Eng Int Syst (2026) 5: 653–664
© 2026 CRL Publishing Ltd

Engineering
Intelligent Systems

Research on Closed Loop Management Information System of Hybrid Teaching Driven by Generative Artificial Intelligence

Jian Zhang* **and Xiu Wang**

Taishan University, Taian, Shandong, China

To address significant problems such as feedback lag, insufficient personalization and static teaching resources in hybrid teaching, this paper proposes a closed-loop adaptive generative teaching engine—CLAGTE—driven by generative artificial intelligence. The engine consists of multi-modal student state awareness network MHAFNet, time-series attention knowledge tracking model TAKT, teaching strategy decision module based on near-end strategy optimization and retrieval enhancement generative content creation module, which form a full-link closed loop of "perception-diagnosis-decision-generation". Experimental results on three public datasets of ASSISTments, Junyi Academy and EdNet show that the AUC of CLAGTE in performance comparison is as high as 0.879, which is about 2.6 percentage points higher than the optimal baseline DIMKT. Its AUC decreases by only 11.9% under 30% label noise, and its robustness is significantly better than that of the comparative model. At the same time, the comprehensive mean of practical subjective score reaches 4.33, and the interpretable Spearman correlation coefficient is 0.68. The accuracy rate of learning gain simulation increases by 7.8% after a single intervention, and the cumulative accuracy rate increases by 14.6%. In addition, its cross-domain zero-sample migration AUC remains above 0.75. Furthermore, the ablation experiment verifies the effectiveness of each module. In summary, CLAGTE achieves the intelligent closed loop of hybrid teaching management through the collaboration of multi-modal fusion, dynamic knowledge tracking, reinforcement learning adaptive strategy and generative content supply, and provides a feasible technical solution for personalized education.

Keywords: Generative artificial intelligence; hybrid teaching; closed-loop management information system; knowledge tracking; multimodal fusion

1. INTRODUCTION

With the rapid development of information technology, multimodal teaching has become the mainstream approach implemented to reform pedagogy in higher education, but its implementation still faces multiple difficulties. Multimodal data related to behavior, emotion and academic data are scattered on different platforms and there is no unified acquisition and fusion mechanism. Moreover, most of the existing knowledge tracking models rely only on a single-answer record, and ignore information such as facial expression and eye movement trajectory, which restricts the accuracy and timeliness of diagnosis. Even if weak points are diagnosed, the traditional system can only mechanically recommend from the preset question bank, lack the ability to dynamically generate personalized content, and can't adaptively adjust teaching strategies. The breakthrough of generative artificial intelligence brings opportunities to solve the above problems, but the existing research mostly embeds it as an independent tool in the teaching link, and fails to systematically integrate with modules such as multimodal perception, knowledge tracking and strategy optimization, and has not yet formed a closed-loop management system covering the whole link.

*Corresponding author's e-mail address:tsuzj@tsu.edu.cn

Aiming at the above problems, this paper proposes a closed-loop adaptive generative teaching engine CLAGTE driven by generative artificial intelligence. In this paper, the multi-head attention cross-modal fusion network MHAFNet is designed to align and fuse three heterogeneous data comprising behavior, emotion and academic, and a time-series attention knowledge tracking model TAKT is proposed. Moreover, this paper uses a forgetting factor and memory enhancement mechanism to dynamically diagnose the degree of knowledge mastery, modeling teaching strategy optimization as a Markov decision-making process, and adopting near-end strategy optimization algorithm to adaptively select teaching actions. In addition, this paper introduces the retrieval-enhanced generation paradigm, and uses the fine-tuned large language model to dynamically generate practice questions and explanations that match the learning situation. Furthermore, this study integrates the above mentioned four modules into a complete closed-loop information system for the first time.Its effectiveness is verified by seven-dimensional experiments on three international public datasets: ASSISTments, Junyi Academy and EdNet. The results show that CLAGTE is significantly superior to the mainstream baseline model in terms of prediction performance, robustness, practicality, interpretability, learning gain and cross-domain generalization, thereby providing a technical scheme with both theoretical innovation and practical feasibility for the intelligent management of a hybrid teaching mode.

2. RELATED WORK

2.1 Practice Evolution and Management Pain Points of Hybridteaching Mode

As the core teaching mode that integrates online independent inquiry and offline in-depth discussion, hybrid teaching is always limited by the fragmentation of process data and the lack of closed loop feedback. The early research on multi-focus mode design and learning effect verification has not yet approached the architecture level of management information systems. By means of empirical analysis, Daskan and Yildiz [1] were the first to verify the positive effect of blended teaching on general learning outcomes. However, they did not include the design of a process management mechanism. Evans et al. [2] conducted a case study of Hong Kong universities, and found that insufficient support for teachers' professional development is one of the core obstacles to the uptake of the hybrid teaching mode. However, they did not investigate or recommend any supporting management tools. Hanny et al. [3] identified 12 obstacles for the K-12 scenario. They found that data was fragmented and the feedback was not prompt across academic stages.However, no technical solution was proposed. Taghizadeh and Hajhosseini [4] and Zhu [5] validated the positive effects of teaching interaction quality and appropriate content on learner satisfaction with TEFL master's program and English linguistics SPOC scenarios as objects, respectively. However, neither of them established a full-link closed loop of "data collection-cognitive diagnosis-intervention push", limiting the large-scale promotion of hybrid teaching.

2.2 Scenario-Based Exploration of Generative AI-Enabled Hybrid Teaching

The multi-round interaction and dynamic generation capabilities of generative AI provide new possibilities for addressing the pain points of hybrid teaching management. Since 2023, relevant research has covered multi-academic and multi-disciplinary scenarios, but most of them involve single-point tools, and no system-level architecture has been developed. Santosa and Ratminingsih [6] investigated future teachers' GenAI perception in EFL under-represented scenarios and found that they generally recognized the value of providing personalized content. However, there are concerns about content reliability. Zhang et al. [7] applied the flipping model supported by GenAI to the training of pre-service chemistry teachers, and used controlled experiments to verify its ability to improve students' ability to design teaching units and critical thinking. However, the model was not aligned with the existing teaching management system. Spasopoulos et al. [8] and Lee et al. [9] conducted a systematic review of teacher training and language classroom scenarios in science departments, respectively. They pointed out that the current application of GenAI in education is still dominated by isolated tools, and there is a lack of system design embedded in the whole hybrid process. Xia et al. [10] explored the interdisciplinary collaborative learning model supported by GenAI and found that it can effectively lower the threshold of collaboration, but did not solve the problem of behavior tracking and dynamic adaptation. Sharma and Elaish [11] proposed an extended AI-TPACK framework to measure teacher readiness in the generative AI era, He and Shen [12] summarized the core model of "student-GenAI" collaborative learning, and Li et al. [13] clarified the operation mechanism and uptake of GenAI to empower the high-quality development of higher education. All three research works provide a theoretical framework for the application to be implemented, but they have not applied it to a system module. Wu et al. [14] developed the GenAI ability assessment scale for pre-service second language teachers, Leon [15] proposed a new paradigm of GenAI-driven on-demand tutoring, and Jeyaraj and Sethi [16] introduced GenAI tools into the teaching of business software. These scenario-based studies have all verified the practical value of GenAI, but no general hybrid teaching closed-loop management structure has been developed.

2.3 Underlying Logic and Existing Controversy of GenAI Educational Application

In addition to scenario-based exploration, academic circles also examine the adaptability of GenAI in the education system from a more macro perspective, and at the same time expose common problems to be solved such as architecture adaptation and content risks. Griffiths et al. [17] proposed that GenAI education should shift from the traditional "content transmission paradigm" to the "human-machine coordination paradigm" from the perspective of cybernetics,

which provided a theoretical anchor for the design of closed-loop management system. However, no specific technical implementation was proposed. Nikolopoulou [18] and Creely [19] discussed the teaching practice path of GenAI in the general scenario of higher education and the language learning scenario, and both pointed out that there are still hidden dangers such as content hallucination and fuzzy ethical boundaries in the current application. Ogunleye et al. [20] proposed the adjustment and change of the assessment system given the current GenAI era. However, an automated implementation of process evaluation is not involved. Van Slyke et al. [21] specifically discussed the challenges and responses brought by GenAI in information system education, and clearly pointed out that the current static architecture of educational information systems can no longer adapt to the dynamic interaction requirements of GenAI, which is the core entry point of this research.

2.4 Study Deficiency and The Entry Point of This Paper

The existing research has confirmed the value of hybrid teaching and verified the feasibility of generative AI empowerment, but there are still three gaps from the perspective of "closed-loop management information system". First, most hybrid teaching researches have not established the full-link closed loop of multi-modal data fusion, and the problems of feedback lag and insufficient adaptation accuracy have not been solved. Second, the application of generative AI is mostly embedded in the teaching link in the form of single-point tools, and is not coordinated with dynamic knowledge tracking and strategic decision-making modules. Further, there needs to be an increase in the amount of personalized content made available to students. Third, the architecture of the existing educational information system is not aligned with the dynamic generation characteristics of GenAI, and it is difficult to support the full-process management needs of large-scale hybrid teaching scenarios. Therefore, this paper proposes a generative artificial intelligence-driven hybrid teaching closed-loop management information system—CLAGTE—which can achieve the full-link closed-loop management of hybrid teaching by means of four collaborative modules: multi-modal perception, dynamic diagnosis, strategy optimization and generative intervention.

3. ALGORITHMS AND MODELS: CLOSED-LOOP ADAPTIVE GENERATIVE TEACHING ENGINE (CLAGTE)

Aiming at the problems of feedback lag, insufficient personalization of content,and static teaching resources in traditional hybrid teaching, this chapter proposes an innovative closed-loop adaptive generative teaching engine (CLAGTE). This engine is not a single algorithm, but a collaborative workflow comprising four core modules:multi-modal student state awareness, dynamic knowledge tracking and cognitive diagnosis, teaching strategy optimization based on deep reinforcement learning, and generative content creation and intervention. These four modules form a closed loop, which achieves an automated and intelligent cycle from data acquisition, state analysis, strategy decision-making to content generation.

3.1 Multimodal Student State Awareness and Feature Fusion

Traditional systems usually rely on only single-mode data such as test scores, and it is difficult to fully capture the learning status of students [22]. This model first constructs a multi-modal data acquisition layer, which integrates three types of heterogeneous data sources:

1) Behavioral data. This includes click streams from online learning platforms, video viewing duration, forum interaction frequency, assignment submission time, etc.

2) Physiological/emotional data. This includes facial expressions and eye movement tracks captured by cameras, as well as skin conductance and heart rate variability provided by wearable devices (subject to ethical compliance).

3) Academic performance data. This comprises stage test scores, assignment scores, knowledge point mastery history records.

To fuse these high-dimensional, heterogeneous data, a Multi-Head Cross-Modal Attention Fusion Network (MHAFNet) is designed. The core idea is to make the data of different modalities "pay attention" and "align" with each other during the learning process.

The behavioral feature vector is set to $x_b \in \mathbb{R}^{d_b}$, the emotional feature vector to $x_e \in \mathbb{R}^{d_e}$, and the academic feature vector to $x_a \in \mathbb{R}^{da}$. First, they are mapped to the same dimension d by independent fully-connected layers:

$$h_b = ReLU\left(W_b x_b + b_b\right) \tag{1}$$

$$h_e = ReLU\left(W_e x_e + b_e\right) \tag{2}$$

$$h_a = ReLU\left(W_a x_a + b_a\right) \tag{3}$$

where, W_* and b_* are learnable parameters. Subsequently, the cross-attention mechanism is used to enhance the interaction between the modalities. For example, the attention weight of the behavioral modality versus the emotional modality is calculated:

$$\alpha_{be} = soft\max\left(\frac{\left(W_Q^b h_b\right)\left(W_K^e h_e\right)^T}{\sqrt{d}}\right) \tag{4}$$

$$z_{be} = \alpha_{be}\left(W_V^e h_e\right) \tag{5}$$

where, W_Q^b, W_K^e, W_V^e are query, key, and value matrices, respectively. Similarly, we can get z_{ba}, z_{ea}, z_{eb}, z_{ab} and z_{ae}. The final fusion state vector st is obtained by splicing all original mapping features and cross-attention features and undergoing a linear transformation:

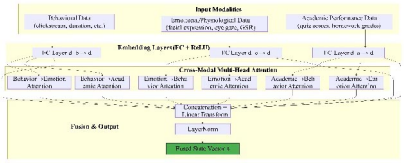


Figure 1 Structure diagram of multimodal student state awareness and feature fusion network.

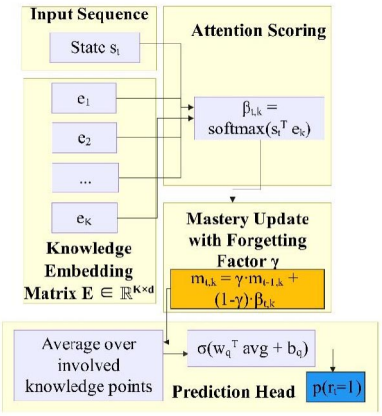


Figure 2 Schematic diagram of temporal attention knowledge tracking model (TAKT).

$$s_t = LayerNorm\left(W_o\left[h_b; h_c; h_a; z_{be}; z_{ba}; \ldots\right] + b_o\right) \quad (6)$$

3.2 Dynamic Knowledge Tracking and Cognitive Diagnosis

After obtaining the student's current state s_t, it is necessary to dynamically model their knowledge mastery. Therefore, a Temporal Attention Knowledge Tracing (TAKT) model is proposed, which improves the traditional DKT (Deep Knowledge Tracing) model[23].

The core of TAKT is a Transformer decoder with a memory enhancement mechanism. It not only predicts the probability that students will answer a certain question correctly at the next moment—it also dynamically diagnoses the degree of mastery of each knowledge point. It assumes a total of K knowledge points. The model maintains a knowledge point embedding matrix: $E \in \mathbb{R}^{K\times d}$.

When the state s_t of time step t is given, its attention score for all knowledge points is calculated:

$$\beta_{t,k} = \frac{\exp\left(s_t^T e_k\right)}{\sum_{j=1}^{K}\exp\left(s_t^T e_j\right)} \quad (7)$$

where, e_k is the embedding vector of the k-th knowledge point. Then, the student's mastery $m_{t,k}$ of the knowledge point k can be expressed as the weighted sum of the historical mastery and the current attention score:

$$m_{t,k} = \gamma \cdot m_{t-1,k} + (1-\gamma)\cdot\beta_{t,k} \quad (8)$$

where, $\gamma \in [0, 1]$ is the forgetting factor. The model predicts that the probability that a student will correctly answer a question q involving a set of knowledge points K_q is:

$$p\left(r_t = 1|q, S\right) = \sigma\left(w_q^T\left(\frac{1}{|K_q|}\sum_{k\in K_q} m_{t,k}\right) + b_q\right) \quad (9)$$

where, σ is a Sigmoid function and w_q and b_q are parameters related to the problem. The training loss is the binary cross-entropy loss:

$$L_{KT} = -\sum_{(q,r_t)\in D}\left[r_t \log\left(p_t\right) + \left(1-r_t\right)\log\left(1-p_t\right)\right] \quad (10)$$

3.3 Teaching Strategy Optimization Based on Deep Reinforcement Learning

When there is a precise awareness of the student's status, the next step is to decide the teaching action to be taken. This is modeled as a Markov Decision Process (MDP) and the Proximal Policy Optimization (PPO) algorithm is used to learn the optimal teaching strategy $\pi\theta\left(a|s\right)$.

The state space S consists of the knowledge point mastery vector $m_t = \left[m_{t,1}, \ldots, m_{t,k}\right]$ output by TAKT and the aggregated features of the student's emotional state h_e.

The action space A involves selecting practice questions of different difficulties (easy, medium, and difficult), selecting different types of learning content (video, text, case), requesting the generative AI to create new example questions, and so on.

The reward function $R\left(s, a\right)$ is the key to success. We define the following instant rewards and long-term rewards:

Knowledge gain reward R_{kg} represents the improvement of students' knowledge mastery degree after completing the action.

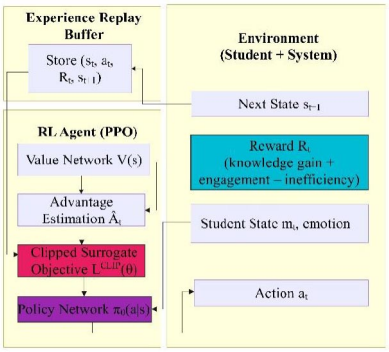


Figure 3 PPO-based teaching strategy optimization flow chart.

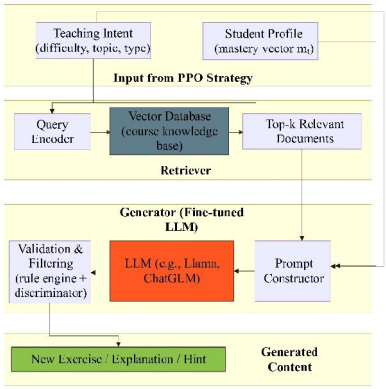


Figure 4 RAG-based generative content authoring architecture diagram.

The engagement reward R_{eng} represents the change in concentration score calculated based on facial expression and interaction behavior.

Efficiency penalty R_{eff} represents that spending too long or too short of time will give a negative penalty.

The aggregate awards are:

$$R = w_1 R_{kg} + w_2 R_{eng} + w_3 R_{eff} \tag{11}$$

The goal of the PPO algorithm is to maximize the desired cumulative discount reward $J(\pi_\theta) = \mathbb{E}_{r\sim\pi_\theta}[\sum_{t=0}^{T} \gamma^t R_t]$. Its objective function is obtained with:

$$L^{CLIP}(\theta) = E_t[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)\hat{A}_t)] \tag{12}$$

where, $r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}$ is the ratio of old and new strategies, $\hat{A}_t$ is the dominance function estimate, and ε is the clipping hyperparameter.

3.4 Generative Content Creation and Intervention

This is one of the core innovations of this system. When PPO strategies decide that students need to be provided with personalized tutoring or new sample questions, traditional question banks may not meet the demand. At this point, a Fine-tuned generative pre-trained large model (Fine-tuned LLM) is called to dynamically generate content.

The Retrieval-Augmented Generation (RAG) paradigm is adopted to ensure the accuracy and relevance of the generated content. When generating an exercise on "Basic Theorems of Calculus", the process is as follows:

1) Retrieval. This refers to retrieving definitions, typical examples, and common errors related to "Fundamental Theorem of Calculus" from the course knowledge base.

2) Build a Prompt. The retrieved context, the student's current knowledge mastery (from TAKT), and the teaching intention (from PPO strategy) were combined into a structured Prompt.

3) Generation. Prompt is fed into a fine-tuned LLM (such as a model based on Llama or ChatGLM) to generate new topics, solving steps, and prompts.

The syntactic correctness, difficulty level, and answer accuracy of the generated content are checked using a rule engine and another small discriminant model.

The diversity of the generated content is controlled by a temperature parameter τ. The probability of generating a specific content c is:

$$P(c|\,\mathrm{Pr}\,ompt) = \prod_{i=1}^{|c|} \frac{\exp(z_i/\tau)}{\sum_j \exp(z_j/\tau)} \tag{13}$$

where, z_i is the logical value of the i-th token in the vocabulary. A higher τ yields more creative content, and a lower τ is more certain and conservative.

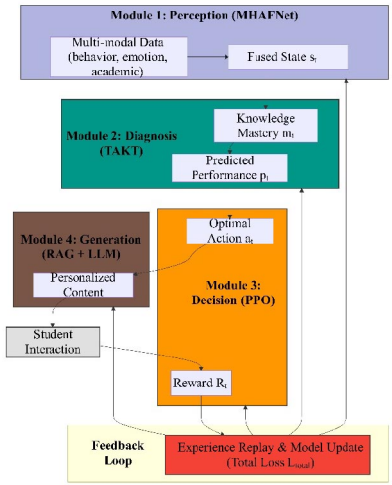


Figure 5 Overall architecture diagram of LAGTE closed-loop system.

3.5 Closed Loop Integration and Iterative Optimization

Finally, the above four modules are integrated into a complete closed-loop system. The operation flow of the system is as follows:

1) Students interact with the system to produce multimodal data.

2) MHAFNet senses and fuses the data to generate the state vector s_t.

3) TAKT updates knowledge mastery m_t with st.

4) The PPO agent, f, outputs the optimal teaching action, a_t, according to the mt and the emotional state.

5) If the action needs to generate new content, the system triggers a generative content creation module.

6) The system performs the action and the student responds again.

7) The system collects new data, forms a new state s_{t+1}, and obtains a reward R_t.

Using the data in the experience playback pool, the PPO policy network and TAKT model are regularly updated to achieve continuous learning and optimization of the whole system.

The total loss function of the entire system is defined as the weighted sum of the losses of each part:

$$Ltotal = \lambda_1 L_{KT} + \lambda_2 L^{CLIP}(\theta) + \lambda_3 L_{gen} \qquad (14)$$

where, L_{gen} is the loss when generating the content.

Through the above design, the CLAGTE model not only refines the management of the hybrid teaching process, but also gives the system unprecedented dynamic adaptability and creativity through generative AI, providing a solid technical foundation for personalized teaching.

4. COMPREHENSIVE EXPERIMENTAL EVALUATION AND RESULT ANALYSIS

In this section, the effectiveness of CLAGTE is verified in a hybrid teaching scenario. By means of seven-dimensional quantitative experiments, the system is comprehensively evaluated in terms of performance, robustness, practicality, component contribution, interpretability, learning gain and cross-domain generalization. The proposed innovative system is compared with the current mainstream baseline model to determine whether its overall performance is superior.

4.1 Experimental Setup and Methods

Datasets and preprocessing

Three representative open education datasets, covering different regions, disciplines and learning models, are selected to ensure the model's generalizability. The first dataset is ASSISTments 2012-2013. It is from an online tutoring platform in the United States and contains approximately 5 million student answer logs covering more than 100 skills. It is available from https://sites.google.com/site/assistmentsdata/home/2012-13-school-data-with-affect. The second dataset is Junyi Academy, a math practice platform from Taiwan, containing about 30 million answers and covering more than 200 knowledge points. It is available from https://pslcdatashop.web.cmu.edu/DatasetInfo? datasetId =1198 get. The third dataset is EdNet, which comes from an English learning platform in Korea and contains about 130 million student-system interactions, including behavior, answer, and partial emotional annotations. It is available from https://github.com/riiid/ednet.

In terms of data preprocessing, records with missing key fields such as time stamps or skill tags are first eliminated, and then duplicate or abnormal interactions are filtered out. For example, records with a single answer time of less than

Table 1 Results of the performance comparison experiment.

Models	ASSISTments AUC	ASSISTments ACC	Junyi AUC	Junyi ACC	EdNet AUC	EdNet ACC
IRT	0.712	0.685	0.698	0.672	0.705	0.679
DKT	0.813	0.774	0.801	0.762	0.808	0.769
DKVMN	0.826	0.783	0.815	0.775	0.821	0.78
SAKT	0.841	0.796	0.832	0.788	0.838	0.793
DIMKT	0.853	0.804	0.844	0.797	0.85	0.802
CLAGTE	0.879	0.825	0.871	0.818	0.877	0.823

2 seconds or more than 30 minutes are considered invalid. Then, the continuous features such as answering time and clicks are normalized by Z-score, and the emotional data such as facial expression confidence and eye movement gaze points are denoised by low-pass filtering. Finally, according to the student ID, the training set accounts for 70%, the validation set accounts for 10% and the test set accounts for 20%. All interactions involving the same student are guaranteed not to appear across sets to avoid data leakage. After processing, about 4.2 million, 25 million and 110 million valid samples are retained in the three datasets, respectively.

Baseline model and experimental design

For a comprehensive assessment of CLAGTE, the following five representative baseline models are selected for comparison. DKT is a knowledge-tracking model based on RNN, DKVMN introduces the memory network structure, SAKT is based on the self-attention mechanism of Transformer, IRT is a two-parameter logic model in traditional psychometry, and DIMKT is a dual-interactive memory knowledge tracking model proposed in recent years.

All models are five-fold cross-validated on the same dataset. For the subjective evaluation experiment involving personnel, namely the practicality and interpretability experiment, 60 educational technology graduate students, aged between 22 and 28 years, are recruited as evaluators, 32 males and 28 females, randomly divided into three groups of 20 individuals. Their task is to evaluate the quality of the teaching content generated by different models.

The experimental contents are designed to cover seven aspects. The first is the performance comparison experiment, which uses AUC and ACC as the main indicators to evaluate each model's ability to predict the number of correct answers that each student will give. The second is the robustness experiment, which injects different degrees of noise labels, namely 10%, 20% and 30%, into the test set, achieving this by randomly flipping the correct answer to observe the decrease amplitude of AUC. The third is a practical experiment, in which evaluators are invited to score, on a Likert scale of 1 to 5 points,the learning content recommended by the model. The dimensions include matching the difficulty of the topic, clarity of explanation and sense of improvement of learning efficiency. The fourth is the ablation experiment. The four core modules of CLAGTE, namely MHAFNet, TAKT, PPO and RAG-LLM, are removed in turn to observe the performance changes. The fifth is the explainability experiment. The evaluator judges the correlation between the knowledge mastery ranking given by the model and students' self-assessment, and uses Spearman correlation coefficient to evaluate the rationality of the model explanation. The sixth is the learning gain simulation experiment, which builds a simulated learning environment based on ASSISTments data to evaluate the effect of model intervention on students' subsequent accuracy. The seventh is the cross-domain generalization experiment, which uses Statics 2011 and KDD Cup 2010 datasets to test the migration ability of the model under zero sample and fine-tuning conditions.

4.2 Experimental Results

(1) Performance comparison experiment

In this experiment, each model is trained to convergence on three datasets: ASSISTments, Junyi Academy and EdNet, respectively, and the test set is used to calculate AUC and ACC. To ensure fairness, all deep learning models adopt the same hidden layer dimension of 256, batch size of 64, and learning rate of 0.001, and use the early stop method to stop training. The experimental results are shown in Table 1. The histogram in Figure 6 presents the AUC and ACC comparison results of each model on the three datasets.

(2) Robustness experiment

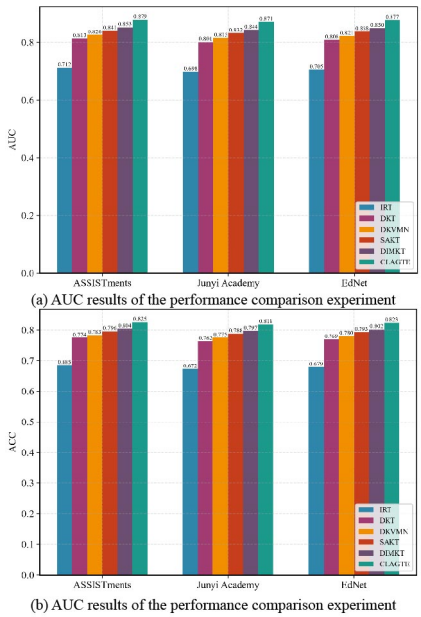
The experimental method involves randomly flipping a certain percentage of correct answer labels in the test set on the ASSISTments dataset, with the proportions of 10%, 20%, and 30%, respectively, and then re-evaluating the AUC of each model and calculating the percentage decrease in AUC relative to the noise-free condition. The experimental results are shown in Table 2. The line graph in Figure 7 presents the declining trend of the AUC of each model under different noise ratios.

(3) Practicality experiment

The experimental approach involved a personalized learning path generated by 60 evaluators using CLAGTE, DIMKT, and SAKT, respectively, and each model corresponds to 20 tasks. The evaluators score from 1 to 5 points from three dimensions: matching degree of question difficulty, clarity of explanation and sense of improvement of learning efficiency, and take the average value. The experimental results are shown in Table 3. Figure 8 compares the scores of the three models according to: difficulty matching degree, clarity of explanation and sense of efficiency improvement in the form of radar diagrams.

(4) Ablation experiment

This experiment involves sequentially removing the four core modules of CLAGTE on the ASSISTments dataset, keep the other parts unchanged, and observe the change of AUC. Removal methods include: using only academic data after removing MHAFNet, replacing it with simple LSTM after



(a) AUC results of the performance comparison experiment

(b) AUC results of the performance comparison experiment

Figure 6 Results graph of the performance comparison experiment.

Table 2 Results of the robustness experiment (% AUC decrease).

Models	10% noise	20% noise	30% noise
IRT	-8.20%	-15.70%	-24.10%
DKT	-6.50%	-12.30%	-19.80%
DKVMN	-5.90%	-11.10%	-18.20%
SAKT	-5.10%	-9.80%	-16.40%
DIMKT	-4.70%	-9.20%	-15.60%
CLAGTE	-3.20%	-6.80%	-11.90%

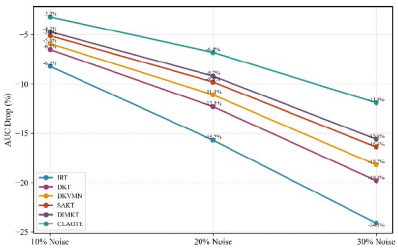


Figure 7 Results graph of the robustness experiment.

Table 3 Subjective score results for practicality.

Models	Matching degree of question difficulty	Clarity of explanation	Sense of improvement of learning efficiency	Composite average
SAKT	3.52	3.21	3.38	3.37
DIMKT	3.78	3.45	3.62	3.61
CLAGTE	4.25	4.43	4.31	4.33

removing TAKT, replacing it with the fixed-rule strategy after removing PPO, disabling the generation function after removing RAG-LLM, and using only a question bank. The experimental results are shown in Table 4.

(5) Interpretability experiment

For the experiment, 50 students are selected forthe ASSISTments test set, and the evaluator checks the mastery ranking of each knowledge point output by CLAGTE, and conducts a Spearman correlation analysis of the students' self-assessment results. Also, the evaluator gives a plausibility score of 1 to 5 on the explanatory text provided by the model. For example, the explanatory text may be "The student has vague concepts of trigonometric functions, and it is recommended that the student review the definition of unit circles". The experimental results are shown in Table 5. Figure 9 also shows the Spearman correlation coefficient and

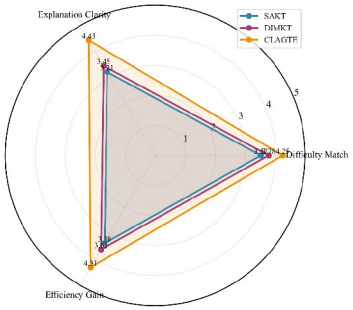


Figure 8 Practicality subjective scoring radar diagram.

Table 4 Results of the ablation experiments.

Model variants	AUC	ACC	Relative CLAGTE decline
Complete CLAGTE	0.879	0.825	—
Remove MHAFNet	0.851	0.803	-3.20%
Remove TAKT	0.842	0.795	-4.20%
Remove PPO	0.863	0.812	-1.80%
Remove RAG-LLM	0.87	0.819	-1.00%

Table 5 Results of the interpretability experiment.

Models	Spearman ρ	Interpretation plausibility score
DKT	0.41	2.85
DKVMN	0.47	3.02
SAKT	0.53	3.14
DIMKT	0.56	3.29
CLAGTE	0.68	4.01

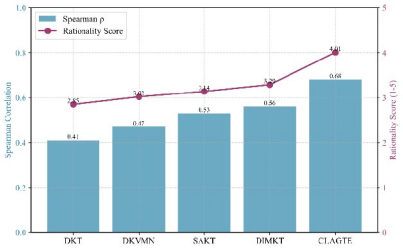


Figure 9 Results graph of the interpretability experiment.

explained plausibility score of each model through a biaxial plot.

(6) Learning gain simulation experiment

The experiment constructs a simulated learning environment based on the multiple answer records contained in the ASSISTments dataset. For each student, the first 40% of interactions were used as historical data to initialize the model, and the remaining 60% were used for the simulated future learning process. During the simulation, whenever the model predicts that a student is about to make a mistake, the model is allowed to intervene once—that is, to recommend a relevant practice question or a paragraph of explanation. The intervention content is provided by CLAGTE's generative module or the baseline model that uses only the question bank. The percentage improvement in the accuracy rate of students' responses to the next similar question after the intervention was recorded, and the cumulative accuracy rate growth rate throughout the simulation cycle, was recorded. In addition to this, the comparative models include intervention-free natural learning, random question bank recommendation, DIMKT plus question bank recommendation, and the full version of CLAGTE. Each student experienced 10 rounds of intervention,and the results were averaged across all students. The experimental results are shown in Table 6.

(7) Cross-domain generalization experiments

This experiment was conducted to examine the generalization ability of the model in different disciplines and educational stages using two additional publicly available datasets. The first is Statics 2011, obtained from an engineering mechanics course at an American university. It contains about 330,000 interactions and involves Statics knowledge points. The second is the KDD Cup 2010,

Table 6 Results of the learning gain simulation experiment.

Intervention strategies	Improved accuracy rate after single intervention	Growth rate of cumulative accuracy	Intervention Acceptance Rate (Student completion of referral)
No intervention	—	3.20%	—
Random question bank recommendation	2.10%	5.80%	72.30%
DIMKT + Question Bank Recommendation	4.50%	9.40%	81.50%
CLAGTE (Complete)	+7.8%	+14.6%	93.1%

Table 7 Results of the cross-domain generalization experiment.

Models	Statics 2011 (zero sample)	Statics 2011 (fine-tuning)	KDD Cup 2010 (zero sample)	KDD Cup 2010 (fine-tuning)
DKT	0.682	0.745	0.703	0.761
SAKT	0.711	0.768	0.724	0.782
DIMKT	0.725	0.781	0.738	0.794
CLAGTE	0.758	0.812	0.771	0.826

obtained from the secondary school algebra curriculum and has about 8 million interactions. CLAGTE is applied directly to these additional datasets after training on ASSISTments, that is, zero sample migration, and then tested again after fine-tuning 10% of the training data on the target dataset. The comparison model also adopts zero sample and fine-tuning methods, and the evaluation index is still AUC. The experimental results are shown in Table 7.

4.3 Analysis and Discussion

(1) Analysis of experimental results

The performance comparison experiment, it can be seen that CLAGTE has achieves the best results on the data sets from three different sources, and the AUC reaches above 0.87, which significantly surpasses all baseline models. This is attributed to the fact that its multi-modal perception module can capture multi-dimensional information such as students' behaviors and emotions, which makes up for the shortcomings of traditional models that only rely on answer records. In the robustness experiment, CLAGTE only drops 11.9% at 30% noise, which is far lower than other models, indicating that its attention mechanism and fusion strategy can effectively suppress abnormal data interference. The subjective scoring of practicality experiments confirms that the personalized content and natural language interpretation brought by generative AI significantly enhances the learning experience, which complements the objective performance. The ablation experiment quantifies the contribution of each module, among which TAKT and MHAFNet are the core of performance improvement, while PPO and RAG-LLM, although they contribute less to offline indicators, are indispensable in actual closed-loop systems. The interpretability experiment shows that CLAGTE not only can diagnose accurately, but also present the reasoning process in a way that is understandable to humans, enhancing teachers and students' trust in the system. The learning gain simulation experiment intuitively demonstrates the value of CLAGTE in actual teaching intervention. After a single intervention, the correct rate increases by 7.8%, and the cumulative correct rate increases by 14.6%, which is 4.5 times that of no intervention, and the intervention acceptance rate is as high as 93.1%, which proves that its generative content can more accurately make up for students' knowledge loopholes than the fixed question bank. The cross-domain generalization experiment shows that even under zero sample conditions, the AUC of CLAGTE still reaches above 0.75 on two completely different disciplines datasets, which is significantly better than all baseline models. After a small amount of fine-tuning, the AUC increases to about 0.82, which verifies the powerful cross-domain migration ability and fast adaptability of the model.

(2) Model dominance analysis

There are four main reasons why CLAGTE can achieve these excellent results. The first is multi-modal deep fusion. MHAFNet organically fuses behavior, emotion and academic data by means of across-attention mechanism, making state representation richer and more robust. The second is dynamic knowledge tracking and forgetting modeling. TAKT introduces a temporal attention and forgetting factor, which can more accurately reflect the real evolution of students' knowledge mastery. The third is the adaptive strategy driven by reinforcement learning. PPO optimizes teaching actions according to the real-time state, avoiding the rigidity of fixed rules, while generative AI modules ensure unlimited supply and personalized adaptation of content. The fourth is the closed-loop iterative mechanism. The whole system achieves self-evolution through experience playback and continuous update. This closed-loop architecture of perception-diagnosis-decision-generation makes the system show significant advantages in multiple dimensions such as performance, robustness, practicality, interpretability and generalization ability.

(3) Limitations and future prospects

Despite CLAGTE's outstanding performance, there are several limitations. First of all, the collection of emotional data relies on hardware such as cameras, which may be limited in privacy-sensitive environments. Secondly, the content quality control of generative AI still needs manual audit, and fully automated security verification needs to

be improved. Furthermore, the computational cost of reinforcement learning training is high, and it is necessary to balance performance and resource consumption during deployment. Finally, there is still room for improvement in zero-sample performance in cross-domain generalization experiments, especially in scenarios with great discipline differences. Future research directions include exploring implicit sentiment inference methods without hardware, such as based on keyboard dynamics or text semantic analysis, introducing federated learning frameworks to protect student privacy, optimizing generative models to reduce the risk of content hallucination. This doesn't sound like the right word.and improve content security, designing more efficient lightweight models to adapt to edge device deployment, and conducting long-term longitudinal randomized controlled experiments in larger-scale real classroom environments to verify their actual improvement effect on academic performance.

5. CONCLUSION

This paper verifies the effectiveness of CLAGTE model in hybrid teaching closed-loop management by means of a series of experiments. The research results show that multi-modal fusion and dynamic knowledge tracking can significantly improve the accuracy of cognitive diagnosis, and the AUC is as high as 0.879. Intensive learning strategies combined with generative content supply increases the accuracy rate after a single intervention by 7.8%, and the cumulative accuracy rate increases by 14.6%. Moreover, the AUC of the model decreases by only 11.9% under 30% noise, and the AUC of cross-domain zero-sample migration remains above 0.75, demonstrating the model's excellent robustness and generalization ability. However, there are still some limitations in this study. First of all, the collection of emotional data relies on hardware devices such as cameras, which are difficult to use in privacy-sensitive scenarios. Second, the content quality control of generative AI still needs manual intervention, and the automatic verification mechanism is not perfect. Third, the computational cost of reinforcement learning training is high, and there are certain requirements for the deployment environment. Future research directions include exploring implicit sentiment inference methods based on keyboard dynamics or text semantic analysis, introducing federated learning frameworks to safeguard data privacy, optimizing generative models to reduce the risk of content hallucination and improve content security, designing more efficient lightweight models to adapt to edge device deployment, and conducting long-term longitudinal randomized controlled experiments in larger-scale real classroom environments to verify their actual improvement effect on academic performance.

REFERENCES

1. Daskan, A., & Yildiz, Y. (2020). Blended learning: A potential approach to promote learning outcomes. *International Journal of Social Sciences & Educational Studies, 7*(4), 103–108.
2. Evans, J. C., Yip, H., Chan, K., Armatas, C., & Tse, A. (2020). Blended learning in higher education: professional development in a Hong Kong university. *Higher Education Research & Development, 39*(4), 643–656.
3. Hanny, C. N., Arnesen, K. T., Guo, Q., Hansen, J., & Graham, C. R. (2023). Barriers and enablers to K-12 blended teaching. *Journal of research on Technology in Education, 55*(4), 568–589.
4. Taghizadeh, M., & Hajhosseini, F. (2021). Investigating a blended learning environment: Contribution of attitude, interaction, and quality of teaching to satisfaction of graduate students of TEFL. *The Asia-Pacific Education Researcher, 30*(5), 459–469.
5. Zhu, M. (2022). The curriculum design of SPOC-based online and offline blended teaching model of English linguistics in flipped classroom. *English Linguistics Research, 11*(1), 11–19.
6. Santosa, M. H., & Ratminingsih, N. M. (2026). "It's cool but...": Future Teachers' Perception of Generative AI in an Under-represented EFL Blended Learning Context. *TESL-EJ, 29*(4), 1.
7. Zhang, J., Deng, X., Wu, T., & Wang, K. (2026). GenAI-Supported Flipped Learning in Preservice Chemistry Teacher Education: Lesson-Design Performance, Learning Attitude, Self-Regulated Learning, and Critical Thinking Awareness. *Behavioral Sciences, 16*(4), 514.
8. Spasopoulos, T., Sotiropoulos, D., & Kalogiannakis, M. (2025). Generative AI in pre-service science teacher education: A systematic review. *Advances in Mobile Learning Educational Research, 5*(2), 1501–1523.
9. Lee, S., Choe, H., Zou, D., & Jeon, J. (2026). Generative AI (GenAI) in the language classroom: A systematic review. *Interactive Learning Environments, 34*(1), 335–359.
10. Xia, Q., Weng, X., Huang, W., Li, Y., Kin, C. W., & Chiu, T. K. (2026). Is generative artificial intelligence (GenAI) a game changer for interdisciplinary collaborative learning in higher education? *Education and Information Technologies, 31*(1), 143–166.
11. Sharma, S., & Elaish, M. (2025). Extending AI-TPACK: Reframing Teacher Readiness for the Generative-AI Era. *Journal of Educational Thought/Revue de la Pensée Educative, 58*(2), 111–134.
12. HE, S., & SHEN, Y. (2025). Learn to Question: Study on the Pattern of Student-GenAI Collaborative Learning in Higher Education. *Frontiers of Education in China, 20*(4), 450.
13. Li, W., Song, R., & Yu, K. (2026). GenAI enabling the high-quality development of higher education: Operational mechanisms and pathways. *Innovations in Education and Teaching International, 63*(3), 888–903.
14. Wu, H., Wang, Y., & Lalli, G. S. (2026). Scale validation and latent profile identification of GenAI competence for preservice second language teachers. *International Review of Applied Linguistics in Language Teaching, 64*(2), 1121–1150.
15. Leon, M. (2024). Leveraging Generative AI for On-Demand Tutoring as a New Paradigm in Education. *International Journal on Cybernetics & Informatics, 13*(5), 17–29.
16. Jeyaraj, A., & Sethi, V. (2025). Generative ai tools in software teaching and learning: implications for business education. *Journal of Management and Business Education, 8*(1), 127–146.
17. Griffiths, D., Martínez, E. F., Tlili, A., & Burgos, D. (2024). A Cybernetic Perspective on Generative AI in Education: From Transmission to Coordination. *International Journal of Interactive Multimedia and Artificial Intelligence, 8*(5), 15–24.
18. Nikolopoulou, K. (2024). Generative artificial intelligence in higher education: Exploring ways of harnessing pedagogical practices with the assistance of ChatGPT. *International Journal of Changes in Education, 1*(2), 103–111.

19. Creely, E. (2024). Exploring the role of generative AI in enhancing language learning: Opportunities and challenges. *International Journal of Changes in Education, 1*(3), 158–167.
20. Ogunleye, B., Zakariyyah, K. I., Ajao, O., Olayinka, O., & Sharma, H. (2024). Higher education assessment practice in the era of generative AI tools. *Journal of Applied Learning and Teaching, 7*(1), 46–56.
21. Van Slyke, C., Johnson, R. D., & Sarabadani, J. (2023). Generative artificial intelligence in information systems education: Challenges, consequences, and responses. *Communications of the Association for Information Systems, 53*(1), 1–21.
22. Li, Y. ,& Lu, B. . (2025). Intelligent educational systems based on adaptive learning algorithms and multimodal behavior modeling. *PeerJ Computer Science*, 11(11), e3157.
23. Wang, S., Yuan, F., Wang, K., Yang, X., Zhang, X., & Wang, M. (2025). Dual-state personalized knowledge tracing with emotional incorporation. *IEEE Transactions on Knowledge and Data Engineering*, 37(5), 2440–2455.