MATHAGENT: Leveraging a Mixture-of-Math-Agent Framework for Real-World Multimodal Mathematical Error Detection

Yibo Yan[1,2,3], **Shen Wang**[1], **Jiahao Huo**[2], **Philip S. Yu**[4], **Xuming Hu**[2,3,†], **Qingsong Wen**[1,†]
[1] Squirrel Ai Learning, [2] The Hong Kong University of Science and Technology (Guangzhou),
[3] The Hong Kong University of Science and Technology, [4] University of Illinois at Chicago
{yanyibo70, qingsongedu}@gmail.com, xuminghu@hkust-gz.edu.cn

Abstract

Mathematical error detection in educational settings presents a significant challenge for Multimodal Large Language Models (MLLMs), requiring a sophisticated understanding of both visual and textual mathematical content along with complex reasoning capabilities. Though effective in mathematical problem-solving, MLLMs often *struggle with the nuanced task of identifying and categorizing student errors in multimodal mathematical contexts*. Therefore, we introduce MATHAGENT, a novel Mixture-of-Math-Agent framework designed specifically to address these challenges. Our approach decomposes error detection into three phases, each handled by a specialized agent: an image-text consistency validator, a visual semantic interpreter, and an integrative error analyzer. This architecture enables more accurate processing of mathematical content by explicitly modeling relationships between multimodal problems and student solution steps. We evaluate MATHAGENT on real-world educational data, demonstrating approximately 5% higher accuracy in error step identification and 3% improvement in error categorization compared to baseline models. Besides, MATHAGENT has been successfully deployed in an educational platform that has served over one million K-12 students, achieving nearly 90% student satisfaction while generating significant cost savings by reducing manual error detection.

1 Introduction

Multimodal Large Language Models (MLLMs) have revolutionized the landscape of artificial intelligence by enabling the integration and understanding of diverse data formats (Wu et al., 2023a; Xie et al., 2024; Yan et al., 2024c). These models have demonstrated remarkable capabilities across various domains, from visual question answering to content generation and complex reasoning tasks

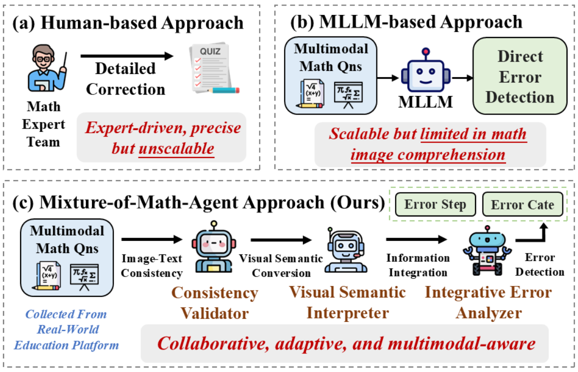

Figure 1: Comparison between previous human-based (a) and MLLM-based (b) paradigms vs. our proposed MATHAGENT framework (c) for multimodal mathematical error detection.

(Yuan et al., 2025). As education increasingly embraces digital transformation (Yan et al., 2025; Ye et al., 2025), the application of MLLMs to mathematical reasoning has emerged as a critical area of research, offering potential solutions to enhance teaching methodologies, provide personalized feedback, and support both educators and students in mathematical learning environments (Küchemann et al., 2025; Wang et al., 2024b; Yan et al., 2024a).

While significant progress has been made in utilizing MLLMs for mathematical problem-solving, a more practical and educationally valuable application lies in **mathematical error detection** (Li et al., 2024d; Song et al., 2025; Yan et al., 2024b; Yang et al., 2024; Zheng et al., 2024a). In real educational settings, identifying and categorizing students' mathematical errors provides deeper insights into their conceptual understanding and learning gaps than merely evaluating final answers (Jiang et al., 2024; Pepin et al., 2025). Error detection is a significantly more challenging task for MLLMs compared to standard problem-solving, as it requires not only understanding the correct solution path but also analyzing the student's flawed reasoning process. This task involves processing multiple inputs: the original problem (which may include multimodal elements), the correct solution, the stu-

†Corresponding authors.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track), pages 69–82
July 28-30, 2025 ©2025 Association for Computational Linguistics

dent's incorrect answer, and their detailed reasoning steps. The expected output comprises both error step identification (pinpointing exactly where the reasoning went wrong) and error categorization (classifying the type of misconception or mistake). This comprehensive analysis enables targeted educational interventions that address specific learning needs (Chu et al., 2025; Yan et al., 2024a).

Existing error detection approaches face significant limitations when applied to real-world multimodal mathematical problems. ❶ As shown in Figure 1(a), traditional human-based approaches rely on expert teams to provide detailed corrections. While precise and pedagogically sound, these methods are inherently unscalable and cannot meet the growing demand for personalized feedback in digital learning environments (Li et al., 2024c). ❷ As illustrated in Figure 1(b), MLLM-centric approaches, despite their computational scalability, exhibit suboptimal performance in mathematical image comprehension. For instance, symbolic representations in diagrams (*e.g.,* misaligned coordinate systems) or mismatched text-image pairs (*e.g.,* inconsistent geometric labels) often evade detection by MLLMs, leading to false predictions in error detection (Lu et al., 2023; Zhang et al., 2024).

To address these challenges, we propose and deploy **MATHAGENT, a novel Mixture-of-Math-Agent framework specifically designed for multimodal mathematical error detection**. Drawing inspiration from expert-guided problem-solving practices (Chen et al., 2025b; Li et al., 2024a), our framework decomposes the error detection workflow into three synergistic agents (refer to Figure 1(c)): an *image-text consistency validator* to detect semantic consistency, a *visual semantic interpreter* to extract structured expression from visual part of the problem, and an *integrative error analyzer* that correlates all text-based inputs to pinpoint error locations and categorize misconception types. By explicitly modeling the interdependencies between textual problem formulations, visual mathematical objects, and solution steps, MATHAGENT overcomes the aforementioned challenges inherent in both human-driven and MLLM-based approaches while maintaining computational tractability for real-world deployment.

Our contributions can be summarized as follows:

❶ We introduce MATHAGENT, the **first agent-based framework specifically designed for multimodal mathematical error detection**. Unlike previous paradigms that struggle with scalability, visual comprehension, and complex reasoning, MATHAGENT leverages a novel mixture-of-agents approach, decomposing the task into multiple subtasks via specialized mathematical agents.

❷ We validate our approach on **data sampled from a real educational platform**, demonstrating performance improvements over baseline models. MATHAGENT achieves approximately 5% higher accuracy in error step identification and 3% higher accuracy in error categorization, confirming its effectiveness in practical educational settings.

❸ MATHAGENT has been successfully **deployed in an educational platform that has served over one million K-12 students**. The system has achieved nearly 90% student satisfaction rates while yielding estimated cost savings of approximately one million dollars by reducing the need for manual error detection, demonstrating both its practical utility and economic value.

2 Related Work

2.1 Mathematical Error Detection

Mathematical error detection has evolved significantly from traditional rule-based systems to more sophisticated AI approaches (Li et al., 2024c; Yan et al., 2024a). Early work focused on predefined error patterns and procedural mistakes in specific mathematical domains, such as arithmetic operations or algebraic manipulations (Rushton, 2018). With the advent of deep learning, researchers develop models capable of identifying more complex conceptual misunderstandings by analyzing student solution processes (Xu et al., 2024a). Recent advances have leveraged LLMs to provide more nuanced error analysis and feedback generation, demonstrating promising results in understanding diverse student reasoning patterns (Gao et al., 2024; Li et al., 2024d, 2025a). However, most existing research has primarily focused on text-based settings, with *limited focus on multimodal contexts* where visual elements play a crucial role in problem representation (Yan et al., 2024b). MATHAGENT extends the frontier of mathematical error detection by specifically addressing the challenges of multimodal mathematical reasoning, introducing a specialized agent-based framework.

2.2 Agent for Mathematical Reasoning

The application of agent-based approaches to mathematical reasoning has gained significant traction in recent years (Chu et al., 2025). Initial efforts focused on single-agent systems that could execute predefined mathematical operations or follow structured solution procedures (Mei et al., 2024; Mitra et al., 2024). As LLMs advanced, researchers developed more sophisticated agents capable of step-by-step reasoning, self-verification, and even multi-step planning for complex mathematical problem-solving (Li et al., 2024b; Wu et al., 2023b; Xiong et al., 2024). Recent work has explored multi-agent frameworks where specialized agents collaborate on different aspects of mathematical reasoning, such as problem decomposition, solution planning, and verification (Gou et al., 2023; Xu et al., 2024b; Zhang et al., 2025). However, existing agent-based systems for mathematical reasoning have *primarily focused on problem-solving rather than error detection*, and few have adequately addressed the unique *challenges posed by multimodal mathematical content*. Our MATHAGENT represents a significant advancement in this domain by introducing a coordinated multi-agent system specifically designed for multimodal mathematical error detection.

See more related work in Appendix A.

3 Our Proposed MATHAGENT

3.1 Task Setting

We evaluate the framework's capability for multimodal error detection. The evaluation set contains N samples. For each sample i, input $\mathcal{I}_i$ includes:

- $Q_{\text{text},i}$: The textual problem statement.
- $Q_{\text{image},i}$: The visual part of the problem.
- $A_{\text{correct},i}$: The correct solution.
- $A_{\text{incorrect},i}$: An incorrect student solution.
- $\{S_{k,i}\}_{k=1}^{n_i}$: A sequence of n_i steps representing the student's step-by-step solution.

We define two subtasks as follows:

Subtask 1: Error Step Identification. The goal is to identify the index, x_i, of the *first* incorrect step in the solution sequence $\{S_{k,i}\}$. Formally:

$$x_i = \arg\min_k \{k \mid S_{k,i} \text{ is incorrect}\}$$

Subtask 2: Error Categorization. The goal is to classify the *type* of error into one of five categories based on the first incorrect step: VIS (Visual Perception), CAL (Calculation), REAS (Reasoning), KNOW (Knowledge), and MIS (Misinterpretation). The error category is denoted as $C_{\text{error},i}$. See details of error categories in Appendix B.

We use accuracy to evaluate performance.

- **Error Step Identification Accuracy:**

$$\text{Acc}_{\text{step}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(x_i = G_{\text{step},i})$$

where $G_{\text{step},i}$ is ground truth index of the first incorrect step, and $\mathbb{I}$ is the indicator function.

- **Error Categorization Accuracy:**

$$\text{Acc}_{\text{cate}} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(C_{\text{error},i} = G_{\text{error},i})$$

where $G_{\text{error},i}$ is ground truth error category.

3.2 Framework Overview

Our MATHAGENT framework is designed for real-world multimodal mathematical error detection. As illustrated in Figure 2, the framework takes as input a multimodal mathematical problem (text and image), a correct answer, a student's incorrect answer, and their solution steps. The output is the identified error step and the corresponding error category. The framework operates in three sequential phases: Image-Text Consistency Verification (Sec.3.3), Question Type-Driven Visual Semantic Conversion (Sec.3.4), and Multimodal Information Integration (Sec.3.5). Each phase employs a specialized agent to perform a specific task.

3.3 Phase 1: Image-Text Consistency Verification

Motivation. Recent studies have demonstrated that MLLMs often exhibit lower performance in multimodal mathematical reasoning tasks when the image and text information are highly redundant (Lu et al., 2023; Zhang et al., 2024). This phenomenon highlights the current limitations of MLLMs in visual understanding and multimodal semantic alignment (Li and Tang, 2024; Wu et al., 2024). Furthermore, in real-world educational settings, adaptively identifying high image-text consistency can improve efficiency, allowing us to bypass subsequent processing steps and directly proceed to error detection for highly overlapping problems.

Methodology. We introduce the *Image-Text Consistency Validator*. This agent takes the image and the textual description of the problem as input. It outputs a binary decision: whether the image and text are highly semantically consistent. The agent automatically determines the extent of semantic similarity between the image and text. Our system defaults to using GPT-4o[1] as the agent

[1]We used gpt-4o-2024-11-20.

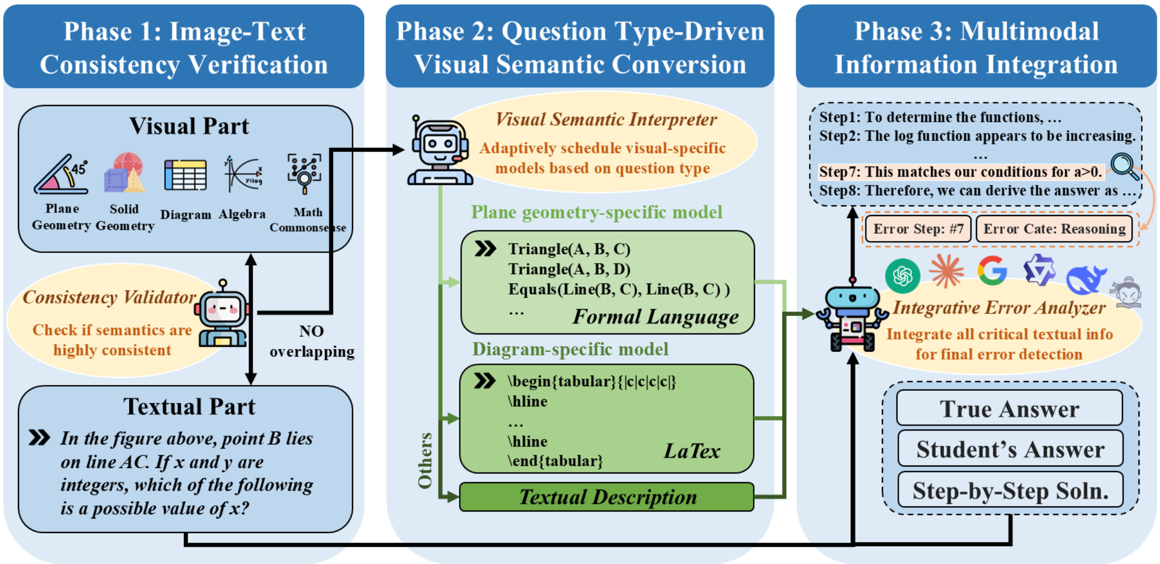

Figure 2: The framework of our proposed Mixture-of-Math-Agent for multimodal mathematical error detection.

for this phase. For example, if the image depicts a triangle with labeled angles and the text describes the same triangle and angles, the validator would output "highly consistent."

3.4 Phase 2: Question Type-Driven Visual Semantic Conversion

Motivation. If the image and text information are not highly overlapping, we need an effective way to extract visual information for subsequent error detection. Inspired by recent advances in symbolic reasoning (Alotaibi et al., 2024; Li et al., 2025b; Sullivan and Elsayed, 2024), we propose that MLLMs can adaptively dispatch specialized visual models based on the question type to convert visual information into a textual format. In particular, multimodal plane geometry problems, with their well-defined geometric relationships, are well-suited for conversion into formal language. Multimodal diagram problems, often involving tables or charts, are best represented using LaTeX. Other types are converted into textual descriptions.

Methodology. We propose the *Visual Semantic Interpreter*. This agent takes the image and the question type as input, and its output is a text-based representation of the visual information, tailored to the specific question type. The agent first determines the question type (*e.g.,* plane geometry, diagram, algebra) and then selects the appropriate conversion method. Our system defaults to using corresponding visual-specific models[2] as the agent for this phase. For instance, if the image is identified as a plane geometry setting, the interpreter might output a formal language representation like "Triangle(A, B, C), Angle(BAC, 45), Line(AB, 5)."

3.5 Phase 3: Multimodal Information Integration

Motivation. Based on the extracted visual information from the previous phase, a comprehensive integration of all available information is crucial for accurate error localization. This phase must combine the problem's content, the student's incorrect answer, and their reasoning steps to pinpoint the cause of the error. The agent in this phase is directly responsible for the output of the two subtasks: error step identification and error categorization. Our system is designed to be compatible with any MLLM for inference, leveraging the increasingly powerful information integration capabilities of modern LLMs (An et al., 2024).

Methodology. We introduce the *Integrative Error Analyzer*. This agent takes as input the problem's textual description, the converted visual information, the true answer, the student's answer, and the student's step-by-step solution. It outputs the identified error step and the error category. The agent first integrates all textual information and then analyzes the student's solution step-by-step, comparing it against the correct solution path. The agent for this phase is a flexibly selectable MLLM. For example, given a student's incorrect calculation in a geometry problem, the analyzer might output "Error Step: #3" and "Error Category: Calculation".

[2]Refer to Appendix C for details.

Table 1: Main result of baseline MLLMs and corresponding MathAgent framework. We denote STEP and CATE for error step identification and error categorization, the two subtasks of error detection, respectively, in Section 4.2.

Model	Error Step Identification	Error Categorization						Average
		VIS	CAL	REAS	KNOW	MIS	Overall	
GPT-4o (OpenAI, 2024)	55.10	46.30	50.40	64.90	9.20	46.30	53.08	54.09
w/ MathAgent	$59.50^{4.4\uparrow}$	$48.40^{2.1\uparrow}$	$55.00^{4.6\uparrow}$	$63.90^{1.0\downarrow}$	$9.50^{0.3\uparrow}$	$54.00^{7.7\uparrow}$	$55.11^{2.0\uparrow}$	$57.30^{3.2\uparrow}$
Gemini-Pro-1.5 (Reid et al., 2024)	52.00	9.10	46.80	62.70	31.90	13.00	44.51	48.26
w/ MathAgent	$57.90^{5.9\uparrow}$	$15.70^{6.6\uparrow}$	$48.50^{1.7\uparrow}$	$61.30^{1.4\downarrow}$	$33.30^{1.4\uparrow}$	$21.00^{8.0\uparrow}$	$46.10^{1.6\uparrow}$	$52.00^{3.8\uparrow}$
Claude-3.5-Sonnet (Anthropic, 2024)	50.20	35.70	48.40	64.80	21.00	11.40	49.50	49.85
w/ MathAgent	$55.10^{4.9\uparrow}$	$40.10^{4.4\uparrow}$	$55.30^{6.9\uparrow}$	$62.70^{2.1\downarrow}$	$24.70^{3.7\uparrow}$	$22.40^{11.0\uparrow}$	$52.63^{3.1\uparrow}$	$53.86^{4.0\uparrow}$
Qwen-VL-Max (Team, 2024)	48.70	15.20	78.90	50.50	14.30	36.60	52.87	50.78
w/ MathAgent	$56.70^{8.0\uparrow}$	$21.70^{6.5\uparrow}$	$81.30^{2.4\uparrow}$	$53.40^{2.9\uparrow}$	$12.80^{1.5\downarrow}$	$36.60^{0.0\uparrow}$	$55.80^{2.9\uparrow}$	$56.25^{5.5\uparrow}$
InternVL2 (Chen et al., 2024)	54.40	33.40	92.40	25.10	10.90	8.10	49.46	51.93
w/ MathAgent	$56.30^{1.9\uparrow}$	$38.80^{5.4\uparrow}$	$85.30^{7.1\downarrow}$	$36.80^{11.7\uparrow}$	$19.00^{8.1\uparrow}$	$13.70^{5.6\uparrow}$	$52.83^{3.4\uparrow}$	$54.57^{2.6\uparrow}$
LLaVA-NEXT (Liu et al., 2024a)	48.44	7.10	86.00	32.00	7.60	0.80	45.08	48.44
w/ MathAgent	$57.60^{5.8\uparrow}$	$15.70^{8.6\uparrow}$	$84.50^{1.5\downarrow}$	$45.10^{13.1\uparrow}$	$8.30^{0.7\uparrow}$	$3.80^{3.0\uparrow}$	$51.05^{6.0\uparrow}$	$54.32^{5.9\uparrow}$
Average Improvement	5.2 ↑	5.6 ↑	1.2 ↑	3.9 ↑	2.1 ↑	5.9 ↑	3.2 ↑	4.2 ↑
Human	81.60	70.30	86.00	63.50	53.40	62.00	72.23	76.91

4 Experiment

4.1 Experiment Settings

Dataset. The dataset consists of a carefully curated collection of 2,500 multimodal mathematical questions sourced from real student problem-solving data on educational platforms. Each entry in this evaluation dataset has been meticulously selected by educational experts to ensure high quality, free from issues such as erroneous question design. The student responses represent the most frequent incorrect answers corresponding to each question. Furthermore, the erroneous steps and error category labels for each question have been determined through discussions among at least three experienced educational specialists. The dataset predominantly features plane geometry problems, supplemented by solid geometry, diagrams, algebra, and mathematical commonsense questions. Refer to Appendix D for more dataset details.

Models. We select representative MLLMs (See sources in Appendix E) that have demonstrated effectiveness in recent studies (Wang et al., 2024a; Yan et al., 2024b; Zhang et al., 2024): InternVL-2 76B (Chen et al., 2024), LLaVA-NEXT 72B (Liu et al., 2024a), Qwen-VL-Max (Team, 2024), Claude-3.5-Sonnet (Anthropic, 2024), Gemini-Pro-1.5 (Reid et al., 2024), and GPT-4o (OpenAI, 2024). These MLLMs are already deployed on the educational platform, allowing for a direct comparison of the gains achieved by MathAgent. In our experiments, directly applying each MLLM to error detection serves as a baseline. We then evaluate the effectiveness of the MathAgent framework by systematically decomposing the complex reasoning task, with the agent in Phase 3 retaining the baseline MLLM. Additionally, we engage evaluators with a background in education to conduct corresponding human evaluations, aiming to assess the gap between MLLM and human-level intelligence.

4.2 Experimental Results & Analysis

Overall Performance Improvement with MathAgent. As shown in Table 1, MathAgent demonstrates significant performance improvements across both STEP and CATE subtasks. When integrated with various baseline MLLMs, MathAgent consistently enhances their error detection capabilities, with an average improvement of 4.2% across all models. Specifically, the framework boosts GPT-4o's performance from 54.09% to 57.30% (3.2% increase) and shows similar improvements for other models. This consistent enhancement across diverse architectures suggests that MathAgent can address inherent challenges in multimodal mathematical error detection by systematically processing multimodal information.

Differential Impact on STEP vs. CATE Tasks. The MathAgent framework yields more substantial improvements in STEP compared to CATE. Across all tested models, MathAgent achieves an average improvement of 5.2% in STEP tasks, while the enhancement for overall CATE tasks is 3.2%. For instance, GPT-4o shows a 4.4% improvement in STEP but only a 2.0% improvement in CATE. This difference likely stems from MathAgent 's information extraction and integration, which particularly benefits the error localization in sequential solution steps, while the more nuanced task of error categorization remains challenging.

Category-Specific Performance Variations.

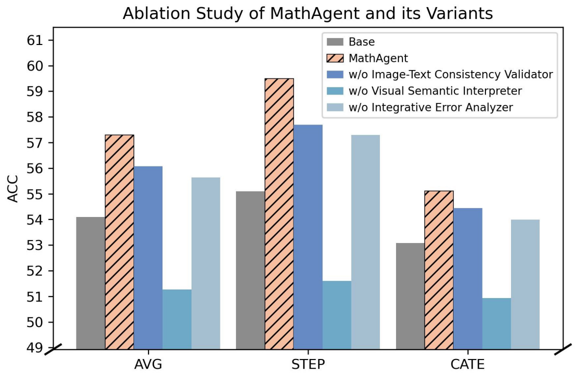

Figure 3: Ablation study of MATHAGENT.

MATHAGENT demonstrates the most significant improvements in detecting VIS and MIS, with average enhancements of 5.6% and 5.9% respectively across all models. For example, Gemini-Pro-1.5 shows a remarkable 6.6% improvement in VIS and 8.0% in MIS categories when augmented with MATHAGENT. In contrast, improvements in CAL, REAS, and KNOW are more modest at 1.2%, 0.4%, and 2.1% respectively. This pattern highlights MATHAGENT 's effectiveness in addressing multimodal integration challenges, as VIS and MIS errors fundamentally involve misalignments between visual information and problem interpretation.

Gap Between MATHAGENT and Human Performance. Despite the notable improvements, MATHAGENT still falls short of human-level performance in mathematical error detection. The best-performing MATHAGENT-enhanced framework (GPT-4o at 57.30%) remains significantly below human performance (76.91%). The persistent performance gap underscores the inherent complexity of mathematical error detection, which requires sophisticated reasoning abilities, domain knowledge, and multimodal understanding.

4.3 Ablation Study

As depicted in Figure 3, we evaluate performance of our MATHAGENT framework and its ablative variants, using GPT-4o with the best overall performance as the base setting. We investigate three variants: (i) *w/o Image-Text Consistency Validator*, which bypasses consistency check and processes all images in Phase 2; (ii) *w/o Visual Semantic Interpreter*, which replaces question type-driven visual model scheduling with a unified captioning approach for all images; and (iii) *w/o Integrative Error Analyzer*, which simply concatenates transcribed image information with student's solution steps and answer, omitting the integration with the problem's textual description. The results demonstrate that MATHAGENT achieves the highest accuracy on both STEP and CATE tasks. Notably, the w/o Visual Semantic Interpreter variant exhibits the lowest performance, presumably because generic descriptions of abstract geometric images may omit crucial details like edge lengths and angle measures. Removing the Image-Text Consistency Validator also leads to a performance drop, suggesting that discrepancies between potentially flawed image transcriptions and textual problem description can introduce contradictory information, negatively impacting the complex reasoning process.

5 Industrial Impact

Error Detection Performance Enhancement in Real-World Educational System. When deployed in educational platforms, MATHAGENT has demonstrated remarkable improvements in error detection performance that directly translate to educational value. As a diagnostic tool, MATHAGENT provides more precise feedback on student work, enabling targeted interventions. Furthermore, MATHAGENT 's adaptive architecture optimizes computational resources by automatically filtering problems based on image-text consistency and selecting specialized visual models according to problem types.

Student Satisfaction Rate Improvement. A/B testing conducted on the educational platform reveals significant improvements in student satisfaction with MATHAGENT-powered feedback systems. In a controlled study involving 10,000 K-12 students, MATHAGENT-enhanced feedback received an over 90% satisfaction rating, compared to 75% for traditional MLLM-based feedback. These improvements in student experience demonstrate MATHAGENT 's effectiveness as a pedagogically valuable tool that enhances the learning process.

We discuss more impact in Appendix F.

6 Conclusion

This paper presented MATHAGENT, a novel and effective framework for multimodal mathematical error detection in real-world educational settings. By leveraging a mixture-of-agent approach, MATHAGENT overcomes the limitations of existing human-based and MLLM-centric methods, achieving superior performance in identifying and categorizing student errors. The successful deployment of MATHAGENT on a large-scale educational platform, with improvements in accuracy, student satisfaction, and cost-effectiveness, underscores its significant technical and practical value.

Limitations

Despite the contributions demonstrated in our work, several limitations remain:

1. The effectiveness of MATHAGENT is contingent on the quality of the multimodal inputs. Poorly formatted or ambiguous problems may lead to inaccurate error detection. We will enhance our engineering pipeline to improve data cleaning and optimization processes, ensuring that input data is standardized and of high quality, which will lead to more accurate error detection.

2. While MATHAGENT improves error detection accuracy, it may still struggle with a broader range of error categories beyond the five specified. We will collaborate with educational experts to develop a more comprehensive framework of error categories that aligns with student needs and encompasses a wider variety of mathematical errors.

3. MATHAGENT does not incorporate recent advancements in o1-like slow-thinking reasoning, which may enhance the depth of error analysis but could impact user feedback time in deployed systems. In the future, we will explore integrating user intent recognition to adaptively schedule fast and slow reasoning modes, providing students with comprehensive and timely error analysis based on their needs.

Acknowledgements

This work was supported by NSF under grants III-2106758, and POSE-2346158; Guangdong Provincial Department of Education Project (Grant No.2024KQNCX028); Scientific Research Projects for the Higher-educational Institutions (Grant No.2024312096), Education Bureau of Guangzhou Municipality; Guangzhou-HKUST(GZ) Joint Funding Program (Grant No.2025A03J3957), Education Bureau of Guangzhou Municipality.

References

Fatimah Alotaibi, Adithya Kulkarni, and Dawei Zhou. 2024. Graph of logic: Enhancing llm reasoning with graphs and symbolic logic. In *2024 IEEE International Conference on Big Data (BigData)*, pages 5926–5935. IEEE.

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, Jian-Guang Lou, and Weizhu Chen. 2024. Make your llm fully utilize the context. *Advances in Neural Information Processing Systems*, 37:62160–62188.

Anthropic. 2024. Claude 3.5.

Simone Bianco, Luigi Celona, Marco Donzella, and Paolo Napoletano. 2023. Improving image captioning descriptiveness by ranking and llm-based fusion. *arXiv preprint arXiv:2306.11593*.

Junkai Chen, Zhijie Deng, Kening Zheng, Yibo Yan, Shuliang Liu, PeiJun Wu, Peijie Jiang, Jia Liu, and Xuming Hu. 2025a. Safeeraser: Enhancing safety in multimodal large language models through multimodal machine unlearning. *arXiv preprint arXiv:2502.12520*.

Justin Chih-Yao Chen, Sukwon Yun, Elias Stengel-Eskin, Tianlong Chen, and Mohit Bansal. 2025b. Symbolic mixture-of-experts: Adaptive skill-based routing for heterogeneous reasoning. *arXiv preprint arXiv:2503.05641*.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wangxiang Che. 2025c. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Zhendong Chu, Shen Wang, Jian Xie, Tinghui Zhu, Yibo Yan, Jinheng Ye, Aoxiao Zhong, Xuming Hu, Jing Liang, Philip S Yu, et al. 2025. Llm agents for education: Advances and applications. *arXiv preprint arXiv:2503.11733*.

Alejandro Cuadron, Dacheng Li, Wenjie Ma, Xingyao Wang, Yichuan Wang, Siyuan Zhuang, Shu Liu, Luis Gaspar Schroeder, Tian Xia, Huanzhi Mao, et al. 2025. The danger of overthinking: Examining the reasoning-action dilemma in agentic tasks. *arXiv preprint arXiv:2502.08235*.

Jingcheng Deng, Zhongtao Jiang, Liang Pang, Liwei Chen, Kun Xu, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2025. Following the autoregressive nature of llm embeddings via compression and alignment. *arXiv preprint arXiv:2502.11401*.

Yifan Du, Zikang Liu, Yifan Li, Wayne Xin Zhao, Yuqi Huo, Bingning Wang, Weipeng Chen, Zheng Liu, Zhongyuan Wang, and Ji-Rong Wen. 2025. Virgo: A preliminary exploration on reproducing o1-like mllm. *arXiv preprint arXiv:2501.01904*.

Bofei Gao, Zefan Cai, Runxin Xu, Peiyi Wang, Ce Zheng, Runji Lin, Keming Lu, Dayiheng Liu, Chang Zhou, Wen Xiao, et al. 2024. Llm critics help catch bugs in mathematics: Towards a better mathematical verifier with natural language feedback. *arXiv preprint arXiv:2406.14024*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280.

Jiahao Huo, Yibo Yan, Boren Hu, Yutao Yue, and Xuming Hu. 2024. Mmneuron: Discovering neuron-level domain-specific interpretation in multimodal large language model. *arXiv preprint arXiv:2406.11193*.

Jiahao Huo, Yibo Yan, Xu Zheng, Yuanhuiyi Lyu, Xin Zou, Zhihua Wei, and Xuming Hu. 2025. Mmunlearner: Reformulating multimodal machine unlearning in the era of multimodal large language models. *arXiv preprint arXiv:2502.11051*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. 2024. Llms can find mathematical reasoning mistakes by pedagogical chain-of-thought. *arXiv preprint arXiv:2405.06705*.

Stefan Küchemann, Karina E Avila, Yavuz Dinc, Chiara Hortmann, Natalia Revenga, Verena Ruf, Niklas Stausberg, Steffen Steinert, Frank Fischer, Martin Fischer, et al. 2025. On opportunities and challenges of large multimodal foundation models in education. *npj Science of Learning*, 10(1):11.

Dawei Li, Zhen Tan, Peijia Qian, Yifan Li, Kumar Satvik Chaudhary, Lijie Hu, and Jiayi Shen. 2024a. Smoa: Improving multi-agent large language models with sparse mixture-of-agents. *arXiv preprint arXiv:2411.03284*.

Hang Li, Tianlong Xu, Kaiqi Yang, Yucheng Chu, Yanling Chen, Yichi Song, Qingsong Wen, and Hui Liu. 2024b. Ask-before-detection: Identifying and mitigating conformity bias in llm-powered error detector for math word problem solutions. *arXiv preprint arXiv:2412.16838*.

Hang Li, Tianlong Xu, Chaoli Zhang, Eason Chen, Jing Liang, Xing Fan, Haoyang Li, Jiliang Tang, and Qingsong Wen. 2024c. Bringing generative ai to adaptive learning in education. *arXiv preprint arXiv:2402.14601*.

Songtao Li and Hao Tang. 2024. Multimodal alignment and fusion: A survey. *arXiv preprint arXiv:2411.17040*.

Xiaoyuan Li, Wenjie Wang, Moxin Li, Junrong Guo, Yang Zhang, and Fuli Feng. 2024d. Evaluating mathematical reasoning of large language models: A focus on error identification and correction. *arXiv preprint arXiv:2406.00755*.

Yiyao Li, Dhanish Musharraf Ubaidali, Lu Wang, and Wenyu Zhang. 2025a. Step-by-step correction of llm-based math word problems solutions. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Zenan Li, Zhaoyu Li, Wen Tang, Xian Zhang, Yuan Yao, Xujie Si, Fan Yang, Kaiyu Yang, and Xiaoxing Ma. 2025b. Proving olympiad inequalities by synergizing llms and symbolic reasoning. *arXiv preprint arXiv:2502.13834*.

Zhong-Zhi Li, Duzhen Zhang, Ming-Liang Zhang, Jiaxin Zhang, Zengyan Liu, Yuxuan Yao, Haotian Xu, Junhao Zheng, Pei-Jie Wang, Xiuyi Chen, et al. 2025c. From system 1 to system 2: A survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024b. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*.

Oscar Mañas, Benno Krojer, and Aishwarya Agrawal. 2024. Improving automatic vqa evaluation using large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4171–4179.

Tula Masterman, Sandi Besen, Mason Sawtell, and Alex Chao. 2024. The landscape of emerging ai agent architectures for reasoning, planning, and tool calling: A survey. *arXiv preprint arXiv:2404.11584*.

Kai Mei, Xi Zhu, Wujiang Xu, Wenyue Hua, Mingyu Jin, Zelong Li, Shuyuan Xu, Ruosong Ye, Yingqiang Ge, and Yongfeng Zhang. 2024. Aios: Llm agent operating system. *arXiv preprint arXiv:2403.16971*.

Arindam Mitra, Hamed Khanpour, Corby Rosset, and Ahmed Awadallah. 2024. Orca-math: Unlocking the potential of slms in grade school math. *arXiv preprint arXiv:2402.14830*.

Minheng Ni, Yutao Fan, Lei Zhang, and Wangmeng Zuo. 2024. Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning. *arXiv preprint arXiv:2410.03321*.

OpenAI. 2024. GPT-4o system card.

Vrajkumar Patel, Aayush Modi, Harsh Mistry, Abhishesh Mishra, Rocky Upadhyay, and Apoorva Shah. 2025. From alt-text to real context: Revolutionizing image captioning using the potential of llm.

Birgit Pepin, Nils Buchholtz, and Ulises Salinas-Hernández. 2025. "mathematics education in the era of chatgpt: Investigating its meaning and use for school and university education"—editorial to special issue. *Digital Experiences in Mathematics Education*, pages 1–8.

Mohaimenul Azam Khan Raiaan, Md Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE access*, 12:26839–26874.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Sheryl J Rushton. 2018. Teaching and learning mathematics through error analysis. *Fields Mathematics Education Journal*, 3(1):1–12.

Minghao Shao, Abdul Basit, Ramesh Karri, and Muhammad Shafique. 2024. Survey of different large language model architectures: Trends, benchmarks, and challenges. *IEEE Access*.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. 2025. Prmbench: A fine-grained and challenging benchmark for process-level reward models. *arXiv preprint arXiv:2501.03124*.

Rob Sullivan and Nelly Elsayed. 2024. Can large language models act as symbolic reasoners? *arXiv preprint arXiv:2410.21490*.

Shiliang Sun, Zhilin Lin, and Xuhan Wu. 2025. Hallucinations of large multimodal models: Problem and countermeasures. *Information Fusion*, page 102970.

Qwen Team. 2024. Introducing qwen1.5.

Omkar Thawakar, Dinura Dissanayake, Ketan More, Ritesh Thawkar, Ahmed Heakl, Noor Ahsan, Yuhao Li, Mohammed Zumri, Jean Lahoud, Rao Muhammad Anwer, et al. 2025. Llamav-o1: Rethinking step-by-step visual reasoning in llms. *arXiv preprint arXiv:2501.06186*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Houxing Ren, Aojun Zhou, Mingjie Zhan, and Hongsheng Li. 2024a. Measuring multimodal mathematical reasoning with math-vision dataset. *Advances in Neural Information Processing Systems*, 37:95095–95169.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024b. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. 2023a. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE.

Tao Wu, Mengze Li, Jingyuan Chen, Wei Ji, Wang Lin, Jinyang Gao, Kun Kuang, Zhou Zhao, and Fei Wu. 2024. Semantic alignment for multimodal large language models. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 3489–3498.

Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. 2023b. Mathchat: Converse to tackle challenging math problems with llm agents. *arXiv preprint arXiv:2306.01337*.

Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. 2024. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*.

Junbin Xiao, Nanxin Huang, Hangyu Qin, Dongyang Li, Yicong Li, Fengbin Zhu, Zhulin Tao, Jianxing Yu, Liang Lin, Tat-Seng Chua, et al. 2025. Videoqa in the era of llms: An empirical study. *International Journal of Computer Vision*, pages 1–24.

Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large multimodal agents: A survey. *arXiv preprint arXiv:2402.15116*.

Wei Xiong, Chengshuai Shi, Jiaming Shen, Aviv Rosenberg, Zhen Qin, Daniele Calandriello, Misha Khalman, Rishabh Joshi, Bilal Piot, Mohammad Saleh, et al. 2024. Building math agents with multi-turn iterative preference learning. *arXiv preprint arXiv:2409.02392*.

Fengli Xu, Qianyue Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, et al. 2025.

Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

Tianlong Xu, Richard Tong, Jing Liang, Xing Fan, Haoyang Li, and Qingsong Wen. 2024a. Foundation models for education: Promises and prospects. *IEEE Intelligent Systems*, 39(3):20–24.

Tianlong Xu, Yi-Fan Zhang, Zhendong Chu, Shen Wang, and Qingsong Wen. 2024b. Ai-driven virtual teacher for enhanced educational efficiency: Leveraging large pretrain models for autonomous error analysis and correction. *arXiv preprint arXiv:2409.09403*.

Yibo Yan and Joey Lee. 2024. Georeasoner: Reasoning on geospatially grounded context for natural language understanding. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 4163–4167.

Yibo Yan, Jiamin Su, Jianxiang He, Fangteng Fu, Xu Zheng, Yuanhuiyi Lyu, Kun Wang, Shen Wang, Qingsong Wen, and Xuming Hu. 2024a. A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. *arXiv preprint arXiv:2412.11936*.

Yibo Yan, Shen Wang, Jiahao Huo, Hang Li, Boyan Li, Jiamin Su, Xiong Gao, Yi-Fan Zhang, Tianlong Xu, Zhendong Chu, et al. 2024b. Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. *arXiv preprint arXiv:2410.04509*.

Yibo Yan, Shen Wang, Jiahao Huo, Jingheng Ye, Zhendong Chu, Xuming Hu, Philip S Yu, Carla Gomes, Bart Selman, and Qingsong Wen. 2025. Position: Multimodal large language models can significantly advance scientific reasoning. *arXiv preprint arXiv:2502.02871*.

Yibo Yan, Haomin Wen, Siru Zhong, Wei Chen, Haodong Chen, Qingsong Wen, Roger Zimmermann, and Yuxuan Liang. 2024c. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*, pages 4006–4017.

Ling Yang, Zhaochen Yu, Tianjun Zhang, Minkai Xu, Joseph E Gonzalez, CUI Bin, and YAN Shuicheng. 2024. Supercorrect: Supervising and correcting language models with error-driven insights. In *The Thirteenth International Conference on Learning Representations*.

Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. 2025a. Towards thinking-optimal scaling of test-time compute for llm reasoning. *arXiv preprint arXiv:2502.18080*.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, et al. 2025b. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *arXiv preprint arXiv:2503.10615*.

Huanjin Yao, Jiaxing Huang, Wenhao Wu, Jingyi Zhang, Yibo Wang, Shunyu Liu, Yingjie Wang, Yuxin Song, Haocheng Feng, Li Shen, et al. 2024. Mulberry: Empowering mllm with o1-like reasoning and reflection via collective monte carlo tree search. *arXiv preprint arXiv:2412.18319*.

Jingheng Ye, Shen Wang, Deqing Zou, Yibo Yan, Kun Wang, Hai-Tao Zheng, Zenglin Xu, Irwin King, Philip S Yu, and Qingsong Wen. 2025. Position: Llms can be good tutors in foreign language education. *arXiv preprint arXiv:2502.05467*.

Yuan Yuan, Zhaojian Li, and Bin Zhao. 2025. A survey of multimodal learning: Methods, applications, and future. *ACM Computing Surveys*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer.

Yi-Fan Zhang, Hang Li, Dingjie Song, Lichao Sun, Tianlong Xu, and Qingsong Wen. 2025. From correctness to comprehension: Ai agents for personalized error diagnosis in education. *arXiv preprint arXiv:2502.13789*.

Jiaxing Zhao, Xihan Wei, and Liefeng Bo. 2025. R1-omni: Explainable omni-multimodal emotion recognition with reinforcing learning. *arXiv preprint arXiv:2503.05379*.

Tianyi Zhao, Liangliang Zhang, Yao Ma, and Lu Cheng. 2024. A survey on safe multi-modal learning systems. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6655–6665.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2024a. Processbench: Identifying process errors in mathematical reasoning. *arXiv preprint arXiv:2412.06559*.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, and Xuming Hu. 2024b. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. *arXiv preprint arXiv:2408.09429*.

Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, et al. 2024. Evaluation of openai o1: Opportunities and challenges of agi. *arXiv preprint arXiv:2409.18486*.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. 2024a. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*.

Pengyuan Zhou, Lin Wang, Zhi Liu, Yanbin Hao, Pan Hui, Sasu Tarkoma, and Jussi Kangasharju. 2024b. A survey on generative ai and llm for video generation, understanding, and streaming. *arXiv preprint arXiv:2404.16038*.

A More Related Work

A.1 Multimodal Large Language Model

Current MLLMs adopt a similar framework, including a vision encoder, a connector, and an LLM backbone, which was initially proposed by LLaVA (Liu et al., 2024b). By training these components via visual instruction tuning, the vision embeddings extracted by the vision encoder are aligned with the word space of LLM through the connector (Raiaan et al., 2024; Shao et al., 2024). Such a framework enables MLLMs to understand visual input such as images and video, while preserving the powerful reasoning and generation abilities of autoregressive LLMs (Deng et al., 2025). As a result, some MLLMs achieve state-of-the-art performance across a wide variety of multimodal tasks such as visual question answering (Mañas et al., 2024; Xiao et al., 2025), image captioning (Bianco et al., 2023; Patel et al., 2025), video understanding (Huang et al., 2024; Zhou et al., 2024b), and more diverse tasks (Huo et al., 2024; Yan and Lee, 2024). On the other hand, with the development of o1-like systems in LLMs (Jaech et al., 2024; Li et al., 2025c; Zhong et al., 2024), there is also a tendency to trigger the slow-thinking potentials of MLLMs (Yang et al., 2025b; Yao et al., 2024; Zhao et al., 2025). For example, Virgo (Du et al., 2025) makes a preliminary exploration of multimodal slow-thinking systems by directly fine-tuning a capable MLLM with a small amount of textual long-form thought data, while Vision-o1 (Ni et al., 2024) proposes a multimodal multi-turn chain-of-thought framework to simulate human reasoning for MLLMs on ambiguous instructions. Furthermore, LlamaV-o1 (Thawakar et al., 2025) uses a multiturn curriculum learning approach to facilitate MLLMs in incremental skill acquisition and problem-solving. Despite these efforts, the development of o1-like multimodal systems is still in its stages (Chen et al., 2025c; Masterman et al., 2024; Xu et al., 2025), with significant problems such as overthinking (Cuadron et al., 2025; Yang et al., 2025a), safety (Chen et al., 2025a; Huo et al., 2025; Zhao et al., 2024), and hallucination (Sun et al., 2025; Zheng et al., 2024b; Zhou et al., 2024a).

B Error Category Details

The discrepancies within the five error categories are delineated as follows:

- ✭ **Visual Perception Errors (VIS)**: These errors arise when there is a failure to accurately interpret the information contained within images or diagrams presented in the question due to visual issues.
- ✭ **Calculation Error (CAL)**: These errors manifest during the calculation process, which may include arithmetic mistakes such as incorrect addition, subtraction, multiplication, or division, errors in unit conversion, or mistakes in the numerical signs between multiple steps.
- ✭ **Reasoning Error (REAS)**: These errors occur during the problem-solving process when improper reasoning is applied, leading to incorrect application of logical relationships or conclusions.
- ✭ **Knowledge Error (KNOW)**: These errors result from incomplete or incorrect understanding of the knowledge base, leading to mistakes when applying relevant knowledge points.
- ✭ **Misinterpretation of the Question (MIS)**: These errors occur when there is a failure to correctly understand the requirements of the question or a misinterpretation of the question's intent, leading to responses that are irrelevant to the question's demands. For instance, if the question asks for a letter and a number is provided, or vice versa.

C Visual-Specific Models

In our deployed system, we employ specialized models tailored to different problem types to ensure optimal performance. For plane geometry problems, we utilize Inter-GPS[3], a groundbreaking geometry problem solver developed by Lu et al. (2021). As the first system capable of automatic program parsing and interpretable symbolic reasoning, Inter-GPS demonstrates its effectiveness through dual-channel processing: it employs rule-based text parsing for textual analysis and neural object detection for diagram interpretation, seamlessly converting problem texts and diagrams into formal language representations. Furthermore, its integration of theorem knowledge as conditional rules enables systematic, step-by-step symbolic reasoning.

When addressing diagram-based problems, particularly those involving tabular data, we imple-

[3] https://github.com/lupantech/InterGPS

ment StructTable-InternVL2-1B[4], a sophisticated model developed by Xia et al. (2024). This end-to-end solution, known as StructEqTable, excels in visual table processing by accurately generating LaTeX descriptions from table images while simultaneously supporting multiple advanced functionalities, including structural extraction and question-answering capabilities, thereby significantly expanding its practical applications.

For general visual content processing beyond these specialized domains, we leverage the vit-gpt2-image-captioning model[5] to generate comprehensive and detailed image captions, ensuring robust performance across diverse visual understanding tasks.

D Dataset Details

D.1 Dataset Statistics

Our evaluation dataset comprises 2,500 multimodal mathematical questions spanning diverse problem types and error categories. As illustrated in Figure 4, the dataset is predominantly composed of Plane Geometry problems (62.4%), followed by Algebra (11.5%), Diagram problems (9.3%), Math Commonsense (9.2%), and Solid Geometry (7.6%). This distribution reflects the prevalence of geometry-based problems in mathematical education that benefit significantly from visual representation and analysis.

The dataset captures a wide spectrum of error categories that students commonly encounter. Reasoning Errors constitute the largest proportion at 38.0%, highlighting the challenges students face in logical deduction and proof construction. Calculation Errors account for 36.5% of the dataset, representing arithmetic mistakes and computational inaccuracies. Visual Perception Errors make up 15.8%, underscoring the importance of correctly interpreting visual elements in mathematical problem-solving. Knowledge Errors and Misinterpretation of Questions represent smaller but significant portions at 4.8% and 4.9% respectively.

The complexity of the problems is reflected in the reasoning steps required for solution, with an average of 7.6 steps per problem, ranging from a minimum of 3 to a maximum of 20 steps. The textual component of the problems varies considerably in length, averaging 168 characters, with the shortest problem containing just 13 characters and the most verbose extending to 719 characters. This variation in problem complexity and presentation provides a robust benchmark for evaluating MATHAGENT 's performance across different mathematical contexts and difficulty levels.

Statistic	**Number**
Total multimodal questions	2,500
Problem Type	
- Plane Geometry	1559 (62.4%)
- Solid Geometry	191 (7.6%)
- Diagram	233 (9.3%)
- Algebra	288 (11.5%)
- Math Commonsense	229 (9.2%)
Error Category	
- Visual Perception Error	395 (15.8%)
- Calculation Error	912 (36.5%)
- Reasoning Error	951 (38.0%)
- Knowledge Error	119 (4.8%)
- Misinterpretation of the Qns	123 (4.9%)
Average Reasoning Step	7.6
Maximum Reasoning Step	20
Minimum Reasoning Step	3
Average Question Length	168
Maximum Question Length	719
Minimum Question Length	13

Figure 4: Key statistics of dataset.

D.2 Data Source

The data used in this study originates from a real-world online education platform, ensuring its relevance and applicability to practical educational scenarios. This dataset is not synthetically generated; instead, it comprises authentic student submissions, including both correct and incorrect solutions. This provides a realistic representation of the types of errors students commonly make in a learning environment. Furthermore, the data includes a diverse range of mathematical problems, reflecting the breadth of topics covered in K-12 mathematics curricula. The use of real-world data enhances the ecological validity of our findings and ensures that the MATHAGENT framework is evaluated on data that closely resembles the challenges encountered in actual educational settings. The platform anonymizes all student data to protect privacy, while preserving the integrity and richness of the information needed for effective error detection and analysis.

[4]https://github.com/Alpha-Innovator/StructEqTable-Deploy

[5]https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

E Model Sources

Table 2 details specific sources for the various MLLMs we evaluate. The chosen MLLMs have been deployed in the educational platform for real-world and real-time evaluation.

MLLMs	Source	URL
InternVL2-76B	local checkpoint	https://huggingface.co/OpenGVLab/InternVL2-Llama3-76B
LLaVA-NEXT-72B	local checkpoint	https://huggingface.co/llava-hf/llava-next-72b-hf
Qwen-VL-Max	qwen-vl-max-0809	https://modelscope.cn/studios/qwen/Qwen-VL-Max
Claude-3.5-Sonnet	claude-3-5-sonnet	https://www.anthropic.com/api
Gemini-Pro-1.5	gemini-1.5-pro-latest	https://deepmind.google/technologies/gemini/pro/
GPT-4o	gpt-4o-2024-11-20	https://platform.openai.com/docs/models/gpt-4o

Table 2: Sources of our evaluated MLLMs.

F More Industrial Impact

We discuss more industrial impact of MATHAGENT as follows:

Cost Savings and Resource Optimization. Based on industry standards where expert mathematical error annotation costs approximately \$1 per problem, MATHAGENT has generated estimated savings of \$1.2 million annually. This calculation is derived from serving approximately 120,000 students, each of whom receives feedback on an average of 10 complex mathematical problems per month. Additionally, the system reduces teacher workload by an estimated 4.7 hours per week, allowing educators to focus on higher-value instructional activities rather than routine error identification. This translates to significant time savings, which can be redirected towards personalized instruction, curriculum development, or professional development. This efficiency gain is particularly important as online learning platforms scale to serve larger student populations.

Learning Outcome Acceleration. Longitudinal studies tracking student performance before and after MATHAGENT implementation show measurable improvements in learning outcomes. Students receiving MATHAGENT-powered feedback demonstrated a 23% faster mastery rate of complex mathematical concepts compared to control groups. This accelerated learning trajectory is attributed to the system's ability to provide immediate, precise feedback on mathematical errors, allowing students to correct misconceptions earlier in their learning process. The educational impact is particularly pronounced in traditionally underserved school districts, where access to expert mathematics teachers is limited, helping to narrow the achievement gap in STEM education.

Teacher Professional Development Enhancement. Beyond student-facing benefits, MATHAGENT serves as a powerful professional development tool for mathematics educators. By analyzing patterns in student errors across classrooms, the system generates insights into common misconceptions and learning obstacles that inform teaching strategies. Teachers report that these insights have transformed their instructional approaches, with 20% indicating they have modified their teaching methods based on MATHAGENT 's analytics. Furthermore, the system serves as a model for teachers to improve their own feedback practices, with educators reporting a 32% increase in confidence when providing mathematical explanations after using the system for one semester. This "teach the teacher" effect creates a virtuous cycle where both student learning and teacher effectiveness continually improve.