scientific reports

Check for updates

OPEN

Evaluating simulated teaching audio for teacher trainees using RAG and local LLMs

Ke Fang[1,2]✉, Ci Tang[3] & Jing Wang[3]

In the training of teacher students, simulated teaching is a key method for enhancing teaching skills. However, traditional evaluations of simulated teaching typically rely on direct teacher involvement and guidance, increasing teachers' workload and limiting the opportunities for teacher students to practice independently. This paper introduces a Retrieval-Augmented Generation (RAG) framework constructed using various open-source tools (such as FastChat for model inference and Whisper for speech-to-text) combined with a local large language model (LLM) for audio analysis of simulated teaching. We then selected three leading 7B-parameter open-source Chinese LLMs from the ModelScope community to analyze their generalizability and adaptability in simulated teaching voice evaluation tasks. The results show that the internlm2 model more effectively analyzes teacher students' teaching audio, providing key educational feedback. Finally, we conducted a system analysis of the simulated teaching of 10 participants in a teaching ability competition and invited three experts to score manually, verifying the system's application potential. This research demonstrates a potential approach to improving educational evaluation methods using advanced language technology.

Keywords Simulated teaching, Open-source tools, RAG framework, Teacher student training, LLMs

Simulated teaching, as an important educational practice, serves not only as a critical means for teacher trainees to exercise and enhance their teaching skills but also as a key process for self-reflection and growth[1,2]. It creates a non-threatening and purposeful environment where students can apply their knowledge, complex skills, and abilities, thereby building confidence and proficiency in teaching[3,4]. This method can replicate various scenarios and provide immediate feedback, thus promoting more learning in a shorter time[5,6].

Behavioral skills such as curriculum planning, classroom management, and communication cannot be fully developed through knowledge-based training methods alone. Practice has shown that these skills are most effectively acquired through hands-on experience[4]. Although traditional lectures and written instruction are effective in conveying facts and conceptual knowledge, student teaching experiences remain the most influential in learning how to teach[7]. Research indicates that students should be active participants in the learning process from the early stages of their internship, engaging in meaningful or relevant contexts[8,9]. Therefore, simulated teaching as a tool can create a more realistic, experience-based learning environment, helping schools and universities address emerging challenges in teacher training[8].

However, traditional evaluations of simulated teaching often require direct teacher involvement and guidance, increasing the teacher's workload and limiting the frequency and opportunities for teacher trainees to practice independently[10–12].

In recent years, artificial intelligence has been increasingly applied in educational instruction and evaluation[13–15].

Zhong et al.[16] conducted a bibliometric analysis using the bibliometrix R package on 970 relevant articles published between 2000 and 2023. Their study revealed the developmental trajectory, major contributors, and emerging trends of machine learning in the education sector, highlighting its potential transformative impact on teaching and learning methods. However, Artificial Intelligence in Education (AIEd) still faces numerous challenges and ethical considerations during its integration, including avoiding bias in algorithm design and application, protecting student privacy, and ensuring fairness and transparency. Zawacki-Richter et al.[17], through a systematic review of 146 publications from 2007 to 2018, pointed out that AIEd research is predominantly focused on computer science and STEM disciplines. It is applied in areas such as academic support, assessment and evaluation, adaptive systems, and intelligent tutoring systems. However, they noted a lack of exploration

[1]Network and Information Center, Chengdu Normal University, Chengdu 610000, China. [2]Sichuan Key Laboratory for the Development and Evaluation of Digital Education, Chengdu 610000, China. [3]Office for the Advancement of Educational Information, Chengdu Normal University, Chengdu 610000, China. ✉email: fk@cdnu.edu.cn

from theoretical pedagogical perspectives and ethical discussions. Crompton and Burke[18] conducted a systematic review further confirming the rapid development of AIEd in higher education, especially the significant increase in research publications from China in recent years. They also noted increased involvement of researchers from educational sectors and emphasized the need for future studies to pay greater attention to emerging tools like ChatGPT.

These studies collectively reveal the broad impact and potential of AI in education while reiterating concerns regarding ethics and bias in artificial intelligence. In the design and implementation of this study, audio data from pre-service teachers were pre-screened and reviewed manually to reduce the risk of introducing content containing obvious discriminatory, hateful, or extreme language into the model's inference process.

Regarding teaching evaluation, Rui Wang and others[19] used small object detection technology to analyze and evaluate students' behaviors in the classroom. TS and Guddeti[20]demonstrated a hybrid convolutional neural network for analyzing students' body postures, gestures, and facial expressions to assess engagement. Guo and others[21]improved the evaluation process of university English teaching by integrating artificial intelligence with the Rasch model. Ngoc Anh and others[22] developed an automated system based on facial recognition technology to assess student behaviors in the classroom. Jingjing Hu[23] significantly enhanced the classification accuracy of the teaching evaluation system using machine learning and artificial intelligence methods, particularly the weighted naive Bayes algorithm, compared to traditional algorithms such as naive Bayes and backpropagation. M Rashmi and others[24] proposed an automated system based on YOLOv3 for locating and recognizing multiple actions of students within a single image frame.

Although these studies have significantly improved the automation level and precision of teaching evaluations by introducing advanced AI technologies such as hybrid convolutional neural networks, facial recognition technology, and machine learning algorithms, they primarily focus on analyzing behaviors, facial expressions, and body postures, with less attention given to language evaluation. Yet, in the evaluation of simulated teaching by teacher trainees, language presentation of classroom plans, classroom management, and experimental design are important dimensions for assessment[25].

LLMs, particularly in the field of natural language processing (NLP), offer new solutions for addressing the above issues[26,27]. LLMs are deep learning models with extensive parameters trained on large datasets to learn rich language features and complex structures[28,29]. These models, pre-trained on vast amounts of text data, have acquired extensive language and world knowledge, enabling them to better understand and process natural language, with excellent generalizability and adaptability[30,31].

Local open-source LLMs, although having fewer parameters and not as universally performant as commercial models, offer advantages of free usage, community support, high flexibility, no dependence on the internet, and complete localization of data and storage. They can also satisfy specific downstream tasks[32].

In view of this, this paper designs a Chinese audio evaluation system for simulated teaching by normal school students. The system utilizes speech-to-text technology and a local large language model for analysis, aiming to explore the feasibility and potential of RAG and local large models in simulated teaching scenarios. As the research focuses more on exploring how AI models provide personalized feedback and suggestions to normal school students in the dimensions of speech transcription and language analysis, this study does not include a control group to systematically compare the effects of traditional evaluation methods and the evaluation system.

The contributions of this study are:

- Introduced a method using open-source tools to build a RAG framework and perform inference analysis with local LLMs.
- Assessed the generalizability and adaptability of three open-source Chinese LLMs in the task of audio evaluation for simulated teaching.
- Analyzed the simulated teaching audio of 10 participants in a teaching ability competition, with three experts invited for manual scoring, demonstrating the application potential of the system.
- Enriched the literature on LLM applications in educational evaluation, an area that still has relatively few studies.

System design and implementation

Integration strategy selection for LLMs

In the current field of artificial intelligence, large language models (LLMs) have become key technologies for achieving natural language understanding and generation[33]. With the expansion of model scale and optimization of algorithms, these models have demonstrated outstanding performance across various downstream tasks[34]. Typical integration strategies for LLMs in downstream tasks include **direct fine-tuning**[35], **zero-shot** and **few-shot learning**[36,37], **knowledge distillation**[38], and **retrieval-augmented generation (RAG)**[39].

Direct fine-tuning is the most straightforward integration strategy, involving the adjustment of pre-trained model weights for specific downstream tasks. This approach is simple and effective but may require a large amount of labeled data. **Zero-shot and few-shot learning** leverage the generalization capabilities of LLMs to perform tasks with no or very few labeled data. While this method achieves significant results on certain tasks, its performance often lags behind fully fine-tuned models. **Knowledge distillation** transfers knowledge from large models to smaller models, aiming to maintain the performance of smaller models while reducing computational costs and improving inference speed. **Retrieval-augmented generation** (RAG) combines retrieval and generation, enhancing generation tasks by retrieving relevant information, thereby improving the model's ability to handle complex queries[39].

This study selects RAG for its advantages:

- Enhanced content understanding: By retrieving relevant documents, RAG models can provide deeper evaluations.
- High adaptability: Due to its reliance on dynamic retrieval, RAG can adapt to various teaching content and styles.
- Reduced data dependency: Compared to strategies like direct fine-tuning, which require large amounts of labeled data, RAG reduces reliance on extensive labeled datasets by utilizing existing data.

System framework and workflow

In constructing the system framework, this study adopts a technical solution based on open-source tools. The system utilizes **NVIDIA Triton Inference Server**[40] for text vectorization and retrieval model deployment, **FastChat**[41] for large language model inference, **Milvus**[42] as the vector retrieval database, **Whisper**[43] for building the speech-to-text service, and the Sanic[44] framework for constructing the intermediate web layer.

NVIDIA Triton Inference Server is deployed using Docker to ensure consistency and scalability of the model deployment environment. **Milvus** is also deployed via Docker, with its official image supporting the rapid deployment of an efficient vector database, providing a stable vector retrieval service. **FastChat,** serving as a large-scale language model inference service, operates independently and is registered as a Sanic service. It communicates with the Sanic backend through internal APIs to handle natural language processing tasks. **Whisper** operates as a speech recognition service, similarly running independently and registered as a Sanic service, for on-demand transcription of user speech input. Sanic processes all requests from the frontend and routes them to the appropriate service based on the request type: speech requests are routed to the Whisper service for recognition and transcription, storage and data retrieval requests are routed to the Milvus service to utilize its vector database for retrieval, and final merged text requests are routed to the FastChat service for processing. All service calls are managed through Sanic's asynchronous processing mechanism. The system architecture is illustrated in Fig. 1.

To fully leverage the performance of the RAG framework, along with support for Chinese natural language processing and NVIDIA RTX 4090 GPU resources, this study selects a series of models to handle tasks such as document-to-text conversion, speech-to-text conversion, text vectorization, re-ranking, and language model generation. These models include Whisper's latest speech-to-text model **Large-v3**, **bce-embedding-base_v1**[45] for text vectorization, **bce-reranker-base_v1**[45] for the re-ranking phase, and models for Chinese natural language generation, such as **Baichuan2-7B-Chat**[46], **Qwen-7b-Chat**[47], and **InternLM2-Chat-7B**[48]. These models are based on large Transformer architectures, specifically designed and optimized for processing Chinese text. They exhibit powerful language understanding and generation capabilities and can operate on NVIDIA RTX 4090 GPUs with 16-bit quantization. Within the RAG framework, these models are responsible for generating text outputs based on retrieved information.

The system workflow is divided into two main stages: the storage process and the retrieval process.

Storage process

During the storage process, audio data is denoised and then converted into text format using Whisper's Large-v3 model. Issues such as repeated phrases, non-Chinese content, background noise, and transcription errors significantly affect text quality during speech transcription. To address these issues:

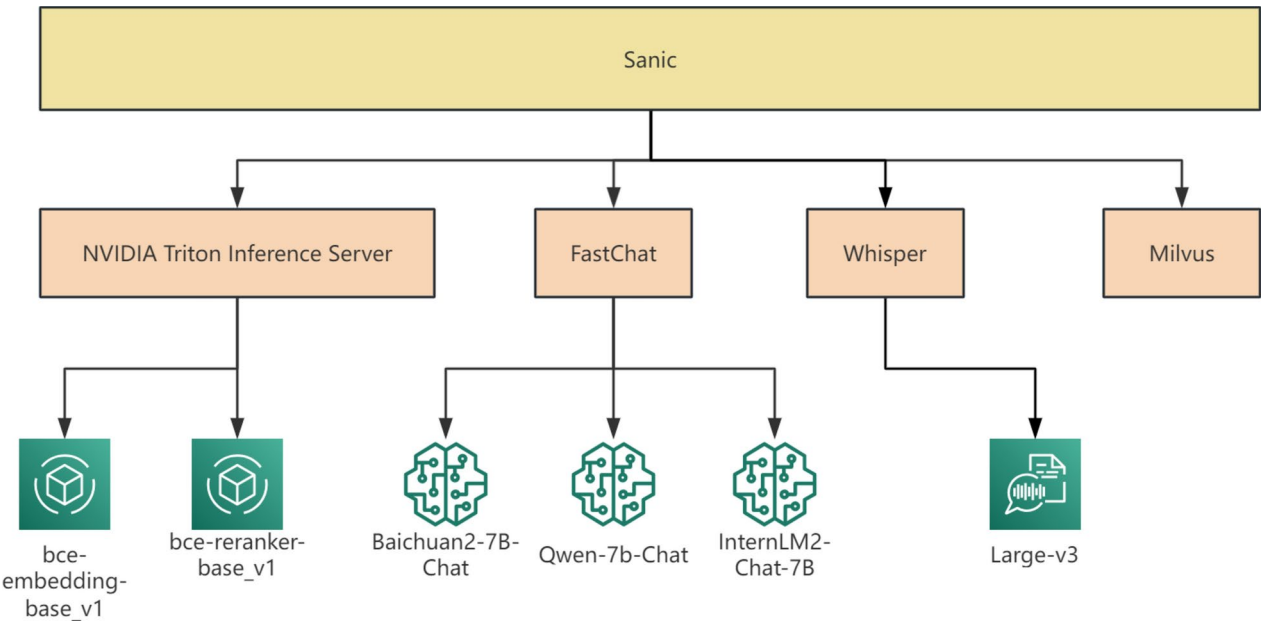

Fig. 1. System architecture diagram.

- Audio denoising and voice enhancement are employed to improve audio clarity.
- Audio slicing is performed to enhance transcription accuracy.
- Dynamic phrase repetition detection algorithms and non-Chinese character ratio detection are used to analyze transcription quality, and poorly transcribed segments are retranscribed.
- Through system-fixed expression adjustments and multiple transcription attempts, the final transcriptions are cleaned and optimized.

The transcribed text data is then input into the bce-embedding-base_v1 model to generate corresponding text embedding representations. Due to the large volume of transcribed text, a segmentation algorithm combined with a dynamic merging strategy is applied to segment the text, generating high-dimensional vector representations to support subsequent semantic search. The generated embeddings are stored in the Milvus vector database.

Retrieval process

During the retrieval process, the query is first converted into a vector representation using the bce-embedding-base_v1 model, following a procedure similar to the text vectorization process in the storage phase. The converted query vector is used to perform similarity search within the Milvus database. Leveraging hybrid search and Milvus's vector retrieval capabilities, the most relevant text data to the query is quickly located.

The retrieved candidate results are then re-ranked using the bce-reranker-base_v1 model. In this step:

- Queries and paragraphs are tokenized, concatenated, and processed.
- For excessively long paragraphs, segmentation and overlap strategies are applied.
- The concatenated sequences are inferred using the Triton Server, with scores calculated via the Sigmoid function.
- For multiple slices of the same paragraph, the maximum score is taken as the final score for that paragraph.

The re-ranked high-relevance results are combined with the user's original query and provided as input to large language models (LLMs) such as Baichuan2-7B-Chat, Qwen-7b-Chat, and InternLM2-Chat-7B. The system workflow is illustrated in Fig. 2, and example code for the core modules can be found in Supplementary Material 1.

Model selection

To assess the models' generalization ability for evaluating simulated teaching texts, this study imported a simulated teaching audio as a reference document. Based on Bloom's taxonomy of educational objectives[49] we designed questions across six dimensions to ask the LLMs. The models' responses were manually analyzed and scored. The scoring standards are provided in Supplementary Material 2.

Questions designed based on Bloom's Taxonomy of Educational Objectives:

- Knowledge level question: Identify the basic knowledge points demonstrated in the student teacher's simulated teaching text and explain their importance.
- Comprehension level question: Explain how the student teacher's simulated teaching text demonstrates an understanding of the teaching content. Provide specific examples.

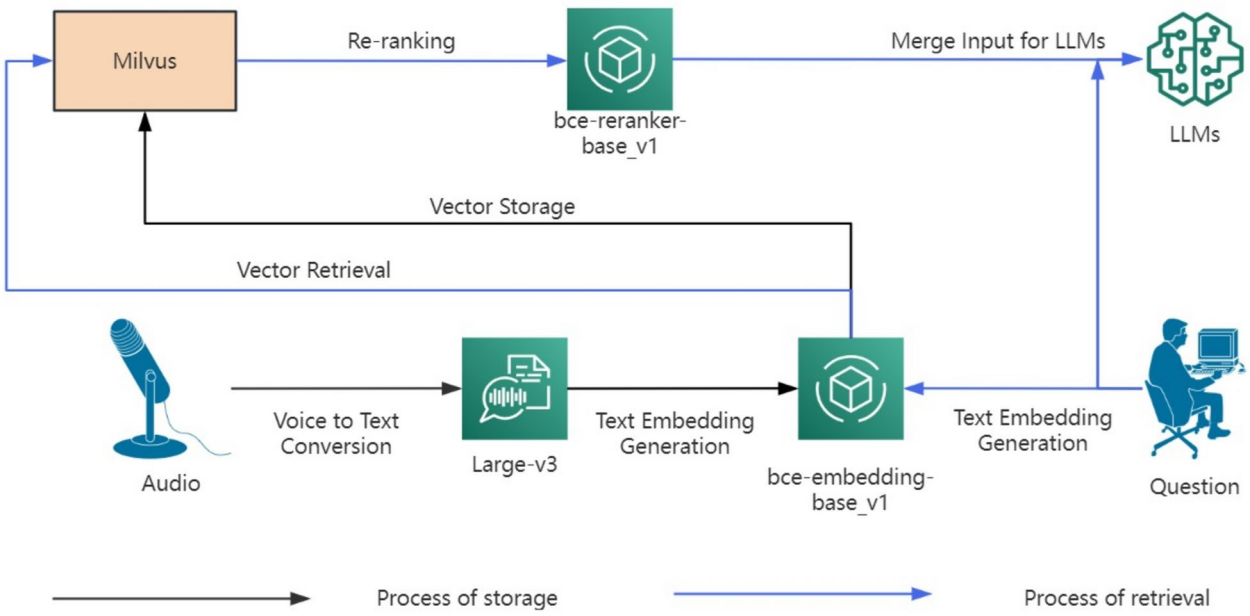

Fig. 2. Workflow diagram.

- Application level question: Analyze how theoretical knowledge is applied to actual teaching activities in the student teacher's simulated teaching text. Provide at least one example of practical application.
- Analysis level question: Evaluate the analytical abilities demonstrated in the student teacher's simulated teaching text when addressing teaching problems, especially how different concepts and theories are distinguished.
- Evaluation level question: Based on the student teacher's teaching text, evaluate their critical thinking regarding teaching methods and strategies.
- Creation level question: Describe how the student teacher's simulated teaching text creatively designs courses or teaching activities to enhance students' learning interest and participation.

The total score for each model across the four dimensions per question (out of a total of 20 points) is shown in Table 1. Detailed responses and analyses from the three LLMs to all questions can be found in Supplementary Material 3.

From the analysis, it can be seen that each model is capable of understanding and describing the questions. The main differences lie in the thoroughness of the answers and the clarity of expression. The InternLM2-Chat-7B model performed best on multiple questions, therefore, we have selected this model for the task.

Experiments and results

After completing the system setup and selecting the LLMs, the study analyzed audio recordings from 10 students who participated in a provincial-level teacher training simulation contest using the LLMs. The results were summarized across four questions:

- "Please analyze the metacognitive skills demonstrated by the teacher trainees in this teaching video, especially how they monitor knowledge and regulate themselves."
- "Evaluate the performance of the teacher trainees in emotional education based on the video content, and provide suggestions for improvement."
- "Assess the diversity and innovation of the teaching strategies used by the teacher trainees, and propose optimization suggestions based on the latest educational theories."
- "Analyze the effectiveness of classroom interaction and student engagement, highlighting strengths and areas for improvement."

These four questions analyze the simulated teaching of preservice teachers from four dimensions: metacognitive skills, emotional education, teaching strategies, and classroom interaction. Metacognition focuses on the professional growth of teachers by analyzing preservice teachers' knowledge monitoring and self-regulation to understand their abilities in self-reflection and improvement[50]. Emotional education reflects the teacher's humanistic qualities, focusing on preservice teachers' responses to students' emotional needs and evaluating the teaching atmosphere and teacher-student relationships[51]. Teaching strategies assess whether preservice teachers can flexibly apply various teaching methods and design and improve their practices based on the latest educational theories[52]. Classroom interaction directly reflects teaching effectiveness, with student engagement and interaction quality being key indicators of successful teaching. This dimension helps teachers concentrate on how to inspire deep learning among students[53].

The model analysis results were scored by three professors specializing in education at our institution. The scoring criteria are available in Supplementary Material 4, with the scoring results shown in Table 2.

The descriptive statistics for the scoring results (Fig. 3) are as follows: Content accuracy: Mean 4.13, Standard deviation 0.51. Depth and detail: Mean 3.70, Standard deviation 0.53. Practicality and innovation: Mean 3.30, Standard deviation 0.47. Logic and organization: Mean 4.50, Standard deviation 0.51. Critical thinking: Mean 3.77, Standard deviation 0.50. Language and terminology: Mean 4.37, Standard deviation 0.49.

The results of the ANOVA (Analysis of Variance, Fig. 4) to examine the differences in evaluation metrics among different courses are as follows: Content accuracy: F-value: 1.33, *P*-value: 0.281. Depth and detail: F-value: 1.24, *P*-value: 0.329. Practicality and innovation: F-value: 1.98, *P*-value: 0.098. Logic and organization: F-value: 0.56, *P*-value: 0.817. Critical thinking: F-value: 2.69, *P*-value: 0.031. Language and terminology: F-value: 0.68, *P*-value: 0.718.

Discussion

Descriptive statistics indicate that content accuracy (4.13) and logical and organization (4.50) received relatively high scores, demonstrating that the model is capable of accurately grasping and reproducing subject matter content while organizing and expressing it in a logically clear manner. Language and terminology (4.37) also performed well, reflecting the model's potential in language expression and the application of domain-specific terminology.

Model	Knowledge-level	Understanding-level	Application-level	Analysis-level	Evaluation-level	Creation-level
Baichuan2-7B-Chat	19	19	18	19	19	18
Qwen-7B-Chat	17	16	16	16	15	16
internlm2-chat-7b	19	20	20	20	20	20

Table 1. Total scores per question for each model.

Subject	Content accuracy	Depth and detail	Practicality and innovation of suggestions	Logicality and organization	Comprehensive assessment and critical thinking	Language expression and use of professional terminology	Expert
Middle school geography	4	3	4	5	3	4	Expert 1
	4	4	3	5	4	4	Expert 2
	4	3	3	5	3	4	Expert 3
Middle school art	4	4	3	5	4	5	Expert 1
	4	3	3	4	4	5	Expert 2
	3	4	3	4	4	4	Expert 3
Middle school math	5	4	3	4	4	5	Expert 1
	4	3	3	5	4	5	Expert 2
	4	3	3	4	4	4	Expert 3
Middle school pe	5	4	3	5	4	4	Expert 1
	4	3	3	4	4	4	Expert 2
	4	4	3	4	4	4	Expert 3
High school biology	4	4	3	5	4	4	Expert 1
	4	3	3	4	3	5	Expert 2
	4	3	3	4	3	4	Expert 3
High school music	5	4	4	5	4	5	Expert 1
	4	3	3	4	3	4	Expert 2
	3	4	4	4	3	4	Expert 3
Elementary school math	5	4	4	5	4	5	Expert 1
	4	3	3	4	3	4	Expert 2
	4	4	3	4	3	4	Expert 3
Elementary school mental health	4	4	3	5	4	4	Expert 1
	5	4	4	5	4	5	Expert 2
	4	4	3	4	4	4	Expert 3
Elementary school english	5	4	4	5	4	4	Expert 1
	5	4	4	5	4	5	Expert 2
	4	4	4	4	4	4	Expert 3
Early childhood education	4	4	3	5	4	5	Expert 1
	5	5	4	5	5	5	Expert 2
	4	4	3	4	4	4	Expert 3

Table 2. Model analysis results scoring.

However, scores for depth and detail (3.70), practicality and innovativeness of suggestions (3.30), and comprehensive assessment and critical thinking (3.77) were relatively low. These results suggest that while the model excels in handling foundational knowledge and language expression, it exhibits limitations in generating in-depth, detailed analysis, innovative content, and critical insights. Improving these areas could be a focus for future model optimization to better support the processing of complex educational content and the cultivation of higher-order thinking skills.

Analysis of variance (ANOVA) results indicate that scores across different courses did not show significant differences in most indicators. Content accuracy (F-value: 1.33, *P*-value: 0.281), depth and detail (F-value: 1.24, *P*-value: 0.329), practicality and innovativeness of suggestions (F-value: 1.98, *P*-value: 0.098), logical and organization (F-value: 0.56, *P*-value: 0.817), and language and terminology (F-value: 0.68, *P*-value: 0.718) suggest that variations in scores among courses are likely influenced more by random factors rather than significant differences in course content itself. However, comprehensive assessment and critical thinking (F-value: 2.69, *P*-value: 0.031) displayed significant differences, particularly in courses such as high school biology, high school music, and elementary mathematics. This suggests that while the model can accomplish foundational knowledge reproduction, it still has limitations in tasks requiring higher-level critical analysis and in-depth evaluation.

Additionally, expert feedback highlighted that the evaluation criteria only addressed limited dimensions of the existing evaluation framework. The system also demonstrated incomplete recognition of more multimodal courses such as music, art, and physical education. Furthermore, the textual output provided by the system was not sufficiently intuitive, requiring teachers to perform secondary interpretation or combine it with other formats for guidance.

This study involves analyzing simulated teaching audio recordings of teacher trainees. All experiments were performed in accordance with the relevant guidelines and regulations.

All audio data were collected with the informed consent of participants, who were explicitly informed that their data might be utilized for future educational research. According to the Ethics Review Committee of Chengdu Normal University (IRB), this study does not require additional ethical approval due to its observational

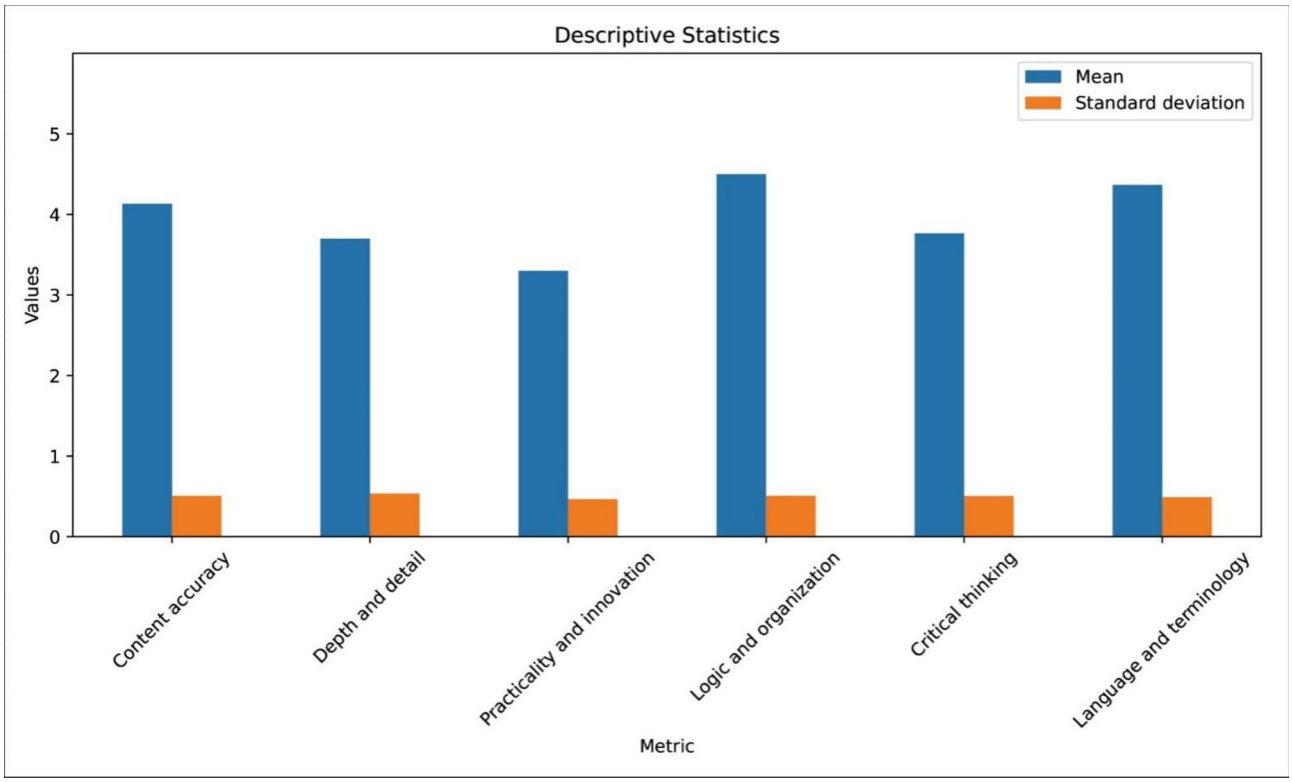

Fig. 3. Descriptive statistical.

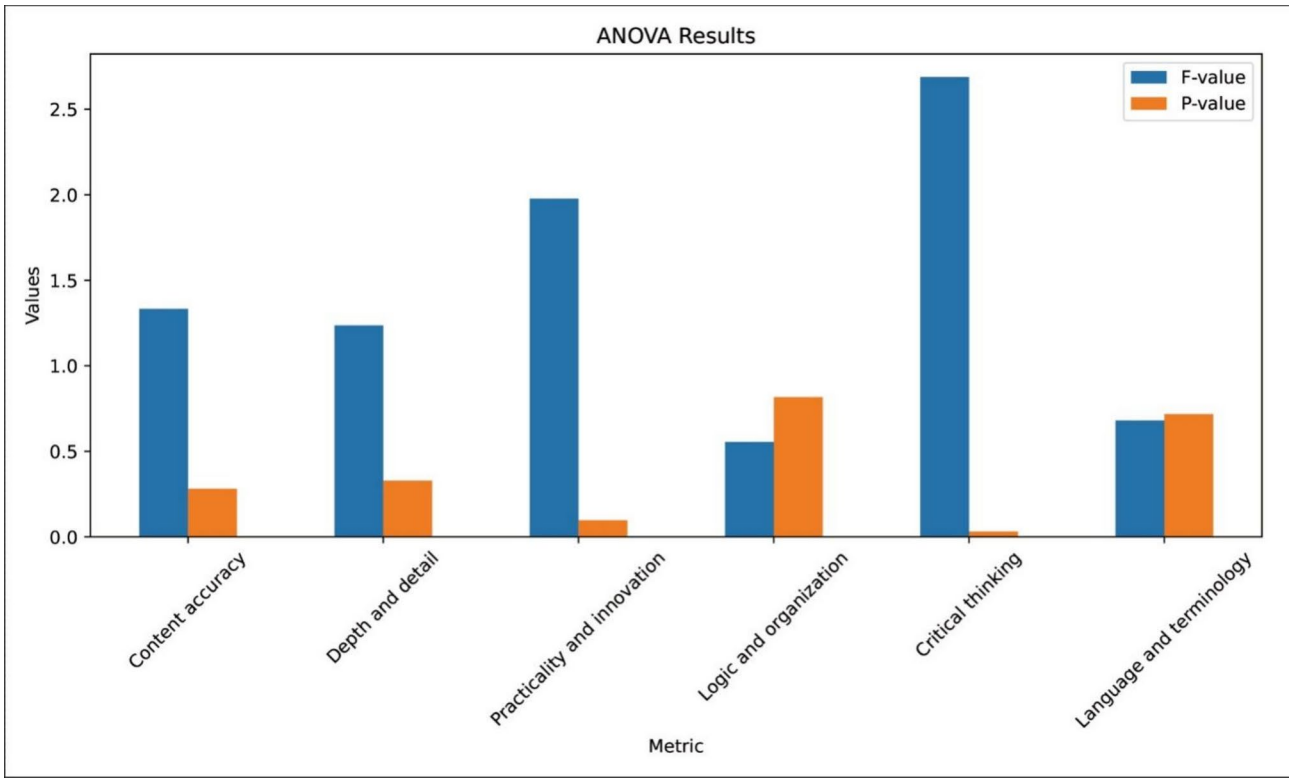

Fig. 4. ANOVA results.

and non-interventional nature. To ensure the personal privacy of the participants, all data were appropriately anonymized prior to analysis.

Limitations

Technical limitations

Although the Whisper model performs excellently in multilingual speech recognition, its sensitivity to specific Chinese syllables is relatively low. This is particularly evident in recognizing sentences with complex contextual dependencies, which may lead to ambiguities. As a result, the model may lose some semantic information when transcribing complex language expressions or specific terms, affecting the accurate understanding of teaching content.

Using the 7B model with 16-bit quantization provides computational efficiency under resource-constrained conditions but inevitably leads to performance loss. During quantization, some precision of the model weights is lost, directly weakening its ability to generate details and limiting its performance in contextual understanding.

This experiment focused solely on speech data, without integrating multimodal information such as video and text. For instance, aspects like the teacher's body language or classroom boardwork, which significantly impact teaching effectiveness, were not included in the analysis. This single-input design fails to provide a comprehensive evaluation of preservice teachers' teaching abilities.

The large-scale model used in the experiment could only output textual evaluations, and the provided textual output was not sufficiently intuitive, still requiring teachers to interpret it or combine it with other formats for use.

Data and scoring limitations

The experimental data included audio samples from only 10 students, which is far from sufficient to reflect the diversity of preservice teachers' teaching behaviors.

The scoring was conducted by three experts with extensive educational experience. However, the limited number of evaluators may restrict the diversity and perspectives in the evaluation process. Additionally, the experts might exhibit some tolerance toward the model's responses during evaluation, introducing a degree of subjective bias into the scoring results.

The experiment did not fully account for other variables that might influence the scoring, such as students' prior knowledge levels, learning styles, subject interests, and age differences.

Insufficient adaptation of the model to teaching scenarios

The current experimental design focuses on standardized simulated teaching scenarios, failing to cover diverse teaching contexts such as special education, cross-cultural classrooms, or rural education.

Teaching is a dynamic interactive process involving students' real-time responses and teachers' immediate adjustments. However, the static analysis approach of this experiment could not capture these real-time interactions, limiting the practical effectiveness of the model's suggestions.

Limitations of the evaluation criteria

The evaluation criteria covered only limited dimensions of the existing evaluation system and lacked thorough modeling and justification, which limited the authority of the scoring standards.

Future work

Technical optimization and enhancement

To address the Whisper model's insufficient contextual understanding of Chinese speech, consider combining it with domain-specific speech transcription models, such as tools optimized for Chinese, to enhance support for complex semantics.

Introduce multimodal fusion technology to combine speech with text, images, and video, capturing more semantic and non-semantic information (e. g., body language, facial expressions, boardwork) during teaching.

Utilize GPUs with larger memory or dedicated inference servers to support higher-parameter models (e. g., 13B or larger), enhancing the model's ability to generate detailed and innovative suggestions. Alternatively, consider using distributed computing frameworks, such as cloud-based inference platforms, to support larger-scale data training and inference.

Fine-tune the model for the education domain to improve its analysis of key teaching points (e.g., critical thinking, student interaction). Employ the latest algorithm optimizations, such as enhanced attention mechanisms or hybrid models, to improve the model's understanding of complex contexts and dynamic teaching scenarios.

Integrate text-to-image, text-to-video, and reinforcement learning-based recommendation systems to create more intuitive multimodal evaluation and guidance tools, enhancing students' efficiency in independent practice.

Data and sample expansion

Expand the experimental sample to a broader group of preservice teachers, covering diverse teaching backgrounds, subject areas, and experience levels.

Collect more student audio, video, and classroom feedback data to build a multimodal dataset, providing richer information for model analysis.

Extend the experimental scope to rural education, cross-cultural classrooms, and special education settings, covering more complex educational contexts and making the model's suggestions more applicable.

Introduce interdisciplinary educational resources (e.g., STEAM teaching content) and real classroom cases to enrich the model's knowledge base and analytical capabilities.

Incorporate more variables that may influence scoring, such as prior knowledge levels, subject interests, and learning styles, into the experimental design. Use multifactor analysis to reduce biases. Standardize the scoring criteria and include specific quantitative indicators.

Practical application and feedback loop

Deploy the model in VR and real teaching environments to observe its actual performance in teacher training, classroom guidance, and teaching design. Collect detailed feedback from teachers and students to analyze the model's strengths and weaknesses in virtual education settings and further adjust its design.

Establish a feedback mechanism to dynamically adjust model parameters and inference processes based on real-world performance and issues, ensuring the model adapts to practical needs.

Continuously update the model's knowledge base by incorporating the latest research findings in education, enabling it to provide cutting-edge educational suggestions.

Conduct long-term follow-up studies to evaluate the model's actual contribution to improving teachers' teaching skills and students' learning outcomes, providing a basis for model improvement. Explore the model's extended applications in education evaluation, curriculum design, and teaching research.

Comprehensive evaluation system development

Customize new scoring standards for the existing teaching evaluation system, using structured modeling approaches and providing thorough quantitative and qualitative justification.

Invite more experts from the field of education to participate in scoring and incorporate the perspectives of teachers and students. Combine expert scoring with automated scoring to explore a human–machine collaborative evaluation model.

Integration of educational philosophy

Incorporate case studies of educational models such as project-based learning (PBL), inquiry-based learning, and flipped classrooms to help the model generate suggestions that better meet practical teaching needs.

Continue exploring the application of metacognitive teaching theories and personalized learning theories in model analysis and suggestions.

Integrate STEAM education principles by combining science, technology, arts, and humanities to enhance the model's interdisciplinary analytical capabilities and provide references for innovative classroom design.

Conclusion

This study analyzed the simulated teaching audio of 10 participating students using a locally deployed RAG-based large model, with the analysis results scored by three education experts. The goal was to evaluate the system's usability and application potential. Overall, the model scored highly in content accuracy, logical coherence and organization, and language expression and use of professional terminology, demonstrating its strengths in processing educational content and precise expression. These findings indicate the potential application value of large models in traditional classrooms and online learning platforms, particularly for handling and presenting complex teaching materials. Additionally, since the model supports multiple languages, it offers useful insights for cross-language contexts such as English teaching analysis.

However, the study also revealed several limitations. First, the sample size was limited to audio data from only 10 students, which constrains the representativeness and generalizability of the findings. Second, the model's understanding of certain courses and complex contexts, as well as its ability to generate innovative content, requires improvement. In more diverse subject areas such as music, art, and physical education, the model's ability to capture teaching key points and nonverbal elements (e.g., body language , music and classroom boardwork) remains inadequate. Furthermore, the scoring relied primarily on experts' subjective evaluations, lacking a refined and quantitative indicator system and control over other factors that may influence teaching behavior (e.g., students' prior knowledge levels, interests, and ages).

Based on these findings, future improvements and expansions should focus on the following aspects:

Technical optimization and multimodal integration: Address the Whisper model's limitations in Chinese context recognition by adopting speech transcription models optimized for Chinese or fine-tuning the model. Incorporate multimodal technologies (text, images, video) to capture a broader range of information in teaching processes.

Sample expansion and multi-scenario coverage: Expand the dataset to include a larger number of preservice teachers from diverse backgrounds, subject areas, and teaching levels, covering varied teaching contexts such as rural education, cross-cultural classrooms, and special education.

Model depth and innovation enhancement: Use larger parameter-scale models or distributed computing frameworks to improve the model's ability to generate detailed, innovative suggestions and analyze critical thinking. Conduct fine-tuning for the education domain to enhance the model's adaptability to dynamic teaching and multidisciplinary content.

Refinement of evaluation standards and feedback mechanisms: Develop more detailed, quantitative, and authoritative scoring standards. Incorporate feedback from teachers, students, and more domain experts to explore hybrid evaluation models combining human and machine assessments. Deploy the model in VR or real classroom settings to collect user feedback for continuous optimization.

Integration of educational philosophy: Incorporate educational approaches such as project-based learning, inquiry-based learning, flipped classrooms, and STEAM education into training and evaluation processes to enrich the model's understanding and application of cutting-edge teaching methodologies.

In summary, the large model used in this study demonstrated potential applications in processing educational content, language expression, and multilingual adaptation. However, continuous efforts are required

in multimodal integration, enhancement of model innovation and critical thinking capabilities, and validation with large-scale samples. Through ongoing technological iteration, data expansion, and feedback loops, the system is expected to provide more valuable intelligent support for developing preservice teachers' teaching abilities and improving classroom quality in diverse educational contexts.

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code availability

The key module code required to reproduce the results (including text transcription, text embedding, text vector storage and retrieval, and re-ranking) is available through Zenodo at https://zenodo.org/records/14699747. Other relevant open-source tools and models can be obtained from the references in the paper. For further inquiries, please contact the corresponding author of this paper.

Received: 27 April 2024; Accepted: 22 January 2025
Published online: 29 January 2025

References

1. Presnilla-Espada, J. An exploratory study on simulated teaching as experienced by education students. *Univ. J. Educ. Res.* **2**(1), 51–63 (2014).
2. Tortorelli, C. et al. Simulation in social work: Creativity of students and faculty during COVID-19. *Soc. Sci.* **10**(1), 7 (2021).
3. Kelleci, Ö. & Aksoy, N. C. Using game-based virtual classroom simulation in teacher training: User experience research. *Simul. Gaming* **52**(2), 204–225 (2021).
4. Salas, E., Wildman, J. L. & Piccolo, R. F. Using simulation-based training to enhance management education. *Acad. Manag. Learn. Educ.* **8**(4), 559–573 (2009).
5. Percival, K. & Jimenez, O. Interactive performance and simulation learning. PARtake J. Perform. Res., **4** (1). (2021).
6. Kumar, A. et al. A short introduction to simulation in health education. *J. Med. Evid.* **4**(2), 151–156 (2023).
7. Caires, S. & Almeida, L. S. Positive aspects of the teacher training supervision: The student teachers' perspective. *Eur. J. Psychol. Educ.* **22**(4), 515–528 (2007).
8. Bell, B. & Kozlowski, S. Active learning: Effects of core training design elements on self-regulatory processes, learning, and adaptability. *J. Appl. Psychol.* **93**, 296–316 (2008).
9. Cannon-Bowers, J. & Bowers, C. Synthetic learning environments: On developing a science of simulation, games and virtual worlds for training. In *Learning, Training, and Development in Organizations*, 229–261. (2009).
10. Cohen, J. et al. Experimental evidence on the robustness of coaching supports in teacher education. *Educ. Res.* **53**(1), 19–35 (2024).
11. Marvi, K. M. H. The anatomization of learning principles administration in secondary school teaching. *Am. J. Educ. Pract.* **7**(1), 82–96 (2023).
12. Burakgazi, S. G. Curriculum adaptation and fidelity: A qualitative study on elementary teachers' classroom practices. *Issues Educ. Res.* **30**, 920–942 (2020).
13. Fang, K. & Wang, J. Interactive design with gesture and voice recognition in virtual teaching environments. *IEEE Access* **12**, 4213–4224 (2024).
14. Sun, S. & Deng, P. Review of Artificial Intelligence Empowerment Teaching Evaluation Based on CiteSpace. In *2023 5th International Conference on Computer Science and Technologies in Education (CSTE)*, 306–311. (2023).
15. Dai, C. P. et al. Improving teaching practices via virtual reality-supported simulation-based learning: Scenario design and the duration of implementation. *Br. J. Educ. Technol.* **54**(4), 836–856 (2023).
16. Zhong, Z., Guo, H. & Qian, K. Deciphering the impact of machine learning on education: Insights from a bibliometric analysis using bibliometrix R-package. *Educ. Inf. Technol.* **29**, 1–28 (2024).
17. Zawacki-Richter, O. et al. Systematic review of research on artificial intelligence applications in higher education–where are the educators?. *Int. J. Educ. Technol. Higher Educ.* **16**(1), 1–27 (2019).
18. Crompton, H. & Burke, D. Artificial intelligence in higher education: The state of the field. *Int. J. Educ. Technol. Higher Educ.* **20**(1), 22 (2023).
19. Wang, R. et al. Post-secondary classroom teaching quality evaluation using small object detection model. *Sci. Rep.* **14**(1), 5816 (2024).
20. Ts, A. & Guddeti, R. M. R. Automatic detection of students' affective states in classroom environment using hybrid convolutional neural networks. *Educ. Inf. Technol.* **25**(2), 1387–1415 (2020).
21. Guo, E. & Sun, L. English-assisted teaching evaluation system based on artificial intelligence and rasch model. *Adv. Multimed.* **2022**, 9550117 (2022).
22. Ngoc Anh, B. et al. A computer-vision based application for student behavior monitoring in classroom. *Appl. Sci.* **9**(22), 4729 (2019).
23. Hu, J. Teaching evaluation system by use of machine learning and artificial intelligence methods. *Int. J. Emerg. Technol. Learn. (iJET)* **16**(05), 87–101 (2021).
24. Rashmi, M., Ashwin, T. & Guddeti, R. M. R. Surveillance video analysis for student action recognition and localization inside computer laboratories of a smart campus. *Multimed. Tools Appl.* **80**(2), 2907–2929 (2021).
25. Petrocelli, E. Pre-service teacher education: Observing senior teachers through the theoretical lens of Ellis's principles of instructed language learning. *EuroAmerican J. Appl. Linguist. Lang.* **8**(1), 20–52 (2021).
26. Wen, B. Research on the applications of natural language processing. *ACE* **2023**(16), 220–227 (2023).
27. Vera, P., Moya, P. & Barraza, L. Rethinking the Evaluating Framework for Natural Language Understanding in AI Systems: Language Acquisition as a Core for Future Metrics. arXiv preprint arXiv:2309.11981, (2023).
28. Patwary, M. Keynote Talk 2 training large language models: Challenges and opportunities. *IEEE Int. Parallel Distrib. Process. Symp. Workshops (IPDPSW)* **2022**, 1245–1245 (2022).
29. Catanzaro, B. Language Models: The Most Important Compute Challenge of Our Time (Keynote). In *Proceedings of the 28th ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, **3**, 2. (2023).
30. Axelsson, A. & Skantze, G. Using large language models for zero-shot natural language generation from knowledge graphs. arXiv preprint arXiv:2307.07312, 2023.
31. Thangarasa, V., Gupta, A., Marshall, W., et al. SPDF: Sparse pre-training and dense fine-tuning for large language models. In *Uncertainty in Artificial Intelligence*, pp. 2134–2146. (2023).

32. Cooper, N., Scholak, T. Perplexed: Understanding When Large Language Models are Confused. arXiv preprint arXiv:2404.06634, (2024).
33. Zhao, W. X., Zhou, K., Li, J., et al. A survey of large language models. arXiv preprint arXiv:2303.18223, (2023).
34. Gadre, S.Y., Smyrnis, G., Shankar, V., et al. Language models scale reliably with over-training and on downstream tasks. arXiv preprint arXiv:2403.08540, (2024).
35. Howard, J. & Ruder, S. Universal language model fine-tuning for text classification. arXiv preprint arXiv:1801.06146, (2018).
36. Romera-Paredes, B. & Torr, P. An embarrassingly simple approach to zero-shot learning. In *International Conference on Machine Learning*. 2152–2161. (2015).
37. Wang, Y. et al. Generalizing from a few examples: A survey on few-shot learning. *ACM Comput. Surv. (csur)* **53**(3), 1–34 (2020).
38. Gou, J. et al. Knowledge distillation: A survey. *Int. J. Comput. Vis.* **129**(6), 1789–1819 (2021).
39. Lewis, P. et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **33**, 9459–9474 (2020).
40. NVIDIA triton inference server[EB/OL]. https://docs.nvidia.com/deeplearning/triton-inference-server/.
41. Zheng, L., Chiang, W.-L. Sheng, Y., et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*. **36**. (2024).
42. Wang, J., Yi, X., Guo, R., Jin, H., Xu, P., Li, S., Wang, X., Guo, X., Li, C., Xu, X. et al. Milvus: A purpose-built vector data management system. In *Proceedings of the 2021 International Conference on Management of Data*, 2614–2627. (2021).
43. Whisper [EB/OL]. https://github.com/openai/whisper.
44. Sanic [EB/OL]. https://github.com/sanic-org/sanic.
45. BCEmbedding [EB/OL]. https://github.com/netease-youdao/BCEmbedding.
46. Yang, A., Xiao, B., Wang, B., et al. Baichuan 2: Open large-scale language models. arXiv preprint arXiv:2309.10305, (2023).
47. Bai, J., Bai, S., Chu, Y., et al. Qwen technical report. arXiv preprint arXiv:2309.16609, (2023).
48. Cai, Z., Cao, M., Chen, H., et al. Internlm2 technical report. arXiv preprint arXiv:2403.17297, (2024).
49. Krathwohl, D. R. A revision of Bloom's taxonomy: An overview. *Theory Pract.* **41**(4), 212–218 (2002).
50. Magno, C. The role of metacognitive skills in developing critical thinking. *Metacognition Learn.* **5**, 137–156 (2010).
51. Martin, B. L. & Reigeluth, C. M. *Affective Education and the Affective Domain: Implications for Instructional-Design Theories and Models* 485–509 (Routledge, 2013).
52. Abdullah, G. et al. Assessing the influence of learning styles, instructional strategies, and assessment methods on student engagement in college-level science courses. *Int. Educ. Trend Issues* **2**(2), 142–150 (2024).
53. Gardner, R. Classroom interaction research: The state of the art. *Res. Lang. Soc. Interact.* **52**(3), 212–226 (2019).

Author contributions

K.F. was responsible for constructing the system framework, data analysis, writing the article, etc., C.T. contribution is the writing, experimental design, and data analysis etc. of the paper, while J.W. was in charge of arranging experiments, inviting experts, and so on.

Funding

Sichuan Province Educational Informatization Application and Development Research Centre, JYXX23-006.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-87898-5.

Correspondence and requests for materials should be addressed to K.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.