Empirical Research on High School English Composition Evaluation via Large Language Models

Siyu Liu
School of Foreign Languages, Northwest Minzu University, Lanzhou, Gansu, China

Abstract: In recent years, artificial intelligence technology has developed rapidly across various domains, with large language models being increasingly widely applied in the field of education. Traditional evaluation of high school English compositions is plagued by issues such as prolonged teacher feedback cycles, strong subjectivity, and insufficient personalized guidance. While some schools have adopted large language models to assist in composition grading, challenges still persist, making it particularly important to test and investigate the evaluation capabilities of these models. To examine the English composition evaluation abilities of large language models, this study selected 100 high school English compositions, designed 20 prompts, and employed two domestic large language models, DeepSeek and Wenxin Large Model, as testing tools to conduct a comprehensive assessment of their capabilities in both scoring and correction. Finally, this paper proposes recommendations for the application of large language models in English education, offering practical foundations for integrating intelligent educational tools with traditional teaching practices.

Keywords: Large Language Model; Senior High School English; Human-Machine Collaboration

1. Introduction

In the wave of digital educational transformation, artificial intelligence technology is profoundly reconstructing traditional teaching practices scenarios. As a core component of language proficiency cultivation, high school English writing instruction not only serves as a crucial method to assess students' comprehensive language abilities through language output but also acts as a key pathway for nurturing logical thinking and innovative expressive skills. Intelligent large language models boast strong language proficiency. Leveraging natural language processing and machine learning technologies, they can rapidly and accurately generate feedback tailored to the needs of teachers and students. The application of large language models can eliminate delays in student learning, precisely pinpoint students' writing weaknesses, enhance expressive skills, provide personalized tutoring, assist teachers in improving grading efficiency, enable self-correction, and boost teaching effectiveness. At the beginning of 2025, DeepSeek broke through technological circles and reached the general public. As a domestic open-source AI large model, DeepSeek is reconstructing the educational landscape and offering innovative solutions to address the challenges in writing instruction, leveraging its robust logical reasoning capabilities, low deployment costs, and localization advantages. Digital tools significantly enhance the role of personalized instruction in high school English writing, improving students' writing proficiency and enriching teaching content through multi-stage application and innovative pedagogical strategies. [1] Given these advantages, an increasing number of schools, teachers, and students are adopting large language models, yet the educational evaluation capabilities of these models require experimental investigation within specific disciplines. To examine the English composition evaluation abilities of large language models, this study selected 100 high school English essays of varying proficiency levels, designed 20 prompts, and employed two domestic large language models DeepSeek and Wenxin Large Model as testing tools to conduct a comprehensive assessment of their capabilities in scoring and feedback provision.

2. Literature Review

The PuTong Gaozhong YingYu KeCheng BiaoZhun[2] promulgated by the Ministry of Education specifies requirements for students' pragmatic knowledge, mandating that students master the selection of appropriate and proper language forms to communicate and respond to others' viewpoints in written communication. Compared with the previous outline, the new standards have raised the requirements for English writing skills, quantifying them into nine proficiency levels with distinct target objectives; the college entrance examination requires achievement of the eighth-level target. Students are expected to paraphrase or summarize read articles, compose short essays or reports based on information provided in text and charts, produce coherent and structurally complete essays to narrate events or express viewpoints and attitudes, and ensure their writing adheres to stylistic norms with grammatically smooth sentences. The evaluation of high school English compositions can be divided into two aspects: scoring and correction. This study focuses on testing the composition evaluation ability of large language models. According to the findings of Wei Shunping, Zhang Yue et al. (2025) [3], large language models perform excellently in grammatical error detection, semantic understanding, and text structure analysis. They can accurately identify and correct errors in compositions while providing targeted improvement suggestions. Research has already been conducted on large language models in the field of composition correction. Wang Fang (2025) pointed out in her article that by integrating DeepSeek technology, teachers can automate and intelligentize composition correction, thereby improving teaching efficiency and quality. At the same time, DeepSeek can also provide personalized guidance for students' learning, helping them better master English writing skills. Automatic assessment technology enabled by large language models has become an inevitable trend[4]. In her paper, Wu Jiahui (2024) conducted an in-depth exploration of the application of automatic assessment systems in English writing learning. The paper focused on key technologies underpinning automatic assessment systems, including natural language processing, machine learning, and deep learning, and designed and implemented an automatic assessment system based on these technologies. [5] For instance, Ma Lijun (2021) pointed out in her article that the application of the "Intelligent Correction" system in high school English composition teaching has achieved remarkable results, as it not only improves correction efficiency but also promotes students' personalized learning. [6]

Based on a comprehensive review of relevant literature, artificial intelligence technology demonstrates significant advantages in composition evaluation and correction. It enhances teaching efficiency and quality while promoting students' writing development. However, existing research also points out that although AIGC technology has a positive impact on students' vocabulary complexity, syntactic complexity and fluency, it has little impact on their accuracy, indicating that the technology may not be able to completely replace educators in terms of ensuring language accuracy.[7] Future research can further explore how to integrate artificial intelligence technology with teachers' professional judgment to achieve more comprehensive and accurate composition evaluation. Additionally, investigations can be conducted into how artificial intelligence technology can be utilized to provide personalized learning support, thereby catering to the diverse learning needs of students.

3. Research Process

Large language models are trained on massive datasets, enabling them to scientifically master disciplinary knowledge. Through natural language processing technology and machine learning algorithms, they can generate feedback quickly and accurately based on the needs of teachers and students, a capability that facilitates their practical application in educational evaluation. To investigate the English composition evaluation abilities of large language models, this study selected 100 high school English essays of varying proficiency levels, designed 20 prompts, and employed two domestic large language models –DeepSeek and Wenxin Large Model – as testing tools to conduct a comprehensive evaluation of their abilities in both scoring and correction.

3.1 Selection of Test Samples

This study selected 100 high school English practical writings of different proficiency levels. Statistics showed an average score of 13.13 with a standard deviation of 2.07. For testing large language models capabilities, 45 high-level essays (scoring 13 points and above), 26 medium-to-low-level essays (scoring 4-12 points), and 29 full-score college entrance examination essays were included. In terms of composition types, all 100 compositions are practical writings, which can better reflect students' understanding and application of the target topic compared with continuation writing.

3.2 Design of Prompts

Prompts are the core control elements for the output of large language models, directly influencing the direction, quality, style, and accuracy of the content generated by the model. The selection of elements for prompts depends on the task that the user intends large language models to accomplish. This study designed 20 prompts from five perspectives: content key points, grammatical structure, vocabulary usage, contextual coherence, and spelling accuracy. To better assist users from different groups, including students, teachers, and examiners, this study designed 6 scenarios where different users have varying requirements for composition correction and tested large language models. For instance, from the examiner's perspective, which demands strict correction, large language models can be utilized to train correction and scoring skills, enabling fast and high-quality scoring.

Table 1. Types and Examples of Prompt Phrases

Prompt Type	Prompt Example		
Scoring Type		q1	This essay should be comprehensively scored from the perspectives of content points, grammatical structure, vocabulary usage, contextual coherence, spelling accuracy, etc., with a full score of 15 points.
		q5	This essay should be graded from the perspectives of content points, grammatical structure, vocabulary usage, contextual coherence, spelling accuracy, etc., with the highest grade being 5.
Commentary Type	Evaluation and Revision	q8	The essay should be evaluated in terms of content points and revision suggestions should be provided.
		q9	The essay should be evaluated in terms of grammatical structure and revision suggestions should be provided.
		q10	The essay should be evaluated in terms of vocabulary usage and revision suggestions should be provided.
		q11	The essay should be evaluated in terms of contextual coherence and revision suggestions should be provided.
		q12	The essay should be evaluated in terms of spelling accuracy and revision suggestions should be provided.
	User Perspective	q15	From the perspective of students' self-improvement, this essay should be scored and corrected, with grammatical errors identified and revised.
		q18	From the perspective of examiners' marking, this essay should be scored with strict adherence to assessment criteria.
		q20	From the perspective of teachers' revision, this essay should be scored and corrected, with grammatical errors identified and revised; encouraging elements should be included, and scores may be awarded appropriately.

4. Calling Large Language Models

For the invocation of large language models, this study first utilized Python to call the API interface of DeepSeek (DeepSeek-V2 employs the Chat Completions API version, with parameters including stream and temperature). A batch questioning test was conducted on 100 compositions in the sample by inputting 20 prompts, resulting in a total of 2000 feedback entries. To examine the evaluation stability of DeepSeek, a second round of testing was performed on 1 scoring prompt (q1) and 1 comment prompt (q8), yielding 200 feedback entries. To assess the score correlation of domestic large language models, Python was again employed to call the API interface of Wenxin Large Model (The API

version is ERNIE 4.0, with parameters being prompt and temperature). The first round of testing was conducted on 2 scoring prompts (q1, q5), generating 200 feedback entries. Upon completion of all tests, a total of 2400 composition evaluation feedback entries related to large language models were collected and organized.

5. Data Analysis

This study employed domestic large language models to conduct two types of tests on essay samples from the high school English essay test set: scoring and correction.

a. For the scoring test, the correlation and stability of the large language models' essay scores were analyzed. Specifically, regarding scoring correlation, SPSS 25 was used to examine the consistency between the scores assigned by the domestic large language models and the original scores. For scoring stability, SPSS 25 was utilized to analyze the consistency between two sets of scores generated by the domestic large language models.

b. For the correction test, the accuracy and stability of the essay comments produced by the domestic large language models were analyzed. To assess comment accuracy, manual sampling of the corrections was performed for verification. To analyze comment stability, prompts were input multiple times, and data was collected to calculate the similarity of the large language models' comments.

6. Score Analysis

6.1 Correlation of Large Language Models Scores

To examine the correlation of large language models scores, this study employed DeepSeek and Wenxin Large Model to evaluate 100 compositions. Specifically, DeepSeek and Wenxin Large Model provided numerical scores (out of 15) and grade ratings (on a 5-grade scale), respectively. Correlation analyses were then conducted between the scores/grades generated by these two large language models and the original human-assigned scores/grades, with Pearson correlation coefficients used to measure the strength of associations. The results are presented in Table 2.

Table 2 shows that for DeepSeek, the Pearson correlation coefficient between its numerical scores and the original scores is r1= 0.67, while that between its grade ratings and the original scores is r2= 0.236. These findings indicate a moderate correlation between DeepSeek's scores/grades and the original scores, suggesting no significant difference between DeepSeek scores and human scores. For Wenxin Large Model, the Pearson correlation coefficient between its numerical scores and the original scores is r3=0.748, and between its grade ratings and the original scores is r4 = 0.609. This indicates a moderate correlation between Wenxin Large Model's evaluations and the original scores, thus revealing no significant difference between Wenxin Large Model scores and human scores.

Table 2. Correlation Analysis of Large Language Models Scores and Original Scores

Scoring Method	Testing Tool	Minimum	Maximum	Average	Standard Deviation	Correlation
Score Evaluation (Full Marks 15)	DeepSeek	5	15	12.075	2.2	0.670
	Wenxin Large Model	4	15	12.055	2.49	0.748
Grade Evaluation (Full Fifth Grade)	DeepSeek	Second Grade	Fifth Grade	4.34	0.92	0.236
	Wenxin Large Model	Second Grade	Fifth Grade	4.52	0.90	0.609

This study also categorized the original scores of 100 essays into groups: 45 high-level essays (scoring 13 points and above), 26 medium-to-low-level essays (scoring 4-12 points), and 29 full-score essays from the college entrance examination. It then conducted a correlation analysis between the scores assigned by two domestic large language models —DeepSeek and Wenxin Large Model—and the original scores for the high-level (including full-score essays) and medium-to-low-level groups. The results are presented in Table 3. Table 3 indicates that when scoring high-level essays, both large language models demonstrated a moderate correlation with the original scores. Specifically, the Pearson correlation coefficient between DeepSeek scores and the original scores is r5=0.511, while that for Wenxin Large Model is r6=0.513. For medium-to-low-level essays, the scores from both domestic large language models also showed a moderate correlation with the original scores. Here, the Pearson correlation

coefficient between DeepSeek scores and the original scores is r7=0.593, and for Wenxin Large Model, it is r8=0.910. It can be seen that the Wenxin Large Model shows the highest correlation between its scores for low-group compositions and the original scores, indicating that it provides the most effective scoring feedback for low-group compositions.

Table 3. Correlation Analysis of the Scores of Compositions in High, Medium, and Low Groups with Original Scores by large language models

Composition Grouping	Assessment Tool	Minimum	Maximum	Average	Standard Deviation	Correlation
High-scoring composition	DeepSeek	6	15	12.59	1.97	0.511
	Wenxin Large Model	7	15	12.52	1.53	0.513
Low-to-medium-scoring composition	DeepSeek	5	14	9.38	2.46	0.593
	Wenxin Large Mode	4	13	9.63	2,15	0.910

6.2 Stability of Scoring

To examine the stability of essay scoring by large language models, this study employed two large language models, namely DeepSeek and Wenxin Large Model, to conduct two rounds of score and grade evaluations on 100 essays each. SPSS 25 was used to analyze the essay scoring data. The descriptive statistics for DeepSeek's score evaluations are presented in Table 4. As shown, there is a significant difference between the average of DeepSeek's two scores and the original scores, while the standard deviation of its two scores exceeded that of the original scores, indicating greater score dispersion in DeepSeek's evaluations compared to the original scores. Furthermore, the maximum score difference between DeepSeek's two evaluations reached 5 points, spanning two grade levels, demonstrating instability in DeepSeek's score assessment of the same essay.

Table 4. Comparative Analysis of DeepSeek Scores and Original Scores

Scoring Method	Minimum	Maximum	Average	Standard Deviation	Correlation
Original Score	4	15	13.13	2.07	
DeepSeek First Scoring	6	15	12.42	2.20	0.735
DeepSeek Second Scoring	5	15	11.73	2.49	0.629

This study continues to apply the large language models Wenxin Large Model to conduct two grading tests on 100 compositions, and uses SPSS 25 to analyze the test data of the compositions. The descriptive statistical results of the Wenxin Large Model score evaluation are shown in Table 5. It can be seen that the difference between the average value of the two scores given by Wenxin Large Model and the original scores is small, while the standard deviation of the two gradings by Wenxin Large Model is greater than that of the original gradings, indicating that the dispersion degree of the Wenxin Large Model gradings is higher than that of the original gradings. In addition, the two gradings of Wenxin Large Model are basically the same, with a maximum span of one grade, which shows that the Wenxin Large Model demonstrates relatively stable grading for the same composition.

Table 5. Comparison and Analysis of Wenxin Large Model Rating and Original Rating

Scoring Method	Minimum	Maximum	Average	Standard Deviation	Correlation
Original Rating Grade	Second Grade	Fifth Grade	4.65	0.64	
Wenxin Large Model First Rating Grade	Second Grade	Fifth Grade	4.3	0.91	0.7
Wenxin Large Model Second Rating Grade	Second Grade	Fifth Grade	4.43	0.89	0.524

6.3 Correction Scores of Large Language Models

To analyze the accuracy of corrections, prompts were input into large language models to generate correction outputs. These outputs were then compared with original manual corrections across five dimensions: content key points, grammatical structure, vocabulary usage, contextual coherence, and spelling accuracy, to identify similarities and differences in their comments. Figures 1 and 2 present a manually corrected practical writing sample and the corresponding correction result from large language models. Through multi-angle analysis, tests revealed a high correlation between the composition scores assigned by large language models and those

from manual corrections.

6.4 Overall Testing and Evaluation of Large Language Models

To analyze the accuracy of corrections, prompt content is input into large language models to generate corrected outputs. These outputs are then compared with original manual corrections across five dimensions: content key points, grammatical structure, vocabulary usage, contextual coherence, and spelling accuracy, identifying similarities and differences in their feedback. Below, a systematic analysis will be conducted using a specific article, comparing the original revisions with the suggestions provided by DeepSeek and Wenxin Large Model. Through multi-angle analysis of the feedback on the same composition, it is observed that large language models can quickly identify basic errors in compositions, such as incorrect verb phrase collocations and spelling mistakes. However, teachers, with their richer grading experience, are better able to quickly assess whether students' compositions address the topic. The scores assigned to compositions by large language models are highly correlated with those revised by human teachers, enabling educators to enhance grading efficiency and pinpoint errors that might otherwise be overlooked during manual assessment.

Our class has planned to organize an activity called agritourism project from 7.a.m to 4.p.m this Sunday. Knowing that you are fascinated by farms, I sincerely invite you to participate in our activity.

As scheduled, we'll gather together at the school gate at 7.a.m. And a medium bus will take us to the farm. The moment we arrive we will be divided into 2 groups with one group picking vegetables like tomatoes and another picking tempting apples and then devote ourselves to picking tempting apples and fresh vegetables including peas and carrots. Dangerous action beans like chasing after others must be avoided not only to protect our safety, but to avoid hurting growing plants. We'll have an orge a totally green and healthy meal on the farm. In the afternoon, we plan to plant sunflowers. What should be kept in mind is that each step ranging from digging to seating must be carefully taken to ensure a good environment for seeds.

Since the project will teach us knowledge about nature and the importance of food, I strongly hope that you'll come. Looking forward to your reply!

In the original correction, the teacher awarded a score of 9, noting the following issues: the word "together" following "gather" in the second paragraph was redundant; "a medium bus took us there" was revised to "we took a bus"; "the moment we arrived" was revised to "upon arrival"; "devote into" was corrected to "devote to"; there was a spelling error with "peans"; and the sentence "What should be kept in mind is that each step ranging from digging to seating must be carefully taken to ensure a good environment for seeds" was deemed by the teacher to have an awkward structure. The summary included: 1. The key points of the topic should be elaborated separately, as the composition structure is somewhat vague. 2. You should first delete some content yourself and outline before writing; otherwise, writing while formulating ideas will lead to a disorganized and lengthy composition. 3. More words and model essays should be memorized after class.

DeepSeek assigned a score of 7, identifying the following issues: "with one group picking... and then devote ourselves" (subject missing) → should be revised to "and we will devote"; preposition error: "devote ourselves into picking" → "to"; non-finite verb misuse: "Dangerous action beans like chasing" → "actions such as". High-frequency spelling errors: sheduled → scheduled • peans → peas • orge → organic • seads → seeds • uneasy (adjective) → uneasiness (noun)

Grammatical omission: "have a totally green meal" (missing article)

DeepSeekOptimized version (meeting the fifth-band standard of the college entrance examination)

Our class will host an agritourism project this Sunday (7 a.m.-4 p.m.). As you're passionate about farms, we sincerely invite you!

We'll assemble at the school gate at 7 a.m. and take a medium bus to the farm. Upon arrival, we'll split into two groups: one will harvest vegetables (tomatoes, peas, carrots), and the other will pick tempting apples. Remember to avoid dangerous behaviors like chasing to protect both ourselves and the crops. A totally organic, healthy lunch will be provided.

In the afternoon, we'll plant sunflowers. Every

step—from digging to seeding—must be done carefully to ensure optimal growth conditions for the seeds. This project will teach us about nature and the hardships of food production—we truly hope you can join us !Summary and Comments: This article meets the third-grade standard (7-9 points) for the following reasons:

1. The content is comprehensive with no omissions.
2. Advanced vocabulary has been attempted (e.g., agritourism, devote, tempting).
3. Complex sentences have been actively utilized (e.g., "What should be kept in mind is that...").

The main reasons for score deduction: Four terminology spelling errors and two critical grammatical mistakes have lowered the scores in vocabulary and structure.

Suggestions: Prior to the exam, special training should be conducted on agricultural-themed vocabulary (e.g., cultivate/seeding/harvest) and participle structure grammar.

Wenxin Large Model has awarded a score of 7 points.

Detailed Analysis:

Content Points (2 points, scored 1 point)

• Strengths: Key information is covered, including the activity time (7 a.m. to 4 p.m. this Sunday), meeting location (school gate), transportation (minibus), activity content (group-based picking of vegetables and fruits, avoiding dangerous behaviors, enjoying green and healthy meals, and planting sunflowers in the afternoon), and activity significance (understanding nature and the hardship involved in food production). The coverage is generally complete. • Shortcomings: Some content is presented in a confusing manner. For example, the section on picking activities lacks logical clarity. It first mentions dividing participants into two groups—one picking tomatoes and the other picking apples—then later refers to "devoting ourselves to picking apples and fresh vegetables including peanuts and carrots." This results in repetitive and disorganized content, leading to insufficient clarity and accuracy in presenting key points. 1 point deducted.

Grammatical Structure (3 points, scored 1.5 points)

• Strengths: The overall sentence structure shows a basic foundation, with the ability to use common sentence patterns to express activity arrangements and other information.

• Shortcomings: There are numerous grammatical errors. For instance, the sentence "The moment we arrive we will be divided into 2 groups with one group picking vegetables like tomatoes and another picking tempting apples and then devote ourselves to picking tempting apples and fresh vegetables including peas and carrots" has a disorganized structure. The phrase "devote ourselves into" is incorrect and should be "devote ourselves to"; "peans" contains a spelling error and should be "peanuts". Additionally, in the sentence "What should be kept in mind is that each step ranging from digging to seating must be carefully taken to ensure a good environment for seeds", there is a spelling error in "seads", which should be "seeds", and "seating" is inappropriately used. Based on the context, which intends to convey a meaning related to "sowing", "sowing" is more appropriate. These grammatical issues affect the accuracy and fluency of the sentence, resulting in a deduction of 1.5 points.

Vocabulary Usage (3 points, scored 1.5 points)

• Strengths: Some activity-related vocabulary is used, such as "agritourism project", "vegetables", "apples", "sunflowers", etc., which can basically convey the content of the activity.

• Weaknesses: The vocabulary used is not sufficiently accurate or rich. For example, while "tempting" can describe apples, the overall vocabulary is relatively ordinary, lacking more expressive words to enhance the quality of the article. Spelling errors such as "peans" and "seads" are also present, resulting in a deduction of 1.5 points.

Coherence of Context (3 points, scored 1.5 points)

• Strengths: Conjunctions like "As scheduled" (should be "As scheduled") and "The moment" are used to connect sentences, which to some extent reflects the logical relationships between them.

• Weaknesses: Insufficient contextual coherence. There are logical gaps between certain sentences; for instance, when describing picking activities, the transitions between sentences lack natural flow, making the content feel disjointed. Additionally, transitions between paragraphs are not clearly

marked, leading to a deduction of 1.5 points.
Spelling Errors (4 points, scored 1 point)
• Weaknesses: The text contains numerous spelling errors, including "peans", "seads", and "seating" (this word usage error is similar to a spelling issue that hinders comprehension). These errors severely compromise the accuracy and professionalism of the text, resulting in a deduction of 3 points.

7. Conclusions and Suggestions

7.1 Advantages of large language models in Correcting High School English Compositions.

In this study, two large language models, DeepSeek and Wenxin Large Model, were employed to evaluate the scores and provide comments on 100 compositions respectively. Regarding scoring: In terms of correlation, the tests revealed that both DeepSeek and Wenxin Large Model scores demonstrated a moderate correlation with the original human scores, with relatively insignificant differences between their scores and the human ratings. Wenxin Large Model shows that there is a high correlation (r = 0.910) for articles at medium and low levels, but DeepSeek has a lower level rating (r = 0.236). In terms of stability, the tests indicated that DeepSeek exhibited good stability in both scoring and rating, whereas Wenxin Large Model showed poor stability in these aspects.
For comments and corrections, DeepSeek's comments were relatively concise and clear, while those from Wenxin Large Model were more detailed. From the stability perspective, the similarity between the comments generated by DeepSeek and Wenxin Large Model for the same composition was found to be low. Overall, large language models are helpful for assisting in high school essay correction. Through extensive testing and prompt refinement, they can be made more intelligent, thereby better facilitating human-machine collaboration.
Test research has revealed that the two large language models exhibit a strong correlation in essay correction and can identify the same error points as manual correction; they are capable of leveraging reinforcement learning techniques to adjust their suggestion strategies based on prior writing data. For instance, when the same essay undergoes multiple rounds of correction, each iteration yields more detailed revision proposals than the previous one, enabling students to further refine their essays; they can deliver personalized scoring and tailored suggestions according to the specific needs of different prompt recipients, aiding various groups in enhancing learning efficiency. Huang Jinchun (2025) pointed out that high school students' English writing abilities see notable improvements with AI technology support. These include enhanced precision in language expression, expanded access to diverse writing materials, stimulated innovative writing thinking, and strengthened autonomous learning effectiveness.[8] Therefore, integrating AI technology appropriately into teaching can promote the development and innovation of English education, helping to cultivate students with an international perspective and cross-cultural communication skills.

7.2 Suggestions

With the rapid development of artificial intelligence technology, the vision of "future education" is gradually being realized. Large models have become capable of accompanying the entire educational process, and human-machine collaboration has emerged as an inevitable trend. Therefore, based on empirical research into large language models evaluation of high school English compositions, the following suggestions are proposed:
During the pre-writing phase, a writing resource library can be established. large language models can be utilized to gather and organize English writing materials across various topics. Furthermore, these models can analyze students' English proficiency and writing interests to design personalized writing tasks for different proficiency levels, thereby motivating students' writing enthusiasm. After finishing their compositions, students may submit their work to large language models for feedback and correction. Large language models do not point out issues such as grammatical errors and spelling mistakes; instead, they evaluate aspects like the structure, logic, and content of the article and provide detailed revision suggestions.
Large language models primarily process

language based on statistical patterns and pattern matching. They struggle with understanding complex semantics, fail to offer targeted correction advice, and require teachers to make revisions independently. Furthermore, large language models lack human emotional and value judgment capabilities. Composition is not merely a combination of words but, more importantly, an expression of emotions. Human emotions are rich, diverse, subtle, and complex, making it difficult for large language models to perceive and accurately evaluate the emotions conveyed in compositions as humans do. After large language models complete their initial revisions, a secondary manual review should be conducted. Drawing on their extensive linguistic expertise and teaching experience, teachers review the model's feedback, correct any potential errors, and conduct in-depth assessments of the composition's semantic coherence, emotional expression, and values. This process yields more comprehensive and accurate feedback, facilitating efficiency.

References

[1] Ruan Fenghui. Personalized Teaching Path of High School English Writing Assisted by Digital Tools [J]. Campus English, 2024, (50): 43 - 45

[2] Formulated by the Ministry of Education of the People's Republic of China, the General high school English curriculum Standard (2017 Edition, Revised in 2020) [S] was published by the People's Education Press in Beijing in 2020.

[3] Wei Shunping, Zhang Yue, Ran Rou. Testing the Chinese Essay Assessment Capabilities of Domestic large language models [J]. Modern Educational Technology, 2025, 35(03): 24-33.

[4] Xia Zhiting. A Study on the Effects and Impacts of Senior High School English Essay Correction [J]. Overseas English, 2021, (17): 89-90+92.

[5] Wu Jiahui. Research on Key Technologies of an Automatic Evaluation System for English Writing Learning [D]. Beijing University of Posts and Telecommunications, 2024.

[6] Wang Fang. Research on Strategies for DeepSeek to Empower high school English teaching [N]. Shanxi Science and Technology News, 2025-03-18(A06).

[7] Liu Zhenghui, Zhao Xiaoyan, Ruan Libin. Empirical Research on the Effect of the Generative AI - Enhanced English Teaching Method Course from the Students' Perspective[J]. Journal of Shanxi Institute of Energy, 2025, 38(03): 96 - 99.

[8] Huang Jinchun. A Study on Cultivating Senior High School Students' English Writing Ability Supported by AI Technology[J]. Education Circle, 2025, (23): 41-43

