DOI: 10.54254/2753-7048/74/2024.BO17887

Research on the Accuracy of Machine Learning-Based AI Grading Systems in Handling High School Math Function Problems: A Comparative Study of MathGPTPro and Zuoyebang

Ziyue Gu[1,a,*]

[1]School of Education, Johns Hopkins University, Baltimore, MD 21211, The United States
a. zgu12@alumni.jh.edu
**corresponding author*

Abstract: As artificial intelligence (AI) technology becomes more prevalent in education, AI automatic grading systems have emerged as essential tools for enhancing homework grading efficiency and alleviating teachers' workloads. The two leading platforms, Zuoyebang and MathGPTPro, are widely utilized in mathematics education. This study employs an experimental research method to compare the performance of these platforms in automatically grading function solution problems. The sample consists of 30 questions from the Function Solution Problems section of China's National College Entrance Examination. The focus will be on four key dimensions: logical steps, final answers, expression symbols, and analysis feedback to assess accuracy rates. A t-test will then examine differences in their handling of complex solution steps. The results show that MathGPTPro can achieve higher accuracy in complex reasoning, and its AI system has more potential to be applied to homework correction and math learning. However, there are still problems of inaccurate identification and wrong judgment in the process of Zuoyebang step recognition. This study offers insights into the application of AI automatic grading systems in education and suggests areas for system optimization.

Keywords: AI automatic grading systems, Function problems, Zuoyebang, MathGPTPro, Accuracy.

1. Introduction

In recent years, the rapid development of educational technology worldwide has induced changes in teaching models, especially in K-12 education. In the teaching process, homework correction is a vital link. Zang, Cao, Zhou, and Zhang pointed out that when teachers correct a large amount of homework, they are easily affected by subjective factors such as fatigue, which will lead to a decrease in the efficiency and accuracy of the correction [1]. With the further integration of artificial intelligence (AI) into education, AI tools can enhance the student learning experience through personalized learning paths and real-time feedback, saving teachers and students a great deal of time [2]. However, despite the growing use of AI in education, the effectiveness of these tools in grading needs to be further explored [3]. Especially when dealing with complex subjective problems, the specific application effect still needs further discussion [4].

AI automatic grading systems generally fall into two broad categories: rule-based scoring systems and machine learning-based scoring systems. According to previous research, rule-based scoring systems rely on preset answer templates, are suitable for standardized questions, and perform poorly on complex mathematical subjective questions [5]. The scoring system based on machine learning can automatically generate assessment criteria by learning a large amount of homework data, analyzing students' problem-solving ideas, and dealing with complex problems with greater flexibility and accuracy [6]. Both MathGPTPro and Zuoyebang, are machine learning-based systems. The important concerns of this study are how effective they are in practical application and whether they can accurately identify error types.

This paper aims to assess the accuracy of machine learning-based AI scoring systems, MathGPTPro and Zuoyebang, in addressing high school mathematical function problems through experimental methodologies. Specific research inquiries include:

1. How accurate are the scores generated by MathGPTPro and Zuoyebang when addressing high school math function-solving problems? Which system exhibits higher accuracy? Can students' problem-solving steps be accurately assessed? Is there a discernible difference?

2. What are AI scoring systems' potential benefits and limitations in practical teaching scenarios?

This investigation not only provides valuable insights for future educational technology research and development, but also offers practical guidance for mathematics educators to effectively utilize AI tools in classroom instruction, while presenting enhanced ideas for software development designers.

2. Research Methodology

2.1. Sample Selection

2.1.1. Scoring system selection

MathGPTPro can handle relatively complex math topics using natural language processing and deep learning techniques [7]. Research has shown that it can provide accurate scoring and instant feedback when dealing with logical reasoning and multi-step problem solving, and is a highly intelligent AI software focused on assessing math problems [8]. Zuoyebang is widely used in China's basic education. It is the only software in China that can do homework correction for all subjects in high school, with relatively mature AI technology and a large stock of questions [9]. Both AI systems support identifying problem-solving steps, final answers, ensuring sufficient grading dimensions for analysis and comparison.

Both softwares are based on machine learning scoring algorithms, and when dealing with subjective questions, both rely on big data training and deep learning models to make scoring decisions, and build up the ability to recognize solving patterns by training a massive math question bank. MathGPTPro scores questions based on the reasonableness and logic of the solving steps. Zuoyebang scores by analyzing students' solution steps and comparing answers. They both use NLP technology to parse students' problem-solving processes. The similarity of core algorithms ensures the comparability of the system's scoring logic.

2.1.2. Selection of Question Samples

The experimental method will be used to collect the national one-volume college entrance examination papers of the past fifteen years and select the function answer questions from them to ensure that the sample is representative. Thirty questions will be collected to ensure the statistical significance of the results.

For selecting function questions, it is more suitable for assessing AI's logical and analytical ability in problem-solving, for example, a question in which students need to use multiple combinations of knowledge points such as the definition and value domains of a function, monotonicity of a function and so on. Function questions have multiple steps and go on to assess the AI scoring system's recognition of steps. In addition, the questions involve many symbols, which can add dimension to assess the AI system. Function questions are also mandatory for many questions in the entrance exam every year.

2.2. Experimental Variables

- Independent variables: AI scoring systems (MathGPTPro and Zuoyebang), function problems.
- Dependent variables: Scoring accuracy (accuracy, precision, recall, F1 score)

2.3. Data Collection and Processing

First, the confusion matrix was utilized to record the scoring results of each question in the two AI systems. The confusion matrix includes:

- TP: recognized correct answer
- TN: recognized wrong answer
- FP: the number of times the wrong answer was judged by the correct answer
- FN: the number of times the correct answer was judged as the wrong answer

Secondly, according to the high school mathematics scoring rules, four scoring dimensions, namely, logic and steps, final answer correctness, expression and notation, and error analysis and feedback, were selected to be recorded for each question, as Table 1 shows.

Table 1: On the development of accuracy scoring rules.

Scoring Dimensions	Weighting	Description
Logic and Steps	40%	Evaluate the logic and steps of the problem-solving process. According to the grading rules of the college entrance examination, the process must be complete and the result correct to get full marks. Even if the result is correct, the missing steps need to be deducted accordingly, so the steps are an important factor in grading. This dimension is set at 40%.
Correctness of final answer	30%	Evaluate the correctness of the answer the student ultimately arrives. According to the grading rules of the high school exam, the process of result pairs earns full marks, and the result undertakes the process of solving the problem and is an important factor in marking. The result follows the problem-solving process and is an essential factor in grading. Therefore, this dimension is set at 30%.
Expression and notation	20%	Evaluate the standardization and correctness of mathematical expressions and symbols used by the students in their answers. According to the marking scheme of the GCE, marks will be deducted for incorrect use of symbols, and inaccurate or missing expressions or omissions will result in the deduction of a certain number of marks. Therefore, the dimension is set at 20%.

Table 1: (continued).

Error analysis and feedback	10%	Evaluate the system's ability to recognize students' errors and provide feedback. In studying the potential and limitations of the two potential and limitations of the two systems, this dimension was able to provide good feedback and play a role in the interpretation of the results. Feedback and plays a role in interpreting the results. Therefore, it was set at 10%.

The scoring of the AI system was recorded in detail based on the four scoring dimensions mentioned above. A scale of 1-5 is set for each dimension and scores are given for each dimension, as Table 2 shows.

Table 2: Scoring rules for each dimension.

Scoring Dimensions	Scoring	Scoring Criteria
Logic and Steps	5 points	Steps are complete and logical, each step is no different from the standard answer.
	4 points	Steps are mostly complete, logic is clear, very few flaws that do not affect the result.
	3 points	Some unclear logic or missing steps, but not seriously affecting the final result.
	2 points	The solution is incoherent, with obvious errors in logic or some steps missing that affect the reasonableness of the solution.
	1 points	The steps in the solution are confusing, most of the reasoning is incorrect, and there are serious errors in logic.
Correctness of Final answer	5 points	The final answer is completely correct and fully consistent with the standard answer.
	4 points	The final answer is generally correct, but there are minor symbolic or numerical errors that do not affect the main solution process.
	3 points	Answer is partially correct but shows some correctness in the solution.
	2 points	The final answer is mostly incorrect, showing only a tendency to approach the correct result.
	1 points	The answer is colored completely incorrectly and is completely inconsistent with the standard answer.
Expression and Notation	5 points	Very clear expression, correct and standardized use of symbols, terminology and formulas, no errors.
	4 points	Somewhat clear expression, symbols and terminology are generally correct, with very few flaws in presentation that do not affect understanding.
	3 points	Some ambiguity in expression, some errors in symbols or terminology, but most of the content can still be understood.
	2 points	The use of symbols is more confusing, the expression is unclear, there are more errors or omissions in symbols or terminology.
	1 points	Expression is very confusing with incorrect use of symbols and a large number of missing terms.

Table 2: (continued).

Error analysis and Feedback	5 points	The system accurately recognizes all errors and provides detailed feedback that helps students understand and improve.
	4 points	The system recognizes most errors and the feedback is highly relevant, but some of the feedback is slightly simplistic.
	3 points	The system recognizes some major errors, but some errors are not recognized and the feedback is more basic.
	2 points	The system is weak in recognizing errors, only recognizing some critical errors, and the feedback is general.
	1 points	The system fails to recognize errors, or there is no feedback.

2.4. Data Analysis

2.4.1. Accuracy Analysis

First, the scores of the four scoring dimensions for each question were weighted by the weighting coefficients to produce a weighted total score that reflects the importance of different dimensions and avoids scoring bias. Further, the accuracy details such as correct identification, incorrect identification, and omission are analyzed using a confusion matrix.

The following key metrics are calculated through the confusion matrix:

- Accuracy: the overall correctness of the system's scoring.
- Precision: The percentage of scores predicted by the system to be "excellent answers."
- Recall: The proportion of excellent answers correctly identified by the system.
- F1 Score: The average of the precision rate and the recall rate.

2.4.2. Analysis of Variance

A t-test was used to compare the difference in performance between MathGPTPro and Zuoyebang in terms of scoring accuracy. This is done as follows:

- Hypothesis formulation:

Accuracy Hypothesis: H0 (Original Hypothesis): There is no significant difference in scoring accuracy between MathGPTPro and Zuoyebang.

H1 (Alternative Hypothesis): The scoring accuracy of MathGPTPro is significantly higher than that of Zuoyebang.

- Calculate the T-test:

First, calculate each AI system's average weighted mean over multiple topics.

Second, the standard deviation of the scoring accuracy of each AI system is calculated to measure the volatility of the accuracy.

Then, the t-test formula was used, and the p-value was calculated. When the p-value was less than 0.05, it was considered that there was a significant difference between the two groups.

3. Results

3.1. Results of Weighted Scoring

According to the criteria developed in the methodology, the formula for calculating the total weighted score is as follows: total weighted score = (Logic and Steps Score × 0.40) + (Correctness of Final

Answer Score × 0.30) + (Expression and Symbol Score × 0.20) + (Error Analysis and Feedback Score × 0.10).

The weighted total scores of MathGPTPro and Zuoyebang Learning Machine after testing on 30 subjective math questions are shown in Table 3 below:

Table 3: Calculation Result Chart.

Systems	Average weighted total score	Standard deviation
MathGPTPro	4.12	0.76
Zuoyebang	3.39	1.04

Based on the above data, the t-test calculates a t-value of:

$$t = \frac{4.12-3.39}{\sqrt{\frac{0.76^2}{30}+\frac{1.04^2}{30}}} = 3.11 \tag{1}$$

The corresponding p-value of 0.0025 was obtained, and since the p-value is less than 0.05, we reject the H0 and accept the H1. That is, MathGPTPro is significantly more accurate in scoring than Zuoyebang. MathGPTPro is significantly higher than Zuoyebang Learning Machine in terms of the average weighted total score, which indicates that the overall score of MathGPTPro is better than that of Zuoyebang Learning Machine in several dimensions, such as Logic and Steps, Correctness of the Final Answer, Expression and Symbolism, and Error Analysis and Feedback.

3.2. Accuracy Analysis

Based on the average criteria shown in the data, a weighted score of ≥ 4.0 was set as an excellent answer, a weighted total score between 3.0 and 4.0 as a qualified answer, and a weighted total score of < 3.0 as a failed answer. The confusion matrix of the two systems can be calculated and compared. The confusion matrix for MathGPTpro is shown in Table 4, and the confusion matrix for the Zuoyebang is shown in Table 5.

Table 4: MathGPTPro's confusion matrix.

MathGPTPro	Predicting excellent solutions (weighted total score ⩾ 4.0)	Predicted Failed Solutions (weighted total score < 4.0)
Actual Excellent Solution	21(TP)	3(FP)
Actual Failed Solutions	2(FN)	4(TN)

Table 5: Zuoyebang's confusion matrix.

Zuoyebang	Predicting excellent solutions (weighted total score ⩾ 4.0)	Predicted Failed Solutions (weighted total score < 4.0)
Actual Excellent Solution	17(TP)	5(FP)
Actual Failed Solutions	4(FN)	4(TN)

Further calculating the values of the MathGPTPro.

Through the formula:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{2}$$

The accuracy is equal to 83.33%.

Through the formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{3}$$

The precision is equal to 87.5%.
Through the formula:

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{4}$$

The recall is equal to 91.3%.
Through the formula:

$$\text{F1} = 2 \times \frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{5}$$

The F1 Score is equal to 89.36%.

The values for each of the Zuoyebang, calculated by the above formula, were 70.0% accuracy, 77.27% precision, 80.95% recall, and 79.07% F1 score. The results show the overall scoring correctness of the MathGPTPro system. A higher percentage of predicted and actual excellent answers were correctly recognized by the system, and the reconciled mean of precision and recall was higher. This indicates that it performs better than the Zuoyebang learning machine in terms of accuracy, precision, recall, and F1 score.

3.3. Feedback Quality Analysis

Through data analysis, combined with weighted scoring and accuracy analysis results, MathGPTPro outperforms the Zuoyebang learning machine in handling subjective high school math problems. Error and feedback data indicate that MathGPTPro provides more specific and constructive feedback, effectively assisting students in correcting thinking errors and incorrect expressions during problem-solving processes. Conversely, Zuoyeabng learning machine's feedback is more standardized, accurately displaying final results without analyzing process steps or providing overall result feedback.

4. Discussion

The average weighted total score of MathGPTPro and Zuoyebang exceeds 4.0, while the average weighted total score of Zuoyebang is lower than 4.0, when dealing with function Problems of Standard Difficulty from the National College Entrance Examination (Gaokao) Mathematics Paper. This shows that MathGPTPro is highly able to identify correct answers and give accurate scores, but the applicability of the Zuoyebang needs further consideration.

MathGPTPro also has a definite advantage over Zuoyebang in dealing with complex problems. The reason may be that MathGPTPro uses more advanced natural language processing and deep learning techniques to understand better and analyze the student's problem-solving process.

5. Conclusion

MathGPTPro has great potential in practical teaching, especially in reducing the burden on teachers and providing timely correction to students. However, there may still be misjudgment or insufficient feedback on some highly subjective topics and require high detail control. Although AI systems can efficiently handle most mathematical subjective problems, whether they can effectively replace manual feedback in the face of open-ended problems needs further consideration. AI may have specific limitations in actual teaching. For example, the analysis of Zuoyebang data showed that some AI systems still could not fully understand the complex solution process, and could not properly deal with inconsistencies in symbols and expressions and give targeted feedback. AI flexibility must be enhanced, especially when teacher experience and judgment are required.

Based on the overall study, the author can analyze its limitations and future direction. This research focuses on function questions, especially Chinese college entrance examination questions. The application of the findings needs to be extended to other areas of mathematics or non-Chinese education systems. Comprehensively evaluate other question types such as inequalities, probabilities. In addition, the scoring process of the AI scoring system needs to be more transparent, and the reasoning behind the system's scoring needs explanation. Future research should explore ways to make the scoring process more transparent and explainable. Longitudinal studies on students' learning effects and teachers' experience of using AI can be studied, and the help of AI feedback to students' problem-solving can be investigated from students' perspective. From teachers' perspective, explore how to combine AI with manual labor better to bring more effective grading results.

References

[1] Zang, N., Cao, H., Zhou, N., & Zhang, X. (2022). Job load, job stress, and job exhaustion among Chinese junior middle school teachers: Job satisfaction as a mediator and teacher’s role as a moderator. Social Psychology of Education, 25, 1003–1030.

[2] Chassignol, M., Khoroshavin, A., Klimova, A., & Bilyatdinova, A. (2018). Artificial intelligence trends in education: A narrative overview. Procedia Computer Science, 136, 16-24.

[3] Van Vaerenbergh, S., & Pérez-Suay, A. (2022). A classification of artificial intelligence systems for mathematics education. Mathematics Education in the Digital Era, 89–106.

[4] Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence unleashed: An argument for AI in education. Pearson Education.

[5] Holmes, W., Bialik, M., & Fadel, C. (2019). Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign.

[6] Kulkarni, C., Bernstein, M., & Klemmer, S. R. (2013). Peer and self assessment in massive online classes. ACM Transactions on Computer-Human Interaction (TOCHI), 20(6), 1-31.

[7] Zhou, S. (2024). Zuoyebang: Pioneering the Future of Intelligent Education Hardware. Advances in Business Strategy and Competitive Advantage, 433–448.

[8] Shute, V. J., & Rahimi, S. (2017). Review of computer-based assessment for learning in elementary and secondary education. Journal of Educational Psychology, 109(4), 605-620.

[9] Patel, A., Bhattamishra, S., & Goyal, N. (2021). Are NLP models really able to solve simple math word problems? Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.