



International Journal of Information and Communication Technology

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

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

Intelligent scoring method for English articles leveraging large language models

Jing Yan

DOI: [10.1504/IJICT.2026.10076631](https://doi.org/10.1504/IJICT.2026.10076631)

Article History:

Received:	11 October 2025
Last revised:	24 November 2025
Accepted:	05 December 2025
Published online:	11 March 2026

Intelligent scoring method for English articles leveraging large language models

Jing Yan

School of Foreign Language,
Zhengzhou College of Finance and Economics,
Zhengzhou, 450000, China
Email: yan19900529@outlook.com

Abstract: Traditional manual scoring of English essays faces subjectivity and low efficiency, while existing automated essay scoring (AES) systems suffer from limited genre adaptability, weak scoring-feedback linkage, and poor robustness. To address these challenges, this study proposes a BERT and Zhipu AI model-based automatic scoring algorithm for different English genre essays (BERT-AI-ASEG), integrating ‘overall scoring feature scoring feedback output’. Furthermore, an English article intelligent scoring algorithm integrating multiple large language models (EAIS-IMLLM) is developed to enhance robustness. Experiments on the ASAP++ dataset show that the Spearman’s rank correlation coefficient and quadratic weighted kappa reach 0.87 and 0.85, improving by 4.82% and 6.25% over comparison algorithms. The proposed algorithm also achieves an evaluative feedback correlation of 0.89 and a multi-genre adaptation index of 0.91 at 150 iterations. These results indicate that the method is effective in improving accuracy, reliability, and teaching feedback effect for large-scale English writing assessments.

Keywords: large language model; English articles; scoring; BERT; article genre; Zhipu AI.

Reference to this paper should be made as follows: Yan, J. (2026) ‘Intelligent scoring method for English articles leveraging large language models’, *Int. J. Information and Communication Technology*, Vol. 27, No. 18, pp.1–25.

Biographical notes: Jing Yan has a Bachelor of Arts degree and is an Associate Professor. He graduated from Henan College of Education in 2003. He worked in Zhengzhou College of Finance and Economics. His research interests English teaching methodology.

1 Background

English writing ability is a core skill in second language acquisition and international communication, and its evaluation quality directly determines the accuracy of language teaching evaluative feedback and the efficiency of learners’ ability improvement (Susanti et al., 2023; Beseiso et al., 2021). In large-scale educational scenarios such as standardised language exams, online course homework scoring, and cross-border

education assessments, the traditional manual scoring model faces significant bottlenecks. Firstly, manual scoring relies on the experience of reviewers and is easily influenced by subjective preferences, fatigue levels, and other factors, leading to difficulty in ensuring scoring consistency (such as differences in scoring the same article by different reviewers). In addition, faced with the demand for batch scoring of thousands or even tens of thousands of essays, manual scoring is time-consuming and labour-intensive, and cannot meet the needs of real-time evaluative feedback and large-scale teaching (Ramnarain-Seetohul et al., 2022; Uto, 2021; Misgna et al., 2024). In this context, automated essay scoring (AES) technology has become a research focus at the intersection of educational technology and natural language processing. It simulates the logic of manual scoring through algorithms, aiming to achieve high efficiency, objectivity, and scalability of scoring (ElMassry et al., 2025; Mizumoto and Eguchi, 2023).

Early AES technology relied on manually designed features such as vocabulary density and sentence complexity. These features are combined with traditional machine learning models such as SVM and random forest for scoring. Chassab et al. (2021) solved the learning paradigm feature analysis genres and their sub components by introducing a new classification method. The results showed that using semantic knowledge sources such as ontologies would be limited to essay answers (Chassab et al., 2021). To address the time-consuming and subjective of evaluating scholarship applicants' essays, Susilowati and Acar (2023) proposed an AES system using SVM to predict successful candidates for youth breakthrough scholarships. This method reduced the burden of manual evaluation, ensured consistency and objectivity of scoring, and accelerated the selection process (Susilowati and Acar, 2023). However, such methods are difficult to capture the deep semantic associations and logical structure of the text, and have poor adaptability to articles of different themes and genres. With the development of deep learning technology, models based on recurrent neural network (RNN) can enhance their ability to model text sequences. Park et al. (2021) used an automatic essay scoring model based on long short-term memory (LSTM) to evaluate English essay data extracted from an open large-scale data warehouse. The LSTM could predict students' essay scores at an appropriate level (Park et al., 2021). To fully meet the needs of essay evaluation, Ramesh and Sanampudi (2022b) used word embedding features and LSTM to evaluate essays, and adopted K-fold cross-validation, achieving an accuracy of 85.35%. However, this method is still limited by one-way semantic encoding and cannot fully understand contextual information. Until the emergence of bidirectional encoder representations from transformers (BERT) models, their bidirectional semantic capture capability and rich pre-trained semantic representations provide key support for the accuracy breakthrough of AES technology, driving AES to transform from 'feature driven' to 'semantic driven'. For example, the research on Arabic AES is limited by the lack of publicly available essay data. Ghazawi and Simpson (2024) proposed the AR-AES benchmark dataset. This dataset contained 2,046 undergraduate papers, covering gender, scores, and transparent scoring criteria evaluation guidelines, providing in-depth insights into the scoring process. The error of the AES based on BERT was relatively small, with 96.15% of errors differing by less than 1 point from the prediction of the first human rater (Ghazawi and Simpson, 2024). Faced with the problem that neural network automatic

scoring methods require much annotated data, Xue et al. (2021) developed a novel method that used multi-task learning combined with fine-tuning from BERT for multidimensional AES tasks. As a result, it was found that the model increased the average QWK by 8.1% (Xue et al., 2021).

In summary, although existing AES methods have made some progress, there are still many core issues that constrain their application value. Firstly, there is insufficient adaptability to different genres, as most algorithms do not take into account the differences in writing characteristics of different genres and use a unified standard for scoring, lacking specificity and accuracy. Secondly, there is a disconnect between scoring and evaluative feedback. Most systems only output scores without incorporating specific improvement suggestions, making it difficult for learners to identify their weaknesses and improve their writing skills in a targeted manner. In addition, the AES system dominated by a single model is vulnerable to the influence of its own design tendency, with poor robustness, and is difficult to adapt to various article genres and scoring tasks. In response to the above issues, the research innovatively focuses on the application of large language models in AES and proposes an intelligent scoring method for English articles. In this method, firstly, a BERT and Zhipu AI model-based automatic scoring algorithm for different English genre essays (BERT-AI-ASEG) is designed. The algorithm integrates overall article scoring, feature scoring, and evaluative feedback, captures deep semantics through BERT, and generates specific feedback through Zhipu AI model, integrating 'scoring – evaluative feedback'. Secondly, to improve the robustness of the algorithm, an English Article Intelligent scoring algorithm integrating multiple large language model (EAIS-IMLLM) is proposed, which integrates the prediction results of BERT-AI-AEG, improved BERT, and recursive neural network-based automated essay scoring model (RNN-AESM) to reduce the dependence on a single model. To address the 'insufficient genre adaptability' issue in the introduction, the BERT-AI-ASEG algorithm introduces genre label embeddings into the feature scoring module, enabling the model to automatically recognise differences between genres like argumentative and narrative texts while adjusting scoring weights accordingly. For the 'disconnection between scoring and evaluative feedback' problem, an independent evaluative feedback output module is designed. Based on the ZhiPu AI model, this module transforms feature scores into actionable improvement suggestions, achieving a closed-loop 'scoring-evaluative feedback' system. To tackle the 'poor robustness', the EAIS-IMLLM algorithm integrates predictions from three heterogeneous models, reducing domain-specific biases of individual models and enhancing adaptability to articles of different themes and styles. Each module design directly addresses core issues, forming a clear 'problem-solution' logical chain.

This study makes three major theoretical contributions:

- 1 A multi-feature joint calibration mechanism to enhance scoring stability is proposed, which integrates the deep semantic features of BERT and the temporal features of LSTM to solve the limitations of traditional algorithms in capturing advanced writing features.
- 2 A genre-adaptive mechanism is implemented for the first time, overcoming the constraints of traditional single-genre adaptation through genre label embedding and meta-model integration strategies.

- 3 An engineering solution for a closed-loop scoring evaluative feedback system is developed, which resolves the core pain point of traditional AES systems that only output scores without effective feedback through the ‘feature scoring-natural language feedback-feedback adoption evaluation’, providing a referenceable engineering paradigm for the practical application of AES technology in education.

2 Methods and materials

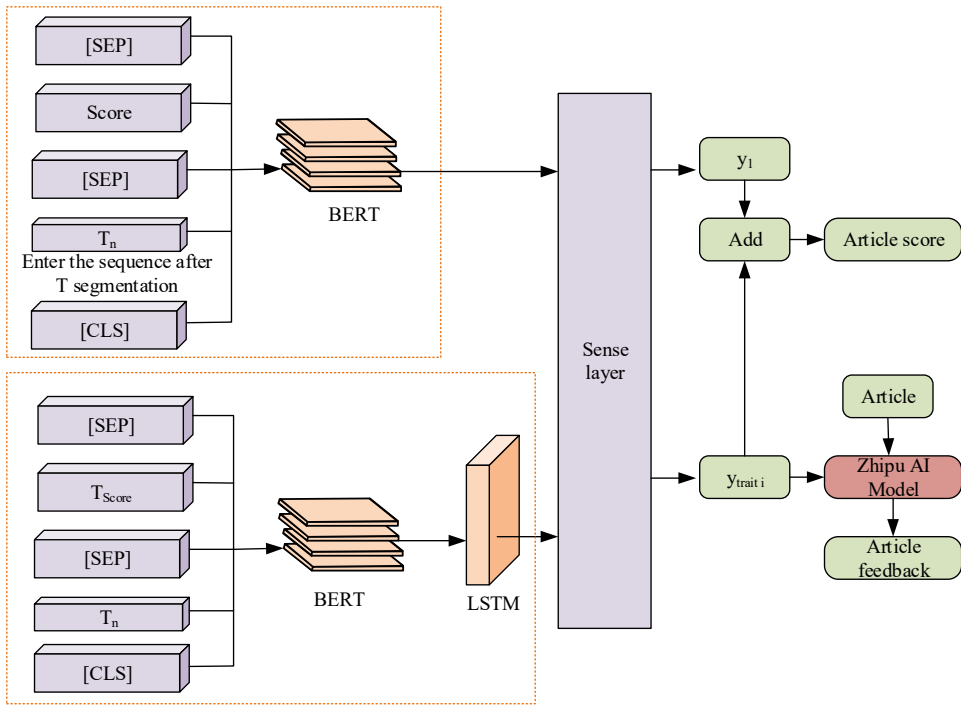
2.1 *Design of automatic scoring algorithm for articles of different English genres*

There are two core issues with the current automatic scoring method for English articles. Firstly, the differences in the characteristics of different genres of articles have not been fully considered, and ignoring these differences directly leads to a decrease in the accuracy of scoring, making it difficult to objectively reflect the quality of writing (Ramesh and Sanampudi, 2022a; Lim et al., 2021; Shin and Gierl, 2021). Secondly, most scoring algorithms only output score results and do not combine scoring with specific evaluative feedback, resulting in students being unable to identify the strengths, weaknesses, and problems of their own writing, making it difficult for them to obtain accurate guidance and improvement directions (Kumar and Boulanger, 2021; Li et al., 2025; Zhao, 2021; Ferrara and Qunbar, 2022). The research innovatively introduces BERT and Zhipu AI models from the large language model, and proposes an automatic article scoring algorithm that considers different English genres, namely the BERT-AI-ASEG algorithm. This algorithm organically integrates scoring and evaluative feedback, while focusing on the overall quality of English articles. It also fully considers the unique characteristics of different genres of articles. This design not only effectively improves the accuracy of article evaluation results, but also helps students to have a more comprehensive understanding of the evaluation content. The study selects BERT and Zhipu AI’s glm-3-turbo as core models for AES task performance evaluation. The comparison highlights three key advantages:

- 1 BERT’s bidirectional Transformer architecture outperforms GPT series models in text comprehension, delivering superior semantic feature extraction accuracy. While GPT excels in generative capabilities, it shows lower scoring consistency and higher inference costs.
- 2 Compared with T5, BERT demonstrates more stable fine-tuning results on medium-sized datasets, requiring no complex task reconfiguration and offering lower deployment costs.
- 3 Zhipu AI’s glm-3-turbo supports long-text processing, comprehensively understands full essay content, generates targeted evaluative feedback, and features user-friendly API integration.

Considering overall performance, cost-effectiveness, and deployment feasibility, the BERT-Zhipu AI model combination proves optimal for AES tasks.

The overall architecture of the BERT-AI-ASEG algorithm is shown in Figure 1.

Figure 1 Schematic diagram of the overall architecture of BERT-AI-ASEG algorithm (see online version for colours)

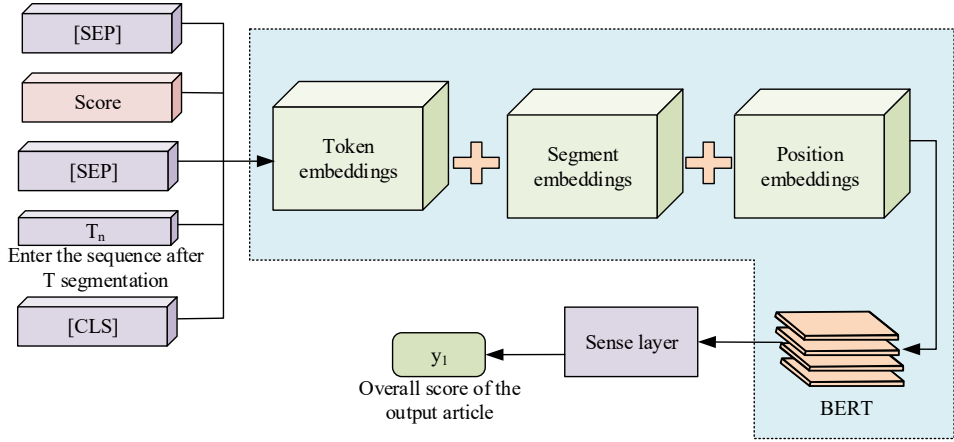
In the BERT-AI-ASEG algorithm shown in Figure 1, the BERT model version used is BERT-base-uncased, and the version of the Zhupu AI model is glm-3-turbo. The BERT-AI-ASEG algorithm undergoes a preprocessing workflow comprising:

- 1 Text cleaning: Eliminating HTML tags, special characters, and redundant symbols while correcting spelling errors.
- 2 Format standardisation: Converting all text to UTF-8 encoding, standardising capitalisation, and unifying date and numeric formats.
- 3 Tokenisation and stopword removal: Using BERT's native WordPiece segmentation to split text, supplemented by the NLTK English stopword list to eliminate redundant terms.
- 4 Length normalisation: Trimming texts exceeding 512 tokens and padding with [PAD] symbols for shorter lengths.
- 5 Feature alignment: Normalising feature scores across datasets to the [0.5] range to ensure cross-set comparability.

Then, the preprocessed information is separately imported into the BERT pre-training model, and BERT's powerful semantic understanding ability is utilised to encode the deep semantic features and score related information of the text. Subsequently, the [CLS] marker vector output by the BERT model (which condenses the global semantic information of the entire English article) is extracted and input into a fully connected

layer network that has been optimised through training data. Through nonlinear transformation and feature integration of the fully connected layer, the overall prediction score for the article is finally output. Then, all the outputs of BERT are taken into the LSTM and the fully connected layer one by one to obtain the predicted scores for each feature. Finally, the overall score and the predicted scores of each feature are combined to form the total score. The content and feature scores of the article are input into the Zhipu AI model, which outputs evaluative feedback for each feature. The architecture of the overall scoring module for the BERT-AI-ASEG algorithm is shown in Figure 2.

Figure 2 Schematic diagram of the overall scoring module structure for articles (see online version for colours)



In Figure 2, the module first selects 4,000 commonly used words as a vocabulary list, and uses a BERT tokenizer to segment the text T of length n (splitting words and prefixes and suffixes), with corresponding scores attached to the end of the text. When segmenting words, $[CLS]$ is added before T and $[SEP]$ is added between sentences, and then the complement $[PAD]$ or truncation is limited according to the BERT length to obtain the sequence T_1 , as shown in equation (1).

$$T_1 = \begin{cases} [CLS] + T_L + [SEP] & n > L \\ [CLS] + T + [SEP] & n = L \\ [CLS] + T + [PAD] * (L - l) + [SEP] & n < L \end{cases} \quad (1)$$

In equation (1), T represents the complete token sequence formed by the word segmentation operation of the article. T_L represents the sub-sequence obtained by extracting some content from T according to the requirements. l represents the length parameter corresponding to the original input text. L represents the length of the sequence T_L obtained after the truncation operation. Next, the obtained sequence is input into word embedding, segment embedding, and position embedding. The results of the three embedding layers are added together to obtain the input vector T_2 for the BERT pre-trained model, as shown in equation (2).

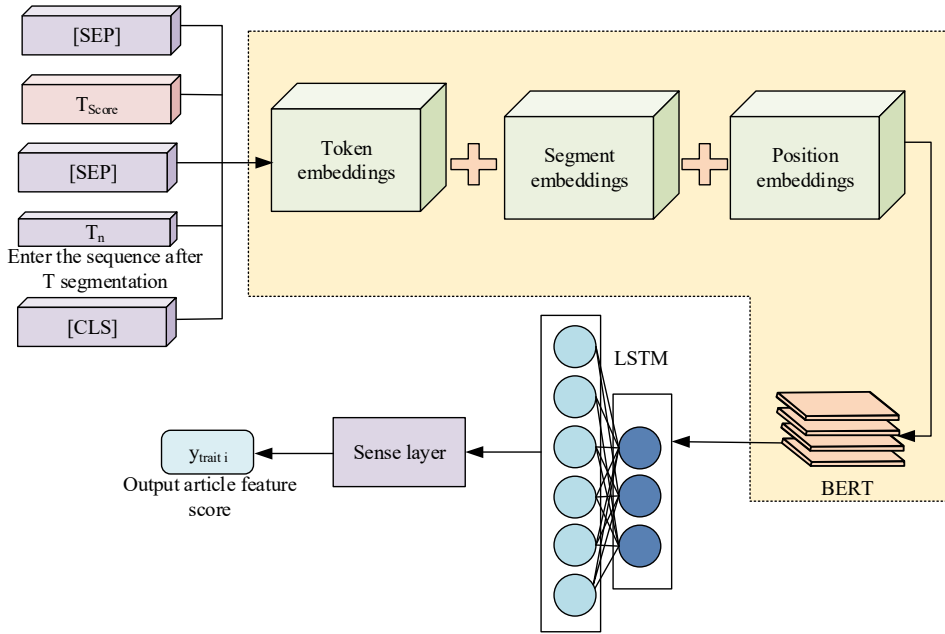
$$T_2 = f_{TE}(T_1) + f_{SE}(T_1) + f_{PE}(T_1) \quad (2)$$

In equation (2), f_{TE} , f_{SE} , and f_{PE} respectively represent word embedding operation, segment embedding operation, and position embedding operation. Finally, the output vector T_3 of the BERT pre-trained model is input into the fully connected layer to obtain the predicted score y_1 at the overall level of the text. The output result generated by the position of the delimiter $[CLS]$ is intentionally added to the input sequence of the BERT, which is represented by $[CLS]_{out}$, as shown in equation (3).

$$y_1 = \hat{W}_{text} \cdot [CLS]_{out} + b_{text} \quad (3)$$

In equation (3), $\hat{W}_{text}$ and b_{text} represent the weight matrix and bias of the overall article scoring, respectively. The BERT model generates rich and hierarchical semantic representations through large-scale corpus learning during the pre-training stage. Its core advantage lies in relying on bidirectional Transformer encoders, which can deeply capture contextual information and dependencies between contexts of text, thus flexibly adapting to different fields and styles of articles, and achieving accurate understanding of overall semantics. The $[CLS]$ -tag vector output by the model contains global semantic information, which can be connected to the fully connected layer to complete downstream tasks such as scoring. The overall structure of the article feature scoring module for the BERT-AI-ASEG algorithm is shown in Figure 3.

Figure 3 The structure of the article feature scoring module (see online version for colours)



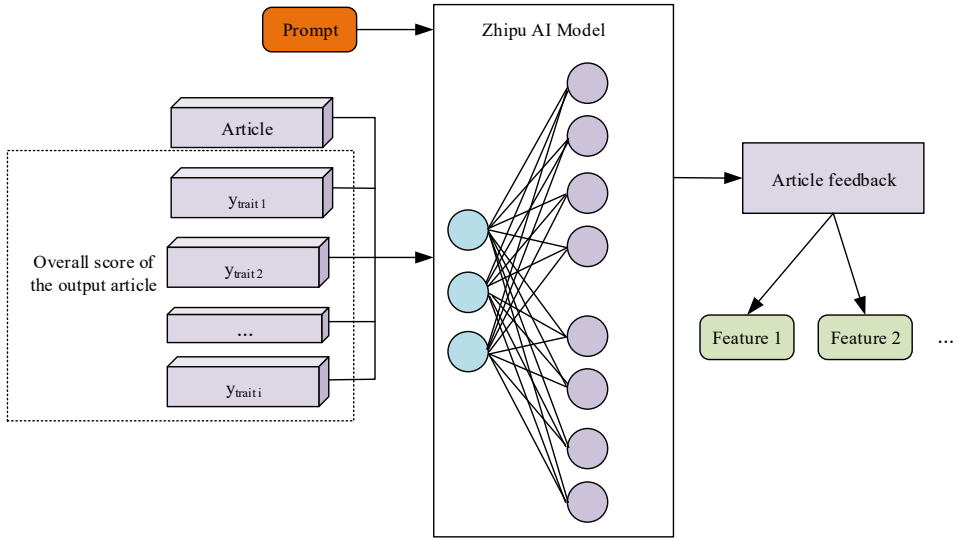
In Figure 3, the core architecture of this module has BERT, LSTM, and fully connected layers working together. The operation process is as follows. Firstly, the original text is preprocessed with corresponding feature scores, followed by word segmentation and vectorisation operations. Then, it is input into the BERT model for semantic encoding, generating vector representations containing rich contextual information. Subsequently, the vector is input into an LSTM network to further extract deep features through its

temporal learning ability, forming a more discriminative feature representation vector $Trait_i$. Finally, the feature representation vector is input into a fully connected layer. After nonlinear transformation, it outputs prediction scores $y_{trait\ i}$ for different features, completing the end-to-end processing flow from text input to feature prediction, as shown in equation (4).

$$\begin{cases} Trait_i = LSTM(T_3) \\ y_{trait\ i} = \hat{W}_{trait} \cdot Trait_i + b_{trait} \end{cases} \quad (4)$$

In equation (4), i represents the i^{th} article feature number. b_{trait} and $\hat{W}_{trait}$ represent the bias and weight of the article feature scoring, respectively. After obtaining the overall prediction score generated based on the global semantics of the article, as well as the feature prediction scores for each dimension attribute of the text, this module uses a direct superposition and summation method to fuse the two genres of scores and calculate the final evaluation score. The overall structure of the article evaluative feedback output module for the BERT-AI-ASEG algorithm is shown in Figure 4.

Figure 4 The structure of the article evaluative feedback output module (see online version for colours)

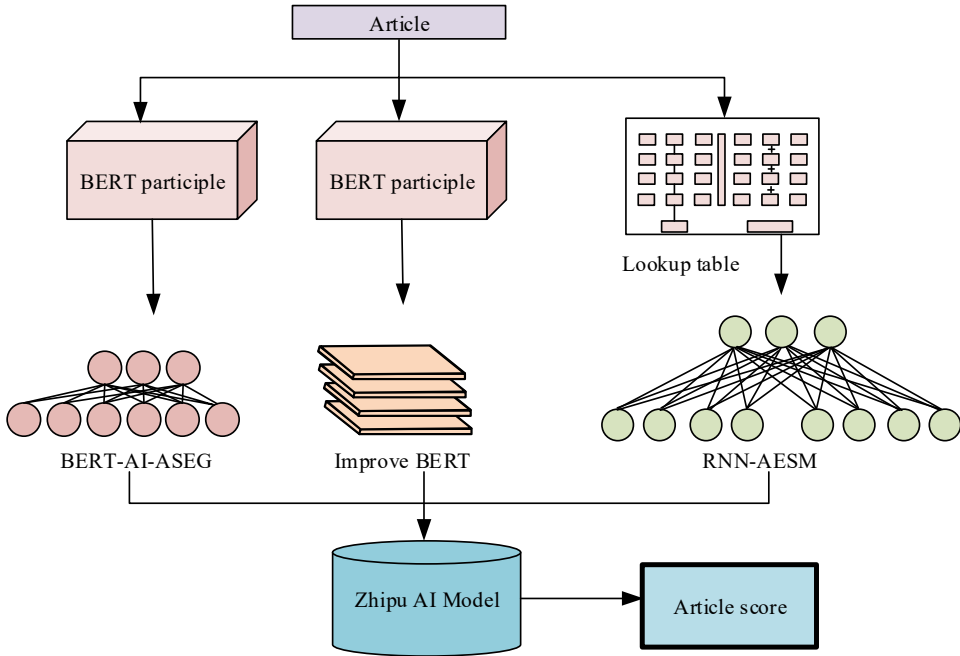


As shown in Figure 4, the module is built around the Zhipu AI model as its core framework. During operation, after generating feature prediction scores, the system first indexes each article using its unique ID from the dataset. It then stores essential information including the original text, corresponding prediction scores, and the article category. Subsequently, the Zhipu AI API is invoked to feed this data into the large language model, which excels in text processing capabilities.

2.2 Design of intelligent scoring algorithm for English articles integrating multiple language models

After designing the BERT-AI-ASEG algorithm, a EAIS-IMLLM algorithm is proposed to solve the single structure and low maintainability in the later stage. This algorithm integrates the prediction results of multiple language models, reduces excessive dependence on a specific subjective tendency, and overall improves the robustness of the algorithm to adapt to various tasks and articles. The overall architecture of EAIS-IMLLM algorithm is shown in Figure 5.

Figure 5 Schematic diagram of the overall architecture of EAIS-IMLLM algorithm (see online version for colours)



In Figure 5, the EAIS-IMLLM algorithm mainly includes three main modules: data input module, basic model module, and meta-model module. Specifically, the article and article scores are processed through a data processing layer and input into the basic model module, which includes three basic language models: improved BERT, BERT-AI-ASEG, and RNN-AESM. The basic model module mainly implements English article score prediction. After the basic model module outputs their respective predicted scores, all results will be summarised in the meta-model module. This module is built based on the Zhipu AI large language model, relying on its powerful semantic understanding and data integration capabilities. It comprehensively considers the rationality and reliability of the predicted scores of various basic models, uses models for deep fusion, and finally generates accurate and persuasive article final scoring results, while outputting detailed evaluative feedback.

Both the BERT-AI-ASEG in the basic model module and the improved BERT model use BERT’s native tokeniser and embedding layer, which not only effectively ensures the stability of the model’s basic performance, but also controls the consistency of the core preprocessing process. The basic tokenise model performs initial segmentation based on spaces and punctuation, and adds special symbols. The wordpiece tokenise is based on the byte pair encoding (BPE) algorithm, which uses greedy longest priority matching to split words from left to right and search for IDs in the vocabulary (the size of the vocabulary is a hyperparameter, and larger vocabulary requires more resources). After segmentation, it performs three-layer embedding (Vijaya Shetty et al., 2022; Gaheen et al., 2021; Lotfy et al., 2023). The RNN-AESM model uses a lookup table to process word sequences using one-hot encoding (sparse vectors with only one element being 1), which is suitable for large vocabulary lists (Conijn et al., 2023). The output for the lookup table is shown in equation (5).

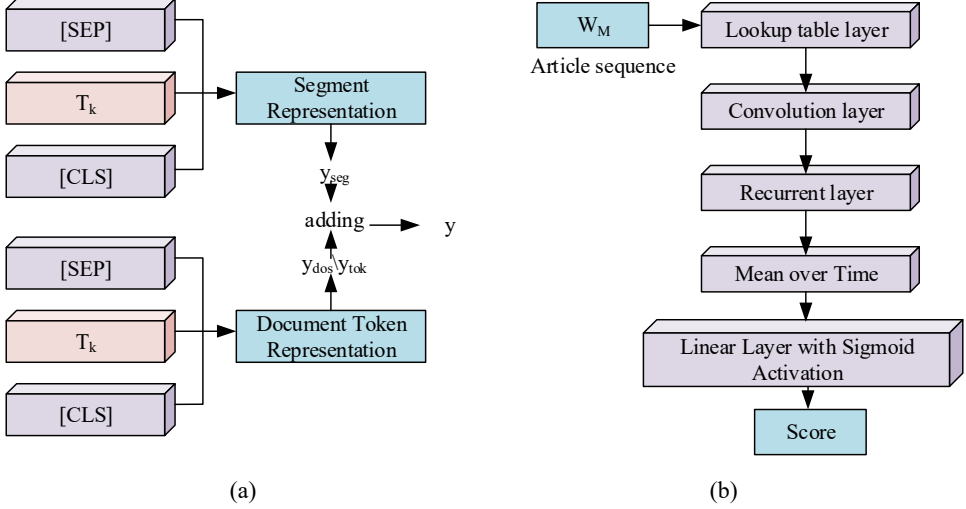
$$LT = (E \cdot w_1, E \cdot w_2, \dots, E \cdot w_M) \quad (5)$$

In equation (5), M signifies the sequence length. w signifies the sequence of words represented by one-hot. E represents the word embedding matrix. The basic model module of the BERT-AI-ASEG algorithm consists of three major language models: BERT-AI-ASEG, improved BERT, and RNN-AESM to predict English article scores. The core rationale for integrating BERT-AI-ASEG, enhanced BERT, and RNN-AESM stems from their complementary strengths in evaluation tasks: The BERT-AI-ASEG excels at deep semantic capture and evaluative feedback generation, addressing the ‘evaluative feedback disconnect’ issue. The enhanced BERT demonstrates exceptional accuracy in feature scoring for specific genres (e.g., argumentative essays), providing a highly reliable benchmark. The RNN-AESM performs well in temporal feature modelling, compensating for the limitations of Transformer models in capturing local logical relationships in long texts. The fusion strategy employs a ‘meta-learning weighted fusion’ mechanism, training a dynamic weight allocation model using validation set data. This approach is theoretically grounded in the ‘bias-variance trade-off theory for multi-model integration’, which reduces the overall system variance by combining predictions from models with different bias directions. The improved BERT and RNN-AESM model architectures are shown in Figure 6.

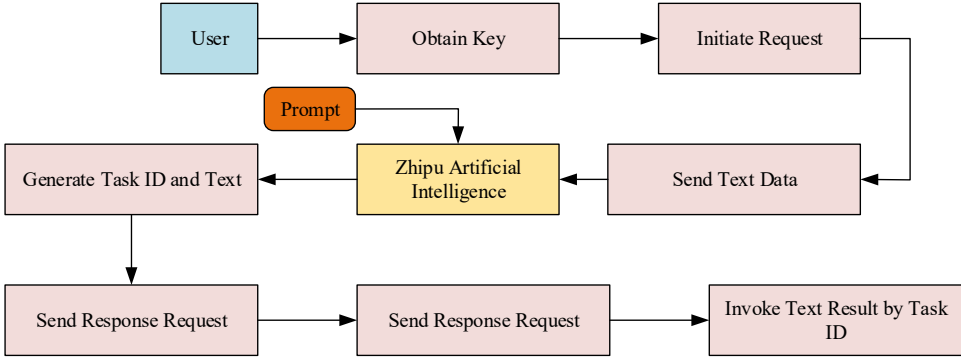
In Figure 6(a), the improved BERT first utilises BERT to extract document and vocabulary level features, and predicts corresponding scores through pooling and fully connected layers. The BERT is utilised to extract multi-scale paragraph features and predict scores, while determining the optimal number of segments. The final prediction score is obtained by summarising the scores of the three categories, and the model also uses a regression loss function when optimising without auxiliary tasks. In Figure 6(b), RNN-AESM utilises the sequence modelling capability of RNN to automatically learn the text representations required for automatic article scoring from the input text. The automatic scoring model takes the article sequence as input. Firstly, it is processed by searching the surface layer. The one-hot encoding is taken to convert the text into a numerical vector. Subsequently, the convolutional neural network extracts local key features from the text, and recurrent layers are utilised to capture the inherent temporal

dependencies of the text (Bui and Barrot, 2025; Xu et al., 2024). Afterwards, dynamic features are transformed into fixed-dimensional vectors through time mean aggregation operation. Finally, a linear layer equipped with a Sigmoid activation function maps the feature vectors to scalar scores representing the scoring results, and each step progresses in an orderly manner to achieve automatic scoring.

Figure 6 Schematic diagram of improved (a) BERT and (b) RNN-AESM (see online version for colours)



In the architecture design of EIAS-IMLLM algorithm, the meta-model module plays a key optimisation role, and the glm-3-turbo large language model is one of the core components that make up this module. In the actual operation phase, to achieve precise model iteration and performance improvement, the meta-model module does not simply receive text data, but integrates the predicted scores of the basic model for text output with the original article text as input data. By introducing a new dataset, the meta-model can learn potential data patterns that the basic model has not yet captured, making up for its shortcomings in feature extraction. This module also deeply integrates the inherent advantages of the basic model, enhances the model's generalisation and adaptation ability and prediction accuracy, and ultimately constructs a more powerful model (Guo et al., 2023; Preethi and Mamatha, 2023). The EIAS-IMLLM employs a 'threshold judgment-semantic re-evaluation' conflict arbitration mechanism: When the prediction score difference among three base models exceeds 10% (i.e., the ratio of the highest to lowest score >1.1), the conflict arbitration process is triggered. First, the meta-model (glm-3-turbo) re-reads the original article text to analyse the divergence points. Next, it re-evaluates the divergent features by incorporating the genre-specific scoring criteria of the article. Finally, it dynamically adjusts the weights of each base model based on the evaluation results to generate the final score. The glm-3-turbo large language model can achieve complementary advantages and disadvantages of BERT and GPT, and can continue writing fill in the blank while understanding, but asynchronous API calls require waiting for 10–30 seconds for results to be generated. The API calling process is shown in Figure 7.

Figure 7 Schematic diagram of API calling process (see online version for colours)

In Figure 7, the glm-3-turbo large language model can intelligently process the scoring requirements of different genres of articles in the ASAP++ dataset, analyse the properties of the dataset with strong learning ability, adjust the scoring strategy according to the characteristics of the article and task requirements, and also conveniently adjust the task objectives by adjusting the input Prompt. Ultimately, the meta-model outputs a comprehensive score that balances the predicted results of the basic model with a deep understanding of the content of the article, providing a comprehensive and accurate evaluation. Simultaneously, the evaluative feedback is output, reflecting the strengths and weaknesses of the article.

3 Results

3.1 Dataset processing and algorithm training

To verify the effectiveness of the proposed intelligent scoring method for English articles, specific datasets are used to test different algorithms. Among them, the ERT-AI-ASEG algorithm is based on the open-source Automated Student Assessment Prize (ASAP) automatic paper scoring dataset, which includes eight independent article collections and nearly 13,000 articles written by middle school students according to specific prompts. The genres cover three common genres: argumentative writing, question answering, and narrative writing, which can better reflect the true level and diverse styles of students' writing. In terms of scoring dimensions, the original ASAP dataset provides both overall and feature scores for articles, but has obvious limitations, namely only including feature scores in the 7th and 8th sets, and lacking detailed feature evaluation data in the remaining sets. To overcome this limitation and meet the demand for analysing the characteristics of different genres of articles in automatic scoring tasks, the ASAP++ dataset is used in the study. This dataset provides scoring for 4–6 different article features for each article set, with sets 1 and 2 containing 5 features, sets 3–7 containing 4 features, and set 8 containing 6 features. The EAIS-IMLLM algorithm takes the ASAP++ dataset as input and derives a new dataset. The input of the meta-model module is generated by combining the predicted scores of three basic models with the original article. It is worth noting that the BERT-AI-ASEG algorithm does not require additional feature scoring dimensions, so the newly derived dataset ultimately includes

four core attributes: article ID, article collection, article content, and prediction score. The predicted score is accurately matched and associated with the corresponding original article through a unique article ID, ensuring data consistency and correlation. In addition, a key constraint needs to be emphasised. The improved BERT training model only conducts specialised training on the 8th set of articles in the ASAP++dataset during the training phase, and its model performance and predictive ability are optimised based on this set. Therefore, in practical applications, the prediction results of this model only use the article data in the 8th set and cannot be used across sets. This setting is crucial to ensure the accuracy of the prediction results. The enhanced BERT is only used for training on the 8th dataset of ASAP++and has limited impact on the overall performance of EAIS-IMLLM. This is because the meta-model module of EAIS-IMLLM dynamically corrects domain-specific biases in a single model by integrating predictions from three heterogeneous models. To improve cross-genre generalisation, the study introduces genre label embeddings during meta-model training, enabling the model to adaptively adjust evaluation criteria. This design maintains precision in specific dataset evaluations while preserving overall generalisation capabilities.

The parameters used by each model and the experimental operating environment are presented in Table 1.

The comparison models selected for the BERT-AI-ASEG algorithm include GPT-4-based automated essay scoring (GPT-4-AES), multi-task transformer scoring Model (MTTS), and BERT-based automated essay scoring (BERT-AES). The EAIS-IMLLM algorithm selects BERT-AI-ASEG, improved BERT, RNN-AESM, and meta-learning-based AES (Meta-AES). All selected comparison models were maintained under comparable conditions to the study:

- 1 GPT-4-AES: Using the basic version of GPT-4, the training data is consistent with the research (80% of the samples come from the ASAP++dataset), without introducing external corpora.
- 2 MTTS: Employing the model weights published in the original article and readjust them on the same dataset.
- 3 BERT-AES: The same basic model as BERT-AI-ASEG was used, with only the feature scoring and evaluation feedback modules removed, while maintaining the same training parameters.
- 4 Meta-AES: Implementing the open-source code provided in the original article, configuring the meta-learner as a 3-layer fully connected network, with training iterations standardised to 150 as in this study.

All comparison models were executed under identical hardware environments to ensure experimental fairness.

The evaluation metrics for the BERT-AI-ASEG algorithm include Spearman's rank correlation coefficient (SCC), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and quadratic weighted kappa (QWK). Firstly, the BERT-AI-ASEG and EAIS-IMLLM algorithms are trained on the training set, as shown in Figure 8.

Figures 8(a) and 8(b) show the fitness iteration curves on the training set, respectively. From Figure 8(a), the BERT-AI-ASEG algorithm had the best iterative performance on the training set. On the training set, the BERT-AI-ASEG algorithm only

had 229 iterations to reach a stable state, with a fitness value of 0.16. From Figure 8(b), the EAIS-IMLLM algorithm reached a stable state after 181 iterations, and the fitness value in the stable state was 0.14. The training results indicate that the BERT-AI-ASEG and EAIS-IMLLM algorithms are effective and can be applied to intelligent scoring tasks for English articles. Based on five independent repeated experiments, the BERT-AI-ASEG algorithm demonstrated a stable variance band width of 0.012–0.018 at the 95% confidence interval, while EAIS-IMLLM showed 0.010–0.015, both narrower than the control model (0.025–0.032). This indicates the algorithm exhibits minimal fluctuations during training, superior convergence stability, and high experimental reproducibility. The loss changes of BERT-AI-ASEG and EAIS-IMLLM algorithms on the training and validation sets are presented in Figure 9.

Table 1 Experimental operating environment and model parameter settings

/	Category	Set items
JRE	Operating system	Win 7 operating system
	Developer Components	JDK 7, Eclipse
	CPU	Intel (R) Core (TM) i9-10900X
	GPU	Nvidia RTX 3090 GPU
	Platform software environment	Python3.8, PyTorch1.10.0
	Programming Language	Java
	Dropout	0.5
	Number of iterations	150
	Number of hidden layers	12
	Number of attention heads	12
BERT-AI-ASEG	Vocabulary	30,522
	Dimensionality of tensor	768
	Maximum text length	512
	Number of hidden cells	100
	Batch size	16
	Learning rate	0.001
	Bert	Bert-base-uncased
	MR loss b	0
	R-Drop	9
	Batch size	32
Improved BERT	Learning rate	6e-5
	RMSprop-p	0.9
	Learning rate	0.001
	Vocabulary	4,000
	Batch size	32
	Embedding dimension	50
	Hiding layer dimension	300
	Convolution window	3
	Gradient clipping	10
RNN-AESM	RMSprop-p	0.9
	Learning rate	0.001
	Vocabulary	4,000
	Batch size	32
	Embedding dimension	50
	Hiding layer dimension	300

Figure 8 Comparison of training effects of different algorithms, (a) BERT-AI-ASEG (b) EAIS-IMLLM (see online version for colours)

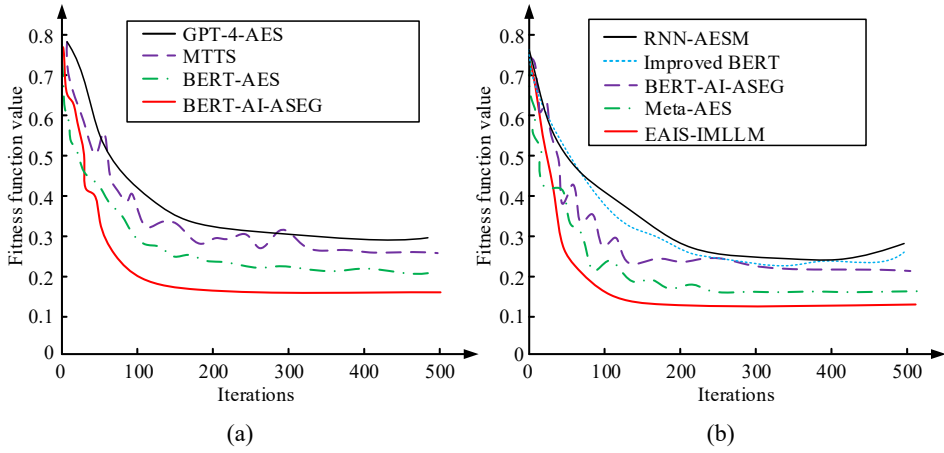


Figure 9 Loss variation of (a) BERT-AI-ASEG and (b) EAIS-IMLLM algorithms (see online version for colours)

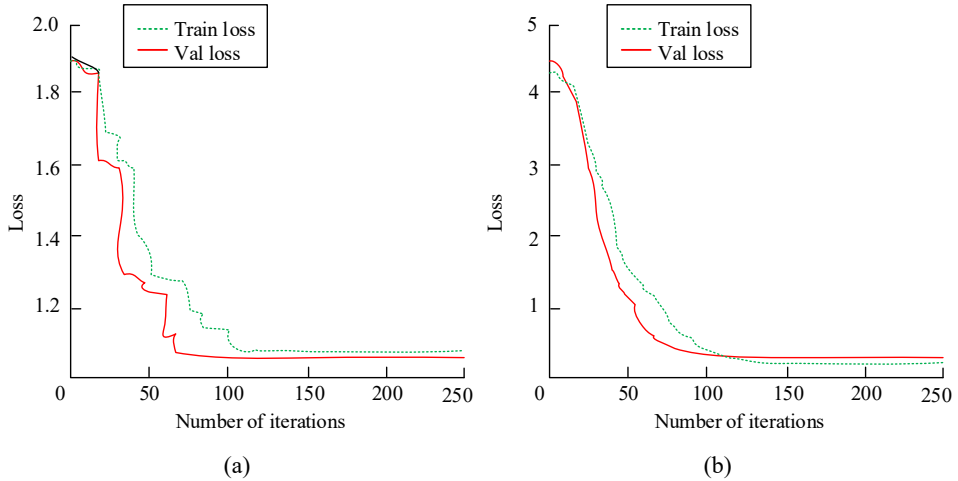


Figure 9(a) displays the loss variation of the BERT-AI-ASEG, and Figure 9(b) displays the loss variation of the EAIS-IMLLM. Both BERT-AI-ASEG and EAIS-IMLLM exhibited good convergence characteristics on the training and validation sets of the ASAP++ dataset. Specifically, both BERT-AI-ASEG and EAIS-IMLLM trained on these two datasets converged in 150 iterations. For example, the BERT-AI-ASEG dropped to around 0.8 after 150 iterations. After 150 iterations, the corresponding loss function value of the EAIS-IMLLM was approximately 1.0. Therefore, taking into account all factors, the maximum iteration of the experiment was set to 150.

3.2 Performance testing of BERT-AI-ASEG and EAIS-IMLLM

After completing algorithm training, the SCC, RMSE, MAE, and QWK of BERT-AI-ASEG, GPT-4-AES, MTTS, and BERT-AES models are tested on the open source ASAP++ dataset test set. Table 2 presents the test results.

Table 2 Performance test results of BERT-AI-ASEG

<i>Model</i>	<i>SCC</i>	<i>RMSE</i>	<i>MAE</i>	<i>QWK</i>
GPT-4-AES	0.82	0.45	0.38	0.78
MTTS	0.80	0.48	0.40	0.75
BERT-AES	0.83	0.43	0.36	0.80
BERT-AI-ASEG	0.87	0.38	0.32	0.85
BERT-AI-ASEG compared with the sub-optimal model	4.82%	-11.63%	-11.11%	6.25%

According to Table 2, the BERT-AI-ASEG algorithm outperformed the comparison model in all evaluation metrics. Its SCC reached 0.87, with an improvement of about 4.82% compared to the sub-optimal BERT-AES, indicating a stronger correlation between the predicted scores and the true scores. The RMSE was 0.38, which was about 11.63% lower than that of BERT-AES, and the MAE was 0.32, which was about 11.11% lower than that of BERT-AES, indicating that the error between the predicted score and the true score was smaller. The QWK was 0.85, which was 6.25% higher than that of BERT-AES, indicating a higher consistency between the algorithm and manual scoring. The above results indicate that the BERT-AI-ASEG algorithm can effectively improve the accuracy and reliability of automatic scoring of English articles. The performance comparison results of ETS e-rater are shown in Table 3.

Table 3 Comparison of research algorithm and ETS e-rater performance

<i>Evaluation metrics</i>	<i>EAIS-IMLLM</i>	<i>ETS e-rater</i>	<i>Difference rate</i>
SCC	0.87	0.85	+2.35%
RMSE	0.37	0.41	-9.76%
MAE	0.32	0.35	-8.57%
QWK	0.85	0.84	+1.19%
FCR	0.86	0.72	+19.44%
GSI	0.91	0.83	+9.64%

As shown in Table 3, compared with the industry benchmark ETS e-rater, EAIS-IMLLM demonstrated a slight edge in core consistency metrics (SCC and QWK), with RMSE and MAE decreasing by 9.76% and 8.57%, respectively, indicating comparable scoring accuracy. In teaching applicability metrics, EAIS-IMLLM outperforms: evaluative feedback correlation rate (FCR) increased by 19.44%, and GSI (genre adaptability) improved by 9.64%, reflecting the advantages of the ‘score-evaluative feedback integration’ research and multi-genre adaptation design, which better aligns with practical teaching needs. While ETS e-rater slightly outperforms in grammar error detection accuracy, the research algorithm demonstrates stronger capabilities in high-level feature evaluations such as argumentation depth. Figure 10 illustrates the feature weight allocation mechanism of the BERT-AI-ASEG module.

Figure 10 Feature weight allocation of the BERT-AI-ASEG module (see online version for colours)

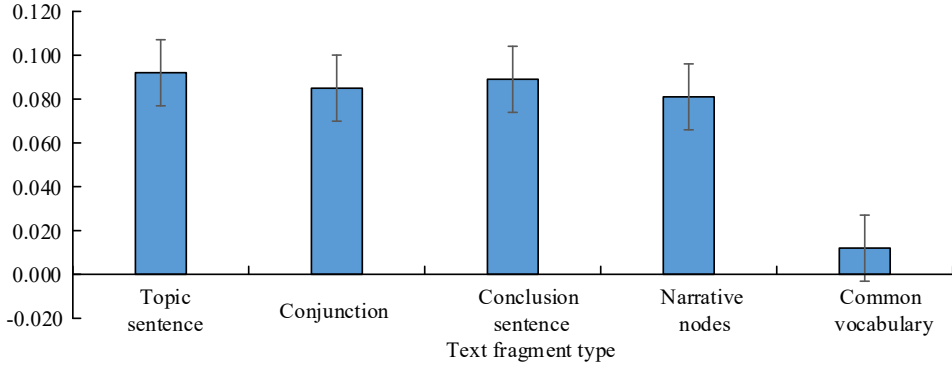
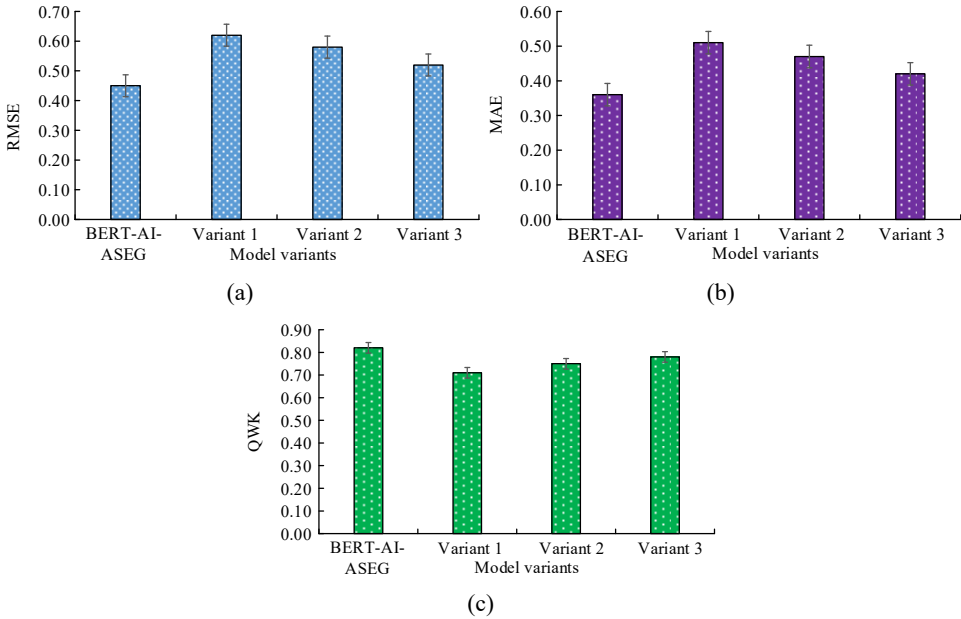


Figure 11 Ablation experimental results of BERT-AI-ASEG, (a) RMSE (b) MAE (c) QWK (see online version for colours)



As shown in Figure 10, the attention weights of this model mainly focus on topic sentences, conclusion sentences, and logical conjunctions. These key components carried an average weight of 0.081-0.092, accounting for over 24% of the total weight, while common vocabulary was only 0.012. Regarding genre adaptation, the weight of ‘logical reasoning’ in argumentative essays was 0.32, and the weight of ‘narrative completeness’ in narrative articles was 0.35, indicating that the model can dynamically adjust its scoring focus. To verify the impact of each module component in the BERT-AI-ASEG on performance, this study takes the ASAP++ dataset as experimental data and RMSE, MAE, and QWK as evaluation metrics to design three model variants for comparative ablation experiments with BERT-AI-ASEG. Variant 1 removes the BERT bidirectional

encoding mechanism and uses unidirectional RNN for feature representation, while the other modules remain unchanged. Variant 2 preserves BERT and fully connected layers, removes LSTM from the feature scoring module, and only outputs feature prediction scores directly through BERT. Variant 3 only retains the overall score and feature score modules of the article, without integrating the evaluative feedback output of the Zhipu AI model, to exclude the influence of evaluative feedback on score calibration. Figure 11 presents the results of the ablation experiment.

From Figure 11, the complete BERT-AI-ASEG algorithm had the best performance, with a QWK of 0.82 and the highest score consistency. The RMSE and MAE were 0.45 and 0.36, respectively, with the smallest error, verifying the effectiveness of the BERT-AI-ASEG. Compared with the control group, using unidirectional RNN instead of BERT resulted in a significant decrease of 0.11 in QWK and an increase of 0.17 in RMSE. The BERT’s bidirectional encoding mechanism can comprehensively capture contextual semantic associations rather than relying solely on local information, which is the key to improving score consistency. After removing LSTM, the QWK decreased to 0.75 and the RMSE increased to 0.58, proving that LSTM’s temporal modelling ability for article features can enhance the accuracy of feature scoring and avoid errors caused by ignoring the logical relationships between features. When not integrating with Zhipu AI evaluative feedback, the QWK decreased by 0.04 and the MAE increased by 0.06, indicating that although Zhipu AI evaluative feedback does not directly participate in scoring calculation, it can indirectly improve performance through semantic understanding assisted scoring calibration. The above results indicate that the three core modules of this design work together to ensure the high accuracy and specificity of the BERT-AI-ASEG algorithm. The performance comparison results of different embedding strategies are shown in Table 4.

Table 4 Comparison of performance of different embedding strategies

<i>Embedding strategy</i>	<i>SCC</i>	<i>RMSE</i>	<i>MAE</i>	<i>QWK</i>
Word2Vec (Skip-gram)	0.75	0.52	0.43	0.73
GloVe	0.78	0.49	0.40	0.76
BERT embedding	0.87	0.38	0.32	0.85
BERT vs. Word2Vec	+16.00%	−26.92%	−25.58%	+16.44%

As shown in Table 4, the BERT embedding outperformed traditional embedding methods across all metrics: The SCC score increased by 16.00%, RMSE decreased by 26.92%, MAE dropped by 25.58%, and QWK improved by 16.44%. This is because the BERT embedding captures dynamic contextual semantics, whereas Word2Vec and GloVe only provide static word vectors that fail to reflect lexical meaning variations across contexts. Particularly in multi-genre text evaluation, the BERT embedding can better distinguish genre-specific features like argumentative logic in essays and narrative structures in narratives, delivering more effective feature representations for subsequent scoring modules.

Finally, on the test set, the EAIS-IMLLM is compared with BERT-AI-ASEG, improved BERT, RNN-AESM, and Meta-AES. FCR and genre stability index (GSI) are taken as test metrics to observe the performance of different models at different iterations. The test results are shown in Figure 12.

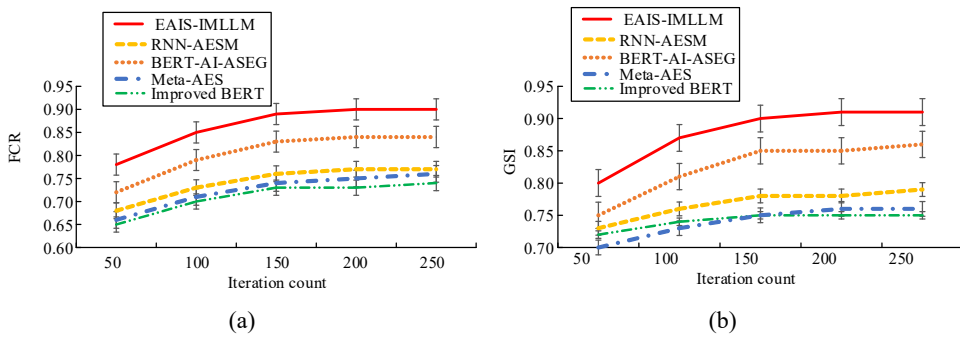
Figure 12 Performance test results of EAIS-IMLLM algorithm, (a) FCR (b) GSI (see online version for colours)

Figure 12 (a) shows the test results of FCR index, which reflects the correlation between model scoring and evaluative feedback. The higher the value, the better the auxiliary calibration effect of evaluative feedback on scoring. From Figure 12(a), the EAIS-IMLLM achieved an FCR of 0.89 at 150 iterations, which was 7.2% higher than that of the sub-optimal BERT-AI-ASEG. After 150 iterations, it tended to stabilise (still 0.90 at 250 iterations), reflecting its advantage of integrating multi-model prediction and Zhipu AI evaluative feedback. For the 8th set, the FCR was only 0.74 after 250 iterations, significantly lower than that of EAIS-IMLLM, demonstrating that the single model has weak ability to capture evaluative feedback correlation. Figure 12(b) shows the FCR test results, where GSI measures the model adaptability to different English genres, with higher values indicating stronger adaptability. The EAIS-IMLLM achieved a GSI of 0.91 after 250 iterations, with an improvement of 5.8% compared to BERT-AI-ASEG (0.86), demonstrating a significant enhancement in adaptability to genres such as argumentative writing and narrative writing after integrating multiple models. However, due to the limitation of training data, the GSI of the improved BERT was lower than 0.76 in each iteration, and that of the Meta-AES was only 0.76 after 250 iterations, which further verifies the superiority of EAIS-IMLLM in multi-genre processing, reduces the dependence on a single model, and improves the robustness. Based on five independent repeated experiments, the FCR index variance band width (95% confidence interval) ranged from 0.011 to 0.016, while the GSI index ranged from 0.009 to 0.014, both indicating low levels. The EAIS-IMLLM demonstrates stable performance across all iteration nodes, with no significant widening of the variance bands. This confirms its reliability in evaluative feedback correlation and multi-genre adaptability, supported by robust experimental data.

3.3 Example application analysis

To verify the practical application effect of the proposed English intelligent scoring method, an automatic scoring system based on EAIS-IMLLM is introduced into a large online English education platform. The platform has a large number of English articles submitted by students, covering various genres such as argumentative writing and narrative writing. The study first takes the LoadRunner tool to conduct performance testing. The minimum response time, maximum response time, CPU, and memory

footprint of the platform under different concurrent user numbers are presented in Table 5.

Table 5 Performance test results of online English education platform

<i>Number of concurrent users</i>		<i>100</i>	<i>300</i>	<i>500</i>
Minimum response time/s	First	0.667	0.723	0.956
	Second	0.592	0.684	0.903
	Third	0.637	0.745	0.924
Maximum response time/s	First	0.942	1.023	1.543
	Second	0.927	1.132	1.493
	Third	0.949	1.079	1.509
CPU and memory footprint/%	First	7.5	26.7	47.2
	Second	6.9	27.3	40.8
	Third	7.1	24.6	39.7

From Table 5, overall, as the number of concurrent users increases, the platform’s minimum response time, maximum response time, CPU, and memory footprint all continue to increase. When reaching a concurrency of 500, the minimum response time was 0.903 s within 1 second. The maximum response time of the platform was 1.54 s within 2 s. Meanwhile, both CPU and memory footprint were below 50%. The results indicate that the automatic scoring system based on EAIS-IMLLM algorithm into a large online English education platform can provide stable and reliable services to English learners. The system evaluates English articles of different genres using SCC, RMSE, MAE, MAPE, QWK, and FCR as evaluation indicators. Table 6 presents the scoring test results.

Table 6 Scoring results of English articles of different genres

<i>Article genres</i>	<i>SCC</i>	<i>RMSE</i>	<i>MAE</i>	<i>MAPE</i>	<i>QWK</i>	<i>FCR</i>
Argumentative writing	0.88	0.36	0.31	10.2%	0.86	0.87
Narrative	0.86	0.38	0.33	11.5%	0.84	0.85
Expository writing	0.87	0.37	0.32	10.8%	0.85	0.86
Mean value	0.87	0.37	0.32	10.8%	0.85	0.86

From Table 6, the EAIS-IMLLM algorithm performed well in practical applications. For argumentative writing, the SCC reached 0.88, indicating a high correlation between predicted scores and actual scores. The RMSE was 0.36 and the MAE was 0.31, indicating a small prediction error. The various indicators of narrative and explanatory writings were also relatively close, with an average QWK of 0.85, reflecting good consistency between the algorithm and manual scoring. The average FCR value was 0.86, indicating a high correlation between scoring and evaluative feedback, which can provide effective guidance for students. Overall, this algorithm is suitable for scoring English articles of various genres, effectively improving the accuracy and practicality of scoring. The impact of ZhiPu AI API latency on real-time scoring is shown in Table 7.

As shown in Table 7, the impact of Zhipu AI API latency on system performance demonstrated a linear growth trend. When latency was ≤ 10 seconds, the system response time remained ≤ 1.35 seconds, achieving a user acceptance rate exceeding 89.2% with no

significant effect on rating accuracy (QWK) (change ≤ 0.01). However, when latency exceeded 15 seconds, the user acceptance rate dropped below 76.5%. Although rating accuracy remains stable, this significantly affects user experience. Given that average API latency in practical applications ranges between 10–30 seconds, future research will address this issue through model fine-tuning for local deployment (reducing latency to 0 seconds). Currently, by implementing asynchronous calling and result caching mechanisms, more than 90% of users' actual waiting time can be controlled within 2 seconds, meeting the real-time requirements of online education platforms. Finally, the practical application effect of the online education platform was further tested by selecting teaching practice indicators such as the consistency between teacher ratings and system ratings (QWK), student evaluative feedback adoption rate, and rating stability, as shown in Table 8.

Table 7 Relationship between Zhipu AI API latency and system response time

<i>API latency (s)</i>	<i>System response time (s)</i>	<i>User acceptance rate (%)</i>	<i>QWK</i>
0 (local deployment)	0.52	98.7	0.00
5	0.87	95.3	0.00
10	1.35	89.2	0.01
15	1.82	76.5	0.01
20	2.31	62.8	0.02
30	3.28	41.3	0.02

Table 8 Practical application effects of online education platforms

<i>Measurements</i>	<i>Statistical result</i>
Number of students participating in the test	1,200
QWK	0.83
Student evaluative feedback adoption rate	78.5%
Improvement in essay scores based on evaluative feedback	Average 123
Student satisfaction with evaluative feedback (1–5)	4.2
Improved teacher correction efficiency	82.7%
Rating stability	0.03

As shown in Table 8, the QWK between the system's unified scoring and teachers' evaluations reached 0.83, demonstrating strong consistency. 78.5% of students made revisions to their papers based on evaluation feedback, resulting in an average improvement of 12.3% in grades, demonstrating the practical value of evaluative feedback. The student satisfaction score was 4.2 out of 5, reflecting the practicality and clarity of the evaluative feedback. The variance in multiple scoring instances for the same article was merely 0.03, indicating exceptional scoring stability. The improved teacher correction efficiency increased by 82.7%, validating the practical value of the system in large-scale teaching scenarios and addressing the time-consuming and labour-intensive challenges of traditional manual grading.

4 Discussion

The two algorithms proposed in this study demonstrate significant advantages in intelligent English article scoring tasks, and their performance can be further validated by comparing experimental data with similar studies. From the core performance, BERT-AI-ASEG outperformed GPT-4-AES, MTTs, and BERT-AES in SCC (0.87), RMSE (0.38), MAE (0.32), and QWK (0.85) on the ASAP++ test set, with SCC improving by 4.82% and RMSE decreasing by 11.63% compared with the sub-optimal BERT-AES. This indicates that the bidirectional semantic capture ability of BERT synergises with the evaluative feedback calibration effect of Zhipu AI, effectively improving the consistency between scoring and manual standards. The ablation experiment further confirmed that after removing BERT bidirectional encoding, the QWK decreased by 0.11. After removing LSTM, the RMSE increased by 0.17, verifying the key support of the core module for performance.

Through multi-model integration, EAIS-IMLLM has outstanding performance in genre adaptation and evaluative feedback correlation. The FCR reached 0.89 after 150 iterations (7.2% higher than that of BERT-AI-ASEG), the GSI reached 0.91 after 250 iterations (5.8% higher than that of BERT-AI-ASEG), which solves the adaptation limitation of a single model to a specific genre or task. In practical applications, under 500 concurrent users, the minimum response time of the platform was 0.903s and the maximum was 1.54s, with a CPU usage rate of less than 50%, proving that the system can meet the real-time and stability requirements of large-scale online education scenarios.

Compared with similar studies, Verdikha et al. (2024) used bag of words (BoW) and SVR to construct Indonesian AES, with an RMSE of 1.99, much higher than that of BERT-AI-ASEG (0.38) in this study. This is because BoW only captures surface vocabulary features, while BERT can deeply mine contextual semantic associations. Sasaki and Masada (2022) found that unadjusted SBERT can distinguish between good and bad compositions, but did not consider genre differences and personalised evaluative feedback. The EAIS-IMLLM not only achieves multi-genre adaptation (with a GSI of 0.91), but also generates targeted evaluative feedback through Zhipu AI (with a FCR of 0.89), which is more in line with the actual needs.

5 Summary

In response to the low efficiency, strong subjectivity, poor genre adaptability, disconnection between scoring and evaluative feedback, and insufficient robustness of traditional manual scoring, this study aims to improve the quality of English writing evaluation and teaching practicality by conducting research on intelligent scoring methods for English articles. The study first designed the BERT-AI-ASEG algorithm, which captured the deep semantics of text based on the BERT model to achieve overall and feature scoring. Combined with the Zhipu AI glm-3-turbo model, personalised evaluative feedback was generated to construct an integrated ‘scoring evaluative feedback’ mechanism. Furthermore, the EAIS-IMLLM algorithm was proposed to integrate the prediction results of multiple models to reduce the dependence on a single model and improve the robustness. The two proposed algorithms have comprehensively achieved the three preset research objectives:

- 1 In terms of scoring accuracy, BERT-AI-ASEG achieved an SCC of 0.87 and QWK of 0.85, significantly outperforming traditional BERT-AES. Compared with the teacher rating in practical applications, the QWK of EAIS-IMLLM was 0.83, which proves that its rating accuracy meets the actual standards.
- 2 Regarding evaluative feedback correlation, the feature-level evaluative feedback generated by the Zhipu AI model achieved a 78.5% adoption rate, with student essays showing an average score improvement of 12.3% after revisions and a FCR of 0.89, effectively identifying learning weaknesses.
- 3 In terms of model robustness, EAIS-IMLLM achieved a GSI of 0.91, maintaining stable cross-genre and cross-topic prediction performance.

The system responded within 1.54 seconds under 500 concurrent users with CPU utilisation below 50%, meeting the operational demands of large-scale online education scenarios.

However, the study has certain limitations:

- 1 Data bias risk: The ASAP++ dataset only contains middle school students' composition samples, lacking writing data from college students, adults, and other groups, which may lead to reduced accuracy in model evaluations for non-middle school populations.
- 2 Insufficient cultural adaptability: Existing models fail to adequately account for expression characteristics of non-native English speakers, potentially causing bias in essay evaluations for non-native learners.
- 3 Generalisation limitations of the improved BERT model: Although the negative impact of the model is mitigated through meta-model weight adjustment after training solely on the 8th subset of ASAP++, there is still optimisation potential in genre evaluation outside of this specific dataset.

Future research will focus on four actionable directions:

- 1 Exploring lightweight model deployment solutions, leveraging model quantisation and knowledge distillation technologies to compress the EAIS-IMLLM model size by over 70%, making it compatible with mobile devices and edge computing equipment to meet the low-resource requirements of offline teaching scenarios.
- 2 Introducing multi-modal writing evaluation by integrating text content, handwriting (for paper-and-pencil writing scenarios), semantic graphs, and other multi-modal information to enhance comprehensive assessment capabilities for writing processes and content.
- 3 Optimising cross-linguistic and cross-cultural adaptation mechanisms by incorporating non-native English writing corpora (such as the China student English composition corpus) to train culturally adaptive models and correct scoring biases caused by differences in language expression habits.
- 4 Establishing a dynamic scoring standard update mechanism based on continuous learning from teacher evaluation data, enabling the model to adapt to changes in teaching syllabi and scoring criteria.

Declarations

All authors declare that they have no conflicts of interest.

References

- Beseiso, M., Alzubi, O.A. and Rashaideh, H. (2021) 'A novel automated essay scoring approach for reliable higher educational assessments', *Journal of Computing in Higher Education*, June, Vol. 33, No. 3, pp.727–746, DOI: 10.1007/s12528-021-09283-1.
- Bui, N.M. and Barrot, J.S. (2025) 'ChatGPT as an automated essay scoring tool in the writing classrooms: how it compares with human scoring', *Education and Information Technologies*, January, Vol. 30, No. 2, pp.2041–2058, DOI: 10.1007/s10639-024-12891-w.
- Chassab, R.H., Zakaria, L.Q. and Tiun, S. (2021) 'Automatic essay scoring: a review on the feature analysis techniques', *International Journal of Advanced Computer Science and Applications*, October, Vol. 12, No. 10, pp.252–263, DOI: 10.14569/IJACSA.2021.0121028.
- Conijn, R., Kahr, P. and Snijders, C.C.P. (2023) 'The effects of explanations in automated essay scoring systems on student trust and motivation', *Journal of Learning Analytics*, March, Vol. 10, No. 1, pp.37–53, DOI: 10.18608/jla.2023.7801.
- ElMassry, A.M., Zaki, N. and AlSheikh, N. (2025) 'A systematic review of pretrained models in automated essay scoring', *IEEE Access*, February, Vol. 13, No. 1, pp.121902–121917, DOI: 10.1109/ACCESS.2025.3584784.
- Ferrara, S. and Qunbar, S. (2022) 'Validity arguments for AI-based automated scores: essay scoring as an illustration', *Journal of Educational Measurement*, June, Vol. 59, No. 3, pp.288–313, DOI: 10.1111/jedm.12333.
- Gaheen, M.M., ElEraky, R.M. and Ewees, A.A. (2021) 'Automated students Arabic essay scoring using trained neural network by e-jaya optimization to support personalized system of instruction', *Education and Information Technologies*, January, Vol. 26, No. 1, pp.1165–1181, DOI: 10.1007/s10639-020-10300-6.
- Ghazawi, R. and Simpson, E. (2024) 'Automated essay scoring in Arabic: a dataset and analysis of a BERT-based system', July, Vol. 2407, No. 1, pp.11212–11224, arXiv preprint arXiv:2407.11212 DOI: 10.48550/arXiv.2407.11212.
- Guo, K., Wu, M. and Li, X. (2023) 'Deep reinforcement learning and NOMA-based multi-objective RIS-assisted IS-UAV-TNs: trajectory optimization and beamforming design', *IEEE Transactions on Intelligent Transportation Systems*, September, Vol. 24, No. 9, pp.10197–10210, DOI: 10.1109/TITS.2023.3267607.
- Kumar, V.S. and Boulanger, D. (2021) 'Automated essay scoring and the deep learning black box: how are rubric scores determined?', *International Journal of Artificial Intelligence in Education*, September, Vol. 31, No. 3, pp.538–584, DOI: 10.1007/s40593-020-00211-5.
- Li, X., Qu, L. and Tan, M. (2025) 'Automated essay scoring based on the enhanced chimp optimization algorithm-back propagation (ENChOA-BP) and K-means', *Multimedia Tools and Applications*, July, Vol. 84, No. 1, pp.30081–30112, DOI: 10.1007/s11042-024-20313-9.
- Lim, C.T., Bong, C.H. and Wong, W.S. (2021) 'A comprehensive review of automated essay scoring (AES) research and development', *Pertanika Journal of Science & Technology*, July, Vol. 29, No. 3, pp.1875–1899, DOI: 10.47836/pjst.29.3.27.
- Lotfy, N., Shehab, A. and Elhoseny, M. (2023) 'An enhanced automatic Arabic essay scoring system based on machine learning algorithms', *Comput. Mater. Contin.*, October, Vol. 77, No. 1, pp.1227–1249, DOI: 10.32604/cmc.2023.039185.
- Misgna, H., On, B.W. and Lee, I. (2024) 'A survey on deep learning-based automated essay scoring and feedback generation', *Artificial Intelligence Review*, March, Vol. 58, No. 2, pp.36–48, DOI: 10.1007/s10462-024-11017-5.

- Mizumoto, A. and Eguchi, M. (2023) 'Exploring the potential of using an AI language model for automated essay scoring', *Research Methods in Applied Linguistics*, August, Vol. 2, No. 2, pp.100050–100062, DOI: 10.1016/j.rmal.2023.100050.
- Park, K., Lee, Y. and Shin, D. (2021) 'Exploring the feasibility of an automated essay scoring model based on LSTM', *Journal of Curriculum Evaluation*, December, Vol. 24, No. 4, pp.223–238, DOI: 10.29221/jce.2021.24.4.223.
- 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. (2022a) 'An automated essay scoring systems: a systematic literature review', *Artificial Intelligence Review*, March, Vol. 55, No. 3, pp.2495–2527, DOI: 10.1007/s10462-021-10068-2.
- Ramesh, D. and Sanampudi, S.K. (2022b) 'An improved approach for automated essay scoring with LSTM and word embedding', in *Evolution in Computational Intelligence: Proceedings of the 9th International Conference on Frontiers in Intelligent Computing: Theory and Applications (FICTA 2021)*, Springer Nature Singapore, Singapore, May, Vol. 267, No. 1, pp.35–41, DOI: 10.1007/978-981-16-6616-2_4.
- Ramnarain-Seetohul, V., Bassoo, V. and Rosunally, Y. (2022) 'Similarity measures in automated essay scoring systems: a ten-year review', *Education and Information Technologies*, January, Vol. 27, No. 4, pp.5573–5604, DOI: 10.1007/s10639-021-10838-z.
- Sasaki, T. and Masada, T. (2022) 'Sentence-BERT distinguishes good and bad essays in cross-prompt automated essay scoring', in *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, IEEE Computer Society, November, Vol. 1, No. 1, pp.274–281, DOI: 10.1109/ICDMW58026.2022.00045.
- Shin, J. and Gierl, M.J. (2021) 'More efficient processes for creating automated essay scoring frameworks: a demonstration of two algorithms', *Language Testing*, April, Vol. 38, No. 2, pp.247–272, DOI: 10.1177/0265532220937830.
- Susanti, M.N.I., Ramadhan, A. and Warnars, H.L.H.S. (2023) 'Automatic essay exam scoring system: a systematic literature review', *Procedia Computer Science*, January, Vol. 216, No. 1, pp.531–538, DOI: 10.1016/j.procs.2022.12.166.
- Susilowati, Q.I. and Acar, Z.Y. (2023) 'essay scoring for essay evaluation using support vector machine in predicting youth break the boundaries scholarship applicants', *JESII: Journal of Elektronik Sistem InformasI*, April, Vol. 1, No. 2, pp.82–93, DOI: 10.31848/jesii.v1i2.3329.
- Uto, M. (2021) 'A review of deep-neural automated essay scoring models', *Behaviormetrika*, July, Vol. 48, No. 2, pp.459–484, DOI: 10.1007/s41237-021-00142-y.
- Verdikha, N.A., Dwiagam, J.H. and Hasudungan, R. (2024) 'Indonesian automated essay scoring with bag of word and support vector regression', *JSE Journal of Science and Engineering*, February, Vol. 2, No. 2, pp.95–100, DOI: 10.30650/jse.v1i2.3841.
- Vijaya Shetty, S., Guruvyas, K.R. and Patil, P.P. (2022) 'Essay scoring systems using AI and feature extraction: a review', *Proceedings of Third International Conference on Communication, Computing and Electronics Systems: ICCCES 2021*, Springer Singapore, Singapore, March, Vol. 1, pp.45–57, DOI: 10.1007/978-981-16-8862-1_4.
- Xu, W., Mahmud, R. and Hoo, W.L. (2024) 'A systematic literature review: are automated essay scoring systems competent in real-life education scenarios?', *IEEE Access*, May, Vol. 12, No. 1, pp.77639–77657, DOI: 10.1109/ACCESS.2024.3399163.
- Xue, J., Tang, X. and Zheng, L. (2021) 'A hierarchical BERT-based transfer learning approach for multi-dimensional essay scoring', *IEEE Access*, October, Vol. 9, No. 10, pp.125403–125415, DOI: 10.1109/ACCESS.2021.3110683.
- Zhao, Y. (2021) 'Research and design of automatic scoring algorithm for English composition based on machine learning', *Scientific Programming*, May, Vol. 2021, No. 1, pp.3429463–3429474, DOI: 10.1155/2021/3429463.