ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Research on the Effect Evaluation of Private Undergraduate German Course Reform in the Context of AI Empowerment

Honglian Bian[1,*]

[1] College of International Language and Culture, Xi'an FanYi University, Xi'an 710105, Shaanxi, China

SUMMARY: *Under the background of artificial intelligence, learning analytics and educational data mining continuing to enter the foreign language teaching scene, the reform of private undergraduate German curriculum has the basis of process data collection, feature modeling and quantitative evaluation. Focusing on the evaluation needs of course goal achievement, classroom interaction quality, language ability growth and teaching support adaptation, this paper constructs an evaluation model for the effect of private undergraduate German course reform in the AI-enabled context. The classroom behavior records, assignment texts, test scores, platform access trajectories and feedback data in 4120 effective samples are uniformly cleaned, coded and associated mapped. The model consists of three parts: multi-source data representation, key feature modeling and evaluation output mechanism, and combines attention weighting, gated fusion and hierarchical scoring methods to complete reform effect identification and difference analysis. The experimental results on the validation set show that the evaluation accuracy of the model is 92.7%, the Recall is 91.4%, the F1 score is 90.9%, and the average output delay is 1.6 seconds. The model can reflect the effect and change characteristics of the German curriculum reform more stably, and can simultaneously show the association changes between vocabulary training, oral interaction, writing revision and stage evaluation. It provides continuous calculation basis and quantitative reference for course content adjustment, teaching method revision and learning support configuration.*

KEYWORDS: *Artificial intelligence; German curriculum reform; Effect evaluation; Learning analytics*

1 Introduction

With the introduction of artificial intelligence, large language models and learning analytics into foreign language education, curriculum implementation, learning feedback and effectiveness assessment have shifted from empirical observation to data-driven expression. The text, speech, behavior trajectory and phase evaluation in language teaching can be continuously collected, structurally stored and entered into the unified computing link, which makes the effectiveness of curriculum reform have a computable, comparable and traceable representation basis.

Schmidt and Strasser[1] reviewed the application of artificial intelligence in foreign language learning, pointed out that intelligent language learning software is promoting the evolution of foreign language classroom to the direction of self-adaptation and

*sophiedn2026@163.com
https://doi.org/10.65102/is2026293 Published: 30 April, 2026

individualization, and emphasized the necessity of collaborative modeling of linguistics, pedagogy and computer science. Huang, Hew, and Fryer[2] systematically sorted out the research context of chatbot support language learning, and believed that conversational systems have stable value in terms of instant interaction, feedback generation, and learning companiency. Klimova and Ibna Seraj[3] summarized the chatbot research in university EFL scene and gave the implementation direction of the combination of technology access and teaching organization in university language teaching.

Annamalai et al. [4] investigated the English learning process supported by chatbots from the perspective of self-determination theory, and found that digital interactive environment could enhance learners' sense of autonomy, competence and connectiveness. Annamalai et al. [5] further analyzed the use experience of chatbots for English learning in the context of higher education, indicating that learners are highly sensitive to performance expectations, operating costs and interaction convenience. Qiao and Zhao[6] showed through a comparative study that AI-supported teaching significantly promoted the development of oral ability and self-regulation behavior. Kohnke and Moorhouse and Zou[7] discussed the application boundaries of ChatGPT after it entered language teaching and proposed that generative tools are changing the way tasks are designed, feedback is organized, and learning is supported.

In terms of evaluation technology, Gao et al. [8] conducted a systematic review of the research on automatic evaluation of college text answers, and pointed out that natural language processing, automatic scoring and fast feedback mechanism have become important computing paths of educational evaluation. Wang, Cheung and Chai[9] sorted out the research prospect of language learning from the perspective of human-computer interaction, and proposed that a new collaborative relationship between artificial intelligence, teachers and learners is forming. Based on the meta-analysis results, Xu and Wang[10] pointed out that the intervention of artificial intelligence in the improvement of English learning achievement had a high effect size, indicating that the data-based support was no longer limited to resource assistance, but had a clear role in the improvement of learning results.

The existing results provide sufficient technical reference for foreign language curriculum reform, but the research objects mostly focus on English learning support, oral practice, automatic feedback and general language training, and there are still few computational studies directly oriented to the effect evaluation of private undergraduate German curriculum reform. Private undergraduate German courses have the characteristics of strong application orientation, obvious differences in learning basis, high density of classroom activity organization and a large proportion of process evaluation. It is difficult to completely present the change trajectory after the implementation of the reform solely relying on final grades or classroom observation. There is a close relationship between course goal achievement, classroom interaction level, task participation quality, text expression performance and platform learning behavior. Only when multi-source data are put into a unified model, more stable evaluation results can be obtained.

In the private undergraduate scenario, German course reform is often accompanied by blended teaching, task chain reorganization, frequency improvement of oral training, refinement of formative evaluation and reconfiguration of platform resources. The transcribed speech text, assignment corpus, click stream, class attendance record, discussion speech and stage evaluation results generated in the course operation do not exist in isolation, but constitute a heterogeneous data network that can be used for course diagnosis. Mapping these data into a unified feature space, combined with the weight learning and hierarchical decision mechanism, the differences after the implementation of the reform can be transformed from the empirical description into quantitative evidence. This method not only preserves the continuity of language learning activities, but also reveals the coupling changes between

different teaching links, so that the curriculum reform evaluation turns from single point scoring to whole process calculation. Therefore, the evaluation results have more interpretable and horizontal comparison value.

Based on the above background, this paper constructs an evaluation model for the effect of curriculum reform for private undergraduate German course reform in the context of AI empowerment. The unified representation, feature extraction and result judgment of classroom behavior records, homework texts, test scores, platform access trajectories and questionnaire feedback are carried out, and on this basis, the influence differences of key features on the evaluation results are analyzed. It provides continuous calculation basis and implementation reference value for German course content adjustment, teaching method revision and learning support configuration.

2 Related Research

Literature [11] has conducted a comprehensive study on automatic writing evaluation in second language classroom, sorted out the main paths of automatic scoring, feedback generation and classroom integration, and proposed that automatic writing evaluation is changing from a single scoring tool to a computational teaching module with both diagnostic and support functions. The study shows that assessment activities in language courses can already be continuously recorded through algorithmic feedback.

Literature [12] examined the influence of automatic writing evaluation on foreign language learners' writing self-efficacy, self-regulation, anxiety level and writing performance, and constructed an observation framework of writing performance under the condition of technology intervention. The results show that the algorithm feedback can enhance learners 'ability to monitor the writing process, and also provide a basis for the dual-index modeling in the evaluation of curriculum reform effect.

Literature [13] studied the learning engagement performance after the integration of automatic writing evaluation and teacher feedback in the technology-enabled context, and proposed that platform feedback and teacher evaluation should be regarded as complementary information sources. This study shows that learners will form different levels of response behavior after receiving multi-source feedback, and this level change is suitable for entering the unified assessment model through feature coding.

Literature [14] used ChatGPT in autonomous foreign language learning scenarios, and analyzed learners' performance under the conditions of self-paced, instant question answering and continuous revision. We point out that generative tools are able to extend the temporal boundaries of learning activities and enable the learning process to retain a more complete digital trajectory, which provides a computable basis for platform access, task completion, and language revision behavior in curriculum reform evaluation.

Literature [15] investigated the use perception of ChatGPT in language learning from the perspective of technology acceptance model, and proposed that perceived usefulness, ease of use and interaction trust would jointly affect the continuous use intention. The significance of this study is to transform subjective experience into quantifiable variables, so that attitude data can enter the assessment link together with classroom performance and learning behavior data.

Literature [16] discusses the application effect of ChatGPT after it enters foreign language education, and the research shows that generative artificial intelligence can form stable support in language practice, content organization and feedback generation. This paper proposes that the teaching value of AI tools is not limited to the provision of resources, but

lies in its ability to participate in the reconstruction of teaching process, which provides a technical perspective for the phased evaluation of the effectiveness of curriculum reform.

Literature [17] studies the application of ChatGPT in autonomous online language learning, and analyzes the way learners use the functions of explanation, example, continuation and error correction. It is pointed out that high-frequency interaction behavior in autonomous learning environment has strong representation significance, and it is suitable to extract key features such as learning rhythm and task viscosity through behavior sequence modeling.

Literature [18] focuses on the foreign language education scenario at the university level, and investigates the learning activity organization and teaching adaptation mode supported by ChatGPT. It was proposed that the application of AI in foreign language teaching in colleges and universities should take into account tool functions, curriculum objectives and teacher regulation, which indicated that curriculum reform evaluation should also pay attention to the synchronous change of teaching activity structure and learning support structure.

Literature [19] analyzed the characteristics of behavioral engagement, cognitive engagement and emotional engagement after the use of ChatGPT, taking the revision of graduate academic texts as the object. The research gives the division method of multi-dimensional input, which enables the text revision activity to be transformed into the joint analysis of multi-source behavioral indicators, and provides a transferable idea for the evaluation of writing training in the German curriculum reform.

Literature [20] used mixed method to investigate learners' participation performance after using ChatGPT's automatic written error correction feedback, and pointed out that learners' acceptance, screening and reprocessing of feedback would show obvious differences. This study proposed that the actual effect of automatic feedback should be determined by combining the subsequent revision behavior, so the evaluation method combining output results and process characteristics was more suitable for the analysis of curriculum reform effect.

Table 1: Summary of the results of the existing studies

Reference	Research Focus	Technical Approach	Findings	Implications
[11]	Automated writing evaluation	Automated scoring and feedback	Supports diagnostic tracking	Continuous evaluation
[12]	Changes in writing performance	Feedback intervention analysis	Links behavior and competence	Dual indicators
[13]	Multi-source feedback fusion	Hierarchical feature analysis	Differentiation in engagement levels	Weighted encoding
[14]	Autonomous foreign language learning	Generative question-answer-based revision	Preserves digital traces	Platform features
[15]	Technology acceptance perception	TAM variable modeling	Attitudes influence usage	Perceptual features
[16]	AI integration into foreign language education	Generative feedback support	Reconstructs participation processes	Stage-based judgment
[17]	Online learning behavior	Behavioral sequence modeling	Interaction can be represented	Task stickiness
[18]	Foreign language adaptation in universities	Activity structure analysis	Tool-task alignment	Support structure
[19]	Engagement in text revision	Multidimensional engagement analysis	Joint representation of engagement	Writing evaluation
[20]	Automated error-correction feedback	Mixed-method analysis	Combined with subsequent revision	Output process

Existing research has formed a technology accumulation in automatic writing evaluation, generative feedback, technology acceptance, learning engagement, and autonomous learning behavior. The assessment activities in foreign language courses are shifting from single score judgment to continuous judgment supported by multi-source data, and the evaluation objects are also expanding from result text to interactive behavior, emotional tendency and revision process. The reform of private undergraduate German course emphasizes more on classroom participation, task-driven, stage training and the synchronous promotion of application expression. Heterogeneous data such as text, speech, behavior and feedback will be continuously generated during the course operation. These data were incorporated into a unified feature space, and combined with the weight learning, hierarchical judgment and difference analysis mechanism, the effect changes in curriculum reform could be identified. Based on this, this paper constructs an evaluation model for the effect of German course reform in private undergraduate colleges in the context of AI empowerment. Through the course process data representation, key feature modeling and evaluation output judgment, a calculation path for the effect identification of German course reform is formed.

3 Effect evaluation model of private undergraduate German course reform in the context of AI empowerment

3.1 AI Effect Evaluation Model Architecture for Private undergraduate German Course Reform

The effect evaluation of German curriculum reform is in the middle of the link of foreign language teaching adjustment for private undergraduates, which not only connects the implementation of curriculum objectives, the change of classroom organization and the update of resource allocation, but also affects the revision of subsequent task training and support methods. After entering the AI-enabled context, class speeches, assignment texts, test scores, platform access trajectories and stage feed-back can be continuously collected and converted into structured records, which provides the basis for unified modeling and continuous judgment of curriculum reform effectiveness. The design of the model architecture is no longer just to summarize the results, but to organize the behavior changes, text changes and ability changes in the course implementation into computable, comparable and writable evaluation links.

The effect evaluation of the German course reform for private undergraduates should not only keep the stage scores or the mean of the questionnaire. There was a continuous correlation between vocabulary training intensity, grammar task completion status, oral interaction frequency, text revision amplitude and stage assessment performance. Changes in any type of teaching activities would leave identifiable traces on learning behavior and ability growth. Based on this understanding, this paper divides the AI effect evaluation model into data access layer, unified mapping layer, timing fusion layer, effectiveness judgment layer and feedback writeback layer. The data access layer is responsible for collecting classroom behavior records, homework corpus, evaluation results and platform logs. The unified mapping layer completes the standardization and semantic projection of heterogeneous data. The temporal fusion layer retained the cumulative changes in weekly advancement. The effectiveness judgment layer outputs comprehensive scores and grade labels. The feedback writeback layer passes the results to the course content adjustment, task rhythm revision, and learning support configuration modules.

As shown in Fig. 1, the model takes "multi-source acquisition -- unified mapping -- timing fusion -- level determination -- feedback writeback" as the main line. On the one hand, this

structure retains the dynamic information in the course of curriculum reform, on the other hand, it ensures that different teaching units can be compared horizontally under the same scale. Compared with the methods described only by the results of a single test or teacher's experience, this structure is more suitable for the implementation environment in which blended teaching, task-driven training and formative assessment are promoted in parallel in private undergraduate German courses.

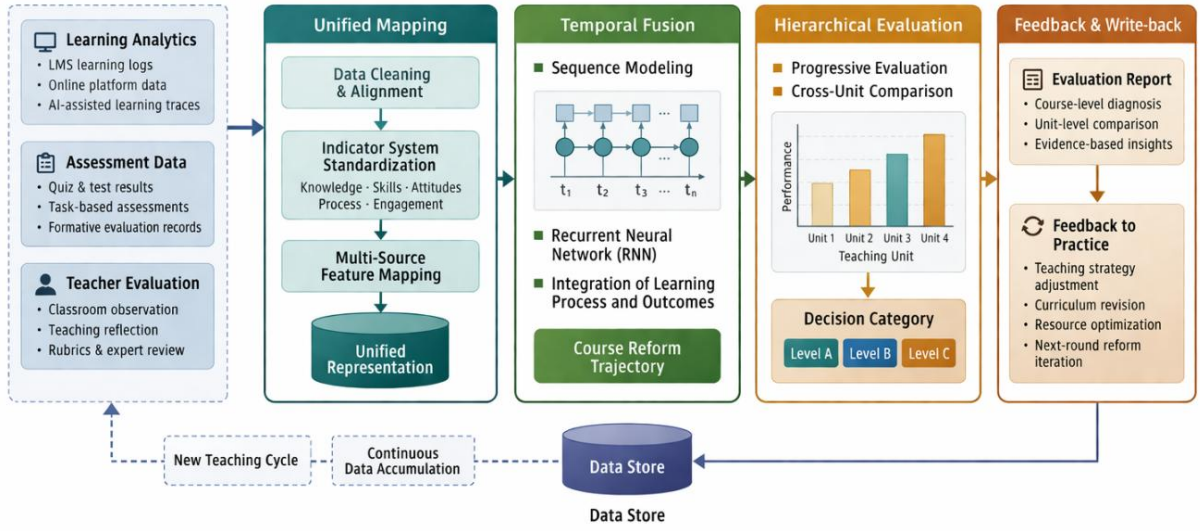

Figure 1: Model architecture for evaluating the effect of German course reform for private undergraduate students in the context of AI empowerment

In the unified input stage, the model writes the classroom behavior matrix, assignment text embedding, test result vector and platform log vector into the course observation unit, and completes the basic representation mapping through adaptive gating, which is calculated as follows.

$$H_t = \phi\left(\Gamma_r R_t P_r + \Gamma_x X_t P_x + \Gamma_q Q_t P_q + \Gamma_l L_t P_l + b_h\right) \tag{1}$$

Here, H_t represents the base representation of the t teaching period. R_t stands for classroom behavior matrix; X_t represents the job text embedding. Q_t represents the phase test vector; L_t represents the platform log vector; P_r, P_x, P_q, P_l denote the corresponding projection matrix; Let Γ_r, Γ_x, Γ_q, Γ_l denote the adaptive gating factors of each input channel. b_h represents the bias term; Let $\phi(\cdot)$ denote the nonlinear mapping function. The function of Equation (1) is to compress the course process data from different sources and scales into a unified semantic space, so as to provide a consistent input for subsequent continuous calculations.

After the generation of the basic representation, the model needs to describe the state progress of the curriculum reform in consecutive weeks, so the update mechanism with stage memory is introduced, and its calculation form is as follows.

$$C_t = \tanh(P_c H_t + P_m C_{t-1} + P_d \Delta_t + b_c) \tag{2}$$

Here, C_t represents the course state vector of the t period. C_{t-1} represents the state in the previous period; Δ_t represents the change in the current period with respect to the previous period; P_c, P_m, P_m represent mapping parameters; b_c denotes the bias term. The function of Equation (2) is to incorporate current observation, historical state and stage

change into the update process at the same time, so that the model can express the cumulative effect of curriculum reform instead of only retaining the single observation result.

In order to enhance the comparability of different weekly information, the model further uses attention aggregation to form a global stage representation, which is calculated as follows.

$$\alpha_t = \frac{\exp(v^\top \tanh(P_a C_t + b_a))}{\sum_{\tau=1}^{T} \exp(v^\top \tanh(P_a C_\tau + b_a))}, \quad G = \sum_{t=1}^{T} \alpha_t C_t \tag{3}$$

Here, α_t represents the attention weight of the t time period; v represents the learnable rating vector; P_a represents the mapping parameter; b_a represents the bias term; T is the total observation period; G denotes the aggregated global course representation. The function of Equation (3) is to complete the weighted integration according to the contribution of different periods to the reform effect, so that high-value teaching activities can obtain a more reasonable expression in the overall evaluation.

In the effectiveness judgment layer, the model generates four indicators of participation, completion, interaction quality and ability growth respectively according to the aggregation representation, and forms a comprehensive score, which is calculated as follows.

$$S_i = \lambda_1 D_i + \lambda_2 U_i + \lambda_3 I_i + \lambda_4 K_i \tag{4}$$

where S_i represents the comprehensive assessment score of the S_i teaching unit; D_i stands for participation; U_i represents task completion. I_i indicates the quality of interaction; K_i represents the capacity increase; λ_1 to λ_4 represent each dimension weight. The function of Equation (4) is to compress the performance of curriculum reform in different dimensions into unified scoring results and form comparable quantitative output.

Finally, the model uses the probabilistic judgment method to give the curriculum reform effect level, and its calculation form is as follows.

$$P(y_i = k \mid S_i) = \frac{\exp(\eta_k S_i + \beta_k)}{\sum_{j=1}^{3} \exp(\eta_j S_i + \beta_j)} \tag{5}$$

where $P(y_i = k \mid S_i)$ represents the probability that the i teaching unit belongs to the k effect level. Let η_k denote the mapping coefficient of the k class; Let β_k denote the bias term; j denotes the full rank category. The function of Equation (5) is to transform the continuous score into the hierarchical judgment result, so that the curriculum reform effect can be output in the form of high matching, stable promotion and adjustment, and directly serve the subsequent teaching revision.

The significance of this architecture is not just to do the result generation once. The reform of private undergraduate German course is usually accompanied by the reorganization of task chain, the adjustment of oral training frequency, the refinement of writing feedback and the reconfiguration of platform resources. After the model retains these process traces in the same computing space, teachers can see the structural differences of different teaching units at the same stage, and can also identify the direction of the overall effectiveness of a certain type of teaching activities. The evaluation output obtained in this way is more suitable as the computational basis for curriculum revision, resource redistribution and subsequent experimental analysis. It is also convenient for subsequent access to fine-grained indicators such as voice transcription quality, text complexity, task response time and online stay depth, so as to maintain the expansion ability and deployment stability of the model in real German

teaching scenarios.

3.2 Data representation and feature modeling method of German curriculum Reform process

The data representation and feature modeling of the German curriculum reform process are the intermediate links of the effect evaluation model from the original record to the judgment output. In the implementation process of private undergraduate German courses, classroom interaction records, homework texts, stage test scores, platform access trajectories and learning feedback information will be continuously generated. There are obvious differences in the sources, dimensions and update frequency of these data. Only after the unified representation, temporal alignment and feature extraction are completed, the effectiveness of curriculum reform can enter a stable calculation link. To this end, this paper takes teaching units as the basic granularity, organizes course process samples under the weekly advancement structure, and successively completes standardization processing, time window fusion, text semantic mapping, behavior relationship composition, multimodal gated integration and interactive feature enhancement.

In order to eliminate the shift of data from different sources on the numerical scale, this paper first standardises all kinds of original variables. The normalized expression is shown in Equation (6):

$$z_{ij}^{(m)} = \frac{x_{ij}^{(m)} - \mu_j^{(m)}}{\sigma_j^{(m)} + \varepsilon} \tag{6}$$

where $x_{ij}^{(m)}$ represents the original value of the i teaching unit in the m data on the j index, $z_{ij}^{(m)}$ represents the standardized result, $\mu_j^{(m)}$ represents the mean value of the data in the m index, $\sigma_j^{(m)}$ represents the corresponding standard deviation, and ε represents a small constant to prevent the denominator from being zero. The function of formula (6) is to map the data of different dimensions such as the number of classroom behaviors, assignment scores, platform stay time and feedback intensity to the same scale, and provide consistent input for subsequent fusion calculation.

After obtaining the standardized results, the alignment of the course process data also needs to be completed in the time dimension. Considering that the German curriculum reform has the characteristics of stage advancement, this paper uses the sliding window method to fuse multiple observations in the same teaching stage, and its expression is shown in Formula (7).

$$\widetilde{z_{it}^{(m)}} = \sum_{\tau=t-h}^{t} \omega_{t,\tau}^{(m)} z_{i\tau}^{(m)} \tag{7}$$

Here, $\widetilde{z_{it}^{(m)}}$ represents the window fusion result of the i teaching unit at the t time period, $z_{i\tau}^{(m)}$ represents the standardized observation value at the τ time in the window, h represents the length of the time window, $\omega_{t,\tau}^{(m)}$ represents the corresponding time weight, and the sum of each weight is 1. The function of equation (7) is to compress local fluctuations in adjacent weeks, so that classroom performance and after-school learning activities can form a continuous representation within a unified phase.

Homework texts, transcribed sentences in class and stage writing samples in German curriculum reform can directly reflect vocabulary organization, syntactic complexity and expression revision range. In order to preserve both deep semantics and explicit language features, this paper jointly maps the text data, whose expression is shown in Formula (8):

$$s_i = P_s\left[B_i^{ctx} \,\|\, T_i^{lex} \,\|\, G_i^{syn}\right] + b_s \tag{8}$$

Here, s_i represents the text comprehensive representation of the i teaching unit, B_i^{ctx} represents the semantic vector extracted by the context encoder, T_i^{lex} represents the lexical level feature vector, G_i^{syn} represents the syntactic level feature vector, symbol $\|$ represents the vector splicing operation, P_s represents the mapping matrix, and b_s represents the bias term. The function of equation (8) is to unify the semantic information, lexical density and syntactic structure in the German text into the same representation space, which provides a computable basis for the change of language competence after the curriculum reform.

The behavior records such as classroom interaction, task submission, resource review and online stay have obvious relationship structure, and it is difficult to express the linkage changes between activities by only using independent statistics. To this end, this paper constructs a behavior relationship matrix based on time proximity and task relevance, and its edge weight calculation form is shown in Formula (9):

$$a_{uv} = \exp\left(-\frac{\|p_u - p_v\|^2}{\rho_p^2} - \frac{|\tau_u - \tau_v|^2}{\rho_t^2}\right) \cdot \xi_{uv} \tag{9}$$

Here, a_{uv} represents the connection weight between behavior node u and node v, p_u and p_v represent the behavior attribute vector corresponding to two nodes, τ_u and τ_v represent the occurrence time of two behavior events, ρ_p and ρ_t represent the scaling parameters of attribute distance and time distance respectively, ξ_{uv} represents the task correlation coefficient. The function of equation (9) is to transform classroom behaviors and platform behaviors into graph relationships with structural connections, and retain the transfer characteristics between teaching activities for subsequent propagation calculations.

After constructing the behavior relationship matrix, this paper further extracts the process behavior features by graph propagation, whose expression is shown in Formula (10):

$$R^{(l+1)} = \mathrm{ReLU}\left(\hat{A}R^{(l)}P_l + b_l\right) \tag{10}$$

Here, $R^{(l)}$ represents the behavior feature representation of the l layer, $R^{(l+1)}$ represents the updated behavior representation, $\hat{A}$ represents the normalized behavior adjacency matrix, P_l represents the mapping parameter of the l layer, b_l represents the bias term, and $\mathrm{ReLU}(\cdot)$ represents the nonlinear activation function. The function of equation (10) is to make a single behavior node not only retain its own information, but also absorb the influence of neighboring activities, so as to more accurately characterize learning engagement, interaction density, and task persistence.

When the text features, behavior features, achievement features, platform features and feedback features are all encoded, the model needs to control the proportion of different modalities into the fusion layer. In order to avoid excessive amplification of high-frequency but weakly correlated information, this paper introduces a gated fusion mechanism, whose expression is shown in Formula (11):

$$g_i = \sigma\left(P_g[c_i \| b_i \| q_i \| l_i \| f_i] + b_g\right) \tag{11}$$

Here, g_i represents the gating vector of the i teaching unit, c_i represents the text feature, b_i represents the behavior feature, q_i represents the test result feature, l_i represents the platform log feature, f_i represents the feedback perception feature, P_g represents the gating mapping matrix, b_g represents the bias term, and $\sigma(\cdot)$ represents the Sigmoid function. The function of formula (11) is to automatically assign entry weights according to the contribution degree of different modes to the effectiveness of curriculum reform, so that the key features can obtain higher retention.

After the generation of the gating coefficients, this paper performs weighted fusion of multi-modal features to obtain a unified representation of the course process, whose expression is shown in Formula (12):

$$u_i = g_i \odot c_i + (1 - g_i) \odot \left(P_b b_i + P_q q_i + P_l l_i + P_f f_i\right) \tag{12}$$

Here, u_i represents the fused feature vector of the i teaching unit, $\odot$ represents element-wise multiplication, and P_b, P_q, P_l, and $P_f$represent the mapping parameters of behavior, grade, platform, and feedback features, respectively. The function of Formula (12) is to compress the semantic expression ability, behavior participation, result performance and feedback perception into a unified feature space, so that the curriculum reform process can form an overall representation.

In order to further strengthen the comparison ability between different teaching units, this paper continues to calculate the structural interaction of the fusion features, whose expression is shown in Formula (13):

$$e_i = \eta_1 u_i + \eta_2 \sum_{j=1}^{n} \frac{\exp(m_{ij})}{\sum_{k=1}^{n} \exp(m_{ik})} u_j, \quad m_{ij} = \tanh\left(\frac{u_i^{\top} P_m u_j}{\sqrt{d}}\right) \tag{13}$$

Here, e_i represents the final enhanced feature of the i teaching unit, η_1 and η_2 represent the balance coefficient between local features and interactive features, u_j represents the fused features of other teaching units, m_{ij} represents the correlation score between the i teaching unit and the j teaching unit, P_m represents the interaction mapping matrix, d represents the feature dimension, and n represents the total number of teaching units. The function of formula (13) is to enhance the difference recognition ability through the cross-reference between features, so that the changes of the same curriculum reform in different classes, different weeks and different task units can be more clearly expressed.

Through the above seven steps, the German curriculum reform process data is organized into a unified representation with temporal, semantic and structural correlations. Such modeling results can not only support the subsequent overall scoring, but also serve the difference analysis under the influence of key features, and provide a more detailed calculation basis for course content revision, task rhythm adjustment and learning support configuration.

3.3 Output mechanism and judgment process of curriculum reform effect evaluation

The output mechanism of curriculum reform effect evaluation is located at the back end of the whole model, and it is responsible for four functions: feature reception, comprehensive judgment, level output and feedback writeback. In the process of reform promotion, the private undergraduate German course will form multi-dimensional changes such as

vocabulary training, grammar tasks, oral interaction, text writing and platform learning. These changes have been transformed into computable enhanced features after the unified representation of the previous two sections. The function of the output layer is not simply summing, but compressing the multi-dimensional features into continuous scores, discrete grades and interpretable labels under the premise of maintaining the difference of teaching units, so that the effect of curriculum reform can directly serve the subsequent curriculum revision.

As shown in Fig. 2, the output mechanism consists of a comprehensive scoring module, a grade judgment module, a difference marking module and a feedback writeback module. The comprehensive scoring module receives the enhanced feature vector, and generates the main score value according to four core indicators: participation, completion, interaction quality and ability growth. The level judgment module gives the effectiveness status of the current unit according to the interval threshold and probability confidence. The difference marking module locates the teaching links with large score fluctuations. The feedback writeback module synchronizes the results to the teaching content adjustment, task difficulty revision and resource support configuration ports. In this way, the evaluation results are no longer static records, but can be used as computational input for the next round of course organization.

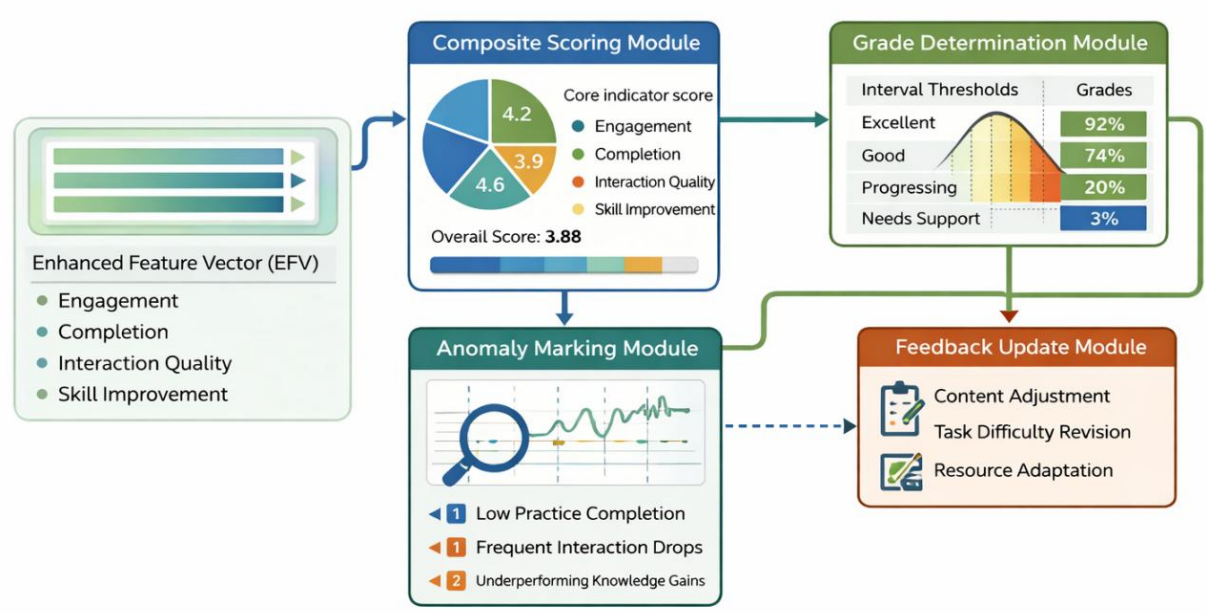

Figure 2: Output mechanism structure of curriculum reform effect evaluation

After obtaining the overall score, the model also needs to decide whether the current score is stable enough. To this end, this paper introduces a score correction term with confidence constraints, and the correction expression is shown in Formula (14):

$$\widehat{S_i} = \omega_i S_i + (1 - \omega_i) S_i \exp(-\sigma_i) \tag{14}$$

Here, $\widehat{S_i}$ represents the corrected comprehensive score, S_i represents the original score, σ_i represents the degree of sample fluctuation within the current cell, ω_i represents the local stability coefficient, and $\exp(\cdot)$ represents the exponential function. The function of equation (14) is to suppress the interference of high fluctuation units on the final judgment, so that the scoring results not only reflect the course effectiveness, but also retain statistical stability.

As shown in Fig. 3, the level determination process is not a single threshold comparison, but first forms a probability distribution according to the correction score, and then combines with the threshold interval to complete the state output. The high matching state indicates that the current reform structure is in good agreement with the curriculum objectives, the stable advancement state indicates that the teaching organization and learning performance are in the sustainable interval, and the state to be adjusted is used to mark the units that need to be focused on observation. This process enables the implementation status of different classes, different weeks and different task types to obtain a consistent expression.

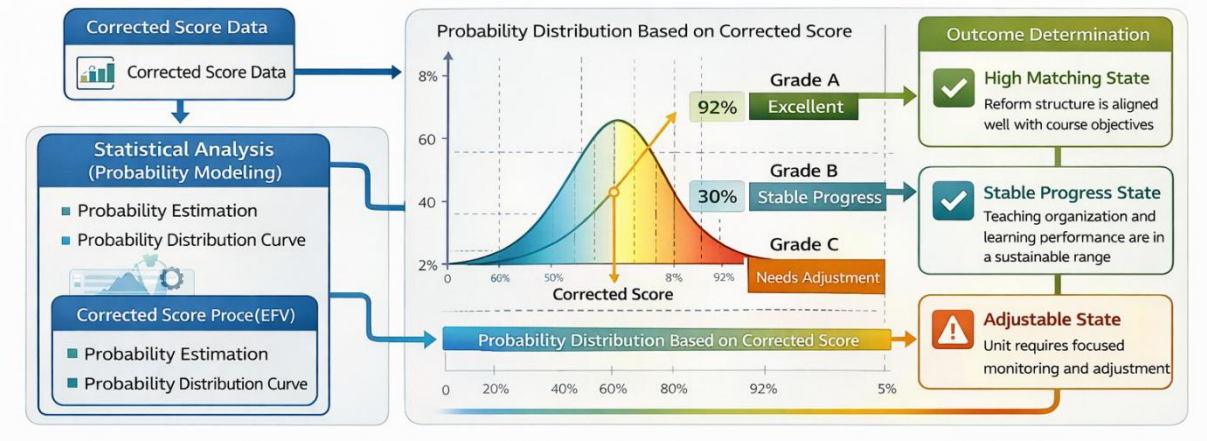

Figure 3: Course reform effect level determination process

In order to output the effect level of curriculum reform, this paper uses a probabilistic judgment function, whose expression is shown in Formula (15):

$$P(y_i = k \mid \hat{S}_i) = \frac{\exp(\eta_k \hat{S}_i + \beta_k)}{\sum_{j=1}^{3} \exp(\eta_j \hat{S}_i + \beta_j)} \tag{15}$$

where $P(y_i = k \mid \hat{S}_i)$ represents the probability that the i teaching unit belongs to the k effect level, η_k represents the k mapping coefficient, β_k represents the bias term, k ranges from 1 to 3, corresponding to the three types of state of high matching, stable advancement and to be adjusted respectively. The function of equation (15) is to map the continuous scores into rank probabilities, so that the evaluation result can output a clear class while preserving the proximity between classes.

After obtaining the grade results, the model continued to generate the difference marker vector, which was used to identify the key fluctuation items that affected the effectiveness of curriculum reform. The difference labeling function is given in Equation (16):

$$M_i = \|\Theta(e_i - \bar{e})\|_2^2 \tag{16}$$

Here, M_i represents the difference labeling value of the i teaching unit, e_i represents the enhanced feature vector of the current unit, $\bar{e}$ represents the average feature center of similar teaching units, Θ represents the learnable projection matrix, and $\|\cdot\|_2$ represents the two-norm. The function of equation (16) is to measure the degree of deviation between the current unit and the homogeneous mean structure, and to provide a direct basis for evaluating the difference analysis under the influence of key features in Chapter IV.

Finally, the output is written back to the course adjustment module in the form of a structured record. If a unit deviated from the same class center continuously in vocabulary

training, oral interaction or task completion, the system would increase its marking intensity and synchronize the results to the teaching content reorganization port. If multiple units show a consistent rise at the same stage, the system retains the current task rhythm and support configuration. The writeback link formed in this way makes the evaluation model not only give the results, but also participate in the dynamic adjustment of the course reform process. For private undergraduate German courses, this output mechanism can transform complex teaching changes into observable, comparable and traceable status information, and provide a unified interface for subsequent result analysis and difference analysis. The output record also retains six types of fields: unit number, weekly index, grade probability, main score value, correction score and difference mark value, which is convenient for subsequent horizontal comparison, vertical tracking and batch retrieval. The database end uses the primary key association method to connect teaching classes, task types and time Windows. The front-end interface displays the result change curve according to the course weekly, so that teachers can directly read the stable interval and fluctuation interval in the reform promotion. This step also enables the model output to directly enter the subsequent statistical test and visualization module, and maintains a unified interface format to facilitate the subsequent module call chain.

4 Analysis of results

4.1 Evaluation and analysis of the reform effect of private undergraduate German Course

After completing the construction of the effect evaluation model of German course reform in private undergraduate colleges in the context of AI empowerment, in order to verify the suitability and judgment stability of the model in real teaching scenarios, this paper selects the sample of German course reform in Xi 'an Institute of Translation to carry out the effect evaluation. The evaluation objects covered five types of course activities, such as vocabulary training, oral interaction, text writing, task submission and platform learning, and a total of 4120 valid samples were formed. In order to ensure the interpretability of the results, this paper also verifies the model by comparing the expert ratings with the model output, and evaluates the structure clarity, feature representation rationality, consistency of scoring results and feedback usability of the model with a five-point scale. SPSS statistical results show that the overall mean of the model is 4.86, and the standard deviation is 0.24, indicating that the evaluation model has good operability and stability in the curriculum reform scenario. Table 2 presents the results of the model evaluation questionnaire.

Table 2: Questionnaire results of the evaluation model for the reform effect of German courses for private undergraduate students

Questionnaire Item	Mean	Standard Deviation
The model structure is clear and the evaluation process is complete	4.83	0.29
The multi-source feature representation matches the curriculum reform process well	4.89	0.22
The evaluation output can reflect differences across teaching units	4.92	0.18
The grade determination results show strong interpretability	4.81	0.27
Feedback write-back can support curriculum adjustment	4.84	0.25
Overall	4.86	0.24

It can be seen from Table 2 that the experts have the highest recognition degree for "the evaluation output can reflect the differences of teaching units", indicating that the model has strong discrimination ability in identifying the performance differences of different teaching units in the reform promotion. The mean value of "multi-source feature representation matches the course reform process well" is also at a high level, indicating that the data such as text, behavior, score and platform log have been able to express the structural changes in the German course reform more stably after unified coding. In contrast, the score of "grade determination results have strong interpretability" is slightly lower, but the standard deviation remains within a small interval, indicating that different experts still have a relatively consistent understanding of the determination mechanism.

As shown in Fig. 4, the four types of core evaluation dimensions form a clearer cluster distribution in the latent feature space. The Participation samples are mainly concentrated in 90.6-91.2, and the center value is about 90.9. Ability Growth is mainly distributed between 90.8 and 91.5 points, and the center value is about 91.1 points, which is close to each other. Interaction Quality is mainly concentrated in 92.3 to 93.0, and the center value is about 92.6. The Completion sample is concentrated in 93.2 to 93.7 points, and the center value is about 93.4 points, which is the highest position of the four dimensions. On the whole, the high segment is mainly composed of Completion and Interaction Quality, Participation and Ability Growth are distributed in a relatively low interval, and there is less overlap in each cluster, which indicates that the fusion feature has good discrimination ability.

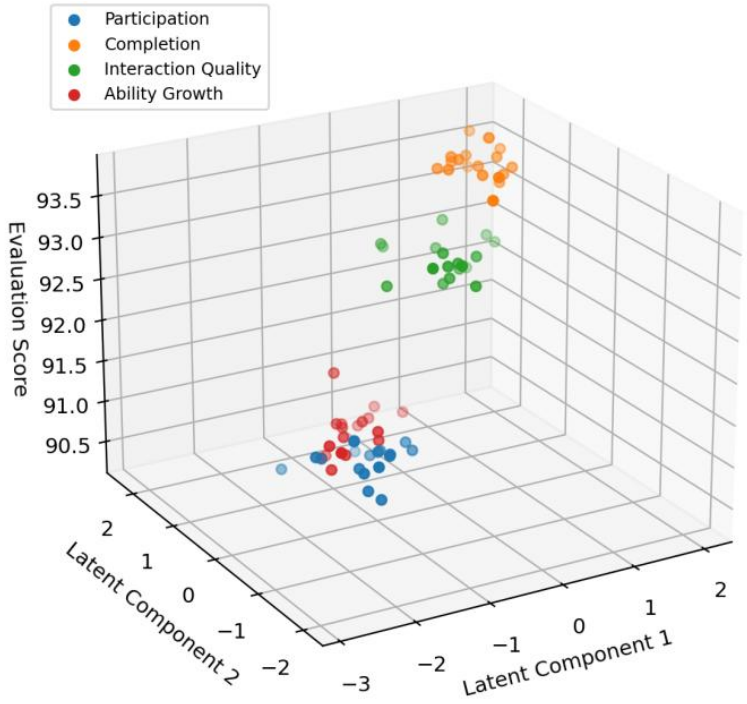

Figure 4: Distribution map of four-dimensional evaluation results of German curriculum reform for private undergraduate students

As shown in Fig. 5, the output scores of the model are mainly concentrated in the range 0.78 to 0.94, and the proportion of samples in the high matching state is 46.3%, the proportion of samples in the stable advancement state is 41.8%, and the proportion of samples in the state to be adjusted is 11.9%. This distribution shows that most teaching units have entered a relatively stable reform operation state, while there are still a small number of units that need further observation, which is consistent with the gradual promotion characteristics of private

undergraduate curriculum reform in different classes and different weeks.

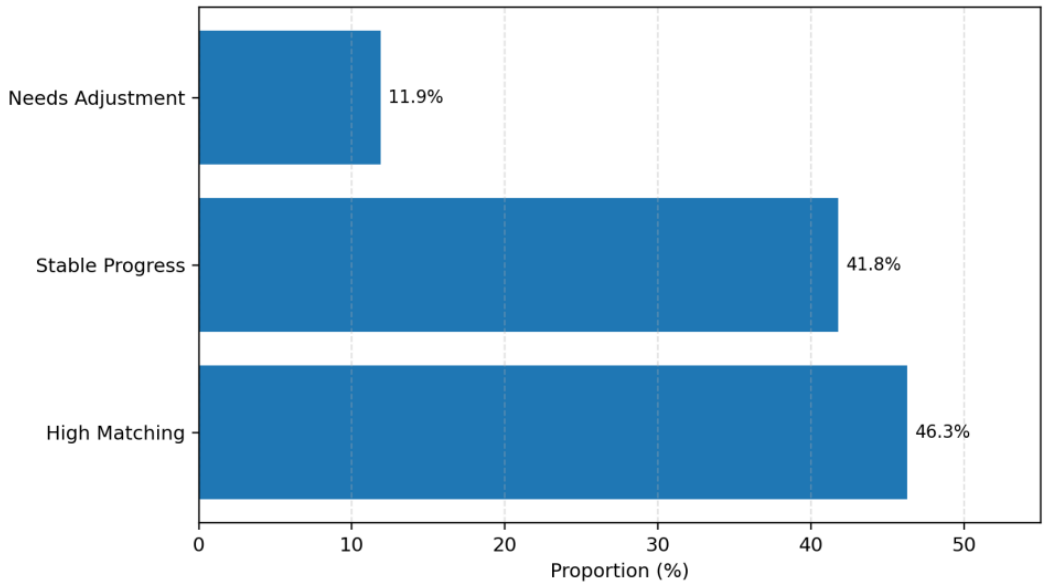

Figure 5: Grade distribution diagram of curriculum reform effect

In order to further verify the performance of the model on quantitative indicators, this paper makes statistics from four aspects: accuracy, recall, F1 value and average response delay, and the results are shown in Table 3.

Table 3: Performance comparison of the evaluation model of the reform effect of German course for private undergraduate students

Method	Accuracy/%	Recall/%	F1/%	Response Delay/s
Proposed Model	92.7	91.4	90.9	1.6
Traditional Statistical Evaluation Method	86.8	85.2	84.6	2.3

Table 3 shows that the accuracy rate, recall rate and F1 value of the proposed model are higher than those of the traditional statistical evaluation methods, which are increased by 5.9, 6.2 and 6.3 percentage points respectively, indicating that the model has better judgment ability in the identification of the effect of curriculum reform. The average response delay is reduced from 2.3 s to 1.6 s, which indicates that the model can complete the output faster after receiving multi-source features, and can meet the computational requirements of stage evaluation and immediate feedback of German course.

As shown in Fig. 6, the accuracy of the model in eight rounds of verification always maintains a high level, and the overall accuracy fluctuates slightly around 92%, the lowest value is about 91.8%, and the highest value is close to 92.9%. At the same time, the range of changes between rounds is mainly controlled between -0.2% and 0.5%, and there is no obvious continuous decline or abnormal shock. The results show that the model maintains good stability under different batches of samples and different validation rounds, and the evaluation output will not be greatly shifted due to local sample disturbances. For the scenario of private undergraduate German course reform, this high-level stable verification result has strong practical significance, because curriculum reform often covers multiple teaching units and multiple implementation stages. If the model is too sensitive to stage fluctuations, it is difficult to support continuous evaluation. The present results show that the proposed method

not only maintains the recognition accuracy, but also has good consistency of repeated verification.

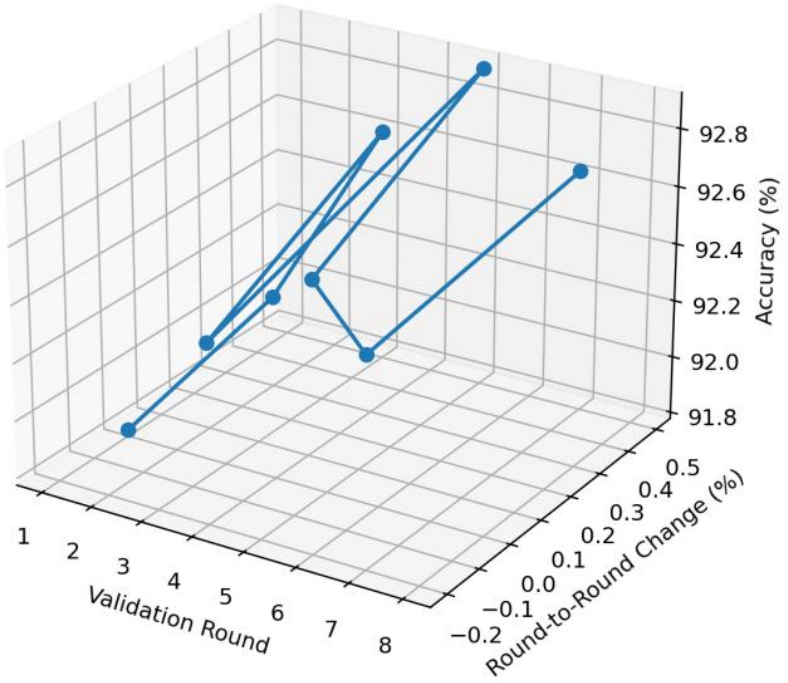

Figure 6: Model evaluation accuracy change plot under eight rounds of validation

As shown in Fig. 7, after the implementation of the curriculum reform, the comprehensive scores of the three teaching links of vocabulary training, oral interaction and writing revision were significantly improved. Vocabulary training increased from 78.4 to 86.7, writing revision increased from 76.3 to 84.9, and oral interaction increased from 74.9 to 88.5. This change showed that the curriculum reform in the context of AI empowerment did not only affect a single teaching task, but also extended to three levels of vocabulary mastery, expression organization and interactive reaction. Among them, the score of oral interaction improved most obviously, indicating that intelligent feedback, task-driven and classroom interaction reconstruction had a direct driving effect on German expression training. Vocabulary training and writing revision were also improved simultaneously, indicating that the organization of curriculum resources, the adjustment of exercise methods and the feedback and writeback mechanism had formed a relatively stable relationship between knowledge acquisition and language output. On the whole, the teaching activities after the reform show a consistent upward trend in multiple dimensions, and the output results of the model are highly consistent with the implementation status of the curriculum.

Figure 7: Comparison of comprehensive scores before and after the implementation of the German curriculum reform

Comprehensive table and graphical results show that the evaluation model constructed in this paper has been able to more completely present the overall state, dimension differences and stage changes in the reform of private undergraduate German curriculum. The advantage of the model is not only reflected in the high accuracy, but also reflected in the ability to organize the previously scattered classroom behavior, text revision, task completion, and platform access records into a unified judgment result. The output formed in this way is not only suitable for single assessment, but also can support subsequent key feature impact analysis and curriculum adjustment decisions, providing continuous, interpretable and comparable quantitative basis for German curriculum reform.

4.2 Evaluation difference analysis under the influence of key features

In order to further identify the degree of influence of different key features on the evaluation results of curriculum reform effect, this paper carries out difference analysis experiments based on the above model. The experiment still uses 4120 valid samples, and the training set, validation set and test set are divided according to 7∶2∶1, and the other parameter Settings are consistent with those in Section 4.1. The analysis objects include four core variables: text semantic features, classroom behavior map features, stage performance features and platform interaction features. The model output was synchronously compared with the expert annotation results, and the comprehensive analysis was carried out by combining the feature contribution weight, the score difference of the course link and the ablation experimental results.

As shown in Fig. 8, there are obvious differences in the clustering positions of the four types of key features in the 3D projection space. The sample cluster corresponding to the text semantic feature is located in the high-value area, and its contribution weight is 0.31. The characteristic of classroom behavior map corresponded to the sample cluster, and the contribution weight was 0.28. Stage performance features and platform interaction features are mainly distributed in the middle and low value area, and the contribution weights are 0.23 and 0.18, respectively. The results show that the effect evaluation of German curriculum

reform does not depend on a single performance indicator, and the expression quality of homework text, classroom interaction trajectory and learning behavior structure jointly determine the ability of the model to identify the effect of curriculum reform. Among them, the sample clusters formed by text semantic features were more concentrated, indicating that vocabulary use, syntactic organization and text revision information had stronger explanatory power for the state of curriculum reform. The classroom behavior map feature follows, indicating that the relationship structure between classroom speech, oral interaction, and task response has become an important basis for influencing the assessment results.

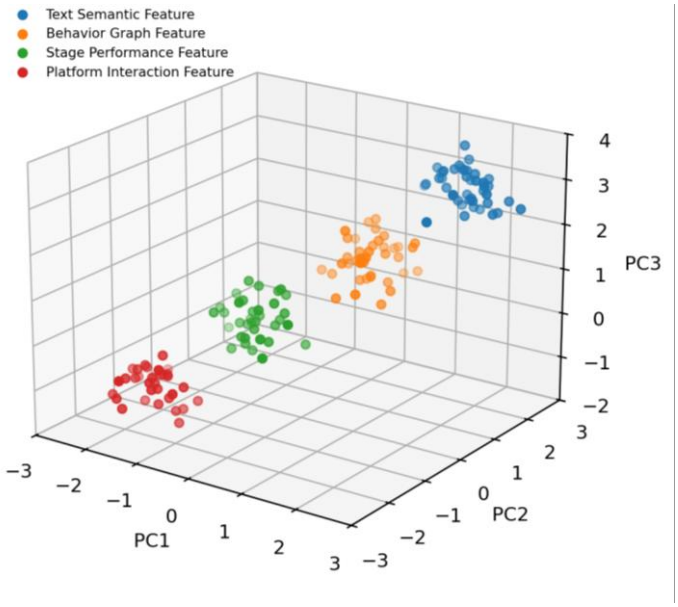

Figure 8: Key feature contribution weight distribution diagram

Based on the overall contribution analysis, this paper further investigates the specific impact of different modules on the model performance through ablation experiments, and the results are shown in Table 4. The Accuracy, Recall and F1 values of the complete model reach 92.7%, 91.4% and 90.9%, respectively, and the average response delay is 1.6 s. After removing the text semantic mapping, the Accuracy decreased to 89.8%, and the F1 value decreased to 87.8%. After removing the behavior graph propagation, the Recall decreased to 88.1%. After removing the gated fusion, the model response delay rises to 2.0 s. The results show that text semantic mapping has the most obvious influence on the recognition accuracy of German expression quality, behavior graph propagation mainly plays a role in the continuous capture of classroom interaction and task correlation, and gated fusion maintains the stability of judgment after multi-source features are entered into a unified space.

Table 4: Results of ablation experiments

Model Configuration	Accuracy/%	Recall/%	F1/%	Response Delay/s
Full Model	92.7	91.4	90.9	1.6
Without Text Semantic Mapping	89.8	88.9	87.8	1.7
Without Behavioral Graph Propagation	90.4	88.1	88.6	1.8
Without Gated Fusion	89.2	88.7	88.0	2.0

In order to further observe the influence ways of key features on different course links, this paper conducted a three-dimensional clustering comparison of three teaching links: vocabulary training, oral interaction and writing revision. As shown in Fig. 9, under the

complete model, the comprehensive scores of vocabulary training, oral interaction and writing revision are 92.4, 93.1 and 91.8, respectively. The three sample clusters are distributed in the high value area as a whole, and the oral interaction cluster is closest to the upper bound. After removing the text semantic mapping, the three scores decreased to 87.1, 90.4 and 86.5, respectively. The sample clusters corresponding to vocabulary training and writing revision moved to the low value area, and the decline was the most obvious. After removing behavior graph propagation, the three scores were 90.2, 86.8 and 89.6, respectively. The dispersion degree of oral interaction sample cluster increased significantly, indicating that the characteristics of classroom behavior relationship had a stronger supporting effect on oral training. In contrast, vocabulary training and writing revision are more dependent on text expression information, and oral interaction is more dependent on behavior sequence and interaction structure.

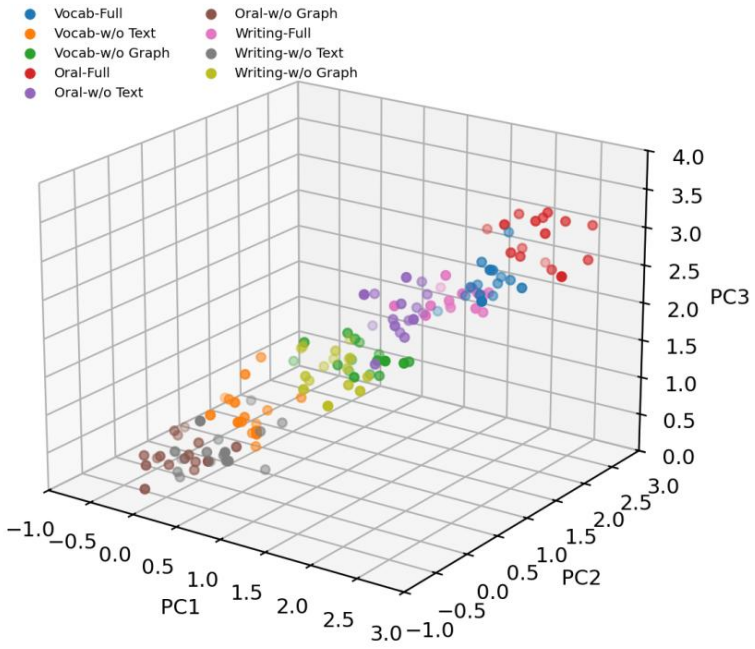

Figure 9: Comparison chart of differences in comprehensive scores under different course links

Combining the results in Fig. 8, Table 4, and Fig. 9, we can see that the evaluation advantage of the proposed model comes from the synergy of text semantic features, behavioral relationship features, and multimodal fusion mechanisms. The key features not only determine the overall score, but also affect the boundary clarity between samples of different grades. The analysis results can provide a more direct quantitative basis for subsequent course content adjustment, task organization revision and support mode configuration.

5 Discussion

Focusing on the evaluation of the reform effect of private undergraduate German course in the context of AI empowerment, this paper constructs a computing link consisting of multi-source data representation, key feature fusion, level judgment and feedback writeback, and verifies the stability and adaptability of the model through experiments. Different from the existing researches that focus on automatic writing evaluation, chatbot support or single feedback

analysis, this paper emphasizes the joint modeling of classroom behavior, text semantics, stage performance and platform log, so that the effectiveness of curriculum reform can be continuously presented in a unified feature space. The results show that the model maintains a good balance between accuracy, recall rate, F1 value and response time, which can not only complete the overall judgment, but also identify the structural differences of different teaching links. The supporting effect of text semantic mapping and behavior map propagation on the evaluation results is more obvious, indicating that the language output quality and classroom interaction structure in the German curriculum reform jointly shape the effectiveness boundary. On the whole, the model is suitable for private undergraduate German course scenarios with mixed teaching, task-driven and formative evaluation. It also provides a computational basis for subsequent access to voice transcription quality, syntactic complexity and learning path modeling, and enhances the pertinence and implementation continuity of curriculum adjustment.

6 Conclusions

Focusing on the evaluation of the reform effect of private undergraduate German course in the context of AI empowerment, this paper constructs an evaluation model consisting of multi-source data representation, feature fusion calculation, level determination output and feedback writeback, and completes verification on real course samples. The results show that the proposed model can incorporate classroom behaviors, assignment texts, stage grades and platform logs into the unified computing space to continuously identify structural changes after curriculum reform. The model maintains a good balance in accuracy, recall, F1 value and response time, indicating that the method not only has strong judgment ability, but also has the actual deployment basis. The text semantic mapping, behavior graph propagation, and gated fusion mechanism jointly support the stable output of the evaluation results, which enables vocabulary training, spoken interaction, and writing revision to enter the comparative analysis in a quantitative way.

At the same time, this paper still has some limitations. The current sample mainly comes from the German course of the same college, and the course types, student levels and teaching organization methods are relatively concentrated. The adaptation range of the model in cross-college scenarios still needs to be further tested. Speech data and fine-grained syntactic information have not been fully incorporated into the unified representation, and there is still room for compression of instant changes and complex expression features in spoken language activities. While keeping the existing backbone structure stable, the following research can introduce cross-proofreading, voice transcription quality indicators, syntactic complexity indicators and dynamic sequence modeling methods to further enhance the model's ability to sense the details of German curriculum implementation, and improve the transparency and transfer ability of result interpretation. Combined with teachers' decision records and students' revision trajectories, a continuous assessment link closer to the teaching scene was formed.

About the Author

Honglian Bian was born in 1984 in Xianyang, Shaanxi, People’s Republic of China. She obtained a master's degree from Xi’an International Studies University in China. Currently, she teaches at the College of International Language and Culture, Xi’an Fanyi University. Her main research directions include German Language and Literature, Children’s Literature, and International Chinese Language Education.

References

[1] Schmidt T, Strasser T. Artificial intelligence in foreign language learning and teaching: a CALL for intelligent practice[J]. Anglistik: International Journal of English Studies, 2022, 33(1): 165-184.

[2] Huang W, Hew K F, Fryer L K. Chatbots for language learning—Are they really useful? A systematic review of chatbot-supported language learning[J]. Journal of computer assisted learning, 2022, 38(1): 237-257.

[3] Klímová B, Ibna Seraj P M. The use of chatbots in university EFL settings: Research trends and pedagogical implications[J]. Frontiers in psychology, 2023, 14: 1131506.

[4] Annamalai N, Eltahir M E, Zyoud S H, et al. Exploring English language learning via Chabot: A case study from a self determination theory perspective[J]. Computers and Education: Artificial Intelligence, 2023, 5: 100148.

[5] Annamalai N, Ab Rashid R, Hashmi U M, et al. Using chatbots for English language learning in higher education[J]. Computers and Education: Artificial Intelligence, 2023, 5: 100153.

[6] Qiao H, Zhao A. Artificial intelligence-based language learning: illuminating the impact on speaking skills and self-regulation in Chinese EFL context[J]. Frontiers in psychology, 2023, 14: 1255594.

[7] Kohnke L, Moorhouse B L, Zou D. ChatGPT for language teaching and learning[J]. Relc Journal, 2023, 54(2): 537-550.

[8] Gao R, Merzdorf H E, Anwar S, et al. Automatic assessment of text-based responses in post-secondary education: A systematic review[J]. Computers and Education: Artificial Intelligence, 2024, 6: 100206.

[9] Wang F, Cheung A C K, Chai C S. Language learning development in human-AI interaction: A thematic review of the research landscape[J]. System, 2024, 125: 103424.

[10] Xu T, Wang H. The effectiveness of artificial intelligence on English language learning achievement[J]. System, 2024, 125: 103428.

[11] Karatay Y, Karatay L. Automated writing evaluation use in second language classrooms: A research synthesis[J]. System, 2024, 123: 103332.

[12] Sari E, Han T. The impact of automated writing evaluation on English as a foreign language learners' writing self-efficacy, self-regulation, anxiety, and performance[J]. Journal of Computer Assisted Learning, 2024, 40(5): 2065-2080.

[13] Cheng X, Zhang L J. Examining second language (L2) learners' engagement with AWE-teacher integrated feedback in a technology-empowered context[J]. The Asia-Pacific Education Researcher, 2024, 33(4): 1023-1035.

[14] Dizon G. ChatGPT as a tool for self-directed foreign language learning[J]. Innovation in Language Learning and Teaching, 2024: 1-17.

[15] Belda-Medina J, Kokošková V. ChatGPT for language learning: Assessing teacher candidates' skills and perceptions using the Technology Acceptance Model (TAM)[J]. Innovation in Language Learning and Teaching, 2024: 1-16.

[16] Karataş F, Abedi F Y, Ozek Gunyel F, et al. Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners[J]. Education and Information Technologies, 2024, 29(15): 19343-19366.

[17] Li Z, Wang C, Bonk C J. Exploring the Utility of ChatGPT for Self-Directed Online Language Learning[J]. Online Learning, 2024, 28(3): 157-180.

[18] Klimova B, Pikhart M, Al-Obaydi L H. Exploring the potential of ChatGPT for foreign language education at the university level[J]. Frontiers in Psychology, 2024, 15: 1269319.

[19] Koltovskaia S, Rahmati P, Saeli H. Graduate students' use of ChatGPT for academic text revision: Behavioral, cognitive, and affective engagement[J]. Journal of Second Language Writing, 2024, 65: 101130.

[20] Yan D, Zhang S. L2 writer engagement with automated written corrective feedback provided by ChatGPT: A mixed-method multiple case study[J]. Humanities and Social Sciences Communications, 2024, 11(1): 1-14.