

A Case Study of AI Agent-Assisted Correction of Intermediate-Level Chinese Writing

Yilan Wu* Phacharaphon Chuenchit** Jiaxin Lin*** Phantira Pangseranee****
and Amika Wongjuntr*****

Department of Global Engagement, Central China Normal University*,***

Faculty of Education, Vongchavalitkul University**,****

Independent Scholar*****

Corresponding author E-mail: phacharaphon_chu@vu.ac.th

(Received: March 4, 2026; Revised: April 8, 2026; Accepted: May 20, 2026)

Abstract

This study aimed to investigate the effectiveness and applicability of an Artificial Intelligence (AI) Agent in assessing intermediate-level Chinese writing produced by learners of Chinese as a second language (L2). The research addressed two primary objectives: first, to examine the degree of consistency between AI Agent scoring and manual scoring conducted by professional teachers, and second, to identify advantages and limitations of using an AI Agent for essay correction in international Chinese language education. The research population consisted of international students in Chinese language programs at Central China Normal University (CCNU). Purposive sampling was employed to collect 31 final examination essays from students taking the Intermediate Chinese Writing II course, comprising 16 essays from the first semester and 15 from the second. All participants passed the HSK Level 4 and represented diverse cultural and linguistic backgrounds. The primary instrument was an AI Agent named “Essay Grading Assistant,” developed on the ByteDance Coze platform using the DeepSeek-V3.2 large language model. This system integrated Optical Character Recognition (OCR), Chinese word segmentation, and document analysis plugins to process handwritten examination scripts. Additionally, two university instructors with over five years of experience independently evaluated the essays using the same analytical scoring rubric.

Data analysis involved descriptive statistics to examine score distributions, Spearman’s rank correlation coefficient to determine the relationship between AI-generated scores and human ratings, and the Intraclass Correlation Coefficient (ICC) using a two-way random effects model with absolute agreement to evaluate inter-rater consistency. The findings revealed a very strong positive correlation between the AI Agent scores and the teachers’ scores, with correlation coefficients of 0.917 and 0.902, respectively ($p < .001$). The ICC value reached 0.964, indicating excellent agreement. The AI Agent significantly improved grading efficiency, requiring approximately 15 seconds per essay, while enhancing scoring objectivity and consistency in evaluating surface-level linguistic features. However, several limitations were identified, particularly in deep semantic interpretation, pragmatic understanding, and the recognition of unconventional learner expressions. The study concluded that AI Agents possess strong potential as supportive tools for international Chinese writing assessment; nevertheless, a human–AI

collaborative scoring approach remains the most appropriate model for ensuring both efficiency and pedagogical quality in second-language writing evaluation.

Keywords: International Chinese Education, Chinese L2 Writing, Automated Scoring, AI Agent

1. Introduction

The State Council's "New Generation Artificial Intelligence Development Plan" explicitly mandates utilizing intelligent technology to reform teaching models and accelerate interactive learning systems, setting the direction for the transformation of International Chinese Education. Cai (2024) notes that 2024 multimodal vision large models, represented by GPT-4o, have pushed the boundaries of multi-sensory human-AI interaction. Applications are evolving from instruction-following "AI Copilots" toward "AI Agents"—intelligent entities integrating Large Language Models (LLM), Planning, Memory, and Tools to achieve autonomous planning and superior task-execution capabilities. This provides new tools for technology-enabled Chinese teaching. Writing is vital to a learner's linguistic competence; however, traditional manual correction is time-consuming, labor-intensive, and subjective, making it difficult to balance multiple scoring dimensions simultaneously. Utilizing technology to assist manual grading and improve scoring efficiency and consistency is now a primary research focus.

Automated Essay Scoring (AES) is one such approach, improving efficiency by extracting shallow textual features—like word count, sentence length, and basic grammatical rules—to evaluate writing via algorithms. In China, English and native Chinese AES systems are maturing, but research tailored for international Chinese essays remains scarce. In existing explorations, Ren (2004) pioneered objectively applying HSK criteria to AES using regression equations to extract key features across genres. Subsequently, Huang et al. (2014) expanded the linguistic feature set via a linear regression model, achieving 84.5% peak accuracy on same-topic essays. Zhang et al. (2020) designed a scoring system for Chinese learners based on four indicators: vocabulary level, topic relevance, sentence pattern difficulty, and expression accuracy. Wang (2022) developed a BERT-based international Chinese AES system incorporating three modules: genre/rhetoric detection, basic information statistics, and automated scoring, demonstrating strong performance and high alignment with manual grading. Additionally, Ma and Xu (2023) developed an intelligent tutoring system utilizing a "shallow textual features + multiple linear regression" approach, though it faces high development costs and promotion difficulties.

Although these AES can assist teaching and facilitate learning, their reliance on professional technical teams increases both the technical learning curve and financial investment for educators. Furthermore, the high technical barriers may limit the application effectiveness of these systems in daily teaching scenarios. In recent years, the emergence of AI Agent technology has provided a new perspective to overcome these dilemmas. Unlike traditional AES systems, essay grading via AI Agents offers advantages such as a low technical barrier, strong task execution capabilities, and excellent scalability. Without prior

programming experience, teachers can leverage platforms like Coze to quickly build personalized agents tailored to the specific needs of various teaching contexts.

However, there remains a gap in current research on AI Agent-based grading specifically designed for international Chinese writing. Ma (2026) developed an intelligent grading system embedded with large language model technology, integrating agent-based grading into the pedagogical model to achieve human-AI collaborative evaluation. Nonetheless, systematic comparative studies between AI agent scoring and human scoring remain insufficient, and the scoring validity within authentic corpora requires further empirical verification. Based on this, the present study intends to construct an AI Agent named "Essay Grading Assistant" using the Coze platform and apply it to grade narrative essays written by international students at Central China Normal University (CCNU). By comparing the scores generated by the AI Agent with those assigned by professional teachers, this study evaluates the efficiency and quality of the AI Agent in essay grading and aims to provide an efficient and easily scalable new approach for grading international students' essays, resulting in advancing the intelligent development of international Chinese education.

2. Research Objective

The research will address the following two core questions:

- 2.1 What is the degree of consistency between the agent's grading and manual grading?
- 2.2 What are the advantages and limitations of using the agent to grade essays?

3. Research Framework

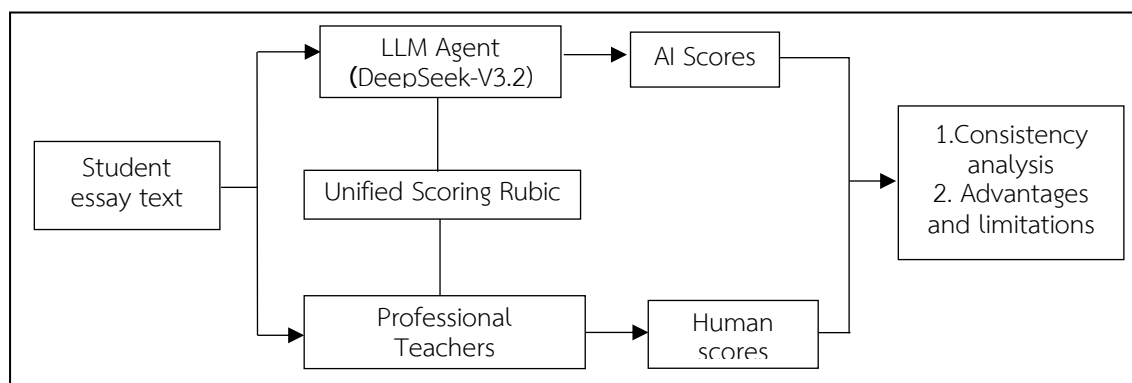


Figure 1 Research Framework

Based on a comparative analysis perspective, this study suggests the research framework illustrated in Figure 1. The research utilizes student essay texts as the input data, which are simultaneously processed by the AI Agent scoring system and evaluated by two professional instructors. Both parties independently generate scores utilizing the identical scoring rubric. This study validates the efficacy of the agent's scoring by comparatively analyzing the degree of consistency between the AI Agent's scores and those of the two human raters. Furthermore, this research adopts the comparative validation paradigm of machine versus human scoring employed by Wang (2022) in AES research. By

substituting the traditional AES system with a DeepSeek-V3.2-based AI Agent, this study evaluates the validity of the agent's scoring through consistency testing and examines its advantages and limitations in comparison to human grading.

4. Methodology

4.1 Source of the corpus and scoring criteria

4.1.1 Source of the corpus

The corpus for this study is derived from the final examination essays of international students enrolled in the Intermediate Writing II class at the International Cooperation Department of CCNU, collected from both the first and second semesters. The dataset comprises a total of 31 essays, amounting to 12,315 characters.

Among these, 16 essays are from the first semester: 15 responded to the prompt "My Room," and 1 responded to the prompt "Beautiful Central China Normal University." The remaining 15 essays are from the second semester, for which the prompt required writing about a specific person, such as "my parents, friends, or teachers." Students were required to illustrate the person's character traits through specific events and were allowed to create their own essay titles. All essays in the dataset belong to the narrative genre. These students in the Intermediate Writing II Class came from seven different countries, including Russia, Saudi Arabia, Hungary, South Korea, Indonesia, etc., with diverse native language backgrounds. There were 7 male students and 15 female students, with ages ranging from 17 to 30 years old. All students passed the HSK Level 4 examination.

4.1.2 Scoring Criteria

This study designed the scoring criteria for the narrative essays of intermediate-level students based on the HSK syllabus, the scoring criteria of Wu et al. (2019), and the scoring indicators of Xu et al. (2015). The scoring was conducted according to four dimensions: Overall Content, Linguistic Richness, Language Accuracy, and Language Fluency. The scores were divided into four grades, and each grade had different score ranges (see Table 1).

Table 1 Scoring Criteria for Intermediate-Level International Students' Narrative Writing

Grade and score	Overall Content	Linguistic Richness	Language Accuracy	Language Fluency
A (85-100)	Exceeds word count; correct formatting; substantial content and complete structure.	Rich vocabulary with accurate usage; demonstrating complex sentence structures and a high frequency of long sentences.	Very few isolated errors in character writing/punctuation; No interference with the narrative.	Smooth and fluent sentences; natural expression with coherent logic; clear and effective storytelling.

Table 1 (continue)

Grade and score	Overall Content	Linguistic Richness	Language Accuracy	Language Fluency
B (75-84)	Near required word count with minor formatting errors; adequate content, and generally complete structure; Main message easily understood.	Relatively diverse vocabulary (some inaccuracies in the meaning); Varied structures, but errors occur in complex sentences.	Small number of grammar, vocabulary, punctuation, and character errors: Fundamentally does not impede the narrative.	Generally fluent sentences, natural expressions; demonstrate basic linguistic coherence for narration.
C (60-74)	Slightly under word count; slightly under word count; multiple formatting errors; lacks substance, though basic meaning is understandable.	Demonstrates basic vocabulary (frequently inaccurate); relies heavily on simple sentences; minimal complex structures.	Frequent grammar, vocabulary, punctuation, and character errors; errors actively impede narrative fluency.	Mediocre fluency; poor coherence and unclear expression requiring reader effort to comprehend.
D (0-59)	Essay is too short and has formatting mistakes; content is not organized; hard to understand overall content.	Use very few words and make many mistakes; sentences are and often not complete.	Serious grammar mistakes, many wrong characters/punctuation errors; very hard to follow the story.	Unsmooth sentence with disconnect with each other; confusing ideas; hard to understand the text

4.2 Construction of the AI Agent

This study utilized the Coze platform to construct the "Essay Grading Assistant" agent. The selection of this platform features several key advantages: first, its no-code nature allows users to configure agents without prior programming experience; second, its extensive feature set supports advanced functionalities such as plugin integration, knowledge base uploads, and workflow design; and third, its ease of deployment ensures that the agent can be rapidly implemented within educational settings upon completion.

Compared to the direct application of general-purpose large language models (LLMs) such as ChatGPT or ERNIE Bot, the Coze platform is considered to be more suitable for the requirements of this study. While generative artificial intelligence (Gen AI) products possess potential for international Chinese language teaching, their lack of domain-specific expertise results in certain deficiencies when evaluating international Chinese essays; notably, some models may produce content where it is difficult to distinguish truth from fiction. Although some chatbots that support knowledge base uploads (e.g., ChatO, Zhipu Qingyan) can be fine-tuned for specific domains, their functionality is largely restricted to conversational interaction, which fails to meet the complexities of comprehensive grading. In contrast, AI Agent technology-characterized by robust task planning, execution capabilities, and high functional scalability-better aligns with the study's requirements for personalized and professionalized essay evaluation.

The AI agent configuration page is found within the "Coze Programming" section. By selecting "AI Agent Development," a personalized AI agent can be configured. The specific development workflow is as follows:

- 1) Defining basic configuration including its name, functional description, persona (role), and response logic.

- 2) Selecting the Appropriate Mode and Large Model, this study selected the Single Agent (LLM mode), which is an agent with a straightforward logical structure. Currently, the Coze platform offers general-purpose models such as Doubao and DeepSeek. Following testing and tuning, this study selected DeepSeek-V3.2 as the foundational large model.

- 3) Adding plugins, a Chinese Word Segmentation plugin, an Optical Character Recognition (OCR) plugin for converting images to text, and a Document Analysis plugin were added. These allow the agent to accurately count words and recognize text from various image and document formats.

- 4) Uploading knowledge base and professional documents, such as essay grading standards and analytical scoring rubrics, these help the model perform accurate analysis and judgment.

- 5) Designing workflow, workflows help standardize tasks and allow for repeatability. The setup involved: A Start Node where a variable was set for the user-uploaded essay; Two LLM Nodes (both powered by DeepSeek-V3.2): one LLM Nodes for composition analysis and the other for essay scoring, lastly, an End Node to output the finalized analysis and scoring results. Linking these four nodes completed the workflow configuration.

- 6) Preview and Debugging: sample essays were uploaded on the orchestration page to preview the results and fine-tuned any areas where the output was suboptimal.

- 7) Publication: The finalized AI Agent was officially deployed for use.



Figure 2 Coze Platform Editing Interface

4.3 Application Procedure

The final examination for international students in the Intermediate Writing II class was conducted as a closed-book test. The exam consisted of three sections: sentence rewriting, paragraph imitation, and a topic-based essay. The total time was 90 minutes, with a maximum score of 100 points. The topic-based essay accounted for 70 points. This study collected the written exam papers of 16 students in the first semester and 15 students in the second semester. After being scanned to photograph, these papers were returned to the instructors. The scanned answer sheets were cropped to isolate the topic-based essays and uploaded to the "Essay Grading Assistant" AI Agent for analysis and scoring. During the application of the AI Agent, it was observed that the OCR and document interpretation plugins displayed certain limitations. Due to the illegible handwriting on some of the answer sheets, the Agent could not completely and accurately recognize and transcribe the original text. To ensure the

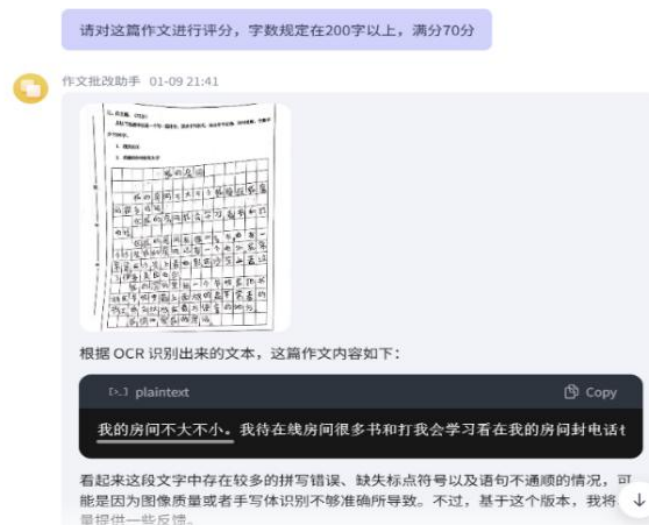


Figure 3 Scoring Interface of the "Essay Grading Assistant"

reliability and accuracy of the final scores, a manual review and error-correction process was implemented following the Agent's initial text recognition. Once the transcriptions were manually verified, they were processed again by the AI Agent for final analysis and scoring, thereby generating the automated grading results.

The study invited two university instructors with over five years of experience teaching Chinese writing courses to serve as a human rater. Referring to the scoring criteria above, raters independently scored the 31 final exam answer sheets. The scoring process was conducted without mutual interference, and the scores were compiled by the researchers upon completion. This study employed Spearman's rank correlation analysis to test the correlation between the agent's scores and the manual scores. The Intraclass Correlation Coefficient (ICC; two-way random effects model, absolute agreement) was used to assess the consistency between the agent's and manual scores.

5. Results

5.1 Consistency Analysis Between Agent and Manual Scores

The study first compiled and analyzed the scores of the 31 final exam papers (with a maximum score of 70) graded by the "Essay Grading Assistant" agent and two professional teachers. There were minor differences between the mean scores of the agent and those of the two professional teachers, with the teachers' scores being slightly higher as shown in Table 2.

Table 2 Descriptive Statistics of AI Agent and Human Ratings

Variable Name(s)	Sample Size	Highest Score	Lowest Score	Mean Score	SD	Median	Variance	Kurtosis	Skewness
AI Agent's grading score	31	65	32	54.032	7.445	55	55.432	2.176	-1.287
Teacher 1's grading score	31	65	35	54.806	7.418	57	55.028	1.558	-1.198
Teacher 2's grading score	31	64	33	54.774	7.5	56	56.247	2.119	-1.312

The three sets of data were subjected to a correlation test using Spearman's rank correlation analysis. The results indicated that the correlation coefficients between the agent and Teacher 1; and between the agent and Teacher 2, were 0.917 and 0.902 respectively ($p < 0.001$), both showing a very strong positive correlation. The correlation coefficient between the two teachers was 0.953 ($p < 0.001$), demonstrating that the scoring trends among the three were highly consistent.

Table 3 Correlation Analysis Between AI Agent and Human Ratings

	AI Agent's grading score	Teacher 1's grading score	Teacher 2's grading score
AI Agent's grading score	1	0.917	0.902
Teacher 1's grading score	0.917	1	0.953
Teacher 2's grading score	0.902	0.953	1

Based on this, to examine the agreement between AI and manual scoring, the study utilized the ICC to conduct an inter-rater reliability analysis for the three sets of scores (AI Agent, Teacher 1, and Teacher 2). Although the score data followed a skewed distribution and did not fully satisfy the normality assumption for analysis of variance, ICC remains a widely used and robust indicator for reliability analysis with non-normal data. A two-way random effects model with absolute agreement was adopted. The results showed that the single measures ICC value for the three sets of scores was 0.964 ($p < 0.001$). According to the evaluation criteria by Cicchetti (1994), where an $ICC > 0.90$ is considered "excellent," the ICC value in this study reached 0.964, indicating excellent consistency between the agent's scores and those of the two teachers.

Table 4 Consistency Analysis Between AI Agent and Human Ratings (ICC)

Term(s)	ICC	95% Confidence Interval		F			
		Lower Limit	Upper Limit	Value	df1	df2	P
Single Measures ICC (1, 1)	0.964	0.94	0.982	86.695	30	62	0.000
Average Measures ICC (1, k)	0.988	0.979	0.994	86.695	30	62	0.000

In summary, the "Essay Grading Assistant" agent developed in this study demonstrates good inter-rater reliability, and its scoring results are highly consistent with manual grading. This finding suggests that the agent has strong potential for application in the grading of intermediate Chinese writing for international students.

5.2 Advantages and Limitations of AI Agent Scoring

5.2.1 Advantages of the "Essay Grading Assistant" AI Agent

5.2.1.1 Enhancing Essay Correction Efficiency

Traditional manual evaluation by instructors is time-consuming and labor-intensive. In contrast, the AI Agent efficiently completes three primary tasks in approximately 15 seconds: computing surface-level linguistic data, assigning specific scores, and formulating targeted writing strategies. By reducing the grading burden, this tool improves overall assessment efficiency and allows educators to focus on personalized writing instruction. Additionally, the Agent's strategy recommendations

serve as a valuable reference for generating tailored feedback, providing students with clear directives to enhance their writing skills.

5.2.1.2 Ensuring Objectivity and Accuracy in Scoring

Manual grading inherently carries some subjectivity and often struggles to evaluate multiple scoring dimensions simultaneously. The AI Agent, on the other hand, follows a systematic approach. It first conducts a detailed analysis of the essay to count errors in vocabulary and sentence structure. Then, it performs a multi-dimensional evaluation based on the scoring rubrics uploaded to its knowledge base and its programmed workflow, ultimately combining these into a total score. This standardized and objective grading process effectively minimizes the impact of personal preferences or subjective human judgment, ensuring the objectivity and accuracy of the results. Additionally, the Agent analyzes the essay from multiple perspectives to guarantee a comprehensive evaluation. For example, it evaluates linguistic accuracy based on surface-level data like the number of incorrect words, pinyin usage, and typos; it judges linguistic fluency based on sentence length and the number of flawed sentences; and it assesses linguistic richness based on the variety of sentences and the use of rhetorical devices.

5.2.1.3 Universal Applicability and Adaptability

For the Coze platform selected for this study, users can easily design an AI Agent tailored to their specific needs without requiring any programming background, making the operation simple and fast. Compared to developing a complete, standalone automated essay correction system from scratch, creating a "Essay Grading Assistant" Agent requires significantly less financial and technical investment. Moreover, to accommodate a wide variety of writing assignments, users simply need to adjust the corresponding variables and parameters on the workflow page or directly input the writing prompt into the chat interface. The AI Agent can quickly adjust to meet these different requirements, demonstrating remarkable flexibility and convenience.

5.2.2 Limitations of the "Essay Grading Assistant" AI Agent

5.2.2.1 Instability of Platform Plugins

Although the Coze platform offers a variety of plugins, they require certain operational costs and carry the risk of being suddenly shut down. When using the "Essay Grading Assistant" to process essays in JPG or PDF formats, the dependent OCR plugin and the Word Reader plugin exhibit unstable performance in transcribing and recognizing the text. For instance, when an essay in the same format is uploaded to the platform multiple times, the plugins might initially recognize the text normally. However, they may subsequently display an error message regard an error in extracting the text from essay and impacts the Agent's efficiency in executing essay correction tasks.

5.2.2.2 Lack of Depth in Content Understanding

While shallow linguistic features can be evaluated for language fluency, accuracy, and richness using designed rubrics, deep semantic understanding cannot be achieved through rubrics alone. Instead, it relies entirely on the interpretation capabilities of the selected LLM, which cannot always

guarantee the quality of the analysis. Moreover, essays written by international students often contain numerous typos and incorrectly used characters, which can further confuse the LLM's understanding of the overall content. For example, a student wrote the phrase "第一种早人房，第二种公人房" (English translation: "the first type is an *early*-person room, the second type is a public-person room") to describe room types. The student's actual intention was to express the choices of a *single* room (单人间) and a double room (双人间). Although the "Essay Grading Assistant" successfully detected the typo "早" (early) and suggested changing it to "单" (single). However, it misunderstood the intended meaning of "公人房" (public-person room) and incorrectly suggested changing it to "公共间" (public space). This suggestion clearly does not align with the practical logic of Chinese usage, reflecting the AI Agent's lack of a deep semantic understanding and pragmatic judgment.

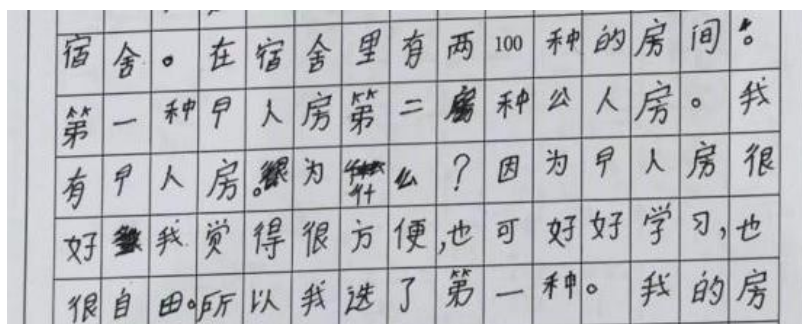


Figure 4 Original Student Essay Excerpt

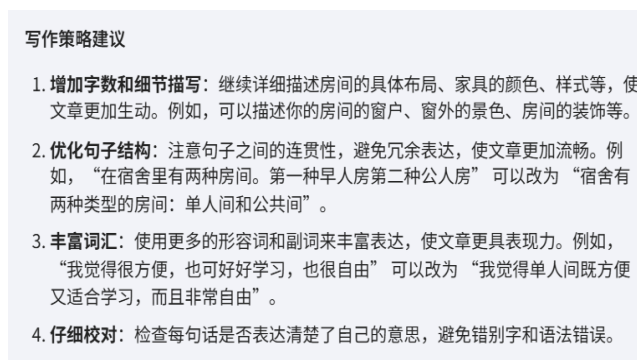


Figure 5 Suggestions provided by the AI Agent

5.2.2.3 The standardized grading approach ignores each student's differences.

When dealing with Chinese language learners from diverse cultural backgrounds and varying learning needs, the "Essay Grading Assistant" applies a uniform standard to correct essays. To a certain extent, this fails to observe the individual differences among learners. For instance, for students from non-character cultural spheres (countries that do not traditionally use Chinese characters), Chinese writing is inherently challenging. Applying strict correction standards to these students can undoubtedly

dampen their enthusiasm for writing and reduce their interest in learning Chinese. Furthermore, learners with different goals require different focal points in essay correction. Relying on a standardized model for grading and evaluation will fail to meet students' personalized developmental needs.

6. Conclusion and Discussion

Analysis of 31 essays scored by the DeepSeek-V3.2 AI agent and two professional instructors reveals extremely high consistency ($ICC = 0.964$). Although the instructors scored an average of 0.758 points higher, the closely aligned variances indicate similar score distributions. This demonstrates the AI's excellent adaptive capabilities in comprehending and effectively applying the uploaded scoring rubrics. The underlying mechanisms can be explained from three perspectives:

First, the application of prompt strategies. Xiang and Huang (2025) demonstrated that few-shot designs using the Chain-of-Thought (CoT) strategy yield the best scoring performance, rivaling or exceeding State-of-the-Art (SOTA) fine-tuning levels. In this study, the agent utilizes a prompt template that requires specific key data outputs (see Figure 6). This approach activates the model's reasoning capabilities, enabling true comprehension of the scoring criteria rather than mere keyword matching, thereby enhancing grading accuracy.

输出格式：
1.作文总字数：
 作文总词数：
2.作文所用到的修辞手法：
3.语法、词汇及汉字书写检查结果：
 错字数量：
 别字数量：
 拼音字数量：
 错词数量：
 离合词数量：
 标点错误数量：
4.内容充实性检测：
 句子数量：
 作文长度：
5.语言通顺性检测：
 错句数量：
 平均句长：
根据评分标准的参考评级：

Figure 6 AI Agent Output Configuration

Second, the implementation of workflows. This study designed a workflow on the Coze platform comprising "Essay Analysis → Essay Scoring → Essay Evaluation." This decomposition of the scoring process into reusable, standardized nodes significantly reduce the randomness of the model's outputs.

Third, the generalization capability of LLMs themselves. Unlike traditional models such as BERT, which require a large amount of labeled data and significant computational power for fine-tuning, LLM-

based agents can demonstrate strong adaptability in few-shot prompts. This implies that in authentic pedagogical settings, instructors can construct high-quality scoring agents by optimizing prompts and establishing workflows, without having to rely on specialized technical teams.

Furthermore, this research found that the AI agent has a distinct advantage in grading efficiency (15 seconds/essay), scoring objectivity, and the identification of shallow linguistic features. These advantages stem from the LLM's parallel computing capabilities and its precise calculation of statistical indicators that human instructors find arduous to accomplish within a brief timeframe.

However, the AI agent still occupies significant limitations in deep semantic understanding and the recognition of unconventional expressions. Ma Ruiling's (2026) research similarly highlighted the agent's deficiencies in providing deep pragmatic guidance. This limitation is primarily correlated with the agent's training corpus: current large models are trained predominantly on standard Chinese texts, lacking sufficient exposure to the interlanguage features of second language learners (e.g., typographical errors, word order confusion, and semantic leaps). Therefore, the model is easily disturbed when processing unconventional expressions. This indicates that AI agents cannot yet entirely replace human scoring; human-AI collaboration represents a more realistic and effective approach at present.

In summary, as a low-barrier, highly efficient assistive tool for essay grading, the AI Agent holds broad application prospects in teaching Chinese as a second language writing. It can effectively alleviate human instructors' burdens, particularly in the automatic identification of and feedback on shallow linguistic features. However, constrained by current technological limitations regarding semantic comprehension and adaptation to individual differences, the AI Agent is better suited as an instructional aide rather than a complete replacement for human grading. Future research should expand the sample size, diversify essay genres, optimize text recognition technologies, and explore the long-term impact of human-AI collaborative scoring on the development of learners' writing proficiency.

7. Recommendations

7.1 Multi-dimensional Human-Machine Collaborative Scoring

AI Agent scoring demonstrates significant advantages in correcting surface-level linguistic features, whereas manual human review excels at grasping the deeper meaning and content of an article. Given their distinct scoring strengths, it is highly recommended to combine the AI Agent with human review to construct a dual-channel collaborative scoring system. Specifically, the AI agent can be assigned to evaluate fundamental linguistic forms and perform basic scoring tasks, thereby liberating teachers from paying too much attention to the basic burdens. This enables teachers to dedicate their efforts to guiding and assessing the cognitive depth and emotional nuances of students' writing and achieving a human-AI collaboration.

7.2 Optimizing the Agent's Scoring Standards and Algorithms

To enhance the Agent's essay correction capabilities, developers should collect diverse writing corpora—including compositions from various cultural backgrounds, proficiency levels, and writing

genres—to enrich the Agent's learning database and strengthen its text recognition and comprehension abilities. The Agent's scoring standards and algorithm settings should be regularly updated and optimized to ensure it can more flexibly adapt to students' personalized developmental needs. Developers might consider adding dynamic model nodes for student essay analysis within the workflow. This would allow the system to track and analyze students' writing portfolios over time, thereby providing teachers with a more targeted, evidence-based foundation for instruction.

7.3 Improving Teachers' Information Literacy

In the era of artificial intelligence, the role of the teacher has undergone profound changes. Teachers are no longer merely transmitters of knowledge; they have become guides for students' autonomous learning, developers of digital educational resources, and practitioners of intelligent teaching. Liu (2024) points out: "The concept of teachers' information literacy has further expanded, encompassing multiple dimensions such as the understanding and application of AI technology, the creation and management of intelligent teaching environments, and the capacity for human-machine collaborative instruction." A teacher's information literacy is the key to driving educational modernization. Therefore, the channels for cultivating this literacy should be broadened to achieve diversified and personalized training pathways. Doing so will enable teachers to better utilize artificial intelligence technology to support their instruction, thereby promoting the innovation of teaching models and the continuous integration and upgrading of educational resources.

7.4 Expanding the Breadth and Depth of Agent Applications

Constrained by natural class sizes, this study utilized convenience sampling to select final essays from intermediate-level classes for analysis, which resulted in a relatively limited sample size. As a preliminary exploration into the application of AI agents for grading Chinese essays, this small sample size presents certain limitations in terms of statistical power. Future research could further expand the sample size by incorporating multi-genre essays written by learners from diverse national backgrounds and proficiency levels. This will help verify the applicability and generalizability of AI agents across varied educational settings.

8. References

- Cai, J. (2024). Zhuliu Meiti AI Agent-xia Yidai Zhineng Quanita Shengchan Chuanshou Tixi De Maodian [Mainstream Media AI Agent: The Anchor of the Next-Generation Intelligent All-Media Production and Communication System]. *Media Science and Technology of China*, 11, 13-17.
<https://doi.org/10.19483/j.cnki.11-4653/n.2024.11.002>
- Huang, Z., Xie, J., & Xun, E. (2014). HSK Zidong Zuowen Pingfen De Tezheng Xuanqu Yanjiu [Study of Feature Selection in HSK Automated Essay Scoring]. *Computer Engineering and Applications*, 50(06), 118-122, 126. <https://bit.ly/4vkqQoX>

- Liu, L. (2024). Rengong Zhineng Shidai Jiaoshi Xinxu Suyang Tisheng Lujing Tansuo [Exploring the Path to Improving Teachers' Information Literacy in the Era of Artificial Intelligence]. *China New Telecommunications*, 26(22), 150-152. <https://bit.ly/4uKlPoN>
- Ma, R. (2026). Da Yuyan Moxing Funeng Zhongwen Eryu Zuowen Zhineng Pinggai De Yingyong Yu Pingjie [Application and Efficacy Assessment for Large Language Models Enabled Intelligent Evaluation & Correction of L2 Chinese Writing]. *Language Teaching and Linguistic Studies*, 47(01), 13-23. <https://bit.ly/3SoO26W>
- Ma, R., & Xu, J. (2024). Shuzhi Shidai Guoji Zhongwen Xiezuozhihui Jiaoxue Ziyuan Chuangxin Yanfa [Innovative Development of International Chinese Writing Smart Teaching Resources in the Digital and Intelligent Era]. *Journal of International Chinese Teaching*, 11(01), 13-23. <https://bit.ly/44sg9EV>
- Ren, C. (2004). HSK Zuowen Pingfen Keguanhua Tantaoh [Exploratory Research on Objective Scoring of HSK Composition]. *Chinese Language Learning*, 24(06), 58-67. <https://bit.ly/4a7YGV5>
- Wang, X. (2022). *Guoji Zhongwen Zuowen Zidong Pigai Ji Fankui Yanjiu* [Research on Automatic Scoring and Feedback of Essays in International Chinese Language Education] [Unpublished master's thesis]. Beijing University of Posts and Telecommunications. <https://bit.ly/4gy3ibF>
- Wu, J., Zhou, W., & Lu, D. (2019). Hanyu Muyu Zhe Hanyu Eryu Xiezuozhihiliang Pinggu Yanjiu-Yi Yuyan Tezheng He Neirong Zhihiliang Wei Celiang Weidu [Assessing Chinese L2 Writing Quality on Basis of Language Features and Content Quality]. *Chinese Teaching in the World*, 33(01), 130-144. <https://link.oversea.cnki.net/doi/10.13724/j.cnki.ctiw.2019.01.013>
- Xiang, J., & Huang, W. (2025). Jiyu Dayuyan Moxing Zhinengti De Zuowen Zidong Pingfen Yanjiu [A Study on Automated Essay Scoring Using LLMs-based Agents]. *Corpus Linguistics*, 12(01), 34-49. <https://bit.ly/3QyluGk>
- Xu, C., Chen, D., Wu, Q. & Xie, Z. (2015). Hanyu Zuowei Di Er Yuyan Zuowen Zidong Pingfen Yanjiu Chutan [On Automated Essay Scoring for Learners of Chinese as a Second Language]. *International Chinese Teaching and Research*, 2(01), 83-89. <https://bit.ly/4eTP43E>
- Zhang, H., Li, D., An, J. & Liu, Y. (2020). Mianxiang Hanyu Xuexi Zhe De Zuowen Zidong Pingfen Xitong Sheji Yu Shixian [Design and Implementation of an Automatic Essay Scoring System for Chinese Learners]. *Electronic Technology & Software Engineering*, 21, 127-130. <https://bit.ly/3ScDbNw>