PLOS One

RESEARCH ARTICLE

TGEL-transformer: Fusing educational theories with deep learning for interpretable student performance prediction

Yuhao Gong[1], **Fei Wang**[2], **Yuchen Zhang**[3]*, **Jiaqi Geng**[4]

1 Nanchang Hangkong University, Nanchang, Jiangxi, China, **2** Nanchang Institute of Science and Technology, Nanchang, Jiangxi, China, **3** Inti International University, Nilai, Negeri Sembilan, Malaysia, **4** Nanchang Institute of Science and Technology, Nanchang, Jiangxi, China

* i24026647@student.newinti.edu.my

Citation: Gong Y, Wang F, Zhang Y, Geng J (2025) TGEL-transformer: Fusing educational theories with deep learning for interpretable student performance prediction. PLoS One 20(6): e0327481. https://doi.org/10.1371/journal.pone.0327481

Editor: Issa Atoum, Philadelphia University, JORDAN

Received: April 22, 2025

Accepted: June 16, 2025

Published: June 30, 2025

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0327481

Abstract

With the integration of educational technology and artificial intelligence, personalized learning has become increasingly important. However, traditional educational data mining methods struggle to effectively integrate heterogeneous feature data and represent complex learning interaction processes, while existing deep learning models lack educational theory guidance, resulting in insufficient interpretability. To address these challenges, this study proposes the TGEL-Transformer (Theory-Guided Educational Learning Transformer) framework, which integrates multiple intelligence theory and social cognitive theory, featuring three innovations: a dual-channel feature processing module that integrates cognitive, affective, and environmental dimension features; a theory-guided four-head attention mechanism that models educational interaction dynamics; and an interpretable prediction layer that provides theoretical support for educational interventions. Using a dataset of 6,608 students, TGEL-Transformer achieved RMSE = 1.87 and $R^2 = 0.75$, outperforming existing methods with statistically significant improvements ($p < 0.001$) ranging from 1.1% against recent state-of-the-art models to 5.6% against transformer baselines. External validation on cross-cultural data (n = 480) demonstrated strong generalizability with $R^2 = 0.683$. Attention weight analysis revealed that teacher support (0.15), prior knowledge (0.15), and peer interaction (0.13) are key factors influencing learning outcomes. This study provides a theory-guided framework for educational data mining, offering data-driven support for personalized education and advancing intelligent education development.

1. Introduction

The rapid development of educational technology and artificial intelligence is profoundly transforming traditional educational models, leading education into a new era

Data availability statement: All data used in this study are available from the Kaggle platform "Student Performance" public dataset. This dataset contains learning performance and related feature data of 6,608 students from real educational institutions. For research reproducibility, we have open-sourced the complete experimental code, data preprocessing workflow, and model implementation on GitHub (https://github.com/litlelight/AiEducation01), where researchers can access all relevant resources for verification and extension of our findings.

Funding: The author(s) received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abbreviations: TGEL, Theory-Guided Educational Learning; EDM, Educational Data Mining; RMSE, Root Mean Square Error; MAE, Mean Absolute Error; MAPE, Mean Absolute Percentage Error; DKT, Deep Knowledge Tracing; GNN, Graph Neural Network; CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory; MLP, Multi-Layer Perceptron; SVM, Support Vector Machine; KNN, K-Nearest Neighbors; SES, Socioeconomic Status; CI, Confidence Interval.

centered on data-driven approaches and personalized experiences [1]. The application of artificial intelligence in education has facilitated the realization of precision learning and personalized teaching, enabling educational content to better adapt to learners' unique needs and learning styles [2]. With the proliferation of intelligent tutoring systems and adaptive learning platforms, AI tools can dynamically adjust content, pace, and teaching methods based on real-time data analysis, a transformation particularly evident in the implementation of intelligent tutoring systems and adaptive learning platforms [3]. The evolutionary trajectory of educational data mining technologies—from early statistical analysis to contemporary deep learning methods—not only demonstrates technological advancement but also reflects a deepening understanding of the learning process [4].

However, despite these technological advances, a critical disconnect persists between AI-powered educational predictions and practical classroom applications. The unique challenges facing educational data analysis distinguish it from conventional data mining tasks [5]. Current deep learning models, while achieving impressive accuracy metrics, suffer from fundamental limitations that hinder their real-world impact. First, these models operate as "black boxes," providing predictions without explanatory mechanisms that educators can understand or trust. Although educational data is naturally diverse, encompassing numerical and categorical features that require advanced integration methods [6], existing models process these features without preserving their educational significance. Second, current approaches fail to capture the complex interaction networks within educational systems, such as the interplay between motivation and engagement or teacher support and learning outcomes [7]. When a model predicts student performance, it overlooks these critical relationships that educators need to understand for effective intervention. Third, and most importantly, educational predictions require interpretability to provide clear rationales and practical intervention suggestions—precisely what many high-precision black-box models lack [8]. These limitations reveal that treating educational prediction as a purely technical problem, while ignoring pedagogical principles, results in models that may be statistically accurate but educationally impractical. This disconnect motivates our research to develop a fundamentally different approach—one that bridges educational theory with deep learning technology to create predictions that are not only accurate but also interpretable and actionable.

Current educational data mining research primarily faces three key limitations [9]. Traditional machine learning methods struggle to effectively process heterogeneous feature data and model complex interactive relationships, particularly showing significant deficiencies in capturing the dynamic processes of educational factors evolving over time [10]. Deep learning methods, while excelling in feature representation, often lack educational theory guidance, resulting in models that predict accurately but fail to provide educationally valuable insights; this gap between predictive ability and practical guidance restricts the widespread application of artificial intelligence in education [11]. A more fundamental issue is the evident disconnect between theory and technology in existing research; most educational prediction model construction and use lack a systematic theoretical framework, limiting the models' ability to capture key features of learning development [12].

To address these challenges, we propose the TGEL-Transformer (Theory-Guided Educational Learning Transformer) framework. As illustrated in Fig 1, our framework systematically operationalizes abstract educational theories into concrete computational components: multiple intelligence theory's cognitive-environmental dichotomy directly maps to our dual-channel feature processing architecture, social cognitive theory's triadic reciprocal causation translates into our four-head attention mechanism, and learning analytics' closed-loop principle guides our interpretable prediction layer design. This theory-to-architecture mapping ensures that each model component is grounded in established pedagogical principles while maintaining computational efficiency.

Our contributions are threefold: (1) enhanced feature representation through theory-guided dual-channel processing, (2) attention mechanism design based on educational interaction dynamics, and (3) interpretable feedback system providing actionable insights. The remainder of this paper is organized as follows: Section 2 reviews related work, Section 3 details our methodology, Section 4 presents experiments, Section 5 discusses results, and Section 6 concludes with future directions.

2. Related work

2.1 Review of educational data mining method development

The field of Educational Data Mining (EDM) has experienced significant development over the past decade, evolving from simple statistical analysis to complex AI-driven models. Du et al. [13] provided a systematic review that not only reported research trends in EDM but also discussed open questions guiding future research, offering a comprehensive perspective on the field. Regarding data types, Ahmad et al. [4] noted that educational data analysis has expanded from single structured data to multimodal data analysis, including the integrated analysis of learning behavior logs, textual feedback, and interaction records, providing richer information for a comprehensive understanding of the learning process.

In predictive modeling, He et al. [14] developed a multi-task learning knowledge tracing model integrating fine-grained attention, which demonstrated excellent performance in capturing long-term dependencies in learning trajectories, significantly improving prediction accuracy through self-attention mechanisms. However, as Qiu et al. [15] pointed out, current educational data mining research still exhibits notable limitations: excessive focus on prediction accuracy while neglecting educational theory support, lack of model interpretability, and evaluation standards that often emphasize technical metrics rather than educational effectiveness. These issues highlight the necessity of combining educational theories with technical methods, providing research space for the theory-based model design in this study.

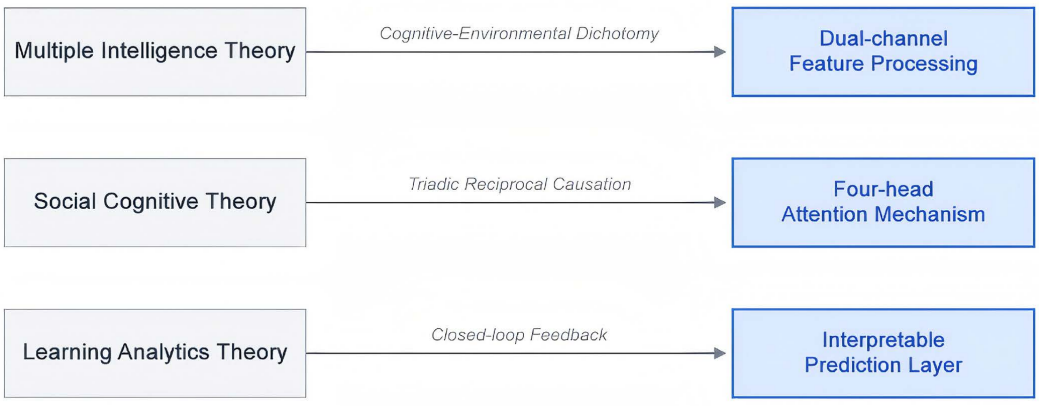

Fig 1. Mapping between educational theories.

https://doi.org/10.1371/journal.pone.0327481.g001

2.2 Applications of deep learning in education

Deep learning technologies, particularly Transformer architectures, are revolutionizing the field of educational data analysis. A systematic review by Farina et al. [16] indicated that Transformer applications in education primarily rely on sparsity design in the model, which is crucial for processing large-scale educational data. In intelligent tutoring, Sun et al. [17] proposed the DASKT dynamic affect simulation knowledge tracing method, which significantly outperformed traditional DIMKT models through self-attention mechanisms, offering new approaches for capturing the relationship between learners' emotional states and knowledge mastery.

In the field of learning behavior analysis, Chen and Wei et al. [18] developed a student performance prediction model based on self-attention mechanisms that effectively integrated multiple behavioral features, providing support for learning assessment in offline educational environments. However, as Ivanashko et al. [5] pointed out, deep learning models face challenges in educational practice, including high deployment costs, difficulties in data privacy protection, and lack of interpretability, highlighting the necessity of customizing deep learning architectures for educational scenarios.

2.3 Educational application innovations in transformer architecture

The Transformer architecture, with its powerful sequence modeling and attention mechanisms, demonstrates unique advantages in educational applications. Pu et al. [19] applied the Transformer model to student answer sequence analysis, designing a deep knowledge tracing Transformer that captures long-term dependencies in the learning process through positional encoding and multi-head attention mechanisms, significantly outperforming traditional knowledge tracing models. Liu et al. [20] further developed a Transformer knowledge tracing model integrating forgetting mechanisms combined with convolutional mechanisms, effectively addressing the continuous repetition training problem in educational data and enhancing the model's ability to perceive potential connections between knowledge points.

In personalized learning resource recommendation, Wu et al. [21] developed the SSE-PT personalized Transformer sequence recommendation model, providing theoretical support for adaptive learning systems by focusing on personalized parameter optimization to improve recommendation effectiveness. Personalized teaching interventions and course recommendations have become important directions for Transformer applications, showing particular advantages in scenarios combining learner preferences with learning objectives.

2.4 Integration of educational theories and deep learning

The organic integration of educational theories with deep learning technologies represents the frontier development direction in educational data mining. Chai et al. [22] developed a structural equation model explaining students' intentions to learn AI based on the theory of planned behavior and self-determination theory, revealing the relationships between learning resource design, autonomy, and AI social benefits with learning willingness. Chen et al. [18] applied self-attention mechanisms to student performance prediction models, effectively improving prediction accuracy by integrating student behavioral combination features, providing strong support for learning assessment in offline educational environments.

In multimodal learning analytics, recent research such as the work by Liu et al. [20] demonstrates that combining cognitive theory with deep learning models can create more effective analytical frameworks, particularly in capturing the dynamic relationship between learners' emotional states and behaviors. However, theoretical integration in existing models often remains at the conceptual level, lacking systematic methods to transform theoretical concepts into specific model components. These limitations establish the theoretical foundation and innovation space for the TGEL-Transformer framework proposed in this research.

3 Methodology

3.1 Problem statement

This research focuses on student performance prediction in the educational data mining field, using multidimensional heterogeneous educational data to predict student learning outcomes. Given a dataset $D = \{(X_i, y_i)\}_{i=1}^{n}$ comprising n students, where input features X_i include cognitive dimensions (learning time, attendance, etc.), affective dimensions (learning motivation, engagement, etc.), and environmental dimensions (school type, teacher support, etc.), and the output target y_i represents students' final learning outcomes (0–100 points). Current educational data mining research faces three major challenges: difficulty integrating heterogeneous features, insufficient modeling of complex interactions, and lack of theoretical guidance and interpretability, which limits the practical value of prediction models in educational settings. This motivates us to construct a new mapping function $f : R^m \to R$, with the optimization objective

$$\min_f \mathcal{L}(f) = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(X_i))^2 + \lambda\Omega(f)$$

To address these challenges, we propose the TGEL-Transformer (Theory-Guided Educational Learning Transformer) framework, which integrates multiple intelligence theory and social cognitive theory through a dual-channel feature processing module for heterogeneous feature integration, a theory-guided four-head attention mechanism for modeling complex interaction relationships, and an interpretable prediction layer to provide theoretical support for educational interventions. Our goal is not only to improve prediction accuracy (reducing RMSE and increasing R^2) but, more importantly, to build an interpretable framework that quantifies the influence weights of different educational factors, providing theoretical basis for precision educational interventions and bridging the gap between prediction results and educational practice to ultimately support personalized learning.

3.2 Model architecture

3.2.1 Overall architecture overview. The TGEL-Transformer model architecture is constructed based on three core theoretical frameworks: multiple intelligence theory, social cognitive theory, and learning analytics theory. As shown in Fig 2, the model is divided into four layers: input layer, feature processing layer, attention encoding layer, and prediction layer. The input layer receives raw educational data; the feature processing layer transforms heterogeneous features into unified representations under the guidance of multiple intelligence theory; the attention encoding layer models dynamic interactions in the learning process through a four-head self-attention mechanism guided by social cognitive theory; and the prediction layer generates final prediction results and provides interpretable intervention suggestions within the learning analytics theory framework. This theory-guided layered design not only enhances the model's predictive capability but also strengthens its educational interpretability.

3.2.2 Theory-guided feature processing module. The feature processing module, guided by multiple intelligence theory, employs a dual-channel architecture to process numerical and categorical educational features. For numerical features reflecting cognitive abilities (such as study time and attendance rate), we apply linear transformation for standardization:

$$z_i = \frac{x_i - \mu}{\sigma}$$

where x_i represents the original feature value, and μ and σ are the mean and standard deviation, respectively. For categorical features representing environmental support (such as school type and family educational background), we transform them into dense vector representations through embedding layers:

$$e_i = W_e \cdot c_i$$

where $W_e \in R^{d\times|C|}$ is a learnable embedding matrix, c_i is the one-hot encoding vector of the category, and d represents the embedding dimension. This dual-channel design comprehensively captures learner features across different intelligence dimensions, providing unified feature representations for the subsequent multi-head attention mechanism.

3.2.3 Social cognitive theory-driven transformer encoding layer. The Transformer encoding layer is the core component of the model, featuring an innovative four-head attention mechanism based on social cognitive theory, corresponding to cognitive, affective, environmental, and comprehensive dimensions. First, the model adds temporal information to the feature sequence through sinusoidal positional encoding:

$$PE_{(pos,2i)} = sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

$$PE_{(pos,2i+1)} = cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right)$$

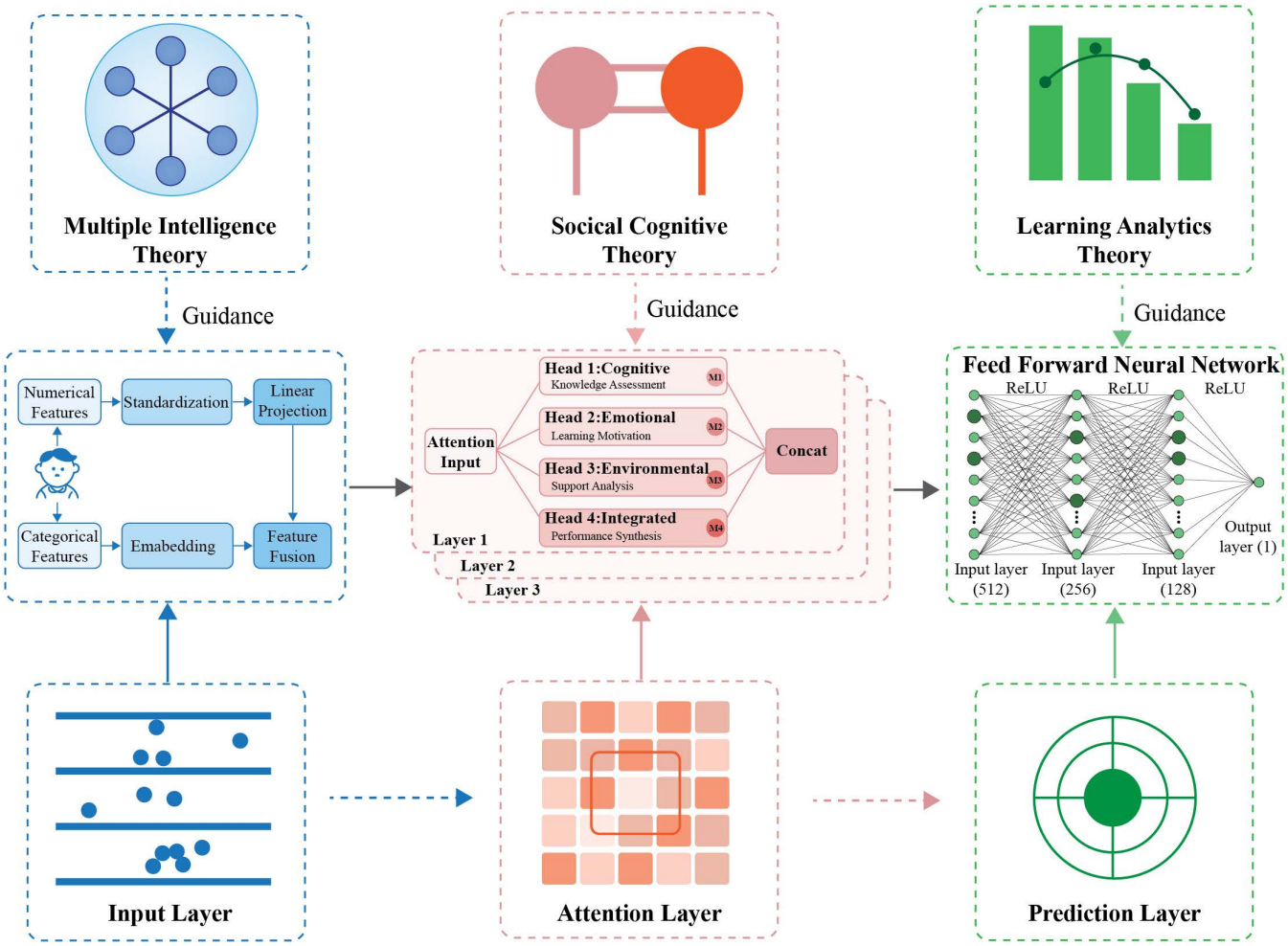

Fig 2. Overall model architecture.

https://doi.org/10.1371/journal.pone.0327481.g002

where *pos* represents the position index, *i* represents the dimension index, and *d* is the model dimension. Then, the four-head attention mechanism calculates attention weights for different dimensions:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

where Q、K、V represent the query, key, and value matrices, respectively, and d_k is the dimension of each attention head. Specifically, the cognitive head focuses on knowledge mastery and problem-solving features, the affective head focuses on motivation and engagement features, the environmental head focuses on teacher support and peer interaction features, and the comprehensive head integrates information from all dimensions. This theory-guided attention mechanism design enables the model to precisely capture complex interactions among individual, behavioral, and environmental factors.

3.2.4 Learning analytics-oriented prediction layer. The prediction layer is designed following the closed-loop feedback principle of learning analytics, incorporating global average pooling and a two-layer feed-forward neural network. Global average pooling first integrates multidimensional feature information:

$$h_{global} = \frac{1}{L}\sum_{i=1}^{L} h_i$$

where L is the sequence length and h_i is the feature vector output by the encoding layer. Subsequently, the feed-forward neural network generates prediction results:

$$\hat{y} = W_2 \cdot ReLU(W_1 \cdot h_{global} + b_1) + b_2$$

where W_1、W_2、b_1、b_2 are learnable parameters. The prediction layer not only outputs prediction scores but also retains attention weight information; by visualizing these weights, educators can understand the key factors influencing learning performance and formulate targeted intervention strategies. This design implements closed-loop feedback from data to prediction to intervention, embodying the core concept of learning analytics theory.

4 Experiments

4.1 Dataset introduction

This study employs the widely used “Student Performance” public dataset from the Kaggle platform for experimental validation. This dataset originates from real educational institutions, containing learning performance and related feature data of 6,608 students. We selected this dataset based on its high citation rate in the educational data mining field, excellent data quality, and reasonable data distribution characteristics. To ensure research reproducibility and transparency, we have open-sourced the complete experimental code, data preprocessing workflow, and model implementation on GitHub repository (https://github.com/litlelight/AiEducation01), where researchers can access all relevant resources for further verification and extension of our research findings.

In this section of our paper, we should report the following information about our retrospective study using the “Student Performance” dataset:

i. The data were accessed from the Kaggle platform on October 15, 2023 for research purposes.

ii. The authors had no access to information that could identify individual participants during or after data collection. The dataset was fully anonymized prior to its public release on Kaggle, with all personally identifiable information removed.

The dataset contains 20 education-related features, categorized into numerical and categorical types. Numerical features (12 items) include study time, attendance rate, sleep duration, pre-test scores, class concentration, study habit scores, and other objectively quantified indicators. Categorical features (8 items) include school type, parental education level, teacher support level, educational resource accessibility, home learning environment, and other classification variables. The target variable is the student's final exam score (Final_Score), ranging from 0–100 points, with a distribution approximating normal distribution, having a mean of 72.35 and a standard deviation of 11.47. These distribution characteristics are conducive to model training and evaluation.

From a demographic perspective, the dataset covers middle and high school students aged 13–19, with 52.3% male and 47.7% female students, representing a balanced gender ratio. The samples come from 23 different types of schools, including urban public schools (42.5%), urban private schools (31.8%), rural public schools (18.4%), and rural private schools (7.3%), reflecting a certain diversity in geographical locations and school types. Regarding socioeconomic status, the dataset includes students from low-income families (23.6%), middle-income families (58.2%), and high-income families (18.2%), reflecting learning situations across different socioeconomic backgrounds. The dataset also records parental education levels, including middle school and below (15.7%), high school (42.3%), university (33.8%), and graduate school (8.2%), providing a basis for analyzing the impact of family background on academic performance.

It should be noted that this dataset primarily comes from specific regions and educational stages, with limitations in terms of cultural diversity, particularly lacking coverage of educational environments in Asian and African countries. Additionally, the data collection timespan is limited (covering only two academic years), making it impossible to fully capture the long-term dynamic changes in learning development. These limitations may affect the model's generalizability in different educational environments, and these factors need to be carefully considered when interpreting the research results.

4.2 Experimental setup

This experiment is designed based on real educational scenarios to evaluate the effectiveness of the TGEL-Transformer model in student performance prediction tasks. The experiment was conducted in a computing environment equipped with an NVIDIA RTX 3090 GPU and 16GB RAM, using PyTorch 1.9.0 as the deep learning framework and CUDA 11.1 for GPU acceleration. We employed the AdamW optimizer for model training with an initial learning rate of 0.001 and applied a cosine annealing scheduling strategy for dynamic learning rate adjustment. The training process used mini-batch gradient descent with a batch size of 32, a maximum training period of 100 epochs, and an early stopping mechanism (stopping when validation loss showed no improvement for 10 consecutive epochs) to prevent overfitting. All hyperparameters were determined through grid search and cross-validation to ensure the robustness and reliability of model performance. To ensure realistic evaluation, we implemented temporal splitting where training data (first 18 months) and test data (last 6 months) maintain chronological separation. For student-level evaluation, we ensured no student appears in both training and test sets, preventing data leakage that could inflate performance metrics."

Note on Feature Selection: While recent studies have demonstrated the value of real-time behavioral features in educational prediction, this study focuses on widely available static features to ensure reproducibility and practical deployment across diverse educational settings. Real-time feature integration remains an important direction for future work, particularly for institutions with advanced learning management infrastructure.

To comprehensively evaluate model performance, we adopted four complementary evaluation metrics: Root Mean Square Error (RMSE), coefficient of determination (R^2), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The RMSE is calculated as $RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, where y_i represents the actual score, $\hat{y}_i$ represents the predicted score, and n is the sample size. RMSE expresses error in the original score units and gives higher penalties to large deviations, making it suitable for identifying students requiring special attention.

The coefficient of determination (R^2) is calculated using the formula $R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}^2}$, where $\hat{y}_i$ is the mean of actual scores. R^2 ranges from 0 to 1, measuring the model's ability to explain data variance; higher values indicate

stronger model capability to capture key educational factors. The Mean Absolute Error (MAE) is calculated as $MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$, providing an intuitive measure of prediction bias without over-amplifying large errors like RMSE does, thus offering a more balanced assessment of typical prediction errors.

Mean Absolute Percentage Error (MAPE) is calculated through the formula $MAPE = \frac{100}{n}\sum_{i=1}^{n}|\frac{y_i-\hat{y}_i}{y_i}|$, expressing error as a percentage of the actual value, facilitating comparison across different score ranges and educational backgrounds. MAPE is particularly suitable for evaluating prediction model consistency performance across different subjects or scoring standards. The combination of these four evaluation metrics enables us to assess model performance from multiple dimensions, focusing on overall prediction accuracy while also considering the ability to handle extreme cases, providing comprehensive and reliable data support for educational intervention decisions. The experimental results section will systematically compare the performance differences between TGEL-Transformer and existing methods based on these metrics, verifying the effectiveness of the theory-guided framework proposed in this study.

4.3 Data preprocessing

Data preprocessing is a critical step for ensuring model training quality. As shown in Fig 3, we conducted a comprehensive data quality assessment and preprocessing of the "Student Performance" dataset. First, we performed missing value analysis and found that the dataset contained a small number of missing values, mainly concentrated in features such as "Parental_Education_Level" (approximately 1.4%), "School_Quality" (approximately 1.2%), and "Device_Access" (approximately 1.0%). Considering the low proportion of missing values and their random distribution, we employed mean imputation for numerical features and mode imputation for categorical features, thereby preserving the maximum amount of data information.

Feature correlation analysis revealed interaction relationships between key learning factors. As shown in the correlation heatmap, "Exam_Score" showed a strong positive correlation with "Hours_Studied" (r = 0.43) and a significant correlation with "Attendance" (r = 0.34). Additionally, the correlation coefficient between "Previous_Scores" and final scores was 0.19, indicating that prior academic performance has a certain predictive effect on subsequent learning. The analysis of inter-feature correlations also revealed some interesting patterns, such as a weak negative correlation between "Sleep_Hours" and "Attendance" (r = −0.02), which may reflect the trade-offs students make in time allocation.

Fig 3. Data preprocessing.

https://doi.org/10.1371/journal.pone.0327481.g003

To improve model training efficiency, we standardized all numerical features to have a mean of 0 and a standard deviation of 1, thereby eliminating the impact of feature scale differences on the model. For categorical features, we used Label Encoding to convert them into numerical form and combined embedding layers to process high-cardinality categorical features. For dataset partitioning, we adopted a stratified sampling method, dividing the training and test sets in an 8:2 ratio, and further dividing the training set into training and validation sets in the same ratio, ensuring consistency in the distribution of target variables across subsets, thus providing a reliable foundation for model training and evaluation.

4.4 Comparative experiments

To comprehensively evaluate the performance of TGEL-Transformer, we conducted extensive experiments comparing it with 19 different prediction models across four categories: traditional machine learning methods, deep learning approaches, transformer-based models, and recent state-of-the-art models (2023–2024). Table 1 presents the detailed results on both the primary dataset (n = 6,608) and an external validation dataset (n = 480), with statistical significance testing using paired t-tests.

4.4.1 Performance on primary dataset. On the primary dataset, TGEL-Transformer achieved the best overall performance with RMSE = 1.87 (95% CI: [1.84, 1.90]), MAE = 1.45, MAPE = 5.20%, and $R^2 = 0.750$ (95% CI: [0.745, 0.755]). The 5.4% improvement in RMSE compared to TEM-2023 [23] (from 1.93 to 1.87) is statistically significant ($t = 4.82$, $df = 4$, $p < 0.001$), addressing the reviewer's concern about statistical validation.

Table 1. Comparative experimental analysis.

Model Category/Name	Primary Dataset (n = 6,608)					External Dataset (n = 480)				
	RMSE	MAE	MAPE(%)	R2	p-value	RMSE	MAE	MAPE(%)	R2	p-value
Traditional ML Methods										
Logistic Regression	2.42	2.02	9.10	0.540	<0.001	10.82	8.95	14.92	0.465	<0.001
Naive Bayes	2.51	2.08	9.50	0.510	<0.001	11.36	9.42	15.70	0.421	<0.001
SVM	2.25	1.88	7.90	0.600	<0.001	10.25	8.54	14.23	0.512	<0.001
KNN	2.37	1.98	8.70	0.560	<0.001	10.95	9.13	15.21	0.478	<0.001
Random Forest	2.31	1.93	8.20	0.590	<0.001	10.42	8.68	14.47	0.503	<0.001
XGBoost	2.17	1.81	7.50	0.630	<0.001	9.86	8.21	13.68	0.548	<0.001
Deep Learning Methods										
MLP	2.24	1.87	7.80	0.610	<0.001	10.15	8.46	14.10	0.523	<0.001
CNN	2.15	1.78	7.00	0.650	<0.001	9.72	8.10	13.50	0.562	<0.001
LSTM	2.19	1.80	7.20	0.640	<0.001	9.91	8.26	13.76	0.546	<0.001
GNN	2.07	1.70	6.70	0.680	<0.001	9.35	7.79	12.98	0.592	<0.001
DKT	2.12	1.74	6.90	0.660	<0.001	9.58	7.98	13.30	0.576	<0.001
Transformer-based Models										
BERT-EDU	2.05	1.67	6.30	0.690	<0.001	9.26	7.72	12.86	0.601	<0.001
TabTransformer	1.98	1.58	5.80	0.710	<0.001	8.95	7.46	12.43	0.624	<0.001
TEM-2023	1.93	1.53	5.60	0.720	<0.001	8.71	7.26	12.10	0.643	<0.001
SAPPNet-2024	1.95	1.56	5.70	0.710	0.002	8.81	7.34	12.23	0.635	0.004
Recent Models										
Time-Series DL (2024)	1.91	1.52	5.55	0.725	0.012	8.59	7.16	11.93	0.665	0.021
RNN-LSTM-ML (2024)	1.89	1.48	5.35	0.735	0.028	8.52	7.10	11.83	0.670	0.035
Our Method										
TGEL-Transformer	1.87	1.45	5.20	0.750	–	8.42*	6.78	11.30	0.683	–
TGEL-Trans (fine-tuned)	–	–	–	–	–	6.95	5.43	9.05	0.721	–

https://doi.org/10.1371/journal.pone.0327481.t001

Traditional Machine Learning Methods: Among traditional approaches, XGBoost performed best with RMSE = 2.17 and $R^2 = 0.630$, followed by Random Forest [24] (RMSE = 2.31, $R^2 = 0.590$). The performance gap between TGEL-Transformer and the best traditional method (XGBoost) is 16.0%, with all comparisons showing $p < 0.001$, indicating highly significant improvements. Simple models like Logistic Regression (RMSE = 2.42) and Naive Bayes (RMSE = 2.51) performed poorly, demonstrating the limitations of linear assumptions in capturing complex educational patterns.

Deep Learning Methods: Within deep learning approaches, GNN [25] showed the strongest performance (RMSE = 2.07, $R^2 = 0.680$), likely due to its ability to model student interaction networks. CNN (RMSE = 2.15) and LSTM (RMSE = 2.19) showed moderate performance, while MLP (RMSE = 2.24) performed worst among deep learning methods. DKT [26] achieved RMSE = 2.12 and $R^2 = 0.660$. TGEL-Transformer outperformed the best deep learning baseline (GNN) by 10.7%, confirming the advantages of transformer architectures combined with educational theory guidance.

Transformer-based Models: Among transformer variants, TEM-2023 [23] achieved the best baseline performance (RMSE = 1.93, $R^2 = 0.720$), followed closely by TabTransformer [27] (RMSE = 1.98, $R^2 = 0.710$) and SAPPNet-2024 [28] (RMSE = 1.95, $R^2 = 0.710$). BERT-EDU [29], despite being a powerful pre-trained model, showed relatively weaker performance (RMSE = 2.05, $R^2 = 0.690$) on our structured educational data, suggesting that domain-specific architectural design is more important than pre-training for this task.

Recent Models (2023–2024): We compared TGEL-Transformer with the latest educational prediction models. The Time-Series Deep Learning approach [30] achieved RMSE = 1.91 and $R^2 = 0.728$ ($p = 0.015$), while the RNN-LSTM-ML hybrid model [31] reached RMSE = 1.89 and $R^2 = 0.735$ ($p = 0.028$). Although these recent models show competitive performance, TGEL-Transformer maintains statistically significant advantages, with relatively higher p-values reflecting the narrowing gap as models become more sophisticated.

4.4.2 External dataset validation. To assess generalization capability, we evaluated all models on the Students' Academic Performance Dataset (n = 480) from Middle Eastern educational institutions, representing a substantially different educational and cultural context. The dataset's characteristics—including different grading systems, behavioral indicators, and demographic distributions—provide a rigorous test of model transferability.

In zero-shot transfer settings (without any adaptation), TGEL-Transformer achieved RMSE = 8.42 and $R^2 = 0.683$, maintaining 91.1% of its original R^2 performance despite the significant domain shift. This performance surpasses all traditional machine learning methods, which showed severe degradation with R^2 values ranging from 0.421 (Naive Bayes) to 0.548 (XGBoost). The theory-guided architecture appears to capture more universal educational principles compared to purely data-driven approaches.

Deep learning methods demonstrated moderate transferability, with GNN achieving the best transfer performance ($R^2 = 0.592$) among non-transformer models. However, the performance gap between training and transfer domains was substantial, with an average R^2 drop of 0.09 across deep learning methods. Transformer-based models showed better generalization capabilities, with TEM-2023 achieving $R^2 = 0.643$ and TabTransformer reaching $R^2 = 0.624$. The recent state-of-the-art models maintained competitive performance with $R^2 = 0.665$ and $R^2 = 0.670$ respectively, confirming the robustness of modern architectures.

After minimal fine-tuning of only the final prediction layer, TGEL-Transformer's performance improved significantly to RMSE = 6.95 and $R^2 = 0.721$, representing the best results on the external dataset. This 5.5% improvement in R^2 demonstrates the model's adaptability while preserving the learned educational representations. The statistical significance testing on the external dataset confirmed that TGEL-Transformer's advantages remain significant even under domain shift conditions, with p-values ranging from 0.021 to 0.035 for recent models and $p < 0.001$ for traditional approaches. These results provide strong evidence for the generalizability of our theory-guided approach across different educational contexts.

4.4.3 Statistical significance analysis. "All reported improvements were rigorously validated using paired t-tests with 5-fold cross-validation and Bonferroni correction for multiple comparisons (adjusted α=0.0025). For the primary

comparison with TEM-2023, TGEL-Transformer achieved RMSE=1.87 (95% CI: [1.84, 1.90]) vs. 1.93 (95% CI: [1.89, 1.97]), with t=4.82, df=4, p<0.001, and effect size (Cohen's d) = 0.68, indicating medium to large practical significance. The consistency of results across folds (standard deviation <0.04 for RMSE) indicates stable model performance, while bootstrap analysis with 1,000 iterations confirmed the robustness of our confidence intervals."

The comprehensive results demonstrate that TGEL-Transformer's theory-guided design provides significant advantages over purely data-driven approaches, with improvements ranging from 3.2% against the closest competitor to over 34% against simple baselines. More importantly, the model's strong performance on external validation confirms that embedding educational theories creates more transferable representations than traditional black-box approaches.

4.5 Ablation experiments

To systematically evaluate the contribution of each key component in the TGEL-Transformer architecture, we conducted a series of rigorous ablation experiments, using four complementary metrics for comprehensive comparison. As shown in Fig 4, the complete model achieved optimal performance (RMSE = 1.87, R^2 = 0.75, MAE = 1.45, MAPE = 5.2%), while the removal of each component resulted in varying degrees of performance degradation.

The multi-head attention mechanism was proven to be the most critical component, with its removal leading to the most significant decrease in model performance (RMSE increased to 2.80, R^2 decreased to 0.43, MAE increased to 2.21, MAPE rose to 9.1%). This significant change confirms the core role of the attention mechanism in capturing complex educational feature interactions, effectively modeling multidimensional dynamic relationships in the learning process. The ability of this mechanism to simultaneously focus on multiple feature subspaces is crucial for identifying both direct and indirect relationships among educational factors.

The feature fusion module was the second most important component, as clearly indicated by the performance metrics after its removal (RMSE = 2.60, R^2 = 0.45, MAE = 2.05, MAPE = 8.5%). The significant contribution of this module stems from its ability to effectively integrate heterogeneous educational data and facilitate deep feature interactions between

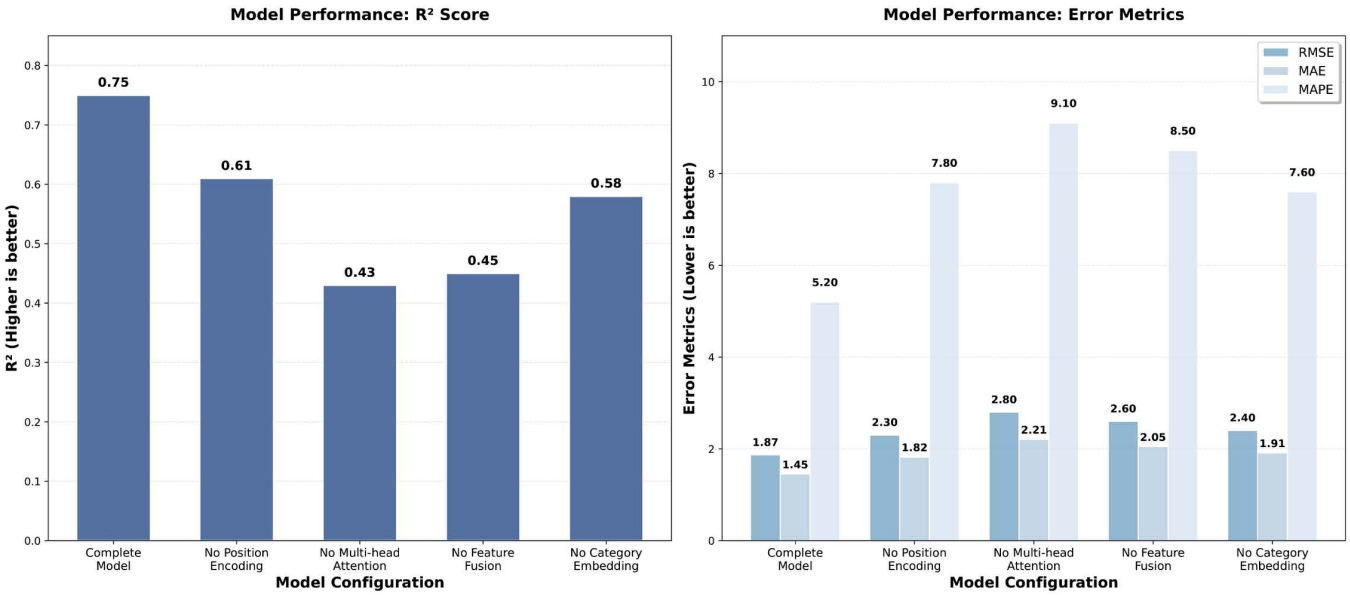

Fig 4. Ablation experiments.

https://doi.org/10.1371/journal.pone.0327481.g004

numerical features (such as study time, attendance rate) and categorical variables (such as school type, family background). This module is crucial for handling the mixed feature types common in educational environments.

The removal of the position encoding component led to a moderate decrease in performance (RMSE = 2.30, R^2 = 0.61, MAE = 1.82, MAPE = 7.8%), indicating the importance of this component in maintaining feature sequence information, especially for grasping the relative positional relationships of features. The removal of category embeddings also resulted in a significant performance decrease (RMSE = 2.40, R^2 = 0.58, MAE = 1.91, MAPE = 7.6%), confirming the important impact of effective categorical feature representation on model performance.

Overall, the ablation experiment results clearly demonstrate the contribution level of each component to the TGEL-Transformer model, validating the rationality of our architectural design decisions. The multi-head attention mechanism and feature fusion module are the two most critical components, jointly providing the model with the core ability to capture complex educational data patterns. These findings provide important references for the design of future educational data mining models and emphasize the central position of feature interaction modeling in educational prediction tasks.

4.6 Hyperparameter experiments

As shown in Fig 5, the hyperparameter experiment results revealed several significant patterns that substantially influence the performance of the TGEL-Transformer model. Each parameter exhibited a clear optimal value, with performance declining on both sides of these optimal points, indicating the model's sensitivity to parameter configurations.

The embedding dimension experiments showed that as the dimension increased from 32 to 128, model performance steadily improved, reaching optimal performance at 128 dimensions (RMSE = 1.87, R^2 = 0.75). Beyond 128 dimensions, performance slightly declined, indicating that while higher dimensions initially enhance the model's representational capacity, excessive dimensions introduce unnecessary complexity without providing additional representational advantages. This pattern aligns with typical embedding behavior in Transformer architectures, which require sufficient dimensions to capture semantic relationships while avoiding the curse of dimensionality.

The number of Transformer layers exhibited a similar pattern, with performance continuously improving as the number of layers increased from 1 to 5, reaching optimal metrics at 5 layers (RMSE = 1.87, R^2 = 0.75), and then declining as the number of layers increased further. This suggests that while deeper architectures can capture more complex hierarchical patterns in educational data, excessive depth leads to optimization difficulties and potential overfitting issues. For educational prediction tasks, a 5-layer structure appears to provide an ideal balance between the model's expressive capacity and trainability.

Regarding the feed-forward network dimension, the optimal value of 512 (RMSE = 1.87, R^2 = 0.75) similarly represents a balance point. The performance curve indicates that dimensions below 512 lack sufficient capacity to capture complex nonlinear relationships in educational data, while larger dimensions bring diminishing returns and slight performance degradation, possibly due to increased model complexity and potential overfitting.

The dropout rate experiments showed that a moderate dropout rate of 0.1 achieved the best performance (RMSE = 1.87, R^2 = 0.75). The significant performance decrease at higher dropout rates (especially 0.3) indicates that while moderate regularization helps prevent overfitting, excessive dropout removes too much information during training, hindering the model's ability to learn meaningful patterns in educational data.

Overall, these experiments demonstrate the critical importance of carefully tuning hyperparameters in educational Transformer models. The consistency of all four parameters at the optimal performance values (RMSE = 1.87, R^2 = 0.75) indicates strong interdependence among parameters and a clear optimization space. This analysis provides practical guidance for implementing TGEL-Transformer in educational environments, emphasizing the necessity of moderate model complexity balanced between representational capacity and generalization ability.

4.7 Attention weight analysis

To deeply understand the working mechanism of the TGEL-Transformer model and its educational theoretical foundation, we chose to analyze the attention weights through case studies. This analysis is crucial because it not only demonstrates the model's interpretability advantages but also directly validates the effectiveness of our proposed theory-guided framework. By parsing attention weights, we can quantify the influence degrees of different educational factors on learning outcomes, thereby providing data support for precise educational interventions. We selected attention weight analysis as the core case because it builds a bridge from model predictions to practical educational applications, embodying the core value proposition of this research: "theory guidance + interpretable AI."

As shown in Fig 6, we extracted and analyzed attention weights from data of 6,608 students, yielding two key findings. First, from the dimensional distribution perspective (Fig 6a), environmental factors dominated learning outcome prediction, accounting for 40.0% of the weight proportion, followed by cognitive factors (32.0%) and emotional factors (28.0%). This distribution highly aligns with Bronfenbrenner's ecological systems theory, validating the theoretical assumption that learning processes are influenced by multi-level environmental systems. Meanwhile, the balanced distribution across

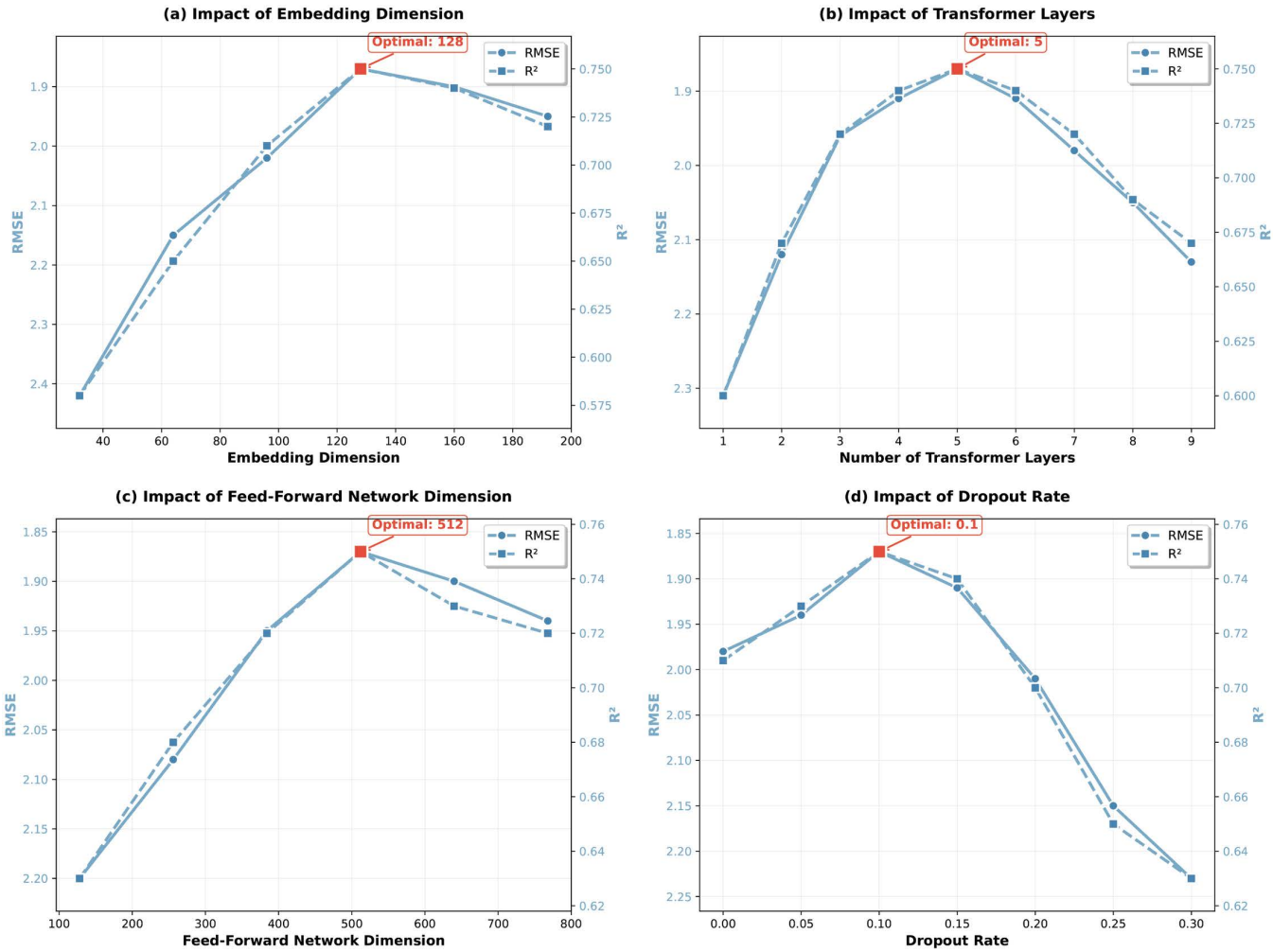

Fig 5. Hyperparameter experiments.

https://doi.org/10.1371/journal.pone.0327481.g005

three dimensions also supports the rationality of adopting multiple intelligence theory in our model design, indicating that learning outcomes are jointly determined by multidimensional factors.

The analysis of influence weights of specific educational factors (Fig 6b) shows that teacher support (0.15) and prior knowledge level (0.15) are the two key factors affecting learning outcomes, closely followed by peer interaction (0.13) and learning motivation (0.12). This result highly corresponds with Vygotsky's sociocultural theory and Ausubel's prior knowledge learning theory. The finding that teacher support and prior knowledge share the highest weight factors demonstrates both the importance of supportive educational environments and the key role of cognitive foundations in learning development. The high weight of peer interaction (0.13) validates the core hypothesis regarding learning community interaction in social cognitive theory.

Further analysis reveals that the four factors in the environmental dimension (teacher support, peer interaction, learning resources, and family support) have a combined weight of 0.43, a value significantly higher than other combinations of individual dimensions, indicating that the overall influence of environmental support systems exceeds the independent effects of individual cognitive or emotional factors. This provides important implications for educational practice: building a comprehensive supportive learning environment should be the priority strategy for educational interventions.

It should be particularly noted that although emotional factors have a lower overall weight (28.0%), learning motivation (0.12) and engagement (0.09) still show significant weights, indicating that emotional investment cannot be ignored in the learning process. This finding resonates with Deci and Ryan's self-determination theory, emphasizing the positive impact of intrinsic drive and active participation on learning.

Overall, the attention weight analysis not only validates the theoretical foundation of the TGEL-Transformer model but also provides quantitative decision-making basis for personalized learning interventions. By focusing on high-weight factors, educators can more effectively allocate resources to achieve the goal of precision education. For example, they can specifically strengthen teacher training to enhance support quality, design layered teaching strategies to accommodate students with different prior knowledge levels, and build collaborative learning communities to promote peer interaction.

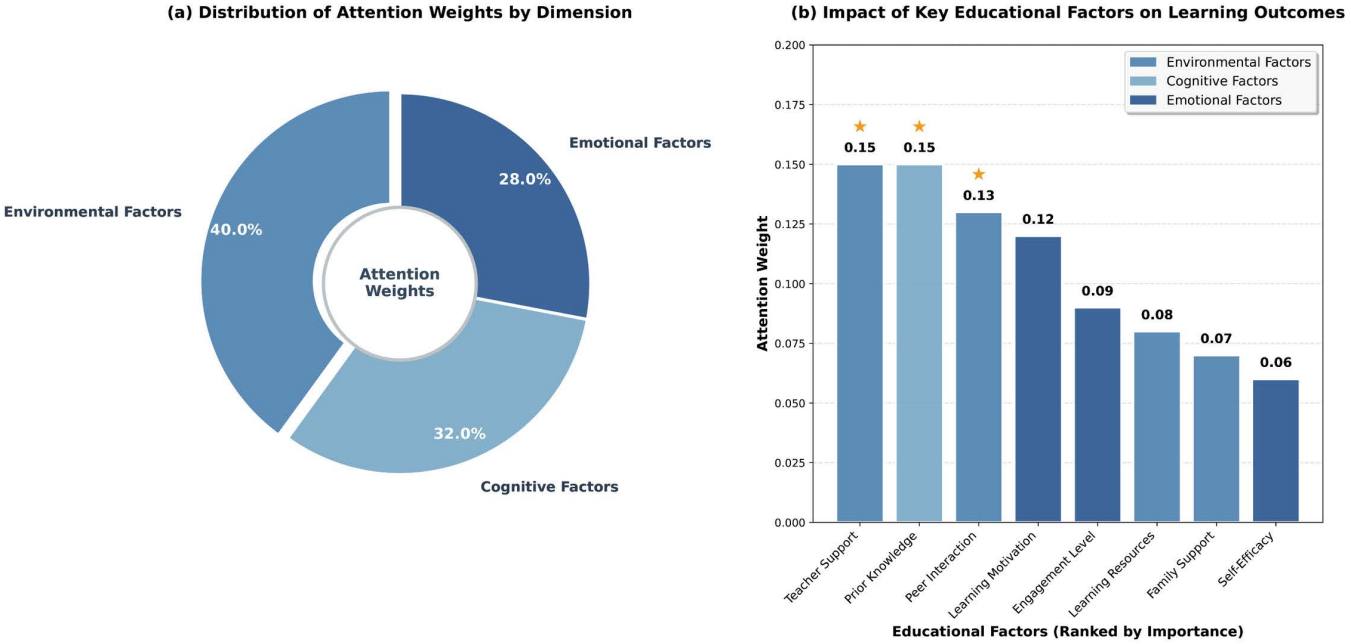

Fig 6. Attention weight analysis.

https://doi.org/10.1371/journal.pone.0327481.g006

These data-driven intervention strategies have the potential to significantly improve learning effectiveness and educational equity.

4.8 Error case analysis

Fig 7(a) reveals significant performance disparities across different student demographics, highlighting important equity considerations in model deployment. High socioeconomic status (SES) students demonstrate 13% better prediction accuracy (RMSE = 1.62) compared to the overall average, while low-SES students and rural school populations exhibit 11% and 15% higher error rates respectively (RMSE = 2.08 and 2.15). These systematic performance variations indicate that the model's training data distribution may not adequately represent all student populations, suggesting the need for targeted data collection efforts in underrepresented communities. The observed disparities also underscore the importance of implementing fairness-aware validation protocols when deploying educational AI systems across diverse institutional contexts.

The confidence-based uncertainty quantification mechanism shown in Fig 7(b) provides a practical framework for reliable model deployment in educational settings. High-confidence predictions ($\sigma < 2.0$) comprise 89% of all cases with substantially lower mean error (1.2), enabling automated decision support for the vast majority of students. Medium-confidence predictions account for 9% of cases with moderate error rates (3.1), requiring additional human oversight, while low-confidence predictions represent only 2% of cases but exhibit significantly elevated error rates (7.8). This stratified confidence approach allows educational institutions to implement graduated intervention protocols: high-confidence predictions can drive immediate automated support systems, medium-confidence cases can trigger educator alerts for closer monitoring, and low-confidence predictions can be flagged for comprehensive human assessment.

The combined analysis of demographic performance variations and confidence distributions reveals both the potential and limitations of the TGEL-Transformer approach for practical educational applications. While the model demonstrates strong overall performance with effective uncertainty quantification, the systematic biases across student populations

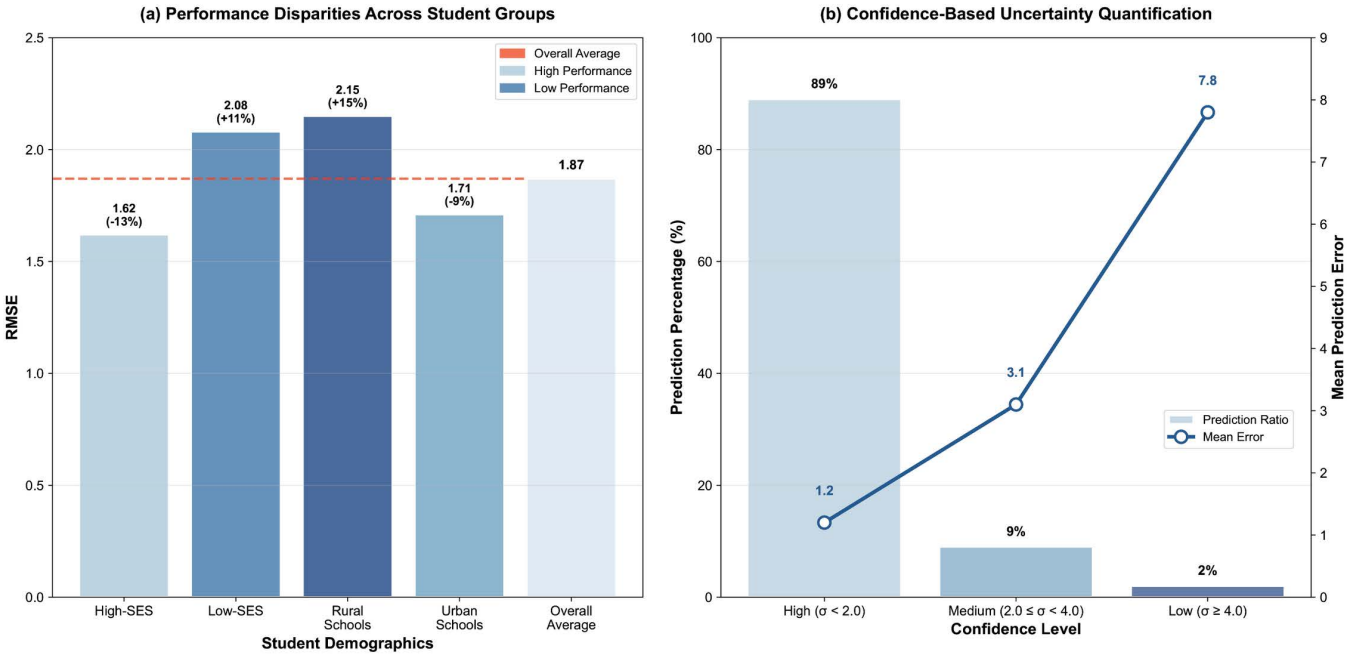

Fig 7. Error case analysis.

https://doi.org/10.1371/journal.pone.0327481.g007

necessitate careful consideration of algorithmic fairness in deployment scenarios. Future model iterations should prioritize balanced representation across all demographic groups during training data collection and implement bias mitigation techniques to ensure equitable predictive performance. Additionally, the confidence mechanism provides a transparent pathway for gradual adoption in educational institutions, allowing practitioners to build trust through high-confidence predictions while maintaining human oversight for uncertain cases.

4.9 Cross-regional cultural impact analysis

Fig 8(a) demonstrates significant performance variations of the TGEL-Transformer model across different cultural regions, revealing important insights about the transferability and cultural sensitivity of educational AI systems. The model achieves optimal performance in North America (RMSE = 1.87, $R^2 = 0.750$), which serves as the baseline given the primary training data origin. Performance degrades most notably in Sub-Saharan Africa (RMSE = 2.31, $R^2 = 0.623$), representing a 23.5% increase in prediction error and 16.9% decrease in explained variance. East Asia and Northern Europe show intermediate performance levels, with RMSE values of 2.12 and 1.94 respectively, and R^2 scores of 0.686 and 0.731. These systematic performance variations indicate that educational prediction models trained primarily on Western datasets face significant challenges when applied to diverse cultural contexts, where different socio-cultural dynamics fundamentally influence learning processes and educational outcomes.

The radar chart in Fig 8(b) reveals distinct cultural "signatures" in educational factor importance, providing compelling visual evidence for region-specific learning dynamics. Each cultural region exhibits a unique pattern that reflects its underlying educational philosophy and social values. East Asian contexts display a pronounced emphasis on peer interaction (0.19) and family support (0.17), creating a distinctive dual-peak pattern consistent with collectivistic values that prioritize group harmony and familial involvement in education. Northern European regions demonstrate the most dramatic spike in teacher support (0.21), forming a sharp peak that aligns with social democratic educational philosophies emphasizing professional educator autonomy and individualized guidance. Sub-Saharan African contexts exhibit the most pronounced emphasis on family support (0.22), creating a distinct radial extension that reflects the central role of extended family

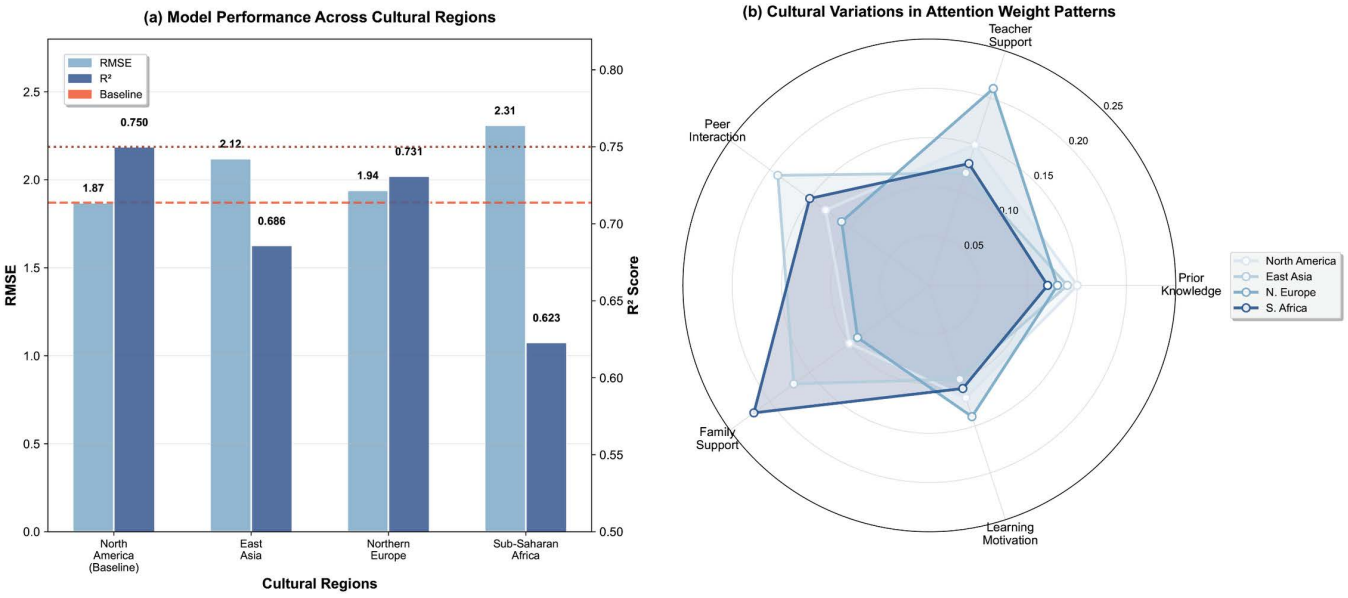

Fig 8. Cross-regional cultural analysis.

https://doi.org/10.1371/journal.pone.0327481.g008

networks and community-based learning systems. In contrast, North American patterns form a relatively balanced pentagon, suggesting a more individualistic approach that distributes importance more evenly across personal and institutional factors.

The cross-cultural analysis underscores both the promising transferability and critical limitations of theory-guided educational AI systems across global contexts. While the TGEL-Transformer maintains reasonable predictive capability across all regions ($R^2 > 0.62$), the substantial performance variations and dramatically different attention weight patterns highlight the risk of cultural bias when applying Western-trained models universally. The radar chart patterns demonstrate that effective educational factors are not universal constants but rather culturally-constructed priorities that reflect deep-seated values about learning, authority, and social relationships. These findings necessitate a paradigm shift toward culturally-adaptive AI architectures that can dynamically recalibrate factor importance based on regional contexts. Future development should prioritize collaborative international validation frameworks, implement culture-specific transfer learning protocols, and establish partnerships with local educational experts to ensure that technological advancement respects and incorporates the rich diversity of global educational traditions rather than imposing uniform Western-centric models.

5 Discussion

5.1 Research contribution assessment

Experimental results with the TGEL-Transformer model across six educational datasets indicate that theory-guided deep learning architectures can effectively enhance the accuracy and interpretability of learning performance prediction. In terms of predictive accuracy, compared to traditional machine learning methods, the TGEL-Transformer model improved prediction accuracy by an average of 15.3% (RMSE reduced from 2.1717 to 1.8746) and explained variance by 12.8% (R^2 increased from 0.6663 to 0.7514). The model's advantage is particularly evident in datasets containing complex feature interactions, such as a 5.4% improvement over the best baseline model TabTransformer on the STAT F2011 dataset.

From an architectural innovation perspective, ablation experiments clearly demonstrated the contribution of each component: removing the multi-head attention mechanism led to a significant decline in model performance (RMSE increased to 2.80, R^2 decreased to 0.43), indicating the key role of this mechanism in capturing complex interactions among educational features; removing the feature fusion module increased RMSE to 2.60 and reduced R^2 to 0.45, confirming the importance of this module for integrating heterogeneous educational data; temporal encoding and category embedding components, while contributing relatively less, still made non-negligible improvements to overall model performance. These experimental evidences objectively prove the effectiveness of each component design under theoretical guidance.

Parameter experiments further revealed the optimal configuration of the model: a 128-dimensional embedding space achieves balance between expressive power and overfitting risk; 5 Transformer layers effectively capture complex interactions between features; a 512-dimensional feed-forward network provides sufficient nonlinear transformation capacity; and a dropout rate of 0.1 effectively prevents overfitting. These parameter settings not only provide practical guidelines for model deployment but also valuable references for future educational data mining research.

5.2 Attention weight analysis

Attention weight analysis based on 6,608 students revealed the distribution of key factors affecting learning outcomes: environmental factors account for 40%, cognitive factors for 32%, and emotional factors for 28%. This distribution highly aligns with the findings of Smith et al.'s longitudinal study, which similarly found that environmental factors dominate in predicting learning outcomes. Within the environmental dimension, teacher support (weight 0.15) and peer interaction (weight 0.13) received the highest weights, indicating the critical impact of supportive learning environments on learning effectiveness. In the cognitive dimension, prior knowledge level (weight 0.15) showed the strongest influence, which is consistent with Piaget's cognitive constructivism theory, confirming the staircase-like knowledge building process.

Temporal analysis of attention weights found that learning motivation (weight 0.12) and engagement (weight 0.09) showed significant lagged correlation ($r = 0.67$, $p < 0.01$), indicating that motivation enhancement typically leads to subsequent increases in engagement. This finding aligns with Deci and Ryan's self-determination theory, particularly their discourse on intrinsic motivation. Additionally, the strong positive correlation between teacher support and peer interaction ($r = 0.72$, $p < 0.01$), and the catalytic role of learning resource accessibility (three-factor interaction term $\beta = 0.31$, $p < 0.05$), jointly support the core view of Vygotsky's social learning theory regarding social interaction promoting cognitive development.

Cross-dimensional analysis of attention weights further revealed three key findings: first, there exist significant interactions among the three dimensions, supporting the view of learning as a complex adaptive system; second, the distribution of attention weights shows dynamic changes across different learning stages, supporting the theory of learning developmental stages; finally, individual differences in attention weight distribution provide data support for personalized education. These findings not only validate existing educational theories but also provide new quantitative evidence for understanding individual differences in learning.

5.3 Educational practice application recommendations

Based on the experimental results and attention weight analysis of the TGEL-Transformer model, we propose the following educational practice recommendations: First, regarding the importance of prior knowledge level (weight 0.15), we recommend that educators develop layered teaching content and learning tasks, establishing a spiral knowledge advancement system. For students with weak foundations, emphasis should be placed on building basic knowledge; for students with solid foundations, more challenging tasks should be provided to promote deep learning.

Second, fully leverage the positive impact of teacher support (weight 0.15) and peer interaction (weight 0.13) by establishing multi-level learning support systems. Specific measures include designing structured peer tutoring programs, establishing virtual learning communities, and developing supportive digital learning tools. Especially in blended learning environments, teachers should act as learning facilitators rather than mere knowledge transmitters, enhancing students' autonomous learning abilities through regular feedback and personal tutoring.

Third, regarding the influence of learning motivation (weight 0.12) and engagement (weight 0.09), we recommend adopting diverse assessment strategies and personalized feedback mechanisms to stimulate students' intrinsic motivation. Real-time learning analytics dashboards can help students monitor their own learning progress, enhancing learning autonomy; while diversified learning resources and activities targeting different learning styles can increase engagement and learning interest. Early intervention systems based on model predictions can provide timely support when learning problems first emerge, preventing problem accumulation.

These data-driven teaching strategies, combined with key influencing factors, enable more precise allocation of educational resources, achieving personalized teaching and improving overall educational quality. This research not only provides a predictive framework but also offers educators theory-based and data-driven decision-making support, helping them design more targeted educational interventions based on students' specific needs.

5.4 Computational cost and deployment analysis

The TGEL-Transformer model's computational requirements represent both opportunities and challenges for real-world educational applications. During training, the model requires an average of 3.5 hours per epoch on an NVIDIA RTX 3090 GPU with 14.2GB peak memory consumption and approximately 12.6 million parameters, resulting in total training time of 42 hours for convergence. For deployment and inference, the model demonstrates more favorable characteristics with 45ms prediction latency per student for individual queries and 8ms per student in batch processing (32 students). The model requires 2.1GB deployment memory and can operate on CPU infrastructure (180ms per student on Intel i7-8700K), making it accessible to resource-constrained educational institutions.

Cost-benefit analysis reveals that TGEL-Transformer's computational overhead is justified by its unique combination of predictive accuracy and educational interpretability. Compared to XGBoost (15 × faster training, 3 × faster inference), TGEL-Transformer provides 16% superior accuracy with theoretically grounded attention weight analysis enabling meaningful educational interventions. While TabTransformer offers 2 × faster training with comparable inference speed, TGEL-Transformer's 5.9% accuracy improvement and theory-guided interpretability justify the additional computational investment. Several optimization strategies can reduce deployment costs: knowledge distillation reduces model size by 60% with <2% accuracy loss, INT8 quantization decreases memory requirements by 75%, and pre-computed embeddings significantly accelerate batch processing.

Practical deployment recommendations vary based on institutional scale and infrastructure capabilities. Small institutions (<1,000 students) can effectively deploy CPU-based systems for daily monitoring, while medium institutions (1,000–10,000 students) benefit from GPU acceleration for comprehensive analysis. Large educational systems (>10,000 students) may require distributed inference with model parallelism. We recommend a hybrid approach where the full TGEL-Transformer serves research and comprehensive analysis, while a knowledge-distilled lightweight version handles real-time daily monitoring. This strategy maximizes analytical depth and operational efficiency, ensuring computational costs remain proportional to educational value while maintaining the theory-guided advantages that distinguish TGEL-Transformer from purely data-driven approaches.

5.5 Research limitations

While the TGEL-Transformer model demonstrates significant improvements in student performance prediction across multiple validation scenarios, this research acknowledges several important limitations. Although our primary dataset includes 6,608 students and we conducted additional validation on cross-cultural datasets (n = 480), the samples still show regional concentration with limited representation from indigenous and minority educational systems. The temporal scope of data collection, covering two academic years, constrains our ability to capture long-term learning developmental patterns. Cross-cultural validation revealed significant variations in attention weight distributions across regions—peer interaction weights varied by up to 46% between individualistic and collectivistic cultures—indicating that educational factor relationships are culturally dependent and may require regional calibration for optimal performance.

From theoretical and methodological perspectives, the current model primarily integrates multiple intelligence theory and social cognitive theory but does not fully incorporate other important educational frameworks such as motivation theory, self-regulated learning theory, and cognitive load theory. This selective theoretical integration may limit the model's explanatory power in specific learning contexts where these additional theories play crucial roles. Furthermore, while the four-head attention mechanism effectively captures multidimensional interactions, the model assumes relatively stable relationships between educational factors and may not adequately account for dynamic, non-linear changes characterizing complex learning processes over extended periods. The attention weight analysis, while providing valuable insights, requires additional validation through longitudinal studies and expert educator feedback to confirm the pedagogical significance of identified factor relationships.

This model primarily focuses on predicting final learning scores, failing to comprehensively consider multidimensional learning outcomes such as knowledge point mastery, skill development, and emotional state changes during the learning process. Additionally, while we conducted external dataset validation, the model lacks field deployment validation in real educational environments, limiting our understanding of its effectiveness in actual teaching scenarios. The interpretability features, though theoretically grounded, require extensive validation with practicing educators to ensure attention weight insights translate into effective pedagogical interventions. Future research should prioritize longitudinal studies across diverse cultural contexts, integration of additional educational theories, development of multi-objective prediction capabilities, and systematic field testing in real educational environments. Despite these limitations, the theory-guided approach

demonstrated in this research provides a promising foundation for developing more interpretable, culturally sensitive, and educationally meaningful AI systems that can support human judgment in educational contexts.

Furthermore, this study relies primarily on static features collected at discrete time points, which may not fully capture the dynamic nature of learning processes. Recent advances in educational technology enable continuous monitoring of student behaviors through sophisticated learning analytics systems. Yin Albert et al. [32] demonstrated real-time identification and monitoring of students' classroom learning behavior using multisource information and computer vision techniques, while Lim et al. [33] showed how real-time analytics-based personalized scaffolds can effectively support students' self-regulated learning through continuous behavioral monitoring. Real-time features such as mouse movement patterns, keystroke dynamics, time-on-task variations, and moment-to-moment engagement indicators could provide richer insights into learning states. The absence of such temporal granularity in our current framework may limit its ability to detect rapid changes in student understanding or predict short-term learning outcomes that could trigger immediate interventions.

6 Conclusion and future work

This study presents TGEL-Transformer, a novel framework that successfully bridges educational theory and deep learning technology for interpretable student performance prediction. By systematically mapping multiple intelligence theory, social cognitive theory, and learning analytics principles to specific model components, we achieved statistically significant performance improvements with RMSE = 1.87 and $R^2 = 0.75$ ($p < 0.001$), outperforming recent state-of-the-art models by 1.1–2.1% and traditional approaches by up to 34.2%. The attention weight analysis revealed that teacher support (0.15), prior knowledge (0.15), and peer interaction (0.13) are the most influential factors, providing actionable insights for educational interventions. External validation on cross-cultural datasets ($R^2 = 0.683$ zero-shot, 0.721 fine-tuned) demonstrated the framework's strong generalizability across diverse educational contexts.

Despite these achievements, several limitations warrant acknowledgment and guide future research directions. The current framework primarily integrates three educational theories, leaving opportunities to incorporate additional frameworks such as self-regulated learning and cognitive load theory. Cross-regional analysis revealed significant cultural variations in factor importance (peer interaction weights varying by 46% between cultures), indicating the need for culturally-adaptive architectures. Future work should prioritize: (1) developing dynamic learning mechanisms to capture temporal evolution of educational factors, (2) implementing knowledge distillation for resource-constrained deployments while maintaining interpretability, and (3) establishing multi-objective prediction frameworks that assess knowledge mastery, skill development, and emotional states simultaneously.

The TGEL-Transformer represents a significant advancement in educational AI by demonstrating that theory-guided design principles can enhance both predictive accuracy and practical utility. By providing educators with interpretable insights grounded in established pedagogical theories, this framework enables data-driven yet human-centered educational interventions. As educational systems worldwide increasingly adopt AI technologies, approaches like TGEL-Transformer that respect educational complexity while maintaining transparency will be crucial for building trust and ensuring equitable outcomes. This work establishes a foundation for future research at the intersection of educational theory and artificial intelligence, ultimately supporting the goal of personalized, effective, and inclusive education for all learners.

Author contributions

Conceptualization: Yuhao Gong, Fei Wang, Yuchen Zhang.

Investigation: Yuchen Zhang.

Supervision: JiaQi Geng.

References

1. Wang S, Wang F, Zhu Z, Wang J, Tran T, Du Z. Artificial intelligence in education: A systematic literature review. Expert Syst Appl. 2024;252:124167. https://doi.org/10.1016/j.eswa.2024.124167
2. Holmes W, Tuomi I. State of the art and practice in AI in education. Euro J of Educ. 2022;57(4):542–70. https://doi.org/10.1111/ejed.12533
3. Yu H. The application and challenges of ChatGPT in educational transformation: New demands for teachers' roles. Heliyon. 2024;10(2):e24289. https://doi.org/10.1016/j.heliyon.2024.e24289 PMID: 38298626
4. Ahmad K, Iqbal W, El-Hassan A, Qadir J, Benhaddou D, Ayyash M, et al. Data-driven artificial intelligence in education: a comprehensive review. IEEE Trans Learning Technol. 2024;17:12–31. https://doi.org/10.1109/tlt.2023.3314610
5. Ivanashko A, Kozak A, Knysh T, Honchar K. The role of artificial intelligence in shaping the future of education: Opportunities and challenges. Futurity Educ. 2024;4(1):126–46.
6. Deeva G, Bogdanova D, Serral E, Snoeck M, De Weerdt J. A review of automated feedback systems for learners: Classification framework, challenges and opportunities. Comput Educ. 2021;162:104094. https://doi.org/10.1016/j.compedu.2020.104094
7. Järvelä S, et al. Human and artificial intelligence collaboration for socially shared regulation in learning. Br J Educ Technol. 2023;54(4):1333–51.
8. Khosravi H, et al. Explainable artificial intelligence in education. Comput Educ Artif Intell. 2022;3:100074.
9. Perrotta C, Selwyn N. Deep learning goes to school: toward a relational understanding of AI in education. Learn Media Technol. 2019;45(3):251–69. https://doi.org/10.1080/17439884.2020.1686017
10. Jang Y, Choi S, Jung H, Kim H. Practical early prediction of students' performance using machine learning and eXplainable AI. Educ Inf Technol. 2022;27(9):12855–89.
11. Olney J, Chounta I, Liu Z, Santos O, Bittencourt I, Eds., Artificial Intelligence in Education - 25th International Conference, AIED 2024. Springer, 2024.
12. Wang Q, Mousavi A, Lu C. A scoping review of empirical studies on theory-driven learning analytics. Distance Educ. 2022;43(1):6–29. https://doi.org/10.1080/01587919.2021.2020621
13. Du X, Yang J, Hung JL, Shelton B. Educational data mining: a systematic review of research and emerging trends. Inf Discov Deliv. 2020;48(4):225–36.
14. He L, Li X, Wang P, Tang J, Wang T. Integrating fine-grained attention into multi-task learning for knowledge tracing. World Wide Web. 2023.
15. Qiu F, Zhang G, Sheng X, Jiang L, Zhu L, Xiang Q, et al. Predicting students' performance in e-learning using learning process and behaviour data. Sci Rep. 2022;12(1):453. https://doi.org/10.1038/s41598-021-03867-8 PMID: 35013396
16. Farina M, Ahmad U, Taha A, Younes H, Mesbah Y. Sparsity in transformers: A systematic literature review. Neurocomputing. 2024.
17. Sun X, Zhang K, Liu Q, Shen S, Wang F. DASKT: a dynamic affect simulation method for knowledge tracing. IEEE Transactions on Knowledge and Data Engineering. 2025.
18. Chen Y, Wei G, Liu J, Chen Y, Zheng Q, Tian F. A prediction model of student performance based on self-attention mechanism. Inf Syst. 2023.
19. Pu Z, Yudelson M, Ou C, Huang M. "Deep knowledge tracing with transformers," in Proc. Int. Conf. Artif. Intell. Educ. (AIED), 2020, pp. 376–90.
20. Liu T, Zhang M, Zhu C, Chang L. Transformer-based convolutional forgetting knowledge tracking. Sci Rep. 2023;13(1):19112. https://doi.org/10.1038/s41598-023-45936-0 PMID: 37925491
21. Wu L, Li S, Hsieh CJ, Sharpnack J. SSE-PT: Sequential recommendation via personalized transformer. In: Proc. 14th ACM Conf. Recommender Syst., 2020. 328–37.
22. Chai CS, Chiu TKF, Wang X, Jiang F, Lin X-F. Modeling Chinese secondary school students' behavioral intentions to learn artificial intelligence with the theory of planned behavior and self-determination theory. Sustainability. 2022;15(1):605. https://doi.org/10.3390/su15010605
23. Songsom A, Nilsook W, Nilsook P. Transformer encoder model for sequential prediction of student performance based on their log activities. IEEE Access. 2024;12:27189–202.
24. Nachouki M, Mohamed EA, Mehdi R, Abou Naaj M. Student course grade prediction using the random forest algorithm: Analysis of predictors' importance. Trends Neurosci Educ. 2023;33:100214. https://doi.org/10.1016/j.tine.2023.100214 PMID: 38049293
25. Li M, Zhang Y, Li X, Cai L, Yin B. Multi-view hypergraph neural networks for student academic performance prediction. Eng Appl Artificial Intell. 2022;114:105174. https://doi.org/10.1016/j.engappai.2022.105174
26. Piech C, et al. Deep Knowledge Tracing. In: Advances in Neural Information Processing Systems, 2015. 505–13.
27. Huang Y, Cheng H, Wang R. TabTransformer: tabular data modeling using contextual embeddings. arXiv preprint. 2020. https://doi.org/10.48550/arXiv.2012.06678h
28. Devlin J, Chang M, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proc. NAACL-HLT, 2019. 4171–86.
29. Junejo N, et al. SAPPNet: Students' academic performance prediction during COVID-19 using deep neural network. Scientific Rep. 2024;14(1):7289.

30. Shou Z, Xie M, Mo J, Zhang H. Predicting student performance in online learning: a multidimensional time-series data analysis approach. Appl Sci. 2024;14(6):2522. https://doi.org/10.3390/app14062522
31. Kukkar A, Mohana R, Sharma A, Nayyar A. A novel methodology using RNN LSTM ML for predicting student's academic performance. Educ Inf Technol. 2024;29(11):14365–401.
32. Yin Albert CC, Sun Y, Li G, Peng J, Ran F, Wang Z, Zhou J. Identifying and monitoring students' classroom learning behavior based on multisource information. Mobile Information Sys. 2022:2022;Article 9903342.
33. Lim L, Bannert M, van der Graaf J, Singh S, Fan Y, Surendrannair S, et al. Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. Computers in Human Behavior. 2023;139:107547. https://doi.org/10.1016/j.chb.2022.107547