AEM JOURNAL .ORG

Fine-Grained Classification of Localized Educational Resources Based on Deep Knowledge Tracing Model

Chunyan Zhang
College of Accounting (College of Wealth Management), Ningbo University of Finance and Economics, Ningbo 315075, Zhejiang, China
Corresponding author: Chunyan Zhang (e-mail: firelight2026@126.com)

ABSTRACT Accurate representation and classification of sequential information are critical for intelligent engineering systems involving complex signal processing and adaptive decision-making. To address the data sparsity problem in fine-grained classification of localized educational resources, this study proposes a contrast-enhanced attentional knowledge tracing framework that combines Attentive Knowledge Tracing (AKT), sliding-window sequence augmentation, and InfoNCE-based contrastive learning. The proposed method extracts cognitive state transition vectors as latent representations of resource functions and performs hierarchical clustering to identify implicit instructional categories without relying on large-scale annotated corpora or textual features. Experimental results demonstrate that the proposed framework achieves a silhouette coefficient of 0.72 under full-data conditions and maintains robust clustering performance under extremely sparse data scenarios, significantly outperforming conventional knowledge tracing models. The learned cognitive representations exhibit weak correlation with coarse-grained subject labels while effectively capturing latent functional characteristics of educational resources. By providing an efficient mechanism for high-dimensional sequential representation learning and feature discrimination, the proposed framework offers valuable insights for educational data mining, adaptive instructional design, and data-driven resource classification strategies that are relevant to intelligent tutoring systems and personalized learning applications.

INDEX TERMS attentive knowledge tracing, contrastive learning, localized educational resources, fine-grained classification, cognitive state transition

I. INTRODUCTION

THE current advancement of digital education has spawned a massive amount of learning resources with regional cultural attributes. These localized educational resources carry the dialectal characteristics, folk traditions, and local knowledge systems of specific regions. However, the effective organization and retrieval of these resources has long been constrained by the bottleneck of classification technology. General subject tags can only indicate their superficial subject affiliation and cannot reveal the cognitive functions actually triggered in the teaching process (Song, et al., 2022; Ma, et al., 2024). The inherent sparsity of interactive data further exacerbates this dilemma. Due to the limited scope of use of individual school-based resources, it is difficult to accumulate enough student answer records to support data-driven fine-grained analysis (Feng, et al., 2022; Khayry, et al., 2024). Solving this problem has urgent engineering value for the intelligent recommendation of regional educational resources, adaptive learning path planning, and digital support for the inheritance of local culture. At the same time, at the theoretical level, it also constitutes a key test for the transfer of cognitive diagnostic models from large-scale standardized assessments to small-sample contextualized learning analysis.

The core challenge of fine-grained classification of localized educational resources stems from a dual scarcity. The first scarcity lies in the volume of interactive data: local resources are often used in single-school or singleclass settings, and the number and length of student response sequences are far lower than the typical training requirements of knowledge-tracking models. Under these conditions, deep neural networks face serious risks of excessively large gradient estimation variance and underfitting decision boundaries (Dai, et al., 2024; Rex, et al., 2022). The second scarcity lies in the granularity of annotation information: the teaching function labels of resources rely on the experience and judgment of frontline teachers, making large-scale fine-grained annotation impractical at the operational level, only yielding coarse, weak supervisory signals (Zhang, et al., 2025). These two scarcity factors

combine to render traditional classification methods relying on content text features ineffective in localized contexts. The error rate of dialect speech transcription leads to unreliable text feature extraction, while general classification models have almost no ability to perceive implicit cultural cognitive associations (Li, et al., 2024; Wassink, et al., 2022). How to construct a fine-grained classification mechanism that is sensitive to resource cognition under extremely small sample conditions has become an unresolved technical problem in this field. To address the bottlenecks in knowledge tracing applications in small-sample scenarios, researchers have explored various approaches. Attention-based knowledge tracing models replace recurrent neural networks with self-attention structures, capturing cognitive dependencies between distant questions within a sequence through a global receptive field, achieving improved data efficiency compared to traditional deep knowledge tracing (Pu, et al., 2022; Xu & Hu, 2024). Advances in contrastive learning paradigms in representation learning have provided another path for small-sample scenarios. Supervised contrastive loss enhances the inter-class separability of representation vectors without increasing training data by bringing similar samples closer together and distancing dissimilar samples further apart in the embedding space (Sabiri, et al., 2024; Hu, et al., 2024). However, each of these methods has inherent limitations: attention-based knowledge tracing focuses on prediction accuracy rather than the clustering structure of the representation space, and its output hidden state vectors lack explicit separation constraints for functional categories (Lu, et al., 2024; Liu, et al., 2026). Supervised contrastive learning, in constructing positive and negative samples, typically requires a high degree of certainty regarding the labels it relies on. However, the weak supervisory signals available in the geochemical resource domain create a tension with this premise (Huang, et al., 2022; Deng, et al., 2024). More critically, existing work rarely involves joint optimization of these two mechanisms for cognitive diagnostic representation. As a result, under conditions of extremely scarce interaction sequences and coarse labeled signals, directly classifying resources into fine-grained cognitive function categories remains a technical gap that few methods have addressed.

This paper addresses the limited scope of existing methods for joint optimization of cognitive diagnostic representations by proposing a contrastive-enhanced attention-based knowledge tracking framework. The core logic of this framework lies in shifting the fine-grained resource classification problem from the content feature space to the cognitive state transition space. This transition is grounded in the premise that while AKT traditionally predicts student mastery, the resulting cognitive state change vector (Δh) acts as a high-dimensional "cognitive fingerprint" of the resource itself. Because functionally similar resources trigger consistent patterns of knowledge state transitions across different learners, these vectors inherently encapsulate the latent semantic and pedagogical properties of the resources, providing a robust basis for fine-grained classification beyond simple performance prediction. The resource cognitive fingerprint Δh is computationally defined as the differential vector in the student's latent state before and after using the resource, $\Delta h = h_{t+\Delta} - h_t$. This vector represents the transition magnitude across the student's multidimensional knowledge mastery space, serving as a quantitative proxy for the resource's instructional impact. To achieve this goal, the framework is specifically designed at three levels: at the data level, a sliding window augmentation strategy is used to address the extreme sparsity of local interaction sequences; at the representation level, the coarse functional labels provided by the teacher are transformed into structured constraints for contrastive learning, guiding similar resources to form compact clusters in the latent space; at the output level, hierarchical clustering replaces the parameterized linear classifier, avoiding the interference of overfitting of the classifier head on the clustering structure under small sample conditions. Under small-scale experimental conditions, the proposed framework effectively classifies three functional orientation resources: native language prosody activation, local empathy arousal, and contextualized knowledge anchoring. The cluster density is better than the baseline model without contrast constraints. The ablation results confirm the independent contribution of each design module. The systematic deviation between the classification boundary and the coarse-grained subject labels indicates that the cognitive diagnosis-driven path captures the implicit teaching attributes that are missed in the content features in the functional dimension.

II. CONTRAST-ENHANCED COGNITIVE TRANSFER METHODS

A. PROBLEM FORMALIZATION AND INTRA-SEQUENCE AUGMENTATION

The core challenge in fine-grained classification of localized educational resources lies in the inherent sparsity of interaction sequences. This problem manifests under small sample conditions as insufficient available training samples to support parameter optimization for deep knowledge tracing models (Bradshaw & McDonald, 2023; Zhao, et al., 2025). To address this challenge, this paper approaches the issue from the perspective of cognitive state transfer, constructing the classification task as a clustering problem of cognitive fingerprint vectors in the latent space. The classification criterion shifts from the content features of resources to the cognitive effects evoked by those resources (Tong & Ren, 2025; Saqr, et al., 2023). Based on this task construction, the formal definition of interaction sequences is first presented below.

Let $S = \{s_1, s_2, \ldots, s_N\}$ be the set of minimal sample interaction sequences, where N is the number of students participating in the experiment. In this scenario, N has a finite number of values. Each sequence s_i records the trajectory of a single student's answers to a number of test questions within a specified time period, $s_i =$

$\{(q_1, a_1), (q_2, a_2), \ldots, (q_T, a_T)\}$, where q_t is the index of the question answered at step t, and $a_t \in \{0, 1\}$ is a binary label indicating whether the answer is correct or incorrect, where $a_t = 1$ indicates correct and $a_t = 0$ indicates incorrect. The sequence length T varies depending on the number of students answering, and the total number of original sequences is N. Since localized resource usage scenarios are inherently niche, the aforementioned sample size is insufficient to directly drive sufficient training of the attention knowledge tracking model, thus highlighting the necessity of sample expansion (Ullah, et al., 2025). To overcome this data bottleneck, a sliding window internal augmentation strategy is introduced. A window length L and a step size d are set. The value of L does not exceed the minimum value of all sequence lengths T to ensure that a single slice is always completed within the sequence boundaries. d satisfies $d < L$, ensuring that there are overlapping intervals between adjacent windows. For the sequence s_i, the window is moved successively from the starting position with a step size d. Each time, a continuous response segment of length L is extracted as an independent subsequence. The subsequence generated by the j-th extraction is denoted as $\tilde{s}_{i,j}$.

$$\begin{aligned}\tilde{s}_{i,j} = \{&(q_1 + (j-1)d, a_1 + (j-1)d),\\ &(q_2 + (j-1)d, a_2 + (j-1)d), \ldots,\\ &(q_L + (j-1)d, a_L + (j-1)d)\}\end{aligned} \tag{1}$$

For an original sequence of length T, the number of subsequences generated is $\lfloor (T-L)/d \rfloor + 1$. After traversing the entire original sequence, the augmented subsequence set S' gathers all the sub-segments generated by each sequence as follows.

$$S' = \bigcup_{i=1}^{N} \{\tilde{s}_{i,1}, \tilde{s}_{i,2}, \ldots, \tilde{s}_{i,k_i}\} \tag{2}$$

In the equation, $k_i = \lfloor (T_i - L)/d \rfloor + 1$ represents the number of subsequences generated after the i-th original sequence is segmented. The total number of samples S' can be obtained by summing the number of subsequences of each sequence.

$$M = \sum_{i=1}^{N} k_i \tag{3}$$

Under typical parameter settings, M can reach several times the level of N, thus expanding the sample size from N to M.

The augmentation operation is strictly limited to a single sequence. Each subsequence inherits the complete response records and their temporal sequence relationships from L consecutive steps in the original sequence.

This process does not introduce synthetic data or external noise, and the cognitive state dependency structure between questions in the original interaction record is preserved in each subsequence (Huang, et al., 2025; Su, et al., 2024). The amplified subsequence set S' is fed into the subsequent attention knowledge tracking model with a larger sample size, thereby alleviating the dilemma of excessively large gradient estimation variance and underfitting decision boundaries in deep models under small sample conditions (Abdelrahman, et al., 2023). The window length L determines the temporal span of the cognitive states carried by a single subsequence, and the step size d controls the degree of information overlap between adjacent subsequences. Together, they regulate the trade-off between the sample augmentation factor and the integrity of sequence information.

B. COGNITIVE STATE ENCODING BASED ON ATTENTION-BASED KNOWLEDGE TRACKING

The sliding window augmentation expands the original sequence set S into a subsequence set S', alleviating the sample size bottleneck for deep model training under small sample conditions. This section deploys an attention-based knowledge tracking model on top of the augmented sequence set S', encodes each subsequence step-by-step, extracts the latent vector representation of the student's knowledge state, and defines the cognitive fingerprint of the resource based on this.

The attention-based knowledge tracking model replaces the loop structure with a self-attention mechanism, directly establishing dependencies between the responses of any two time steps in the sequence (Khosla, et al., 2026; Xiao, et al., 2023). For a subsequence $\tilde{s}_{i,j}$ in the subsequence set S', the model jointly embeds each question index q_t and the corresponding response a_t into an input vector x_t. x_t is generated by a linear combination of the question embedding matrix and the response embedding matrix, carrying information about the question identity and the correctness of the answer at that time step. The input vector sequence $\{x_1, x_2, \ldots, x_L\}$ calculates the attention weight of each position to all positions in the sequence through a self-attention layer. The attention weight $\alpha_{t,\tau}$ of step t to step τ is obtained by scaling the inner product of the query vector Q_t and the key vector K_τ and then normalizing it using softmax.

$$\alpha_{t,\tau} = \frac{\exp(Q_t \cdot K_\tau / \sqrt{D})}{\sum_{j=1}^{L} \exp(Q_t \cdot K_j / \sqrt{D})} \tag{4}$$

In the equation, D is the dimension of the attention head. The context vector c_t at each time step is obtained by weighted summation of the attention weights over the value vector V_τ at all positions.

$$c_t = \sum_{\tau=1}^{L} \alpha_{t,\tau} V_\tau \tag{5}$$

The context vector c_t encapsulates all historical information in the sequence that has cognitive associations with the

answer at step t. The distribution of attention weights $\alpha_{t,\tau}$ directly reflects the model's strength in capturing cognitive dependencies between distant questions within the sequence.

To separate the influence of question difficulty and student ability on the answering results, the model introduces a Rasch embedding layer (Cheng, et al., 2022). The question difficulty parameter d_{q_t} and the student ability parameter θ participate in knowledge state updates through learnable scalar embeddings. The latent knowledge state vector h_t at step t is jointly determined by the context vector c_t, the latent vector h_{t-1} from the previous time step, the question difficulty d_{q_t}, and the student ability θ.

$$h_t = \mathrm{FFN}(c_t + h_{t-1} + d_{q_t} \cdot \theta) \tag{6}$$

FFN stands for Feed-Forward Network. h_t encodes the student's multidimensional cognitive state after answering the question at step t, with each dimension corresponding to the mastery strength of different knowledge components. At each time step, the model simultaneously outputs the predicted probability $p_t = \sigma(W_o \cdot h_t + b_o)$ for answering the next question correctly, and calculates the prediction loss L_{pred} to drive parameter optimization.

$$L_{\mathrm{pred}} = -\sum_{t=1}^{L-1} \left[a_{t+1} \log p_t + (1 - a_{t+1}) \log(1 - p_t)\right] \tag{7}$$

Optimization of the prediction loss enables the latent vector h_t to accurately capture the dynamics of students' knowledge acquisition.

After completing the time-step encoding of the subsequences, the two latent vectors adjacent to the resource usage event in each subsequence are extracted as snapshots of the cognitive state before and after use. The resource cognitive fingerprint Δh is defined as the change in the student's latent state before and after using the resource, $\Delta h = h_{t+\Delta} - h_t$, where t corresponds to the last time step before resource use, and $t + \Delta$ corresponds to the first time step after resource use. Δh encodes the multidimensional cognitive structure changes triggered by the resource in vector form. Resources with different functional orientations will generate cognitive fingerprints with differences in both direction and magnitude (Li, et al., 2025). The cognitive fingerprint vectors of all resources constitute a unified set of input representations, providing a data foundation for subsequent comparative optimization and fine-grained functional clustering.

C. COGNITIVE REPRESENTATION OPTIMIZATION BASED ON CONTRAST CONSTRAINTS

In the pre-encoding stage, all localized educational resources have been mapped to a cognitive fingerprint vector set $\Delta h = \{\Delta h_1, \Delta h_2, \ldots, \Delta h_K\}$, where K is the total number of resources. These vectors directly carry the changes in students' cognitive states before and after using the resources. However, due to the small sample size of the interaction sequences, the estimation accuracy of individual vectors fluctuates randomly (Wang, et al., 2024). To enhance the functional discriminative power of cognitive fingerprints under conditions of limited samples, this section introduces contrastive constraints to apply structured regularization to the vector space.

The approximate functional category labels $y_k \in \{1, 2, 3\}$ of resources are obtained from the teacher's weak supervision signals. These three categories respectively point to native language prosody activation, local empathy evoked, and contextualized knowledge anchoring. Resources with similar labels share similar cognitive effects, while resources with dissimilar labels elicit different cognitive transfer patterns. Based on this prior, a positive sample set $P(i) = \{j \mid y_j = y_i,\ j \neq i\}$ is defined for each resource, representing other resources with the same label as resource i; and a negative sample set $N(i) = \{j \mid y_j \neq y_i\}$. The division of positive and negative sample pairs does not depend on the textual content or subject attributes of the resources, but is only constrained by the teacher's weak classification signals at the functional level. The similarity between two vectors is defined as cosine similarity in the cognitive fingerprint space.

$$\mathrm{sim}(\Delta h_i, \Delta h_j) = \frac{\Delta h_i \cdot \Delta h_j}{\|\Delta h_i\| \|\Delta h_j\|} \tag{8}$$

The contrast loss is expressed in the form of InfoNCE, which calculates the normalized logarithm loss of the similarity between anchor point i and its positive samples relative to the similarity of all non-anchor point samples for each batch of resources. Assuming the current batch contains B resource indexes, the InfoNCE loss is as follows.

$$L_{\mathrm{cl}} = -\frac{1}{B} \sum_{i=1}^{B} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathrm{sim}(\Delta h_i, \Delta h_p)/\tau)}{\sum_{j \neq i} \exp(\mathrm{sim}(\Delta h_i, \Delta h_j)/\tau)} \tag{9}$$

In the equation, τ is a temperature coefficient that adjusts the concentration of the similarity distribution. The numerator of the InfoNCE loss encourages anchor points and their positive samples to move closer together in the latent space, while the denominator forces anchor points to move further apart from all non-anchor samples, achieving intra-class clustering and inter-class separation through a normalization competition mechanism (Zhang, et al., 2023; Hu, et al., 2025).

By jointly weighting the contrastive loss and the prediction loss L_{pred} from the attention knowledge tracking model, the end-to-end optimization objective is as follows.

$$L = L_{\mathrm{pred}} + \lambda L_{\mathrm{cl}} \tag{10}$$

In the equation, λ is the contrastive loss weight coefficient, which adjusts the proportion of the InfoNCE regularization term in the joint optimization. During the joint optimization process, the predictive loss L_{pred} ensures the

fitting accuracy of the latent vectors to the answer results, while the contrastive loss L_{cl} reorganizes the geometric distribution of the vectors at the functional category level. It is crucial to distinguish the teacher's "weak signals" from the model's final functional labels: the former act as a coarse structural prior to regularize the embedding space, whereas the latter are emergent patterns derived from the high-dimensional variance in students' actual cognitive state transitions (Δh). The contrastive loss guides the optimization trajectory, but the model must still reconcile these priors with the objective response data, preventing simple label memorization. The synergistic effect of the two types of signals enables the model to extract functionally sensitive cognitive representations from a limited number of interaction sequences even with extremely small sample settings (Xu, et al., 2023). When λ is appropriately chosen, the cognitive fingerprint vectors of similar resources cluster tightly under cosine distance, while the vectors of dissimilar resources are far apart. This latent space structure directly corresponds to the clustering assumption of fine-grained functional classification, providing input features with natural separability for subsequent clustering algorithms.

D. FINE-GRAINED FUNCTIONAL CLUSTERING BASED ON COGNITIVE FINGERPRINTS

After the contrastive constraint optimization, the cognitive fingerprint vector set Δh exhibits a geometric structure of intra-class clustering and inter-class separation in the latent space. This section builds upon this by performing fine-grained functional clustering, transforming the geometric distribution in the continuous vector space into discrete functional category labels.

The clustering algorithm employs hierarchical agglomerative clustering, using the cosine distance between cognitive fingerprint vectors as a measure of dissimilarity. Hierarchical agglomerative clustering does not require pre-defined cluster centers; it merges the closest vector pairs layer by layer from bottom to top, and its merging trajectory naturally matches the local density structure formed by the contrastive constraints (Guan & Yuen, 2022; Ran, et al., 2023). Let the cosine distance between resource i and resource j be represented as follows:

$$d(\Delta h_i, \Delta h_j) = 1 - \text{sim}(\Delta h_i, \Delta h_j) \tag{11}$$

In the equation, $\text{sim}(\Delta h_i, \Delta h_j)$ represents the cosine similarity between two vectors.

In the initial state of clustering, each resource is treated as an independent cluster $C_i = \{i\}$. Subsequently, inter-cluster distance calculation and merging operations are performed iteratively. The inter-cluster distance adopts the average-linkage criterion, which is the mean of the cosine distances of all vectors within two clusters.

$$d_{\text{avg}}(A, B) = \frac{1}{|A| \cdot |B|} \sum_{i \in A} \sum_{j \in B} d(\Delta h_i, \Delta h_j) \tag{12}$$

The iterative process continues until the number of available clusters is reduced to a preset value G, where G is set to 3, consistent with the number of functional categories in the teacher's weak supervision signal. After clustering terminates, G clusters are output, each cluster corresponding to a type of local educational resource with similar cognitive effects.

The mapping of clustering results to fine-grained functional labels is determined based on the distribution of teacher weak supervision labels for resources within the cluster. Let C_g be the set of resource indexes contained in cluster g. The frequency distribution of teacher weak supervision labels $y \in \{1, 2, 3\}$ within this cluster is statistically analyzed, and the cluster is named according to the functional category corresponding to the majority of labels. Three functional label categories are used as previously defined: native language prosody activation, local empathy arousal, and contextualized knowledge anchoring. Specifically, "native language prosody activation" refers to resources such as audio recordings of Shaanxi dialect poetry recitations that trigger phonological awareness; "local empathy arousal" includes visual materials like documentaries on the Xi'an ancient city wall to evoke cultural identity; and "contextualized knowledge anchoring" involves localized problem sets, such as mathematical word problems centered on regional culinary heritage (e.g., calculations involving the history of Xi'an Roujiamo), to bridge abstract concepts with local life. Defining the mapping rule $\phi : \{1, 2, 3\} \rightarrow \{$"Mother tongue prosody activation", "Local empathy awakening", "Contextualized knowledge anchoring"$\}$, the final classification label of each cluster g is $\phi(\text{majority}(\{y_i \mid i \in C_g\}))$.

III. EXPERIMENTAL SETUP

A. EXPERIMENTAL DATA AND ACQUISITION PROCESS

The experiment selects eight localized educational resources as research subjects, classifying them into three functional orientations based on teachers' weak categorization signals at the functional level. Each orientation contained three, three, and two resources, respectively. The "Mother Tongue Prosody Activation" category includes three resources: audio recordings of dialect classical poetry recitations, videos of dialect children's songs, and dialect tone comparison practice courseware. The "Local Empathy Awakening" category includes three resources: short documentaries of local traditional festivals, videos demonstrating local handicrafts, and audio picture books of local folk tales. The "Contextualized Knowledge Anchoring" category includes two resources: interactive courseware on chronological sequencing of local historical events and a collection of mathematical word problems related to local food culture. These eight resources encompass multiple modalities in presentation, including audio, video, interactive courseware, and picture books, and covered different subjects such as economics, management, finance, and accounting.

The experiment involves sixty undergraduate students

majoring in finance and economics from the local region, all of whom are native speakers of the local dialect and had no prior systematic exposure to the eight localized educational resources mentioned above. The experiment employs a pre-test-view-post-test repeated measurement design. In the pre-test phase, all participants complete a knowledge test corresponding to the eight resources. The test includes four question types: dialect pronunciation recognition, knowledge of regional industrial development, comprehension of local business culture, and application questions concerning the economic impact of local food industries. The "empathy" dimension is thus quantified by the statistically significant gain in accuracy on questions requiring an understanding of local cultural values and narrative empathy. There are two questions of each type, for a total of eight questions, and the results are recorded using binary judgment. In the viewing phase, participants independently accessed all eight resources in Latin square order. After viewing or using each resource, they immediately complete two post-test questions directly related to that resource. The post-test questions assess the same knowledge constructs as the corresponding question types in the pre-test, but the wording is modified to capture changes in cognitive state before and after resource use. All answer records are exported from the online experimental platform as a five-column table, with column attributes as follows: Student (ID), Resource ID, Question ID, Correct/Incorrect Answer Mark, and Answer Timestamp. The dataset contains 1,920 original answer records from 60 undergraduate students for two questions before and after each of the eight resources. The average length of a single sequence is 32 steps, which forms the data foundation for training the attention knowledge tracking model.

B. BASELINE MODEL AND ABLATION DESIGN

Three baseline methods are set up for comparison. Separated Self-Attentive Neural Knowledge Tracing (SAINT) employs a separate self-attention encoder to process question embeddings and answer embeddings separately before cross-fusion, representing a relatively mature variant of Transformer-type knowledge tracing models. Learning Process-consistent Knowledge Tracing (LPKT) explicitly models the absorption and forgetting state transitions during the learning process within the Long Short-Term Memory (LSTM) framework, using a gating mechanism to separate knowledge retention and decay at different stages. Convolutional Knowledge Tracing (CKT) replaces the recurrent structure with a one-dimensional convolutional neural network to capture sequence dependencies, extracting local and global features from answer patterns by progressively expanding the receptive field through multiple convolutional kernels. All three baselines are trained solely using answer prediction loss, without introducing contrastive constraints or cognitive fingerprint mechanisms.

The ablation design comprises three operations: removing the InfoNCE contrastive loss to test the necessity of contrastive regularization; replacing the AKT encoder with a standard LSTM encoder to test the contribution of the attention mechanism to capturing long-range cognitive dependencies; and proportionally splitting the full training data into four portions (100%, 50%, 25%, and 10%) to test the robustness of the method under data sparsity conditions. All hyperparameters remained consistent across ablation experiments except for the manipulated variable.

C. EXPERIMENTAL ENVIRONMENT AND HYPERPARAMETER CONFIGURATION

This paper's framework is implemented using PyTorch 1.13 and the pyKT 0.2 toolkit. Experiments are conducted on an NVIDIA GeForce RTX 3060 GPU with 12GB of VRAM and CUDA version 11.7. The window length L of the sliding window augmentation is set to 10, and the step size d is set to 3. This combination balanced the information integrity and augmentation factor of the subsequences under the condition of an average sequence length of 32 steps. In the attention knowledge tracking model, the attention head dimension D is set to 64, the number of self-attention layers is 2, and the Rasch embedding dimension is 32. The temperature coefficient τ of the InfoNCE contrastive loss is fixed at 0.1. The contrastive loss weight coefficient λ is set to a default value of 0.1 after parameter sensitivity analysis, which represents the optimal balance between prediction accuracy and representation separability under different data scales. The model uses the Adam optimizer with a learning rate of 1×10^{-3}, a batch size of 32, a maximum training epoch of 200, and an early stopping strategy, terminating training if the loss does not decrease for 15 consecutive epochs. In the ablation experiments, the AKT w/o Contrastive variant set λ to zero to remove the contrast constraint, and the DKT w/o Attention & Contrastive variant replaced the encoder with a single-layer LSTM and also set λ to zero. The remaining hyperparameters remained consistent with the full framework.

IV. RESULTS ANALYSIS

A. VISUALIZATION OF COGNITIVE FINGERPRINT VECTOR DIMENSIONALITY REDUCTION AND CLUSTERING STRUCTURE PRESENTATION

To examine the effectiveness of the contrast-enhanced attention knowledge tracking framework in differentiating cognitive function categories of localized educational resources, hierarchical agglomerative clustering is performed on the cognitive fingerprint vectors output by the framework for eight resources, and the correspondence between the merged trajectories and resource function labels is presented in Figure 1. The eight resources are coded according to their function categories as follows: Category A represents the native language prosody activation category, including dialect ancient poetry recitation audio A1, dialect children's song learning video A2, and dialect tone comparison exercise courseware A3; Category B represents the local empathy awakening category, including local traditional

festival documentary short film B1, local handicraft demonstration video B2, and local folk tale audio picture book B3; Category C represents the contextualized knowledge anchoring category, including local historical event chronological sequencing interactive courseware C1 and a set of mathematical word problems in local food culture C2.

Figure 1 illustrates the hierarchical aggregation paths of the three types of functional resources along the cosine distance dimension. All three resources of type A and type B complete the merging of similar resources within a low distance range below 0.15. The two resources of type C aggregate into an independent sub-cluster at 0.13. The distance span traversed by merging resources of the same type is narrower than the span of subsequent merging of resources of different types, indicating that similar resources have formed a compact local neighborhood in the latent space. After each of the three sub-clusters closes, type A and type C merge at 0.42, and this merged cluster connects with type B at 0.58 to form a root node. The cosine distance required for merging resources of different types is approximately three to five times that of merging resources of the same type. This difference stems from the combined effect of the contrast constraint continuously narrowing the distance between positive sample pairs and widening the distance between negative sample pairs during the optimization process, causing the representational boundaries between different functional categories to exhibit a jump in cosine distance after convergence. At the truncation threshold of 0.35, the three clusters are completely separated. The cluster division corresponded to the teacher's weak supervision label and there is no cross-cluster intersection. This verifies that the cognitive diagnostic representation spontaneously formed a clustering architecture consistent with the boundary of the resource teaching function under the contrast constraint.

B. FINE-GRAINED CLUSTERING QUALITY ASSESSMENT

To verify the advantages of the contrast-enhanced attention knowledge tracking framework over the relatively independent baseline model in the fine-grained clustering task of resources, the proposed method and three baseline models, SAINT, LPKT, and CKT, are run on eight localized educational resources. The resource cognitive representations output by each model are extracted and hierarchical agglomerative clustering is performed. The cluster density is quantified by the silhouette coefficient. The comparison results under the two training conditions of full data and 10% data are shown in Figure 2.

Figure 2 shows that the silhouette coefficients of the proposed method reached 0.72 and 0.68 under the full data and 10% data conditions, respectively, both higher than the best performance of the three baseline models. SAINT, by using separate self-attention to encode the question embedding and the answer embedding separately, captured some cognitive dependencies within the sequence, and the silhouette coefficients are 0.65 and 0.60 under the two conditions, respectively, which is relatively high among the three baselines. LPKT is limited by the gated forgetting mechanism of LSTM and has insufficient ability to model long-range implicit cultural cognitive associations, with silhouette coefficients dropping to 0.58 and 0.50. CKT, due to the difficulty of the local receptive field of the convolution kernel in covering cognitive transfer information between distant questions, had silhouette coefficients of only 0.52 and 0.42, ranking last among the three baselines. The silhouette coefficient decay of the three baselines is significantly greater than that of the proposed method when the data is reduced from the full data to 10%, indicating that the latent space structure of the pure prediction-driven model lacking contrast constraints and cognitive fingerprinting mechanisms is more prone to rapid degradation due to the decrease in sample size under small sample conditions. The proposed method achieves leading cluster density under both abundant and extremely sparse data conditions, verifying the regularization effect of contrastive constraints on the functional separability of latent space and the gain of cognitive fingerprint differential coding on resource functional representation.

C. DEVIATION ANALYSIS BETWEEN MODEL CLASSIFICATION AND COARSE-GRAINED LABELS

To reveal the essential difference between the fine-grained functional classification output in this paper and the traditional coarse-grained subject classification, Figure 3 presents the attribution relationship of eight localized educational resources under the two labeling systems. The original coarse-grained subject labels carried by the eight resources are Economics, Management, Finance, and Accounting, while the fine-grained functional labels are categories A, B, and C.

In Figure 3, resources under the "Economics" category flow into categories A and B respectively; the single resource under the "Management" category flows into category A; resources under the "Finance" category flow into category B; and resources under the "Accounting" category flow into category C. Resources within the same subject label belong to different categories in terms of function, reflecting that coarse-grained subject classification cannot uniformly characterize the cognitive-evident effects of resources. The Normalized Mutual Information (NMI) is 0.42, far below the theoretical maximum of 1.0, indicating only a weak correlation between the model's output functional classification and subject classification. The transformation between the two label systems confirms that fine-grained classification based on cognitive state transfer captures implicit teaching functional attributes that subject labels cannot represent, and the classification boundary driven by cognitive diagnosis is independent of the traditional content classification system.

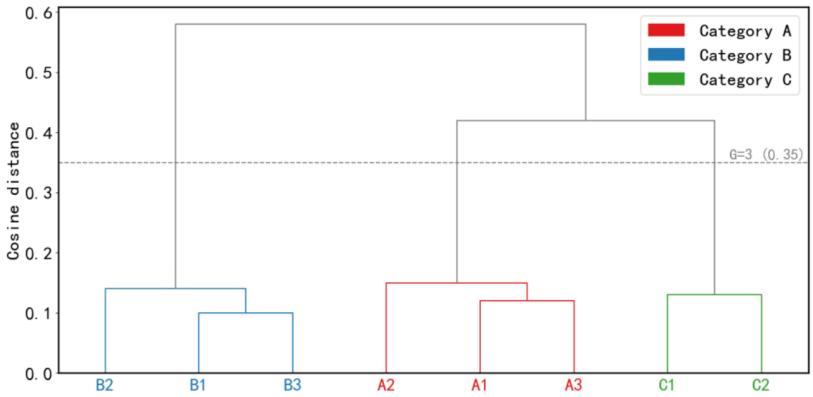


FIGURE 1. Hierarchical clustering tree diagram of eight localized educational resource cognitive fingerprints

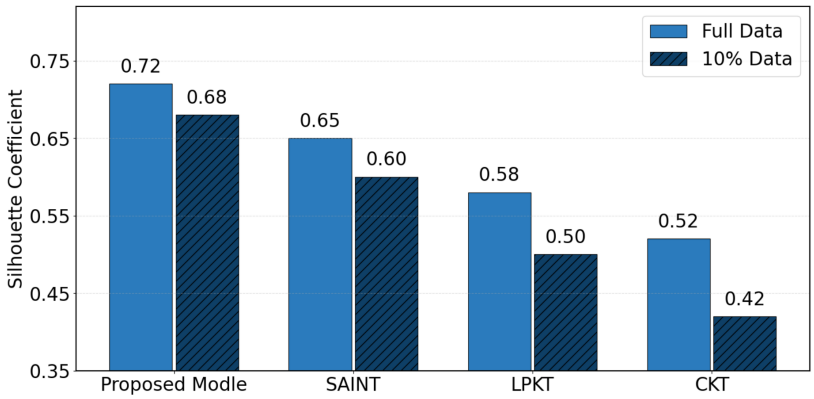


FIGURE 2. Fine-grained clustering profile coefficients of different models under full and 10% data conditions

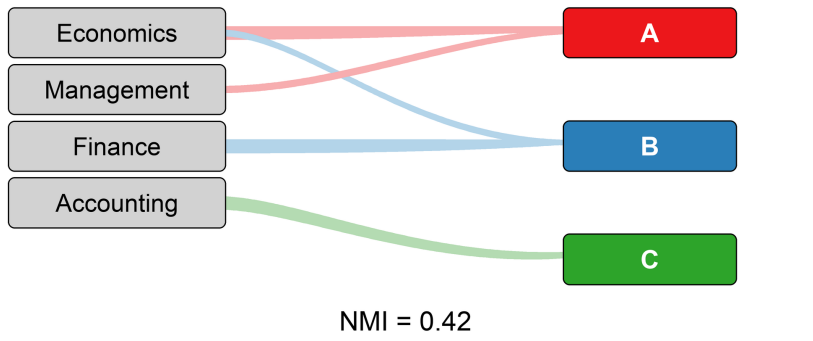


FIGURE 3. Flow of resources from coarse-grained subject tags to fine-grained functional tags

D. ROBUSTNESS VERIFICATION OF THE METHOD UNDER DATA ABLATION

To test the robustness of our proposed method under sparse data conditions, the training data is successively reduced at four ratios: 100%, 50%, 25%, and 10%. At each ratio, the model is retrained, and fine-grained clustering is performed on eight localized educational resources. The silhouette coefficient is used to measure the degradation trend of clustering quality with decreasing data volume. The comparison curves include our proposed method, the variant AKT with contrastive loss module removed, and the variant DKT with attention & contrastive loss removed simultaneously. The silhouette coefficient changes for the three configurations under the four data ratios are shown in Figure 4.

Figure 4 shows that the silhouette coefficients of the proposed method are 0.72, 0.71, 0.70, and 0.68 for the

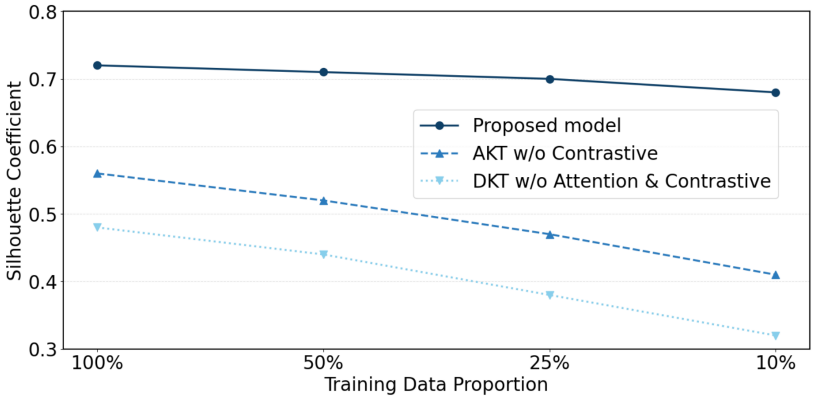


FIGURE 4. Decrease curves of fine-grained clustering silhouette coefficients under different training data ratios

four data proportions, respectively. The curve remains high throughout and the decay is extremely gentle, decreasing by only 0.04 from the full data set to 10%. AKT w/o Contrastive has a silhouette coefficient of 0.56 at 100% data, decreasing to 0.52 at 50%, further decreasing to 0.47 at 25%, and finally dropping to 0.41 at 10%, with a decay of 0.15. The curve begins to decline significantly from 50%. DKT w/o Attention & Contrastive has a silhouette coefficient of only 0.48 at 100% data, subsequently decreasing to 0.44, 0.38, and 0.32, with a decay of 0.16. The curve remains low throughout and drops sharply as the data decreases. The comparison of the three sets of curves separates the small-sample gains of the two components: the introduction of the attention encoder improves the silhouette coefficient by 0.08 under abundant data conditions, and maintains a difference of approximately 0.09 under extremely sparse data conditions; the superposition of the contrastive loss module provides an additional improvement of 0.16 under 100% data and an additional improvement of 0.27 under 10% data. Both differences widen with data reduction, indicating that the regularization effect of the contrastive constraint is more prominent when the sample size decreases. The proposed method outperforms both ablation variants under all data proportions, validating the supporting role of joint optimization of attention encoding and contrastive constraints in the robustness of fine-grained clustering of localized educational resources.

E. IMPACT OF CONTRAST LOSS WEIGHTS ON CLASSIFICATION PERFORMANCE

To examine the moderating effect of the contrast loss weight coefficient λ on fine-grained clustering performance, λ is set to five values (0.01, 0.1, 0.5, 1.0, and 2.0) under four different training data ratios. The framework in this paper is trained and the clustering silhouette coefficients are calculated. The distribution of silhouette coefficients under different conditions is shown in Figure 5.

Figure 5 illustrates the systematic variation of the silhouette coefficient with λ and data proportion. Under 100% data conditions, the silhouette coefficient reaches a peak of 0.72 when λ is 0.1. When λ decreases to 0.01 or increases to 2.0, the coefficient drops to 0.68 and 0.58, respectively. Under 50% data conditions, the peak remains at $\lambda = 0.1$, with a silhouette coefficient of 0.71. At $\lambda = 0.01$ and $\lambda = 2.0$, the coefficients decrease to 0.67 and 0.56, respectively. Under 25% data conditions, the peak remains at 0.70, with the coefficient decreasing to 0.64 and 0.52 when λ moves towards either end. Under 10% extremely sparse conditions, the peak of 0.68 still falls at $\lambda = 0.1$. At $\lambda = 0.01$, the coefficient decreases to 0.57, and at $\lambda = 2.0$, it decreases to 0.44, with the decrease further intensifying compared to when data is abundant. The optimal λ for each data proportion row is concentrated at 0.1, indicating that this value maintains a stable regularization strength as the data size changes. When λ is too small, the contrastive constraints are insufficient, limiting the separation between classes in the latent space; when λ is too large, the contrastive loss dominates, causing the latent vectors to deviate from the cognitive diagnostic objective, resulting in accelerated degradation of clustering quality at the data sparsity end. Selecting the contrastive loss weight around 0.1 can balance prediction accuracy and representation separability under different data conditions, providing a clear range of values for the hyperparameter selection of the framework in practical deployment.

V. CONCLUSION

This paper proposes a contrast-enhanced attentional knowledge tracking framework, which shifts the fine-grained classification of localized educational resources from the content feature space to the cognitive state transfer space. The framework uses attentional knowledge tracking as its encoding backbone, extracting latent vectors of student knowledge states from augmented interaction sequences within a sliding window. A cognitive fingerprint is constructed based on changes in latent states before and after resource use. InfoNCE contrastive constraints, guided by weak teacher supervision signals, apply structured regularization to the cognitive fingerprint space. Finally, hierarchical agglomerative clustering outputs fine-grained categories corresponding to the implicit teaching functions of the resources.

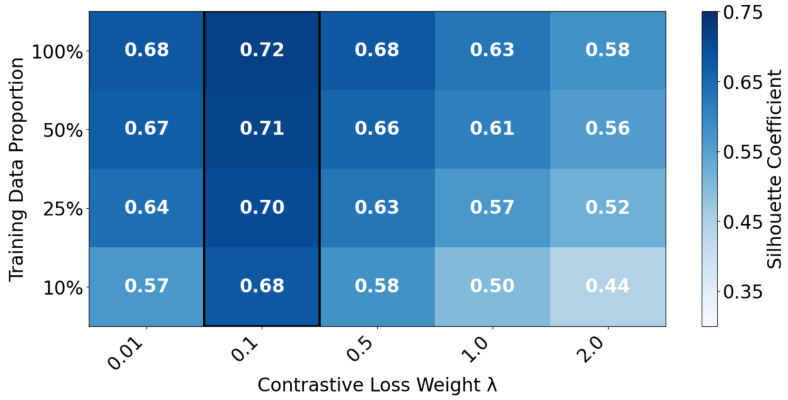


FIGURE 5. Heatmap of cluster profile coefficients for comparing loss weights under different data scales

Experiments are conducted on eight localized resources covering three functional orientations. The results show that contrastive constraints drive the formation of compact clusters of similar resources in the latent space. The proposed method exhibits superior cluster density compared to three independent baseline models. The joint design of attentional encoding and contrastive loss maintains a gradual decline in clustering performance even when the training data is reduced to 10%. The functional classification output by the model shows a weak correlation with coarse-grained subject labels, and the classification boundary driven by cognitive diagnosis is independent of the content feature system. The experimental resources in this paper all originated from a single dialect region. The generalization performance of cross-regional cultural resources needs to be tested, and the adaptability of the three-category classification of weak teacher supervision signals to resources with ambiguous boundaries needs further investigation. Future work will expand to multi-regional resources, examine the robustness of the method under conditions of weak supervision signal degradation, and study the distillation strategy of cognitive encoders to lightweight content-side models to reduce the dependence of new resource classification on interactive data.

AUTHOR CONTRIBUTIONS

Chunyan Zhang designed, collected and analyzed the data, and drafted the manuscript. Chunyan Zhang conducted the study, critically revised the manuscript for important intellectual content, and gave final approval of the version to be published. Chunyan Zhang participated fully in the work, take public responsibility for appropriate portions of the content, and agreed to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

CONFLICTS OF INTEREST

The author declares no conflict of interest.

FUNDING

This work was supported by the 2026 Higher Education Research Project of Zhejiang Association of Higher Education “Research on the Localized Model of AI Empowering the Cultivation of Health and Wellness Talents in Zhejiang” (Grant No. KT2026160); the Higher Education Research Project of Ningbo Association of Higher Education “Research on the Mechanism and Path of Interdisciplinary Curriculum for Ningbo’s ‘361’ Industries from the Perspective of Industry-Education Symbiosis”; and the Ministry of Education Industry-University Cooperative Education Project “Construction and Practice of a Comprehensive Practice Curriculum System for Digital and Intelligent Tax Planning” (Grant No. 241100353214029).

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

[1] G. Abdelrahman, Q. Wang, and B. Nunes, “Knowledge tracing: A survey,” ACM Computing Surveys, vol. 55, no. 11, pp. 1-37, 2023, doi: 10.1145/3569576.

[2] E. D. Bradshaw and J. K. McDonald, “Informal practices of localizing open educational resources in Ghana,” International Review of Research in Open and Distributed Learning, vol. 24, no. 2, pp. 18-36, 2023, doi: 10.19173/irrodl.v24i2.7102.

[3] Y. Cheng, G. Wu, H. Zou, P. Luo, and Z. Cai, “A knowledge query network model based on rasch model embedding for personalized online learning,” Frontiers in Psychology, vol. 13, pp. 846621-846628, 2022, doi: 10.3389/fpsyg.2022.846621.

[4] H. Dai, Y. Yun, Y. Zhang, R. An, W. Zhang, and X. Shang, “Self-paced contrastive learning for knowledge tracing,” Neurocomputing, vol. 609, pp. 128366-128371, 2024, doi: 10.1016/j.neucom.2024.128366.

[5] Q. Deng, Y. Guo, Z. Yang, H. Pan, and J. Chen, “Boosting semi-supervised learning with Contrastive Complementary Labeling,” Neural Networks, vol. 170, pp. 417-426, 2024, doi: 10.1016/j.neunet.2023.11.052 .

[6] G. Feng, M. Fan, and Y. Chen, “Analysis and prediction of students’ academic performance based on educational data mining,” IEEE Access, vol. 10, pp. 19558-19571, 2022, doi: 10.1109/ACCESS.2022.3151652.

[7] C. Guan and K. K. F. Yuen, “The cognitive comparison enhanced hierarchical clustering,” Granular Computing, vol. 7, no. 3, pp. 637-655, 2022, doi: 10.1007/s41066-021-00287-x.

[8] H. Hu, X. Wang, Y. Zhang, Q. Chen, and Q. Guan, “A comprehensive survey on contrastive learning,” Neurocomputing, vol. 610, pp. 128645-128652, 2024, doi: 10.1016/j.neucom.2024.128645.

[9] M. Hu, H. Chang, S. Shan, and X. Chen, “Inference calibration of Vision-Language foundation models for Zero-Shot and Few-Shot learn-

ing," Pattern Recognition Letters, vol. 192, pp. 15-21, 2025, doi: 10.1016/j.patrec.2025.03.010.
[10] B. Huang, Y. Lin, and C. Xu, "Contrastive label correction for noisy label learning," Information sciences, vol. 611, pp. 173-184, 2022, doi: 10.1016/j.ins.2022.08.060.
[11] C. Huang, W. Jiang, K. Li, J. Wu, and J. Zhang, "Enhancing learning process modeling for session-aware knowledge tracing," Knowledge-Based Systems, vol. 309, pp. 112740-112747, 2025, doi: 10.1016/j.knosys.2024.112740.
[12] D. Khairy, N. Alharbi, M. A. Amasha, M. F. Areed, S. Alkhalaf, and R. A. Abougalala, "Prediction of student exam performance using data mining classification algorithms," Education and Information Technologies, vol. 29, no. 16, pp. 21621-21645, 2024, doi: 10.1007/s10639-024-12619-w.
[13] A. Khosla, H. Singh, A. Aggarwal, and J. Bedi, "A knowledge tracing based intelligent framework for formative assessment of students skills," Complex & Intelligent Systems, vol. 12, no. 1, pp. 22-29, 2026, doi: 10.1007/s40747-025-02149-4.
[14] Q. Li, Q. Mai, M. Wang, and M. Ma, "Chinese dialect speech recognition: a comprehensive survey," Artificial Intelligence Review, vol. 57, no. 2, pp. 25-33, 2024, doi: 10.1007/s10462-023-10668-0.
[15] Q. Li, X. Yuan, J. Yue, X. Shen, R. Liang, S. Liu, and Z. Yan, "Dual-view multi-scale cognitive representation for deep knowledge tracing," Knowledge-Based Systems, vol. 310, pp. 113010-113017, 2025, doi: 10.1016/j.knosys.2025.113010.
[16] Y. Liu, Z. Wu, K. Deng, Z. Zheng, and C. Huang, "DisenKT: A variational attention-based approach for disentangled crossdomain knowledge tracing," Information Processing & Management, vol. 63, no. 3, pp. 104582-104587, 2026, doi: 10.1016/ j.ipm.2025.104582.
[17] Y. Lu, L. Tong, and Y. Cheng, "Advanced Knowledge Tracing: Incorporating Process Data and Curricula Information via an Attention-Based Framework for Accuracy and Interpretability," Journal of Educational Data Mining, vol. 16, no. 2, pp. 58-84, 2024, doi: 10.31219/osf.io/3k7z2.
[18] F. Ma, C. Zhu, and D. Liu, "A deeper knowledge tracking model integrating cognitive theory and learning behavior," Journal of Intelligent & Fuzzy Systems, vol. 46, no. 3, pp. 6607-6617, 2024, doi: 10.3233/JIFS-235723 .
[19] Y. Pu, W. Wu, T. Peng, F. Liu, Y. Liang, X. Yu, and P. Feng, "Embedding cognitive framework with self-attention for interpretable knowledge tracing," Scientific Reports, vol. 12, no. 1, pp. 17536-17542, 2022, doi: 10.1038/s41598-022-22539-9.
[20] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, "Comprehensive survey on hierarchical clustering algorithms and the recent developments," Artificial Intelligence Review, vol. 56, no. 8, pp. 8219-8264, 2023, doi: 10.1007/s10462-022-10366-3.
[21] C. E. S. Rex, J. Annrose, and J. J. Jose, "Comparative analysis of deep convolution neural network models on small scale datasets," Optik, vol. 271, pp. 170238-170244, 2022, doi: 10.1016/j.ijleo.2022.170238.
[22] B. Sabiri, A. Khtira, B. El Asri, and M. Rhanoui, "Investigating contrastive pair learning's frontiers in supervised, semi-supervised, and self-supervised learning," Journal of Imaging, vol. 10, no. 8, pp. 196-203, 2024, doi: 10.3390/jimaging10080196.
[23] M. Saqr, W. Matcha, N. A. A. Uzir, J. Jovanovic, D. Gašević, and S. López-Pernas, "Transferring effective learning strategies across learning contexts matters: A study in problem-based learning," Australasian Journal of Educational Technology, vol. 39, no. 3, pp. 35-57, 2023, doi: 10.14742/ajet.8303.
[24] X. Song, J. Li, T. Cai, S. Yang, T. Yang, and C. Liu, "A survey on deep learning based knowledge tracing," Knowledge-Based Systems, vol. 258, pp. 110036-110045, 2022, doi: 10.1016/j.knosys.2022.110036.
[25] J. Su, X. Yu, X. Wang, Z. Wang, and G. Chao, "Enhanced transfer learning with data augmentation," Engineering Applications of Artificial Intelligence, vol. 129, pp. 107602-107609, 2024, doi: 10.1016/j.engappai.2023.107602.
[26] C. Tong and C. Ren, "Deep knowledge tracing and cognitive load estimation for personalized learning path generation using neural network architecture," Scientific Reports, vol. 15, no. 1, pp. 24925-24933, 2025, doi: 10.1038/s41598-025-10497-x.
[27] F. Ullah, I. Ullah, K. Khan, S. Khan, and F. Amin, "Advances in deep neural network-based hyperspectral image classification and feature learning with limited samples: a survey," Applied Intelligence, vol. 55, no. 6, pp. 370-376, 2025, doi: 10.1007/ s10489-024-06139-w.
[28] Y. Wang, Y. Huo, C. Yang, X. Huang, D. Xia, and F. Feng, "Knowledge ontology enhanced model for explainable knowledge tracing," Journal of King Saud University-Computer and Information Sciences, vol. 36, no. 5, pp. 102065-102072, 2024, doi: 10.1016/j.jksuci.2024.102065.
[29] A. B. Wassink, C. Gansen, and I. Bartholomew, "Uneven success: automatic speech recognition and ethnicity-related dialects," Speech Communication, vol. 140, pp. 50-70, 2022, doi: 10.1016/j.specom.2022.03.009.
[30] Y. Xiao, R. Xiao, N. Huang, Y. Hu, H. Li, and B. Sun, "Knowledge tracing based on multi-feature fusion," Neural Computing and Applications, vol. 35, no. 2, pp. 1819-1833, 2023, doi: 10.1007/s00521-022-07834-w.
[31] J. Xu and W. Hu, "An enhanced deep knowledge tracing model via multi-band attention and quantized question embedding," Applied Sciences, vol. 14, no. 8, pp. 3425-3433, 2024, doi: 10.3390/app14083425 .
[32] L. Xu, H. Xie, Z. Li, F. L. Wang, W. Wang, and Q. Li, "Contrastive learning models for sentence representations," ACM Transactions on Intelligent Systems and Technology, vol. 14, no. 4, pp. 1-34, 2023, doi: 10.1145/3593590.
[33] A. Zhang, L. Sheng, Z. Cai, X. Wang, and T. S. Chua, "Empowering collaborative filtering with principled adversarial contrastive loss," Advances in Neural Information Processing Systems, vol. 36, no. 1, pp. 6242-6266, 2023, doi: 10.52202/075280-0273.
[34] Y. Zhang, X. Cao, W. Li, and L. Zhang, "A fine-grained course session recommendation method based on knowledge point pruning," Scientific Reports, vol. 15, no. 1, pp. 13164-13169, 2025, doi: 10.1038/s41598-025-95395-y.
[35] W. Zhao, Z. Xu, and L. Qiu, "BPSKT: Knowledge Tracing with Bidirectional Encoder Representation Model Pre-Training and Sparse Attention," Electronics, vol. 14, no. 3, pp. 458-465, 2025, doi: 10.3390/electronics14030458 .