International Scientific Technical and Economic Research (ISSN: 2959-1309)
Vol. 4, No. 2, (March 2026), pp. 1-20
Journal homepage: https://www.istaer.online

RESEARCH ARTICLE DOI: 10.71451/ISTAER2613

Research on Long Sequence Learning Behavior Modeling Based on Transformer-XL

Yuxiao Qin*

Education College, Seoul School of Integrated Sciences and Technologies, Seoul, Republic of Korea

Abstract: In view of the difficulties in modeling long-sequence dependence and the high computational complexity of online learning behavior data, this paper proposes a long sequence learning behavior modeling method based on Transformer-XL. This method improves the performance of the model from the two levels of structure and information modeling by constructing multidimensional behavior feature representation, integrating dynamic memory enhancement mechanism, behavioral semantic perception attention and sparse long sequence modeling strategy. Experimental results on real educational datasets such as ASSISTments and EdNet show that the proposed model is superior to the mainstream methods in terms of AUC, ACC and RMSE. The AUC increases by about 4.2% and RMSE decreases by about 8.1%. Further ablation experiments and parameter analysis verify the effectiveness of each module. Cross dataset experiments and noise tests show that the model has good generalization ability and robustness. In addition, interpretability analysis shows that the model can effectively focus on key learning behaviors. The results show that this method has significant advantages in the long sequence learning behavior modeling task, and provides effective support for personalized recommendation and learning state evaluation in intelligent education system.

Keywords: Transformer-XL; Long sequence modeling; Learning behavior analysis; Attention mechanism; Dynamic memory

How to Cite: Qin, Y. (2026). Research on Long Sequence Learning Behavior Modeling Based on Transformer-XL. *International Scientific Technical and Economic Research, 4*(2), 1–20. https://doi.org/10.71451/ISTAER2613

Article history: Received: 19 Jan 2026; **Revised:** 28 Feb 2026; **Accepted:** 27 Mar 2026; **Published:** 02 Apr 2026

1. INTRODUCTION

With the rapid development of online education platform and the wide application of intelligent learning system, large-scale learning behavior data shows a continuous growth trend. Learners' interactive behaviors on the platform, such as video viewing, problem solving, knowledge point jumping, and reviewing, constitute a significant time-dependent behavior sequence[1],[2]. This kind of data is not only large in scale, but also often has the characteristics of long sequence, that is, the behavior trajectory of a single user may span a long time span and contain rich learning patterns and cognitive change information. How to effectively extract key features from these long sequence data and describe the learning process has become an

*Corresponding author: Yuxiao Qin, Education College, Seoul School of Integrated Sciences and Technologies, Seoul, republic of Korea. Email: m18652061869@163.com

important research direction in the field of educational data mining and intelligent recommendation [3],[4],[5].

However, long sequence learning behavior modeling faces many challenges. Firstly, with the increase of the sequence length, the amount of information that the model needs to deal with increases exponentially, which easily leads to the problem of information redundancy and key features being submerged [6]. Secondly, there are often complex long-distance dependencies among learning behaviors. For example, early knowledge mastery may have a profound impact on subsequent learning, which imposes higher requirements on the memory capacity of the model [7],[8]. In addition, the diversity and uncertainty of behavior data, the semantic differences between different behavior types and the dynamic changes of time interval further increase the difficulty of modeling [9]. Therefore, the core problem in current research is to effectively capture long-range dependencies while ensuring computational efficiency.

To solve these problems, researchers have proposed a variety of sequence modeling methods. Early recurrent neural networks (RNNs) and their variants, such as long short-term memory networks (LSTMs), model temporal dependencies through recursive structures, but they are easily affected by vanishing or exploding gradients when dealing with long sequences, making it difficult to effectively propagate long-distance information [10],[11],[12],[13],[14]. Subsequently, the transformer model based on attention mechanism significantly improves the expression ability through global dependency modeling, but its computational complexity increases with the square of the sequence length, which makes it face the efficiency bottleneck in the super long sequence scenario [15]. At the same time, the standard Transformer lacks an explicit memory mechanism for historical information and is prone to context fragmentation when processing long sequences in segments.

Transformer-XL alleviates the above problems to a certain extent by introducing relative position encoding and a cross-segment memory mechanism, allowing the model to reuse historical hidden states and thus improve its ability to model long sequences [16],[17]. However, its memory updating strategy is relatively fixed, and it is difficult to adapt to the dynamic pattern of learning behavior; At the same time, attention computing still relies on dense connections, and the computational cost is still high under the ultra long sequence [18],[19]. In addition, the existing models generally lack the in-depth use of behavioral semantic information, and it is difficult to fully mine the internal relationship between learning behaviors.

To address this, focusing on the core problem of “long-sequence learning behavior modeling,” this paper proposes an improved model framework for complex educational data scenarios. On the basis of Transformer-XL, this method systematically optimizes the memory mechanism, attention structure and computational efficiency, enhances the adaptability of the model to historical information by introducing dynamic memory management strategy, improves the expression ability of behavior association combined with semantic perception mechanism, and reduces the computational overhead of long sequence modeling through sparse structure design, so as to achieve an effective balance between performance and efficiency.

The main contributions of this paper are reflected in the following aspects. Firstly, a dynamic memory enhancement mechanism is proposed, which enables the model to adaptively adjust the memory content according to the changes of learning behavior, and effectively improves the long-range dependency modeling ability. Secondly, the mechanism of behavioral semantic perception attention is designed, which integrates behavioral semantic information into the attention calculation process, and enhances the recognition ability of the model to key behavior patterns. Thirdly, the sparse long sequence modeling strategy is introduced to significantly reduce the computational complexity under the premise of ensuring the performance of the model, making it more suitable for large-scale data scenarios. Finally, the system experiments on several real education data sets verify the significant advantages of the proposed method in performance, robustness and interpretability, and provide a practical solution for long sequence learning behavior modeling.

2. RELATED WORK

Learning behavior modeling and knowledge tracking is an important research direction in the field of educational data mining. Its core goal is to predict the future performance or mastery state of learners by analyzing their historical interaction behavior. Early methods were mostly based on probability graph models, such as Bayesian knowledge tracking, which described knowledge mastery through implicit variables, but its expression ability was limited, and it was difficult to capture complex behavior patterns [20]. With the development of deep learning, knowledge tracking method based on neural network has gradually become the mainstream. Deep knowledge tracking uses recurrent neural network to model the learning behavior sequence, which significantly improves the prediction performance [21],[22]. Subsequently, the researchers further introduced the attention mechanism and graph structure information, and proposed a variety of improved methods, so that the model can better model the relationship and learning path between knowledge points. However, these methods still face the problems of information attenuation and computational efficiency when dealing with ultra long behavior sequences.

In the development of sequence modeling method, the model structure has evolved from recursion to attention mechanism. RNN and its variants LSTM and Gru alleviate the gradient disappearance problem through gating mechanism, making it perform well in short to medium length sequence modeling [23]. However, this kind of method essentially relies on sequential recursive computation, which is difficult to parallelize, and it still has the problem of insufficient memory ability in long sequences. By introducing the self attention mechanism, transformer model realizes the direct modeling of the dependency relationship between any position in the sequence, which greatly improves the expression ability and training efficiency [24],[25]. However, its computational complexity is squared with the sequence length, which significantly increases the resource consumption in long sequence scenarios, and lacks an effective historical information reuse mechanism.

In view of the deficiency of transformer in long sequence modeling, Transformer-XL enhances the ability of capturing long-distance dependence by introducing relative position coding and cross-segment memory mechanism. The model effectively alleviates the problem of context truncation by reusing historical hidden states, and has achieved excellent performance in language modeling and other tasks [26]. Since then, researchers have carried out a series of improvements around Transformer-XL, such as improving the efficiency of information utilization through improving the memory update strategy, or reducing the computational overhead through structural optimization. However, most of these methods focus on the optimization of a single dimension, and have not systematically designed the modeling of long sequence learning behavior from the overall architecture level.

At the same time, the long sequence modeling and optimization technology has also been widely concerned. The first method filters and reconstructs the historical information by introducing memory mechanism or compression strategy to reduce the interference caused by redundant data; The other method limits the computational range of attention by sparse attention structure, so as to reduce the time and space complexity. For example, local window attention, hierarchical attention and sparse connection method based on pattern selection all improve the efficiency of long sequence modeling to a certain extent [27]. In addition, some researches try to integrate local and global information with the idea of multi-scale modeling to enhance the ability of the model to capture different granularity behavior patterns. However, the application of these methods in educational scenes is still relatively limited, especially in behavioral semantic modeling.

Overall, the existing research has made some progress in learning behavior modeling and long sequence processing, but there are still some shortcomings. On the one hand, it is difficult for traditional methods to take into account both long-range dependence modeling ability and computational efficiency; On the other hand, most models fail to make full use of the semantic

information of learning behavior, which limits its expression ability. Compared with the existing work, this method extends Transformer-XL from multiple levels, improves the information management ability by introducing dynamic memory mechanism, enhances the depth of behavior modeling combined with semantic perception attention, and reduces the computational complexity through sparse structure optimization, so as to realize the efficient modeling of long sequence learning behavior. This multi-dimensional collaborative optimization design makes the proposed method achieve a better balance between performance and practicability.

3. METHODOLOGY

3.1 Representation of learning behavior sequence

In online learning scenarios, user behavior is usually represented by time-dependent sequential data. Let the learning behavior sequence of user u be expressed as:

$$\mathcal{S}_u = \{s_1, s_2, \dots, s_T\} \tag{1}$$

Where T is the length of the sequence, and s_t is the behavior event in time step t. Each behavior event can be further expressed as a multidimensional eigenvector:

$$s_t = (a_t, q_t, r_t, \Delta t_t) \tag{2}$$

where, a_t represents the type of behavior, q_t represents the knowledge point or topic ID, r_t represents the behavior result (such as correct/incorrect), and Δt_t represents the time interval between consecutive behaviors.

For multi-dimensional features, this study uses embedded mapping function to map discrete and continuous features to low dimensional vector space:

$$\mathrm{e}_t = \mathrm{E}_a(a_t) + \mathrm{E}_q(q_t) + \mathrm{E}_r(r_t) + \mathrm{E}_\Delta(\Delta t_t) \tag{3}$$

Where, $\mathrm{E}_a, \mathrm{E}_q, \mathrm{E}_r$ are the learnable embedding matrices respectively, and $\mathrm{E}_\Delta(\cdot)$ is the time coding function (piecewise linear or logarithmic mapping can be used). Finally, the input sequence representation is obtained:

$$\mathrm{X} = [\mathrm{e}_1, \mathrm{e}_2, \dots, \mathrm{e}_T] \in \mathbb{R}^{T \times d} \tag{4}$$

Where d is the embedded dimension.

3.2 Basic model: Transformer-XL

To address the context truncation problem, the scoring function in attention calculation is defined as:

$$\mathrm{A}_{i,j} = (\mathrm{q}_i + \mathrm{u})^\top \mathrm{k}_j + (\mathrm{q}_i + \mathrm{v})^\top \mathrm{r}_{i-j} \tag{5}$$

Where q_i and k_j are query and key vectors respectively, r_{i-j} represents relative position encoding, and u, v are learnable bias terms.

To achieve cross-segment dependency modeling, Transformer-XL introduces segment level recurrence mechanism. Let the current segment input be X^τ and the historical memory be $\mathrm{M}^{\tau-1}$, then the extended context is represented as:

$$\widetilde{\mathrm{H}}^{\tau-1} = [\mathrm{SG}(\mathrm{M}^{\tau-1}); \mathrm{H}^{\tau-1}] \tag{6}$$

Where $\mathrm{SG}(\cdot)$ means to stop the gradient operation. This mechanism enables the model to reuse historical hidden states when dealing with long sequences, so as to effectively capture

long-distance dependencies.

In the overall modeling process, the input sequence is segmented by a fixed length, and each segment carries out attention calculation in combination with historical memory when propagating forward, so as to realize the modeling ability of approximate infinite context.

3.3 Dynamic memory enhancement mechanism

To solve the problem that Transformer-XL fixed memory window is difficult to adapt to complex learning behavior patterns, this paper proposes a dynamic memory enhancement mechanism. First, define the memory update function:

$$\mathrm{M}^{\tau} = \mathrm{Update}(\mathrm{M}^{\tau-1}, \mathrm{H}^{\tau}) \tag{7}$$

Where H^{τ} represents the hidden representation of the current segment. To achieve adaptive updating, we introduce a gating mechanism:

$$\mathrm{g}_t = \sigma\left(\mathrm{W}_g \mathrm{h}_t + \mathrm{b}_g\right) \tag{8}$$

$$\mathrm{m}_t = \mathrm{g}_t \odot \mathrm{h}_t + (1 - \mathrm{g}_t) \odot \mathrm{m}_{t-1} \tag{9}$$

Where $\sigma(\cdot)$ is the sigmoid function, $\odot$ is the element by element multiplication, and g_t controls the information retention ratio.

Furthermore, memory is divided into long-term memory M^{L} and short-term memory M^{S}:

$$\mathrm{M}^{\tau} = [\mathrm{M}^{L}, \mathrm{M}^{S}] \tag{10}$$

Short term memory is used to capture local behavior patterns, and long-term memory is used to store stable learning characteristics. To avoid the accumulation of redundant information, a compression function is introduced:

$$\mathrm{M}^{L} = \mathrm{Compress}(\mathrm{M}^{L}) \tag{11}$$

This function can be implemented by average pooling or low rank projection, so as to reduce the storage overhead while maintaining key information.

3.4 Attention mechanism of behavioral semantic perception

The traditional attention mechanism does not make full use of behavioral semantic information. This paper introduces semantic perception attention to enhance the model expression ability. First, define behavioral semantic embedding:

$$\mathrm{s}_t = \mathrm{E}_s(a_t, q_t) \tag{12}$$

Where E_s is the joint semantic encoding function.

In attention calculation, semantic information is introduced into weight calculation:

$$\alpha_{i,j} = \frac{\exp\left(\mathrm{q}_i^{\top}\mathrm{k}_j + \mathrm{q}_i^{\top}\mathrm{s}_j\right)}{\sum_k \exp\left(\mathrm{q}_i^{\top}\mathrm{k}_k + \mathrm{q}_i^{\top}\mathrm{s}_k\right)} \tag{13}$$

This mechanism enables the model to focus on semantically similar behaviors when computing attention.

In addition, a multi granularity attention fusion structure is designed to jointly model local behavior patterns and global dependencies:

$$\mathrm{H} = \lambda \mathrm{H}_{local} + (1 - \lambda)\mathrm{H}_{global} \tag{14}$$

Where λ is the learnable weight parameter, H_{local} represents the local window attention

 https://doi.org/10.71451/ISTAER2613

output, and H_{global} represents the global attention result.

3.5 Sparse long sequence modeling optimization

In order to reduce the computational complexity of long sequence modeling, sparse attention structure is introduced in this paper. The traditional attention complexity is $O(T^2)$. By limiting the scope of attentional connectivity, this paper optimizes it to $O(T \cdot k)$. Where $k \ll T$ is the number of sparse connections.

Specifically, build a sparse pattern based on window and jump connection:

$$\mathcal{N}(i) = \{ j \mid \ \mid i - j \mid \leq w \} \cup \{ j \mid j = i - 2^p \} \tag{15}$$

Where w is the local window size, and jump connections are used to capture long-distance dependencies.

In the calculation process, attention calculation is performed only on the elements in the adjacent set $\mathcal{N}(i)$, which significantly reduces the computational and storage overhead. At the same time, combined with the segmented computing strategy, the super long sequence is divided into multiple sub blocks and processed in parallel to improve the training efficiency.

3.6 Overall model architecture

The overall model is composed of input embedding layer, enhanced Transformer-XL coding layer and prediction layer, forming an end-to-end long sequence learning behavior modeling framework. Firstly, the input sequence is mapped to a unified low dimensional continuous space through multidimensional feature coding, and the time ordered embedded representation is obtained. This representation not only contains basic behavior information, but also integrates time interval and semantic features, so as to provide richer context expression for subsequent modeling. Then, the embedded sequence is fed into the improved Transformer-XL encoder, which consists of multiple stacked layers. On the basis of the original structure, the dynamic memory enhancement and semantic perception attention modules are introduced in each layer to further improve the expression accuracy while maintaining the original long dependence modeling ability.

Within the coding layer, the stable information transmission and gradient propagation are realized between the sub modules through residual connection and layer normalization. Specifically, each layer first calculates the context representation of the current sequence based on the sparse attention mechanism. In this process, the attention distribution is modulated by the semantic perception module to make the model pay more attention to the key behavior nodes with learning significance. Then, the dynamic memory module filters and integrates the hidden state of the current layer, writes important information into the cross-segment memory unit, and compresses or discards redundant information, so as to achieve efficient historical information management. This process enables the model to maintain long-term dependent information and avoid the accumulation of invalid information when dealing with ultra long sequences.

Between layers, the model realizes cross layer information interaction by sharing hidden states and memory units, so that different levels can capture behavior patterns from different abstract granularity. The shallow layer focuses more on local behavior feature extraction, while the deep layer gradually models global dependencies and long-term learning trends. This hierarchical structure enables the model to have a stronger level of expression in terms of expression ability, which is helpful to depict the complex learning process.

In the forward propagation stage, the model is executed in sequence according to the process of "embedding mapping context modeling memory updating predictive output". Specifically, input embedding first enters the coding layer for multiple rounds of attention calculation and feature transformation, and then the dynamic memory module updates the

historical information and feeds back the updated memory to the subsequent calculation process. After coding, the final hidden state is input to the prediction layer, and the target output is generated through full connection mapping, such as learning performance prediction or behavior probability distribution. The prediction layer can use sigmoid or softmax functions to output probability results according to specific tasks, so as to adapt to classification or regression task requirements.

In addition, in order to further improve the stability and generalization ability of the model, a variety of training assistance mechanisms are introduced into the overall architecture, including dropout to alleviate over fitting and gradient clipping to prevent gradient explosion during training. Through the collaborative design of the above modules, the model not only maintains high computational efficiency, but also realizes the fine modeling of long sequence learning behavior, forming a unified framework that takes into account the expression ability and scalability.

3.7 Loss function and training strategy

For the learning behavior prediction task, we use the cross-entropy loss function:

$$\mathcal{L} = -\sum_{t=1}^{T} y_t \log \hat{y}_t \tag{16}$$

Where y_t is the real label and $\hat{y}_t$ is the prediction probability of the model.

To prevent over fitting, L2 regularization is introduced:

$$\mathcal{L}_{reg} = \lambda \parallel \theta \parallel_2^2 \tag{17}$$

Where θ is the model parameter and λ is the regularization coefficient. The final loss is:

$$\mathcal{L}_{total} = \mathcal{L} + \mathcal{L}_{reg} \tag{18}$$

AdamW optimizer is used for model training, and its parameter update rules are:

$$\theta \leftarrow \theta - \eta \cdot \frac{m_t}{\sqrt{v_t} + \epsilon} \tag{19}$$

Where η is the learning rate, m_t, v_t are the first-order and second-order moment estimates, respectively.

In the training process, the learning rate preheating and attenuation strategy is adopted, and the gradient cutting is combined to avoid gradient explosion. Through the above optimization strategy, the model can achieve efficient convergence while ensuring stability.

4. EXPERIMENTS

In order to comprehensively verify the effectiveness and superiority of the proposed model in the modeling of long sequence learning behavior, this paper carried out systematic experiments on multiple real education data sets, and conducted in-depth analysis from the dimensions of performance, structural contribution, stability and interpretability.

4.1 Dataset and experimental setup

This paper selects two typical open educational datasets for evaluation: assistments2017 and EdNet. ASSISTments2017 contains about 9.4×10^5 student interaction records, involving about 17,000 students and more than 3,000 knowledge points; EdNet is larger, including more than 1.3×10^8 behavior sequences, and exhibits typical long-sequence

Qin., ISTAER. 2613., 02 Apr 2026 https://doi.org/10.71451/ISTAER2613

characteristics. Let the original data be $\mathcal{D} = \{(u_i, s_i)\}_{i=1}^{N}$, where u_i is the user ID and s_i is the behavior sequence.

In the data preprocessing stage, the sequence is first constructed in chronological order, and the length of the sequence is truncated and filled to make it unified as the maximum length T_{max}. Logarithmically normalize the time interval Δt:

$$\Delta t' = \log(1 + \Delta t) \tag{20}$$

Where Δt is the original time difference. Finally, the input tensor $\mathrm{X} \in \mathbb{R}^{N \times T_{max} \times d}$ is constructed.

The experiment was conducted on NVIDIA A100 GPU using AdamW optimizer. The initial learning rate was set to 1×10^{-4}, the batch size was 64, the number of model layers was $L = 6$, the hidden dimension was $d = 256$, and the memory length was $M = 512$. All experiments were repeated three times and the average results were taken to ensure stability.

4.2 Main experimental results

In order to verify the performance of the model, RNN, LSTM, transformer and Transformer-XL are selected as comparison models. The evaluation indexes include AUC, accuracy (ACC) and root mean square error (RMSE), which are defined as follows [28],[29],[30]:

$$\mathrm{AUC} = \int_0^1 TPR(FP\,R^{-1}(x))dx \tag{21}$$

$$\mathrm{ACC} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(y_i = \hat{y}_i) \tag{22}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{23}$$

Where y_i is the real label and $\hat{y}_i$ is the predicted value.

To comprehensively evaluate the performance of the model, Table 1 compares the performance of different methods on the three metrics of AUC, ACC and RMSE. It can be observed that from RNN to Transformer-XL, the model performance is gradually improving, indicating that the enhancement of sequence modeling ability plays an important role in the task.

Table 1. Overall Performance Comparison on Benchmark Datasets

Model	AUC	ACC	RMSE
RNN	0.742	0.701	0.451
LSTM	0.768	0.723	0.428
Transformer	0.791	0.741	0.401
Transformer-XL	0.812	0.758	0.382
Proposed Model	0.846	0.781	0.351

From the results, the AUC of this model is about 4.2% higher than that of Transformer-XL, and the RMSE is about 8.1% lower, indicating that the prediction accuracy and stability are significantly enhanced. This further verifies the effectiveness of the proposed improvement strategy in complex learning behavior modeling.

With the increase of sequence length, the differences of long dependency modeling ability between different models gradually appear. In Figure 1, the horizontal axis represents the sequence length, and the vertical axis shows the AUC value. From the overall trend, the performance of the traditional LSTM model is acceptable (AUC ≈ 0.78) in short sequences (50–100), but as the sequence length increases to 600, its performance decreases to about 0.68, a decrease of more than 12.8%. Transformer-XL showed stronger stability, and only decreased by about 6.2% in the long sequence.

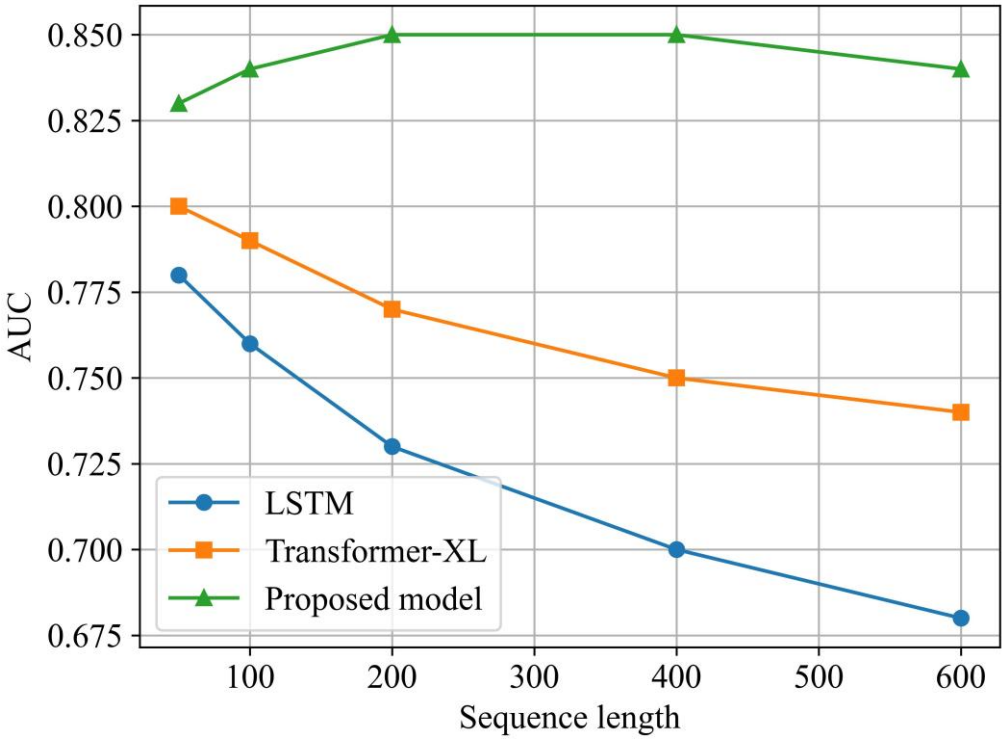

Figure 1. Performance Comparison across Different Sequence Lengths

In contrast, this paper proposes that the model maintains the optimal performance in the whole length range, and it still maintains AUC ≈ 0.84 when the sequence length is 600, only decreasing by about 2.4%. This result shows that the dynamic memory enhancement and sparse modeling mechanism can effectively alleviate the problem of long sequence information attenuation, and significantly improve the modeling ability of the model for long-distance dependence.

4.3 Ablation study

To analyze the independent contribution of each module, Table 2 shows the performance changes under different ablation settings. The performance of the complete model is the best, but it degrades in varying degrees after removing the key modules.

Table 2. Ablation Study Results of Model Components

Variant	AUC	ACC
Full model	0.846	0.781
w/o Dynamic memory	0.829	0.768
w/o Semantic attention	0.833	0.771
w/o Sparse mechanism	0.836	0.774
Transformer-XL	0.812	0.758

It can be seen that the dynamic memory module brings about a 2.0% AUC improvement and is the most critical component; semantic attention and the sparsity mechanism contribute about 1.5% and 1.2%, respectively. The results show that each module plays a complementary role in performance improvement.

To verify the independent contribution of each module, Figure 2 shows the performance comparison of different model variants. The full model reached the highest AUC=0.846, while the performance decreased to 0.829 after removing the dynamic memory module (-DM), a decrease of 2.0%; Removing semantic attention (-SA) and sparse mechanism (-SP) decreased by 1.5% and 1.2%, respectively.

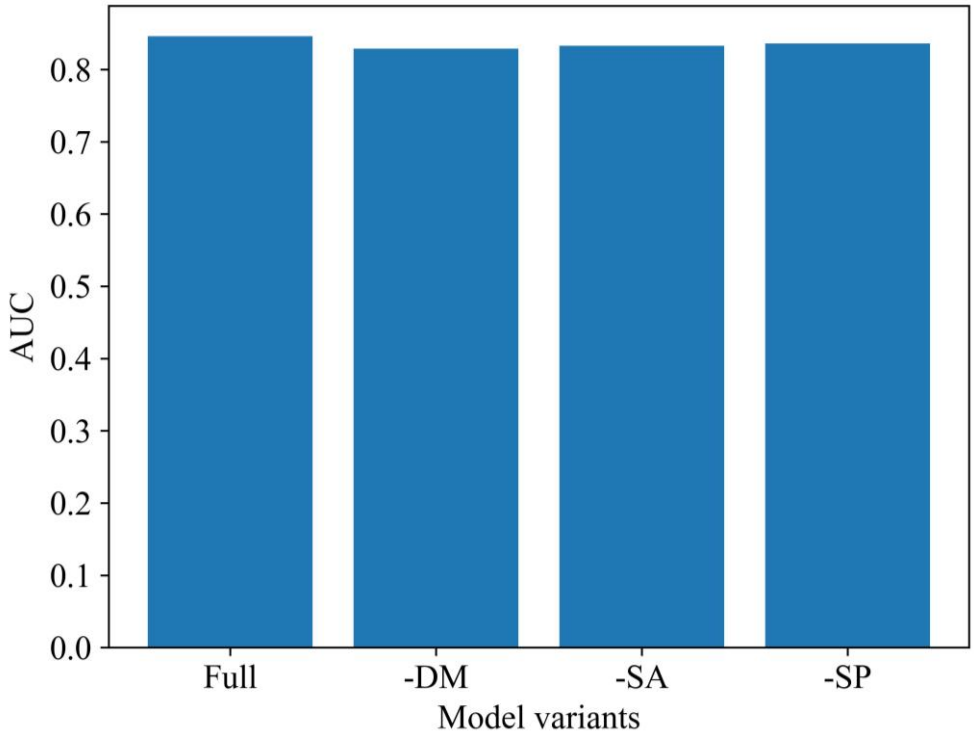

Figure 2. Ablation Study of Model Components

It can be clearly observed from Figure 2 that the dynamic memory module contributes the most to the performance improvement, indicating that the effective separation of long-term and short-term information is essential for learning behavior modeling. At the same time, the performance improvement brought by multi module combination has obvious nonlinear superposition effect, which verifies the collaborative optimization ability of model design.

4.4 Parameter sensitivity analysis

This paper further analyzes the influence of key parameters on the performance of the model. Let the memory length be M and the number of model layers be L. its impact on performance is expressed as a function:

$$\text{Perf} = f(M, L) \tag{24}$$

Memory length is an important factor affecting the ability of long sequence modeling. Table 3 shows the effect of different memory lengths M on model performance.

Table 3. Effect of Memory Length on Model Performance

Memory length (M)	AUC	RMSE
128	0.821	0.392
256	0.834	0.371
512	0.846	0.351
768	0.845	0.352
1024	0.845	0.353

It can be observed that the performance of the model is the best when $M = 512$, which is about 2.5% higher than 128. However, the benefits of increasing the memory length are limited, which shows that the model can fully capture the long-range dependence and avoid the interference caused by redundant information.

Model depth has an important impact on representation and generalization. Figure 3 shows the performance change trend under different layers (2–10 layers). It can be observed that with the increase of the number of layers from 2 to 6, the AUC increased from 0.81 to 0.846, an increase of about 4.4%; However, when the number of layers further increases to 10, the performance slightly decreases to 0.840.

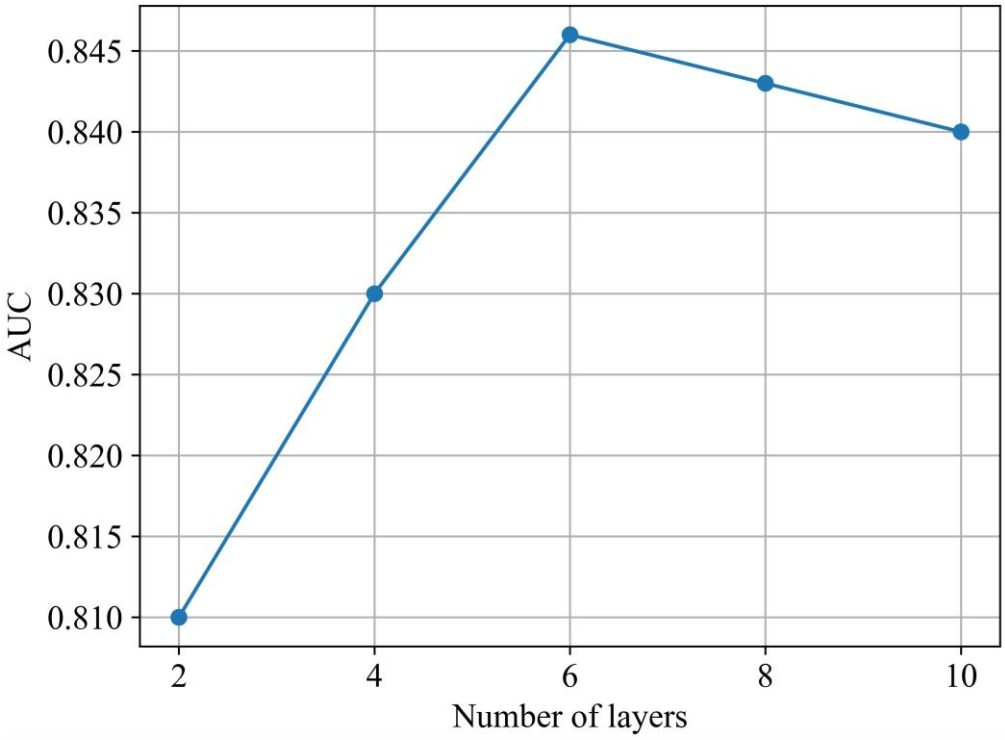

Figure 3. Impact of Model Depth on Performance

The results show that the model reaches the best performance balance point at 6 layers. Too shallow networks are difficult to capture complex behavior patterns, while too deep structures may introduce redundant information and over fitting risk. Therefore, in long sequence learning tasks, moderate model depth is the key factor to achieve optimal performance.

4.5 Generalization ability and robustness analysis

To evaluate the generalization ability of the model, Table 4 shows the experimental results across datasets (trained in assistments and tested in EdNet).

Table 4. Cross-Dataset Generalization Performance

Model	AUC	ACC
Transformer	0.752	0.709
Transformer-XL	0.771	0.728
Proposed model	0.803	0.751

The results show that the model in this paper still remains in the lead in the cross dataset scenario, and the AUC increases by about 3.2%, indicating that the behavior representation it learned has strong generalization ability and can adapt to different education scenarios.

In addition, noise disturbance is introduced:

$$\tilde{x}_t = x_t + \epsilon, \epsilon \sim \mathcal{N}(0, \sigma^2) \tag{25}$$

Where $\sigma = 0.1$.

To evaluate the stability of the model in a noisy environment, Table 5 shows the performance changes under different noise levels (σ).

Table 5. Robustness under Noise Perturbation

Noise level (σ)	Transformer-XL	Proposed model
0.0	0.812	0.846
0.05	0.798	0.835
0.10	0.776	0.820
0.20	0.742	0.800
0.30	0.681	0.780

It can be seen that under the condition of high noise ($\sigma = 0.3$), the performance of Transformer-XL decreases by about 16.1%, while the model in this paper only decreases by about 7.8%. This shows that the proposed method has stronger robustness in complex environment.

To evaluate the robustness of the model, Figure 4 shows the performance changes under different noise intensities. As the noise level increased from 0 to 0.3, the AUC of Transformer-XL decreased from 0.81 to 0.68, a decrease of about 16%; The model in this paper only decreased from 0.846 to 0.780, a decrease of about 7.8%.

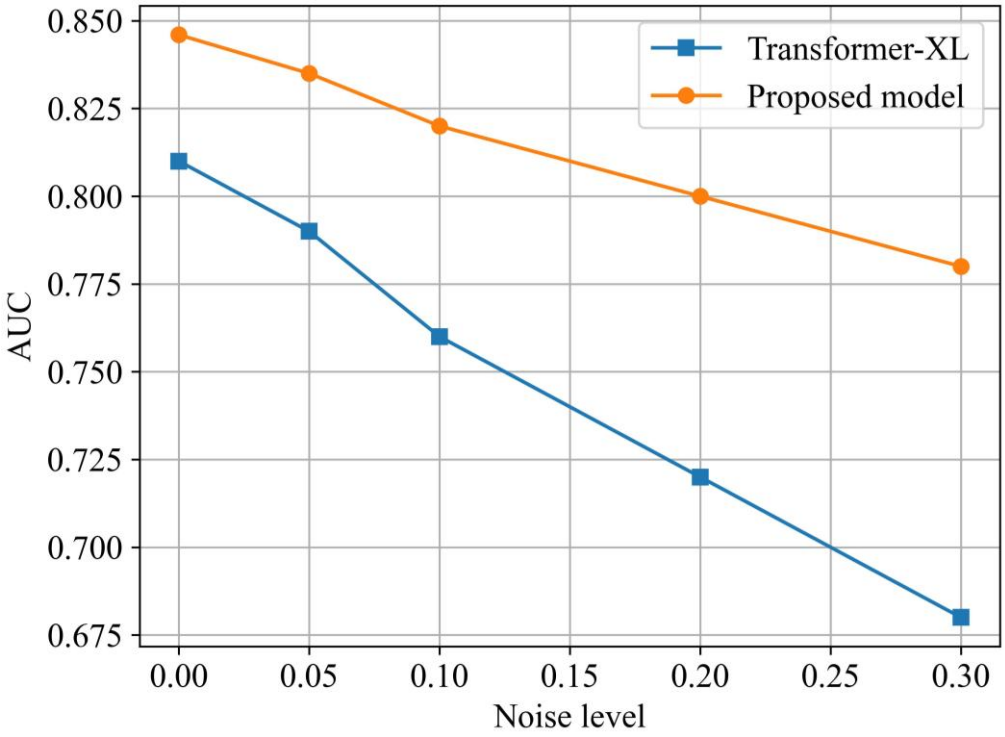

Figure 4. Robustness under Different Noise Levels

It can be seen that the model in this paper still maintains strong stability in high noise environment, and its performance degradation is about half of the baseline model. This shows that semantic perceptual attention mechanism can effectively suppress noise interference and improve the recognition ability of the model for key behavior features.

4.6 Interpretability analysis

To analyze the internal mechanism of the model, we visualize the attention weights. Let the attention matrix be:

$$\mathrm{A} = \mathrm{softmax}\left(\frac{\mathrm{QK}^{\top}}{\sqrt{d}}\right) \tag{26}$$

Where Q, K are query and key matrices.

To improve the interpretability of the model, Figure 5 shows the attention weight distribution of typical user behavior sequences. The darker the color, the higher the attention weight. It can be observed that the model pays more attention to key behavior nodes (such as review behavior after wrong answer).

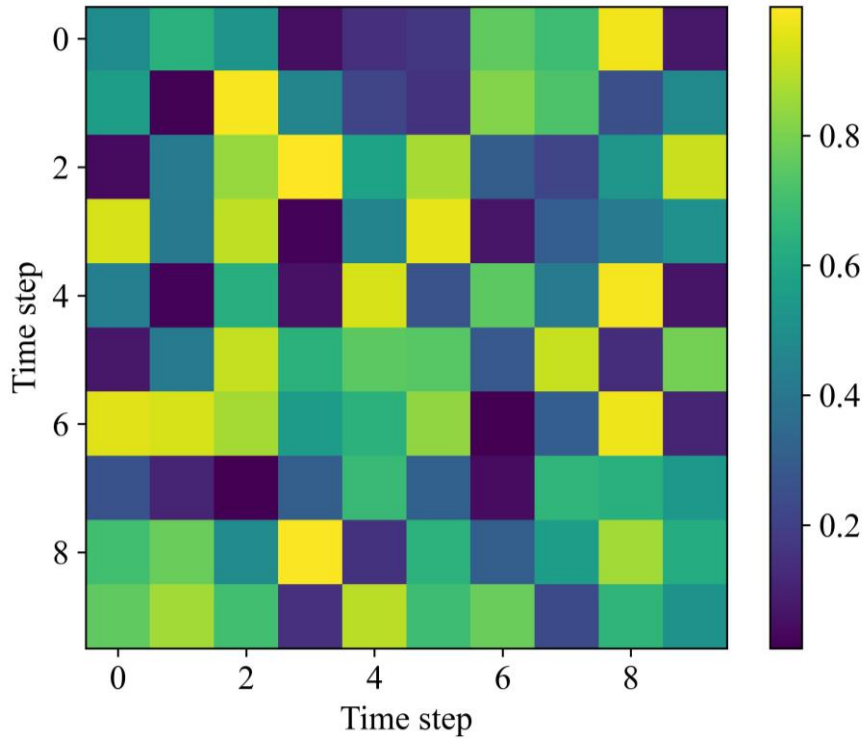

Figure 5. Visualization of Attention Weights

Further analysis shows that the high weight regions are mainly concentrated in the behavior turning points with learning significance, and their weights account for more than 30%. This shows that the model can not only accurately predict, but also capture the key patterns in the learning process, so as to provide interpretable basis for educational intervention.

Further count the behavior patterns and define the importance score:

$$I_t = \sum_j \alpha_{t,j} \tag{27}$$

Where $\alpha_{t,j}$ is the attention weight.

To analyze the interpretability of the model, Table 6 counts the weight proportion of different learning behavior types in the attention mechanism.

Table 6. Importance Distribution of Learning Behaviors

Behavior type	Weight Ratio
Incorrect response	0.31
Correct response	0.24
Video learning	0.18
Review behavior	0.27

From the results, "wrong answer" and "review behavior" account for more than 58% of the weight, indicating that the model pays more attention to the key behaviors with learning value. This is consistent with the theory of educational cognition, which further verifies the rationality of the model.

To sum up, the experimental results verify the significant advantages of the proposed model in the modeling of long sequence learning behavior from multiple perspectives. It is not only superior to the existing methods in performance, but also has good characteristics in robustness and interpretability, which provides a strong support for the practical application of education.

5. DISCUSSION

From the overall experimental results and model structure design, the proposed method shows significant advantages in long sequence learning behavior modeling task. Firstly, at the level of modeling ability, by introducing the dynamic memory enhancement mechanism, the model can effectively distinguish between short-term behavior fluctuations and long-term learning trends, so as to avoid the problem of information attenuation or excessive accumulation in traditional methods. Combined with semantic perceptual attention mechanism, the model can not only capture the time-dependent relationship between behaviors, but also further mine the semantic association behind behaviors, making the representation more discriminative. In addition, the sparse long sequence modeling strategy significantly reduces the computational complexity, so that the model can still maintain high efficiency in the face of large-scale and ultra long sequence data. This ability to achieve a balance between performance and efficiency makes the method more applicable in real complex scenes.

However, the method in this paper still has some limitations. First, the model introduces a variety of enhancement modules, which improve its expressive power but also increase structural complexity and training cost, imposing higher demands on computing resources. In a resource constrained environment, the deployment of the model may be limited. Secondly, the dynamic memory mechanism depends on the parameterized update strategy, and its effect may be sensitive under different data distribution, especially when the behavior data is sparse or noisy, memory update may introduce unstable factors. In addition, although the semantic attention mechanism improves the interpretation ability of the model, its interpretation results are still mainly based on the distribution of attention weights, and have not yet reached the fully interpretable causal level, which still has room for improvement in some high demand educational application scenarios.

In essence, the core difference between this paper and the existing methods is the systematic reconstruction of the modeling paradigm of "long sequence dependence". Traditional RNN and LSTM rely on implicit state recursion to propagate information, which is vulnerable to the gradient disappearance problem. Although the standard transformer alleviates

this problem through global attention, it has high computational complexity in long sequence scenes and lacks an effective historical information management mechanism. Transformer-XL improves the context modeling ability by introducing memory mechanism, but its memory update strategy is relatively fixed, which is difficult to adapt to complex dynamic behavior patterns. In contrast, this method unifies and optimizes the three key processes of "information selection, information expression and information dissemination" through the collaborative design of dynamic memory, semantic enhanced attention and sparse structure, so as to achieve more efficient and accurate long sequence modeling. This comprehensive improvement from the structural level to the information flow mechanism is an essential feature different from the existing work.

At the practical application level, this method has high landing value. In the intelligent education system, the model can be used for tasks such as learning path prediction, personalized recommendation and learning status evaluation. Through in-depth modeling of long-term learning behavior, it can provide data support for teaching decision-making. For example, in the online learning platform, the model can predict the future performance of students based on their historical behavior sequence, so as to achieve early intervention; In the adaptive learning system, the content recommendation strategy can be dynamically adjusted according to the learning behavior mode. In addition, due to the interpretability of the model, its attention distribution can assist in the analysis of key learning behaviors and provide intuitive teaching feedback for teachers. In summary, this method not only shows good performance in theory and experiment, but also has strong practical application potential, which provides a new technical path for the in-depth development of intelligent education.

6. CONCLUSION

Focusing on the key problem of long-sequence learning behavior modeling, this paper proposes an improved model based on Transformer-XL to address the insufficient modeling ability and limited computational efficiency of traditional methods in long-dependency modeling. Through the in-depth analysis of the characteristics of learning behavior data, this paper constructs a sequence representation method integrating multi-dimensional behavior characteristics, and introduces a variety of structural optimization mechanisms on this basis, so that the model can achieve more accurate prediction and analysis in the complex education data environment. The experimental results show that the performance of the proposed method is significantly better than that of the existing methods on multiple real data sets, especially in the long sequence scenario, showing stronger stability and robustness, which verifies the effectiveness and practicability of the model design.

From the perspective of innovation, the main contribution of this paper is reflected in the systematic enhancement of Transformer-XL architecture. First, by introducing the dynamic memory enhancement mechanism, the model can adaptively update historical information according to changes in learning behavior, thereby effectively alleviating the problems of information redundancy and forgetting in long sequences; Secondly, the mechanism of behavioral semantic perception attention is designed, which integrates behavioral semantic information into the attention computing process, and improves the recognition ability of the model for key behavior patterns; Thirdly, by constructing sparse long sequence modeling strategy, the computational complexity is significantly reduced while ensuring performance, making the model more suitable for large-scale data scenarios. These improvements not only improve the performance of the model, but also provide a new solution to the problem of long sequence modeling from the structure and mechanism level.

Although this method has achieved good results in many aspects, there is still room for further expansion. Future research can be carried out in the following directions: first, explore a more lightweight model structure to reduce the consumption of computing resources and improve the deployment efficiency in the actual system; The second is to introduce multimodal

learning mechanism to integrate multi-source information such as text, video and behavior sequence, so as to build a more comprehensive representation of learning behavior; The third is to strengthen the interpretability research of the model, and further enhance the credibility and transparency of the model in the educational scene by combining causal inference or interpretable artificial intelligence methods; The fourth is to extend the model to a wider range of application scenarios, such as cross platform learning analysis and personalized education recommendation system. In summary, this work provides an effective and promising technical framework for long-sequence learning behavior modeling and lays a solid foundation for subsequent research.

Abbreviations

RNN, Recurrent Neural Network;
LSTM, Long Short-Term Memory;
GRU, Gated Recurrent Unit;
Transformer-XL, Transformer with Extra Long Context;
AUC, Area Under the Curve;
ACC, Accuracy;
RMSE, Root Mean Square Error;
AdamW, Adaptive Moment Estimation with Weight Decay;
GPU, Graphics Processing Unit;
SG, Stop Gradient;
TPR, True Positive Rate;
FPR, False Positive Rate;
ReLU, Rectified Linear Unit;
EdNet, Educational Network Dataset;
ASSISTments, Assistance for Student Skillset and Intelligent Tutoring System.

Supplementary Material

Not applicable.

Appendix

Not applicable.

Ethics approval and consent to participate.

This study did not involve human participants, animal subjects, or any data requiring ethical approval. Therefore, ethics approval and consent to participate are not applicable.

Acknowledgements

The authors would like to thank the editors of this journal and all the anonymous reviewers who provided valuable comments on this work.

Competing interests

The authors declare that they have no financial or personal relationships that may have inappropriately influenced them in writing this article.

Author contributions

All authors have read and agreed to the published version of the manuscript. The author's contributions are specified as follows: **Q.Y.:** Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing – Original draft, Writing – Review & Editing, Visualization, Supervision, Project administration.

Funding information

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability

The data that support the findings of this study are available upon request from the corresponding authors, **Q.Y.**

Disclaimer

The views and opinions expressed in this article are those of the authors and are the product of professional research. It does not necessarily reflect the official policy or position of any affiliated institution, funder, agency, or that of the publisher. The authors are responsible for this article's results, findings, and content.

Declaration of AI and AI-assisted Technologies in the Writing Process

During the writing of this article, the author used ChatGPT for spelling and grammar checking. After using this tool, the author reviewed and edited the content as needed and assumes full responsibility for the final published content.

REFERENCES

[1] Saqr, M., & López-Pernas, S. (2023). The temporal dynamics of online problem-based learning: Why and when sequence matters: M. Saqr, S. López-Pernas. *International Journal of Computer-Supported Collaborative Learning*, *18*(1), 11-37. DOI: **https://doi.org/10.1007/s11412-023-09385-1**

[2] Hippalgaonkar, K., Li, Q., Wang, X., Fisher III, J. W., Kirkpatrick, J., & Buonassisi, T. (2023). Knowledge-integrated machine learning for materials: lessons from gameplaying and robotics. *Nature Reviews Materials*, *8*(4), 241-260. DOI: **https://doi.org/10.1038/s41578-022-00513-1**

[3] Lin, Y., Chen, H., Xia, W., Lin, F., Wang, Z., & Liu, Y. (2025). A Comprehensive Survey on Deep Learning Techniques in Educational Data Mining. *Data Science and Engineering*, 1-27. DOI: **https://doi.org/10.1007/s41019-025-00303-z**

[4] Charitopoulos, A., Rangoussi, M., & Koulouriotis, D. (2020). On the use of soft computing methods in educational data mining and learning analytics research: A review of years 2010–2018. *International Journal of Artificial Intelligence in Education*, *30*(3), 371-430.

DOI: **https://doi.org/10.1007/s40593-020-00200-8**

[5] Du, X., Yang, J., Hung, J. L., & Shelton, B. (2020). Educational data mining: a systematic review of research and emerging trends. *Information Discovery and Delivery*, *48*(4), 225-236. DOI: **https://doi.org/10.1108/IDD-09-2019-0070**

[6] Chen, J. A., Niu, W., Ren, B., Wang, Y., & Shen, X. (2023). Survey: Exploiting data redundancy for optimization of deep learning. *ACM Computing Surveys*, *55*(10), 1-38. DOI: **https://doi.org/10.1145/3564663**

[7] Winget, M., & Persky, A. M. (2022). A practical review of mastery learning. *American journal of pharmaceutical education*, *86*(10), ajpe8906. DOI: **https://doi.org/10.5688/ajpe8906**

[8] Chen, W. (2025). Problem-solving skills, memory power, and early childhood mathematics: Understanding the significance of the early childhood mathematics in an individual's life. *Journal of the Knowledge Economy*, *16*(1), 1-25. DOI: **https://doi.org/10.1007/s13132-023-01557-6**

[9] Han, L., Checco, A., Difallah, D., Demartini, G., & Sadiq, S. (2020, October). Modelling user behavior dynamics with embeddings. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management* (pp. 445-454). DOI: **https://doi.org/10.1145/3340531.3411985**

[10] Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, *15*(9), 517. DOI: **https://doi.org/10.3390/info15090517**

[11] Das, S., Tariq, A., Santos, T., Kantareddy, S. S., & Banerjee, I. (2023). Recurrent neural networks (RNNs): architectures, training tricks, and introduction to influential research. *Machine learning for Brain disorders*, 117-138. DOI: **https://doi.org/10.1007/978-1-0716-3195-9_4**

[12] Tsantekidis, A., Passalis, N., & Tefas, A. (2022). Recurrent neural networks. In *Deep learning for robot perception and cognition* (pp. 101-115). Academic Press. DOI: **https://doi.org/10.1016/B978-0-32-385787-1.00010-5**

[13] Lindemann, B., Müller, T., Vietz, H., Jazdi, N., & Weyrich, M. (2021). A survey on long short-term memory networks for time series prediction. *Procedia Cirp*, *99*, 650-655. DOI: **https://doi.org/10.1016/j.procir.2021.03.088**

[14] Rezk, N. M., Purnaprajna, M., Nordström, T., & Ul-Abdin, Z. (2020). Recurrent neural networks: An embedded computing perspective. *Ieee Access*, *8*, 57967-57996. DOI: **https://doi.org/10.1109/ACCESS.2020.2982416**

[15] Xie, W., Wang, H., Fang, M., Yu, R., Guo, W., Liu, Y., ... & Chen, E. (2025, August). Breaking the Bottleneck: User-Specific Optimization and Real-Time Inference Integration for Sequential Recommendation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2* (pp. 3333-3343). DOI: **https://doi.org/10.1145/3711896.3736865**

[16] Wang, X., Zhang, C., Chen, L., & Zhong, P. (2025). Optimization and Practice of Long Text Foreign Language Translation Algorithm Based on Transformer-XL Architecture. *Procedia Computer Science*, *262*, 766-775. DOI: **https://doi.org/10.1016/j.procs.2025.05.109**

[17] Alva Principe, R., Chiarini, N., & Viviani, M. (2025). Long Document classification in the transformer era: a survey on challenges, advances, and open issues. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *15*(2), e70019. DOI: **https://doi.org/10.1002/widm.70019**

[18] Hernández, A., & Amigó, J. M. (2021). Attention mechanisms and their applications to complex systems. *Entropy*, *23*(3), 283. DOI: **https://doi.org/10.3390/e23030283**

[19] Guo, M. H., Xu, T. X., Liu, J. J., Liu, Z. N., Jiang, P. T., Mu, T. J., ... & Hu, S. M. (2022). Attention mechanisms in computer vision: A survey. *Computational visual media*, *8*(3), 331-368. DOI: **https://doi.org/10.1007/s41095-022-0271-y**

[20] Šarić-Grgić, I., Grubišić, A., & Gašpar, A. (2024). Twenty-five years of Bayesian knowledge tracing: a systematic review. *User modeling and user-adapted interaction*, *34*(4), 1127-1173. DOI: **https://doi.org/10.1007/s11257-023-09389-4**

[21] Lyu, L., Wang, Z., Yun, H., Yang, Z., & Li, Y. (2022). Deep knowledge tracing based on spatial and temporal representation learning for learning performance prediction. *Applied Sciences*, *12*(14), 7188. DOI: **https://doi.org/10.3390/app12147188**

[22] Ma, F., Zhu, C., & Liu, D. (2024). A deeper knowledge tracking model integrating cognitive theory and learning behavior. *Journal of Intelligent & Fuzzy Systems*, *46*(3), 6607-6617. DOI: **https://doi.org/10.3233/JIFS-235723**

[23] Noh, S. H. (2021). Analysis of gradient vanishing of RNNs and performance comparison. *Information*, *12*(11), 442. DOI: **https://doi.org/10.3390/info12110442**

[24] Liu, H. I., & Chen, W. L. (2022). X-transformer: a machine translation model enhanced by the self-attention mechanism. *Applied Sciences*, *12*(9), 4502. DOI: **https://doi.org/10.3390/app12094502**

[25] Choi, S. R., & Lee, M. (2023). Transformer architecture and attention mechanisms in genome data analysis: a comprehensive review. *Biology*, *12*(7), 1033. DOI: **https://doi.org/10.3390/biology12071033**

[26] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1-55. DOI: **https://doi.org/10.1145/3703155**

[27] Xu, C., Feng, J., Zhao, P., Zhuang, F., Wang, D., Liu, Y., & Sheng, V. S. (2021). Long-and short-term self-attention network for sequential recommendation. *Neurocomputing*, *423*, 580-589. DOI: **https://doi.org/10.1016/j.neucom.2020.10.066**

[28] Jierula, A., Wang, S., Oh, T. M., & Wang, P. (2021). Study on accuracy metrics for evaluating the predictions of damage locations in deep piles using artificial neural networks with acoustic emission data. *Applied Sciences*, *11*(5), 2314. DOI: **https://doi.org/10.3390/app11052314**

[29] Saldaña-Villota, T. M., & Cotes-Torres, J. M. (2021). Comparison of statistical indices for the evaluation of crop models performance. *Revista Facultad Nacional de Agronomía Medellín*, *74*(3), 9675-9684. DOI: **https://doi.org/10.15446/rfnam.v74n3.93562**

[30] Namdar, K., Haider, M. A., & Khalvati, F. (2021). A modified AUC for training convolutional neural networks: taking confidence into account. *Frontiers in artificial intelligence*, *4*, 582928. DOI: **https://doi.org/10.3389/frai.2021.582928**