2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Automatic Scoring for Elementary Mathematics Solutions Based on Bi-LSTM and Attention Mechanism[1]

Ke Xu†[1,3], Yuan Sun†[2], Xi Zhang[3], Yan Huang[4], Songfeng Lu*[4]

and Yongqiang Zhang*[5]

[1] Research Institute of Huazhong University of Science and Technology in Shenzhen, Shenzhen, China
[2] College of Pharmacy, Hubei University of Chinese Medicine, Wuhan, China
[3] College of Computer Science, South-Central Minzu University, Wuhan, China
[4] School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan, China
[5] Wuhan Dameng Database Co., Ltd, Wuhan, China

Abstract. Automated scoring for mathematical subjective responses remains challenging due to the inherent complexity of integrating symbolic notations, procedural reasoning, and natural language explanations. Traditional NLP approaches fail to capture domain-specific mathematical semantics and logical dependencies between problem-solving steps. This study proposes a novel neural architecture named BiLSTM-AM that synergizes Bi-LSTM with hierarchical attention mechanisms to tackle pivotal challenges in the automated scoring of mathematical subjective responses. The architecture effectively models the conjunction of formulaic expressions and textual narratives through a dual-channel embedding strategy, dynamically allocates weights to pivotal procedural elements using step-level attention, and automates the alignment of student-generated solutions with knowledge graphs constructed from expert solutions. Evaluated on the dataset of 1120 elementary mathematics solutions, this paper achieves 90.52% scoring accuracy, outperforming state-of-the-art baselines by 8.17%. The novelty of this research is underscored by its interpretable attention mechanisms, which offer quantitative means to track the propagation of errors throughout the steps of mathematical solutions. This study contributes to the field of AI-enhanced educational evaluation by introducing a scalable and curriculum-sensitive framework designed for the assessment of open-ended mathematics problems.

Keywords: automatic essay scoring; elementary mathematics questions; problem solving; bidirectional LSTM; attention mechanisms

† co-first authors: K. Xu, Y. Sun, email: cole.xu@gmail.com, 2889@hbucm.edu.cn

* Corresponding author: K. Xu, Y. Sun, S.F. Lu, Y.Q. Zhang

1 Introduction

Automatic essay scoring systems (AES) have received much attention in the field of educational technology due to their high efficiency in improving student assessment efficiency and objectivity. A variety of technical means have been proposed and applied in practice, those researchers have achieved many results in the field of AES study [1, 2]. Cao studied the adaptability of AES in different fields and improved the accuracy and versatility of model evaluation [3]. Chen studied the PMAES model for cross-prompt automatic essay scoring and solved the problem of different evaluation consistency and accuracy in different situations [4]. The automatic scoring of solutions to elementary mathematics problems remains challenging due to the diverse approaches involved in deducing problem-solving strategies.

AES has seen successful implementation across multiple disciplines [5], but its application to evaluating problem-solving skills in elementary mathematics remains limited. Elementary mathematics problem solving involves a variety of logical reasoning, the study investigates automated scoring methods tailored specifically for elementary mathematics solutions. Refer.[6] fine-tunes pre-trained language models and integrates regression and ranking techniques to improve the performance of automated essay scoring. Hendrycks utilizes the MATH dataset to measure mathematical problem-solving abilities, providing a foundational framework for assessing students' mathematical competencies [7]. Refer.[8] explores an unsupervised automated essay scoring method, demonstrating the feasibility of scoring without extensive labeled data. Those study deepen our understanding of the elements and skills involved in mathematical problem-solving and offer valuable insights for advancing research in the automated assessment of elementary mathematics solutions.

Although AES are well-established in the domain of natural language processing (NLP), their implementation for automated assessment of elementary mathematics problem-solving continues to present persistent challenges [9–11]. A primary issue lies in adaptability, models trained on general datasets often underperform in mathematics-specific contexts due to the unique logical structure of mathematical problem-solving [12–14]. The availability of comprehensive, diverse, and domain-specific datasets for elementary mathematics problem-solving is still limited, which constrains the training and performance optimization of automated scoring models [7]. Existing studies struggle to extract critical details from mathematical problem-solving processes, such as computational accuracy, logical coherence, and reasoning chains [15, 16]. These limitations hinder the further improvement of accuracy and reliability in automated scoring systems for elementary mathematics.

In this paper, a novel method named BiLSTM-AM that combines Bi-LSTM networks and attention mechanisms is presented for automated scoring of elementary mathematics solutions. The method focuses on accurately capturing contextual information and key features in mathematical answers to provide accurate and reliable scoring results. The specific contributions of this research are as follows:

-This study takes advantage of Bi-LSTM to process sequence data, and uses an attention mechanism to focus on key aspects of the problem-solving process to improve performance in an elementary mathematics problem-solving assessment.

-The BiLSTM-AM method of applying the AES system to elementary mathematics problems, providing a new research direction for future elementary mathematics problem solving assessment.
-This research has to some extent reduced the workload of teachers in marking and improved the fairness and objectivity of marking.

Those investigation offer fresh perspectives and innovative methods to the field of automated scoring, and establishes a robust base for creating more precise and flexible solutions.

2 Related works

The research on Automatic Essay Scoring (AES) attracts much attention in the field of educational assessment in recent years, providing more research directions for student essay evaluation [17, 18]. This research promotes the development of unsupervised AES models, which have improved AES research and the ability to score various essays [8, 19]. This paper studies the current status of AES research, the application of the Bi-LSTM model, and the evaluation of math problem-solving results.

Traditional AES is mainly used to evaluate the quality of essays, it is gradually being extended to other application areas. LaVoie studies latent semantic analysis (LSA) and gives some results on the evaluation of different types of questions using automatic evaluation technology [12]. The Coh-Metrix model is used to automatically score multilingual essays and study its accuracy in multilingual tests [13]. The development and integration of multiple models and algorithms has made essay evaluation no longer a problem.

AES has various potential development directions beyond essay evaluation, expanding its applicability to broader educational assessments. Refer.[10] systematically describes AES and summarizes the current research status and future development trends in this field. Li researches the advantages and challenges of AES and proposes future research directions for this model [20]. Those researches point out the potential of AES to evaluate many aspects of learning.

Deep learning has contributed greatly to the development and application of AES. Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) networks play an important role [21]. Liu proposes a model that combines Bi-LSTM and attention mechanism for text classification, which can efficiently extract key information of text data [22]. The classification model is studied by integrating the attention mechanism on the basis of Bi-LSTM, and it is proposed that the attention mechanism has a strong ability in extracting text-related information [23, 24]. Lu combines deep learning and mathematical logic reasoning and proposes a model that can be applied to elementary and secondary mathematical logic reasoning [25]. These studies provide many directions for conducting research on logical reasoning and evaluation of mathematical problems.

This paper investigates the problem of scoring mathematical problem solving questions and extends the AES approach to the automatic assessment of mathematical problem solving ability [26]. The introduction of the MATH dataset [7], which covers

problems from mathematics competitions such as AMC 10, AMC 12, and AIME, provides a sample basis for the research of mathematical logic reasoning evaluation models. These studies provide methods for evaluating logical reasoning in mathematical problems. Refer.[15] investigates a method (MathScale) for the adaptation of mathematical reasoning instruction, which constructs concept maps by extracting topics and knowledge points from mathematical problems to generate new problems. Jie approaches mathematical word problems as complex relation extraction tasks and propose a method for learning deductive reasoning [16]. These studies provide valuable methodologies for logical inference in mathematical problem-solving. This paper proposes a new approach to the assessment of elementary mathematical problem solving by applying these methods and models.

3 Methodology

3.1 Data Preprocessing and Standardization

Linearized Representation of Mathematical Symbols. This study develops rules for liberalizing mathematical expressions based on LaTeX typesetting conventions. The goal of these rules is to linearize two-dimensional mathematical symbols into text sequences while preserving semantics. For example, the symbol θ is transformed into "theta" and the symbol α is transformed into "alpha". Table 1 gives the definitions of some symbol transformation rules.

Normalization of Formula Expressions. A standardized encoding scheme for mathematical expressions is determined to address syntax differences that occur in solving elementary math problems. This progress has effectively solved the problem of different symbols in math problems due to differences in writing styles, and ensures the uniformity of mathematical symbols.

Table 1. lllustrative Examples of Linearization Protocols

Mathematical symbols	Converted symbols	Mathematical symbols	Converted symbols
$\perp$	perpendicular to	θ	\theta
$\surd$	square root	α	\alpha
$\cap$	intersect at	λ	\lambda
$\in$	belong to	$\circ$	degree
$\subset$	contained in	n	vector n
$\because$	because	Δ	triangle
$\therefore$	so	$\angle$	angle

Variable-length Text Alignment. Mathematical problem solving is characterized by uncertain text length, and this study developed a specific length alignment framework

for this purpose. This framework utilizes a dynamic fillingtruncation hybrid mechanism and comprehensively considers the structural and semantic features in the process of mathematical problem solving.

Let $X = \{x_1, x_2, \ldots, x_n\}$ denote an input sequence of mathematical operations, where n varies across solutions.

Our alignment framework is implemented as follow:

$$L = \lceil \mu + 2\sigma \rceil \tag{1}$$

where μ and σ represent the mean and standard deviation of solution lengths in the training set.

Sequences shorter than L receive context-aware padding:

$$P(X) = X \oplus \{[PAD]_k | TF - IDF(X) \cdot W_p\}_{k=1}^{L-n} \tag{2}$$

where $W_p \in \mathbb{R}^d$ denotes learnable padding embeddings weighted by the solu- tion's TF-IDF significance.

Those mathematical expressions are decomposed through LaTeX parse trees *T*(*X*) generating structure-aware masks:

$$M_s = BiLSTM\left(GCN(T(X))\right) \odot \sigma(W_m \cdot X) \tag{3}$$

where *GCN* denotes graph convolutional network operations over equation syntax trees.

Implement hybrid attention for cross-step alignment:

$$\alpha_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{L} exp(e_{ik})},\ e_{ij} = h_i^{\top} W_a h_j \tag{4}$$

where h_i, h_j are the Bi-LSTM hidden states, and W_a is an alignment matrix.

The context vector is defined as follows:

$$c_i = \sum_{j=1}^{L} \alpha_{ij}\, h_j \oplus \underset{1 \le k \le L}{max}(h_k) \tag{5}$$

in which it combines soft alignment with local salience detection.

3.2 Analysis of Multiple Problem-solving Approaches

This paper focuses on the subjective problems in elementary mathematics. Elementary mathematics problems themselves have the characteristics of multiple methods of solving problems. For example, in solving solid geometry problems, comprehensive analysis or vector method can be used to solve the problem. In the same problem-solving method, individual steps may use a combination of multiple problem-solving techniques.

Analysis of Key Points in Problem-solving Steps. This study system-atically identifies the key steps in solving specific mathematical problems by analyzing a large number of problem-solving samples. This method conducts in- depth semantic analysis and logical reasoning to extract the key features in each problem-solving step.

Supplementary Reference Answers for Novel Problem-solving Approaches. In order to deal with the unconventional problem-solving methods that students may use in the process of solving problems due to their innovative thinking, this study expands the optional problem-solving methods by using basic theoretical reasoning and expert-driven evaluation techniques to evaluate the correctness of these non-standard answers.

Candidate Reference Answer Matching. This study proposes the Dynamic Reference Answer Matching (DREAM), which integrates semantic pattern recognition and morphological analysis to solve the problem of multiple solution methods for subjective mathematics questions.

Let $R = \{r1, r2, \ldots, rn\}$ denote the set of reference answers.

For each r_i , we extract distinctive key phrases $K_i = \{k_{i1}, k_{i2}, \ldots, k_{im}\}$ through the following formula:

$$TF\text{-}IDF(k_{ij}) = \frac{f(k_{ij}, r_i)}{|r_i|} \times log\left(\frac{n}{1+\sum_{q=1}^{n} \mathbb{I}(k_{ij} \in r_q)}\right) \tag{6}$$

where $II(\cdot)$ is the indicator function. Phrases with TF-IDF scores exceeding threshold τ are retained as solution-specific features.

The paper utilizes Bi-LSTM with attention to encode student answer:

$$\overrightarrow{h_t} = LSTM\left(e(s_t), \overrightarrow{h_{t-1}}\right) \tag{7}$$

$$\overleftarrow{h_t} = LSTM\left(e(s_t), \overleftarrow{h_{t+1}}\right) \tag{8}$$

where s_t represents the t - th token (e.g., word) in the student's answer sequence, $e(\cdot)$ denotes word embedding.

The attention weights αt are computed as follows:

$$\alpha_t = \frac{exp(e_t^T u_a)}{\sum_{k=1}^{T} exp(e_k^T u_a)} \tag{9}$$

where $e_t = \tanh(W_a h_t + b_a)$.

The final representation is defined by the following equation:

$$v_S = \sum_{t=1}^{T} \alpha_t \, h_t \tag{10}$$

The matching score between student answer S and reference r_i is calculated.

$$MatchScore(S, r_i) = \frac{\sum_{k \in K_i} sim(v_S, v_k)}{|K_i|} \times \left(1 - \frac{ED(S, r_i)}{max(|S|, |r_i|)}\right) \tag{11}$$

where $sim(\cdot)$ denotes cosine similarity and $ED(\cdot)$ is normalized edit distance.

The optima reference is selected as:

$$r^* = arg \max_{r_i \in R} MatchScore(S, r_i) \tag{12}$$

This study uses keyword extraction and fuzzy matching methods to compare the similarity between student answers and reference answers, and select the most appropriate reference answers for students. The pseudo-code of this method is shown in Algorithm 1.

Algorithm 1: Dynamic Reference Answer Matching

Input: *A*: Preprocessed student answer text
R: {*r1, r2 ,..., rn*} (Set of reference answers)
Output: *best_reference*: Matched reference answer with highest confidence
Initialize:
max_score ← *0*
best_reference ← *null;*
Train the prompt discriminator through *Eq. (10)*;
for *each reference answer ri* ∈ *R* **do**
    Ki ← extract_key_phrases(ri);
    for each key_phrase kij ∈ Ki do
        Uij ← *tokenize(kij);*
        UA ← *tokenize(A);*
        $match\_core_{ij} \leftarrow \frac{|U_A \cap U_{ij}|}{|U_A \cup U_{ij}|}$;
        $total\_match\_core$ += $match\_core_{ij}$;
    end
    avg _score ← $\frac{\text{totalmatchscore}}{|K_i|}$;
    if *avg_score* > *max_score* **then**
        max_score ← *avg_score*;
        best_reference ← r_i ;
    end
end
return *best_reference*;

3.3 Sequence Feature Learning Based on Bi-LSTM

Context-dependent Capture. This study employs a Bi-LSTM network to analyze the sequential patterns in elementary mathematics solutions, capturing context-dependent information that is crucial for understanding the flow of mathematical reasoning.

The Bi-LSTM architecture models sequential dependencies in mathematical solutions. The final representation concatenates both directions:

$$h_t = \left[\overrightarrow{h_t}; \overleftarrow{h_t} \right] \tag{13}$$

where $h_t \in \mathbb{R}^{2d}$ captures contextual patterns centered at position *t*.

Error Pattern Detection. This paper introduces an error pattern detection method that utilizes features extracted based on Bi-LSTM networks. The method analyzes the temporal dynamics encoded in the hidden states and the discriminative patterns within their output representations. The attention layer computes significance scores for each position.

The context vector $\mathrm{c} = \sum_{\mathrm{t}=1}^{\mathrm{T}} \alpha_{\mathrm{t}}\, \mathrm{h}_{\mathrm{t}}$ is then fed into a softmax classifier for error type prediction:

$$P(y|X) = softmax(W_c c + b_c) \tag{14}$$

where $y \in \{Calculation, Conceptual, Logical\}$ denotes error categories. This dual-directional analysis effectively detects error propagatior patterns in multistep solutions.

3.4 BiLSTM-AM (Bi-LSTM with Attention Mechanisms) Framework

This paper proposes a novel architecture named BiLSTM-AM, which combines Bi-LSTM with hierarchical attention mechanisms to address key challenges in the automated scoring for mathematical questions, the architecture is illustrated in Figure 1.

The overall framework of the BiLSTM-AM model can be described as follows:

-Input layer: The input sequence is embedded in a higher dimensional space.

-Bi-LSTM layer: The embedded sequence is processed by a Bi-LSTM layer to capture context-dependent information.

-Attention layer: Attention weights are computed and applied to the hidden states to obtain a context vector.

-Fusion layer: Symbolic embeddings are fused with the context vector to improve reasoning capabilities.

-Output layer: The final output is computed using a fully connected layer.

This framework effectively captures the complex interactions and dependencies in mathematical solutions, leading to more accurate and reliable automated scoring.

Step-level Attention. The BiLSTM-AM framework addresses the combination of mathematical formulae and text in student solutions through a dual-channel representational learning paradigm that dynamically assigns weights to key procedural elements using step-level attention.

For each solution step s_{i} containing n tokens $\{\mathrm{w}_1, \ldots, \mathrm{w}_{\mathrm{n}}\}$, we construct dual-channel embeddings: where M denotes mathematical symbol set, $\phi(\cdot)$ maps symbols to type-specific embeddings.

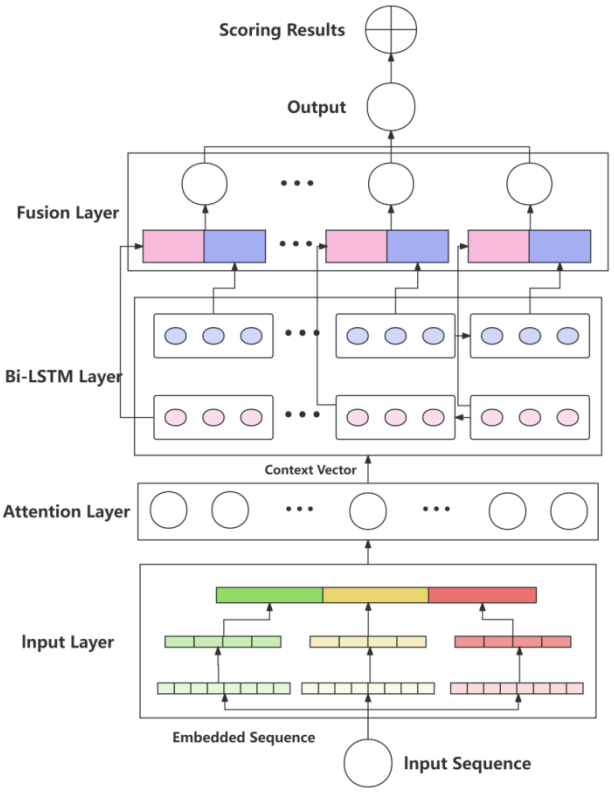

Fig. 1. Overview of the general framework diagram.

The step representation derives from attention-weighted fusion:

$$\beta_i = \sigma\left(W_\beta\left[E_{math}^{(i)}; E_{text}^{(i)}\right] + b_\beta\right) \tag{15}$$

Symbolic Semantic Fusion. The BiLSTM-AM framework combines the syntactic configurations and semantic embeddings of mathematical symbols, effectively modelling operator-operand dependencies.

The dependency between the operator and the operand can be defined by symbolic graph convolution:

$$G_i = SyntaxTree(s_i) \tag{16}$$

$$Z_i = GCN(G_i, \{z_w | w \in s_i\}) \tag{17}$$

$$\hat{h}_i = LayerNorm(W_z Z_i + W_h h_i) \tag{18}$$

where z_w denotes symbol embeddings, GCN propagates syntactic constraints through adjacency matrix $A_{ij} = \mathbb{I}(\text{node}_i \leftrightarrow \text{node}_j)$.

3.5 Adaptive Scoring Based on BiLSTM-AM

This paper proposes an adaptive scoring method that incorporates multi-criteria similarity fusion, taking into account formula structures, conceptual representations and error propagation patterns in mathematical problem solving processes.

The final score aggregates step-level predictions through temporal attention.

$$u_t = tanh\left(W_u \hat{h}_t\right) \tag{19}$$

$$\gamma_t = \frac{exp(u_t^T u_\gamma)}{\sum_{k=1}^{m} exp(u_k^T u_\gamma)} \tag{20}$$

$$Score = \sum_{t=1}^{m} \gamma_t \cdot \left(W_c \beta_t \tilde{h}_t + b_c\right) \tag{21}$$

Multi-strategy Similarity Fusion. The multi-criteria similarity fusion scoring method achieves the evaluation of elementary mathematical solutions by combining syntactic structure alignment, semantic equivalence assessment and error pattern correlation analysis.

Scoring of Key Steps. This paper presents a scoring method based on step-level attention, which assigns different scores to key points in problem-solving steps according to their importance and difficulty.

Summary of Final Scores. The method aggregates the scores of the individual steps and the similarity metrics in order to obtain the final score of the student's answer.

Algorithm 2: Final Scoring Based on BiLSTM-AM

Input: S : student's answer sequence (already processed through previous steps)
m: Number of steps in the answer
Pre-trained model parameters: $W_u, W_e, W_B, W_h, b_c, b_\theta$
Output: Score: Final score of the student's answer

Initialize: Score ← 0;
for t = 1 **to** m **do**
    Calculate $u_t = \tanh(W_u h_t)$
    %where h_t is the hidden state from BiLSTM for step t
    Calculate $\gamma_t = \frac{\exp(u_t^T u_\gamma)}{\sum_{k=1}^{m} \exp(u_k^T u_\gamma)}$
    Calculate $\beta_t = \sigma\left(W_\beta \left[F_{\text{math}}^{(t)}; F_{\text{text}}^{(t)}\right] + b_\theta\right)$
    %assuming $F_{\text{math}}^{(t)}$ and $F_{\text{text}}^{(t)}$ are already computed
    Calculate intermediate_score = $\gamma_t \cdot (W_c \beta_t + h_t \cdot W_h + b_c)$
    Score = Score + intermediate_score
end
return Score;

4 Results and Discussion

4.1 Analysis of Multiple Problem-solving Approaches

Learning Agency Lab-Automated Essay Scoring 2.0. The competition dataset comprises about 24000 student-written argumentative essays on Kaggle. Each essay was scored on a scale of 1 to 6. The goal is to predict the score an essay received from its text. The dataset is organized as three attributes which are the essay ids, the essays, and their corresponding scores, and 17308 rows in total, used as training data and fine-tuning for model development.

Elementary Mathematics Geometry Questions. The experimental data came from the final mathematics exam of the second semester (spring 2022) of the second year of the second grade at Townsend Lake School in East Lake High-tech Development Zone, Wuhan, where a total of 11 sub-topics of the three-dimensional geometry type were selected from the exam questions, making a total of 1020 questions. The total score for each question ranged from 3 to 5 points, and multiple solutions and a step-by-step scoring rubric were manually developed.

4.2 Data Preprocessing

Learning Agency Lab. The preprocessing of the ASAP dataset for LSTM, Bi-LSTM, and BERT models comprises several critical steps to ensure data is properly formatted for training and fine-tuning. Given that neural networks require fixed-length inputs, text sequences are padded for shorter texts or truncated for longer ones to ensure uniform input size.

Mathematics Geometry Questions. In this experiment, the total score of 5 points of the topic, the scoring score is the same or the difference of 1 point, that is, the scoring results were found to be accurate, and the other points of the topic in accordance with the proportion of discount. In this paper, the tenfold cross-validation method is used, and 1/10 of the samples (102 entries) are taken in turn each time as the test set, and the other responses are used as the training set. Statistics on mathematics exam questions and solution-related information are shown in Table 2, such as number of solutions, number of solution steps, number of keywords, average length of reference answer, average length of student answers and number of student answers.

This experiment converts mathematical symbols such as mathematical formulae, expressions and two-dimensional matrices in elementary mathematics into Chinese text strings, and some examples of linearised conversions are shown in Figure 2.

Table 2. Statistics on test questions and solution-related information

ID	score	Solns.	Sol.Step	K.P	Avg.R.A.Len.	Avg.S.A.Len.	S.A
F0901	3	1	3	7	189	231	132
F0902	4	2	3	8	152	165	127
F0903	3	1	2	4	155	210	129
F1001	4	1	2	7	100	159	127
F1002	5	3	3	7	428	325	128
F1003	4	1	4	8	529	369	123
F1201	5	5	3	10	278	375	48
F1202	5	3	5	5	480	450	48
F1501	4	3	2	7	150	120	52
F1502	5	2	3	4	328	330	54
F1503	5	3	2	5	226	186	52

Mathematical symbols	*Converted texttext*
四棱锥 E-ABCD 的体积: $V_{E-ABCD}=\frac{1}{3}\times S_{\cdot ABCD}\times EO=\frac{1}{3}\times\frac{5\times5}{2}\times6=25$	四棱锥 E-ABCD 的体积 V 右下 (E-ABCD)=(1/3)*S 右下(平行四边形 ABCD)*EO=(1/3)*(5*5/2)*6=25
$\begin{cases}x+\sqrt{5}z=0\\2x-4z=0\end{cases}$	{x+(根号 5)*z=0，2x-4z=0}
$\begin{vmatrix}x & y & z\\-2 & 0 & 6\\1 & 0 & -3\end{vmatrix}$	|（x，y，z），(-2，0，6），(1，0，-3）|

Fig. 2. Examples of Linearisation Conversion.

4.3 Candidate Reference Answer Matching

The DREAM method combines semantic pattern recognition and morphological analysis to automatically match the optimal reference answer to each student answer. For example, there is a question with two solutions, its two reference answers and the student's answer, as shown in Figure 3.

Answer Type	*Description of the answer in Chinese text*
Student Answer	由于 AC 平行 GF，DE 又垂直于 GF，故 AC 垂直于 DE 得证。
Reference Answer 1	角 BAC 等于角 DCE、AC 垂直于 DE ……
Reference Answer 2	AC 平行于 FG、DE 垂直于 FG ……

Fig. 3. Example of multiple solutions matching.

This experiment computes the key phrases of the reference answers of the two solution methods with the same student answer using single character participle, ordinary participle and DREAM respectively. According to the calculation of Algorithm 1 in this paper, the matching rate of Solution 2 is higher than that of Solution 1. Therefore, Solution 2 is selected as the optimal reference answer for scoring the student's answer.

4.4 Experimental results on Learning Agency Lab

The comparative study of the LSTM, Bi-LSTM, and BiLSTM-AM models for automated essay scoring provides significant insights into their respective capabilities and limitations.

Figure 4 presents the trends of the folded lines, which offer a comprehensive perspective on the performance and evolution of the three models, namely LSTM, Bi-LSTM, and BiLSTM-AM, during the training process over ten epochs. For the LSTM model, the training loss drops remarkably from 2.9842 to 0.2313. This significant reduction indicates that the model has strong learning capabilities and an improved accuracy on the training data. The Bi-LSTM model also demonstrates a substantial decrease in training loss, from 2.1316 to 0.2908. Regarding the BiLSTM-AM model, its training loss decreases notably from 1.8174 to 0.2207. Although the rate of decrease is slightly slower compared to the LSTM and Bi-LSTM models, it still reflects effective learning during the training process.

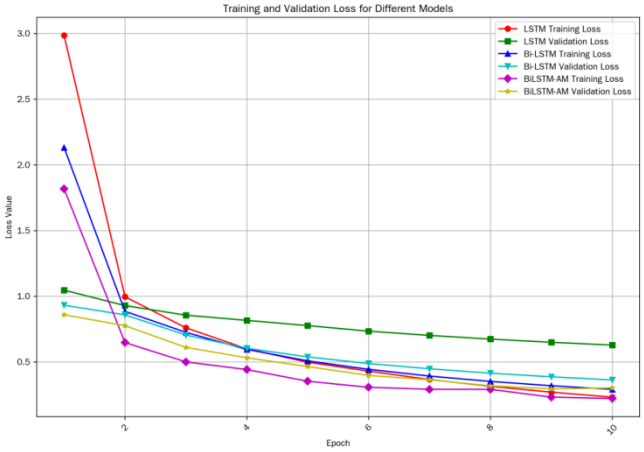

Fig. 4. Training and Validation Loss for Different Models.

Comparative analysis of the LSTM, Bi-LSTM, and BiLSTM-AM models used for automatic scoring can provide insights into their peculiarities and limiting factors. The dot-line graph in Figure 5 divides the performance indicators of each model, and their comparative advantages and disadvantages can be clearly seen.

As shown in Table 3, the BiLSTM-AM model shows stronger performance in all metrics compared to the LSTM and Bi-LSTM models. Specifically, the BiLSTM-AM model has the lowest MAE of 0.4153, which indicates a higher prediction score accuracy, better than the LSTM model's MAE of 0.5145 and the Bi-LSTM model's MAE of 0.4257. Similarly, the BiLSTM-AM model has the lowest MSE of 0.2990, which is significantly lower than the LSTM model's MSE of 0.4893 and the Bi-LSTM model's MSE of 0.3225, indicating that the model can more effectively reduce the prediction error.

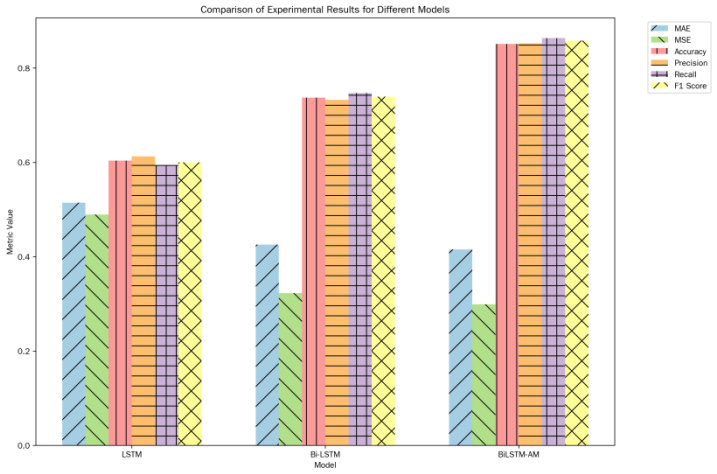

Fig. 5. Experimental Results for Different Models

In addition, the accuracy of LSTM and Bi-LSTM is 0.6035 and 0.7371 respectively, but the accuracy of BiLSTM-AM model is as high as 0.8513. The higher accuracy reflects that BiLSTM-AM model has a stronger ability to accurately classify articles into corresponding score categories.

The BiLSTM-AM model's strong performance is mainly due to its transformer-based architecture with a bidirectional attention mechanism. Compared with the sequential processing methods adopted by LSTM and Bi-LSTM models, this design can more effectively capture complex contextual relationships. This comparative study shows that in the field of automatic essay scoring, improved transformer models such as BiLSTM-AM have advantages over traditional sequential models, especially in achieving higher performance indicators.

Table 3. Evaluation metrics for Models

Model	MAE	MSE	Accuracy	Precision	Recall	F1-Score
LSTM	0.5145	0.4893	0.6035	0.6125	0.5942	0.5998
Bi-LSTM	0.4257	0.3225	0.7371	0.7324	0.7468	0.7395
BiLSTM-AM	0.4153	0.2990	0.8513	0.8523	0.8634	0.8577

4.5 Model Performance on Elementary Mathematics

This section evaluates the performance of the BiLSTM-AM model in automatic scoring of elementary mathematics solutions and compares it with the performance of two baseline models, LSTM and Bi-LSTM. The evaluation is performed on a dataset of elementary school mathematics geometry problems, which consists of 1020 problems with varying degrees of difficulty and multiple solutions.The obtained performance indicators such as accuracy, precision, recall, and F1-Score are shown in Table 4, and the intuitive graph is shown in Figure 6.

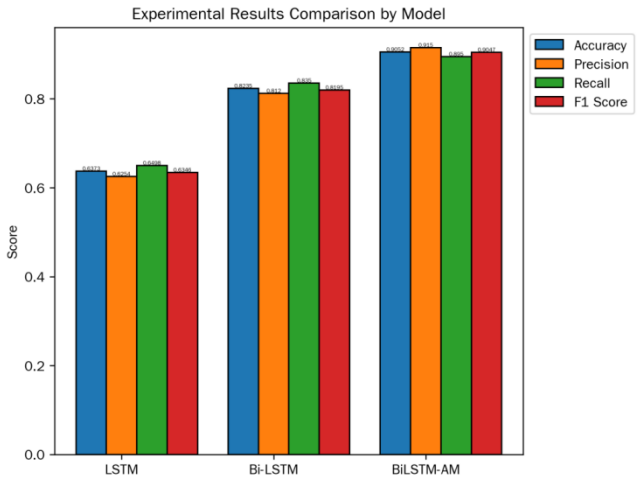

Fig. 6. Performance on Elementary Mathematics

Table 4 clearly shows that the LSTM model evaluation has an accuracy of 0.6373, precision of 0.6254, recall of 0.6498, and F1-Score of 0.6346. The performance of the LSTM model is relatively average, which may be due to its sequential processing characteristics, resulting in it may not be able to fully extract the complex semantic and syntactic relations in the math problem solutions.

Table 4. Performance on Mathematical dataset

Model	Accuracy	Precision	Recall	F1-Score
LSTM	0.6373	0.6254	0.6498	0.6346
Bi-LSTM	0.8235	0.8120	0.8350	0.8195
BiLSTM-AM	0.9052	0.9150	0.8950	0.9047

Compared with the LSTM model, the Bi-LSTM model has improved accuracy to 0.8235, precision to 0.8120, recall to 0.8350, and F1-Score to 0.8195. The bidirectional structure of Bi-LSTM enables it to capture both forward and backward contextual information. This enables the model to better understand the logical relationship of the mathematical solution and improve performance.

The BiLSTM-AM model outperforms the LSTM and Bi-LSTM models, achieving an accuracy of 0.9052, precision of 0.9150, recall of 0.8950, and an F1-Score of 0.9047. The BiLSTM-AM model integrates an attention mechanism that selectively focuses on different parts of the input sequence. This is very beneficial when dealing with mathematical solutions, as it can highlight the key steps and logical relationships in the problem-solving process.

In summary, the results demonstrate the effectiveness of the BiLSTM-AM model in automatic grading of elementary math solutions. The attention mechanism plays a key role in improving the performance of the model, enabling it to better capture important information in the solution. These findings provide a good research direction

for future research in the field of automatic grading of math problems, and may be applied to other related fields.

5 Conclusion

This study conducted a comprehensive evaluation of the BiLSTM-AM model for automatic scoring of essays and mathematics solutions. On the Learning Institution Laboratory–Essay Automatic Scoring 2.0 dataset, the BiLSTM-AM model is evaluated with a very high accuracy of 85.13%. The model also achieves an accuracy of 90.52% on the primary school geometry problem dataset. The results emphasize the strong performance of the BiLSTM-AM model compared to traditional LSTM and Bi-LSTM models, particularly in identifying complex contextual relationships and achieving high accuracy. This study highlights the efficiency of attention mechanisms in improving the performance of automatic scoring models, particularly in the evaluation of mathematical solutions. The results of this study provide valuable insights into the field of automatic scoring and have the potential to be applied to other fields, thereby promoting more efficient and accurate assessment of student performance.

Acknowledgments. This research is supported by the Natural Science Foundation of Hubei Province (2024AFD342), the traditional Chinese Medicine scientific research projects for 2025-2026 of Health Committee of Hubei Province (ZY2025M029), the Wuhan Key Research and Development Program Projects (2023010402010614), the Wuhan East Lake High-tech Development Zone 'Revealing the List and Taking the Challenge' Projects (2023KJB204) and the Educational Science Planning Projects for 2019 of Department of Education of Hubei Province (Research on the Subject Knowledge Q&A Service Model of Future Schools in the Context of Artificial Intelligence, Grant Number: 17). This research is supported by "the Fundamental Research Funds for the Central Universities" South-Central MinZu University (CZY23008).

References

1. Jin, C., He, B., Hui, K., Sun, L.: TDNN: A Two-stage Deep Neural Network for Prompt-independent Automated Essay Scoring. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1088–1097. Association for Computational Linguistics (2018)
2. Li, X., Chen, M., Nie, J.Y.: SEDNN: Shared and Enhanced Deep Neural Network Model for Cross-prompt Automated Essay Scoring. Knowledge-Based Systems 210, 106491 (2020)
3. Cao, Y., Jin, H., Wan, X., Yu, Z.: Domain-Adaptive Neural Automated Essay Scoring. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 1011–1020. ACM (2020)
4. Chen, Y., Li, X.: PMAES: Prompt-mapping Contrastive Learning for Cross-prompt Automated Essay Scoring. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1489–1503. Association for Computational Linguistics (2023)
5. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Multi-grained Attention Network for Aspect-level Sentiment Classification. In: Proceedings of the 2018 Conference on Empirical Methods in

Natural Language Processing, pp. 3433–3442. Association for Computational Linguistics (2018)
6. Yang, R., Cao, J., Wen, Z., Wu, Y., He, X.: Enhancing Automated Essay Scoring Performance via Fine-tuning Pre-trained Language Models with Combination of Regression and Ranking. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1560–1569. Association for Computational Linguistics (2020)
7. Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., Steinhardt, J.: Measuring Mathematical Problem Solving With the MATH Dataset. In: Proceedings of the 34th International Conference on Neural Information Processing, pp. 1–22 (2021)
8. Wang, C., Jiang, Z., Yin, Y., Cheng, Z., Ge, S., Gu, Q.: Aggregating Multiple Heuristic Signals as Supervision for Unsupervised Automated Essay Scoring. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 13999–14013. Association for Computational Linguistics (2023)
9. Uto, M.: Neural Automated Essay Scoring Incorporating Handcrafted Features. Journal of Natural Language Processing 28(2), 716–720 (2021)
10. Ramesh, D., Sanampudi, S.K.: An Automated Essay Scoring Systems: A Systematic Literature Review. Artificial Intelligence Review 55(3), 2495–2527 (2022)
11. Li, X., Yang, H., Hu, S., Geng, J., Lin, K., Li, Y.: Enhanced Hybrid Neural Network for Automated Essay Scoring. Expert Systems 39(10), e13068 (2022)
12. LaVoie, N., Parker, J., Legree, P.J., Ardison, S., Kilcullen, R.N.: Using Latent Semantic Analysis to Score Short Answer Constructed Responses: 19 Automated Scoring of the Consequences Test. Educational and Psychological Measurement 80(2), 399–414 (2020)
13. Latifi, S., Gierl, M.: Automated Scoring of Junior and Senior High Essays Using Coh-Metrix Features: Implications for Large-scale Language Testing. Language Testing 38(1), 62–85 (2021)
14. Shin, J., Gierl, M.J.: More Efficient Processes for Creating Automated Essay Scoring Frameworks: A Demonstration of Two Algorithms. Language Testing 38(2), 247–272 (2021)
15. Tang, Z., Zhang, X., Wang, B., Wei, F.: MathScale: Scaling Instruction Tuning for Mathematical Reasoning. In: Proceedings of Machine Learning Research, pp. 47885–47900 (2024)
16. Jie, Z., Li, J., Lu, W.: Learning to Reason Deductively: Math Word Problem Solving as Complex Relation Extraction. In: Proceedings of the Annual Meeting of the Association for Computational Linguistics, pp. 5944–5955. Association for Computational Linguistics (2022)
17. Jiang, Z., Liu, M., Yin, Y., Yu, H., Cheng, Z., Gu, Q.: Learning from Graph Propagation via Ordinal Distillation for One-Shot Automated Essay Scoring. In: Proceedings of the Web Conference 2021, pp. 2347–2356. ACM (2021)
18. Jiang, Z., Gao, T., Yin, Y., Liu, M., Yu, H., Cheng, Z., Gu, Q.: Improving Domain Generalization for Prompt-Aware Essay Scoring via Disentangled Representation Learning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 12456–12470. Association for Computational Linguistics (2023)
19. Li, X., Chen, Y.: PLAES: Prompt-generalized and Level-aware Learning Framework for Cross-prompt Automated Essay Scoring. In: 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, pp. 12775–12786. ELRA (2024)
20. Li, S., Ng, V.: Automated Essay Scoring: Recent Successes and Future Directions. In: Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, pp. 8114–8122. IJCAI (2024)

21. Xu, K., Zou, K., Huang, Y., Yu, X., Zhang, X.: Mining Community and Inferring Friendship in Mobile Social Networks. Neurocomputing 174, 605–616 (2015)
22. Liu, G., Guo, J.: Bidirectional LSTM with Attention Mechanism and Convolutional Layer for Text Classification. Neurocomputing 337, 325–338 (2019)
23. Liu, Y., Zhang, Y.: Attention-based Bidirectional Long Short-term Memory Networks for Relation Classification. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(11), 1983–1995 (2018)
24. Huang, Y., Sun, H., Xu, K., Lu, S., Wang, T., Zhang, X.: Corelate: Learning the Correlation in Multi-fold Relations for Knowledge Graph Embedding. Knowledge-Based Systems 213, 106601 (2021)
25. Lu, P., Qiu, L., Yu, W., Welleck, S., Chang, K.W.: A Survey of Deep Learning for Mathematical Reasoning. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 14605–14631. Association for Computational Linguistics (2023)
26. Westera, W., Dascalu, M., Kurvers, H., Ruseti, S., Trausan Matu, S.: Automated Essay Scoring in Applied Games: Reducing the Teacher Bandwidth Problem in Online Training. Computers & Education 123, 212–224 (2018)