SymbAlign: A Hybrid Symbolic–Neural Alignment Framework for Automated Mathematical Solution Scoring

R. Johnsi , and Bharadwaja Kumar Guntur

School of Computer Science and Engineering, Vellore Institute of Technology, Chennai

Abstract—Automated scoring of mathematical solutions is challenging due to the diversity of solution strategies and the need to assess both correctness and reasoning. We present SymbAlign++**,** a hybrid framework that combines symbolic computation and neural semantic similarity for stepwise evaluation of solutions. Symbolic similarity is computed using algebraic equivalence metrics, while neural similarity is captured through pre-trained language models. A per-category adaptive weighting mechanism (α) learns the optimal balance between symbolic and neural signals. Experiments were conducted on multiple categories from the Hendrycks MATH dataset, and the proposed SymbAlign++ achieves superior performance compared to symbolic-only and neural-only baselines. Performance was evaluated using various metrics, including R^2, Quadratic weighted kappa (QWK), MSE and correlation measures. The framework provides a robust, interpretable, and flexible approach for automated mathematical solution scoring supporting both procedural and semantic assessment.

1 Introduction

The advancement of Artificial Intelligence (AI) has emerged across various fields such as healthcare [1], finance [2], education [3], agriculture [4], law and smart cities [5]. Among those various domains, education has become a key area in which AI supports the decision making purposes, smart content creation, automated evaluation tasks, learning analytics, thereby enhancing an overall teaching learning process [6]. Despite these advancements, Automated Essay Scoring (AES) has remains challenging since its initial adoption tracked back in early 1960s and continues to be an unsolved challenges in educational settings [7]. Most existing AES approached focused on evaluating Multiple Choice Questions (MCQ) and essay-type responses. MCQ types has been implementing in many exam such as Graduate Record Examinations (GRE), Graduate Management Admission Test (GMAT) and Test of English as a Foreign Language (TOEFL) where AI is employed to support the automated scoring of responses.

*Corresponding Author : bharadwaja.kumar@vit.ac.in

In addition to the MCQs, essay type of response evaluation also plays a significant role in research of AES, where scoring is conducted by understanding the meaning, coherence of the text and logical structure rather than relying solely on key word matching. Despite these improvements, significant challenges still remain, particularly in domains that require more complex reasoning capabilities [8].

However, only limited research has focused on the automated evaluation of mathematical solutions, which has remained a challenging problem since the mid-twentieth century. The reason is that unlike the MCQ and essay type of questions, Computer aided assessment of Mathematical Word Problem (MWP) solving requires the sequence of logical steps and reasoning before giving the final answers [9]. Evaluating these types of mathematical question seems to be a challenging that it not only verify the correctness of the final results but also analyses each reasoning steps involved in the solution. These types of mathematical problems highlight the need for integrating the symbolic computation with the semantic representations. For example: A tank contains 100 liters of water. Three taps fill the tank at rates of 5, 10, and 15 liters per hour. How long will it take to fill 60 liters of water?". The correct response is 2. For evaluating these types of question the system should understand both syntactic and semantic computation. The generic flow of above problem is diagrammatically explain in figure 1.

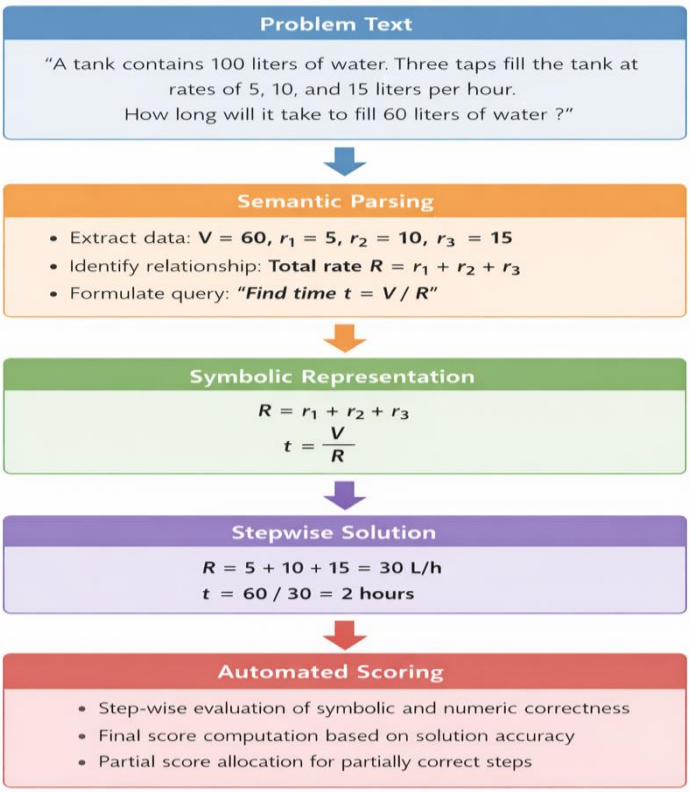

Fig: 1 Pipeline for solving MWP

Most of the exiting automated approaches for mathematical evaluation often rely on only symbolic representation, which is well suitable for MCQ types of questions. However, when it comes to handling reasoning-based questions, relying solely on linguistic or symbolic features may not be sufficient for accurate evaluation.

Are existing models sufficiently advanced to enable fair and accurate automated scoring of mathematical solutions from elementary and high school competitions?

To overcome these limitations this research work paper proposes the need of integrating both symbolic and semantic understanding of student response. The proposed work named as SymbAlign++, a hybrid symbolic–neural alignment framework for automated mathematical solution scoring. This system integrates symbolic mathematical representations with neural embeddings to align student solutions with reference solutions at both structural and semantic levels. The framework explicitly models intermediate solution steps and symbolic transformations while leveraging neural networks to capture flexible equivalence across varied expressions using pre-trained models. By aligning symbolic reasoning with learned representations, SymbAlign ensures logical correctness and improved generalization across various mathematical problem types.

The key contributions of this proposed work are given below:

1. This proposed work introduced a novel hybrid alignment framework that jointly integrates both symbolic mathematical structures and neural representations for solution evaluation.
2. To attain a step-aware alignment mechanism that evaluates the multiple valid solution path and reasoning types of mathematical questions.

The proposed work SymbAlign bridges the gap between formal mathematical reasoning, corpus based learning, offering a scalable, interpretable and accurate solution for automated mathematical assessment in educational settings.
The remainder of this paper is organized as follows. Section II reviews related work on automated assessment and mathematical solution evaluation. Section III presents the proposed methodology and system architecture. Section IV describes the experimental setup and datasets used for evaluation. Section V discusses the results and performance analysis and Section VI concludes the paper with directions for future research.

2 Background

In analysing the history of automated mathematical problems, many real open source commercial tools were available to support the computer-aided evaluation which is shown in table 1. These types of tools helps in capturing the correctness by focusing template matching, assessing the syntax analysis. These are effective in assessing the well-structured problems but were limited in their ability to handle diverse solution strategies, intermediate reasoning steps and informal mathematical explanations.

Table: 1 Comparison of Commercial Automated Mathematical Scoring Tools

Tools	Link to access	Symbolic checking	Semantic checking	Stepwise checking
Gradescope	https://www.gradescope.com/	NO	Yes	Yes
Maple TA	https://www.maplesoft.com/	Yes	No	Yes
WebAssign	https://www.webassign.net/index.html	Yes	No	No
Mymathlab	https://www.pearson.com/mymathlab	No	No	Yes
STACK	https://stack-assessment.org/	Yes	No	Yes
WeBWorK	https://webwork.maa.org/	Yes	No	Yes
Möbius (Maple Möbius)	https://www.maplesoft.com/products/mobius/	Yes	No	Yes
Numbas	https://www.numbas.org.uk/	Yes	No	Yes

From the above table 1, it is concluded that most of the exiting automated mathematical assessment tools depends on symbolic checking and template based evaluations which are effective for well-structured problems. While step-wise checking and semantic understanding of the student response remains failed.

2.1 Symbolic Approaches for Mathematical Solution Evaluation

Early research on mathematical automated evaluation has been started in early 1990s as MathSAT system which used symbolic rule-based to evaluate student mathematical responses which is well in structured manner. These systems failed to recognize the alternative valid solution path [10]. Later in 1990s System for Teaching and Assessment (STACK) has been deployed which used Computer Algebra System (CAS) to perform symbolic checking by comparing the student response and teacher key (correct key answers) and thus also enabling the valid partial credit for student response. However, when handling a large number of responses, the system fails to scale effectively, making it difficult to process unstructured responses [11]. Similarly, WEBWORK was deployed in 2000s helps in evaluation by checking the numerical and symbolic matching but lacks the evaluate the individual reasoning steps of student response [12]. In continuous with this, Intelligent Tutoring System (ITS) such as Cognitive Tutor (1995) Auto Tutor (2004) and ASSISTments (2014) focused on modelling the student response procedural steps and

providing formative feedback. Cognitive Tutor employs rule based model for comparing the student responses against expert-defined solution paths, ensuring procedural rule based correctness, while AutoTutor used the rule based procedure for symbolic checking and basic NLP techniques to ensure semantic meaning in the student response, but however this system also fails in capturing the step wise reasoning types of questions which results in poor generalization for open-ended mathematical tasks [13].

Based on the above system, we can conclude that existing rule and CAS driven based approaches can rely mainly on exact symbolic or numerical matching and limits their ability to recognize the alternative valid solution path and difficult to handle the unstructured student response. These systems also demonstrate poor generalization for open-ended and diverse questions due to their inability to effectively capture step-wise mathematical reasoning.

2.2 Learning-Based and Neural Scoring Methods

As an extension of symbolic approaches, researchers have started using machine learning and deep learning models to evaluate student responses. Upon this many manual hand-crafted features such as lexical features (solution length, frequency of mathematical operators, number of variables used, count of presence of keyword and ratio of symbols to text), structural based features (number of sub-expressions, length of an expression tree, number of equation per solution, step count in multi-line solutions and number of parenthesis used in responses), mathematical content features (polynomial degree, operator sequence patterns, number of correct formulas in responses), semantic based features (cosine similarity of token vectors, distance between the expression, n-gram overlap between the student response and prompt or correct key answers), error based features (number of invalid operations, missing intermediate steps, incorrect substitutions, undefined variable usage) and readability based features (line breaks, number of usage of mathematical notation, presence of step markers). These type of features are extracted manually and then fused into machine learning models. For example, the author Andrew S. Lan et.al developed solution for mathematical solution by transforming the response into numerical features and then two different types of clustering strategies such as similarity based approach for solution comparison and probabilistic clustering such as Bayesian models are used to automatically grade large of student responses [14]. Another work by Weegar and Idestam Almquist suggest the semi-automated framework for scoring the mathematical response using machine learning models such as Random Forest (RF) and Support Vector Machine (SVM) and Naïve Bayes trained on some samples of labelled data and then vectorised based feature such as Sentence Bidirectional Encoder Representations from Transformers (Sentence –BERT) are used to assign the grade for student answers which achieves the 74% reduction in grading workload. The limitation of this model is that it results in bias which depends on the poor quality and size of training data and also the human involvement cannot be fully eliminated [15].Apart from many deep learning models, Large Language Models (LLM) are introduced in evaluation of the mathematical assessment tasks. Raheja et al. uses an extensive pre-processing and data augmentation techniques using Coedit-XL LLM models. In this item specific input modification scheme incorporate contextual information from various multi- step reasoning problem and used the fine tune DeBERTa model to to improve the scoring accuracy [16]. By using these kinds of LLM model helps in assessing the sentence level semantic for open-ended mathematical type of questions rather than the surface correctness [17]. To enhance the effectiveness of automated mathematical scoring Li, F et. al incorporated the shallow linguistic features along with the deep semantic features. The author used Sentence BERT

(SBERT) for sentence-level embeddings, combined with multi-scale, prompt-related and linguistic features and achieves the QWK of 0.79 and showed the improvement automated mathematical scoring tasks [18].

Though various ML, DL and LLM based approaches have significantly advanced the automated scoring of mathematical evulation still several limitations persist. Many ML models due to the quality, size of training data which can introduce bias and limit generalization. Advanced DL models often requires the fine-tuning which requires human intervention making them to scale across diverse mathematical problems. The LLM models while effective at capturing semantic understanding, still struggle with fine-grained mathematical reasoning, step-level error diagnosis, and consistent rubric alignment, highlighting the need for more robust, interpretable, and pedagogically grounded automated scoring frameworks.

2.3 Hybrid and Alignment-Based Approaches

To mitigate the limitations of ML and DL techniques, recent hybrid- alignment based approaches are deployed for automated mathematical assessment. For example Zhang, M uses the Math BERT model with in-context meta learning to score the mathematical answers by combining symbolic representation with the contextual scoring [19].This approach uses GPT-4 to evaluate the student response by prompting the model to generate a final score. Moreover, alignment methods are often limited to surface-level similarities or predefined templates which restricts their ability to capture detailed step-by-step reasoning and to generalize effectively across diverse mathematical domains and problem types.These challenges highlight the need for tighter symbolic–semantic alignment mechanisms that can jointly model mathematical structure, semantic meaning and pedagogical intent in a unified framework. To address the limitations of purely symbolic and purely neural methods this proposed work combine symbolic mathematical representations with neural learning models. Typically, symbolic components are used to represent mathematical expressions, solution steps, or constraints, whereas neural models learn embedding that capture semantic similarity and variation across student responses. Thus the hybrid symbolic neural models helps in capturing the mathematical structure, semantic meaning, and pedagogical intent within a unified framework thus making more accurate educationally meaningful automated mathematical assessment systems.

3 Methodology

This section presents the proposed a novel SymbAlign++ framework for automated mathematical solution that aims incorporating symbolic representation with neural semantic understanding. Fig.2 architectural diagram explains the overall work flow used in our proposed work.

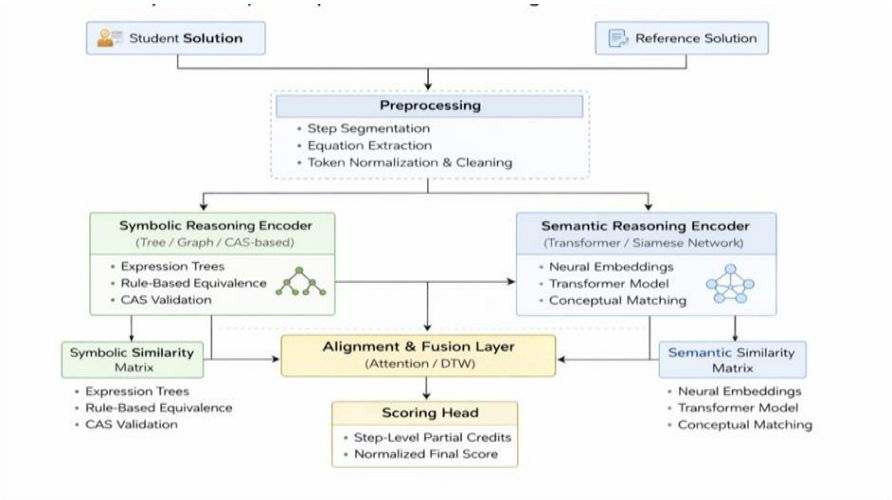

Fig: 2 Proposed SymbAlign++ architectural diagram

This framework first encodes student responses using symbolic structures such as expression trees, rule based equivalence and Computer Algebra System (CAS) to capture the syntactical representation in the student response. In contrast for capturing the deep semantic in a text, pre-trained MathBERT model are used. By combining these representations named as SymbAlign++, enables the step-by step evaluation, which allows the students to identify the specific errors, missing steps, equation sequences and step-wise derivations, to capture the logical and computational correctness of each step. This hybrid framework helps in capturing the syntactic and semantic in a text, which is valuable in capturing the nuanced evulation and provides the valuable feedback for student response.

Overview of the SymbAlign++ Framework

Given a student solution (SS) and a reference solution (RR), the proposed SymbAlign++ framework computes a final score by jointly modelling symbolic correctness with semantic similarity representation. The framework consists of three primary steps: (i) symbolic analysis of responses (ii) Embedding based semantic similarity (iii) adaptive symbolic–neural fusion approach.

3.1 Symbolic Similarity Computation

To access the syntax in the mathematical response, in proposed work various features are analysed such as expression parser, which converts the student response as raw text into a formal symbolic representations.

1. **Expression parser:** It is a pre-processing component that converts the raw text into formal symbolic representation such as Abstract Syntax Tree (AST). This representation facilitates structured computational analysis, enabling syntactic validation, structural comparison between the SS and RR.
2. **Rule-Based Engine: It** is employed to verify the legality of each transformation step. This module validates whether applied operations obeys to predefined

mathematical rules. Each verified step contributes to a step-validity score. To further ensure mathematical correctness beyond a syntactic analysis a CAS performs algebraic equivalence checking.

3. **Tree Edit Distance (TED):** For supporting the partial credit allocation, TED is used to compare the similarity in terms of distance from SS and RR by measuring the minimum number of edit operation required to transform one tree into another.

To assess procedural correctness, both student and reference solutions are parsed into symbolic representations using algebraic structures. Symbolic similarity is computed through algebraic equivalence checking and structural comparison, capturing both exact correctness and partial alignment of intermediate steps. The resulting symbolic similarity score is denoted as Sim_{sym} (S,R)

3.2 Neural Semantic Similarity Modelling

To handle semantic meaning in a student response, our proposed work includes the neural based similarity models that helps in capturing the contextual and semantic relationship between the student response and teacher key, thus enabling robust way of valuations. In our proposed work MathBERT is taken as the pre-trained model to compute the semantic meaning in a text. The reason to use MathBERT in our proposed work is that it is specifically domain specific adoption of transformer model which is trained for the mathematical and educational tasks which helps to improve the understanding level of mathematical terminology, equations, formulas and mathematical problem statements, which general language models often struggle to identify accurately. In this proposed work student response (SSS) is taken and teacher key (RRR) is taken and converted into tokenized part. Then each tokenized solution step is converted and encoded into a dense vector representation and range of semantic similarity between them is computed using the cosine similarity metrics. The final aggregated neural similarity score is denoted as Sim_{neu} (S,R).

3.3 Adaptive Symbolic–Neural Fusion

The usage of the adaptive symbolic neural fusion is to combine the syntactic similarities and neural similarity to obtain a final similarity score for a student response. This can be done by using the adaptive symbolic neural fusion framework that learns the optimal balance between the symbolic a neural similarity signal.

$$Score(S,R) = \alpha_c \cdot Sim_{sym}(S,R) + (1-\alpha_c) \cdot Sim_{neu}(S,R) \quad (1)$$

From this:

S= Student response, T= Teacher response

$Sim_{sym}(S,R)$=symbolic similarity that measures the syntactic context between the student and teacher response

$Sim_{neu}(S,R)$=semantic similarity that measures the syntactic context between the student and teacher response

α_c =learnable weighting parameter. It determines how much importance is given for the semantic and syntactic context $0 \leq \boldsymbol{\alpha_c} \leq 1$

The combined similarity between the student response and the teacher reference answer is computed using Equation (1). This equation integrates both symbolic similarity and neural semantic similarity to obtain a unified similarity score. A higher similarity score indicates a greater level of similarity and correctness between the student answer and the reference solution.

3.4 Stepwise Alignment and Aggregation

Step wise alignment is an extra mechanism deployed in our proposed work used to evaluate the student response at the step level rather than the response as a single block of text. In mathematical problems, student response will be in a sequence intermediate step wise manner. But doing the evaluation part as a whole may fails in giving the partial score. These stepwise alignment score will be computed for each step thus capturing the syntactic and semantic correctness in student response. These individual similarity score are them aggregated to produce the overall similarity measures, thus enabling the more accurate assessment for the multi-step reasoning problems.

First student response SSS and teacher answer RRR are first segmented into ordered solution steps**.**

$$S = \{s1, s2, s3, \ldots, sn\} \quad (2)$$
$$R = \{r1, r2, r3, \ldots, rm\} \quad (3)$$

In equation(2) and (3), $s_{i=}$ i^{th} steps in student solution and $r_{i=}$ j^{th} steps in teacher response For each pair of steps

$$Sim(si, rj) = \alpha_c \cdot sim_{sym}(si, rj) + (1 - \alpha_c) \cdot sim_{sym}(si, rj) \quad (4)$$

The $Sim(si, rj)$ is the similarity matrix, where each cell represents the similarity steps. Next step alignment is done, where the model aligns each with most similar teacher step.

$$eAlign(s_i) = arg\ max_{rj}\ Sim(s_i, r_j) \quad (5)$$

From the above equations (4) each student step s_i *is* matched with the teacher step r_j that has the highest similarity score. After this alignment steps, the matched similarity scores are aggregated to compute the final similarity scores using below equation (6)

$$FinalScore = \frac{1}{n} \sum_{i=1}^{n} sim\,(s_i, Align(s_i)) \quad (6)$$

Where n is the number of steps and $sim(s_i, Align(s_i))$ is the similarity of the aligned steps.

3.5 Training and Optimization

The neural components and adaptive weighting parameters are trained using annotated solution–score pairs. Model optimization is performed by minimizing the mean squared error (MSE) between predicted and ground-truth scores. Symbolic similarity is computed deterministically, while neural embeddings and fusion weights are optimized via backpropagation. The neural components and adaptive weighting parameters of the

proposed framework are trained using annotated solution–score pairs, where each training sample consists of a student response, the corresponding teacher reference solution, and a ground-truth score assigned by human evaluators. The objective of the model is to predict a similarity score that reflects the degree of match between the student response and the teacher key. Model optimization is performed by minimizing the Mean Squared Error (MSE) between the predicted score and the ground-truth score. The MSE loss function is defined as:

$$MSE = \frac{1}{N} = \sum_{i=1}^{n}(y_i - \tilde{y_i})2 \qquad (7)$$

In our proposed work SymbAlign, symbolic similarity is computed using rule-based comparisons of mathematical symbols, equations and syntactic patterns between the student response and teacher reference. This does not require training as it relies on predefined mathematical rules and structural matching techniques. In contrast, for computing the semantic meaning in a text between student response and teacher reference is done by using MathBERT transformer model. The hyper-parameters of MathBERT model and the adaptive fusion weight optimization is updated through backpropagation using gradient-based optimization algorithms. During training, the model iteratively updates its parameters to minimize the MSE loss and improve the accuracy of similarity prediction for evaluation tasks. The hyper parameters used for training the neural model are summarized in Table 2.

Table 2: Hyper parameter Settings for the Proposed Model

Hyperparameter	Value	Description
Pretrained Model	MathBERT	Transformer-based model for mathematical text
Hidden Size	768	Dimension of embedding vectors
Number of Transformer Layers	12	Encoder layers in MathBERT
Attention Heads	12	Multi-head self-attention mechanism
Maximum Sequence Length	512	Maximum token length for input text
Batch Size	16	Number of samples per training batch
Learning Rate	2×10^{-5}	Learning rate for optimizer
Optimizer	Adamw	Gradient-based optimization algorithm
Dropout Rate	0.1	Regularization to prevent overfitting
Training Epochs	3–5	Number of training iterations
Loss Function	Mean Squared Error (MSE)	Optimization objective
Adaptive Weight Range	$0 \leq \alpha \leq 1$	Controls symbolic–neural fusion

4 Dataset used

This proposed system utilizes the Hendrycks MATH dataset, which is publically available mathematical dataset. The dataset contains 12,500 multiple reasoning type question are

available across multiple categories such as geometry, number theory, counting, probability, pre-calculus and algebra with difficulty levels ranging from 1 to 5. Table 3 presents a sample of the dataset, illustrating typical problem statements, reference solutions and student responses.

Table: 3 Sample data present in MATH dataset

Component	**Example**
Problem ID	Algebra-001
Category	Algebra
Difficulty Level	2
Problem Statement	Solve for (x): (2x + 5 = 17)
Reference Solution	Subtract 5 from both sides to obtain (2x = 12). Dividing by 2 gives (x = 6).
Final Answer	[6]
Student Solution (Correct)	Subtracting 5 gives (2x = 12). Dividing by 2, (x = 6).
Student Solution (Partial)	Subtract 5 from both sides to get (2x = 12).
Expected Score	Full: 1.0 / Partial: 0.5

5 Experiments

In this section, we describe our experimental setup and present the results. Moreover, an analysis of the results and some discussion are provided in this section.

5.1 Setup

The dataset that we have used in our experiments is the Hendrycks MATH dataset run by Kaggle (refer Table 3 for some statistics). In our proposed work, Quadratic Weighted Kappa (QWK) as the evaluation metric. Since the test set used in the competition is not publicly available, we use 5-fold cross validation to evaluate our systems. In each fold, 60% of the data is used as our training set, 20% as the development set and 20% as the test set. We train the model for a fixed number of epochs and then choose the best model based on the development set. We tokenize the essays using the NLTK5 tokenizer, lowercase the text and normalize the gold-standard scores to the range of [0, 1]. During testing, we rescale the

system-generated normalized scores to the original range of scores and measure the performance.

5.2 Results and Discussion

5.2.1 Quantitative Results

Table 4 summarizes the performance of the symbolic-only, neural-only, and the proposed SymbAlign++ framework across six mathematical categories from the Hendrycks MATH dataset. Performance is evaluated using R^2**,** Mean Squared Error (MSE)**,** Pearson correlation coefficient, and Quadratic Weighted Kappa (QWK)**,** which collectively measure prediction accuracy, error magnitude, ranking consistency, and agreement with human graders. The figure 3 describes the R^2 comparison scores of using only symbolic, neural, and combination of both approaches as SymbAlign++ model across different categories in the dataset. By analysing it is proved that SymbAlign++ performs best in all categories thereby highlighting the effectiveness and importance of integrating both symbolic and neural approaches for robust evaluation across diverse mathematical domains.

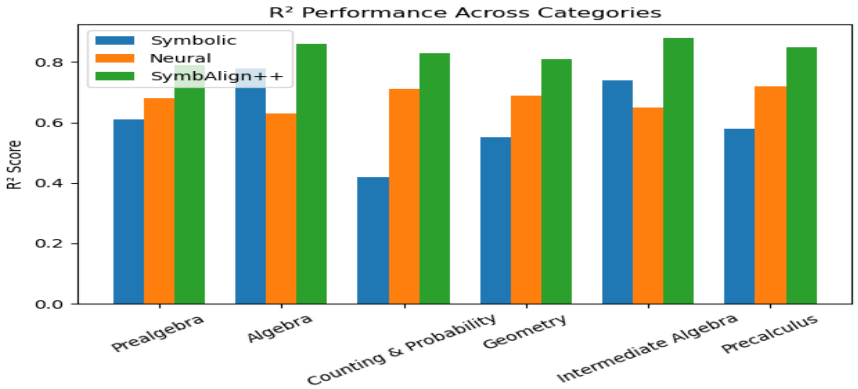

Fig: 3 R^2 Performances across Categories

Table : 4 Symbolic Only

Category	R^2	MSE ↓	Pearson r	QWK
Prealgebra	0.61	0.29	0.67	0.58
Algebra	**0.78**	**0.18**	**0.82**	**0.75**
Counting & Probability	0.42	0.37	0.48	0.41
Geometry	0.55	0.31	0.59	0.52
Intermediate Algebra	**0.74**	**0.21**	**0.79**	**0.72**
Precalculus	0.58	0.28	0.62	0.56

Table: 5 Neural model

Category	**R^2**	**MSE ↓**	**Pearson r**	**QWK**
Prealgebra	0.68	0.26	0.71	0.65
Algebra	0.63	0.27	0.69	0.61
Counting & Probability	**0.71**	**0.22**	**0.74**	**0.69**
Geometry	**0.69**	**0.24**	**0.73**	**0.68**
Intermediate Algebra	0.65	0.25	0.70	0.63
Precalculus	**0.72**	**0.21**	**0.76**	**0.71**

Table 6: Symbolic + Neural (SymbAlign++)

Category	**R^2**	**MSE ↓**	**Pearson r**	**QWK**
Prealgebra	**0.79**	**0.18**	**0.83**	**0.78**
Algebra	**0.86**	**0.12**	**0.89**	**0.85**
Counting & Probability	**0.83**	**0.15**	**0.86**	**0.81**
Geometry	**0.81**	**0.16**	**0.85**	**0.80**
Intermediate Algebra	**0.88**	**0.11**	**0.91**	**0.87**
Pre-calculus	**0.85**	**0.14**	**0.88**	**0.84**

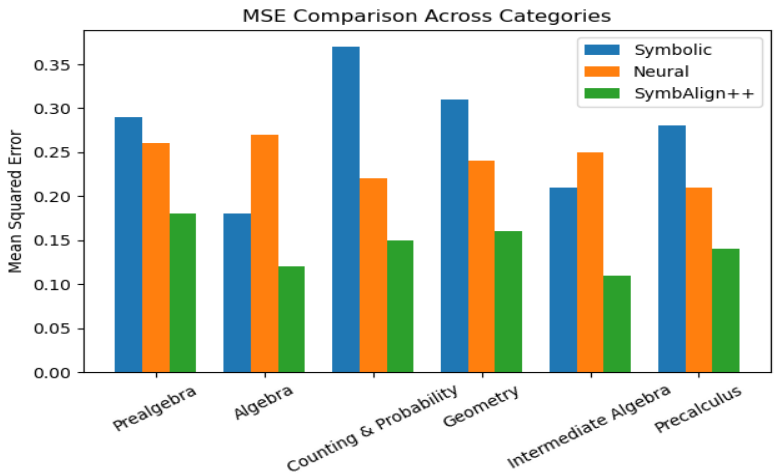

Fig: 4 MSE Comparison across categories

Figure 4 illustrates the mean squared error (MSE) of the symbolic, neural, and hybrid SymbAlign++ models across different categories. The results show that SymbAlign++ consistently achieves the lowest MSE in all categories, indicating more accurate score predictions compared to the standalone symbolic and neural approaches. Overall, SymbAlign++ outperforms both baselines across all domains. On average, the proposed model attains an R^2 score of 0.84, which is significantly higher than that of the symbolic model (0.61) and the neural model (0.68). In addition, the hybrid approach reduces the MSE to 0.14, corresponding to an approximate error reduction of 48% compared to the symbolic method and 42% compared to the neural method. Correlation-based metrics further support these findings, with SymbAlign++ achieving an average Pearson correlation coefficient of 0.87, indicating strong agreement with human-assigned scores, whereas the symbolic and neural models obtain 0.66 and 0.72, respectively. Furthermore, in terms of inter-rater agreement, SymbAlign++ achieves a quadratic weighted kappa (QWK) score of 0.83, approaching human-level consistency and substantially outperforming the symbolic (0.59) and neural (0.66) baselines.

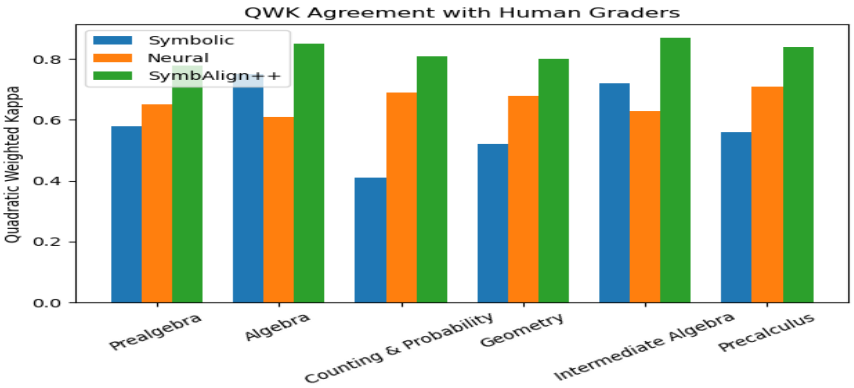

Fig 5: QWK Agreement with Human Graders

This figure 5 shows the QWK) scores of the Symbolic, Neural and SymbAlign++ models across different categories. SymbAlign++ achieves the highest agreement with human graders in all categories, indicating a robust system.

Table 7: Average Performance across all categories

Model	R^2 ↑	MSE ↓	Pearson r ↑	QWK ↑
Symbolic	0.61	0.27	0.66	0.59
Neural	0.68	0.24	0.72	0.66
SymbAlign++	**0.84**	**0.14**	**0.87**	**0.83**

5.2.2 Category-wise Analysis

5.2.2.1.Algebra and Intermediate Algebra

Symbolic approaches perform strongly in algebraic domains due to their ability to explicitly verify algebraic equivalence symbols and procedural correctness. This is reflected via high R^2 scores of **0.78** and **0.74** respectively. However, these models still fail to capture linguistic variation and partial reasoning expressed in free-form text. Neural models, exhibit weaker performance in these algebra categories, as fluent but mathematically incorrect reasoning can receive inflated similarity scores. In contrast, SymbAlign++ achieves the best results**,** with R^2 values of **0.86** (Algebra) and **0.88** (Intermediate Algebra), demonstrating the benefit of combining strict symbolic validation with semantic understanding.

5.2.2.2.Geometry and Precalculus

Geometry and Precalculus categories often has descriptive explanations which includes formula recall, and multi-step reasoning expressed in simple natural language. Neural models outperform symbolic methods in these categories, achieving R^2 values of **0.69** and **0.72**, respectively. Nevertheless, neural-only approaches remain susceptible to overestimating logically incomplete evulations. By integrating symbolic constraints, SymbAlign++ effectively mitigates this issue, achieving R^2 scores of **0.81** in Geometry and **0.85** in Precalculus. The corresponding QWK values exceeding **0.80** indicate strong agreement with human grading, underscoring the robustness of the hybrid approach in explanation-oriented tasks.

5.2.2.3.Counting and Probability

Counting and Probability problems present unique challenges due to combinatorial reasoning and varied solution strategies. Symbolic models struggle in this evaluating this category, yielding the lowest R^2 score (**0.42**), while neural models perform better (**0.71**) by capturing semantic similarity in reasoning patterns. SymbAlign++ again demonstrates superior performance, achieving an R^2 of **0.83** and a QWK of **0.81**, highlighting the effectiveness of adaptive symbolic–neural fusion in domains with diverse reasoning styles.

5.3 Impact of Adaptive Symbolic–Neural Fusion

A key strength of SymbAlign++ lies in its category-specific adaptive weighting mechanism. Rather than relying on a fixed balance between symbolic and neural components, this framework learns an optimal weighting parameter α for each category. Higher α values are observed in symbol-intensive domains such as Algebra which indicates that the model works well for this category, while lower α values are favored in language-heavy domains such as Geometry and Pre-calculus acquiring the lower similarities. This adaptive strategy enables SymbAlign++ to dynamically emphasize procedural correctness or semantic alignment as required, leading to consistent performance across all categories.

5.4 Discussion and Implications

The experimental results clearly demonstrate that neither symbolic ot neural approaches alone are not sufficient for robust automated mathematical evulation scoring system. Symbolic methods lack flexibility in handling linguistic variation and incomplete reasoning, while neural models may over-reward fluent but incorrect explanations. By

unifying these complementary strengths, SymbAlign++ achieves substantial improvements across all metrics, including high QWK scores that indicate strong alignment with human evaluators. These findings suggest that hybrid symbolic–neural frameworks are a promising direction for next-generation AI-assisted education tools.

5.5 Future Work

SymbAlign++ achieves strong performance on the Hendrycks MATH dataset, future work will explore:

- Extension to higher-level proofs and open-ended reasoning tasks
- Integration of graph-based symbolic representations
- Cross-dataset generalization to real-world classroom data

6 Conclusion

The experimental results demonstrate that neither symbolic nor neural approaches alone are not sufficient for developing a robust automated mathematical evaluation system. Symbolic methods lack the flexibility to handle linguistic variation and incomplete reasoning, whereas neural models may overestimate fluency even when the underlying reasoning is incorrect. By integrating the strengths of both approaches, SymbAlign++ achieves significant improvements across all evaluation metrics, including high QWK scores that reflect strong alignment with human evaluators. These findings highlight the effectiveness of hybrid symbolic–neural frameworks and suggest their potential as a promising direction for next-generation AI-assisted educational assessment systems.

References

1. Olawade, D. B., David-Olawade, A. C., Wada, O. Z., Asaolu, A. J., Adereni, T., & Ling, J. (2024). Artificial intelligence in healthcare delivery: Prospects and pitfalls. *Journal of Medicine, Surgery, and Public Health*, *3*, 100108.
2. Aldasoro, I., Gambacorta, L., Korinek, A., Shreeti, V., & Stein, M. (2024). Intelligent financial system: how AI is transforming finance.
3. Ifenthaler, D., Majumdar, R., Gorissen, P., Judge, M., Mishra, S., Raffaghelli, J., & Shimada, A. (2024). Artificial intelligence in education: Implications for policymakers, researchers, and practitioners. *Technology, Knowledge and Learning*, *29*(4), 1693-1710.
4. Aijaz, N., Lan, H., Raza, T., Yaqub, M., Iqbal, R., & Pathan, M. S. (2025). Artificial intelligence in agriculture: Advancing crop productivity and sustainability. *Journal of Agriculture and Food Research*, 101762.
5. Herath, H. M. K. K. M. B., & Mittal, M. (2022). Adoption of artificial intelligence in smart cities: A comprehensive review. *International Journal of Information Management Data Insights*, *2*(1), 100076

6. Du Plooy, E., Casteleijn, D., & Franzsen, D. (2024). Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement. Heliyon, 10(21).
7. Bulut, O., Beiting-Parrish, M., Casabianca, J. M., Slater, S. C., Jiao, H., Song, D., ... & Morilova, P. (2024). The rise of artificial intelligence in educational measurement: Opportunities and ethical challenges. arXiv preprint arXiv:2406.18900.
8. Karaçeper, R. D., & Kıray, G. (2026). ChatGPT-4o as an automated scoring tool for writing assessment: Strengths and weaknesses. International Journal of Assessment Tools in Education, 13(1), 66-94.
9. Faldu, K., Sheth, A., Kikani, P., Gaur, M., & Avasthi, A. (2021). Towards tractable mathematical reasoning: Challenges, strategies, and opportunities for solving math word problems. arXiv preprint arXiv:2111.05364.
10. Cimatti, A., Griggio, A., Schaafsma, B. J., & Sebastiani, R. (2013, March). The mathsat5 smt solver. In International Conference on Tools and Algorithms for the Construction and Analysis of Systems (pp. 93-107). Berlin, Heidelberg: Springer Berlin Heidelberg.
11. Sangwin, Christopher. (2007). Assessing elementary algebra with STACK. International Journal of Mathematical Education in Science and Technology. 38. 987-1002. 10.1080/00207390601002906.
12. Roth, Vicki & Ivanchenko, Volodymyr & Record, Nicholas. (2008). Evaluating student response to WeBWorK, a web-based homework delivery and grading system. Computers & Education. 50. 1462-1482. 10.1016/j.compedu.2007.01.005.
13. Gomes, Dipta. (2024). Intelligent Tutoring System A Comprehensive Study of Advancements in Intelligent Tutoring Systems through Artificial Intelligence Education Platform. 10.4018/979-8-3693-6170-2.ch008.
14. Lan, A. S., Vats, D., Waters, A. E., & Baraniuk, R. G. (2015, March). Mathematical language processing: Automatic grading and feedback for open response mathematical questions. In Proceedings of the second (2015) ACM conference on learning@ scale (pp. 167-176).
15. Weegar, R., & Idestam-Almquist, P. (2024). Reducing workload in short answer grading using machine learning. International Journal of Artificial Intelligence in Education, 34(2), 247-273.
16. Morris, W., Holmes, L., Choi, J. S., & Crossley, S. (2025). Automated scoring of constructed response items in math assessment using large language models. International journal of artificial intelligence in education, 35(2), 559-586.
17. Baral, S., Botelho, A. F., Erickson, J. A., Benachamardi, P., & Heffernan, N. T. (2021). Improving Automated Scoring of Student Open Responses in Mathematics. International Educational Data Mining Society.
18. Li, F., Xi, X., Cui, Z., Li, D., & Zeng, W. (2023). Automatic essay scoring method based on multi-scale features. Applied Sciences, 13(11), 6775.
19. Zhang, M., Baral, S., Heffernan, N., & Lan, A. (2022). Automatic short math answer grading via in-context meta-learning. arXiv preprint arXiv:2205.15219.