CHECK-MAT: A Benchmark for Evaluating Vision-Language Models on Grading Hand-Written Mathematical Solutions

Author

Ruslan Khrulev

Affiliations

Lomonosov Moscow State University, Huawei

Publication

MathNLP Workshop at EMNLP 2025

While modern Vision-Language Models (VLMs) excel at *solving* math problems, their ability to *grade* human-written solutions remains underexplored. We introduce **CHECK-MAT**, a new benchmark of 122 hand-written exam solutions with official expert grades, to evaluate a VLM's capacity for nuanced error identification and rubric alignment. Our findings reveal that even state-of-the-art models struggle with the complexities of human reasoning, highlighting a critical gap for future AI-assisted educational tools.

1 Introduction

Current benchmarks like MATH and GSM8K test a model's ability to generate a correct final answer. However, in real-world education, the focus is on assessing the entire problem-solving process. **A critical capability gap exists**: can models *understand, analyze, and grade human-generated solutions according to a predefined rubric*?

Our Contribution: We address this gap by proposing a new task - **Solution Assessment,** and a benchmark to measure it, using high-stakes exam materials from the Russian Unified State Exam (EGE).

2 Objective

To create and validate a new benchmark, CHECK-MAT, for evaluating the ability of modern Vision-Language Models to perform **automated, rubric-aligned grading of hand-written mathematical solutions**.

3 Methodology

- **Dataset**: Compiled *122 problem solutions* from the official EGE expert guide, each with a scanned image, problem statement, and official expert score
- **Evaluated Models**: Arcee AI Spotlight, Google Gemini series (2.0 Flash, 2.5 Flash Preview), OpenAI o4-mini, and Qwen 2.5 VL 32B.
- **Evaluation Modes**: • *Baseline*: Solution Image + Problem • *With Answer*: Baseline + Correct Final Answer • *With True Solution*: Baseline + Full Correct Solution

4 Results

OpenAI o4-mini consistently achieved the highest performance, with a peak accuracy of **56.56%** in the "With Answer" mode. Providing the correct final answer (Mode 2) generally improves performance across models. However, **overall accuracy remains modest**, indicating significant challenges in aligning with human grading criteria.

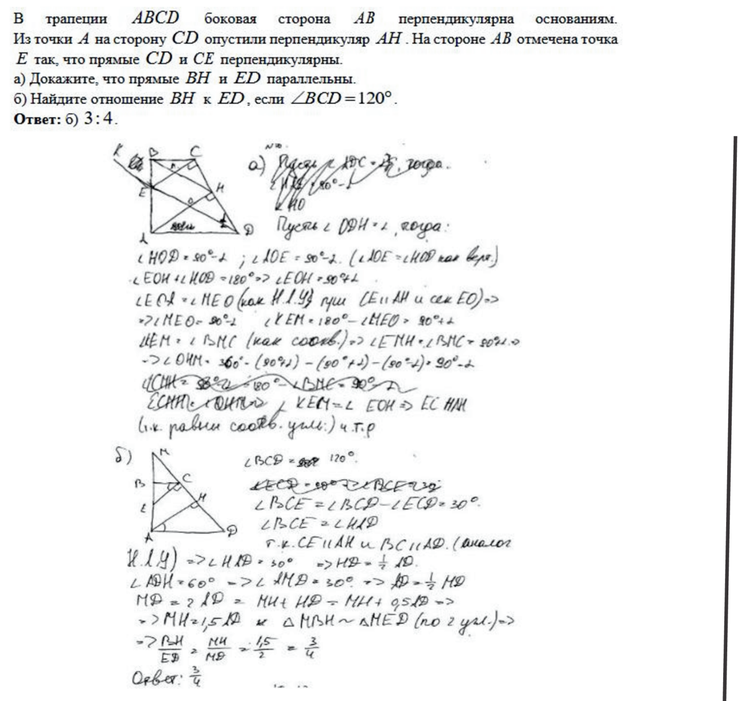
Model	Mode	Acc. (%)	Qual. (%)	Dist.
Arcee AI Spotlight	Without Answer	27.87	64.48	1.04
Google Gemini 2.0 Flash	With Answer	47.54	74.04	0.75
Google Gemini 2.0 Flash Lite	With True Solution	38.52	70.22	0.84
Google Gemini 2.5 Flash Preview	With True Solution	45.90	71.35	0.79
Google Gemini 2.5 Flash Preview:thinking	With True Solution	43.44	65.92	0.99
OpenAI o4-mini	With Answer	**56.56**	**78.17**	**0.60**
Qwen 2.5 VL 32B	With True Solution	43.44	70.49	0.81

Table 1: Overall performance of all models across three evaluation modes

5 Analysis

Task ID	Domain	Count	Score Range
13	Trigonometric equations	21	0–2
14	Stereometry	18	0–3
15	Logarithmic inequalities	19	0–2
16	Financial mathematics problems	17	0–2
17	Planimetry	15	0–3
18	Parameterised equations	16	0–4
19	Number theory/ combinatorics	16	0–4

Table 2: Benchmark breakdown by task type.



В трапеции $ABCD$ боковая сторона AB перпендикулярна основаниям. Из точки A на сторону CD опустили перпендикуляр AH. На стороне AB отмечена точка E так, что прямые CD и CE перпендикулярны.
а) Докажите, что прямые BH и ED параллельны.
б) Найдите отношение BH к ED, если $\angle BCD = 120°$.
Ответ: б) 3:4.

Figure 2: Random sample from dataset.

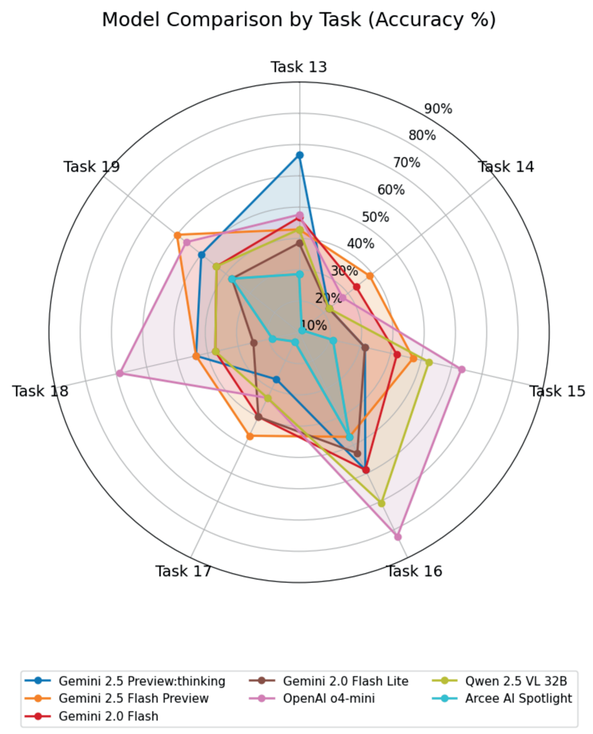


Figure 1: Radar chart showing model Accuracy across all seven task types.

6 Conclusion

- Current VLMs **are not** yet robust enough for reliable, automated grading of complex, hand-written mathematical solutions.
- Models **struggle to align** with detailed grading rubrics and often fail to identify subtle yet critical errors.
- **Future Work**: This benchmark opens new avenues for research in AI-assisted assessment, focusing on improving fine-grained reasoning and alignment with human-centric evaluation criteria.

Related literature

- Hendrycks, D., et al. (2021). Measuring Mathematical Problem Solving with the MATH Dataset.
- Cobbe, K., et al. (2021). Training Verifiers to Solve Math Word Problems.
- Kasneci, G., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education.