MathMistake Checker: A Comprehensive Demonstration for Step-by-Step Math Problem Mistake Finding by Prompt-Guided LLMs

Tianyang Zhang[1]*, Zhuoxuan Jiang[2]*, Haotian Zhang[1]*, Lin Lin[3], Shaohua Zhang[2]
[1]Learnable.ai, Shanghai, China
[2]Shanghai Business School, Shanghai, China
[3]UCloud Technology Co. Ltd., Shanghai, China
{tianyang.zhang,haotian.zhang}@learnable.ai, {jzx,zhangsh}@sbs.edu.cn, lesley.lin@ucloud.cn

Abstract

We propose a novel system, MathMistake Checker, designed to automate step-by-step mistake finding in mathematical problems with lengthy answers through a two-stage process. The system aims to simplify grading, increase efficiency, and enhance learning experiences from a pedagogical perspective. It integrates advanced technologies, including computer vision and the chain-of-thought capabilities of the latest large language models (LLMs). Our system supports open-ended grading without reference answers and promotes personalized learning by providing targeted feedback. We demonstrate its effectiveness across various types of math problems, such as calculation and word problems.

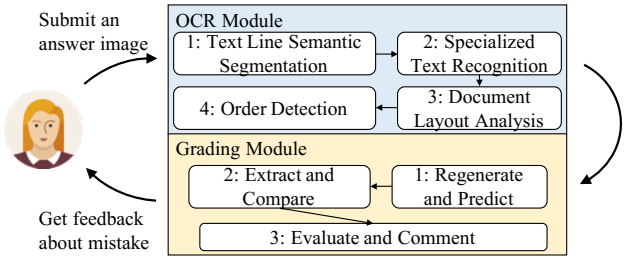

Figure 1: Architecture of MathMistake Checker.

Introduction

In mathematics education, accurately assessing students' problem-solving methods has long been a critical challenge, particularly for word problems that require step-by-step verification of a student's answer (Black and Wiliam 1998; Yavuz, Çelik, and Yavaş Çelik 2024). Unlike multiple-choice or fill-in-the-blank questions that demand only brief responses, effectively assessing the reasoning between adjacent steps in longer answers is pedagogically significant, as it measures student understanding, informs instruction, and provides meaningful feedback (Hattie and Timperley 2007).

Checking each step of student answers is a challenging task. Traditionally, it has been time-consuming for human teachers to read every answer step when grading student homework or examinations. Reference answers are often helpful for teachers to quickly identify mistakes and expedite the grading process. Consequently, some automated grading systems have been developed to verify student answers against a reference answer.

However, traditional grading systems have inherent limitations. They typically depend on predefined reference answers, which restrict their ability to recognize and evaluate the diverse approaches students may employ. This can result in a narrow interpretation of student competence, overlooking alternative methods and mathematically valid creative solutions. As educational practices shift toward promoting higher-order thinking and problem-solving skills, there is an increasing need for grading systems that can manage this complexity, providing more detailed and flexible assessments that align with the various ways students approach problems (Black and Wiliam 2009; Popham 2009).

To develop a grading system that operates without reference answers, two essential capabilities are required: Optical Character Recognition (OCR) and decision-making for mistake finding. The former handles the interpretation of image-based inputs, while the latter focuses on grading the answers. With the recent advancements in large models, Large Vision-Language Models (LVLMs) (Adewumi et al. 2024) can be utilized for end-to-end grading of mathematical answers. However, current LVLMs primarily emphasize visual understanding, often overlooking the reasoning aspect of natural language, particularly when mathematical texts demand more precise multi-modal comprehension. In this paper, we propose a two-stage approach to enhance interpretability and facilitate model training. As shown in Figure 1, the OCR Module in Stage 1 involves transforming image-based inputs into textual information, while the Grading Module in Stage 2 focuses on step-by-step mistake finding.

Specifically, we propose a demo system called MathMistake Checker, which integrates several cutting-edge technologies. In Stage 1, the system focuses on the precise extraction and correction of mathematical content from images submitted by users. This stage includes a pipeline with several phases: text line semantic segmentation (Chen et al. 2018; Xie et al. 2021), specialized text recognition (Shi, Bai, and Yao 2016; Li et al. 2023), document layout analysis (Huang et al. 2022), and order detection (Vinyals, For-

*The authors contributed equally to this work. Zhuoxuan Jiang is the corresponding author.

tunato, and Jaitly 2015). In Stage 2, we utilize the latest Large Language Models (LLMs) to serve as the step-by-step grader for the recognized text. Building on existing research that highlights LLMs' reasoning capabilities (Ishida et al. 2024; Lundgren 2024; Xie et al. 2024), we specifically incorporate the Pedagogical Chain-of-Thought (PedCoT) prompting strategy (Jiang et al. 2024), which effectively identifies logical mistakes in students' answers step-by-step. Notably, our system can support a variety of LLMs and prompting strategies.

Our demo system provides detailed feedback on the student's steps by systematically analyzing their problem-solving approach, identifying key steps and potential missteps, and providing targeted feedback that addresses specific areas of misunderstanding or mistake. It is designed to enhance pedagogical significance by aligning with the student's thought process and providing feedback that is both relevant and constructive.

System Overview

The proposed two-stage MathMistake Checker comprises two primary modules: OCR module and grade module.

OCR Module

The OCR Module employs a multiphase pipeline to process handwritten mathematical contents.

Text Line Semantic Segmentation This phase receives raw images with printed questions and handwritten answers, using semantic segmentation models (Chen et al. 2018; Xie et al. 2021) to separate text lines into printed text, handwritten text, and equations. This segmentation enhances accuracy in subsequent steps.

Specialized Text Recognition During this phase, each segmented text line is processed individually, with the appropriate model selected based on content type. Vision encoder-decoder transformer-based models (Li et al. 2023) are used for handwritten text and equations, while a convolutional recurrent neural network (Shi, Bai, and Yao 2016) handles printed text. This multi-model approach ensures high throughput and precision across diverse text formats.

Document Layout Analysis After recognizing text lines, a multi-modal transformer (Huang et al. 2022) is used for document layout analysis to understand spatial and logical relationships. This step reconstructs structures like tables and diagrams, preserving their original format.

Order Detection This phase is for reconstructing the original writing sequence of the handwritten content. By utilizing a Pointer Network (Vinyals, Fortunato, and Jaitly 2015), the correct order of text lines based on positional data can be determined, particularly in free-form layouts. This ensures an accurate logical step flow of the student's responses.

Grading Module

The Grading Module functions in three phases to evaluate a student's answer by using LLMs. For details on prompt strategies, refer to the original papers (Jiang et al. 2024).

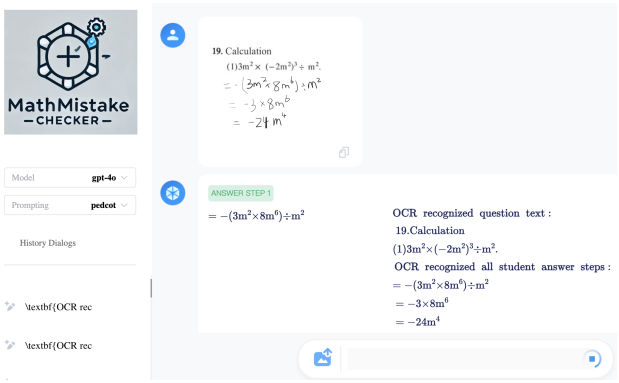

Figure 2: Demo interface for grading the math answer.

Regenerate and Predict The system prompts the selected LLM with the problem statement and the initial steps of the student's solution using a specified prompting strategy. The LLM then generates the expected mathematical concepts, problem-solving approaches, and calculations for the next step without accessing the student's actual response.

Extract and Compare In this phase, the LLM receives the student's actual response. It extracts the key elements of the answer and compares them with the predictions from previous phase. The system labels any discrepancies as mistakes, categorizing them according to predefined criteria such as correctness and alignment.

Evaluate and Comment In this phase, the system prompts the LLM to evaluate the student's reasoning up to the current step. The LLM provides a concise assessment, indicating whether the current step is correct, incorrect, or partially correct, and offers targeted feedback on the student's approach.

Case Studies

We apply our demo system to several cases. Figure 2 illustrates the demo interface for grading the answer to a math calculation problem. The left section of drop-down boxes provides options for various LLMs and prompting strategies, while the right section features a dialog interaction window that displays the step-by-step grading process. For more detailed information about the demos, please refer to the video.

Conclusion

In this paper, we present MathMistake Checker, a demo system that automates the grading of mathematical problems by integrating advanced technologies such as OCR and the chain-of-thought approach in the latest large language models (LLMs). The system delivers accurate and open-ended grading while supporting personalized learning by providing targeted feedback. MathMistake Checker streamlines the grading process and enhances learning experiences from a pedagogical perspective. Future work will focus on extending the system's capabilities to other subjects and improving the quality of feedback.

Acknowledgments

This work is partially supported by International Science and Technology Cooperation Program of Shanghai (No. 24170790602). We thank all the anonymous reviewers for their insightful and constructive comments.

References

Adewumi, T.; Alkhaled, L.; Gurung, N.; van Boven, G.; and Pagliai, I. 2024. Fairness and bias in multimodal ai: A survey. *arXiv preprint arXiv:2406.19097*.

Black, P.; and Wiliam, D. 1998. Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, 5(1): 7–74.

Black, P.; and Wiliam, D. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of personnel evaluation in education)*, 21: 5–31.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Hattie, J.; and Timperley, H. 2007. The power of feedback. *Review of educational research*, 77(1): 81–112.

Huang, Y.; Lv, T.; Cui, L.; Lu, Y.; and Wei, F. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4083–4091.

Ishida, T.; Liu, T.; Wang, H.; and Cheung, W. K. 2024. Large Language Models as Partners in Student Essay Evaluation. *arXiv preprint arXiv:2405.18632*.

Jiang, Z.; Peng, H.; Feng, S.; Li, F.; and Li, D. 2024. LLMs can Find Mathematical Reasoning Mistakes by Pedagogical Chain-of-Thought. *arXiv preprint arXiv:2405.06705*.

Li, M.; Lv, T.; Chen, J.; Cui, L.; Lu, Y.; Florencio, D.; Zhang, C.; Li, Z.; and Wei, F. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 13094–13102.

Lundgren, M. 2024. Large Language Models in Student Assessment: Comparing ChatGPT and Human Graders. *arXiv preprint arXiv:2406.16510*.

Popham, W. J. 2009. Assessment literacy for teachers: Faddish or fundamental? *Theory into practice*, 48(1): 4–11.

Shi, B.; Bai, X.; and Yao, C. 2016. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE transactions on pattern analysis and machine intelligence*, 39(11): 2298–2304.

Vinyals, O.; Fortunato, M.; and Jaitly, N. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34: 12077–12090.

Xie, W.; Niu, J.; Xue, C. J.; and Guan, N. 2024. Grade Like a Human: Rethinking Automated Assessment with Large Language Models. *arXiv preprint arXiv:2405.19694*.

Yavuz, F.; Çelik, Ö.; and Yavaş Çelik, G. 2024. Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments. *British Journal of Educational Technology*.