International Journal for Housing Science and Its Applications
Publish August 10, 2025. Volume 46, Issue 4 Pages 1385-1395

https://doi.org/10.70517/ijhsa464126

Research on Automatic Grading Algorithm for College English Teaching Based on Generative Artificial Intelligence

Yao Ma[1,*]
[1] Geely University of China, Chengdu, Sichuan, 610000, China
Corresponding authors: (e-mail: emmayang223@163.com).

Abstract With the emergence of generative artificial intelligence such as GPT, it provides an important tool to support the automatic study of English. This paper mainly proposes an automatic scoring method for college English based on generative AI-ChatGPT fine-tuning model. Through ChatGPT's powerful language understanding and generative capabilities, intelligent automatic grading of students' answers is carried out by scoring logic and algorithms. Comparative experiments are conducted on different datasets, and the conclusions show that the loss function curve of the automatic scoring algorithm proposed in this paper reaches the optimum at about 30 iterations, and its convergence speed is faster and the model performance is better. Compared with LSTM, GRU and other control algorithms, the accuracy of the method proposed in this paper reaches 89.62%, which is higher and the average absolute error is smaller. The method in this paper achieved an average score of 0.7964 on the average QWK scores of the eight topics. The attitudes of the students of the experimental and control versions towards automated scoring have mutual in difference with Sig value of 0.001.

Index Terms generative artificial intelligence, ChatGPT, fine-tuning model, automatic scoring

I. Introduction

With the rapid development of science and technology, especially the advancement of artificial intelligence technology, the field of education has also ushered in unprecedented changes [1], [2]. As an important branch of artificial intelligence, generative artificial intelligence (GAI) is gradually becoming a hot research topic in the field of educational technology by learning a large amount of data to generate new content with similar characteristics to the training data [3]-[5]. In university English teaching, the application of GAI can not only enrich teaching resources and improve teaching efficiency, but also provide a more personalized learning experience for students' individual differences, while automated grading of teaching is an important driver for optimizing English teaching [6]-[9].

GAI can generate all kinds of assessment tools, such as questionnaires, test forms, etc., to help teachers score the effectiveness of teaching, and it enables teachers to efficiently and quickly identify problems by providing detailed feedback reports, and communicate these problems and corrections to students in a timely manner, as well as providing a basis for teachers to adjust their teaching strategies [10]-[13]. For example, if the report shows that most students are struggling with a particular knowledge point, teachers may need to reconsider their teaching methods for that section or add exercises to deepen understanding [14], [15]. By consistently using the data and analysis provided by GAI, teachers can continuously optimize course design to ensure that teaching activities are more relevant to students' needs, ultimately achieving the goal of improving overall teaching quality [16]-[18]. However, GAI is not perfect, and there are some challenges and limitations. The content generated by GAI may lack depth and diversity, and requires secondary processing and optimization by teachers [19]-[21]. Therefore, teachers need to be technologically and creatively aware in order to fully utilize its potential [22].

Aiming at the traditional manual grading method which is not only inefficient and other problems, this paper proposes an automatic grading method for college English based on generative artificial intelligence-ChatGPT fine-tuning model mainly using natural processing technology, machine learning and so on. The students' English answers are first preprocessed to standardize the text, which is convenient for the subsequent extraction of key features from the preprocessed answers. The Transformer architecture is selected, and the scoring logic and algorithm are the core of the automatic scoring method. The architecture incorporates a multi-layer self-attention mechanism and a feed-forward neural network to assess the quality of candidates' answers. Entering the fine-tuning phase, supervised learning is used to measure the difference between the model's predicted scoring and the actual scoring to improve the model's performance. A large number of rules and weights are utilized to design the scoring algorithm. In this paper, a large amount of real exam data is used to verify the accuracy and reliability of the model.

II. Research on automatic scoring method based on ChatGPT

II. A.Data pre-processing

Data preprocessing is a crucial aspect. Its main purpose is to transform the original subjective question answer text into a standardized and structured format for subsequent analysis and scoring by the ChatGPT model. Through text preprocessing, we can effectively improve the accuracy and efficiency of scoring. Text preprocessing includes: text cleaning, text standardization, word and sentence division and other processes.

II. A. 1) Text cleaning

Text cleaning is the first step of text preprocessing, the main purpose of which is to remove irrelevant characters, special symbols and noisy data from the text. These irrelevant elements may cause interference to the analysis of ChatGPT and reduce the accuracy of scoring. In the text cleaning process, we usually use tools such as regular expressions to identify and remove these unnecessary elements.

II. A. 2) Text standardization

Text standardization is another important step in text preprocessing. In this stage, we mainly carry out uniform formatting of the text, including the unification of letter case, the standardization of punctuation, and the deletion of redundant spaces. These operations help to improve ChatGPT's comprehension of the text, which in turn improves the accuracy of scoring. The main operations of this standardization, including case conversion, punctuation normalization, and deletion of redundant spaces [23].

II. A. 3) Textual clauses and subclauses

Segmentation is the last key step in text preprocessing. For Chinese and other languages without obvious lexical boundaries, the operation of word segmentation is especially important. Segmentation can divide the continuous text into independent lexical units, which is convenient for ChatGPT to analyze more deeply. At the same time, clause segmentation helps ChatGPT to better understand the structure and semantic relationship of the text [24].

II. B.Feature Extraction and Representation

The textual dimensions that feature extraction focuses on are shown in Figure 1. In the process of feature extraction, we mainly focus on three dimensions of the text: semantic, structural and syntactic. Semantic features: mainly reflect the relevance between students' answers and the requirements of the questions. This includes keyword identification, conceptual understanding, and analysis of the fit between the answer and the topic of the question. Structural features: reflect the logic and organization of students' answers. This includes: paragraph placement, logical structure and organization. Grammatical features: reflecting the correctness of students' language use, including: spelling check, grammatical error check and language fluency check. Feature representation refers to the conversion of raw data (e.g., text, images, etc.) into a numerical form that can be understood and processed by the model. In natural language processing, feature representation usually involves converting textual data into points in a vector space so that the model can capture the semantic relationships between words, phrases and sentences [25].

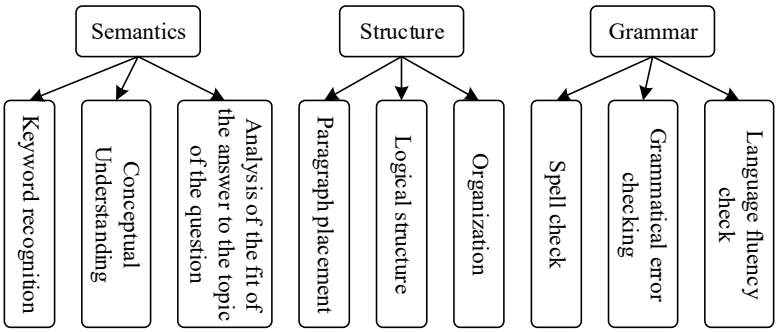


Figure 1: The text of the map is the text level that is concerned

II. C.Model training and fine-tuning

When choosing the model architecture, we mainly consider the Transformer architecture, especially the GPT family, because it has shown excellent performance in natural language processing tasks [26].The architecture of Transformer is shown in Fig. 2.Transformer is an attention mechanism based on the attention mechanism proposed by a research team of Google in 2017 Sequence transformation model, which abandons Recurrent Neural Network

(RNN) and Convolutional Neural Network (CNN), and realizes the global dependency between input and output only through the attention mechanism. The architecture of Transformer includes encoder and decoder, the main function of encoder is to transform the input into word vectors, and the decoder generates the translation result based on the output of encoder. Each encoder and decoder consists of multiple identical layers, each containing self-attention mechanisms and feed-forward neural networks. This architecture allows Transformer to improve the performance of the model by capturing both local and global dependencies in the sequence while processing the sequence transformation task.The Transformer model first transforms the text sequence into machine-understandable vectors through embedding and captures the word order information through positional encoding. The encoder consists of multiple layers, which include a multi-head attention mechanism and a feed-forward neural network to capture long-distance dependencies in the text.

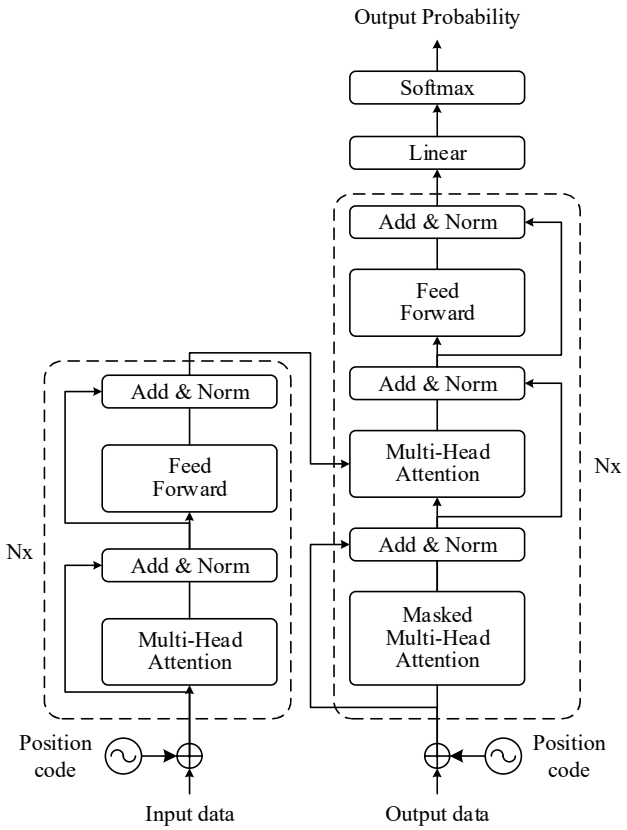


Figure 2: transformer structure

(1) Attention mechanism

Attention mechanism is a mechanism that focuses on local information, which belongs to deep learning technology, it can help the model to better understand the input data and extract useful information from it, its basic idea is to enable the model to ignore irrelevant information and pay more attention to the key information that should be paid attention to, for example, the attention mechanism when dealing with the task of a sentence of text, the area of attention will be placed on the semantics of this sentence most important words [27]. Specifically, the attention mechanism can calculate the attention distribution based on the correlation between the input data, and then calculate the weighted average of the input information based on the attention distribution, so that limited information can be allocated to the important parts, and the results obtained by using the attention mechanism are usually expressed in the form of probability maps or probability feature vectors. The attention mechanism has fewer parameters and lower complexity, which effectively alleviates the limitations of the current neural network's lack of computational power, and at the same time, the attention mechanism can do parallel processing, i.e., in each step of the computation does not depend on the previous step of the computation. After the introduction of the attention mechanism, the neural network model has the ability of “memory”, which can alleviate the limitations of the optimization algorithm in the model. The disadvantage of the attention mechanism is that it requires a large amount

of data, because the attention mechanism is to capture the key information and ignore the unimportant information, so when the data is small, the model is not effective.

(2) Multi-head Attention Mechanism

The multi-head attention mechanism is an improvement of the traditional attention mechanism, which splits the input sequence into multiple sub-sequences and applies the attention mechanism on each sub-sequence individually, i.e., it allows the model to compute the attention-weighted averages from different representation spaces separately, and then stitch these attention weights together to get the final output [28]. Compared with the traditional attention mechanism, the model using the multi-head attention mechanism can focus on the information between multiple positions in the sequence at the same time, and the attention-weighted average of the sequence under different linear transformations can be obtained by applying multiple linear transformations to the input feature matrix, so as to obtain comprehensive feature information. The steps for calculating the weighted average of multiple attention are as follows:

Step1: Initialize three vectors, the query vector Query, the key vector Key, and the value vector Value, the initial value is obtained by the product of the vector X corresponding to each character in the input sequence and the weight matrix W_q, W_k, and W_v:

$$Query = W_q X \tag{1}$$

$$Key = W_k X \tag{2}$$

$$Value = W_v X \tag{3}$$

Step2: Calculate Attention Score and Softmax Score, which reflects the correlation degree between this character and other characters, and also reflects the “attention degree” to other positions. The Attention Score is scaled and normalized to get the Softmax Score, where d_k denotes the dimension:

$$AttentionScore = Query \cdot Key \tag{4}$$

$$Soft\max Score = soft\max\left(\frac{AttentionScore}{\sqrt{d_k}}\right) \tag{5}$$

Step3: Finally multiply each Value vector by Softmax Score to get the weighted sum, which is the first input Attention Value:

$$AttentionValue = Value \cdot Soft\max Score \tag{6}$$

Then do multiple linear mappings on the original vectors $Query, Key, Value$, and map each time the result into multiple spaces. Repeat the above process, each time the result is called a “head”, and each “head” is put together to get the weights of multi-head attention. In short, the mechanism of multi-head attention is to let each attention to focus on different parts of the input information, and then combine the parts.

II. D.Scoring Logic and Algorithm

The scoring algorithm is designed based on a series of rules and weights to ensure the objectivity and fairness of scoring [29].

The idea of automatic marking is to compare the similarity between standard answers and students' answers, assign different weights to different words according to their different levels of importance, and finally give marks based on the similarity and weights. In order to reduce the complexity of marking points, marking requires pre-processing the student's answers with the standard answers in the database, including word splitting, removing useless words and punctuation, and so on. The next step, then, is to represent the reference and student answers as sets of words, $B = \{Wb1, Wb2, \cdots, Wbn\}$ with $D = \{Wd1, Wd2, \cdots, Wdn\}$. In addition, the similarity is recorded as $S = (Wbi, Wdj)$, which gives the similarity matrix between the student's answer and the reference answer as:

$$M(B,D) = \begin{bmatrix} S(W_{b1}, W_{d1}), S(W_{b2}, W_{d2}), \ldots, S(W_{bn}, W_{dn}) \\ S(W_{bm}, W_{d1}), S(W_{bm}, W_{d2}), \ldots, S(W_{bm}, W_{dn}) \end{bmatrix} \tag{7}$$

Then the similarity between the word strings Wbi and Wdj is calculated as:

$$S\left(W_{bi},W_{dj}\right)=\alpha\times\left(\frac{Match\left(W_{bi},W_{di}\right)}{Num\left(W_{bi}\right)}+\frac{Match\left(W_{bi},W_{di}\right)}{Num\left(W_{di}\right)}\right)\Big/2$$
$$+\beta\times\gamma\times\left(\sum\frac{Match\left(W_{bi},k\right)}{\sum_{i}^{m}k}+\sum\frac{Match\left(W_{dj},l\right)}{\sum_{i}^{n}l}\right) \tag{8}$$

α : it means the weight of semantic similarity when the two word strings contain the same number of elements, the value can be set to 0.7 in the system, of course, the value can be adjusted manually according to the situation;

β : it means the relationship between the position of the same element in the word string on the influence of semantic similarity weight, in the system can set its value to 0.3, of course, the value can be adjusted manually according to the situation;

γ : the positional coefficient expressed, calculated as:

$$\gamma=Min\left(Num\left(W_{bi}\right)\Big/Num\left(W_{dj}\right),Num\left(W_{dj}\right)\Big/Num\left(W_{bi}\right)\right) \tag{9}$$

$Match\left(W_{bi},W_{dj}\right)$ represents the number of two words containing the same element; and $\sum\frac{\mathrm{Match}\left(W_{bi},k\right)}{\sum_{l}^{m}k}$ and $\sum\frac{Match\left(w_{dj},l\right)}{\sum_{l}^{n}l}$ match the weight sum of the location in Wbi and Wdj, respectively, then according to the above equation we can get the formula for calculating the similarity between the candidate's answer and the given reference answer:

$$S\left(B,D\right)=\frac{\sum_{i=1}^{n}\max\left(s\left(W_{bi},W_{dl}\right),\ldots,s\left(W_{bi},W_{dn}\right)\right)}{\max\left(m,n\right)} \tag{10}$$

III. Performance analysis of generative AI auto-scoring algorithms

III. A. Simulation environment and performance testing

The simulation compilation environment is python language and the deep learning network model is built based on Tensor-flow framework. The hardware environment for simulation is Win 7, 64-bit operating system Lenovo server with 2.30GHz CPU, 16GB RAM, and NVIDIA RTX1080 with 12GB graphics card.

This section utilizes the data collected from the online English test to measure the experiments conducted, and the final dataset related statistics are 2930 for the training set, 731 for the validation set, and 1827 for the test set.

Firstly, the data in the training, validation and test sets are normalized while all the texts are converted to lowercase letters. Some of the parameters during network training are shown in Table 1.

Table 1: Network training parameters

Parameter	Learning rate	Learning rate decay cycle	Learning rate attenuation multiple	BatchSize
Value	10^{-8}	22	0.2	33
Parameter	Training frequency	Dropout rate	α	
Value	154	0.6	0.6	

Under the same network conditions, the training effects of different loss functions are first compared. The training performance curves of different loss functions are shown in Figure 3. The network is trained using the loss function proposed in this paper and the traditional cross-entropy loss function respectively, and the training phase is iterated. It can be seen that the model trained with the loss function of this paper's method basically reaches the optimal iteration in about 30 simulation steps, and the final loss is 0.1154, while the traditional cross-entropy loss obtains the optimal iteration in about 50 simulation steps, and the final loss is 0.1367. Thus, it can be seen that the loss function of the proposed method makes the training model converge faster, and the performance of the model is more optimal.

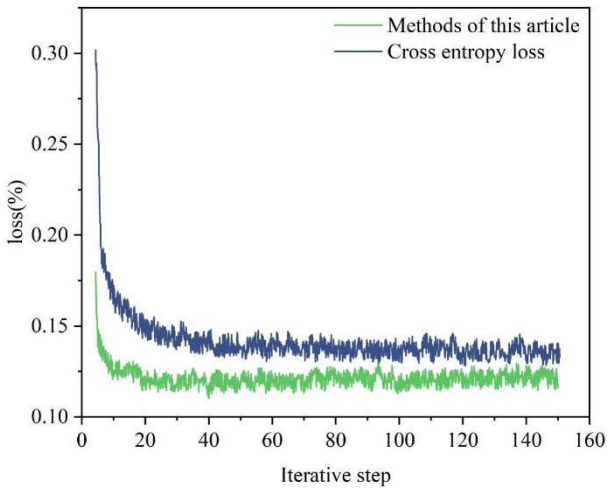


Figure 3: Simulation of different loss functions

The results of the accuracy comparison between the method proposed in this paper and the LSTM and GRU methods are shown in Fig. 4. It can be seen that the performance of the method proposed in this paper has been significantly improved, and the accuracy rate reaches 89.62%, which is higher than the accuracy rate of LSTM and GRU methods.

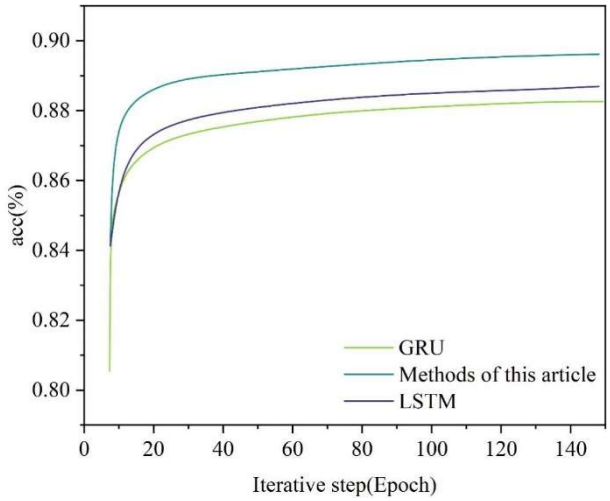


Figure 4: Comparison results of different methods of accuracy

In this subsection the experimental results of the generative AI based automatic scoring models are presented as shown in Fig. 5, which is the QWK scores of the different models on the eight topics of the ASAP dataset. The models presented in this section are represented by ChatGPT. Among the compared models, the classical RNN, CNN, LSTM and also the statistical based EASE model are selected in this paper.SKIPFLOW is a recently proposed neural network model obtained based on the modification of LSTM, which achieves more than all the previous methods. The results in the table show that the basic RNN model does not perform as well as the GRU and LSTM models, the reason for this is due to the fact that RNNs are weak in dealing with long term dependencies, whereas GRUs and LSTMs are able to obtain better results by designing their storage units specifically to address this problem.EASE is an automatic scoring system based on feature engineering, which is done by constructing handcrafted features, and feeding the features into a classifier for model training. In this paper, we have selected results from Support Vector Regression (SVR) and Bayesian Linear Regression (BLRR) methods for comparison. Compared with other models, ChatGPT achieved better results on five of the themes. For topic I, ChatGPT scores 0.874, which is 0.022 higher than the optimal SKIPFLOW model.For topic II, ChatGPT scores 0.748, which is 0.021 higher than the optimal LSTM model.For topic III, ChatGPT scores 0.717.For topic IV, ChatGPT scores 0.844, which is 0.844 higher than the optimal SKIPFLOW model by 0.026.For topic VI, the ChatGPT score is 0.850, which is 0.02

higher than the optimal SKIPFLOW model.Finally, the ChatGPT model proposed in this paper achieves a score of 0.7964 in terms of average QWK scores over the eight topics.

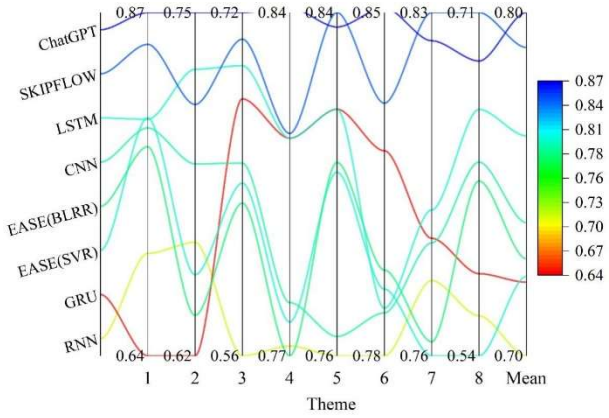
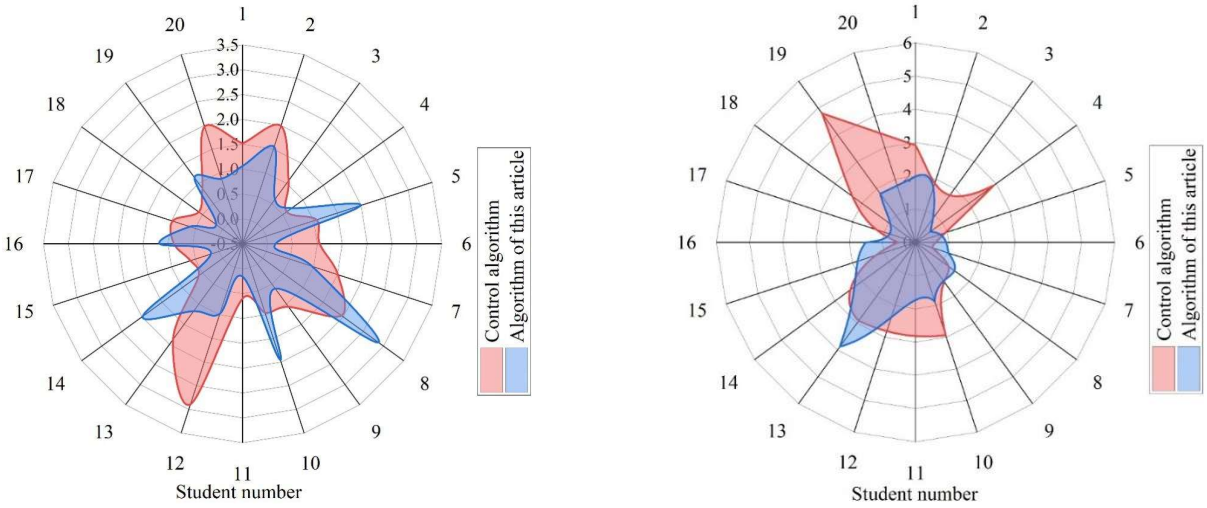


Figure 5: Results on the ASAP dataset

In order to verify the effectiveness of the module proposed in this paper, ablation experiments are conducted in this paper. The results of its experiments are shown in Table 2: It can be seen that the best results can be achieved by combining Transformer,scoring algorithm, and multi-head attention mechanism.

Table 2: Ablation study

Name	1	2	3	4	5	6	7	8	Mean
Transformer	0.832	0.701	0.695	0.788	0.789	0.788	0.784	0.684	0.757625
Scoring Algorithm	0.845	0.708	0.704	0.805	0.814	0.820	0.794	0.662	0.769
Multi-focus mechanism	0.857	0.711	0.700	0.806	0.810	0.836	0.794	0.678	0.774
Whole	0.855	0.721	0.714	0.822	0.814	0.830	0.795	0.682	0.779125



(a)Topic 1 (b)Topic 2

Figure 6: Score error contrast

III. B. Accuracy studies

The experimental data selected for this experiment were the assessment questions of the English course, which consisted of short-answer questions, of which a total of about 190 students answered each question. In addition,

two of the questions were selected for systematic testing, in which 20 student answers were selected for each question to test the automatic scoring based on the method. The test was conducted using the generative artificial intelligence-based automatic scoring algorithm proposed in this paper, the control algorithm was based on the Jaccard similarity algorithm, and the experimental evaluation was conducted using the size of the absolute value of the difference between the true scores of the students' answers and the scores predicted by the system as a measure. The results of the 20 samples tested in the two subjective questions are shown in Fig. 6. Compared with the control algorithm, the error between the predicted score and the true score of the algorithm proposed in this paper is relatively smaller compared to that of the control algorithm, and the model predicts a smaller error in the scores in the majority of the students' answers compared to the control algorithm. For example, the model prediction error is smaller than the control algorithm for 13 answers in question one, and the model prediction error is smaller than the control algorithm for 12 answers in question two.

In order to better measure the error between the scores predicted by the model and the actual scores, the average absolute error was used to measure the effectiveness of the model, and the results of the experiment using five questions are shown in Figure 7. Each question was answered by 190 students respectively using the control algorithm and the algorithm proposed in this paper for scoring comparisons, using the average absolute error for evaluation, this algorithm in the five test questions the average absolute error is smaller than the control algorithm, which can be seen that the scores predicted by the model are closer to the real scores in general.

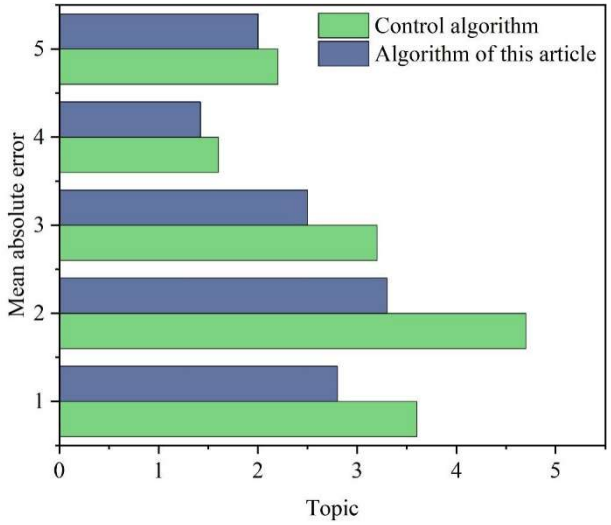


Figure 7: Mean absolute error

IV. Analysis of data from pre- and post-test questionnaires

The subjects of this study are the students of the third and sixth classes of sophomore English majors in a college in S city, totaling 100 students, and the two classes adopt the same lecturer and keep the same teaching method, teaching progress and teaching duration, in which the third class of sophomore English is taken as the experimental class and the sixth class of sophomore English is taken as the control class. A total of questionnaires were distributed: “Questionnaire on the Current State of English” and “Questionnaire on Perceptions and Attitudes toward the Feedback Method of the Automatic Grading System”. A total of 50 questionnaires were distributed, and at the end of the experiment, the same questionnaires as before the experiment were distributed again to the experimental class, “Questionnaire on College Students' Perceptions and Attitudes toward Automatic Grading Methods” with a total of 48 questionnaires. Using the questionnaire, the changes in English proficiency of the two classes were compared and analyzed under the influence of the research intervention.

The Cronbach's value of the scale is 0.763, which is greater than 0.7. From this, we can conclude that the measures of the research variables have high reliability and the survey data are reliable, i.e., the reliability is high. The Cranach's value reaches 0.897, which indicates that the overall measures have high reliability. In addition, we can also find that the reliability of the survey data is relatively high, which further proves that the questionnaire is highly reliable. The KMO value of 0.797 was obtained, which is greater than 0.7. The Sig value of Bartlett's sphere test is 0.000, which is less than 0.001 and reaches the level of significance, indicating that this questionnaire is set up and has high validity.

By collecting and organizing the Questionnaire on the Status of College Students' English Writing in the experimental and control classes, the results are shown in Table 3, and it can be found that the Sig value of each question of the questionnaire is greater than 0.05, which does not reach the level of significance. It means that the status of English writing between the experimental class and the control class does not have too obvious difference and is similar.Q1~Q15: I am more satisfied with the automatic scoring. The scores given are fair. More time-saving and labor-saving. More love for English. Increased awareness of revision. Revisions given are easy to understand. Helps to improve word accuracy. Reduces errors in writing specification. Reduces errors in vocabulary use. Reduces errors in grammar. Proficient use of complex sentence patterns. Revises according to automatic scoring. Will revise repeatedly according to feedback. Preferred. Looking forward to continued use.

Table 3: Questionnaire matching sample T test

	Class	N	M	SD	T	Sig.
Q1	Experiment class	50	3.72	1.054	0.132	0.916
	Control class	50	3.68	1.182		
Q2	Experiment class	50	4.56	0.506	0.411	0.679
	Control class	50	4.52	0.512		
Q3	Experiment class	50	2.04	1.064	-0.714	0.487
	Control class	50	2.22	1.185		
Q4	Experiment class	50	1.84	0.978	0.478	0.652
	Control class	50	1.73	0.984		
Q5	Experiment class	50	2.25	1.122	0.984	0.337
	Control class	50	1.98	1.014		
Q6	Experiment class	50	4.71	0.663	0.202	0.848
	Control class	50	4.69	0.722		
Q7	Experiment class	50	4.68	0.214	-1.526	0.153
	Control class	50	5.01	0.002		
Q8	Experiment class	50	5.11	0.284	-0.436	0.662
	Control class	50	4.94	0.245		
Q9	Experiment class	50	4.98	0.994	-0.514	0.628
	Control class	50	1.82	1.175		
Q10	Experiment class	50	1.97	0.485	-0.362	0.728
	Control class	50	4.69	0.479		
Q11	Experiment class	50	4.71	1.046	0.704	0.516
	Control class	50	3.34	1.136		
Q12	Experiment class	50	3.16	0.867	-0.758	0.467
	Control class	50	3.56	0.728		
Q13	Experiment class	50	3.68	0.504	1.036	0.316
	Control class	50	4.65	0.510		
Q14	Experiment class	50	4.23	0.472	-0.587	0.564
	Control class	50	4.25	0.463		
Q15	Experiment class	50	1.45	0.622	-0.035	0.986
	Control class	50	1.45	0.546		

The results of the paired t-test comparing the pre- and post-tests of the questionnaire survey of the experimental class are shown in Table 4, the corresponding significance Sig values of 0.001 in terms of students' views on automatic grading, students' attitudes toward the modifications given by automatic grading, and students' handling of the feedback from automatic grading are all less than 0.05, which has reached the level of significance, and the mean value of the post-tests of the questionnaire survey of the experimental class is significantly higher than that of the pre-tests. This shows that the attitude of the students in the experimental class towards automatic scoring has gradually changed.

Table 4: Match sample T test

		M	N	SD	T	Sig.
Pair1	The students measured the automatic scoring system	4.738	50	0.2472	19.24	0.001

	The students measured the automatic scoring system	3.025	50	0.6674		
Pair2	The students measured the attitude of the automatic grading system	4.712	50	0.2143	14.257	0.001
	The students measured the attitude of the automatic grading system	3.367	50	0.5894		
Pair3	The students measured the treatment of automatic scoring system feedback	4.879	50	0.1758	12.187	0.001
	The students measured the response of automatic scoring system feedback	3.417	50	0.8479		

V. Conclusion

The continuous development of big modeling and artificial intelligence technology has brought new challenges to English automatic scoring. In this paper, we design an automatic scoring method for college English based on ChatGPT fine-tuning model. It combines the natural language processing technology, scoring logic and algorithm of ChatGPT, so as to score accurately according to the scoring criteria. By comparing with other scoring algorithms, the following conclusions can be drawn:

(1) The loss function of the university English automatic scoring method based on ChatGPT fine-tuning model converges faster and has better performance. It is more accurate than LSTM and GRU methods. Taking the ASAP dataset as the dataset, the mean QWK score of the automatic scoring method proposed in this paper achieves a score of 0.7964 on 8 topics, which achieves better results.

(2) The mean absolute errors of this paper's methods are all smaller than those of the control algorithm, which are closer to the real scores.

(3) After the practice of English teaching, the views and attitudes of students in the experimental class towards automatic scoring are gradually getting better.

Funding

This work was supported by Hunan Province Social Science Achievements Evaluation Committee in 2025: Research on Constructing Personalized Teaching of College English with Generative Artificial Intelligence (NO: XSP25YBC560).

References

[1] Wan, Y., Wan, X., Huang, L., & Zhao, Y. (2025). Artificial Intelligence Empowered Reforms in Economics Education. International Journal of Computer Science and Information Technology, 5(1), 1-15.

[2] Ling, X. (2020, December). Research on University Education Reform in the Era of Artificial Intelligence. In 2020 International Conference on Information Science and Education (ICISE-IE) (pp. 561-564). IEEE.

[3] Sengar, S. S., Hasan, A. B., Kumar, S., & Carroll, F. (2024). Generative artificial intelligence: a systematic review and applications. Multimedia Tools and Applications, 1-40.

[4] Su, J., & Yang, W. (2023). Unlocking the power of ChatGPT: A framework for applying generative AI in education. ECNU Review of Education, 6(3), 355-366

[5] Tafazoli, D. (2024). Exploring the potential of generative AI in democratizing English language education. Computers and Education: Artificial Intelligence, 7, 100275.

[6] Vo, A., & Nguyen, H. (2024). Generative artificial intelligence and ChatGPT in language learning: EFL students' perceptions of technology acceptance. Journal of University Teaching and Learning Practice, 21(6), 199-218.

[7] Nash, B. L., Hicks, T., Garcia, M., Fassbender, W., Alvermann, D., Boutelier, S., ... & Young, C. (2023). Artificial intelligence in English education: Challenges and opportunities for teachers and teacher educators. English Education, 55(3), 201-206.

[8] Liao, H., Xiao, H., & Hu, B. (2023). Revolutionizing ESL teaching with generative artificial intelligence—Take ChatGPT as an example. International Journal of New Developments in Education, 5(20), 39-46.

[9] Esfandiari, S., & Ghamari, P. (2025). A Systematic Review of the Effects of Generative Artificial Intelligence Models (ChatGPT& Gemini) on English Language Teaching (ELT): Opportunities & Challenges. Journal of Foreign Language Research, 14(4), 611-641.

[10] Sharples, M. (2023). Towards social generative AI for education: theory, practices and ethics. Learning: Research and Practice, 9(2), 159-167.

[11] Burbano G, D. C., & Ibarra C, J. F. (2024, November). The Role of Generative Artificial Intelligence in Educational Innovation. In International Congress of Telematics and Computing (pp. 283-297). Cham: Springer Nature Switzerland.

[12] Cubillos, C., Mellado, R., Cabrera-Paniagua, D., & Urra, E. (2025). Generative Artificial Intelligence in Computer Programming: Does it enhance learning, motivation, and the learning environment?. IEEE Access.

[13] Lim, W. M., Gunasekara, A., Pallant, J. L., Pallant, J. I., & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or reformation? A paradoxical perspective from management educators. The international journal of management education, 21(2), 100790.

[14] Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI, 7(1), 52-62.

[15] Chiu, T. K. (2024). The impact of Generative AI (GenAI) on practices, policies and research direction in education: A case of ChatGPT and Midjourney. Interactive Learning Environments, 32(10), 6187-6203.

[16] Simelane, P. M., & Kittur, J. (2025). Use of Generative Artificial Intelligence in Teaching and Learning: Engineering Instructors' Perspectives. Computer Applications in Engineering Education, 33(1), e22813.

[17] Monzon, N., & Hays, F. A. (2025). Leveraging Generative Artificial Intelligence to Improve Motivation and Retrieval in Higher Education Learners. JMIR Medical Education, 11(1), e59210.

[18] Li, Y., Ji, W., Liu, J., & Li, W. (2024, March). Application of Generative Artificial Intelligence Technology in Customized Learning Path Design: A New Strategy for Higher Education. In 2024 International Conference on Interactive Intelligent Systems and Techniques (IIST) (pp. 567-573). IEEE.
[19] Yan, L., Greiff, S., Teuber, Z., & Gašević, D. (2024). Promises and challenges of generative artificial intelligence for human learning. Nature Human Behaviour, 8(10), 1839-1850.
[20] Kapterev, A. I. (2023). Challenges of generative artificial intelligence for the higher education system. RUDN Journal of Informatization in Education, 20(3), 255-264.
[21] Samala, A. D., Rawas, S., Wang, T., Reed, J. M., Kim, J., Howard, N. J., & Ertz, M. (2025). Unveiling the landscape of generative artificial intelligence in education: a comprehensive taxonomy of applications, challenges, and future prospects. Education and Information Technologies, 30(3), 3239-3278.
[22] Verma, G., Campbell, T., Melville, W., & Park, B. Y. (2023). Navigating opportunities and challenges of artificial intelligence: ChatGPT and generative models in science teacher education. Journal of Science Teacher Education, 34(8), 793-798.
[23] Jacob S Berkowitz,Apoorva Srinivasan,Jose Miguel Acitores Cortina,Yasaman Fatapour & Nicholas P Tatonetti. (2025) .Biomedical Text Normalization through Generative Modeling. .medRxiv : the preprint server for health sciences.
[24] Shengyang Cheng,Jianyong Huang,Xiaodong Wang,Lei Huang & Zhiqiang Wei. (2025) .Image–text aggregation for open-vocabulary semantic segmentation. Neurocomputing,630,129702-129702.
[25] Hong WeiTing,Whyte Jennifer & Xue Jin. (2025). A Natural Language Processing–Driven Framework for Policymaking in Infrastructure Development. Journal of Construction Engineering and Management,151(5).
[26] Shuai Wang,Zechen Zheng,Kuan Diao & Xiaojun Liu. (2025) .Enhancing resolution uniformity of spectral line confocal 3D sensors through integration of vision transformer with perpendicular attention-augmented parallel convolutional neural networks. Measurement,249,116980-116980.
[27] Shanshan Ding,Renwen Chen,Hao Liu,Fei Liu & Junyi Zhang. (2024) .ASG-HOMGAT: a high-order multi-head graph attention network with adaptive small graph structure for rolling bearing fault diagnosis. Measurement Science and Technology,35(6).
[28] Haitao Wang,Yongyang Xu,Anna Hu,Xuejing Xie,Siqiong Chen & Zhong Xie. (2025) .Building pattern recognition by using an edge-attention multi-head graph convolutional network. International Journal of Geographical Information Science,39(4),732-757.
[29] Michael McGuire & Jenifer Larson Hall. (2025) .Assessing Whisper automatic speech recognition and WER scoring for elicited imitation: Steps toward automation. Research Methods in Applied Linguistics,4(1),100197-100197.