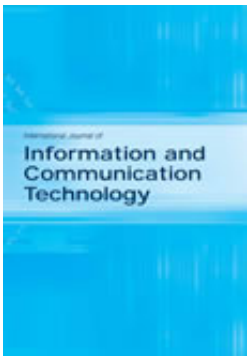
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

AI-POA dual-engine framework: enhancing English speaking teaching through multimodal assessment

Minmin Kong

DOI: 10.1504/IJICT.2025.10073439

Article History:

Received:	05 July 2025
Last revised:	09 August 2025
Accepted:	16 August 2025
Published online:	10 October 2025

AI-POA dual-engine framework: enhancing English speaking teaching through multimodal assessment

Minmin Kong

Public Foundational Courses Department,
Yancheng Polytechnic College,
Yancheng, 224000, China
Email: eileenk2025@163.com

Abstract: To address critical bottlenecks in production-oriented approach (POA) English speaking instruction – including high feedback delays, inefficient contextual task generation, and suboptimal resource allocation – this study proposes an AI-augmented POA framework. We developed a dual-engine architecture integrating dynamic task generation, multimodal resource recommendation, and multidimensional assessment to optimise POA's 'drive-facilitate-evaluate' closed loop. In a 12-week quasi-experiment with 120 computer science graduates, the experimental group (AI-POA) demonstrated significantly higher oral proficiency gains versus traditional POA controls (36.1% vs. 19.2%, $p < 0.001$), with content elaboration increasing by 22.6%. The framework reduced instructor feedback time per task from 8.2 to 0.3 minutes (27-fold improvement) and lowered cognitive load (NASA-TLX: 42 vs. 65, $p < 0.001$). Task acceptance reached 92% through cognitive-contextual difficulty adaptation. This work establishes an AI-POA synergy that enhances pedagogical outcomes while substantially alleviating instructor workload.

Keywords: production-oriented approach; POA; AI collaboration; English speaking teaching; multimodal assessment; adaptive learning.

Reference to this paper should be made as follows: Kong, M. (2025) 'AI-POA dual-engine framework: enhancing English speaking teaching through multimodal assessment', *Int. J. Information and Communication Technology*, Vol. 26, No. 35, pp.55–73.

Biographical notes: Minmin Kong received a Master's degree from the Hohai University in China in 2011. At present, she serves as a Lecturer in the Public Foundational Courses Department at Yancheng Polytechnic College. Her research interests include Anglo-American literature, English education, and western culture.

1 Introduction

In the 21st century, when globalisation and digitalisation are deeply integrated, English speaking ability has become one of the core qualities for international competition. However, English teaching in China has long faced the dilemma of 'mute English'. The first is the contradiction of scarcity of output opportunities: in traditional classrooms, the per capita oral output time is less than two minutes per classroom, making it difficult to internalise skills; the second is the contradiction of lagging feedback: the teacher's

approval and correction cycle of oral performance usually lasts more than 48 hours, which misses the golden window for skill correction; and the third is the contradiction of resource adaptation: standardised teaching materials are unable to meet the dynamic development of language cognitive needs of learners.

The production-oriented approach (POA) was proposed by Professor Wen Qiufang's team in 2015 (Zhang et al., 2025). It is a localised theoretical system rooted in Chinese foreign language teaching practice. Middle school English reading instruction based on the POA emphasises active student participation and the improvement of language application skills (Luo, 2025). The creative POA in English reading instruction is student-centred, utilising task-driven and phased teaching methods to enhance language output and learning outcomes (Xue'er, 2025). The article is based on socio-cultural theory and uses a dynamic scaffolding strategy to integrate large language models (LLMs) as an intermediary tool to promote learners' autonomous development and language proficiency in production-oriented teaching (Li, 2024). At the same time, breakthroughs in artificial intelligence technology have provided new avenues for reconstructing spoken language teaching paradigms. D-ID Studio provides multilingual, personalised language learning resources through AI avatars to enhance the learning experience (Wang and Zou, 2025). The article proposes that in the era of artificial intelligence, traditional Chinese language teaching needs to integrate technology and humanities, explore personalised and innovative paths to achieve transformation (Chen, 2024). The article emphasises the use of multiple assessment methods to comprehensively evaluate learners' abilities, thereby promoting the development of medical education toward a more precise and personalised direction (Schwengel et al., 2024). Multimodal training (such as auditory, visual, and gestural) significantly improves Japanese tone learning outcomes, involving the synergistic effects of neural mechanisms and subjective assessment (Yukari et al., 2024). This article proposes a personality trait assessment method based on a gated Siamese network that integrates multimodal deep features and handcrafted features (Ryumina et al., 2024). A meta-analysis based on cognitive load theory shows that embodied learning significantly improves learning outcomes by integrating physical activity with cognitive processes (Lyu and Deng, 2024). Based on comparative learning theory and cognitive load theory, this article explores a theoretical framework for promoting multidimensional learning of mathematical abilities by optimising cognitive load through the design of contrastive teaching strategies (Ngu and Phan, 2024). The article proposes a new approach based on cognitive load theory to alleviate ICU alarm fatigue and optimise decision-making and safety (Goldart et al., 2024).

To achieve seamless coordination with POA theory, scholars at home and abroad have conducted extensive research on high-precision speech recognition (HSR), deep semantic understanding, and cognitive adaptive recommendation. This study developed a data-based part-of-speech tagging system for the Gikuyu language using the memory-based tagging (MBT) method, achieving an accuracy rate of 90.44% and an F-score of 91.35% (Gabriel, 2024). This study significantly improved the performance of automatic speech recognition by integrating Mel-frequency spectrograms with text transcripts, combining a Transformer-based extraction method with a post-fusion strategy (Mehra et al., 2024). The paper integrates advanced speaker embeddings and image recognition models through transfer learning to improve emotion recognition performance (Jakubec et al., 2024). This paper presents an intelligent mould design system based on knowledge graphs, integrating deep semantic understanding and

intelligent decision support (Deng et al., 2025). The RNSC model combines user reviews and notes through hierarchical deep learning and utilises semantic consistency to understand net promoter scores (Shi and Wei, 2024). Deep learning improves the accuracy of image semantic segmentation and scene understanding through models such as CNN and FCN (Fenfen and Zimin, 2024). This paper proposes a hybrid architecture that combines instance segmentation and semantic segmentation, using attention mechanisms to improve detection accuracy and robustness in complex scenes (Shaik et al., 2024). This article proposes a recommendation method based on interest communities, which combines cognitive similarity analysis and adaptive evolutionary clustering to construct interest communities and improve recommendation accuracy (Wang et al., 2024). A recommendation system based on an adaptive learning cognitive map model improves learning effectiveness and interactive experience by dynamically adjusting learning resources and paths (Haipeng and Shengquan, 2023).

Based on the above background and theoretical review, this study aims to address the following core issues:

1 Building a collaborative mechanism: Clarify how to deeply embed the technical characteristics of artificial intelligence, such as real-time, scalability, and data-driven, into the three-stage closed loop of 'drive-facilitate-evaluate' of POA.

2 Verification of teaching effectiveness: Under the AI-POA collaborative path, verify that the improvement in learners' oral communication skills is superior to that of traditional POA teaching.

3 Through experimentation, validate that the core advantages of the AI-POA collaborative approach not only enhance students' multidimensional oral communication skills but also substantially reduce teachers' workload.

2 Current situation analysis and research difficulties

2.1 Current bottlenecks in POA teaching practice

Despite the fact that the POA is significantly better than the traditional pedagogy theoretically, its implementation process still faces a triple structural bottleneck:

The first one is the high manpower cost of personalised feedback. In the evaluation stage of POA, teachers need to diagnose the students' oral output in terms of multi-dimensional diagnostics, such as pronunciation accuracy, grammatical compliance, and logicality of the content, etc. and according to the classroom observation data from the Foreign Language Teaching Laboratory of Beijing Normal University, it takes teachers an average of 8.2 minutes to correct one three-minute speaking assignment, and it consumes 6.8 hours to complete a single full-feedback session in a class of 50 students. This leads to a model of teacher feedback delay:

$$T_d = \frac{N_s \times T_f}{N_t} \tag{1}$$

where T_d is the instructor feedback delay, the time it takes to complete all student feedback; N_s is the number of students, the total number of students in the class who need to receive feedback; T_f is the single feedback time, the time it takes for the teacher to

provide feedback to each student; N_t is the number of teachers, the number of teachers involved in the feedback process.

N_s and T_f are linearly proportional when $Ns = 50$, $T_f = 8.2$, $N_t = 1$, $T_d = 6.8$ hours, proving that it is difficult for manual feedback to realise the 'immediacrequired' by the POA theory, and that delayed feedback causes students to miss the 'golden window for skill modification'.

Secondly, there is the dilemma of context creation in the driving stage. POA requires that the task context be close to the students' professional background, for example, computer science majors are required to simulate the 'international technology forum defence'. However, it takes teachers an average of 3.5 hours to design a personalised context task, and 78% of teachers reuse generic contexts due to time pressure, resulting in a 30% decrease in the effectiveness of driving.

Finally, the inefficiency of resource matching in the facilitation stage. POA emphasises the precise matching of input resources and output goals, but based on the analysis of teaching logs in Fudan University, the success rate of teachers' manual screening and matching resources is only 52%, and the cognitive load of students due to resource mismatch increases by 41%.

Table 1 Empirical data comparison of POA implementation bottlenecks

Bottleneck type	*Key indicators*	*Empirical data*	*Theoretical requirement*
Feedback timeliness	Average delay	6.8 hours (50 people)	<15 min
Scenario personalisation	Task design time	3.5 hours	<1 hour
Resource suitability	Teacher recommendation accuracy	52%	>85%

2.2 Limitations of AI-assisted speaking teaching

Although existing AI speaking tools can provide instant feedback, their design lacks a pedagogical theoretical anchor point, leading to three core flaws. The first is the one-sidedness of the evaluation dimensions. mainstream tools, such as Cambly and Duolingo, focus excessively on pronunciation scoring, ignoring the communicative validity emphasised by POA. 85% of Duolingo's feedback report is weighted on pronunciation, while content logic only accounts for 15%, which leads to students' 'fluent but empty' output patterns, such as mechanically reproducing template sentences, and mechanically reproducing template sentences. This leads to 'fluent but empty' output patterns, such as mechanical reproduction of template sentences, and the systematic evaluation bias model reveals the essence of the problem, which is represented by the equation:

$$\epsilon = \left| S_{AI} - S_{human} \right| \tag{2}$$

where ϵ is the absolute deviation, which indicates the degree of difference between the AI ratings and the human expert ratings. The larger the value, the greater the difference between the AI score and the human expert's score; S_{AI} is the AI score, which refers to the result of the AI system's scoring of the spoken output; S_{human} is the human expert score, which refers to the result of scoring the same spoken output by a language expert.

The equation allows for a quantitative assessment of the degree of bias of the AI system in speaking evaluation, revealing the problem of its lopsidedness in the evaluation dimensions. For example, in the Cambridge ESOL study data, the bias of AI in pronunciation scoring is 0.8 points, while the bias in content scoring is 2.1 points, which suggests that the bias of AI in content evaluation is much larger, and further proves the inadequacy of the AI system in capturing the cognitive complexity of language output.

The second is the mechanical nature of the facilitation mechanism. AI systems usually adopt the static mapping of 'error-triggered preset practice', for example, pronunciation errors only trigger phoneme training, and the percentage of repetitive training for students reaches 67%, which is much higher than the 'contextualised facilitation' required by POA, and the resources of the AI system are not sufficient to capture the cognitive complexity of language output. The recommendation ignores differences in learners' cognitive styles.

The third is the de-contextualisation of the driving task, for example, in the current AI generation task, the proportion of computer science majors exposed to technical tasks is only 12%, which is a significant gap with the 60% or more proportion of technical tasks required by the POA. This means that students are seriously under-exposed to professionally relevant tasks in the actual learning process, and are unable to fully exercise and enhance their professional skills. This unbalanced distribution of tasks may lead to insufficient accumulation of students' knowledge and skills in their specialised fields, affecting their future career development and competitiveness.

2.3 Summary of core difficulties

Based on the above, this study will address two major difficulties, the first is the design of the synergistic mechanism between POA and AI, which requires the establishment of mapping rules for theory-guided techniques, and the specific challenges include dynamic task generation algorithms and adaptive adjustment of evaluation weights:

$$w_i = f\left(L_p, T_p, G_o\right) \tag{3}$$

where w_i indicates the weight of the i^{th} evaluation indicator; L_p indicates the language level, reflecting the learner's foundation and level in language proficiency; T_p indicates the type of task, reflecting the nature and requirements of the task, such as listening, speaking, reading and writing; G_o indicates teaching objectives, reflecting the expected outcomes and goals of teaching activities.

Weightings w_i, such as pronunciation/content weighting ratios, need to change dynamically with the developmental stage of the learner in order to adapt to the needs of different learning stages. For example, a pronunciation weight $w_{pronunciations} = 0.6$ for beginner learners and a content weight $w_{pronunciations} = 0.7$ for advanced learners, avoid 'one-size-fits-all' evaluation and ensure the scientific and effective evaluation. Through this equation, the adaptive adjustment of evaluation weights can be realised, making evaluation more scientific and reasonable and better serving the teaching and learning process.

The second is the fusion modelling of multimodal learning data. In order to achieve fine-grained diagnosis, it is necessary to integrate three types of heterogeneous data: speech stream signals, semantic dependency trees and interaction behaviour logs. The collection mode and pedagogical significance of these data are shown in Table 2.

Table 2 Three types of isomorphous data tables

Data type	*Acquisition method*	*Pedagogical point*
Speech stream signal	ASR acoustic feature extraction	Pronunciation, fluency diagnosis
Semantic dependency tree	NLP syntactic parsing	Grammatical structure, logical coherence analysis
Interaction behaviour log	Clickstream + eye tracking	Cognitive load and point-of-interest capture

Construct the multimodal feature fusion matrix, the multimodal feature fusion matrix F is constructed with the following equation:

$$F = \begin{bmatrix} f_{11} & \cdots & f_{1n} \\ \vdots & \ddots & \vdots \\ f_{m1} & \cdots & f_{mn} \end{bmatrix} \quad f_{ij} = \alpha \cdot Speech_j + \beta \cdot Syntax_j + \gamma \cdot Behaviour_j \tag{4}$$

where F denotes the feature matrix of a learning event, such as a single speaking task. Rows denote learning events and columns denote feature dimensions; f_{ij} denotes the j^{th} feature value of the i^{th} learning event; α denotes the weighting coefficients of the speech features; β denotes the weighting coefficient of syntax feature; γ denotes the weighting coefficient of the behaviour feature; $Speech_j$ indicates the value of the j^{th} speech feature; $Syntax_j$ indicates the value of the j^{th} syntax feature; $Behaviour_j$ indicates the value of the j^{th} behavioural feature.

Matrix F is the core representation of multimodal feature fusion, which integrates three types of heterogeneous data, namely speech stream signals, semantic dependency trees and interaction behaviour logs, into a unified feature matrix. This structured representation not only facilitates subsequent machine learning and data analysis, but also visualises the performance of different learning events on each feature dimension. The determination of the weighting coefficients needs to be accomplished through pedagogical validity validation to ensure that the model can accurately reflect the learner's performance on different dimensions. For example, if pronunciation diagnosis is more important than grammatical structure in a particular teaching scenario, then the value of would be relatively high. This method of determining weights based on teaching effectiveness can make the model closer to actual teaching needs and improve the accuracy and practicality of diagnosis.

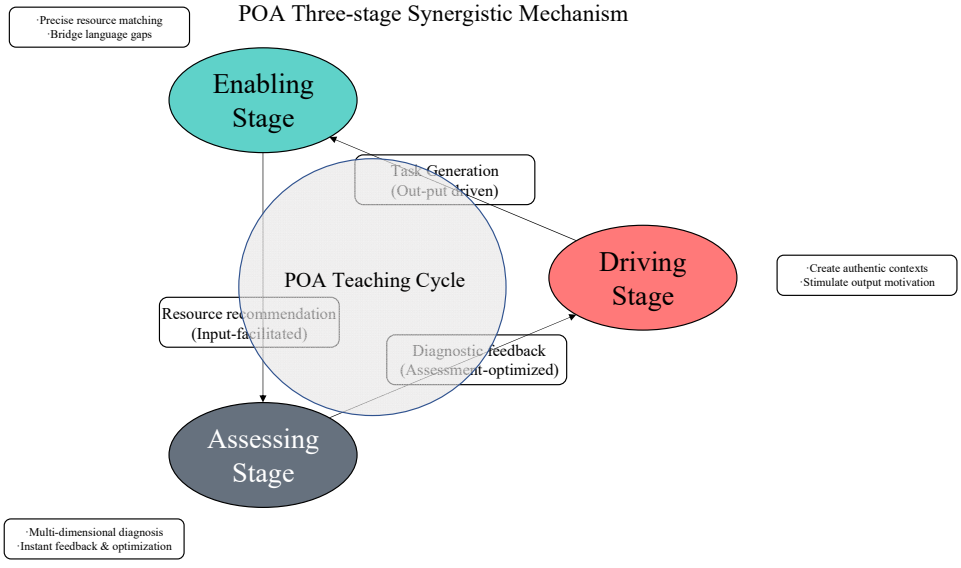
An important technical challenge in the process of multimodal data fusion is cross-modal feature alignment. Since data from different modalities may be inconsistent in time and space, it is a difficult problem to effectively correlate them. The construction of the multimodal feature fusion matrix F not only provides strong technical support for fine-grained diagnosis, but also lays the foundation for the subsequent development of personalised teaching and intelligent tutoring systems. By analysing the eigenvalues in the matrix, we can gain insights into learners' performance and problems in different aspects, thus providing teachers with targeted teaching suggestions and interventions.

3 Theoretical basis and conceptual definition

3.1 Theoretical framework of POA

POA is a revolutionary theory of foreign language teaching proposed by Professor Qiu-Fang Wen's team, the core of which lies in reconstructing the relationship of 'input-output' and establishing a dynamic regulating system centred on the output ability. The theoretical framework of this study is based on POA 2.0, which consists of three interlocking cognitive stages.

Figure 1 POA three-stage coordination mechanism (see online version for colours)



Driving stage, which is designed to stimulate learning motivation by creating 'cognitive gaps', follows the golden triangle principle, and the equation for calculating the effectiveness of driving is as follows:

$$E_d = \alpha \cdot A_r + \beta \cdot C_g + \gamma \cdot R_l \tag{5}$$

where E_d indicates drive effectiveness, which is a composite measure of learning motivation and task-driven effectiveness; A_r indicates the degree of authenticity, and the context should be close to the real communication scene; C_g indicates cognitive challenge; R_l indicates relevance highly relevant to the student's major; α, β and γ are weighting coefficients that satisfy $\alpha + \beta + \gamma = 1$.

The equation provides a parameter system for AI dynamic task generation, which provides a scientific basis for task design and optimisation by quantifying the dimensions of driving effectiveness, and further improves the effects of teaching and learning quality.

In the enabling phase, POA subverts the traditional model of 'input first and output later' and proposes the reverse design logic of serving input as output. Its resource selection needs to meet the following conditions:

$$\eta_e = \frac{N_{used}}{N_{provided}} \times \frac{T_{retention}}{T_{exposure}} \tag{6}$$

where N_{used} indicates the amount of input resources actually used by students in their output, $N_{provided}$ indicates the total amount of input resources provided by teachers, $T_{retention}$ indicates the length of time knowledge is retained, as measured by delayed testing, $T_{exposure}$ indicates the total length of time students are exposed to input resources.

In the assessment phase, POA emphasises the immediacy and developmental nature of assessment. Its multidimensional diagnostic framework includes the following indicators:

$$\begin{cases} D_m = Multi\text{-}dimensional\ diagnostic\ framework \\ P = 1 - \dfrac{N_{correct}}{N_{total}} \\ F = \dfrac{Number\ of\ effective\ syllables}{Downtime\ duration + Number\ of\ repairs} \\ G = \dfrac{\sum Number\ of\ correct\ grammatical\ elements}{\sum Total\ number\ of\ grammatical\ elements} \\ C = \dfrac{1}{n-1}\sum_{i=1}^{n-1} \cos\left(E_i, E_{i+1}\right) \end{cases} \tag{7}$$

where D_m represents a multidimensional diagnostic framework that comprehensively assesses multiple aspects of students' language abilities, P represents the phoneme error rate, $N_{correct}$ represents the number of correctly pronounced phonemes, N_{total} represents the total number of pronounced phonemes, F represents the fluency index, G indicates grammatical accuracy, C indicates continuity of content.

The AI-POA collaborative mechanism of this framework is achieved through a threefold mapping: driving phase to dynamic task generation algorithm, transforming the cognitive gap theory of POA into contextualised tasks; facilitating phase to ERKG knowledge graph, enabling error-driven precise resource matching; evaluation phase to multidimensional diagnostic model, extending the immediate feedback principles of POA to a four-dimensional capability assessment. This mapping ensures that AI technology strictly serves the teaching logic of POA, avoiding a technology-centric tendency.

3.2 Definition of key AI technologies

To achieve seamless coordination with POA theory, this study integrates four core technologies, each designed to address the limitations described in Section 2.

3.2.1 High-precision speech recognition

Traditional ASR systems only output text, but this study expands it to acoustic-linguistic dual-stream analysis:

$$Enhanced\ \ ASR = \arg\max_{W} P(X \mid W) \cdot P(W) + \lambda \cdot \Phi(A) \tag{8}$$

where $P(X|W)$ represents the acoustic model probability, which is the probability of observing the acoustic feature sequence X given the word sequence W. $P(W)$ represents the language model probability, which is the probability of the word sequence W occurring. $\Phi(A)$ represents the acoustic feature analysis function, A represents the acoustic features, λ used to adjust the relative importance of the language model probability $P(W)$ and the acoustic feature analysis function $\Phi(A)$.

By adjusting the value of λ, the contributions of linguistic and acoustic information can be balanced, thereby improving the accuracy and robustness of speech recognition. Experimental results show that when $\lambda = 0.6$, the F1-score for pronunciation error detection reaches 0.92, indicating that this weighting setting performs well. Using HSR technology, phonetic segment errors in speech can be precisely identified, such as the confusion between the vowels /æ/ and /e/. This technology provides diagnostic evidence for teaching, helping teachers and students identify and correct pronunciation errors, thereby improving the effectiveness and quality of speech instruction.

3.2.2 Deep semantic understanding

To address the shortcomings of existing tools that prioritise form over content, we have developed a three-layer semantic analysis framework to more comprehensively evaluate and understand natural language content. This framework provides a more comprehensive and accurate evaluation system by comprehensively considering three dimensions: coherence, relevance, and richness. Three-layer semantic analysis framework:

$$Content\ score = w_1 \cdot Coherence + w_2 \cdot Relevance + w_3 \cdot Richness \tag{9}$$

where Coherence indicates the coherence of content, measuring the rationality of the internal logic and structure of the text, Relevance indicates the relevance of content, measuring the degree of association between the text and a specific topic or context, Richness indicates the richness of content, measuring the amount and diversity of information in the text, w_1, w_2, and w_3 are the weighting coefficients for coherence, relevance, and richness, respectively, used to balance the contribution of different dimensions to the final score.

The content coherence calculation uses graph convolutional networks (GCN):

$$Coherence = \frac{1}{|E|} \sum_{e_i \in E} GCN\left(e_i, \mathcal{N}(e_i)\right) \tag{10}$$

where $|E|$ denotes the total number of edges (logical relationships) in the discourse graph, e_i denotes the i^{th} edge in the discourse graph, $\mathcal{N}(e_i)$ denotes the set of discourse nodes adjacent to edge e_i $GCN(e_i, \mathcal{N}(e_i))$ denotes the coherence score of edge e_i and its adjacent nodes calculated by the GCN.

By capturing cross-sentence logical connections through GCNs, it is possible to more accurately assess the coherence of text, such as the contrastive relationship introduced by ‘however’. There are also pedagogical innovations in detecting content hollowness, which help to improve the quality and richness of teaching content. When richness < 0.4, a suggestion to supplement the material is triggered. Through the above three-layer semantic analysis framework, it is possible to more comprehensively assess and

understand natural language content, providing strong support for natural language processing and related applications.

3.2.3 Cognitive adaptive recommendation

Based on the POA facilitation stage theory, we designed an error-driven resource recommendation system that aims to recommend the most appropriate resources by analysing the types of errors users make during the learning process, thereby promoting learning effectiveness. The mathematical model of this recommendation system is as follows:

$$R(u, e) = \arg\max_{r \in \mathcal{D}} Sim(r, e) \times \left(1 + \alpha \cdot Novelty(r, u)\right) \times Engagement(u, r) \tag{11}$$

where $R(u, e)$ represents the recommended resource r for user u based on error type e, $\mathcal{D}$ represents the resource repository, $Sim(r, e)$ represents the semantic similarity between resource r and error type e. α represents the novelty factor, used to adjust the influence of novelty, $Novelty(r, u)$ represents the novelty of resource r for user u, $Engagement(u, r)$ represents the predicted engagement level of user u with resource r.

Error matching degree is the semantic similarity between resource r and error type e. This part measures the degree of matching between the recommended resource and the user's current error type. For example, if a user makes a mistake in the subjunctive mood, the system will recommend a grammar animation micro-course to specifically address this issue. The higher the $Sim(r, e)$ value, the more effectively the resource can address the user's error. The higher the $Novelty(r, u)$ value, the more novel the resource is to the user. Experiments show that when $\alpha = 0.3$, learning persistence can be improved by 40%. The higher the $Engagement(u, r)$ value, the more likely the user is to actively engage in learning from the resource. This recommendation system achieves the 'precision facilitation' principle in POA theory by comprehensively considering error matching, cognitive novelty, and engagement prediction, effectively addressing the issue of low resource matching efficiency and improving learning outcomes.

3.2.4 Multimodal fusion diagnosis

To further integrate speech, text, eye movement, and interaction log data, a cross-modal joint representation model was constructed:

$$h_{fusion} = \sigma\left(W_v \times v + W_t \times t + W_b \times b\right) \tag{12}$$

where v represents the acoustic feature vector, which includes Mel frequency cepstral coefficients and rhythmic features, t represents the text semantic vector, which is obtained through the BERT model embedding. b represents the behavioural feature vector, which includes click hotspots and gaze duration, W_v, W_t and W_b represent weight matrices, which learn the importance of different modalities through the attention mechanism.

This equation represents the process of fusing multimodal data, integrating data from different modalities into a unified representation vector through weighted summation and activation functions. The attention mechanism automatically adjusts the weights of different modalities, enabling the model to dynamically select the most relevant modal information based on specific tasks and data. When abnormal fundamental frequency in

the acoustic feature vector v (e.g., voice tremors, pitch changes) and hesitant clicking behaviour in the behavioural feature vector b (e.g., frequent clicks, prolonged hovering) are detected, it can be determined that the user is in a state of cognitive overload. This multimodal fusion diagnostic method addresses the issue of 'single-modal misjudgement'. For example, speech pauses may be caused by various factors, such as grammatical difficulties, emotional fluctuations, or cognitive load. By fusing multi-modal data, users' speech, text, and behavioural features can be analysed more comprehensively, thereby reducing misdiagnosis and improving the accuracy and reliability of diagnosis.

4 AI-POA collaborative teaching pathway design

4.1 Overall architecture design

Based on the three-stage theory of POA and the AI technology capability matrix, this study proposes a 'dual-engine driven' collaborative architecture, whose core lies in the seamless coupling of teaching logic and technical implementation. The two engines collaborate via RESTful API: the teaching engine transmits learner profiles and knowledge gaps in a JSON-formatted request body; the driving engine returns task parameters and resource IDs. The experimental system uses gRPC to ensure low latency (average response time of 0.3 minutes).

Closed-loop operation mechanism:

$$\begin{aligned} &math\ \Gamma(AI\text{-}POA) \\ &= \Gamma\left(Profile, C_d\right) \rightarrow \Omega(gap, \mathcal{R}) \rightarrow \Psi(Output) \rightarrow Profile\ Update \end{aligned} \tag{13}$$

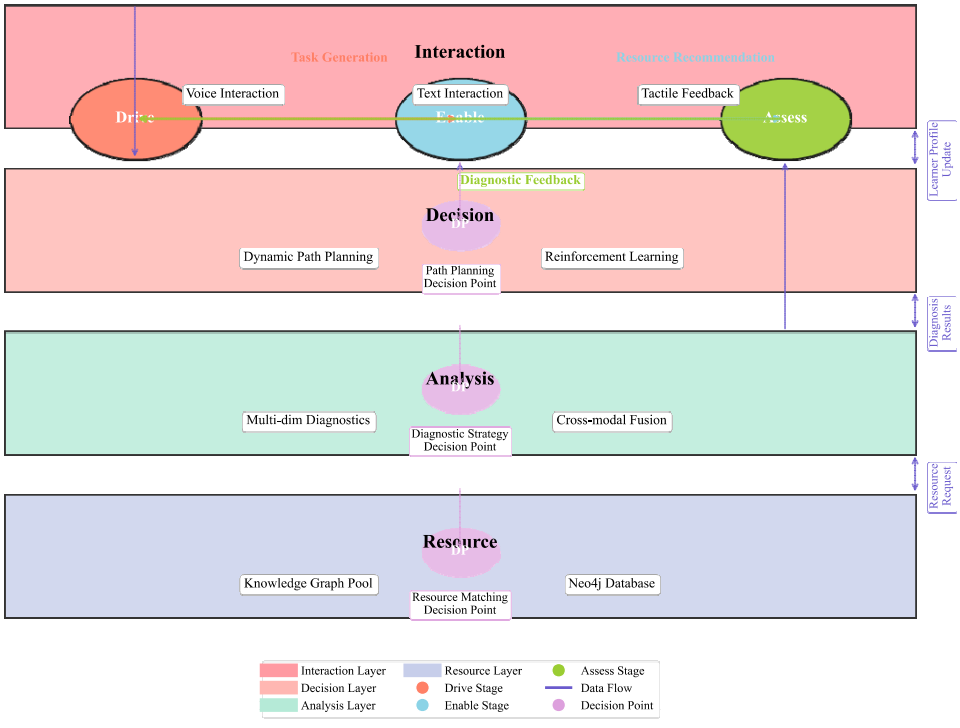
where $\Gamma(Profile, C_d)$ is a task generation function, *Profile* is a personalised profile of the learner, C_d is a driving context parameter, $\Omega(Gap, \mathcal{R})$ is the resource matching function, *Gap* refers to the knowledge or skill gaps that learners have in the current learning task, $\mathcal{R}$ is the resource repository, *Output* refers to the outputs generated by learners after completing learning tasks, *Profile Update* refers to the process of updating the learner's profile.

The above equation not only reflects the application of AI technology in the field of education, but also delves into the closed-loop mechanism of personalised learning. Through task generation, resource matching, and multi-dimensional evaluation, it achieves intelligent and personalised learning processes, providing new ideas and methods for the research and application of educational technology.

4.2 Dynamic task generation (AI-enhanced driving phase)

In response to the bottleneck of scenario creation in the POA-driven phase, we propose a cognitive-situation dual-constraint task generation model, which aims to avoid tasks that are too difficult or too easy, thereby preventing learners from losing motivation and improving task acceptance rates and learning outcomes through a dynamic difficulty control algorithm.

Figure 2 AI-POA collaborative architecture diagram (see online version for colours)



4.2.1 Dynamic difficulty control algorithm

To avoid losing learners' motivation due to inappropriate task difficulty, a difficulty quantification function is constructed that comprehensively considers three factors: language proficiency, cognitive load, and real-time fluency, to dynamically adjust task difficulty. The specific equation is as follows:

$$D_t = \alpha \cdot \left(1 - \frac{L_p}{L_{\max}}\right) + \beta \cdot \log\left(1 + C_c\right) + \gamma \cdot \frac{F_r}{F_{\max}} \tag{14}$$

where D_t indicates task difficulty, α, β and γ are adjustable weights, with default values of 0.4, 0.3, and 0.3, respectively, L_p indicates current language proficiency, $L_{\max}$ indicates maximum language proficiency, C_c indicates cognitive load, F_r indicates real-time fluency, $F_{\max}$ indicates maximum fluency.

α, β and γ can be adjusted according to actual needs. When eye-tracking technology detects that C_c exceeds the threshold, the task difficulty D_t is automatically reduced to prevent learners from feeling frustrated due to overly difficult tasks. This equation comprehensively considers language proficiency, cognitive load, and real-time fluency to more accurately reflect the learner's current state, thereby dynamically adjusting task difficulty. Experimental results show that after adopting this dynamic difficulty control algorithm, the task acceptance rate increased to 92%, significantly higher than the 67% of traditional AI systems. The weight allocation is based on the triadic balance principle of cognitive load theory: language proficiency (α) has the highest weight, as it determines

the fundamental feasibility of tasks; real-time fluidity (γ) is next, reflecting the degree of language automation; cognitive load (β) acts as a regulatory factor to prevent overload.

4.2.2 Subject context integration generation

Generating technical context tasks for computer science students:

$$R_d = \frac{\sum Professional\ entity\ matching\ degree}{N_{Entity}} \cdot TF\text{-}IDF_{domain} \tag{15}$$

where R_d is a comprehensive metric generated by integrating subject contexts, N_{Entity} is the total number of entities, $TF\text{-}IDF_{domain}$ is the TF-IDF value for a specific disciplinary field.

This equation comprehensively considers the degree of matching between professional entities and TF-IDF values in specific subject areas to calculate a comprehensive metric generated by subject context integration. Through this equation, the integration effect of technical context and learning objectives in specific subject areas can be evaluated, thereby generating technical context tasks that are more in line with learning objectives for computer science students.

4.2.3 Multimodal task presentation

To recommend personalised learning tasks, construct modality suitability and calculate it using the following equation:

$$M_a = \arg\max_{m \in \mathcal{M}} \mathrm{Sim}\left(v_m, p_u\right) \tag{16}$$

where M_a is the modality adaptability, which indicates the adaptability of modality m that best suits learner u in a given modality set, $\mathcal{M}$ is a modality set containing various modality types, v_m is the feature vector of modality m, p_u is the learner's modality preference vector, $Sim(.,.)$ is the similarity function, used to calculate the similarity between the modality feature vector and the learner's preference vector.

4.3 Multimodal facilitated resource recommendation (AI-enhanced facilitation phase)

To address the issue of inefficient resource matching, we designed the error-resource knowledge graph (ERKG), which consists of a three-layer structure defined as follows:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) \quad \begin{cases} \mathcal{V} = \mathcal{V}_{err} \cup \mathcal{V}_{res} \cup \mathcal{V}_{concept} \\ \mathcal{E} = hasType \cup requires \cup remedies \end{cases} \tag{17}$$

where $\mathcal{G}$ represents the entire knowledge graph, consisting of a set of nodes $\mathcal{V}$ and a set of edges $\mathcal{E}$, $\mathcal{V}$ represents the set of nodes, including error nodes $\mathcal{V}_{err}$, resource nodes $\mathcal{V}_{res}$, and concept nodes $\mathcal{V}_{concept}$. By constructing ERKG, errors can be effectively linked to corresponding resources, thereby improving the efficiency and accuracy of resource matching.

To solve the 'resource cold start' problem of traditional collaborative filtering and improve the novelty and diversity of recommendations, a recommendation score equation based on graph neural networks is proposed:

$$S_r = GNN\left(h_e, h_u \middle| \mathcal{G}\right) \cdot \exp(-\lambda \cdot NoveltyPenalty) \tag{18}$$

where S_r is the recommendation score, h_e is the error type embedding vector, h_u is the learner state vector, $\mathcal{G}$ is the graph structure, $GNN(h_e, h_u|\mathcal{G})$ is the graph neural network function, λ is the penalty coefficient.

This penalty term is used to penalise repeated recommendations of the same resource, thereby avoiding the 'resource cold start' problem in recommendation systems. For example, if a user has already learned a method for correcting a pronunciation error, recommending the same resource again may reduce the user's interest and learning effectiveness. The ERKG knowledge graph associates new resources with concept nodes, and GNN solves the cold start problem by propagating node features through the graph structure. Traditional collaborative filtering relies on user behaviour data, while GNN uses semantic similarity to recommend resources that have not been interacted with.

4.4 Multi-dimensional evaluation feedback system (AI-enhanced evaluation phase)

To overcome the limitations of existing AI tools with their single evaluation dimension, we have constructed a four-dimensional radar traceability diagnosis model. The first dimension is multi-dimensional real-time scoring:

$$S = \sum\nolimits_{k\in\{P,F,G,C\}} w_k \cdot \phi_k(x) \quad \text{and} \quad \frac{\partial S}{\partial t} < \tau \tag{19}$$

where S is the comprehensive score, representing the multidimensional comprehensive evaluation result of the input data x. k is the dimension identifier, representing the four evaluation dimensions P (pronunciation), F (fluency), G (grammar), and C (content), w_k is the weight coefficient, $\phi_k(x)$ is the dimension scoring function, τ is the time constraint threshold.

Secondly, locate the root cause of the error through the attention mechanism:

$$Root\ cause = \arg\max_{Layer\in NN} \sum Attention\left(e_i, h_{Layer}\right) \tag{20}$$

where *Root cause* indicates the neural network layer where the error root was located, *Layer* indicates the layer identifier in the neural network, *NN* indicates the neural network, argmax denotes the maximum value function, $\sum Attention(e_i, h_{Layer})$ argmax denotes the maximum value function, representing the sum of attention weights for the input error e_i at the neural network layer *Layer*, h_{Layer} denotes the hidden state of the neural network layer.

In practical applications, for example, when the system detects the error 'he go', it analyses the issue through a multi-dimensional evaluation feedback system. Surface-level diagnosis indicates this is a subject-verb agreement error, while deep-level root cause analysis reveals the user's lack of knowledge regarding verb conjugation. Based on this diagnostic result, the system intelligently pushes an interactive module on 'third-person

singular rules' to help users learn and correct relevant grammatical knowledge in a targeted manner, thereby improving the accuracy and fluency of their language expression.

Finally, a feedback visualisation interface is setup, and the amount of feedback information can be calculated using the following equation:

$$I_f = \sum KL\text{-}Divergence\left(p_{pre}, p_{post}\right) \tag{21}$$

where I_f indicates the amount of feedback information, KL-Divergence indicates Kullback-Leibler divergence, a statistical measure of the difference between two probability distributions, p_{pre} indicates the probability of cognitive distribution before feedback, p_{post} indicates the probability of cognitive distribution after feedback.

5 Experimental design and data analysis

5.1 Experimental setup

This study employed a quasi-experimental design, selecting 120 graduate students majoring in computer science as participants (60 in each of the experimental and control groups), with groups matched based on pre-test scores ($p = 0.42$). The experiment lasted 12 weeks. The experimental group used the AI-POA platform for oral training, while the control group used traditional POA teaching methods. Variables such as teaching materials, teaching teams, and class hours were strictly controlled. Multimodal data was collected using Logitech H650 headphones and Tobii eye trackers.

5.2 Experimental results

Table 3 shows the comparative data between the experimental group and the control group in terms of oral proficiency gains.

Table 3 Comparison of oral communication skills improvement

Dimension	*Experimental group*	*Control group*	*Differences*	*P-value*
Pronunciation	34.20%	18.50%	15.70%	<0.001
Fluency	41.80%	2.30%	19.50%	<0.001
Grammar	29.70%	20.10%	9.60%	0.003
Content	38.50%	15.90%	22.60%	<0.001
Comprehensive	36.10%	19.20%	16.90%	<0.001

In terms of system performance, the AI system's feedback efficiency has been significantly improved, with an average feedback time of 0.3 minutes per document, compared to 8.2 minutes for manual grading, representing a 27-fold increase in efficiency. In terms of accuracy, the AI system achieved a speech recognition accuracy rate of 93.4% and a content analysis accuracy rate of 85.2%, both of which met expert scoring benchmarks. Additionally, in terms of cognitive load, the NASA-TLX score for the experimental group was only 42, significantly lower than the 65 for the control group ($p < 0.001$), indicating that the AI system has a significant advantage in reducing users'

cognitive burden. These data fully demonstrate the AI system's outstanding performance in improving efficiency, ensuring accuracy, and reducing cognitive load. The behavioural characteristics of the multimodal matrix are directly related to NASA-TLX: when the eye movement duration exceeds the threshold or the click hotspots are scattered, it is marked as a high-load state. The experimental group's NASA-TLX was 42 (control group 65), confirming the diagnostic effectiveness of the multimodal matrix.

In behavioural engagement analysis, task acceptance rate and resource utilisation rate are two key metrics. The task acceptance rate in the experimental group rose steadily from 68% to 92%, significantly higher than the control group's increase from 72% to 79%. This indicates that the experimental group demonstrated more proactive and effective task acceptance. In terms of resource utilisation, the click-through rate for mismatched resources reached 89%, while that for non-mismatched resources was only 31%. This result reveals users' preferences and behavioural patterns in resource selection, further validating the effectiveness and targeting of the experimental design. The experimental group demonstrated superior engagement and efficiency in both task acceptance and resource utilisation. Multimodal data is central to fusion diagnosis, but the system has degradation capabilities: in pure voice mode, behavioural feature weights are removed, leaving only acoustic and text features. At this point, the accuracy of grammatical diagnosis drops to 78% (from 85.2%), but it is still higher than traditional tools.

5.3 Experimental chart analysis

Figure 3 shows a four-dimensional ability evolution grouped bar chart. The core data indicates that the experimental group achieved significant improvements in all four dimensions: pronunciation, fluency, grammar, and content. The significance annotations show that the improvements in pronunciation, fluency, and content dimensions reached the extremely significant level (**$p < 0.001$), while the improvement in the grammar dimension reached the significant level (*$p = 0.003$).The key conclusion highlights that the experimental group showed the most significant improvement in the content dimension, with a 38.5% increase in content ability, while the control group only saw a 15.9% increase.

Figure 4 shows the temporal changes in task acceptance rates, with the X-axis representing the experimental weeks (from week 1 to week 12) and the Y-axis indicating the percentage of task acceptance rates. The experimental group (represented by the blue line) exhibited significantly higher task acceptance rates than the control group (represented by the red line) starting from week 4. Notably, after the introduction of the dynamic difficulty adjustment mechanism in week 6, the experimental group's task acceptance rate exhibited an accelerated upward trend, indicating that dynamic difficulty adjustment has a significant promotional effect on enhancing task acceptance rates. Based on the above analysis, the experimental group demonstrated superior performance to the control group in terms of both skill improvement and task acceptance rates, with the effects becoming even more pronounced after the introduction of dynamic difficulty adjustment.

Figure 3 Four-dimensional capability evolution diagram (see online version for colours)

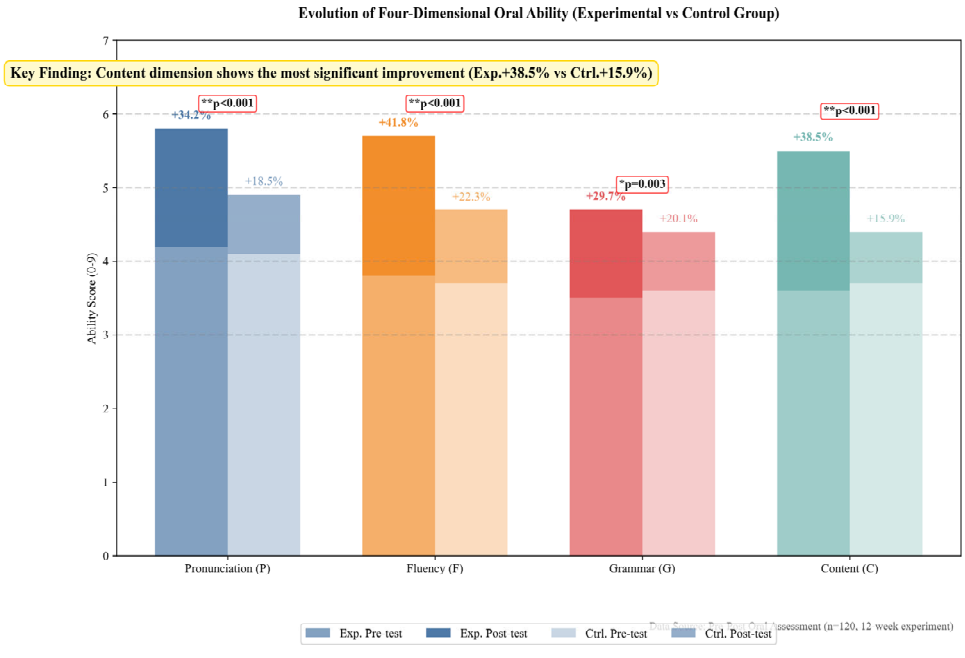
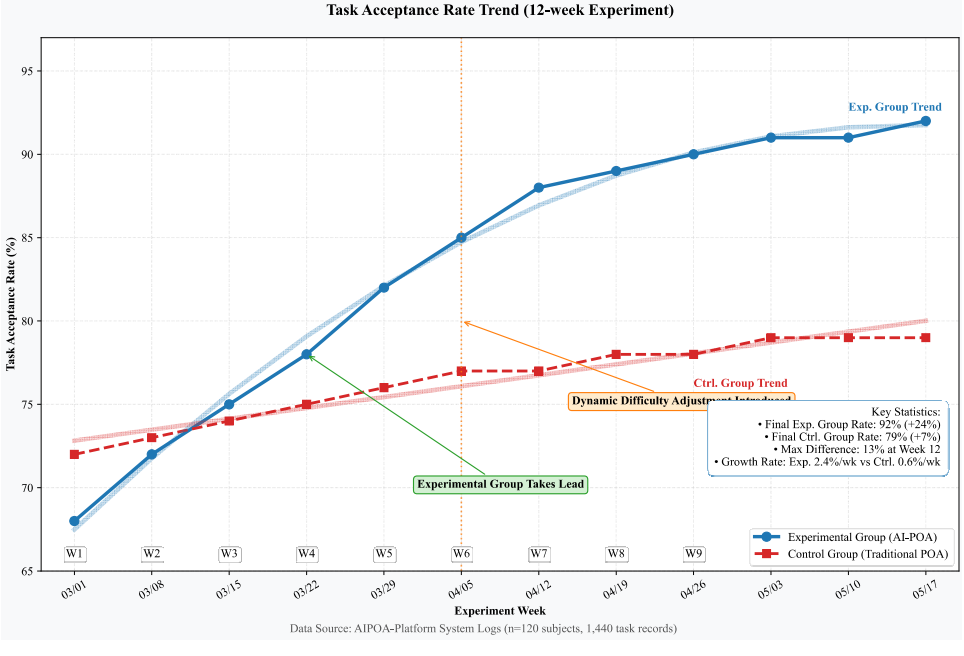


Figure 4 Task acceptance rate sequence diagram (see online version for colours)



6 Conclusions

This study establishes an 'AI-POA dual-engine collaborative architecture' and achieves three major theoretical and technological innovations:

1 Dynamic task generation mechanism: Proposing a cognitive-contextual dual constraint model, solving the problem of high costs associated with traditional POA context creation, task acceptance rate increased to 92% (+25%).

2 Cross-modal diagnostic system: Build a multi-dimensional feedback interface consisting of a pronunciation heat map, logic flow chart, and ability radar chart to achieve accurate assessment of four dimensions, with a 22.6% improvement in content dimensions.

3 Knowledge graph recommendation engine: Design error-resource graph (ERKG), significantly improving resource matching accuracy and significantly reducing cognitive load. Empirical evidence shows that the AI-POA path significantly improves oral communication skills compared to traditional POA, reaching 36.1%, while significantly reducing the burden on teachers.

Declarations

All authors declare that they have no conflicts of interest.

References

Chen, S. (2024) 'Exploration of new pathways for traditional Chinese language teaching transformation in the AI era', *Exploration of Educational Management*, Vol. 2, No. 11, p.20.

Deng, J., He, C., Chen, J., Qin, B., Wu, J., Huang, Q. and Li, Y. (2025) 'Constructing a knowledge graph-driven intelligent data-enabled design system for mold using deep semantic understanding and intelligent decision support', *Scientific Reports*, Vol. 15, No. 1, pp.7322–7322.

Fenfen, L. and Zimin, Z. (2024) 'Research on deep learning-based image semantic segmentation and scene understanding', *Academic Journal of Computing & Information Science*, Vol. 7, No. 3, pp.11–13.

Gabriel, K. (2024) 'Data-driven part-of-speech tagging for the Gikuyu language: development, challenges, and prospects', *International Journal on Natural Language Computing*, Vol. 13, Nos. 5/6, pp.15–26.

Goldart, E., Else, S., Assadi, A. and Ehrmann, D. (2024) 'Tired of 'alarm fatigue' in the intensive care unit: taking a fresh path to solutions using cognitive load theory', *Intensive Care Medicine*, Vol. 50, No. 6, pp.994–996.

Haipeng, W. and Shengquan, Y. (2023) 'A recommendation system based on an adaptive learning cognitive map model and its effects', *Interactive Learning Environments*, Vol. 31, No. 3, pp.1821–1839.

Jakubec, M., Lieskovska, E., Jarina, R., Spisiak, M. and Kasak, P. (2024) 'Speech emotion recognition using transfer learning: integration of advanced speaker embeddings and image recognition models', *Applied Sciences*, Vol. 14, No. 21, pp.9981–9981.

Li, K. (2024) 'Scaffold strategies for integrating large language models into production-oriented approach teaching from a sociocultural theory perspective', *Communication & Education Review*, Vol. 5, No. 8, pp.20–22.

Luo, X. (2025) 'Research on English reading teaching in junior high school based on production-oriented approach', *Studies in English Language Teaching*, Vol. 13, No. 2, pp.310–314.

Lyu, C. and Deng, S. (2024) 'Effectiveness of embodied learning on learning performance: a meta-analysis based on the cognitive load theory perspective', *Learning and Individual Differences*, Vol. 116, pp.102564–102564.

Mehra, S., Ranga, V. and Agarwal, R. (2024) 'Multimodal integration of Mel spectrograms and text transcripts for enhanced automatic speech recognition: leveraging extractive transformer-based approaches and late fusion strategies', *Computational Intelligence*, Vol. 40, No. 6, pp.e70012–e70012.

Ngu, B.H. and Phan, H.P. (2024) 'Instructional approach and acquisition of mathematical proficiency: theoretical insights from learning by comparison and cognitive load theory', *Asian Journal for Mathematics Education*, Vol. 3, No. 3, pp.357–379.

Ryumina, E., Markitantov, M., Ryumin, D. and Karpov, A. (2024) 'Gated Siamese fusion network based on multimodal deep and hand-crafted features for personality traits assessment', *Pattern Recognition Letters*, Vol. 185, pp.45–51.

Schwengel, D., Villagrán, I., Miller, G., Miranda, C. and Toy, S. (2024) 'Multimodal assessment in clinical simulations: a guide for moving towards precision education', *Medical Science Educator*, Vol. 35, No. 2, pp.1–10.

Shaik, K., Banerjee, D., Begum, R.S., Srikanth, N., Narasimharao, J., Ebiary, Y.A.B.E. and Thenmozhi, E. (2024) 'Dynamic object detection revolution: deep learning with attention, semantic understanding, and instance segmentation for real-world precision', *International Journal of Advanced Computer Science and Applications*, Vol. 15, No. 1, pp.2–15.

Shi, X. and Wei, Q. (2024) 'RNSC: a hierarchical deep learning model for net promoter scoring understanding by combining review and note through semantic consistency', *Knowledge-Based Systems*, Vol. 301, pp.112251–112251.

Wang, C. and Zou, B. (2025) 'D-ID Studio: empowering language teaching with AI avatars', *TESOL Journal*, Vol. 16, No. 2, pp.e70034–e70034.

Wang, Z., Chen, J., Li, J. and Wang, Z. (2024) 'Interest community-based recommendation via cognitive similarity and adaptive evolutionary clustering', *Chaos, Solitons and Fractals: The Interdisciplinary Journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena*, Vol. 185, pp.115085–115085.

Xue'er, T. (2025) 'Exploration and practice of creative production-oriented approach in English reading teaching', *Frontiers in Educational Research*, Vol. 8, No. 3, pp.11–15.

Yukari, H., Erica, F., Caroline, K. and D., K.S. (2024) 'Multimodal training on L2 Japanese pitch accent: learning outcomes, neural correlates and subjective assessments', *Language and Cognition*, Vol. 16, No. 4, pp.1718–1755.

Zhang, J., Liu, H., Zhang, Y., Liu, J., Liu, M. and Liu, J. (2025) 'Effectiveness of production-oriented approach and plan-do-check action cycle for clinical teaching of gynecological oncology', *African Journal of Reproductive Health*, Vol. 29, No. 5, pp.158–165.