Can MLLMs Read Students' Minds? Unpacking Multimodal Error Analysis in Handwritten Math

Dingjie Song[1], Tianlong Xu[2], Yi-Fan Zhang[4], Hang Li[5], Zhiling Yan[1], Xing Fan[3], Haoyang Li[3], Lichao Sun[1], and Qingsong Wen[2⋆]

[1] Lehigh University, Bethlehem, PA, USA
{dis724,zhy423,lis221}@lehigh.edu
[2] Squirrel Ai Learning, Bellevue, WA, USA
txu0915@gmail.com, qingsongedu@gmail.com
[3] Squirrel Ai Learning, Shanghai, China
fanxing@songshuai.com, derek@squirrelai.com
[4] Institute of Automation, Chinese Academy of Sciences, Beijing, China
yifanzhang.cs@gmail.com
[5] Michigan State University, East Lansing, MI, USA
lihang4@msu.edu

Project Page Dataset Code

Abstract. Assessing student handwritten scratchwork is crucial for personalized educational feedback but presents unique challenges due to diverse handwriting, complex layouts, and varied problem-solving approaches. Existing educational NLP primarily focuses on textual responses and neglects the complexity and multimodality inherent in authentic handwritten scratchwork. Current multimodal large language models (MLLMs) excel at visual reasoning but typically adopt an "examinee perspective", prioritizing generating correct answers rather than diagnosing student errors. To bridge these gaps, we introduce **ScratchMath**, a novel benchmark specifically designed for explaining and classifying errors in authentic handwritten mathematics scratchwork. Our dataset comprises 1,720 mathematics samples from Chinese primary and middle school students, supporting two key tasks: Error Cause Explanation (ECE) and Error Cause Classification (ECC), with seven defined error types. The dataset is meticulously annotated through rigorous human-machine collaborative approaches involving multiple stages of expert labeling, review, and verification. We systematically evaluate 16 leading MLLMs on ScratchMath, revealing significant performance gaps relative to human experts, especially in visual recognition and logical reasoning. Proprietary models notably outperform open-source models, with large reasoning models showing strong potential for error explanation. All evaluation data and frameworks are publicly available to facilitate further research.

Keywords: Multimodal Large Language Models · Mathematical Reasoning · Error Diagnosis · Multimodal Systems · Benchmarking · Handwritten Recognition.

⋆ Corresponding author.

1 Introduction

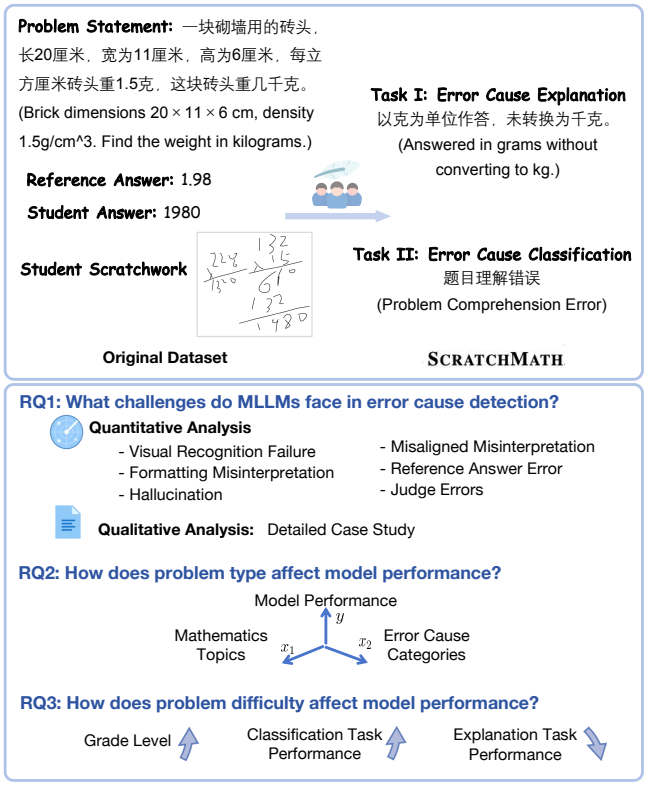


Fig. 1: Overview of this work. (Top) An example illustrating two tasks proposed. (Bottom) Summary of the three research questions addressed in this study.

Automatically analyzing student work to provide precise, personalized feedback is critical in educational AI [33,12,8]. Teachers often diagnose misconceptions and errors by examining students' handwritten scratchwork [5]. Authentic scratchwork reflects individual cognitive processes but introduces unique challenges: ambiguity in symbol recognition (e.g., confusion between "1," "l," and "|"), complex spatial layouts (e.g., fractions, superscripts), and personalized problem-solving strategies [5]. Accurate automated analysis of scratchwork can significantly enhance personalized teaching interventions [30].

Previous educational NLP studies utilized rule-based systems or machine learning classifiers for error detection [4,17], but these approaches lack generalizability and rely heavily on expert-defined error types. Recent work involving fine-grained LLM-based analyses using cognitive theory-guided strategies [11] or

iterative feedback loops [6] mostly address textual answers, neglecting multimodal inputs such as handwritten scratchwork.

While multimodal large language models (MLLMs) [14,2] excel at visual reasoning tasks, they primarily adopt an "examinee perspective," focusing on generating correct answers rather than analyzing student solutions to diagnose errors—a perspective analogous to that of an educator or examiner [32,26,30]. Additionally, recent multimodal benchmarks, such as ErrorRadar [30], and MathAgent [31], often utilize structured data, limiting their effectiveness in capturing the complexity of authentic handwritten scratchwork and focusing mostly on error classification rather than detailed explanations.

To address these gaps, we introduce **ScratchMath**, a novel benchmark specifically designed for explaining and classifying errors in authentic handwritten scratchwork. Our dataset comprises 1,720 Chinese mathematics samples from primary and middle school students, covering five critical mathematical topics: Numbers and Expressions, Equations and Functions, Geometry and Measurement, Applied Mathematics, and Statistics and Probability. The dataset supports two essential tasks: **Error Cause Explanation (ECE)** and **Error Cause Classification (ECC)**. Based on student scratchwork, we define seven student error types, including *Problem Comprehension Error* and *Calculation Error*. The annotation process employs a human-machine collaborative approach, initially using MLLM for preliminary annotations, followed by multiple stages of expert labeling, review and verification to ensure accuracy and reliability.

We systematically evaluate 16 leading MLLMs (e.g., o4-mini [10], QVQ [24]) on ScratchMath with extensive analysis (see Figure 1). Results reveal significant gaps compared to human experts, particularly in correcting visual recognition errors and understanding logical transitions in multi-step solutions. Notably, proprietary models significantly outperform open-source models, and large reasoning models show promising capabilities, especially on the explanation task. Our primary contributions are threefold:

1. Introducing a novel multimodal error-detection and explanation benchmark task, specifically tailored for educational settings;
2. Developing and publicly releasing the first high-quality, multimodal dataset of authentic student handwritten scratchwork, annotated via rigorous human-machine collaboration;
3. Conducting the first evaluation of state-of-the-art MLLMs on this task, including detailed analyses highlighting their capabilities and limitations.

2 Related Work

2.1 LLMs and MLLMs for Education

Research on LLMs as AI tutors prioritizes pedagogical alignment and practical feedback [27]. For example, a LLaMA model was fine-tuned using GPT-4-based rubrics [13]. Studies also demonstrate that adaptive LLM-generated feedback effectively boosts student motivation [12], and multimodal LLMs (MLLMs) can

effectively summarize diverse learner data to aid teachers' assessments [8]. However, existing MLLM-based grading methods for handwritten student solutions [5,15] often struggle due to the complexity of authentic scratchwork. Despite advances in automatic scoring and feedback generation, few studies focus explicitly on pinpointing and explaining the precise reasoning failures within authentic handwritten scratchwork.

2.2 LLMs and MLLMs for Mathematical Reasoning

Beyond text-based mathematical reasoning, recent studies highlight challenges faced by MLLMs in interpreting diagrams, handwritten derivations, and visual reasoning tasks [29]. Benchmarks such as MathVerse [32], MATH-V [26], and MileBench [20] reveal that even advanced models overlook crucial visual details. Specialized methods like Math-LLaVA [19] and LLaVA [14] have not fully resolved these issues. Recent multimodal benchmarks, including ErrorRadar [30] and MathAgent [31], mainly use structured or semi-structured inputs, emphasizing error localization or classification rather than detailed explanations. Moreover, cognitive theory-guided approaches [11] and iterative feedback strategies [6] remain limited to text-only contexts.

Our work is also related to Handwritten Mathematical Expression Recognition (HMER), which converts handwritten mathematical notation into machine-readable formats. The CROHME competition series [16] has served as the primary benchmark for this task, driving progress from structural approaches to neural encoder-decoder and transformer-based models [25]. However, HMER focuses on symbol-level recognition accuracy, whereas ScratchMath targets a fundamentally different goal: diagnosing the *reasoning errors* behind student solutions, which requires understanding both the visual content and the underlying mathematical logic. Our work addresses these complementary gaps by explicitly evaluating multimodal error detection and detailed explanation within authentic handwritten student scratchwork.

3 The ScratchMath Benchmark

3.1 Task Definition and Taxonomy.

Our goal is to evaluate MLLMs' ability to detect and explain errors in student solutions to math problems. A primary challenge is interpreting student scratchwork, which often combines diverse elements (e.g., handwritten text, symbolic notation, drawings) and requires integration with logical mathematical reasoning. Figure 2 provides an overview of our task setup and evaluation framework.

Formal Task Definition Formally, each instance is defined by the following tuple:

$$(Q, A_{\text{ref}}, S, A_{\text{stud}}, I)$$

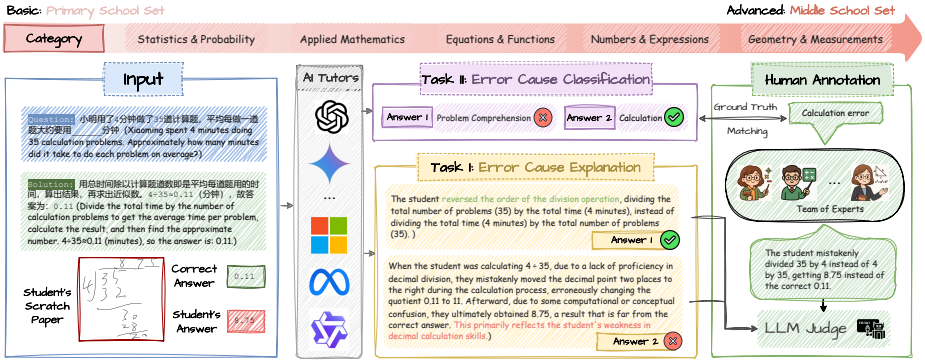


Fig. 2: Overview of ScratchMath (with models' answer and human labeled answer translated in English). It illustrates dataset structure, tasks (ECE and ECC), multimodal model predictions, and expert human annotations.

where Q is the problem statement, A_{ref} is the reference (correct) answer, S is the reference solution, A_{stud} is the student's provided answer, and I is an image of the student's scratchwork.

Our dataset is specifically structured to support two critical tasks: **Error Cause Explanation (ECE)**: An open-ended explanation describing the specific reason for the student's error, denoted as E. **Error Cause Classification (ECC)**: A categorical classification identifying the type of error from a predefined taxonomy, denoted as C.

Error Cause Taxonomy The taxonomy was systematically constructed through iterative expert reviews and educational theory-driven analysis on a much larger corpus of educational data, resulting in seven distinct error categories: *Procedural Error, Calculation Error, Logical Reasoning Error, Transcription Error, Problem Comprehension Error, Conceptual Knowledge Error, and Attention and Detail Error*, with the quantity distribution shown in Figure 4. Notably, all error types are represented across both primary and middle school problems in our dataset.

Evaluation Metrics We employ two evaluation approaches aligned with the dual outputs required: **Error Cause Explanation (ECE)**: We use an LLM-as-a-Judge framework, which assesses the semantic alignment of model-generated explanations with ground truth. **Error Cause Classification (ECC)**: Error classification is evaluated using accuracy (Acc), strictly considering correct only those cases where the predicted class exactly matches the annotated class. This strict criterion emphasizes precision in classification performance.

3.2 Dataset Construction

Recent studies have shown that data contamination is a prevalent concern in MLLM benchmarks [21]; our dataset mitigates this risk as it consists entirely of

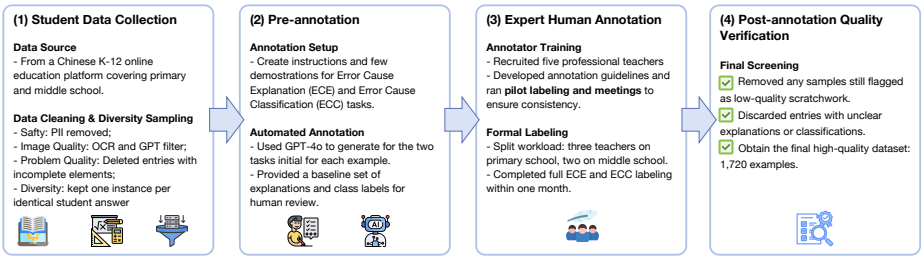


Fig. 3: An overview of the ScratchMath benchmark construction pipeline.

original, unpublished student scratchwork. As shown in Figure 3, the dataset construction consists of four parts.

Part I. Student Data Collection **Data Source.** Student data were sampled from an online education platform, covering primary (grades 1-6) and middle school (grades 7-9) math questions. Students completed teacher-assigned tasks and received feedback. **Data Cleaning.** For the data safety, sensitive personally identifiable information (PII) was removed, retaining only relevant content related to answering questions. We also applied dual filtering using OCR tools and the GPT-4o-mini model to remove low-quality scratchwork images, such as illegible text or significant blurring. Entries with incomplete questions or missing answers were deleted, and text formatting was corrected. Questions containing images were excluded to simplify the initial recognition task. **Diversity Sampling.** To maintain diversity, only one instance per identical student answer for the same question was retained, resulting in around 3,400 distinct questions from an initial pool of about 1.1 million entries.

Part II. Answer Pre-annotation To reduce human workload and accelerate annotation efficiency, we adopted a human-computer collaborative approach for data annotation, referencing methods from previous works [32,34]. Leveraging the gpt-4o-2024-05-13 model, known for its robust performance in generating preliminary annotations, we created initial Error Cause Explanation and Error Cause Classification responses for each question.

Part III. Expert Human Annotation Our labeling pipeline combined advanced automated methods with expert human validation to ensure high annotation quality. We engaged five professional mathematics teachers based in Beijing, each possessing over three years of teaching experience at primary and middle school levels. Teachers were remunerated at a rate of at least 60 RMB per hour. The annotation workload was strategically divided, with three teachers focusing on primary-level questions and two dedicated to middle-school-level queries. The annotation procedure was systematically structured into three core stages:

Stage 1: Human Annotation Training. Annotators were extensively trained by the researchers to revise and validate GPT-generated annotations. Training

Table 1: Dataset statistics for the ScratchMath

Metric	Primary	Middle
Total Samples	1479	241
Grade Distribution	1-2: 23.0% 3-4: 46.3% 5-6: 30.7%	7: 31.5% 8: 33.6% 9: 34.9%
Problem Q and Solution S		
Unique Problem	1279	241
Avg Q Token	61.1	61.5
Avg S Token	139.4	175.6
Error Cause Explanation E		
Avg E Token	53.4	48.4
Min E Token	4	26
Max E Token	162	116

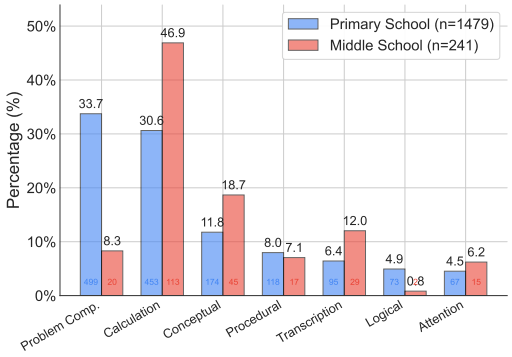


Fig. 4: Distribution of error cause classification labels for primary school and middle school problems

sessions included detailed guidance and annotation rules clearly articulated through example image prompts.

Stage 2: Trial Annotation. Moreover, annotators underwent trial annotation sessions using a standardized set of 30 questions. Post-session discussions facilitated clarification, resolution of uncertainties, and refinement of annotation guidelines, a process that was iterated until an inter-annotator agreement (IAA) of over 90% was achieved by the annotators on this standardized set, ultimately ensuring consistency and accuracy in labeling.

Stage 3: Formal Annotation. Following two comprehensive team meetings to finalize annotation protocols, annotators commenced formal labeling. The annotation process for all 3,400 questions was completed within one month.

Part IV. Post-annotation Quality Verification To further enhance dataset quality, post-annotation verification involved two additional screening phases. Firstly, scratchwork entries identified by annotators as low-quality were eliminated. Secondly, entries where the error cause or classification was indeterminate were discarded, culminating in the final, high-quality dataset with 1,720 entries.

3.3 Data Statistics

Our dataset includes 1,720 math problems, spanning 1,479 primary and 241 middle school problems, carefully selected to ensure rich coverage and representativeness. Detailed statistics, including precise grade distributions and comprehensive token counts for questions, solutions, and error explanations, are concisely presented in Table 1. A notable diversity is evident in error distributions across educational levels (see Figure 4), highlighting distinct challenges students encounter at different stages of learning. To ensure educational relevance and alignment, mathematical topics are categorized according to the authoritative Chinese *Compulsory Education Curriculum Plan and Standards (2022 edition)* (Table 2). Primary-level

Table 2: Distribution and Examples of Mathematical Topics by Category in ScratchMath (translated in English)

Category	Pri. (%)	Mid. (%)	Description and Examples
Numbers and Expressions	45.8	30.7	Arithmetic operations, fractions, decimals, algebraic identities. *e.g., Multiples of 3, Simplified addition*
Equations and Functions	1.9	43.6	Equations, inequalities, function analysis. *e.g., Quadratic equations, Linear functions*
Geometry and Measurement	25.0	20.2	Areas, perimeters, volumes, geometry. *e.g., Isosceles trapezoid area, Pythagorean theorem*
Applied Mathematics	25.4	1.6	Practical math, unit conversions, financial calculations. *e.g., Cost calculations, Average speed*
Statistics and Probability	1.9	4.0	Data collection, probability estimation. *e.g., Statistical tables, Frequency probability*

questions predominantly address foundational areas such as numbers, expressions, geometry, and applied mathematics, while middle-school-level problems delve deeper into equations, functions, and advanced algebraic concepts.

4 Experiments

4.1 Experiment Setup

Evaluated Models We selected 16 representative MLLMs for benchmarking on our dataset, covering a wide spectrum of model sizes and architectures. The evaluated models include 10 open-source models: Qwen2.5-VL (7B, 72B) [3], DeepSeek-VL2 [28], Phi-4-Multimodal [1], Llama-3.2-Vision (11B, 90B) [9], Gemma-3 [23], Skywork-R1V [18], QVQ [24], and InternVL2.5 [7]; as well as 6 proprietary models: Gemini 2.0 Flash (Flash-Lite, Flash Thinking) [22], GPT-4o (GPT-4o mini, o4-mini[6]) [10].

Prompting and Hyperparameters For consistency and fairness in comparison, we standardized the prompting approach across all tested models. Specifically, we utilized a structured prompt during testing. To further assess prompting effects, we conducted additional Chain-of-Thought (CoT) experiments, revealing some improvements in ECC task performance. To ensure reproducibility and comparability, we set the generation temperature to 0 (greedy decoding), the maximum output length to 2048 tokens, and evaluated open-source models using NVIDIA A800-80G GPUs. For proprietary reasoning models, we adopted their recommended default temperature settings, and additional experiments confirmed minor performance fluctuations when employing higher temperature settings.

Validation of LLM-as-a-judge As described in §3.1, we employed the LLM-as-a-judge metric to evaluate the Error Cause Explanation (ECE) task. To validate its reliability, we conducted an experiment using 70 randomly sampled ECE cases,

[6] https://openai.com/index/introducing-o3-and-o4-mini/

Table 3: **Performance of state-of-the-art MLLMs on ScratchMath**. We use weighted-average accuracy for the ECC task. **Boldface** and underline mark the best and second-best results for each metric, reported *separately* for proprietary vs. open-source groups. †: activated parameters of MoE model. * indicates *reasoning* models.

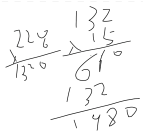
Model	#Params	Error Cause Explanation			Error Cause Classification			Average	Avg Rank
		Primary	**Middle**	**Rank**	**Primary**	**Middle**	**Rank**		
Random Guessing	–	*0.0*	*0.0*	–	*12.5*	*12.5*	–	*6.25*	–
Human Performance	–	*89.3*	*86.2*	–	*78.4*	*81.5*	–	*83.9*	–
Proprietary Models									
Gemini 2.0 Flash	–	52.2	46.9	4	38.6	**49.0**	2	46.7	3
Gemini 2.0 Flash Lite	–	36.0	32.0	8	34.6	46.5	5	37.2	5
Gemini 2.0 Flash Thinking*	–	65.9	61.0	2	**43.9**	47.3	1	54.5	2
GPT–4o	–	47.7	44.8	5	26.1	22.0	11	35.2	9
GPT–4o mini	–	3.5	1.2	16	20.1	14.1	13	9.8	16
o4-mini*	–	**71.8**	**69.7**	1	40.1	47.3	3	57.2	1
Open-Source (< 10 B)									
Qwen2.5-VL	7 B	15.6	12.4	15	21.0	11.6	14	15.2	15
DeepSeek-VL2	4.5 B†	20.9	25.7	12	16.6	7.9	16	17.8	14
Phi-4-Multimodal	5.6 B†	12.0	18.3	14	28.9	32.8	9	23.0	12
Open-Source (10-40 B)									
Llama-3.2-Vision	11 B	13.4	20.3	13	17.6	26.1	12	19.4	13
Gemma-3	27 B	38.9	26.1	9	32.2	46.1	6	35.8	7
Skywork-R1V*	38 B	37.5	33.6	7	27.7	43.2	8	35.5	8
Open-Source (> 40 B)									
QVQ*	72 B	**57.5**	**56.8**	3	12.7	17.0	15	36.0	6
Qwen2.5-VL	72 B	40.0	34.0	6	**32.5**	**49.4**	4	39.0	4
InternVL2.5	78 B	27.1	24.5	11	30.7	44.8	7	31.8	10
Llama-3.2-Vision	90 B	27.7	26.1	10	15.9	45.6	10	28.8	11

evaluated by the advanced LLM, o3-mini. Manual verification showed the judge's accuracy reached 88.6%, close to human-human inter-annotator agreement of 91.4%, confirming its suitability for our evaluation. The accuracy was below 100% primarily because the judge occasionally identified plausible yet unannotated error reasons as mismatches. We selected o3-mini due to its optimal trade-off between accuracy and evaluation cost, with the total cost of evaluating the entire benchmark (1,720 cases) being less than 10 USD.

4.2 Main Results

Table 3 summarizes the performance of MLLMs on our benchmark. Key findings include: **(1) Proprietary Models Outperform Open-source Models.** Proprietary models consistently outperform open-source models even at similar parameter scales, likely benefiting from more diverse training data. However, a considerable gap remains compared to human performance, emphasizing the benchmark's inherent challenge. **(2) Scaling Law on Both Tasks and Rea-**

Table 4: Examples (translated in English) illustrating typical model error types in ECE. "Ref. Ans." and "Stu. Ans." denote the reference and student answers, respectively. Bold text highlights the erroneous parts of the prediction.

Scratchwork	Problem Statement	Ref. Ans.	Stu. Ans.	Ground Truth	Model Prediction
Visual Recognition Failure					
	Solve for x: $4x - 3(20 - x) = 6x - 7(9 - x)$	$\frac{1}{2}$	$-\frac{1}{2}$	Student solved $-6x = -3$ as $x = -\frac{1}{2}$ instead of $\frac{1}{2}$.	...Finally $x = \frac{-3}{-6} = \frac{1}{2}$. However, **the student miscalculated** $-63 + 60$ **as** $+3$ **at this step, so they wrote** $-6x = 3$, leading to $x = \frac{3}{-6} = -\frac{1}{2}$.
Formatting Misinterpretation					
	If the two roots of $a(x + m)^2 + b = 0$ are -1 and 4, then the solutions to $a(x + m - 3)^2 + b = 0$ are $x =$_____.	2 or 7	2 and 7	Formatting error ("and" vs. "or"); also mis-shifted the quadratic by m.	... The student, in their scratchpad, **mistakenly treated the second equation** $a \cdot (x + m - 3)^2 + b = 0$ **as** $a \cdot (x - 3)^2 + b = 0$. They then directly set $x - 3 = -1$ and $x - 3 = 4$, thus obtaining $x = 2, 7$...
Misaligned Misinterpretation					
	Brick dimensions $20 \times 11 \times 6$ cm, density 1.5 g/cm^3. Find the weight in kilograms.	1.98	1980	Answered in grams without converting to kg.	...When calculating by hand, **he even converted "**1320×1.5**" to "**132×15**"**, and although the result was still 1980, **the division by 10 and multiplication by 10 in the process didn't correspond to the units**...

soning Model Superiority in ECE. Performance generally follows scaling laws, with larger models showing better results. Reasoning models specifically excel in Error Cause Explanation (ECE), highlighting their advantage in tasks demanding deeper semantic understanding. Conversely, Error Cause Classification (ECC) remains significantly more challenging across all models. **(3) Elementary Tasks Not Necessarily Easier.** While models typically perform better on primary tasks in the ECE task, primary-level performance unexpectedly falls below middle school performance in the ECC task. This could stem from less structured and harder-to-interpret handwriting in primary-level scratchwork, complicating precise error classification.

5 Further Analysis

We analyze three key research questions to deepen our understanding of model performance:
RQ1: What challenges do MLLMs face in error cause detection?
RQ2: How does problem type affect model performance?
RQ3: How does problem difficulty affect model performance?

5.1 RQ1: Challenges in Error Identification

Qualitative Analysis We conducted detailed case studies to further illustrate typical errors made by the o4-mini. Table 4 presents three representative cases categorized by the type of error: Visual Recognition Failure, Formatting Misinterpretation, and Misaligned Misinterpretation. These examples highlight specific

aspects that need improvement, such as visual processing accuracy, proper understanding of formatting requirements, and accurate inference of the student reasoning processes.

Quantitative Anlysis To explore the difficulties faced by current MLLMs, we conducted an error analysis of 100 randomly selected cases in which the strongest model (o4-mini) failed on the Error Cause Explanation task. As shown in Figure 5, we categorize these errors into six types. Key findings include: (1) The most frequent errors were related to OCR and image recognition, often stemming from unclear handwriting. (2) Models struggled significantly with accurately reconstructing students' reasoning processes, indicating limitations in logical inference. (3) Many errors involved over-inference or speculative reasoning by the models, suggesting tendencies to extrapolate beyond available evidence. In addition, to understand error patterns in smaller, open-source models, we conducted a similar analysis on Qwen2.5-VL-7B. We found a higher incidence of hallucination errors (22%) and a new category, "Model Calculation Error" (17%), indicating arithmetic reasoning difficulties specific to smaller models.

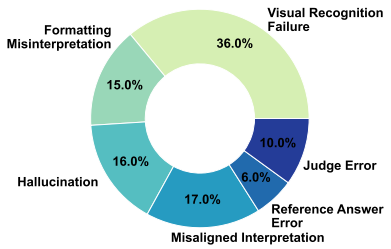


Fig. 5: Failure Cases Distribution of o4-mini on ECE

5.2 RQ2: Impact of Problem Type on Performance

Performance Across Error Cause Categories We investigated how problem categories influence model performance on ECC Task, averaging scores across primary and middle school datasets. Several insights emerged from the results shown in Figure 6: (1) The top-performing models, o4-mini and Gemini 2.0 Flash Thinking, notably excelled in most error categories, except Logical Reasoning and Calculation Errors, which are harder due to implicit reasoning steps and compounded errors in visual number recognition and multi-step arithmetic. (2) Many mod-

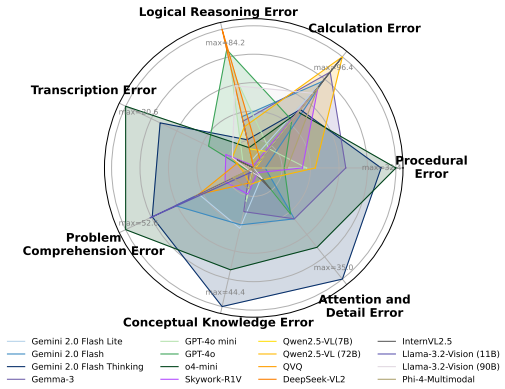


Fig. 6: Models' Performance on different ECC classes

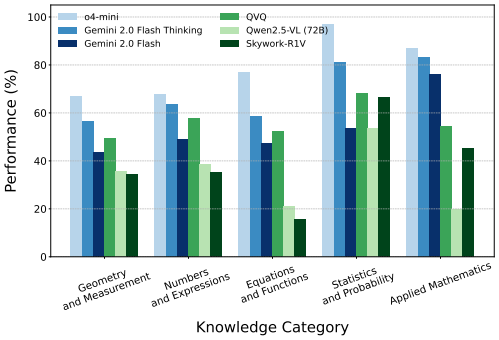


Fig. 7: Model Performance Across Math Topics in ECE Tasks (Top 3 Open-source and Proprietary Models)

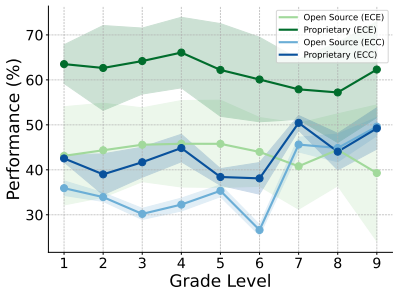


Fig. 8: Performance across grade levels, averaged for top-3 open-source and proprietary models on two tasks.

els demonstrated potential overfitting to specific error categories, particularly Logical Reasoning and Calculation Errors, indicating specialized rather than generalized error detection capabilities. (3) Procedural and Transcription errors generally posed significant challenges to all models, highlighting areas for further targeted development. Performance disparities across error types suggest varied levels of complexity inherent in different problem categories, reflecting a nuanced interaction between model architecture and problem characteristics.

Performance Across Mathematics Topics We also analyzed model performance based on the topics of math problems (introduced in Table 2). Figure 7 illustrates several notable findings: (1) Proprietary models consistently showed strong and stable performance across all knowledge categories, with o4-mini significantly outperforming others. (2) Open-source models exhibited varied performance, with Skywork-R1V notably stronger in Statistics and Probability and Applied Mathematics, yet weaker in Equations and Functions. The disparity in open-source model performance indicates a potential specialization or bias in training data, highlighting the importance of diverse and comprehensive training datasets.

5.3 RQ3: Impact of Difficulty on Model Performance

Additionally, we examined the impact of educational grade level on model performance by selecting the top three open-source and proprietary models from ECC and ECE tasks. The analysis, depicted in Figure 8, suggests several trends: (1) On the ECE task, model performance exhibits a slight downward trend as grade level increases, indicating greater complexity or ambiguity in higher-grade solutions. (2) Conversely, performance on the ECC task generally improves with increasing grade levels, possibly due to clearer and more structured scratchwork provided by older students. Through detailed sampling, we observed that middle-school scratchwork contains clearer sequential steps and standardized notation

compared to elementary-level responses. These structured presentations facilitate error classification. (3) Proprietary models consistently outperform open-source models across all grades, underscoring the potential advantages of more diverse training data.

6 Conclusion and Future Work

In summary, ScratchMath advances educational AI by introducing a comprehensive multimodal benchmark that exposes the limitations of current MLLMs in diagnosing student errors. It highlights the urgent need for developing models that better align with educators' analytical processes. A limitation of this work is that all samples were collected from Chinese students using a single online education platform, which may constrain generalizability to other languages, demographics, and educational contexts. Future work includes enhancing model training by incorporating explicit error-type predictions, integrating advanced visual recognition techniques, and exploring step-by-step reasoning alignment to improve model interpretability. Expanding the dataset across diverse populations and educational settings, and leveraging cross-cultural comparisons, could yield deeper insights into universally effective educational strategies.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Abouelenin, A., Ashfaq, A., Atkinson, A., Awadalla, H., Bach, N., Bao, J., Benhaim, A., Cai, M., Chaudhary, V., Chen, C., et al.: Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. arXiv preprint arXiv:2503.01743 (2025)
2. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv preprint arXiv:2308.12966 (2023)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Botelho, A.F., Baral, S., Erickson, J.A., Benachamardi, P., Heffernan, N.T.: Leveraging natural language processing to support automated assessment and feedback for student open responses in mathematics. Journal of Computer Assisted Learning (2023)
5. Caraeni, A., Scarlatos, A., Lan, A.: Evaluating gpt-4 at grading handwritten solutions in math exams. arXiv preprint arXiv:2411.05231 (2024)
6. Chen, S., Li, B., Niu, D.: Boosting of thoughts: Trial-and-error problem solving with large language models. arXiv preprint arXiv:2402.11140 (2024)
7. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)

8. Davalos, E., Zhang, Y., Srivastava, N., Salas, J.A., McFadden, S., Cho, S.J., Biswas, G., Goodwin, A.: Llms as educational analysts: Transforming multimodal data traces into actionable reading assessment reports. arXiv preprint arXiv:2503.02099 (2025)
9. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
10. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
11. Jiang, B., et al.: Llms can find mathematical reasoning mistakes by pedagogical chain-of-thought. In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) (2024)
12. Kinder, A., Briese, F.J., Jacobs, M., Dern, N., Glodny, N., Jacobs, S., Leßmann, S.: Effects of adaptive feedback generated by a large language model: A case study in teacher education. Computers and Education: Artificial Intelligence **8**, 100349 (2025)
13. Lee, J., Baraniuk, R., Lan, A.S.: Training llm-based tutors to improve student learning outcomes in dialogues. arXiv preprint arXiv:2503.06424 (2025)
14. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. arXiv preprint arXiv:2304.08485 (2023)
15. Liu, T., Chatain, J., Ruzika, S., Kuhn, J., Küchemann, S.: Ai-assisted automated short answer grading of handwritten university level mathematics exams. arXiv preprint arXiv:2408.11728 (2024)
16. Mahdavi, M., Zanibbi, R., Mouchère, H., Viard-Gaudin, C., Garain, U.: Icdar 2019 crohme + tfd: Competition on recognition of handwritten mathematical expressions and typeset formula detection. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR). pp. 1533–1538 (2019)
17. McNichols, H., Zhang, M., Lan, A.: Algebra error classification with large language models. In: Proceedings of the 24th International Conference on Artificial Intelligence in Education (AIED) (2023)
18. Peng, Y., Wang, X., Wei, Y., Pei, J., Qiu, W., Jian, A., Hao, Y., Pan, J., Xie, T., Ge, L., et al.: Skywork r1v: Pioneering multimodal reasoning with chain-of-thought. arXiv preprint arXiv:2504.05599 (2025)
19. Shi, W., Hu, Z., Bin, Y., Liu, J., Yang, Y., Ng, S.K., Bing, L., Lee, R.K.W.: Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. In: Findings of EMNLP (2024)
20. Song, D., Chen, S., Chen, G.H., Yu, F., Wan, X., Wang, B.: Milebench: Benchmarking MLLMs in long context. In: COLM (2024)
21. Song, D., Lai, S., Chen, S., Sun, L., Wang, B.: Both text and images leaked! a systematic analysis of data contamination in multimodal LLMs. In: Findings of EMNLP (2025)
22. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
23. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
24. Team, Q.: Qvq: To see the world with wisdom (December 2024), https://qwenlm.github.io/blog/qvq-72b-preview/

25. Truong, T.N., Nguyen, C.T., Zanibbi, R., Mouchère, H., Nakagawa, M.: A survey on handwritten mathematical expression recognition: The rise of encoder-decoder and GNN models. Pattern Recognition **153**, 110531 (2024)
26. Wang, K., Pan, J., Shi, W., Lu, Z., Ren, H., Zhou, A., Zhan, M., Li, H.: Measuring multimodal mathematical reasoning with the math-vision dataset. In: NeurIPS Datasets and Benchmarks (2024)
27. Wang, S., Xu, T., Li, H., Zhang, C., Liang, J., Tang, J., Yu, P.S., Wen, Q.: Large language models for education: A survey and outlook. IEEE Signal Processing Magazine **42**(6), 51–63 (2026)
28. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024)
29. Yan, Y., Su, J., He, J., Fu, F., Zheng, X., Lyu, Y., Wang, K., Wang, S., Wen, Q., Hu, X.: A survey of mathematical reasoning in the era of multimodal large language model: Benchmark, method & challenges. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 11798–11827 (2025)
30. Yan, Y., Wang, S., Huo, J., Li, H., Li, B., Su, J., Gao, X., Zhang, Y.F., Xu, T., Chu, Z., et al.: Errorradar: Benchmarking complex mathematical reasoning of multimodal large language models via error detection. arXiv preprint arXiv:2410.04509 (2024)
31. Yan, Y., Wang, S., Huo, J., Yu, P.S., Hu, X., Wen, Q.: MathAgent: Leveraging a mixture-of-math-agent framework for real-world multimodal mathematical error detection. In: ACL (2025)
32. Zhang, R., Jiang, D., Zhang, Y., Lin, H., Guo, Z., Qiu, P., Zhou, A., Lu, P., Chang, K.W., Gao, P., Li, H.: MathVerse: Does your multi-modal llm truly see the diagrams in visual math problems? In: ECCV (2024)
33. Zhang, Y.F., Li, H., Song, D., Sun, L., Xu, T., Wen, Q.: From correctness to comprehension: Ai agents for personalized error diagnosis in education. arXiv preprint arXiv:2502.13789 (2025)
34. Zhou, J., et al.: Mathscape: Evaluating mllms in multimodal math scenarios through a hierarchical benchmark. arXiv preprint arXiv:2408.07543 (2024)