



**International Journal of Continuing Engineering Education
and Life-Long Learning**

ISSN online: 1741-5055 - ISSN print: 1560-4624

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

**Intelligent English composition scoring method based on
transfer paragraph segmentation and Pearson detection**

Haibei Chen

DOI: [10.1504/IJCEELL.2026.10079803](https://doi.org/10.1504/IJCEELL.2026.10079803)

Article History:

Received: 17 September 2025

Accepted: 30 November -0001

Published online: 01 October 2026

Intelligent English composition scoring method based on transfer paragraph segmentation and Pearson detection

Haibei Chen

School of Culture and Tourism,
Guangdong Engineering Polytechnic,
Guangzhou, 510520, China
Email: haibeich@outlook.com

Abstract: With the growing demand for intelligent education, automated English composition scoring has gained significant research attention. Addressing limitations in cross-prompt generalisation and feature extraction, this study proposes a transfer learning-based method incorporating paragraph segmentation and Pearson detection. The approach combines BERT encoding, convolutional and bidirectional LSTM networks, augmented with attention mechanisms and cross-topic feature sharing. In performance evaluation, the proposed model achieved a weighted kappa coefficient of 0.866 and a Pearson correlation coefficient of 0.79 on the test set, significantly better than the comparison model in terms of consistency and correlation. It also showed lower prediction bias in error metrics, demonstrating higher fitting accuracy and rating stability. In three typical application scenarios, the model consistently outperformed the baseline method in terms of off topic detection accuracy and F1 score, and could more effectively capture semantic and structural features. In addition, in terms of resource consumption, the model parameter count was only 16.7M, the inference time was 28.9 ms, and the maximum processor utilisation rate was only 73.6%, demonstrating good deployment adaptability. The experimental results show that this method has good accuracy and generalisation ability in multi-scenario automatic scoring tasks, and has potential value for deployment and application in educational scenarios.

Keywords: English composition scoring; transfer paragraph segmentation; semantic modelling; off topic detection; feature sharing.

Reference to this paper should be made as follows: Chen, H. (2026) 'Intelligent English composition scoring method based on transfer paragraph segmentation and Pearson detection', *Int. J. Continuing Engineering Education and Life-Long Learning*, Vol. 36, No. 13, pp.1–25.

Biographical notes: Haibei Chen graduated from Wuhan University in Wuhan, China with a Bachelor's degree in English in 2002. She obtained her Master's degree in Software Engineering from Guangdong University of Technology in Guangzhou, China in 2005. Presently, she is working as an Associate Professor of English at the School of Culture and Tourism of Guangdong Engineering Polytechnic, Guangzhou. She has published over 30 papers in public and presided more than ten provincial-level research projects. Her research interests include English language testing, English language teaching, and British and American literature.

1 Overview

In the context of the continuous digitisation and intelligence of the current education evaluation system, automated composition scoring technology is gradually becoming an important component of intelligent education (Shen and Coker, 2023). The traditional manual scoring method is not only costly and inefficient, but also has significant shortcomings in subjectivity and consistency (Wolters and Kim, 2024). Therefore, automatic scoring systems based on natural language processing (NLP) and machine learning technologies have received widespread attention, especially in scenarios such as English writing training and exam evaluation, with broad application prospects. In recent years, multiple research teams have made positive progress in the field of automatic scoring of English compositions. Hu (2024) designed an English text automatic scoring algorithm based on graph neural networks that combines deep syntactic smoothing and semantic matching. By conducting comparative experiments to score these data, the results showed that the algorithm outperformed traditional methods in terms of scoring accuracy, could achieve better scoring results, and better handled automatic paper scoring problems. The actual mean value was 0.790, the smoothness algorithm was 0.768, and the text matching vector was 0.759. Martín-Monje and Barcena (2024) adapted the G-Rubric system to the tourism English distance education scenario and explored its potential to replace or assist human mentors in essay evaluation. The system provided dual feedback of digital scoring and form/function, with content evaluation relying on pattern matching and machine reasoning from professional corpora. By comparing the performance differences between the system and human mentors, it was pointed out that further exploration of the optimal integration path of human-machine collaboration was needed to provide empirical direction for the educational application of intelligent evaluation tools. Nath et al. (2025) developed a modular framework called Assamese translation with modular transfer (AstraMT) for achieving small sample translation from Assamese to English. This framework integrates the following key components: context aware prompt selector, syntactic prompt template, multi-translation reordering mechanism based on bilingual evaluation replacement score and cross language optimisation translation evaluation indicators, and a lightweight post editing module to correct named entity errors and complete translation omissions. Experimental results showed that AstraMT can adapt to different large language models without the need for additional fine-tuning. Jiang (2025) conducted work based on transformer pre-trained language models, utilising their superior contextual understanding and automatic feature extraction capabilities to optimise the model's handling of complex syntactic structures and long-distance dependencies. This study significantly enhanced the generalisation and robustness of the model by combining large-scale diversified datasets with data augmentation techniques.

With the widespread application of deep learning (DL) in NLP tasks, the combination of transfer learning and neural network structures has become a key path to improve text modelling ability and scoring accuracy (García-Méndez et al., 2023). Especially in scenarios where data such as English essay grading is scarce and tasks are heterogeneous, models have higher requirements for cross task semantic transfer and structural generalisation ability (Jin et al., 2023). In the last few years, transfer learning has been broadly utilised in automatic scoring systems to address the issues of insufficient annotated samples and domain adaptation. Mezzadri et al. (2024) proposed a statistical framework for category learning data, using customised segmentation/clustering

techniques to adapt transfer models. By applying the generalised context model in three sequential effect experiments, the framework successfully detected the performance difference caused by the rule order advantage in two-thirds of the experiments. The adjusted framework enabled the transfer model to adapt to learning data characteristics while maintaining pattern capture capability (Mezzadri et al., 2024). De Nardin et al. (2024) investigated the effectiveness of DL model initialisation strategies through pre-training on cross-domain and intra-domain datasets. The experiment used two publicly available datasets with different structures and language features, and the results showed that combining cross-domain and intra-domain transfer learning methods could significantly improve model performance and accelerate convergence. In terms of semantic modelling, bidirectional long short-term memory (BiLSTM) networks have been broadly utilised in text analysis due to their ability to capture contextual dependencies (Li et al., 2023). Madhubala et al. (2024) proposed a medical information retrieval system that integrates DL and knowledge base management. The system is based on BiLSTM to construct a context representation retrieval network, and combines the knowledge re-ranking (KR-R) algorithm to optimise document relevance. By integrating medical record semantic understanding technology, the accuracy of biomedical retrieval has been significantly improved (Madhubala et al., 2024). Islam et al. (2024) designed a knowledge-based hybrid attention capsule deep learning method (KHACDD) for sentiment classification tasks. This method integrates a hierarchical capsule architecture with attention mechanism, bidirectional gated recurrent unit, BiLSTM, extended convolutional neural networks (CNNs) and domain knowledge, significantly enhancing feature representation and sentiment classification capabilities (Islam et al., 2024).

In summary, in existing research, most scoring models focus on overall text modelling, ignoring paragraph structure, inter sentence logic, and writing variability under different themes, making it hard to accurately depict the hierarchical nature of composition content and the multidimensional distribution of scoring elements. In addition, when facing real-world application scenarios such as cross task and cross topic scenarios, models often suffer from insufficient generalisation ability and weak feature transfer ability, making it hard to satisfy the multiple requirements for stability and scalability in educational practice. To this end, the research designs an intelligent scoring method for English compositions based on transfer paragraph segmentation and Pearson detection, which innovatively integrates transfer learning, paragraph semantic modelling, and Pearson correlation detection. By combining shallow language features with deep semantic representation, two-stage modelling is used to extract shared features unrelated to prompts and scoring features related to prompts. Combined with the bidirectional encoder representation from transformers (BERT) word vector and BiLSTM-attention mechanism, paragraph level modelling and scoring of essays are achieved.

The innovation of the research is mainly reflected in the following three aspects:

- 1 A two-stage transfer learning framework is proposed, which obtains stable cross topic features through paragraph level semantic alignment in the irrelevant stage, and achieves topic adaptation ability through cross topic feature sharing in the relevant stage, effectively improving the cross topic generalisation performance of the essay automatic scoring model.

- 2 A hybrid scoring network that combines deep semantic modelling with shallow feature enhancement is constructed. This network is based on BERT deep semantic encoding and improves the robustness of the model in complex discourse structures through a regularised shallow feature guidance mechanism, enabling the scoring results to better reflect the language quality and discourse organisation of the text.
- 3 A Pearson correlation detection mechanism is introduced for out-of-topic diagnosis, to assess the linear consistency of the model under different prompts, improve the stability and cross-topic consistency of the scoring process, and provide observable explanatory dimensions for model behaviour.

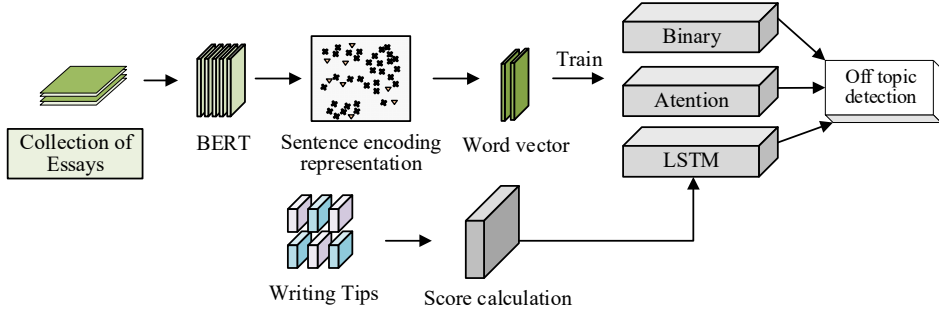
2 Experiment design

2.1 *Transfer paragraph segmentation mechanism based on cross-topic structure alignment*

In contemporary English writing teaching, the rationality and logical coherence of students' composition structure not only reflect their language proficiency, but also reflect the maturity of cross-cultural discourse awareness and thinking organisation ability. Especially in a second language writing environment dominated by non-native English speakers, students often encounter problems such as unclear paragraph division, logical breaks, or theme shifts due to a lack of deep understanding of English discourse structure. These structural deficiencies directly affect the evaluation of essay quality and the effectiveness of teaching feedback. Therefore, accurately constructing and identifying paragraph structures has become a key issue in writing evaluation research that combines linguistic and educational significance. In the last few years, as the artificial intelligence and NLP technology continuously develop, the use of related technologies to achieve automated analysis of students' writing logic and expression structure has become a hot topic of concern in academia. In this context, transfer learning, as an important means of cross task knowledge transfer, provides a new technical path for solving the problems of structural judgement bias and grading adaptation in English writing structure recognition. To enhance the understanding and adaptability of the model to English composition structure and grading, a transfer paragraph segmentation mechanism is introduced to transfer the structure recognition ability learned by the pre-trained language model in general text to the composition grading task. By using semantic boundary recognition and sentence vector aggregation, automatic segmentation and hierarchical modelling of essay paragraphs can be achieved, improving the expression effect of structural dimensions (Purohit and Dave, 2023). The study models paragraph recognition as a sequence annotation task based on sentence vector boundary prediction. The pre-trained language model has transferred the discourse structure clues learned from large-scale corpora, such as topic progression, inter sentence cohesion, and semantic inflection, to be used for essay paragraph judgement. When processing compositions, BERT first generates sentence level semantic representations, followed by BiLSTM-attention capturing discourse structure signals such as inter sentence continuity, topic progression, and semantic inflection, and using boundary classifiers to predict whether sentences constitute paragraph start and end points. The generated paragraph level representation is further input into the off topic detection and scoring module, allowing the paragraph

structure to participate as transferable mid-level discourse features in subsequent inference, enhancing the model's cross topic generalisation ability. This strategy effectively strengthens the model's ability to identify semantic and structural deviations in text, and the overall modelling process is shown in Figure 1.

Figure 1 Off topic scoring process (see online version for colours)



In Figure 1, the composition collection is first encoded by BERT to form sentence level semantic representations, which are then converted into sentence vectors and input into the paragraph segmentation module. This module consists of binary, attention and LSTM, and generates paragraph vectors through training. At the same time, the question prompts are also encoded as topic vectors, which participate in the calculation of off topic questions together with paragraph vectors in subsequent stages. Finally, the model outputs off topic detection results based on the degree of matching between paragraph semantics and topic semantics. The BERT model demonstrates outstanding performance in the field of NLP. The study used BERT encoding to obtain text features, and generated word feature representations by summing and averaging the vector representations of each word in the sentence. The specific calculation is denoted in equation (1) (Ganai and Khursheed, 2023).

$$\begin{cases} v_i = BERT^{represent}(c_i) \\ represent = \frac{1}{3} \sum_k layer_k \end{cases} \quad (1)$$

In equation (1), v_i represents the word vector representation of word c_i , c_i represents the i^{th} word in the input sentence, and $layer_k$ represents the output of the k^{th} layer of the BERT model. The study used the BiLSTM-attention model to capture text sequence features, and dynamically evaluated the contribution of keywords to sentence coherence through BERT word vector input combined with attention mechanism. The temporal encoding process of BiLSTM is shown in equation (2) (Chabalala et al., 2025).

$$r_t = BiLSTM(v_t) \quad (2)$$

In equation (2), r_t represents the output vector of BiLSTM at time t , and v_t represents the word vector input at time t . At time step t , BiLSTM encodes the input to obtain r_t , and adjusts the weights through attention mechanism. The calculation process is shown in equation (3) (Singh et al., 2025).

$$\begin{cases} h_t = \tanh(W_v \cdot R_t + b) \\ a_t = \frac{\exp(h_t^T \cdot h_v)}{\sum_t \exp(h_t^T \cdot h_v)} \end{cases} \quad (3)$$

In equation (3), h_t is the hidden representation after nonlinear mapping, W_v is the trainable weight matrix, R_t denotes the feature vector of the input sequence at time t , and b is the bias vector utilised to adjust the projection result. a_t represents attention weight, $h_t^T \cdot h_v$ represents similarity score, which measures the correlation between the features of t at time and the global vector. After processing by the attention layer, the final vector representation is output as denoted in equation (4).

$$u_t = a_t \cdot r_t \quad (4)$$

In equation (4), u_t means the input vector weighted by attention. After the attention mechanism generates word vectors, they are concatenated and input into a fully connected layer (FCL) for SoftMax classification. The concatenated vector representation is shown in equation (5).

$$\begin{cases} u' = fc(u) \\ y' = softmax(u' \cdot W_s + b_s) \end{cases} \quad (5)$$

In equation (5), u denotes the concatenation vector of the sentence, u' is the output vector of the FCL, and $fc(\cdot)$ denotes the FCL. y' is the final prediction result, W_s represents the weight matrix of the classification layer, and b_s represents the bias vector of the classification layer. The BiLSTM model adopts a cross-entropy loss function, and its batch calculation formula is shown in equation (6) (Shou et al., 2024).

$$L(Y, Y') = -\frac{1}{n} \sum_{i=1}^n (y_i \cdot \log(y'_i) + (1 - y_i) \cdot \log(1 - y'_i)) \quad (6)$$

In equation (6), L means the loss function, Y and Y' denote the true label vector and predicted value vector, y_i means the true label of the i^{th} sample, y'_i indicates the predicted probability of the i^{th} sample, and n indicates the amount of samples in a batch. The quality of semantic vector representation directly affects the similarity calculation between content and prompts (Zhu et al., 2022). The study maps the content and prompts of a composition to the same semantic space. Assuming the composition is divided into m segments, the segment vectors and topic prompt word vectors are obtained through vectorisation processing. The semantic similarity between the two is calculated, and the final off topic score is obtained. The calculation process is denoted in equation (7).

$$\begin{cases} S_i(T_i, P) = \frac{\langle T_i, P \rangle}{|T_i| |P|} \\ S_E(E, P) = \frac{\langle E, P \rangle}{|E| |P|} \end{cases} \quad (7)$$

In equation (7), S_i represents cosine similarity, S_E represents cosine similarity between the entire essay and the title, T_i represents semantic vector of the i^{th} paragraph of the essay, P

means semantic vector of the title, and EV represents overall semantic vector of the entire essay. The calculation of off topic detection score is shown in equation (8).

$$\begin{cases} Y = \frac{1}{m+2} \left(\left(\sum_{i=1}^m S_i \right) + \beta + S_e \right) \\ \beta = \frac{1 - e^{-n}}{1 + e^{-n}} \end{cases} \quad (8)$$

In equation (8), Y represents the final off topic detection score, β is the paragraph number correction coefficient, controlled by a Sigmoid like function, and gradually approaches 1 as the number of paragraphs m increases. The study used Pearson correlation coefficient (PCC) to analyse the correlation between the degree of digression and manual scoring. The essence of Pearson test is to measure the linear consistency between the output results of the model and the manual scores. After completing the detection of off topic scores or predicting the total score of an essay online, the PCC between the predicted score sequence and the manual scoring sequence is calculated as a fitting evaluation indicator for the quality of the scoring. Given the M essay dataset, let the manual scoring set be R and the off topic detection score set be D . The correlation calculation is shown in equation (9) (Dufera et al., 2023).

$$Per_{Y,Y_T} = \frac{E_V(Y \cdot Y_T) - E_V(Y)E_V(Y_T)}{\sqrt{E_V(Y^2) - E_V^2(Y)} \cdot \sqrt{E_V(Y_T^2) - E_V^2(Y_T)}} \quad (9)$$

In equation (9), Per_{Y,Y_T} represents the PCC between the off topic detection score and the total score of the composition, E_V denotes the expected value, and Y_T denotes the actual total score of the essay.

2.2 Construction of a scoring network integrating semantic modelling and shallow feature enhancement

On the basis of implementing automatic paragraph division and semantic structure modelling in composition, the model needs to further complete a comprehensive evaluation of the overall quality of the text. To fully capture the multi-level features of English compositions in content, language, and structure, a scoring network integrating deep semantic modelling with shallow feature enhancement is designed and used as the key scoring module following the paragraph modelling stage. To this end, a two-stage cross prompt task English composition scoring model is constructed by combining shallow features designed manually with DL features, and integrating off topic detection modules. The complete process is shown in Figure 2.

In Figure 2, the scoring model adopts a two-stage transfer learning framework of ‘prompt irrelevant prompt relevant’. The study first constructs a feature generator in the prompt irrelevant stage, which extracts unified semantic and writing style features from compositions under different topics through shared embedding layers, convolutional layers, and BiLSTM structures, to obtain shared representations that are irrelevant to the topic. In this stage, to eliminate the influence of topic variables on feature distribution, the topic column ablation operation is introduced, which removes topic related fields from the training input and only retains the text ontology, allowing the model to focus on

stable writing features that exist across topics. After entering the relevant prompting stage, the model is reintroduced with the question prompts and the shared representations learned in the first stage are used to combine and reorganise the semantic vectors of different questions into a combined representation to generate pseudo-data to expand the training samples, thereby enhancing the model’s generalisation ability in cross-question evaluation tasks. On the one hand, it generates pseudo data to expand the training set, and on the other hand, it models and learns the input information through a hierarchical neural network, ultimately achieving score prediction for the target essay. To further enhance the fine-grained discriminative ability of the model, a rule-based shallow text feature extraction module is also introduced in the study. This module uses statistical methods to design features at the lexical and syntactic levels, with a focus on the author’s mastery of vocabulary and sentence writing ability. The process of extracting shallow features is shown in Figure 3 (Liu et al., 2023).

Figure 2 Scoring model two-stage cross prompt framework (see online version for colours)

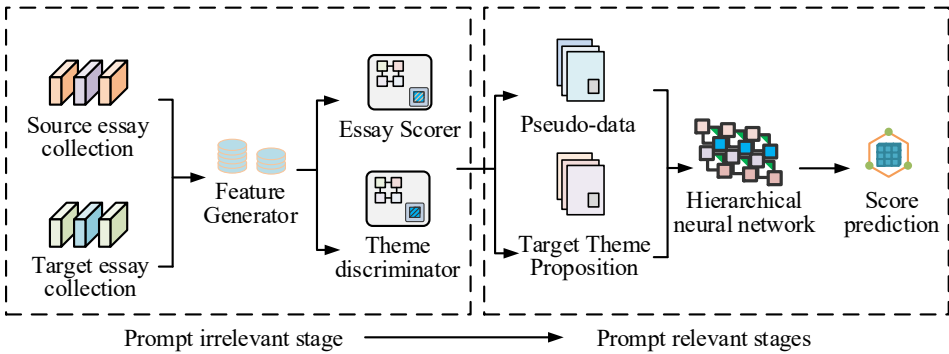
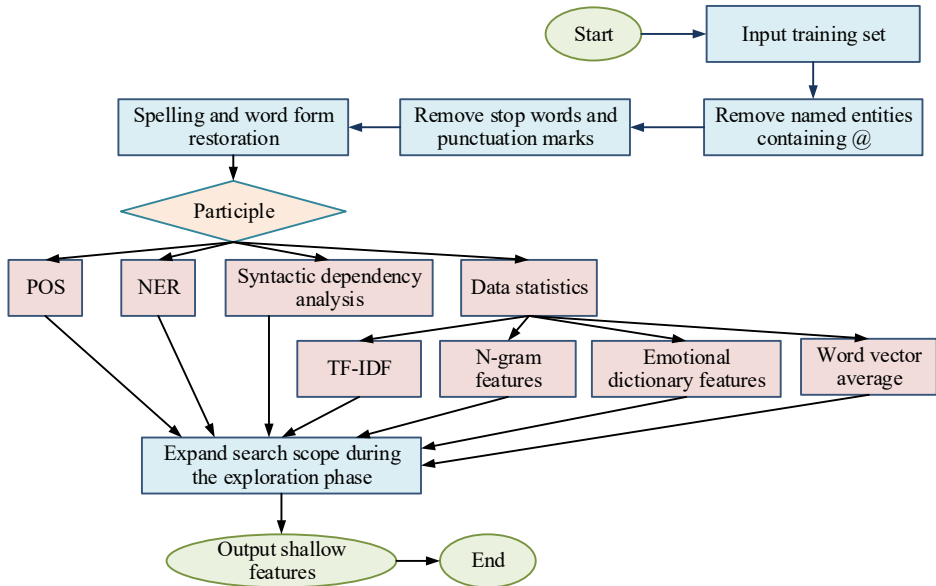


Figure 3 Shallow text feature extraction process (see online version for colours)



In Figure 3, text feature extraction first inputs the training set and sequentially performs preprocessing operations, including removing @ named entities, stop words, and punctuation marks, followed by spelling correction and word form restoration. Subsequently, the segmented text will undergo multiple feature extraction branches, such as part-of-speech tagging (POS), named entity recognition (NER), syntactic dependency analysis, and basic data statistics. These syntax and statistical information can be further used to construct semantic layer features such as TF-IDF vectors, N-gram features, sentiment dictionary features, and word vector averaging (Sheridan and Onsjö, 2024). Finally, all extracted features are integrated and processed through feature fusion to generate the final output features for model training or prediction. In the deep semantic modelling section, a hybrid structure combining CNN and long short-term memory (LSTM) network is designed to capture local language patterns and global coherence features in compositions. This DL module serves as the core component of deep feature extraction in the scoring model, and is input into the fusion layer together with shallow artificial features to strengthen the perception ability of multidimensional scoring factors. The overall model structure is shown in Figure 4.

Figure 4 Two-stage shared deep semantic modelling module (see online version for colours)

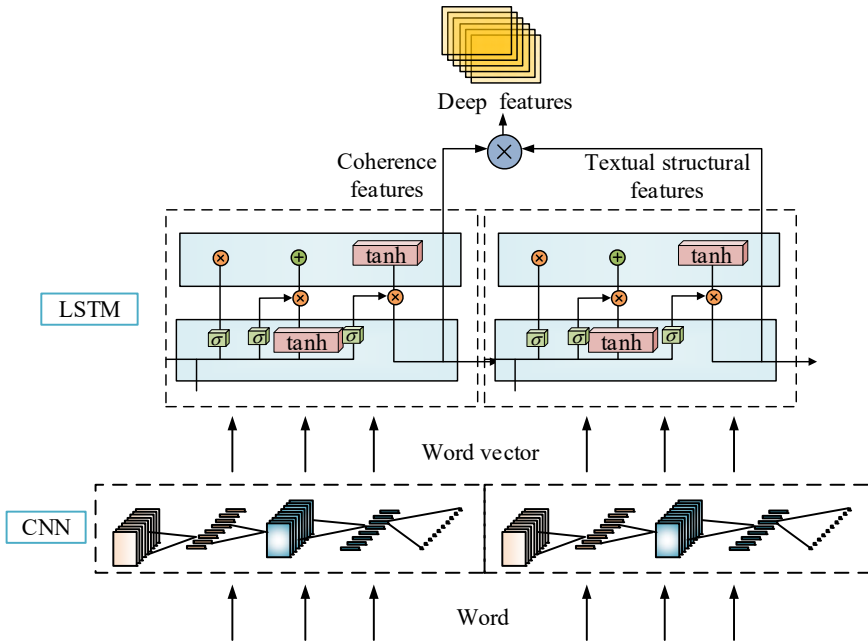


Figure 4 shows the deep semantic modelling module shared by both stages. This module uses Word2Vec to generate k dimensional word vectors. Starting from the word vectors in the input layer, each word is first represented as a sequence of embedding vectors. These vectors are then fed into the CNN to capture local textual features. The resulting representations are subsequently processed by the LSTM to model temporal dependencies in the context. Together, they construct a deep feature representation of the text (Dornaika and Hoang, 2023). After fusing the deep features with the shallow text features, they are simultaneously input into the rater and discriminator. The rater

optimises prediction accuracy, while the discriminator interferes with the source judgement of prompts. The two improve feature generalisation ability through adversarial training. The loss function is shown in equation (10).

$$L_f = L_y - \alpha L_d \quad (10)$$

In equation (10), L_f represents the total loss function of the feature generator, L_y represents the loss of the essay evaluator, L_d represents the loss of the topic discriminator, and α is the balance coefficient used to adjust the weight of the topic discriminator loss in overall optimisation. The scoring network adopts a multi-task learning structure, which takes into account both overall scores and local feature scores, improving the stability of the model and the richness of evaluation dimensions (Uto et al., 2023). The final two-stage scoring system is constructed, and its overall structure is shown in Figure 4.

Figure 5 The structure of relevant scoring stages (see online version for colours)

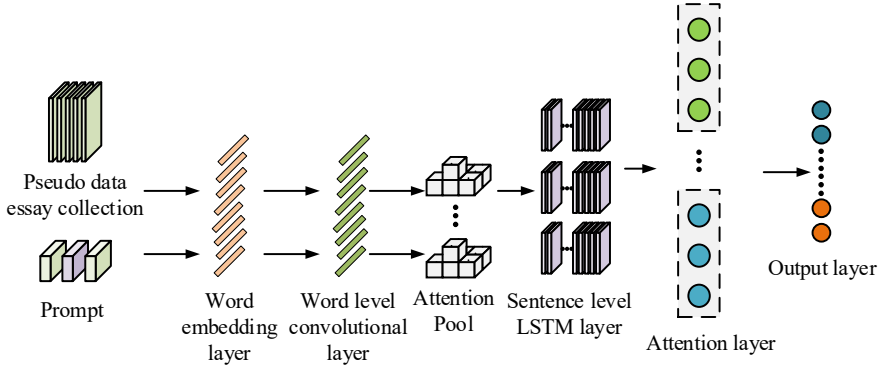


Figure 5 shows a two-stage scoring structure, in which the topic representation, cross topic feature sharing results, and paragraph level semantic vectors are fused and input into a deep scoring network to obtain the final essay score and off topic detection results. The two-stage scoring structure begins with inputting pseudo data into a text set and prompts. The text is then transformed into a vector representation through a word embedding layer, and local semantic features are extracted through a word level convolutional layer. Subsequently, it enters an attention pool to aggregate information from different positions. On this basis, the model constructs a sentence level LSTM layer to capture long-range dependencies, and overlays an attention layer on top of it to highlight the semantic contribution of key sentences. Different task branches share the underlying embedding and convolutional features, and separate output layers are connected after sentence level LSTM and attention mechanisms to achieve multi-task parallel prediction. The study uses one-dimensional CNN to extract part of speech features of words and generates sentence representations through attention pooling. Convolutional feature extraction is shown in equation (11) (Preethi and Mamatha, 2023).

$$F_i = f(W_F \cdot [w_i; w_{i+H_w-1}] + b_F) \quad (11)$$

In equation (11), F_i represents the convolutional feature of a word, W_F indicates the weight matrix of the convolutional kernel, H_w indicates the size of the convolutional kernel, b_F is the bias term of the convolutional feature, and w_i represents the word in the

sentence. Further through linear mapping and tanh activation function, the feature map is obtained as shown in equation (12) (Ma et al., 2024).

$$a_i = \tanh(W_a \cdot F_i + b_a) \quad (12)$$

In equation (12), a_i represents the attention of the word after mapping, W_a represents the weight matrix used for feature mapping, and b_a is the bias term for attention mapping. By using softmax normalisation, the importance of all words is distributed within the $[0, 1]$ interval to obtain the attention weight of each word, as denoted in equation (13).

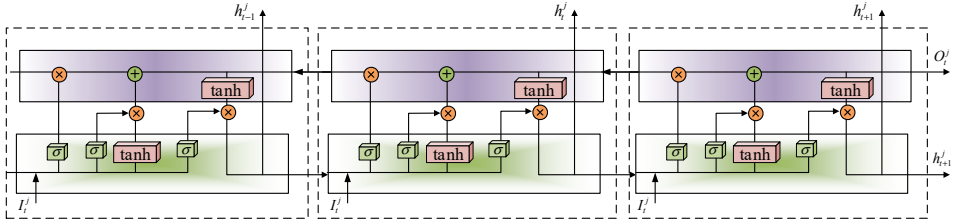
$$\omega_i = \frac{e^{w_a \cdot a_i}}{\sum_j e^{w_a \cdot a_j}} \quad (13)$$

In equation (13), ω_i represents the attention weight of each word, and w_a is the trainable attention parameter vector. The ω_i is multiplied by the corresponding F_i , and the results for all words are summed to obtain the sentence-level representation as shown in equation (14).

$$Q = \sum \omega_i F_i \quad (14)$$

In equation (14), Q is the final sentence representation. Given the crucial role of text order in semantic evaluation, BiLSTM is adopted in the study, with its structure shown in Figure 6 (Alawi and Bozkurt, 2025).

Figure 6 BiLSTM structure (see online version for colours)



In Figure 6, I_t^j indicates the input vector of the current time t , h_t^j represents the hidden state vector of the j^{th} task at time t , h_{t-1}^j and h_{t+1}^j are the hidden state vectors of the previous and next time respectively, O_t^j represents the output gate vector of the j^{th} task at time t , and σ represents the sigmoid activation function. BiLSTM integrates temporal information to generate comprehensive representations at time step t , thereby comprehensively capturing contextual semantics. For task j , BiLSTM is used to extract sentence level semantic features, as shown in equation (15) (Li et al., 2024).

$$\begin{cases} O_t^j = \sigma(W_o^j \cdot I_t^j + U_o^j \cdot h_{t-1}^j + b_o^j) \\ h_t^j = O_t^j \cdot \tanh(k_t^j) \end{cases} \quad (15)$$

In equation (15), O_t^j represents the output gate vector of the j^{th} task at time t , W_o^j represents the weight matrix between the input and output gates, I_t^j indicates the input

vector of the current time t , h_t^j represents the hidden state vector of the j^{th} task at time t , h_{t-1}^j indicates the hidden state vector of the previous time, k_t^j represents the unit state of the j^{th} task at time t , U_o^j indicates the weight matrix of the hidden state h_{t-1}^j and output gate at the previous time, and b_o^j indicates the bias term of the output gate. The final vector representation of the composition is obtained through attention pooling processing, and the calculation is shown in equation (16).

$$O^j = \omega_t^j h_t^j \quad (16)$$

In equation (16), O^j represents the final article vector. Based on the article vector representation, the rating prediction within the range of 0–1 is output through the sigmoid activation function, and the calculation is shown in equation (17).

$$Y' = \text{sigmoid}(W_y^j O^j + b_y^j) \quad (17)$$

In equation (17), Y' represents the final predicted score, W_y^j indicates the weight vector, and b_y^j indicates the bias vector, utilised to adjust the translation of the prediction results.

3 Results

3.1 Model performance validation and ablation experimental analysis

To validate the effectiveness of the discourse aware transfer segmentation scoring model (DTS Score), the ASAP dataset and composition dataset were selected as the evaluation corpus and broken into training and testing sets in a 7:3 ratio. Text preprocessing included basic cleaning, segmentation, and subtyping, with input length uniformly truncated/filled. BERT used the average of vectors from layers 9–11 to enhance transfer robustness, and then integrated BiLSTM and attention module to complete discourse modelling. The experimental configuration and parameter settings are denoted in Table 1.

Table 1 Experimental configuration parameter table

<i>Parameter category</i>	<i>Parameter description</i>	<i>Value/setting</i>
Hardware environment	CPU	Intel Core i9-13900K @ 3.0 GHz
	GPU	NVIDIA GeForce RTX 4090 (24 GB)
	Memory	128 GB DDR5
	OS	Ubuntu 22.04 LTS
Software environment	Programming language	Python 3.10
	DL framework	PyTorch 2.2 + CUDA 12.1
	NLP Toolkit	Transformers (HuggingFace)
Model parameters	Embedding dimension (BERT)	768
	BiLSTM hidden size	256 (per direction)
	Attention mechanism dimension	128

According to the configuration in Table 1, an ablation experiment was designed and conducted to conduct unified component analysis on paragraph semantic alignment, cross topic feature sharing, attention mechanism, CNN-BiLSTM, BERT encoder, and Pearson detection component. Each time only a single module was removed or shielded, and the rest of the network structure, parameter initialisation, and training super parameters remained unchanged. All variants were trained from scratch on the same training set, and the same verification set and early stop strategy were used for model selection, so as to ensure the comparability of different model results under the same experimental conditions. The changes in model performance under different combinations in terms of weighted kappa (QWK), PCC, mean absolute error (MAE), and root mean square error (RMSE) are observed. The findings are denoted in Table 2.

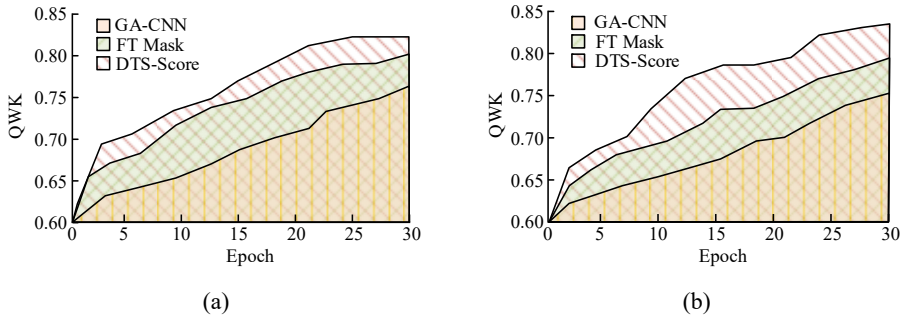
Table 2 Results of ablation experiment

<i>Model</i>	<i>QWK</i>	<i>PCC</i>	<i>MAE</i>	<i>RMSE</i>
Full model	0.726	0.713	0.392	0.521
w/o segment alignment	0.709	0.698	0.417	0.548
w/o cross-topic sharing	0.701	0.694	0.428	0.559
w/o attention	0.694	0.682	0.439	0.571
w/o CNN-BiLSTM	0.703	0.691	0.431	0.564
w/o BERT	0.705	0.693	0.434	0.566
w/o Pearson detection	0.715	0.695	0.409	0.536

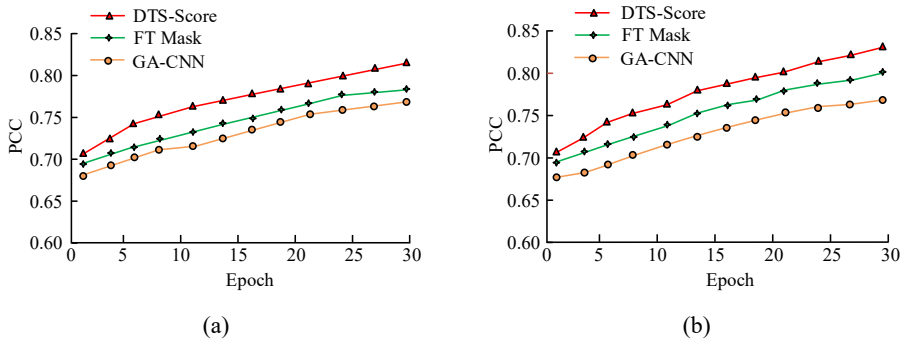
According to Table 2, the DTS core model achieved the best performance in all four indicators, with QWK and PCC being 0.726 and 0.713, respectively, and MAE and RMSE being 0.392 and 0.521, respectively. This indicates that when the modules of paragraph alignment, cross topic feature sharing, attention mechanism, and Pearson detection collaborate, the model is most superior in terms of evaluation consistency and error control. After removing paragraph semantic alignment or cross topic feature sharing, QWK decreased to 0.709 and 0.701, respectively, and MAE increased to 0.417 and 0.428, indicating that these two parts contribute significantly to the stability and generalisation ability of cross topic scoring. When the attention mechanism was removed, QWK further decreased to 0.694 and RMSE increased to 0.571, indicating that weighted modelling of key semantic segments is particularly important for fine scoring. When removing CNN-BiLSTM or BERT, QWK remained in the range of 0.703–0.705 and MAE increased to 0.431–0.434, indicating that local grammar patterns and deep semantic representations are equally fundamental for constructing high-quality essay vectors. In contrast, the performance degradation caused by removing Pearson detection is relatively small, with QWK only slightly reduced to 0.715 and MAE increased to 0.409, but still inferior to the complete model. This indicates that although Pearson correlation is not the main loss term, it provides useful auxiliary constraints in model selection and scoring stability. To further compare the feasibility of the designed model, it was compared and tested with optimisation fusion algorithms based on genetic algorithm-convolutional neural network (GA-CNN) (Wang, 2023), as well as segmentation algorithms based on feature transfer mask (FT Mask) (Liu et al., 2024). GA-CNN is a typical fusion architecture of convolutional neural networks and optimisation algorithms, which can simulate shallow semantic modelling methods mainly based on local language patterns. It

is often used as a traditional deep learning baseline in automatic essay grading tasks to evaluate the improvement of feature extraction efficiency and grading stability. FT Mask is an emerging text modelling method based on transfer learning and feature reassignment. Through cross task feature transfer and masking mechanisms, it enhances the recognition ability of discourse structure and text boundaries, making it suitable as a strong baseline model in the field of transfer learning. These two methods represent the main directions of non-transfer-based deep feature modelling and transfer-based feature enhancement modelling, respectively, and can provide stable and comparable performance references for cross topic essay scoring models from different dimensions. The variation results of different models on the dataset using QWK as the indicator are shown in Figure 7.

Figure 7 QWK results, (a) training set (b) test set (see online version for colours)



Figures 7(a) and 7(b) show the variation trends of QWK metrics with iteration times for three models on the training and testing sets, respectively. In Figure 7(a), in the initial stage, the QWK of DTS-Score in the fifth round was close to 0.70, significantly higher than the GA-CNN model's about 0.64 and the FT Mask model's about 0.67. As the training epochs increased, the gap gradually widened. In the 20th round, the QWK of DTS-Score reached about 0.83, FT Mask was 0.77, and GA-CNN was only 0.72. The curve growth of the designed model was smoother and converged faster, indicating its stronger learning ability and rating level modelling ability during the training process, which can better fit the training data. In Figure 7(b), starting from the 5th round, the QWK of DTS-Score exceeded 0.68, while FT Mask was 0.65 and GA-CNN was only 0.63. By the 15th round, the model designed by the research broke through 0.78, while the comparison models were 0.72 and 0.69, respectively, and the performance gap continued to widen. At the end of the 30th round, DTS-Score achieved a QWK of 0.83 on the test set, significantly better than FT Mask and GA-CNN, fully demonstrating its stronger scoring stability and consistency when facing unseen data. From this, the model designed by the research outperforms GA-CNN and FT Mask methods in both the training and testing sets in terms of QWK metrics, indicating that it not only has good training fitting ability, but also stronger generalisation performance, making it a more effective automatic scoring modelling scheme. To further examine the consistency between the model's predicted scores and manual scoring, the study used PCC as the evaluation metric and compared the performance of different methods on cross prompt essay scoring tasks. The results are shown in Figure 8.

Figure 8 PCC results, (a) training set (b) test set (see online version for colours)

Figures 8(a) and 8(b) respectively show the trend of PCC of our research model and two comparative methods on the training and testing sets with the number of iterations. In Figure 8(a), DTS-Score maintained a leading position in all rounds, reaching 0.64 in the fifth round, while GA-CNN and FT Mask were 0.59 and 0.56, respectively, indicating weak initial learning ability. As the training progressed, the PCC values of all models steadily increased, but there were differences in the growth rate. In the 10th round, DTS-Score reached 0.76, the GA-CNN model was 0.71, and the FT Mask was 0.68. By the 20th round, DTS-Score had finally stabilised at 0.86, while the comparison models were 0.80 and 0.77, respectively, indicating that DTS-Score had more advantages in feature fitting ability and convergence efficiency. In Figure 8(b), the generalisation ability of DTS-Score was still the most superior, reaching 0.61 in the fifth round, significantly higher than the GA-CNN model's 0.56 and FT Mask's 0.52. As the amount of training epochs increased, the synchronisation of PCC in each model's test set improved. In the 10th round, the model's PCC was 0.72, the GA-CNN model was 0.67, and the FT Mask method was 0.63. By the 20th round, the PCC of DTS-Score reached 0.79, while the GA-CNN and FT Mask methods were 0.72 and 0.70, respectively. The curve of the model designed by the research always maintained a steady upward trend with small fluctuations, demonstrating good stability and robustness. Using MAE, RMSE, and the coefficient of determination R^2 as metrics, it evaluated the model's predictive accuracy and fitting performance for both the overall score and individual dimension scores in the essay grading task. The findings are denoted in Table 3.

Table 3 Analysis of prediction accuracy and fitting effect

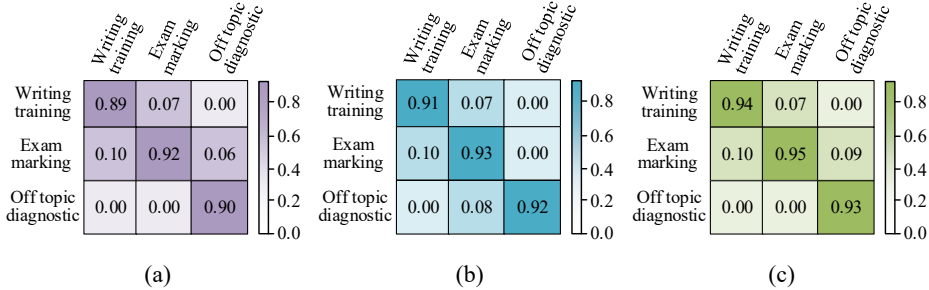
Dataset	Model	MAE	RMSE	R^2
Training set	GA-CNN	0.512	0.643	0.781
	FT Mask	0.436	0.575	0.832
	DTS-Score	0.392	0.521	0.883
Test set	GA-CNN	0.537	0.668	0.764
	FT Mask	0.458	0.593	0.815
	DTS-Score	0.408	0.544	0.866

From the findings in Table 3, DTS-Score performed better than the compared GA-CNN and FT Mask methods on both the training and testing sets. On the training set, the MAE and RMSE of the designed model were 0.392 and 0.521, respectively, which were reduced by 0.044 and 0.054 compared to FT Mask, indicating that it has more advantages in fitting accuracy and error control. Meanwhile, R^2 reached 0.883, significantly higher than FT Mask's 0.832 and GA-CNN's 0.781, indicating a stronger fit of the model to the scoring labels. On the test set, DTS-Score still maintained a leading position, with MAE and RMSE of 0.408 and 0.544 respectively, both lower than FT Mask's 0.458 and 0.593, and far better than GA-CNN's 0.537 and 0.668, demonstrating strong generalisation ability. In addition, its R^2 on the test set was 0.866, which was significantly better than FT Mask's 0.815 and GA-CNN's 0.764, further verifying the stability and reliability of the model in practical applications. Overall, the model designed by the research has achieved comprehensive improvement in error indicators and fitting ability, and has good application prospects.

3.2 *Evaluation of the effectiveness of the automatic scoring model for English compositions*

To comprehensively verify the application value of the proposed cross-prompt English composition automatic scoring model in practical educational evaluation scenarios, a scoring task simulation environment based on an intelligent writing assistance platform was studied. Three typical application scenarios, namely writing training, exam marking, and off topic detection, were selected for comparative analysis. The off topic detection rates of different models in three scenarios are shown in Figure 9.

Figure 9 Off topic detection accuracy (see online version for colours)



Figures 9(a), 9(b) and 9(c) respectively demonstrate the classification accuracy of the three models in three typical English writing application scenarios. In Figure 9(a), the classification accuracy of GA-CNN was 0.89 in the writing training scenario, 0.92 in the exam marking assistance, and 0.90 in the off topic detection, indicating its ability to recognise the main task to some extent. However, there was significant confusion at the category boundary, especially in the 10% cross judgement between writing training and exam marking, and it had some shortcomings in off topic detection. In Figure 9(b), the classification performance of the FT Mask model was slightly improved compared to GA-CNN in all scenarios, with an accuracy of 0.93 for exam marking, 0.92 for off topic detection, and 0.91 for writing training scenarios. The overall misjudgement rate decreased, especially in off topic tasks where misclassification for writing training has

been basically eliminated, demonstrating stronger boundary preservation ability. In Figure 9(c), the accuracy of DTS-Score in writing training, exam marking, and off topic detection tasks was 0.94, 0.95, and 0.93, respectively, which was significantly better than the comparison model. Additionally, the cross item misjudgement rate was extremely low, not exceeding 0.10. Especially in terms of off topic detection and task differentiation, the model achieved almost error free classification performance, indicating its more accurate scene perception and task differentiation ability, verifying its robustness and practicality in complex, multi-task automatic scoring scenarios. To further investigate the adaptability of the model to students with different writing levels in writing training scenarios, the composition samples were divided into three levels based on the artificial total score: high score (highest 30%), medium score (middle 40%), and low score (lowest 30%), and the PCC and MAE of the three models were calculated at each level. The experimental results are shown in Table 4.

Table 4 Consistency analysis for different score levels in the writing training scenario

<i>Score level</i>	<i>Model</i>	<i>PCC</i>	<i>MAE</i>
High	GA-CNN	0.72	0.45
	FT Mask	0.78	0.39
	DTS-Score	0.83	0.34
Medium	GA-CNN	0.69	0.53
	FT Mask	0.74	0.44
	DTS-Score	0.80	0.39
Low	GA-CNN	0.65	0.57
	FT Mask	0.72	0.54
	DTS-Score	0.76	0.45

According to Table 4, the linear consistency of the three models in high, medium, and low scoring compositions slightly decreased with decreasing writing proficiency, but overall remained in a relatively stable range. Among them, the PCC of the DTS core model reached 0.83, 0.80, and 0.76 at three levels, which were significantly higher than the GA-CNN's 0.72, 0.69, and 0.65, as well as the FT Mask's 0.78, 0.74, and 0.70. The corresponding MAE was always the lowest among the three, controlled between 0.34–0.45. In contrast, GA-CNN had an MAE of 0.57 in low scoring texts, indicating significant fluctuations in its scores on texts with weak language or unstable structure. Overall, the DTS Score model shows more robust consistency in scoring at different levels of writing proficiency. It can accurately characterise the structure and language advantages of high scoring essays, while maintaining good sorting and discrimination abilities for low to medium scoring essays. This is more conducive to supporting hierarchical feedback and personalised guidance in writing training scenarios. The precision-recall (PR) and F1-score results of different models in essay grading tasks are shown in Figure 10.

Figures 10(a) and 10(b) denote the PR curves and F1-score trends of the three models in the off topic detection task, respectively. In Figure 10(a), from the curve shape, the designed model maintained a high accuracy throughout the entire recall range, indicating its significant advantage in false alarm control. Especially in the section with a recall rate greater than 0.5, the accuracy of DTS-Score was much higher than that of FT Mask and

GA-CNN, indicating that it could effectively avoid false positives while detecting comprehensiveness. The FT Mask curve was relatively flat, located at the middle level, while GA-CNN performed weakly overall with a low curve, indicating significant shortcomings in balancing recall and accuracy. In Figure 10(b), the F1-score of the designed model remained around 92% throughout the entire iteration process, with minimal fluctuations, reflecting extremely high result stability and overall performance level. In contrast, the F1-score of the FT Mask model remained stable at 84%, although it had some competitiveness, there was still a significant gap compared to the model designed by the research. The F1-score of the GA-CNN model consistently hovered around 73%, significantly lower than the other two, and showed significant fluctuations in multiple training stages, indicating weak robustness during the iteration process. Based on the results of the two subgraphs, the designed model not only has better recognition accuracy and comprehensiveness in off topic detection tasks, but also exhibits stable and reliable training performance, significantly better than FT Mask and GA-CNN, making it a more suitable solution for practical deployment. In off topic detection tasks, it is particularly important to avoid misjudging essays that are highly related to the topic as off topic. Therefore, in addition to overall accuracy, it is also necessary to examine the balance between precision and recall. Therefore, based on the same test set as Figure 9, the precision, recall and F1-scores of three models in the off topic detection task were statistically analysed, and the results are shown in Table 5.

Figure 10 (a) PR curve and (b) F1-score results (see online version for colours)

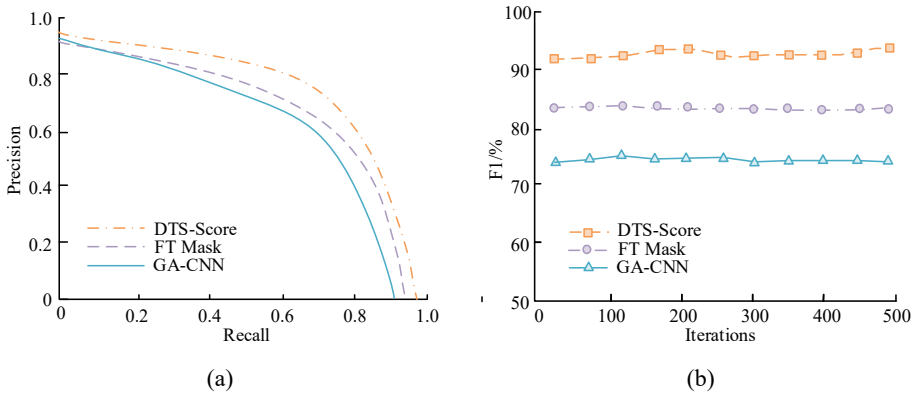


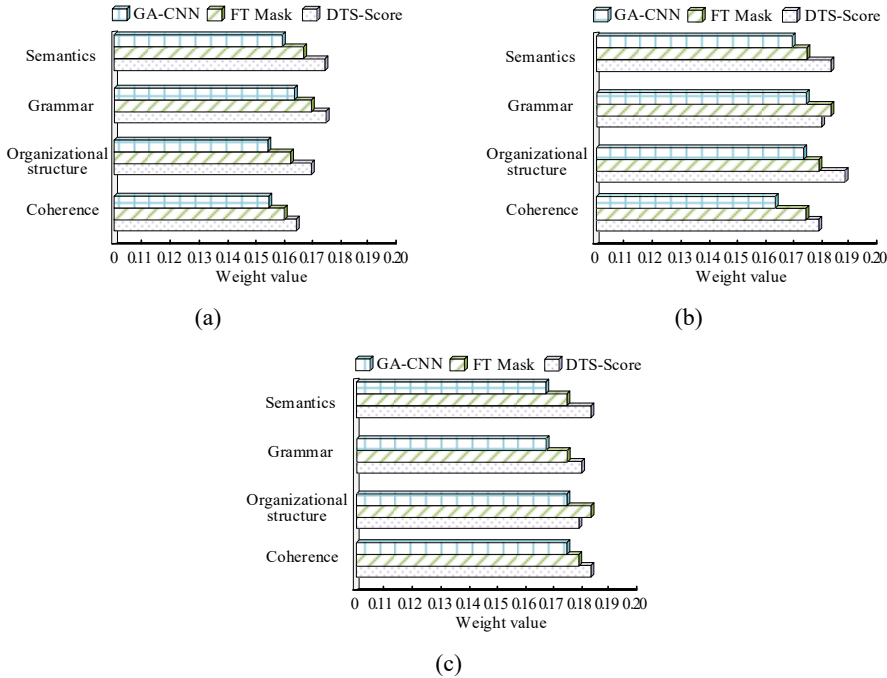
Table 5 Precision-recall comparison in the off-topic diagnostic scenario

Model	Precision	Recall	F1
GA-CNN	0.78	0.71	0.74
FT mask	0.87	0.82	0.84
DTS-Score	0.94	0.91	0.92

From Table 5, the precision and recall of GA-CNN in off topic detection tasks were 0.78 and 0.71, respectively, with an F1-score of 0.74, indicating a high risk of false positives and false negatives. FT Mask has improved in two indicators, with a precision of 0.87, a recall of 0.82, and an F1 score of 0.84. In contrast, the proposed model achieved the best results in both accuracy and recall, reaching 0.94 and 0.91 respectively, with a

corresponding F1-score of 0.92. This indicates that the model can effectively avoid misjudging normal essays that are closely related to the topic as off topic while identifying true off topic essays. It significantly reduces the risk of unfair scoring caused by misjudgement while ensuring detection coverage, making it more suitable for deployment in large-scale automatic grading and online learning platforms. The feature weights of semantic, syntactic, organisational structure, and coherence dimensions of different models in three application scenarios are shown in Figure 11.

Figure 11 Four dimensional feature weight results, (a) writing training (b) exam marking (c) off topic diagnostic (see online version for colours)



Figures 11(a), 11(b) and 11(c) respectively show the feature weight distribution of three models on four language dimensions in three typical application scenarios. In Figure 11(a), DTS-Score exhibited slightly higher feature weights in all dimensions, with semantic weights reaching 0.170, syntactic weights at 0.175, organisational structure weights at 0.166, and coherence weights at 0.152. In contrast, GA-CNN and FT Mask had slightly lower weights in the semantic and grammatical dimensions, with GA-CNN having a semantic weight of 0.162 and FT Mask having a weight of 0.168, indicating that DTS-Score focused more on deep understanding of content and sentence structure. This higher level of language attention helps the model accurately capture the expression details of students' compositions in writing training scenarios, thereby improving the accuracy of grading. In Figure 11(b), the feature weights of the designed model were higher than those of other models, with syntax and organisational structure weights of 0.178 and 0.173, respectively, significantly higher than GA-CNN and FT Mask. This indicated that the model proposed by the research placed more emphasis on syntactic structure and paragraph logic in grading, which met the grading requirements for

standardisation and organisation in exam scenarios. In addition, the designed model had a coherence dimension weight of 0.154, which was also better than other models, indicating that it paid higher attention to the fluency of the article in the scoring. In Figure 11(c), the semantic weight of DTS-Score was 0.198, which was the highest among all scenarios, while GA-CNN and FT Mask were 0.189 and 0.193, respectively. At the same time, in the two structural dimensions of organisational structure and coherence, the designed model had the lowest weights, which were 0.110 and 0.122 respectively, indicating that the model relied more on semantic matching rather than discourse structure in off topic detection. This difference is also in line with the characteristics of off topic detection tasks, which focus on determining whether the content of the article revolves around the topic, followed by the structure. The resource consumption results of different models are denoted in Table 6.

Table 6 Resource consumption results

<i>Scene category</i>	<i>Model</i>	<i>Inference time (ms)</i>	<i>Memory usage (MB)</i>	<i>Parameter quantity (M)</i>	<i>CPU utilisation rate (%)</i>
Writing training	GA-CNN	41.7	1,582	20.5	82.4
	FT Mask	35.4	1,420	18.2	76.3
	DTS-Score	28.9	1,345	16.7	71.2
Exam marking	GA-CNN	42.3	1,601	20.5	83.4
	FT Mask	36.8	1,443	18.2	78.1
	DTS-Score	30.2	1,358	16.7	72.5
Off topic detection	GA-CNN	44.1	1,625	20.5	84.7
	FT Mask	38.5	1,457	18.2	79.4
	DTS-Score	31.5	1,371	16.7	73.6

From Table 6, in all three scenarios, DTS-Score outperformed GA-CNN and FT Mask models in all resource consumption indicators, demonstrating higher computational efficiency and system friendliness. In terms of inference time, the designed model consistently maintained the lowest level, requiring only 28.9 ms in writing training tasks, which was 6.5 ms shorter than FT Mask and nearly 13 ms shorter than GA-CNN. Similar trends also appeared in exam marking and off topic detection, with DTS-Score completing tasks in 30.2 ms and 31.5 ms, respectively, demonstrating overall more efficient response capabilities. In terms of memory usage, the model proposed by the research also maintained a low memory requirement, with a maximum memory usage of 1,371 MB in the three scenarios, significantly lower than FT Mask and GA-CNN, demonstrating stronger adaptability, especially suitable for resource constrained deployment environments. In terms of model parameter quantity, DTS-Score parameter quantity was 16.7M, which was compressed compared to FT Mask and GA-CNN. This lightweight structure not only reduced the computational burden, but also provided convenience for subsequent model training and deployment. In addition, the CPU utilisation index further validated the system friendliness of DTS-Score. In all tasks, the CPU usage designed by the research was lower than the two comparison models, with the lowest value being 71.2% and the highest not exceeding 73.6%. In contrast, the CPU usage of GA-CNN remained above 82%, and FT Mask also exceeded 76%, indicating that the research proposed model was more energy-efficient and efficient in multi-threaded environments.

4 Discussion

Aiming at the shortcomings of automated composition scoring systems in semantic understanding and discourse logic modelling in the development of intelligent education, this study proposed an English composition intelligent scoring model based on transfer paragraph segmentation and Pearson detection. This model constructed a two-stage scoring network framework, with the first stage extracting irrelevant features and the second stage focusing on topic relevance judgement through semantic alignment mechanism. Integrating modules such as BERT, BiLSTM, attention mechanisms, and paragraph transfer segmentation, it achieved comprehensive representation of essays and scoring decisions across multiple semantic layers. The findings denoted that on the ASAP public dataset, the designed model outperformed the comparison model in mainstream scoring metrics such as QWK, PCC, MAE and RMSE. Among them, the QWK of the test set reached 0.866 and the PCC reached 0.79, indicating that the model had good rating consistency and linear correlation. In terms of error, the MAE and RMSE of the model on the test set were 0.408 and 0.544, respectively, which decreased by 0.067 and 0.079 compared to the baseline model, reflecting higher fitting accuracy and prediction stability. Compared with existing research, the designed method achieved significant improvements in both model structure design and feature construction. In reference Martín-Monje and Barcena (2024) introduced professional corpora and inference mechanisms into the G-Rubric system, attempting to replace some manual scoring processes with formal functional analysis and scoring feedback systems to promote practical exploration of human-machine collaboration paths. Although the system enhanced the educational adaptability of automatic feedback, the modelling of cross topic understanding and semantic transfer ability was still relatively limited. In addition, in reference (Nath et al., 2025), the multidimensional quantitative scoring system constructed by Zhang et al. used word frequency inverse document frequency technology as the core, combined with fuzzy clustering and rough set theory, to enhance the quantitative evaluation ability of essays from the perspectives of vocabulary distribution and text similarity, significantly improving the efficiency and consistency of traditional manual grading. However, this method still focused on shallow text features and lacked the ability to model the hierarchical semantic structure of the text. From the perspective of task scenario adaptability, DTS-Score performed stably in three typical applications: writing training, exam grading, and off topic detection. In terms of F1-score, the writing training scenario achieved 92.1%, significantly better than GA-CNN's 73.6%, indicating that the model could more effectively capture the overall quality and structural features of the composition in real scoring scenarios. At the same time, the PR curve showed that the model maintained high accuracy while improving recall, demonstrating good robustness and misjudgement control capability. Although this method has achieved good performance in multiple dimensions such as cross topic scoring, paragraph level discourse modelling, and off topic detection, there are still certain limitations. Firstly, the paragraph segmentation module relies on BERT's sentence level representation, which may lead to boundary blurring issues when dealing with extremely long texts, colloquial expressions, or non-standard writing, thereby affecting the stability of paragraph level semantic modelling. Secondly, the model currently mainly targets monolingual English composition scenarios, and its structural transfer ability may be limited in multilingual writing and discourse with significant cross-cultural expression differences. In addition,

although the pseudo data generation strategy enhances cross topic generalisation ability, it may still generate biased estimates in cases where the topic span is too large or the writing types are significantly different. In the future, further optimisation is needed in areas such as long text modelling, multilingual migration, interpretable mechanisms, and data quality control to enhance the robustness of methods in real educational scenarios.

5 Conclusions

The study designed an automatic scoring method for English compositions that integrates transfer learning, paragraph semantic modelling, and shallow feature enhancement, aiming to enhance the accuracy and adaptability of the model in multi-scenario and multi-dimensional scoring tasks. The findings denoted that the model achieved excellent performance on both the training and testing sets, especially in core indicators such as QWK and PCC, which always led the comparison methods. It also demonstrated good fitting and generalisation abilities in terms of model convergence speed and curve stability. Specifically, DTS-Score could achieve high accuracy in the initial iteration stage and maintain performance advantages in subsequent training, verifying the effectiveness and robustness of the method in scoring tasks. In the three typical application scenarios of writing training, exam marking, and off topic detection, the model demonstrated excellent task identification ability, indicating that the model could not only accurately identify various writing task features, but also had significant advantages in false alarm control. Feature attention analysis further revealed that the model maintained a balanced modelling of core language dimensions such as semantics, grammar, and structure in writing training and exam grading, while paying more attention to semantic information in off topic detection tasks, reflecting the model's adaptive modelling ability in different tasks. In addition, from the perspective of system resource consumption, DTS-Score significantly outperformed the comparative methods in response speed, video memory usage, model size, and CPU resource usage, demonstrating good deployment adaptability. Especially in application environments with limited computing power such as educational platforms and mobile devices, it had more practical value. By refining multidimensional feature attention and introducing attention visualisation mechanisms, the model not only enhanced the rationality of scoring, but also provided a data foundation and theoretical support for subsequent scoring interpretation and teaching interventions. In summary, the proposed scoring method achieves notable improvements in precision, efficiency, and generalisation. It provides a solid foundation for developing reliable and scalable intelligent writing evaluation systems. It also shows broad potential for practical application in real educational settings. From a broader perspective of intelligent education, the proposed cross topic composition automatic scoring framework not only improves the model's generalisation performance at the technical level, but also provides a feasible path for large-scale online writing evaluation. The transfer modelling and off topic detection mechanism based on paragraph discourse structure can support personalised writing feedback in practical teaching scenarios, such as automatically identifying weak links in article structure, detecting topic shifts, and generating targeted writing suggestions. Meanwhile, in large-scale exams or online classrooms, this model can serve as an auxiliary grading system, significantly reducing the burden of manual grading and improving grading consistency and efficiency. With the continuous popularisation of

intelligent education platforms, research methods have the potential to become a key component of writing teaching and evaluation systems, and lay the foundation for building a fairer, efficient and scalable automated writing evaluation ecosystem.

Declarations

The author declares no conflicts of interest.

References

- Alawi, A.B. and Bozkurt, F. (2025) 'Performance analysis of embedding methods for deep learning-based Turkish sentiment analysis models', *Arabian Journal for Science and Engineering*, Vol. 50, No. 10, pp.7299–7321, DOI: 10.1007/s13369-024-09360-4.
- Chabalala, S., Ojo, S.O. and Owolawi, P.A. (2025) 'Transfer learning for named entity recognition in Setswana language using CNN-BiLSTM model', *International Journal of Advanced Computer Science and Applications*, Vol. 16, No. 2, pp.472–481, DOI: 10.14569/IJACSA.2025.0160249.
- De Nardin, A., Zottin, S., Piciarelli, C., Foresti, G.L. and Colombil, E. (2024) 'In-domain versus out-of-domain transfer learning for document layout analysis', *International Journal on Document Analysis and Recognition*, Vol. 28, No. 2, pp.164–175, DOI: 10.1007/s10032-024-00497-4.
- Dornaika, F. and Hoang, V.T. (2023) 'Deep data representation with feature propagation for semi-supervised learning', *International Journal of Machine Learning and Cybernetics*, Vol. 14, No. 4, pp.1303–1316, DOI: 10.1007/s13042-022-01701-9.
- Dufera, A.G., Liu, T. and Xu, J. (2023) 'Regression models of Pearson correlation coefficient', *Statistical Theory and Related Fields*, Vol. 7, No. 2, pp.97–106, DOI: 10.1080/24754269.2023.2164970.
- Ganai, A.F. and Khursheed, F. (2023) 'Computationally efficient recognition of unconstrained handwritten Urdu script using BERT with vision transformers', *Neural Computing and Applications*, Vol. 35, No. 34, pp.24161–24177, DOI: 10.1007/s00521-023-08976-1.
- García-Méndez, S., de Arriba-Pérez, F., Barros-Vila, A., et al. (2023) 'Automatic detection of relevant information, predictions and forecasts in financial news through topic modelling with latent Dirichlet allocation', *Applied Intelligence*, Vol. 53, No. 16, pp.19610–19628, DOI: 10.1007/s10489-023-04452-4.
- Hu, L. (2024) 'Performance comparison of multiple scoring algorithms: a study of automatic scoring of English text', *Intelligent Decision Technologies*, Vol. 18, No. 1, pp.397–406, DOI: 10.3233/IDT-230305.
- Islam, M.S., Ghani, N.A., Zamli, K.Z., Hasan, M.M. and Lokman, A.S. (2024) 'KHACDD: a knowledge-based hybrid method for multilabel sentiment analysis on complex sentences using attentive capsule and dual structured recurrent network', *Neural Computing and Applications*, Vol. 36, No. 29, pp.18065–18086, DOI: 10.1007/s00521-024-09934-1.
- Jiang, J.H. (2025) 'Developing an artificial intelligence model for English grammar correction: a computational linguistics approach', *Journal of Computational Methods in Sciences and Engineering*, Vol. 25, No. 4, pp.2992–3006, DOI: 10.1177/14727978251318799.
- Jin, X., Bu, J., Yu, Z., Zhang, H. and Wang, Y. (2023) 'FedCrack: federated transfer learning with unsupervised representation for crack detection', *IEEE Transactions on Intelligent Transportation Systems*, Vol. 24, No. 10, pp.11171–11184, DOI: 10.1109/TITS.2023.3286439.

- Li, H., Peng, Q., Wang, X., Mou, X. and Wang, Y. (2023) 'Sehf: a summary-enhanced hierarchical framework for financial report sentiment analysis', *IEEE Transactions on Computational Social Systems*, Vol. 11, No. 3, pp.4087–4101, DOI: 10.1109/TCSS.2023.3323885.
- Li, M., Huang, R. and Qi, X. (2024) 'Chinese clinical named entity recognition using multi-feature fusion and multi-scale local context enhancement', *Computers, Materials & Continua*, Vol. 80, No. 2, pp.2283–2299, DOI: 10.32604/cmc.2024.053630.
- Liu, B., Yi, R., Yu, Z., Wang, S., Yang, X. and Wang, F. (2024) 'Semi-supervised instance segmentation algorithm based on transfer learning', *Nondestructive Testing and Evaluation*, Vol. 39, No. 1, pp.185–203, DOI: 10.1080/10589759.2023.2274013.
- Liu, H., Duan, X., Chen, H., Lou, H. and Deng, L. (2023) 'DBF-YOLO: UAV small targets detection based on shallow feature fusion', *IEEE Transactions on Electrical and Electronic Engineering*, Vol. 18, No. 4, pp.605–612, DOI: 10.1002/tee.23758.
- Ma, Y., Zou, X., Pan, Q., Yan, M. and Li, G. (2024) 'Target-embedding autoencoder with knowledge distillation for multi-label classification', *IEEE Transactions on Emerging Topics in Computational Intelligence*, Vol. 8, No. 3, pp.2506–2517, DOI: 10.1109/TETCI.2024.3372693.
- Madhubala, P., Ghanimi, H.M.A., Sengan, S. and Abhishek, K. (2024) 'RETRACTED: Bridging the gap in biomedical information retrieval: harnessing machine learning for enhanced search results and query semantics', *Journal of Intelligent & Fuzzy Systems*, Vol. 46, No. 4, pp.9491–9510, DOI: 10.3233/JIFS-237056.
- Martín-Monje, E. and Barcena, E. (2024) 'Tutor vs. automatic focused feedback and grading of student ESP compositions in an online learning environment', *Journal of Research in Applied Linguistics*, Vol. 15, No. 2, pp.22–42, DOI: 10.22055/rals.2024.45636.3198.
- Mezzadri, G., Laloë, T., Mathy, F., et al. (2024) 'How to fit transfer models to learning data: a segmentation/clustering approach', *Behavior Research Methods*, Vol. 56, No. 3, pp.2549–2568, DOI: 10.3758/s13428-023-02166-6.
- Nath, B., Sarkar, S. and Mukhopadhyay, S. (2025) 'AstraMT: instruction-tuned few-shot Assamese-English translation with context-aware prompting and reranking', *Computación y Sistemas*, Vol. 29, No. 3, pp.1167–1178, DOI: 10.13053/CyS-29-3-5886.
- Preethi, P. and Mamatha, H.R. (2023) 'Region-based convolutional neural network for segmenting text in epigraphical images', *Artificial Intelligence and Applications*, Vol. 1, No. 2, pp.119–127, DOI: 10.47852/bonviewAIA2202293.
- Purohit, J. and Dave, R. (2023) 'Leveraging deep learning techniques to obtain efficacious segmentation results', *Archives of Advanced Engineering Science*, Vol. 1, No. 1, pp.11–26, DOI: 10.47852/bonviewAAES32021220.
- Shen, Y. and Coker, D.L. (2023) 'Exploring reading-writing connections in native English speakers and English language learners', *Reading and Writing*, Vol. 36, No. 8, pp.2099–2117, DOI: 10.1007/s11145-022-10355-6.
- Sheridan, P. and Onsjö, M. (2024) 'The hypergeometric test performs comparably to TF-IDF on standard text analysis tasks', *Multimedia Tools and Applications*, Vol. 83, No. 10, pp.28875–28890, DOI: 10.1007/s11042-023-16615-z.
- Shou, Z., Liu, Y., Li, D., Mo, J. and Zhang, H. (2024) 'A bloom cognitive hierarchical classification model for Chinese exercises based on improved Chinese-RoBERTa-WWM and BiLSTM', *International Journal of Advanced Computer Science and Applications*, Vol. 15, No. 3, pp.479–489, DOI: 10.14569/IJACSA.2024.0150349.
- Singh, K., Tata, G., Oeveren, E.V. and Ray, N. (2025) 'Unpaired document image denoising for OCR using BiLSTM enhanced CycleGAN', *International Journal on Document Analysis and Recognition*, Vol. 28, No. 2, pp.207–224, DOI: 10.1007/s10032-024-00499-2.
- 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 Transactions on Learning Technologies*, Vol. 16, No. 6, pp.983–1000, DOI: 10.1109/TLT.2023.3253215.

- Wang, L. (2023) 'An evaluation method of English composition automatic grading based on genetic optimization algorithm and CNN model', *International Journal of Advanced Computer Science and Applications*, Vol. 14, No. 10, pp.56–66, DOI: 10.14569/IJACSA.2023.0141007.
- Wolters, A.P. and Kim, Y.G. (2024) 'Crosslinguistic influence on spelling in written compositions: evidence from English-Spanish dual language learners in primary grades', *Reading and Writing*, Vol. 37, No. 5, pp.1059–1078, DOI: 10.1007/s11145-023-10416-4.
- Zhu, G., Huang, X., Yang, R. and Sun, R. (2022) 'Relationship extraction method for urban rail transit operation emergencies records', *IEEE Transactions on Intelligent Vehicles*, Vol. 8, No. 1, pp.520–530, DOI: 10.1109/TIV.2022.3160502.