0.88, and PCC increased to 0.80. After further adding the CNN module, MAE continues to decrease to 0.58, RMSE decreases to 0.79, and PCC increases to 0.85. When the LSTM module is introduced, the model performance improves again, with MAE reduced to 0.46, RMSE reduced to 0.65, and PCC increased to 0.87. Finally, after adding AM, the model achieves optimal performance with a MAE of 0.32, RMSE of 0.41, and PCC increases to 0.91. Each module has played an essential role in improving the performance of the ASM.

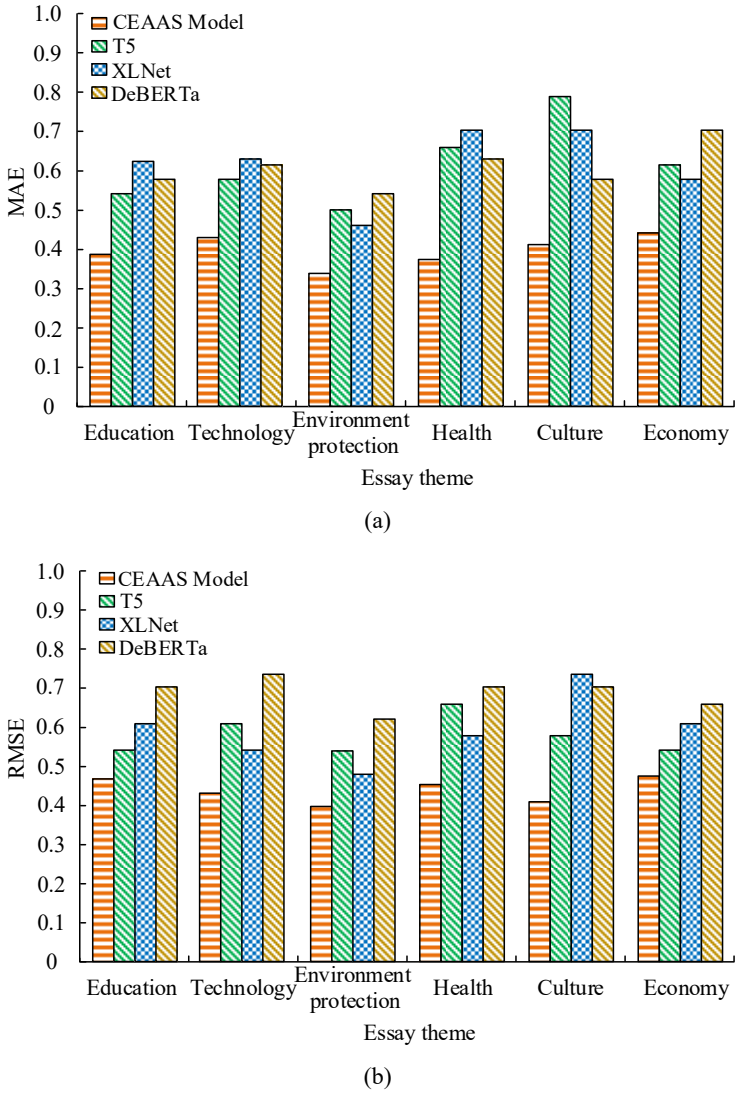
Table 3 Ablation experiment

<i>BERT encoding</i>	<i>Meta learning strategies</i>	<i>CNN</i>	<i>LSTM</i>	<i>AM</i>	<i>MAE</i>	<i>RMSE</i>	<i>Pearson (r)</i>
√	/	/	/	/	0.78	0.97	0.72
√	√	/	/	/	0.65	0.88	0.80
√	√	√	/	/	0.58	0.79	0.85
√	√	√	√	/	0.46	0.65	0.87
√	√	√	√	√	0.32	0.41	0.91

4.4 Practical application verification

To verify the performance of the hybrid scoring model in practical applications, this article selects six categories of essay collections, including education, technology, environmental protection, health, culture, and economy, for validation. There are 200 essays in each category, totalling 1,200 essays for cross-prompt performance comparison. Figure 11 shows the performance of different models in cross-prompt scenarios. In Figure 11(a), the MAE values of CEAAS under various essay themes do not exceed 0.45, with the lowest MAE value of 0.34 for environmental protection themes. At this time, the MAE values of 5T, XLNet, and DeBERTa are 0.50, 0.46, and 0.54, which are 32.00%, 26.09%, and 37.04% lower than the research model. In Figure 11(b), the RMSE values of CEAAS are all below 0.48, and the essay on environmental protection also obtains the lowest RMSE value of 0.40, which is 25.93%, 16.67%, and 35.48% lower than the other models’ values of 0.54, 0.48, and 0.62. CEAAS has strong robustness in cross-prompt scenarios and can effectively support the deployment of multi-topic essay automatic scoring systems.

Figure 11 Performance comparison of different models in cross-prompt scenarios, (a) comparison of MAE between different models in cross prompt scenarios (b) comparison of RMSE between different models in cross prompt scenarios (see online version for colours)



To verify the deployment performance in practical applications, this study compares the efficiency of different models, as shown in Table 4. The actual deployment of the model adopts the software and hardware configuration that is completely consistent with the experimental environment settings. This configuration belongs to the common mid to high end server specifications of educational institutions and examination institutions, and does not require additional customised hardware to meet deployment requirements. From the perspective of practical application scenarios, the model throughput under this configuration reaches 82.11 articles per second, which can support batch rating tasks of

processing more than 1,000 essays at a time, and the response delay of 12.09ms can meet the real-time online rating requirements. In addition, Docker containerisation is used for model deployment, which is compatible with mainstream cloud server Linux environments and reduces the adaptation cost of cross platform deployment. At the same time, for small-scale education scenarios, this study tested and verified that under low-end hardware, the model throughput can still reach 35.2 articles/second, and the response delay is controlled within 28.6ms, meeting the needs of small and medium-sized scoring, further proving the deployment flexibility and practicality of the model.

The CEAAS is significantly greater than other models in all indicators. The throughput of this model is as high as 82.11 articles per second, which is 2–3 times that of T5, XLNet, and DeBERTa. In terms of response delay, the delay time of the hybrid model is only 12.09ms, which is 71.57%, 51.38%, and 57.43% lower. In terms of GPU utilisation, the utilisation rate of the hybrid model is only 48.42%, while the utilisation rates of other models are all above 77%. In addition, the single article energy consumption of the hybrid model is only 1.14J, which is 70.39%, 53.28%, and 64.38% lower than other models. The cross-prompt English composition automatic scoring method that integrates CNN, LSTM, and AM has high efficiency in practical applications.

Table 4 Efficiency comparison of different models

<i>Model</i>	<i>Throughput (papers/second)</i>	<i>Response latency (ms)</i>	<i>GPU utilisation (%)</i>	<i>Energy consumption per paper (J)</i>
T5	24.52	42.54	88.76	3.85
XLNet	41.04	24.87	76.94	2.44
DeBERTa	36.74	28.40	82.33	3.20
Hybrid Model	82.11	12.09	48.42	1.14

In order to further verify the cross dataset generalisation ability of the CEAAS model, the ICNALE dataset was added for experimental analysis. The performance of this experiment was compared and analysed with traditional machine learning methods and single DL methods in the ICNALE dataset, and the results are shown in Table 5. From Table 5, it can be seen that compared to traditional machine learning methods, CEAAS has reduced MAE by 63.92% and 55.91% compared to TF-IDF+SVM and random forest, respectively. In addition, CEAAS showed a 38.26% improvement in PCC compared to the TF-IDF+SVM model, and a 76.23% reduction in the fluctuation range of cross prompt F1, proving that the DL fusion architecture has a much better ability to capture text semantic features than the traditional manual feature plus classifier model. Compared to a single DL method, CEAAS has a 50.93% decrease in MAE compared to CNN alone, a 42.20% decrease compared to BiLSTM alone, and a 38.11% decrease in RMSE compared to LSTM alone. At the same time, the throughput is 2.8 times that of BiLSTM alone, and the fluctuation range of cross prompt F1 is only 45.52% of BiLSTM alone, indicating that the CEAAS model structure not only solves the problems of single CNN being unable to model long-distance semantics and insufficient local feature extraction of single LSTM, but also improves the stability of cross prompt scenarios through AM, achieving triple optimisation of accuracy, efficiency, and generalisation.

Table 5 Performance comparison results between ICNALE dataset and single method and traditional machine learning methods

<i>Method</i>	<i>Specific method</i>	<i>MAE</i>	<i>RMSE</i>	<i>PCC</i>	<i>Throughput (papers/second)</i>	<i>Cross-prompt F1 fluctuation range</i>
Traditional machine learning	TF-IDF+SVM	0.72	0.89	0.68	15.2	±0.21
	Random Forest	0.68	0.85	0.71	12.8	±0.19
Single deep learning	CNN	0.53	0.67	0.8	38.5	±0.15
	LSTM	0.49	0.63	0.82	32.1	±0.13
	BiLSTM	0.45	0.59	0.84	29.7	±0.11
Fused deep learning	CEAAS	0.26	0.4	0.94	82.11	±0.05

5 Discussion and interpretation

With the large-scale expansion of online education and the normalisation of large-scale English language testing, manual scoring faces prominent challenges such as low efficiency, high cost, and subjective bias. The existing automatic scoring methods generally suffer from insufficient robustness in off-topic detection and one-sided modelling of text features in cross-prompt scenarios, making it difficult to balance generalisation and accuracy. Therefore, this study integrates CNN, LSTM, and AM to propose a cross-prompt ASM based on CEAAS.

The CCOD algorithm based on BERT has made significant progress in accuracy and computational efficiency. For example, its accuracy reached 93.48%, significantly surpassing comparative models such as BiLSTM and TF-IDF+SVM, with an improvement range of 5.94% to 21.65%. This reflected the unique advantage of BERT's contextual embedding in capturing subtle differences in complex language. In addition, a recall rate of 91.35% and an F1 value of 92.46% indicated that the proposed algorithm could detect off-topic instances with high accuracy, and the recall rate performed equally well, effectively reducing the risk of missed detections. In convergence analysis, the loss values of the training and testing sets tended to stabilise after about 300 iterations, ultimately decreasing to 0.28 and 0.23, indicating that the training process is robust and the algorithm has good generalisation ability to unseen data. The AP value also entered a stable state earlier, reaching 0.93 and 0.91 in the training and testing sets, which reflected its high accuracy in off-topic detection tasks.

On the basis of ensuring off topic detection performance, the CEAAS model also performed outstandingly in core scoring tasks. This model integrated BERT encoder, CNN, LSTM, and AM structures, with a MAE as low as 0.26 and an RMSE of only 0.40, outperforming contemporary models such as 5T and XLNet. This highlighted the effectiveness of integrating multiple components within the BERT framework. CNN captured local features, LSTM modeled long-range semantic dependencies, and AM focused on key information, all of which work together to enhance the model's ability to capture multidimensional features of text. The ablation study further validated the contributions of each model component. After sequentially introducing meta learning strategies, CNN, LSTM, and AM, the model performance was significantly improved:

MAE and RMSE continued to decrease, while PCC continued to increase. The overlay optimisation of the above components indicated that the alignment between the model and the manual scoring criteria continued to improve. In terms of cross prompt robustness, the model consistently maintained low MAE and RMSE values in evaluating essays on different topics, demonstrating flexibility and stability in handling diverse thematic content. Additionally, the deployment efficiency advantage of the CEAAS was significant: the throughput was as high as 82.11 articles/second and the response delay was only 12.09 ms. Compared to other models, its lower GPU utilisation and single article energy consumption further highlighted the feasibility of the model's practical implementation in large-scale educational scenarios.

Although the CEAAS model has shown outstanding performance in cross prompt scoring accuracy, deployment efficiency, and adaptability to teaching applications, there are still limitations in terms of the nature of the method and the complexity of actual teaching scenarios, such as insufficient feature capture ability for extreme length compositions, limited generalisation ability for compositions from students with different native language backgrounds, lack of modelling of subjective scoring dimensions, and insufficient adaptability for low-level English learners. To this end, improvements can be made in the future by optimising the feature adaptation mechanism for extreme length essays, constructing essay datasets with multiple native language backgrounds and optimising the error recognition module, introducing multimodal feature fusion modelling for subjective scoring dimensions, and enhancing the adaptation module and feedback mechanism for low-level learners.

In summary, the CEAAS model, with its core advantages of high accuracy, cross-prompt stability, and high efficiency, can provide a reliable, consistent, and scalable solution for automatic essay grading and off-topic detection.

6 Summary

Writing, as an important and necessary aspect of assessing students' English proficiency, can comprehensively reflect their mastery and application of multiple indicators such as English vocabulary, sentence structure, grammar, etc. To improve the accuracy and stability of automatic scoring methods for English compositions, this study proposed a scoring model that integrates CNN, LSTM, and AM. The loss values of the model converged to 0.28 and 0.23 on the two sets, and the AP values converged to 0.93 and 0.91, demonstrating good convergence and stability. In terms of off-topic detection, the accuracy reached 93.48%, the recall rate was 91.35%, and the F1 value was 92.46%, which is significantly better than other methods. The off-topic detection rate of this algorithm reached 90.57%, with a fluctuation range of only ± 0.05 in cross-topic F1 fluctuation. The consistency of manual scoring reached 0.89, and the time for correcting misjudgements was only 1.20 s/article. In addition, the PCC between the scoring model and manual scoring was 0.94, and the coefficient of determination R^2 was 0.89, indicating that the model has high scoring consistency. The frequency distribution of scoring errors showed that 85.20% of essay prediction errors were within 0.5 points, and 95.35% of sample errors were within 1.0 points. Under large sample sizes, the MAE and RMSE of the CEAAS were 0.26 and 0.40. The ablation experiment showed that the final MAE and RMSE of the complete model reached 0.32 and 0.41, and the PCC increased to 0.91. Practical application verification showed that in environmental protection themes,

MAE was reduced to 0.34 and RMSE was reduced to 0.40. The throughput of the model was as high as 82.11 articles/s, with a response delay of only 12.09 ms. The GPU utilisation and single article energy consumption were also significantly lower than other models. In summary, the CEAAS effectively broke through the limitations of CPOTD, achieved collaborative optimisation of scoring accuracy and deployment efficiency, and provided a feasible solution for automatic scoring of multi-scenario essays. However, the scoring dimensions and standards of different standardised exams such as CET-4 and CET-6, TOEFL, etc. are different, while this study mainly focuses on the overall scoring of essays. Therefore, future work should further explore the specific scoring dimensions and standards of different examination systems to further improve and enhance the applicability of the model. In addition, the CEAAS model proposed in the study can be adapted to mainstream English writing teaching scenarios such as K12 high-order, higher education, and large-scale testing, and has practical value in cross prompt scoring accuracy and deployment efficiency. However, it is necessary to acknowledge its limitations in extreme length essay adaptation, generalisation of multiple native language backgrounds, and modelling of subjective dimensions. Therefore, in the future, improvements can be made through Transformer long sequence modelling, construction of multi native language datasets, and subjective feature fusion. The model will further cover low-level English learning, creative writing evaluation, and multi native language background teaching scenarios, ultimately achieving an upgrade from automated scoring tools to personalised writing teaching assistance systems, providing more accurate ability assessment and guidance for English learners at different stages and backgrounds.

Declarations

All authors declare that they have no conflicts of interest.

References

- Abdul Salam, M., El-Fatah, M.A. and Hassan, N.F. (2022) 'Automatic grading for Arabic short answer questions using optimized deep learning model', *PLOS ONE*, August, Vol. 17, No. 8, pp.272269–272276, DOI: 10.1371/journal.pone.0272269.
- Ala'raj, M., Abbod, M.F., Majdalawieh, M. and Jum'a, L. (2022) 'A deep learning model for behavioural credit scoring in banks', *Neural Comput. Appl.*, January, Vol. 34, No. 8, pp.5839–5866, DOI: 10.1007/s00521-021-06695-z.
- Bai, H., Wang, P., Zhang, R. and Su, Z. (2023) 'Segformer: a topic segmentation model with controllable range of attention', *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37, No. 11, pp.12545–12552, DOI: 10.1609/aaai.v37i11.26477.
- Bhosle, K. and Musande, V. (2023) 'Evaluation of deep learning CNN model for recognition of Devanagari digit', *Artif. Intell. Appl.*, February, Vol. 1, No. 2, pp.114–118, DOI: 10.47852/bonviewAIA3202441.
- Chen, H. and Pan, J. (2022) 'Computer or human: a comparative study of automated evaluation scoring and instructors' feedback on Chinese college students' English writing', *Asian-Pacific Journal of Second and Foreign Language Education*, Vol. 7, No. 1, pp.34–41, DOI: 10.1186/s40862-022-00171-4.
- Dawn, N., Ghosh, S., Saha, A., Chatterjee, S., Ghosh, T., Guha, S., Sarkar, S., Mukherjee, P. and Sanyal, T. (2024) 'A review on digital twins technology: a new frontier in agriculture', *Artif. Intell. Appl.*, March, Vol. 2, No. 4, pp.250–262, DOI: 10.47852/bonviewAIA3202919.

- di Summa, M., Mosca, N., Colella, L., Bianco, V., Ferraro, P. and Stella, E. (2023) 'Realistic interactions with micro-objects in a virtual environment', *AIA*, July, Vol. 12621, No. 7, pp.206–210, DOI: 10.1117/12.2678665.
- Ferrara, S. and Qunbar, S. (2022) 'Validity arguments for AI-based automated scores: essay scoring as an illustration', *J. Educ. Meas.*, June, Vol. 59, No. 3, pp.288–313, DOI: 10.1111/jedm.12333.
- Firoozi, T., Mohammadi, H. and Gierl, M.J. (2023) 'Using active learning methods to strategically select essays for automated scoring', *Educ. Meas.: Issues Pract.*, December, Vol. 42, No. 1, pp.34–43, DOI: 10.1111/emip.12537.
- Gombert, S., Fink, A., Giorgashvili, T., Jivet, I., Di Mitri, D., Yau, J. and Drachsler, H. (2024) 'From the automated assessment of student essay content to highly informative feedback: a case study', *Int. J. Artif. Intell. Educ.*, January, Vol. 34, No. 4, pp.1378–1416, DOI: 10.1007/s40593-023-00387-6.
- Haruna, A.A., Muhammad, L.J. and Abubakar, M. (2022) 'Novel thermal-aware green scheduling in grid environment', *Artif. Intell. Appl.*, November, Vol. 1, No. 4, pp.244–251, DOI: 10.47852/bonviewAIA2202332.
- Hasanvand, M., Nooshyar, M., Moharamkhani, E. and Selyari, A. (2023) 'Machine learning methodology for identifying vehicles using image processing', *AIA*, April, Vol. 1, No. 3, pp.170–178, DOI: 10.47852/bonviewAIA3202833.
- Katkade, S.N., Bagal, V.C., Manza, R.R. and Yannawar, P.L. (2023) 'Advances in real-time object detection and information retrieval: a review', *Proc. Artif. Intell. Appl.*, March, Vol. 1, No. 3, pp.139–144, DOI: 10.47852/bonviewAIA3202456.
- Kim, S. and Jo, M. (2024) 'Is GPT-4 alone sufficient for automated essay scoring? A comparative judgment approach based on rater cognition', *Proc. ACM Conf. Learn. Scale*, July, Vol. 12, No. 1, pp.315–319, DOI: 10.1145/3657604.3664703.
- Kumar, S. and Solanki, A. (2023) 'An abstractive text summarization technique using transformer model with self-attention mechanism', *Neural Comput. Appl.*, June, Vol. 35, No. 25, pp.18603–18622, DOI: 10.1007/s00521-023-08687-7.
- Li, W. and Liu, H. (2024) 'Applying large language models for automated essay scoring for non-native Japanese', *Humanities and Social Sciences Communications*, Vol. 11, No. 1, pp.1–15, DOI: 10.1057/s41599-024-03209-9.
- Li, X., Yang, H., Hu, S., Geng, J., Lin, K. and Li, Y. (2022) 'Enhanced hybrid neural network for automated essay scoring', *Expert Syst.*, June, Vol. 39, No. 10, pp.13068–13072, DOI: 10.1111/exsy.13068.
- Ma, Y., Huang, Z., Su, J., Shi, H., Wang, D., Jia, S. and Li, W. (2023) 'A multi-channel feature fusion CNN-Bi-LSTM epilepsy EEG classification and prediction model based on attention mechanism', *IEEE Access*, June, Vol. 11, pp.62855–62864, DOI: 10.1109/ACCESS.2023.3287927.
- Misgna, H., On, B.W., Lee, I. and Choi, G.S. (2024) 'A survey on deep learning-based automated essay scoring and feedback generation', *Artif. Intell. Rev.*, December, Vol. 58, No. 2, pp.36–42, DOI: 10.1007/s10462-024-11017-5.
- Poddar, A.K. and Rani, R. (2023) 'Hybrid architecture using CNN and LSTM for image captioning in Hindi language', *Procedia Comput. Sci.*, January, Vol. 218, No. 1, pp.686–696, DOI: 10.1016/j.procs.2023.01.049.
- Preethi, P. and Mamatha, H.R. (2023) 'Region-based convolutional neural network for segmenting text in epigraphical images', *Artif. Intell. Appl.*, September, Vol. 1, No. 2, pp.119–127, DOI: 10.47852/bonviewAIA2202293.
- Ramesh, D. and Sanampudi, S.K. (2022) 'An automated essay scoring systems: a systematic literature review', *Artif. Intell. Rev.*, September, Vol. 55, No. 3, pp.2495–2527, DOI: 10.1007/s10462-021-10068-2.

- Ren, P., Yang, L. and Luo, F. (2023) 'Automatic scoring of student feedback for teaching evaluation based on aspect-level sentiment analysis', *Educ. Inf. Technol.*, July, Vol. 28, No. 1, pp.797–814, DOI: 10.1007/s10639-022-11151-z.
- Schaap, M.J., Cardozo, N.J., Patel, A., de Jong, E.M., van Ginneken, B. and Seyger, M.M. (2022) 'Image-based automated psoriasis area severity index scoring by convolutional neural networks', *J. Eur. Acad. Dermatol. Venereol.*, October, Vol. 36, No. 1, pp.68–75, DOI: 10.1111/jdv.17711.
- Sharma, A., Katlaa, R., Kaur, G. and Jayagopi, D.B. (2023) 'Full-page handwriting recognition and automated essay scoring for in-the-wild essays', *Multimed. Tools Appl.*, March, Vol. 82, No. 23, pp.35253–35276, DOI: 10.1007/s11042-023-14558-z.
- Sharoufi, H. (2022) 'Towards a unified English technology-based writing curriculum in the Arabian Gulf countries: the case of Oman', *Language Testing in Asia*, Vol. 12, No. 1, pp.33–36, DOI: 10.1186/s40468-022-00178-1.
- Singh, S., Pupneja, A., Mital, S., Shah, C., Bawkar, M., Gupta, L.P. and Shah, R.R. (2023) 'H-AES: towards automated essay scoring for Hindi', *Proc. AAAI Conf. Artif. Intell.*, July, Vol. 37, No. 13, pp.15955–15963, DOI: 10.1609/aaai.v37i13.26894.
- Thompson, I., Koenig, N., Mracek, D.L. and Tonidandel, S. (2023) 'Deep learning in employee selection: evaluation of algorithms to automate the scoring of open-ended assessments', *J. Bus. Psychol.*, March, Vol. 38, No. 3, pp.509–527, DOI: 10.1007/s10869-023-09874-y.
- Usman, A.M. and Abdullah, M.K. (2023) 'An assessment of building energy consumption characteristics using analytical energy and carbon footprint assessment model', *Green Low-Carbon Econ.*, March, Vol. 1, No. 1, pp.28–40, DOI: 10.47852/bonviewGLCE3202545.
- Uto, M. and Okano, M. (2022) 'Learning automated essay scoring models using item-response-theory-based scores to decrease effects of rater biases', *IEEE Trans. Learn. Technol.*, January, Vol. 14, No. 6, pp.763–776, DOI: 10.1109/TLT.2022.3145352.
- Uto, M., Aomi, I., Tsutsumi, E. and Ueno, M. (2023) 'Integration of prediction scores from various automated essay scoring models using item response theory', *IEEE Trans. Learn. Technol.*, March, Vol. 16, No. 6, pp.983–1000, DOI: 10.1109/TLT.2023.3253215.