Towards Accessible Education: A Benchmark for AI-Powered Assistive Grading

Khrulev Ruslan
Lomonosov Moscow State University
Russia, Moscow
ra.khrulev@gmail.com

Abstract

AI-powered assistive technologies hold immense potential to create more accessible and equitable educational environments. A key barrier, however, is the laborious and subjective process of grading complex, handwritten assignments, which limits the accessibility of timely and consistent feedback for students and overwhelms educators. While Vision-Language Models (VLMs) are a promising solution, their readiness to function as a reliable assistive tool must be rigorously evaluated to prevent unfair outcomes. To this end, we introduce CHECK-MAT, the first benchmark designed to assess the capabilities of VLMs as an assistive technology for grading handwritten, multi-step mathematical solutions from a real-world national exam. Our benchmark is composed of 122 scanned solutions from the Russian Unified State Exam (EGE), complete with official expert grades, providing a realistic testbed for this accessibility challenge. We evaluate seven modern VLMs and find their performance remains significantly below the level required for reliable use, especially in understanding the logical steps of human reasoning. Our findings chart a path for future research, highlighting the core challenges that must be overcome to develop the next generation of trustworthy, fair, and genuinely assistive AI technologies that can empower both educators and learners. You can find code in https://github.com/Karifannaa/Auto-check-EGE-math.

1. Introduction

AI-powered assistive technologies hold immense potential to make education more accessible, yet a significant frontier remains in automatically assessing complex, handwritten mathematical solutions. Specifically, they can empower educators by automating the laborious grading process, freeing up their time for more personalized student interaction. For students, especially those with learning disabilities or requiring more frequent feedback, such tools can provide instant, consistent, and private evaluations, fostering a more inclusive learning environment. While numerous benchmarks test a model's ability to solve mathematical problems—such as *MATH* [4] and *GSM8K* [3]—the crucial task of grading human solutions is less explored. Recent works like *Fermat* [8] and *MathCCS* [7] have made progress in error diagnosis, but often rely on synthetic data or focus on error classification rather than applying a holistic, multi-point grading rubric from a real-world, high-stakes exam. To fill this gap and create a tool for developing more equitable assistive technologies, we introduce CHECK-MAT, a novel benchmark derived from the Russian Unified State Exam (EGE).

Our benchmark is uniquely composed of 122 real, scanned handwritten solutions from the EGE, paired with official, multi-point scores assigned by expert human graders. The Russian Unified State Exam (EGE) was specifically chosen for several reasons. Firstly, its official expert guides are publicly available and provide a rich collection of real, diverse student solutions. Secondly, and most critically, these guides contain highly detailed, multi-point grading rubrics and expert justifications for each score. This granular level of detail provides a unique and rigorous ground truth for evaluating the nuanced assessment capabilities of VLMs, a feature not as readily available for many other standardized exams. This setup challenges Vision-Language Models (VLMs) to move beyond mere problem-solving and emulate the nuanced assessment process of experts—a vital capability for any true assistive grading tool. We evaluate seven state-of-the-art VLMs and find that their performance is currently insufficient for reliable deployment, highlighting systematic weaknesses in handling geometric reasoning and complex handwritten notation. CHECK-MAT thus provides a vital diagnostic tool for the community to measure progress and pave the way for more robust, fair, and genuinely assistive educational technologies.

2. Benchmark Design and Dataset

Our benchmark is designed to evaluate Vision-Language Models (VLMs) on their ability to assess handwritten mathematical solutions, a task that requires a deep understanding

Table 1. Benchmark breakdown by task type.

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

of both visual information and mathematical reasoning. The core of our benchmark is a unique dataset derived from the Russian Unified State Exam (EGE), specifically focusing on the second part of the mathematics exam, where students provide detailed, handwritten solutions.

2.1. Dataset Sourcing and Characteristics

The dataset comprises 122 problem solutions, meticulously sourced from the official EGE expert guide. This guide provides a rich collection of real student solutions, along with expert-assigned grades and detailed justifications for those grades. Each entry in our dataset includes:

- **Scanned Handwritten Solution:** An image of the students complete handwritten solution, often spanning multiple pages, capturing the nuances of human handwriting, diagrams, and mathematical notation.
- **Problem Statement:** The original text of the mathematical problem, providing context for the solution.
- **Expert Grade:** The official score assigned by human experts according to the EGE grading criteria.
- **Reference-Based Expert Evaluation:** Includes the final score assigned by a human expert. The assessment is based on a provided *gold-standard* solution and a granular grading rubric, which are available for each task to ensure a transparent and replicable evaluation process.

The solutions cover a range of mathematical topics typically found in EGE, including algebra, geometry, trigonometry, and calculus, ensuring a diverse set of challenges for the evaluated models. The handwritten nature of the solutions introduces significant variability in terms of handwriting styles, penmanship, and layout, requiring robust VLM capabilities for accurate interpretation.

2.2. Mathematical Domains and Task Types

Each task corresponds to a standard EGE problem type requiring a written solution with reasoning. Table 1 provides an overview of the tasks, including their domain, a brief description, the number of solution samples in our dataset, and the score range (points) for each task.

2.3. Grading Criteria and Assessment Focus

The central point of the EGE assessment process is the clearly defined grading criteria for each task. These criteria specify how points are awarded or deducted based on the correctness of the solution steps, the validity of the reasoning, and the accuracy of the final answer. Our benchmark leverages these criteria as the ground truth for evaluation. The primary focus is not on whether the model can solve the problem itself, but rather on its ability to:

- **Understand the Solution Flow:** Comprehend the logical progression of the students solution, including intermediate steps and derivations.
- **Identify Errors:** Accurately pinpoint mathematical errors, logical flaws, or omissions within the handwritten solution.
- **Apply Grading Rubrics:** Assess the identified errors and correct parts of the solution against the specific EGE grading criteria to assign an appropriate score.

This emphasis on assessment rather than problem-solving distinguishes our benchmark from many existing math-focused datasets and provides a more realistic evaluation of AI potential in educational grading scenarios.

3. Experimental Setup

We evaluated seven diverse Vision-Language Models (VLMs): Arcee AI Spotlight [1], Google's Gemini series (2.0 Flash, 2.0 Flash Lite, 2.5 Flash Preview, and its "thinking" variant) [6], OpenAI o4-mini [5], and Qwen 2.5 VL 32B [2]. The evaluation was structured around three modes to assess capabilities under different levels of context: **(1) Without Answer**, where models received only the problem and the handwritten solution; **(2) With Answer**, where the correct final answer was also provided; and **(3) With True Solution**, where a complete, ”gold standard” reference solution was included. For each mode, models were given tailored prompts including the EGE grading criteria and were instructed to output a score in a structured format for automated analysis. Full details on the data curation and prompt templates are available in our public repository.

4. Results

Our evaluation of seven Vision-Language Models across three distinct evaluation modes provides insights into their capabilities in assessing handwritten mathematical solutions.

4.1. Metrics

We report three complementary metrics:

Accuracy (Exact Match:) Percentage of cases where the predicted score exactly matches the expected score:

$$\text{Accuracy} = \frac{\text{Exact Matches}}{\text{Total Evaluations}} \times 100\%.$$

Quality Score: Normalized closeness between predicted and expected scores:

$$\text{Quality Score} = 100\% \times \left(1 - \frac{|S_{\text{pred}} - S_{\text{true}}|}{S_{\max}}\right),$$

where $S_{\max} \in \{2, 3, 4\}$ is the task-specific maximum.

Average Score Distance:

$$\text{Avg. Distance} = \frac{1}{n} \sum_{i=1}^{n} \left| S_{\text{pred},i} - S_{\text{true},i} \right|.$$

4.2. Performance Analysis

As can be seen from Table 2, OpenAI o4-mini consistently demonstrates the highest performance across all evaluation modes, achieving the best Accuracy (56.56% with Answer) and Quality Score (78.17% with Answer), and the lowest Average Score Distance (0.60 with Answer). This suggests that OpenAI's model possesses superior capabilities in understanding handwritten solutions and applying grading criteria compared to other evaluated models.

Among other models, Google Gemini 2.0 Flash also shows strong performance, particularly in the **With Answer** and **With True Solution** modes, indicating its ability to effectively leverage additional context. Models like Arcee AI Spotlight and Qwen 2.5 VL 32B exhibit lower accuracy and higher score distances, suggesting that while they can process the visual input, their mathematical reasoning and grading alignment are less precise. The *thinking* variant of Google Gemini 2.5 Flash Preview, despite its higher cost and longer average time, does not consistently outperform its non-*thinking* counterpart, raising questions about the efficacy of its enhanced reasoning capabilities for this specific task.

A detailed breakdown of performance by task type, illustrated in Figure 1, reveals significant variations. It is evident that algebraic tasks (13 and 15) are handled more effectively by most models. In contrast, both geometry categories (14 — stereometry, 17 — planimetry) consistently yield poorer agreement with human graders. We hypothesise that current VLMs still struggle to map free-hand diagrams onto the rigorous spatial reasoning chains required by the EGE rubric. The full per-task scores for all models can be found in repository.

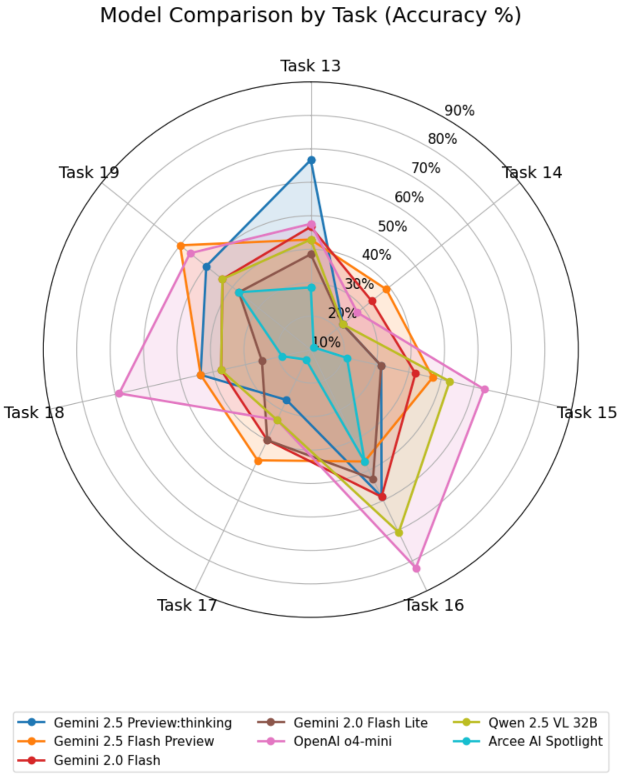

Figure 1. Radar chart showing model Accuracy (%) in the With True Solution mode across all seven task types. The outer edge represents a perfect score. This visualization highlights the models' strengths and weaknesses on different mathematical domains.

4.3. Impact of Evaluation Modes

One of the most interesting findings is the varied impact of the evaluation modes on model performance. For some models, providing additional context (correct answer or true solution) significantly improved their performance. For instance, Google Gemini 2.0 Flash showed a notable increase in Accuracy when provided with the correct answer (from 36.89% to 47.54%). This suggests that these models can effectively leverage external information to refine their assessment, indicating a capacity for conditional reasoning. However, this improvement was not universal; Arcee AI Spotlight, for example, saw a slight decrease in performance with additional context, which might indicate issues with how it integrates or prioritizes external information versus its internal analysis of the handwritten solution.

The **With True Solution** mode, while providing the most comprehensive context, did not consistently lead to the best performance across all models. This could be attributed to several factors: the models might struggle with effectively comparing a student's potentially divergent solution path with a provided reference solution, or lack the complexity sufficient to fully leverage the detailed informa-

Table 2. Overall performance of all models across three evaluation modes. The best result for each combination of mode and metric is shown in bold, and the second best result is underlined.

Model	Provider	Mode	Acc. (%)	Qual. (%)	Avg. Dist.	Cost ($)	Time (s)
Arcee AI Spotlight	Arcee AI (via OpenRouter)	Without Answer	27.87	64.48	1.04	**0.01**	8.80
		With Answer	26.23	63.18	1.09	**0.01**	6.99
		With True Solution	25.41	59.22	1.16	**0.01**	6.98
Google Gemini 2.0 Flash	Google	Without Answer	36.89	71.04	0.84	0.14	4.56
		With Answer	47.54	74.04	0.75	0.14	4.82
		With True Solution	46.72	75.82	0.71	0.21	3.13
Google Gemini 2.0 Flash Lite	Google	Without Answer	31.97	64.96	1.00	0.04	**3.08**
		With Answer	35.25	67.83	0.90	0.04	**3.13**
		With True Solution	38.52	70.22	0.84	0.04	**3.09**
Google Gemini 2.5 Flash Preview	Google	Without Answer	44.26	71.04	0.81	0.32	16.08
		With Answer	40.98	70.49	0.82	0.30	14.92
		With True Solution	45.90	71.35	0.79	0.34	11.67
Google Gemini 2.5 Flash Preview:thinking	Google	Without Answer	40.16	64.30	1.05	0.60	39.48
		With Answer	42.62	66.44	0.99	0.62	39.98
		With True Solution	43.44	65.92	0.99	0.78	47.59
OpenAI o4-mini	OpenAI	Without Answer	**55.74**	**75.55**	**0.66**	2.18	39.62
		With Answer	**56.56**	**78.17**	**0.60**	2.02	32.94
		With True Solution	**54.10**	**76.16**	**0.66**	2.28	58.47
Qwen 2.5 VL 32B	Alibaba Cloud (via OpenRouter)	Without Answer	31.15	62.09	1.09	0.46	22.97
		With Answer	30.33	61.95	1.08	0.46	23.27
		With True Solution	43.44	70.49	0.81	0.63	27.55

tion in a reference solution when the student's approach deviates significantly. This highlights a crucial area for future research: developing VLMs that can perform robust comparative analysis between a student's solution and a reference, even when the two solution paths differ.

5. Discussion and Conclusion

Our evaluation provides a unique perspective on VLM capabilities, revealing a substantial gap between current model performance and human expert-level grading in a real-world assessment scenario (highest accuracy: 56.56%). The primary limitations stem from two interconnected challenges: flawed visual interpretation of complex handwriting, which leads to error propagation, and a lack of the deep symbolic reasoning required to align with nuanced grading rubrics. While our work relies on zero-shot prompting, future directions should include fine-tuning on larger datasets and developing more robust methods for contextual reasoning. Furthermore, we acknowledge a potential limitation regarding the linguistic and educational context. The benchmark is based on the Russian EGE, and the evaluated VLMs are predominantly trained on English-language data. Future research should extend this benchmark to include assignments from different linguistic and educational systems to assess the cross-cultural generalizability and fairness of these assistive technologies.

As a step towards creating AI-powered assistive technologies for education, CHECK-MAT serves a crucial diagnostic role. Our results highlight the primary risk of premature deployment—unfair student outcomes—and position our benchmark as a tool for auditing the fairness and reliability of these systems. We advocate for a "human-in-the-loop" approach where AI assists, rather than replaces, human experts. This human-in-the-loop model is central to accessibility; it ensures that the technology serves as a genuine aid, reducing grading bottlenecks and potential biases, thus allowing educators to focus on the pedagogical needs of every student. Ultimately, the path to genuinely accessible educational tools requires not only improving accuracy but also enhancing the explainability and transparency of model decisions, which remains a key challenge for future work.

6. Conclusion

Our work provides a realistic evaluation of current VLMs as an assistive technology for education. While they are not yet ready for autonomous deployment, our benchmark, CHECK-MAT, lays the groundwork for developing the next generation of fair, reliable, and truly accessible AI tools that can support both students and educators.

7. License

The source code and dataset for this research are available under the MIT License. This permissive license allows for reuse, modification, and distribution, both in academic and commercial settings, provided that the original copyright and license notice are included.

References

[1] Arcee.ai. Arcee blog. https://www.arcee.ai/blog, 2025. Accessed: 2025-07-06. 2

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang,

Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 2

[3] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021. 1

[4] Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021. 1

[5] OpenAI. Introducing o3 and o4-mini. https://openai.com/index/introducing-o3-and-o4-mini/, 2025. Accessed: 2025-07-06. 2

[6] Gemini Team et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[7] Yitong Wu, Yuan Li, Yujun Li, and Wang Zhou. Mathccs: A new benchmark for mathematical classification and constructive suggestions. *arXiv preprint arXiv:2405.17642*, 2024. 1

[8] Zhen Yuan, Yifan Zhang, Jing Liu, Yuxiang Wang, Jie Zhang, Hanwang Liu, and Tat-Seng Chua. Fermat: A benchmark for evaluating vlm's ability in factual error correction of handwritten math solutions. *arXiv preprint arXiv:2405.10100*, 2024. 1