CHECK-MAT: Probing the Mathematical Reasoning and Rubric-Alignment of Vision-Language Models on Handwritten Solutions

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

Abstract

The application of contemporary NLP models for inference over mathematical text remains a critical and under-explored area. While Vision-Language Models (VLMs) have shown promise, a significant gap exists in their ability to perform nuanced, rubric-based assessment of handwritten mathematical arguments, a task requiring the joint interpretation of visual, textual, and symbolic modalities. This paper directly addresses the need for robust evaluation tasks in this domain. This paper introduces CHECK-MAT, a new benchmark and methodology for the automated, rubric-based assessment of handwritten mathematical solutions using Vision-Language Models (VLMs). Composed of 122 real-world solutions from a high-stakes national exam, CHECK-MAT evaluates the capacity of VLMs to emulate expert graders by identifying logical flaws and applying detailed grading rubrics. Our systematic evaluation of seven state-of-the-art VLMs serves as a direct instance of probing the mathematical understanding of state-of-the-art models. We reveal key limitations in their ability to parse complex notation and align with human grading rubrics, which we frame as a challenge in understanding the linguistic analysis of mathematical discourse. Our work contributes a robust benchmark to the NLP community and offers critical insights for developing models with more sophisticated mathematical reasoning capabilities. You can find code in https://github.com/Karifannaa/Auto-check-EGE-math.

1 Introduction

The articulation of mathematical arguments is a fundamental part of scientific reasoning and communication. As Large Language Models (LLMs) and Vision-Language Models (VLMs) become more capable, their application to understanding and evaluating these arguments is a key area of research in NLP. However, a significant gap exists in their ability to perform inference over complex, multimodal mathematical text, such as handwritten student solutions. While benchmarks like Fermat (Yuan et al., 2024) focus on error detection and localization, and MathCCS (Wu et al., 2024) on systematic error analysis for feedback, CHECK-MAT addresses a different, complementary challenge: **rubric alignment**. Our benchmark tests a VLM's ability to not just identify flaws, but to map the overall quality of a complex, multi-step solution onto a granular, official scoring rubric, emulating the holistic judgment of a human grader.

Instead of merely testing a model's problem-solving capabilities, our benchmark probes its capacity for nuanced understanding of human thought processes, error identification, and adherence to structured assessment rubrics. This is crucial for developing AI systems that can genuinely assist in educational assessment, providing scalable support for tasks that are currently bottlenecked by manual expert grading. Our work makes the following contributions:

- We introduce CHECK-MAT, a new public benchmark that addresses the need for evaluation tasks requiring the joint interpretation of different modalities (handwritten notation, natural language problems, and equational rubrics) in mathematical text.
- We provide a comprehensive evaluation of seven state-of-the-art VLMs, offering a rigorous analysis that directly probes the mathematical understanding of state-of-the-art models and identifies specific weaknesses, particularly in geometric reasoning.
- We frame the task of rubric-based grading as a form of linguistic analysis of mathematical discourse, providing a challenging testbed for developing models that can follow complex, human-defined argumentation relations.

Proceedings of The 3rd Workshop on Mathematical Natural Language Processing (MathNLP 2025), pages 77–94
November 9, 2025 ©2025 Association for Computational Linguistics

2 Related Work

Our work is situated at the intersection of automated mathematical assessment, multimodal reasoning, and diagnostic evaluation. The field has evolved significantly from its early reliance on deterministic systems to the current exploration of sophisticated AI models.

2.1 Automated Assessment as an NLP Task

The evolution of automated assessment in mathematics education has been marked by a shift from answer verification to process analysis. Early systems, built on Computer Algebra Systems (CAS) like STACK (Sangwin, 2014), excelled at verifying the symbolic equivalence of a final answer. However, they were fundamentally “correctness-focused” and could not evaluate the student’s reasoning process.

This limitation spurred the development of process-focused assessment. This modern paradigm is embodied in intelligent tutoring systems and benchmarks such as MathCCS (Wu et al., 2024), which uses real-world student data for systematic error analysis as a foundation for generating pedagogically useful feedback. Our work directly extends this trajectory by leveraging large vision-language models (VLMs) to perform diagnostic assessment on authentic, handwritten student work, posing a new, challenging task for the mathematical NLP community.

2.2 The Multimodal Challenge of Handwritten Solutions

Assessing handwritten mathematics is an inherently multimodal task. A specialized field, Handwritten Mathematical Expression Recognition (HMER), has focused on the modular task of transcribing visual notation into a structured format like LaTeX (Deng et al., 2017). This is a non-trivial problem due to the two-dimensional structure of mathematical expressions and visual ambiguity between symbols.

In parallel, the rise of end-to-end Vision-Language Models (VLMs) like GPT-4o (OpenAI, 2024) has introduced a more integrated paradigm. However, studies applying general-purpose VLMs to grade handwritten assignments have consistently highlighted the problem of **error propagation**: inaccurate Optical Character Recognition (OCR) of handwriting leads to faulty input for the reasoning module, resulting in incorrect grades (Kasneci et al., 2023). This suggests that a VLM’s generalist vision encoder may be less robust for the specific domain of mathematical notation than a specialized HMER model, making this a critical area for benchmarking.

2.3 The Paradigm Shift to Diagnostic Assessment

The most advanced research in this domain has shifted from simply assigning a score to performing diagnostic assessment—identifying the specific nature of a student’s error. This requires a model to parse a multi-step solution, compare it to valid reasoning pathways, and classify deviations into a meaningful taxonomy of error types.

This paradigm shift is embodied by the recent development of specialized benchmarks. The **Fermat** benchmark (Yuan et al., 2024) was explicitly designed to evaluate a VLM’s ability to perform error detection, localization, and correction on handwritten solutions containing synthetically generated, human-verified errors. Similarly, the **MathCCS** benchmark (Wu et al., 2024) uses real-world student data to focus on systematic error analysis as a foundation for generating pedagogically useful feedback. The creation of these rigorous benchmarks signifies a maturation of the field, moving the central question from "Can the model get the score right?" to "Can the model identify and explain the error?" Our work aligns with this trajectory by requiring models to assess solutions against a multi-point rubric that implicitly requires error diagnosis.

2.4 Benchmarks for Mathematical Reasoning in NLP

The performance of any NLP system for mathematical tasks is capped by the reasoning power of its underlying models. While text-based benchmarks like *MATH* (Hendrycks et al., 2021) and *GSM8K* (Cobbe et al., 2021) drove initial progress, they suffer from issues like data contamination and an inability to penalize flawed reasoning that leads to a correct answer.

The need to evaluate reasoning in visual contexts has led to more robust multimodal benchmarks. **MathVista** (Lu et al., 2023) and **R2-MultiOmnia** (Ranaldi et al., 2025), for example, provides a comprehensive suite of problems requiring the interpretation of charts, diagrams, and figures. The significant performance gap between state-of-the-art models and humans on such benchmarks demonstrates that visually-grounded mathematical reasoning remains a formidable challenge. This provides essential context for our work, as it establishes a realistic upper bound on the expected performance of VLMs on the even more complex task of grading, which requires not only solving a problem but diagnosing errors in another agent's solution.

3 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 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.

3.1 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. It is important to note that all handwritten solutions and the original problem statements are in **Russian**, reflecting the source of the data. 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.

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

The dataset was sourced from the official EGE expert guide, a publicly available resource for training human graders. All solutions are anonymized and published for educational purposes, thus no ethical clearance was required for their use in this research. The guide does not provide demographic details such as the number of unique students or their grade levels. These are authentic solutions written by students under real high-stakes exam conditions, providing a realistic and challenging distribution of writing styles and error types.

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.

3.2 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.

3.3 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.

4 Experimental Setup

To evaluate the performance of Vision-Language Models on our EGE-Math Solutions Assessment Benchmark, we conducted experiments with seven different state-of-the-art models. The evaluation was structured around three distinct procedures, or "modes", designed to assess the models' capabilities under different levels of contextual information. This required a meticulous data curation process where a specific version of the dataset was prepared for each mode.

4.1 Evaluated Models

We selected a diverse set of VLMs to cover a range of architectures and capabilities:

- **Arcee AI Spotlight:** A model from Arcee AI, accessed via OpenRouter. (Arcee.ai, 2025)
- **Google Gemini 2.0 Flash:** Google's VLM, known for its multimodal capabilities (Team et al., 2023).
- **Google Gemini 2.0 Flash Lite:** A lighter version of Google Gemini 2.0 Flash.
- **Google Gemini 2.5 Flash Preview:** A preview version of Google's next-generation VLM.
- **Google Gemini 2.5 Flash Preview:thinking:** A variant of Google's Gemini 2.5 Flash Preview with enhanced reasoning abilities.
- **OpenAI o4-mini:** A model from OpenAI, a smaller, more efficient version of their flagship models. (OpenAI, 2025)
- **Qwen 2.5 VL 32B:** A large Vision-Language Model from Alibaba Cloud, accessed via OpenRouter (Bai et al., 2025).

We selected a diverse set of VLMs to cover a range of architectures and providers. The inclusion of multiple models from the Google Gemini family allows for a direct comparison of performance trade-offs within a single model lineage, evaluating the impact of model size and specialized tuning (e.g., the 'thinking' variant) on this nuanced task.

Each model was prompted to analyze the handwritten solution image and provide an assessment based on the EGE grading criteria. The output format was standardized to facilitate automated comparison with expert grades.

4.2 Evaluation Procedure and Data Curation

To thoroughly test the model's understanding and reasoning, we designed and prepared data for three evaluation modes. This approach allows for a granular analysis of how additional context influences the models' assessment performance.

- **Mode 1: Without Answer.** In this mode, the model receives only the handwritten solution image and the problem statement. To facilitate this, we prepared a **baseline dataset** where each entry consisted of a pre-processed image and the problem text. The image preprocessing involved standardizing dimensions and resolution to ensure consistent input quality across all experiments. This mode assesses the model's ability to assign a grade based solely on the provided content and its internal understanding of the EGE grading rubric.

- **Mode 2: With Answer.** For this mode, the model receives the handwritten solution, the problem statement, and the correct final numerical answer. To enable this, the baseline dataset was **augmented** by appending the correct final answer for each of the 122 problems, sourced from official EGE materials. This mode assesses whether the model can leverage the correct outcome to better identify errors or confirm the correctness of the student's solution steps.

- **Mode 3: With True Solution.** This is the most informative mode, where the model is given the handwritten solution, the problem statement, and a complete, correct reference solution. The dataset for this mode was **further enriched** with a transcribed, step-by-step "gold standard" solution from the EGE expert guide. This allows us to evaluate the model's ability to compare the student's approach with a known correct method and identify deviations or errors more precisely.

4.3 Prompt Templates and Score Extraction Methodology

For each evaluation, the models were provided with specific prompt templates tailored to the task and evaluation mode. These templates included the problem description, the student's handwritten solution (as an image), and the relevant grading criteria. For the **With Answer** and **With True Solution** modes, the correct answer or reference solution was also incorporated into the prompt. The models were instructed to output their assessment in the structured format, including the analysis of the solution, the final score, and the justification for that score. This structured output facilitated automated extraction of the assigned scores for quantitative analysis. The full prompt templates used for all evaluation modes are available in the project's public repository.

5 Results

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

5.1 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:

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

5.2 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.

To illustrate the models' reasoning process, consider a solution for a parameterised equation (Task 18), which an expert graded as 2 out of 4 points. The student correctly found all potential roots but made a mistake when combining the final intervals, omitting a valid range. OpenAI o4-mini correctly identified this as a computational error deserving 2 points, aligning with the rubric. In contrast, Qwen 2.5 VL 32B failed to spot the missing interval and incorrectly assigned a perfect score of 4, demonstrating a lack of attention to detail. A full analysis of a representative example is provided in Appendix B.

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.

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. (%)	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

5.3 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 information 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.

6 Limitations

Our evaluation provides a unique perspective on VLM capabilities in a real-world assessment scenario. The results highlight a substantial gap between current model performance and human expert-level grading, with the highest accuracy at 56.56%. This indicates ample room for improvement in nuanced mathematical reasoning and the precise application of grading rubrics.

Several factors contribute to these performance limitations and pave the way for future research:

- **Visual Interpretation and Error Propagation:** The diversity in student handwriting, penmanship, and layout poses a significant challenge for accurate visual interpretation. This often leads to *error propagation*, where inaccuracies in the initial visual recognition are passed downstream to the reasoning module, causing incorrect assessments. Future work could explore hybrid approaches, combining general VLM perception with specialized Handwritten Mathematical Expression Recognition (HMER) models to mitigate this issue.

- **Deep Reasoning and Rubric Alignment:** Assessing complex solutions requires deep symbolic and logical reasoning, especially for non-standard solution paths or subtle errors. Models often struggle to translate qualitative grading criteria into quantitative scores, sometimes misinterpreting the severity of an error or failing to identify all relevant mistakes.

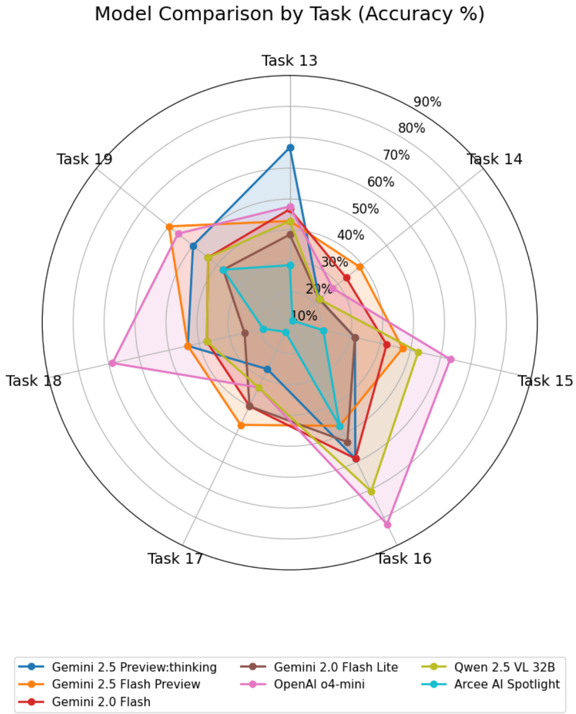


Figure 1: Radar chart showing model Accuracy (%) in the With True Solution mode across all seven task types. The outer edge represents a perfect score. This visualization highlights the models' strengths and weaknesses on different mathematical domains.

- **Dataset and Fine-Tuning:** The current benchmark utilizes 122 solutions. A larger, more diverse dataset would enable more comprehensive evaluation. Furthermore, our study primarily relies on zero-shot prompting; fine-tuning VLMs on this specific assessment task could significantly improve their performance and alignment with the specific curriculum and rubrics.

- **Contextual Reasoning:** While some models effectively use additional context (like a correct answer), others struggle to integrate this information. This highlights a need for more robust mechanisms for conditional reasoning and information fusion in VLMs.

- **Monolingual and Cultural Focus:** The dataset is exclusively in Russian and sourced from a single national curriculum (the Russian EGE). Educational practices and reasoning styles can have cultural specificities. The performance of VLMs may vary on similar benchmarks from different linguistic and educational contexts. Future work could involve extending CHECK-MAT to other languages and curricula.

- **Explainability and Future Directions:** The transparency and interpretability of the models' reasoning processes remain a challenge. Developing more explainable AI is crucial for building trust and utility in educational assessment tools. Future work could also explore interactive assessment scenarios or adapt the benchmark to other global curricula to test for generalization.

6.1 Implications for Mathematical NLP

Our findings have direct implications for the Mathematical NLP community. The CHECK-MAT benchmark provides a challenging new evaluation task for researchers. The systematic failures we observed, particularly in geometric reasoning, highlight a critical area for future work: developing multimodal models that can better understand the interplay between visual diagrams and symbolic reasoning. This aligns with the need for techniques for the joint interpretation of different modalities present in mathematical text. Furthermore, the difficulty models had in applying rubrics points to the need for new neuro-symbolic architectures or fine-tuning strategies to better capture the argumentation relations in the context of mathematical text. This work serves as a call for a deeper focus on these complex, multimodal reasoning challenges.

6.2 Cost and Efficiency

Beyond performance, our study also sheds light on the practical considerations of deploying such models for automated assessment. The significant variation in total cost and average evaluation time across models (e.g., OpenAI o4-mini being considerably more expensive and slower than Google Gemini 2.0 Flash Lite) highlights a trade-off between performance and operational efficiency. For large-scale deployment in educational settings, cost-effectiveness and speed are critical factors that need to be balanced against grading accuracy. Larger frontier models (e.g. OpenAI `o3` or Google `gemini-2.5-pro`) were not included in this benchmark due to computational and budgetary constraints; their evaluation remains future work.

6.3 Future Directions in Knowledge-Intensive Reasoning

Our work highlights several key challenges for the future of knowledge-intensive reasoning. The primary difficulty for current VLMs lies in the robust fusion of perceptual data (handwriting) with a symbolic knowledge base (the grading rubric). The CHECK-MAT benchmark serves as a tool to measure progress in this area. We advocate for future research into hybrid neuro-symbolic architectures and methods that improve the explainability of the model's reasoning process, ensuring that their application of knowledge is both accurate and transparent.

7 Conclusion

In conclusion, this paper introduced CHECK-MAT, a novel benchmark designed to probe the mathematical understanding of Vision-Language Models on the complex, multimodal task of grading handwritten solutions. Our findings demonstrate that while state-of-the-art VLMs can perform this complex task to some degree, they exhibit significant weaknesses in applying the required domain knowledge, particularly for geometric reasoning. This research contributes a valuable diagnostic tool for evaluating models on mathematical discourse and multimodal reasoning. We hope CHECK-MAT will spur the development of the next generation of models that can better handle the joint interpretation of visual, symbolic, and natural language, a key challenge for the field of Mathematical NLP.

8 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

Arcee.ai. 2025. Arcee blog. `https://www.arcee.ai/blog`. Accessed: 2025-07-06.

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, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

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

Yuntian Deng, Anssi Kan, Fei Yin, and Zhaoyang Zhang. 2017. Watch, attend and parse: An end-to-end neural network based approach to handwritten mathematical expression recognition. *Pattern Recognition*, 71:196–206.

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

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, and 1 others. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274.

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*.

OpenAI. 2024. Gpt-4o.

OpenAI. 2025. Introducing o3 and o4-mini. `https://openai.com/index/introducing-o3-and-o4-mini/`. Accessed: 2025-07-06.

Leonardo Ranaldi, Federico Ranaldi, and Giulia Pucci. 2025. R2-MultiOmnia: Leading multilingual multimodal reasoning via self-training. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8220–8234, Vienna, Austria. Association for Computational Linguistics.

Christopher J Sangwin. 2014. Computer-aided assessment of mathematics using stack. *ZDM*, 46(2):307–320.

Gemini Team and 1 others. 2023. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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

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

A Per-Task Performance Data

This appendix provides the detailed per-task scores for a selection of the evaluated models.

Table 3: Per-task scores — `openai_o4-mini`.

Task	Examples	Accuracy	Average Score	Expected Score	Cost
13	21	47.6%	1.48	0.95	$0.4259
14	18	27.8%	1.72	1.28	$0.3465
15	19	**63.2%**	1.68	1.11	$0.3115
16	17	**82.4%**	1.24	1.29	$0.2957
17	15	33.3%	1.20	1.20	$0.2560
18	16	**68.8%**	2.12	2.38	$0.3543
19	16	**56.2%**	1.75	2.06	$0.2879

Table 4: Per-task scores — `qwen-2.5-vl-32b`.

Task	Examples	Accuracy	Average Score	Expected Score	Cost
13	21	42.9%	1.62	0.95	$0.1095
14	18	22.2%	2.17	1.28	$0.0999
15	19	**52.6%**	1.58	1.11	$0.0875
16	17	**70.6%**	1.41	1.29	$0.0783
17	15	33.3%	1.73	1.20	$0.0753
18	16	37.5%	2.75	2.38	$0.0970
19	16	43.8%	2.06	2.06	$0.0868

Table 5: Per-task scores — `arcee-ai_spotlight`.

Task	Examples	Accuracy	Average Score	Expected Score	Cost
13	21	28.6%	0.95	0.95	< $0.01
14	18	11.1%	2.72	1.28	< $0.01
15	19	21.1%	0.74	1.11	< $0.01
16	17	47.1%	1.47	1.29	< $0.01
17	15	13.3%	2.67	1.20	< $0.01
18	16	18.8%	2.12	2.38	< $0.01
19	16	37.5%	2.62	2.06	< $0.01

Table 6: Per-task scores — `gemini-2.5-flash-preview`.

Task	Examples	Accuracy	Average Score	Expected Score	Cost
13	21	42.9%	1.48	0.95	$0.0493
14	18	38.9%	1.00	1.28	$0.0616
15	19	47.4%	1.79	1.11	$0.0401
16	17	47.1%	1.41	1.29	$0.0419
17	15	46.7%	0.73	1.20	$0.0387
18	16	43.8%	1.81	2.38	$0.0713
19	15	**60.0%**	1.40	2.13	$0.0414

Table 7: Per-task scores — `gemini-2.5-flash-preview_thinking`.

Task	Examples	Accuracy	Average Score	Expected Score	Cost
13	21	**66.7%**	1.57	0.95	$0.1096
14	18	22.2%	1.33	1.28	$0.1043
15	19	31.6%	1.21	1.11	$0.0998
16	17	**58.8%**	1.53	1.29	$0.0964
17	15	26.7%	0.73	1.20	$0.0908
18	16	43.8%	2.50	2.38	$0.1037
19	16	50.0%	2.44	2.06	$0.1787

B Representative Example Analysis

This appendix provides a detailed analysis of a representative example from our benchmark to illustrate the evaluation process and the typical performance patterns of the models.

Table 8: Per-task scores — `gemini-2.0-flash-001`.

Task	Examples	Accuracy	Average Score	Expected Score	Cost
13	21	**61.9%**	1.48	0.95	$0.0295
14	18	33.3%	1.61	1.28	$0.0284
15	19	42.1%	1.42	1.11	$0.0276
16	17	**58.8%**	1.47	1.29	$0.0270
17	15	40.0%	0.93	1.20	$0.0254
18	16	37.5%	2.50	2.38	$0.0286
19	16	43.8%	2.31	2.06	$0.0392

Table 9: Per-task scores — `gemini-2.0-flash-lite-001`.

Task	Examples	Accuracy	Average Score	Expected Score	Cost
13	21	**57.1%**	1.38	0.95	$0.0059
14	18	22.2%	1.50	1.28	$0.0056
15	19	31.6%	1.26	1.11	$0.0052
16	17	**52.9%**	1.29	1.29	$0.0051
17	15	40.0%	0.87	1.20	$0.0048
18	16	25.0%	2.62	2.38	$0.0054
19	16	37.5%	2.06	2.06	$0.0049

B.1 Analysis of Solution 18.3.3

B.1.1 Problem Statement

Find all values of parameter a for which the equation has exactly three distinct roots:

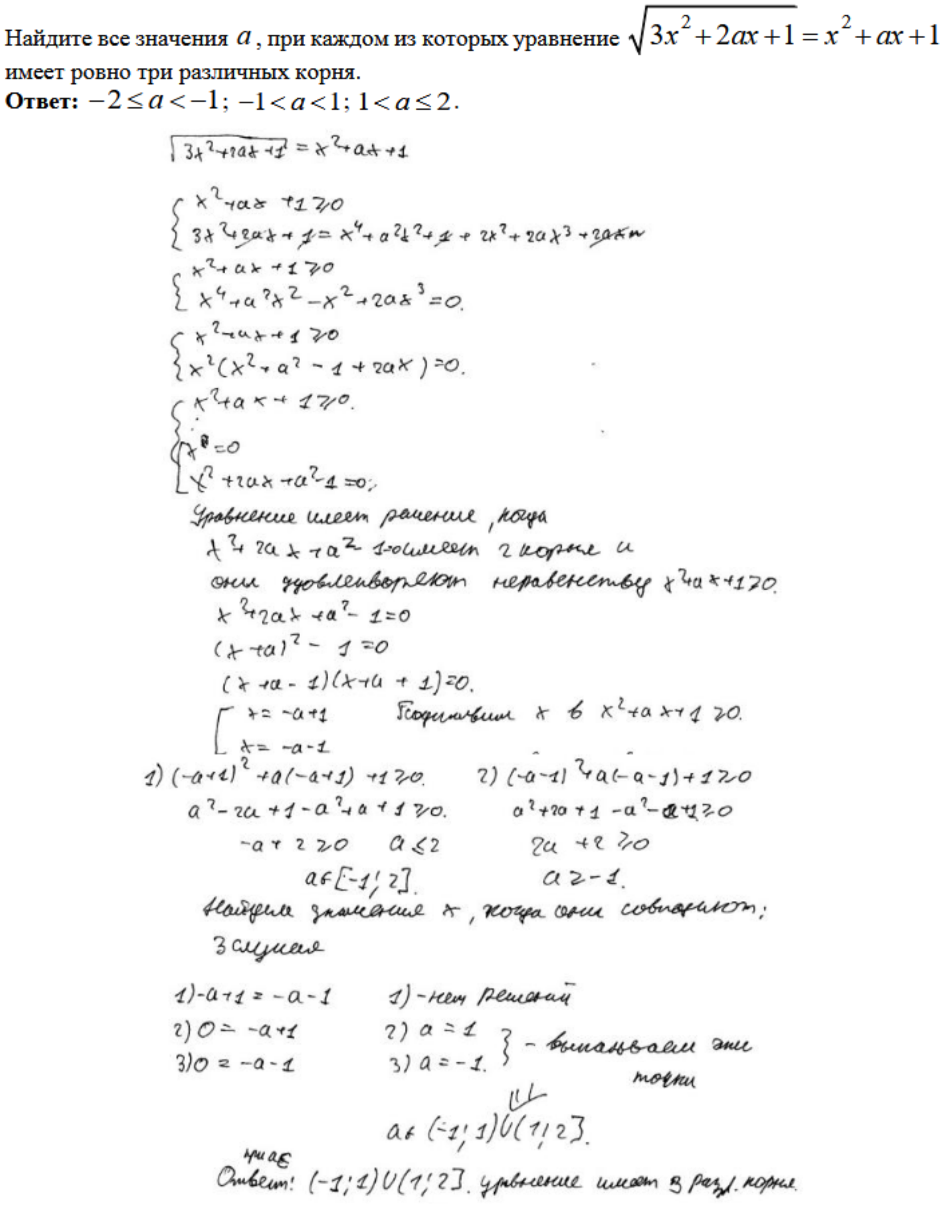
$$\sqrt{3x^2 + 2ax + 1} = x^2 + ax + 1$$

B.1.2 Official Grading Criteria (Task 18)

- **4 points:** A well-reasoned, correct solution is provided.
- **3 points:** A set of parameter values is obtained that differs from the correct set only by the inclusion/exclusion of boundary points.
- **2 points:** A correct interval of parameter values is obtained (possibly with incorrect boundary points), OR an incorrect answer is obtained due to a computational error, but all logical steps are correct.
- **1 point:** The roots of the equation are found, and the problem is correctly reduced to investigating these roots under the given condition(s).
- **0 points:** The solution does not meet any of the criteria above.

B.1.3 Visual Materials

Figure 2 shows the student's handwritten solution, and Figure 3 shows the official correct solution provided in the EGE expert guide.



Найдите все значения a, при каждом из которых уравнение $\sqrt{3x^2+2ax+1} = x^2+ax+1$ имеет ровно три различных корня.

Ответ: $-2 \leq a < -1$; $-1 < a < 1$; $1 < a \leq 2$.

$\sqrt{3x^2+2ax+1} = x^2+ax+1$

$\begin{cases} x^2+ax+1 \geq 0 \\ 3x^2+\cancel{2ax}+\cancel{1} = x^4+a^2x^2+\cancel{1}+2x^2+2ax^3+\cancel{2ax} \end{cases}$

$\begin{cases} x^2+ax+1 \geq 0 \\ x^4+a^2x^2-x^2+2ax^3=0. \end{cases}$

$\begin{cases} x^2+ax+1 \geq 0 \\ x^2(x^2+a^2-1+2ax)=0. \end{cases}$

$\begin{cases} x^2+ax+1 \geq 0. \\ \left[\begin{array}{l} x^2=0 \\ x^2+2ax+a^2-1=0; \end{array}\right. \end{cases}$

Уравнение имеет решение, когда

$x^2+2ax+a^2-1=0$ имеет 2 корня и

они удовлетворяют неравенству $x^2+ax+1 \geq 0$.

$x^2+2ax+a^2-1=0$

$(x+a)^2-1=0$

$(x+a-1)(x+a+1)=0.$

$\left[\begin{array}{l} x=-a+1 \\ x=-a-1 \end{array}\right.$ Подставим x в $x^2+ax+1 \geq 0$.

1) $(-a+1)^2+a(-a+1)+1 \geq 0.$ 2) $(-a-1)^2+a(-a-1)+1 \geq 0$

$a^2-2a+1-a^2+a+1 \geq 0.$ $a^2+2a+1-a^2-a+1 \geq 0$

$-a+2 \geq 0$ $a \leq 2$ $2a+2 \geq 0$

$a \in [-1;2].$ $a \geq -1.$

Найдем значение x, когда они совпадают:

3 случая

1) $-a+1=-a-1$ 1) - нет решений

2) $0=-a+1$ 2) $a=1$

3) $0=-a-1$ 3) $a=-1.$

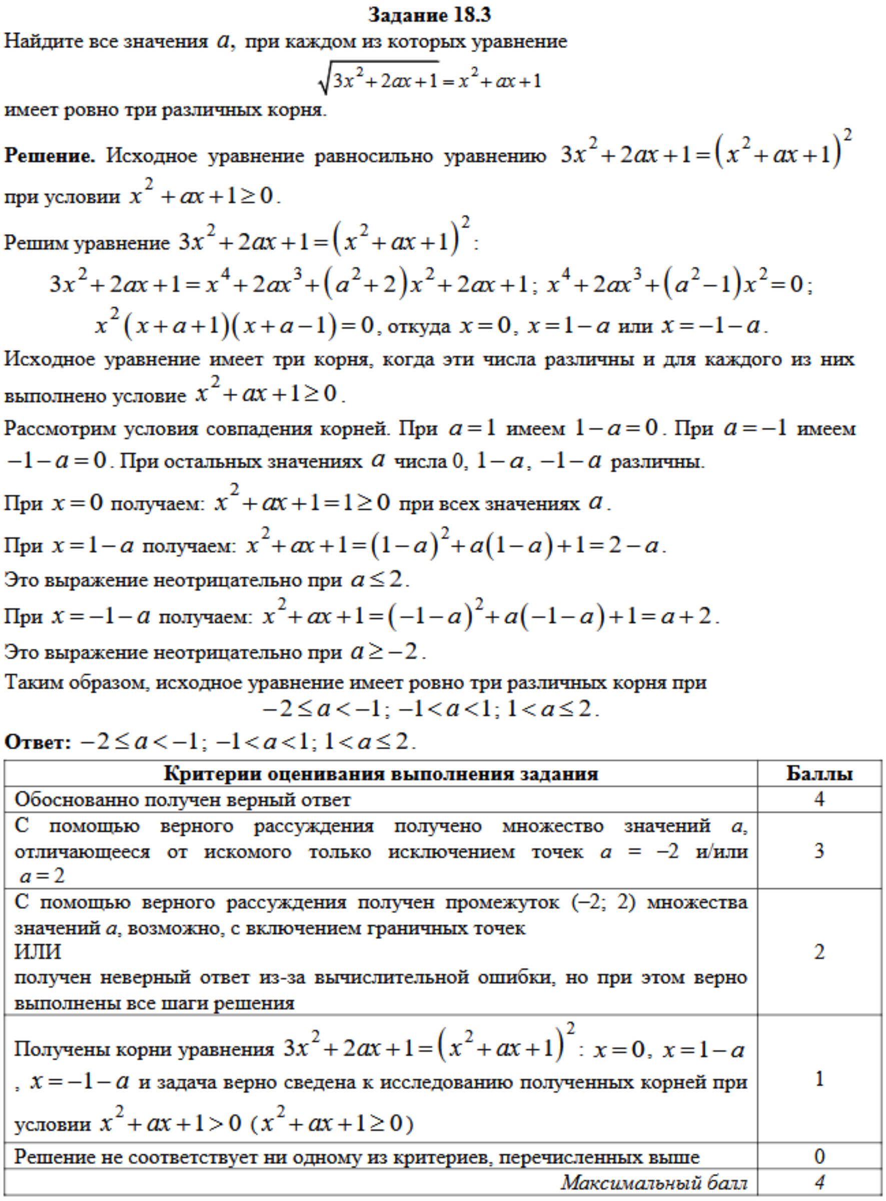
$\left.\begin{array}{l} \\ \end{array}\right\}$ - выкалываем эти точки

$\Downarrow$

$a \in (-1;1) \cup (1;2].$

Ответ: при $a \in$ $(-1;1) \cup (1;2]$. уравнение имеет 3 разл. корня.

Figure 2: Student's handwritten solution for problem 18.3.3. The expert-assigned score is 2.



Задание 18.3

Найдите все значения a, при каждом из которых уравнение

$$\sqrt{3x^2+2ax+1}=x^2+ax+1$$

имеет ровно три различных корня.

Решение. Исходное уравнение равносильно уравнению $3x^2+2ax+1=\left(x^2+ax+1\right)^2$ при условии $x^2+ax+1\ge 0$.

Решим уравнение $3x^2+2ax+1=\left(x^2+ax+1\right)^2$:

$$3x^2+2ax+1=x^4+2ax^3+\left(a^2+2\right)x^2+2ax+1;\ x^4+2ax^3+\left(a^2-1\right)x^2=0;$$

$$x^2\left(x+a+1\right)\left(x+a-1\right)=0,\ \text{откуда}\ x=0,\ x=1-a\ \text{или}\ x=-1-a.$$

Исходное уравнение имеет три корня, когда эти числа различны и для каждого из них выполнено условие $x^2+ax+1\ge 0$.

Рассмотрим условия совпадения корней. При $a=1$ имеем $1-a=0$. При $a=-1$ имеем $-1-a=0$. При остальных значениях a числа 0, $1-a$, $-1-a$ различны.

При $x=0$ получаем: $x^2+ax+1=1\ge 0$ при всех значениях a.

При $x=1-a$ получаем: $x^2+ax+1=\left(1-a\right)^2+a\left(1-a\right)+1=2-a$.

Это выражение неотрицательно при $a\le 2$.

При $x=-1-a$ получаем: $x^2+ax+1=\left(-1-a\right)^2+a\left(-1-a\right)+1=a+2$.

Это выражение неотрицательно при $a\ge -2$.

Таким образом, исходное уравнение имеет ровно три различных корня при

$$-2\le a<-1;\ -1<a<1;\ 1<a\le 2.$$

Ответ: $-2\le a<-1$; $-1<a<1$; $1<a\le 2$.

Критерии оценивания выполнения задания	**Баллы**
Обоснованно получен верный ответ	4
С помощью верного рассуждения получено множество значений a, отличающееся от искомого только исключением точек $a = -2$ и/или $a=2$	3
С помощью верного рассуждения получен промежуток (–2; 2) множества значений a, возможно, с включением граничных точек ИЛИ получен неверный ответ из-за вычислительной ошибки, но при этом верно выполнены все шаги решения	2
Получены корни уравнения $3x^2+2ax+1=\left(x^2+ax+1\right)^2$: $x=0$, $x=1-a$, $x=-1-a$ и задача верно сведена к исследованию полученных корней при условии $x^2+ax+1>0$ ($x^2+ax+1\ge 0$)	1
Решение не соответствует ни одному из критериев, перечисленных выше	0
Максимальный балл	*4*

Figure 3: Official correct solution for problem 18.3.

B.2 Model Assessment Results

The table 10 summarizes the scores assigned by different models in the **With True Solution** mode. The expected score was 2.

Table 10: Model scores for solution 18.3.3.

Model	Assigned Score	Expected	Result
OpenAI o4-mini	2	2	Correct
Qwen 2.5 VL 32B	4	2	Overestimated
Google Gemini 2.0 Flash	2	2	Correct
Google Gemini 2.0 Flash Lite	2	2	Correct
Google Gemini 2.5 Flash Preview	2	2	Correct
Google Gemini 2.5 Flash Preview Thinking	2	2	Correct
Arcee-AI Spotlight	0	2	Underestimated

B.2.1 Key Observations

- **High Accuracy of Most Models:** The majority of the models (5 out of 7) successfully handled the task, assigning the correct score of 2. This group included OpenAI o4-mini and all the tested models from the Google Gemini family.

- **Divergent Errors:** Two models evaluated the solution incorrectly, and their errors were opposites. Qwen 2.5 VL 32B significantly overestimated the score (**4**), while Arcee-AI Spotlight failed to produce a final answer: it became stuck in a loop of writing out equations, which resulted in a score of (**0**).

- **Distinct Failure Modes:** The errors highlight very different failure modes. One model overestimated the score, while the other failed to complete the task entirely. This points to unique flaws in the logic of each model rather than a shared, systematic bias.

B.3 Full Model Responses and Prompt (Translated to English)

B.3.1 Prompt Used for Evaluation (With True Solution)

Analyze the solution of task 18

(an equation, inequality, or system thereof with a parameter) and evaluate it according to the criteria.

Task

{task description}

Assessment criteria for task 18

- **4 points:** A correct answer is obtained with justification.

- **3 points:** A set of parameter values is obtained through correct reasoning, differing from the required set only by the exclusion of boundary points or the inclusion of points not belonging to the answer.

- **2 points:** An interval of the set of parameter values is obtained through correct reasoning, possibly including boundary points, OR an incorrect answer is obtained due to a computational error, but all steps of the solution are correctly performed.

- **1 point:** The roots of the equation are found, and the problem is correctly reduced to the investigation of these roots under the given condition(s).

- **0 points:** The solution does not meet any of the criteria listed above.

IMPORTANT: Assessment Principles

- Evaluate the solution **STRICTLY** according to the criteria.
- Pay attention to mathematical correctness, not the presentation.
- Compare the student's solution with the correct solution provided as a reference.
- Check if the student has correctly performed all key steps of the solution.
- If the student used a different approach, evaluate its correctness and compliance with the criteria.
- Problems with parameters allow for various solution methods: algebraic, geometric, functional.
- Pay **SPECIAL ATTENTION** to the correctness of handling boundary points and the completeness of considering all cases.
- When assessing for 3 points: check that the difference from the correct answer is **ONLY** in the boundary points, not in the main intervals.
- For a geometric approach: check the correctness of the interpretation and the justification of all geometric statements.

IMPORTANT: Instructions for working with the correct solution and the student's solution

You are provided with:

1. The correct solution to the task - use it as a reference for comparison.
2. The student's solution - this is what you must evaluate.

During the analysis:

- Compare each step of the student's solution with the corresponding step of the correct solution.
- Note all deviations and errors.
- Check if the intermediate and final results match.
- Pay special attention to the handling of boundary points and the completeness of considering all cases.

Instructions for checking the solution of a problem with a parameter

1. Check the solution method:
    - Correctness of the chosen approach (algebraic, geometric, functional).
    - Correctness of applying formulas and theorems.
    - Completeness of considering all cases.
2. Check mathematical correctness:
    - Correctness of algebraic transformations.
    - Correctness of working with inequalities.
    - Correctness of finding the domain of permissible values (ODZ).
3. Check the handling of boundary points:
    - Correctness of determining the boundary points.
    - Correctness of including/excluding boundary points in the answer.
4. For a geometric approach, check:
    - Correctness of the geometric interpretation of the conditions.
    - Justification of all geometric statements.
    - Completeness of the analysis of all possible relative positions of geometric objects.

CRITICALLY IMPORTANT: Immediately compare the student's answers with the correct ones!

- **FIRST AND FOREMOST**, check if the student's answer matches the correct answer.
- If the student's answer is **INCORRECT**, this **MUST** be taken into account in the assessment.
- Even if all transformations are performed correctly, but the answer is wrong due to a non-computational error - this must affect the score.
- Do not forget to note all discrepancies between the student's answer and the correct answer.

IMPORTANT: Distinguish between computational and conceptual errors

- **Computational errors:** errors in arithmetic operations, simplifying expressions, calculating values.
- **Conceptual errors:** incorrect application of formulas, wrong solution method, errors in understanding the properties of the parameter.

If a student made only computational errors, but the solution method is correct - this may correspond to the 2-point criterion. If a student made conceptual errors - this usually corresponds to a lower score criterion.

Assessment Examples

Example 1 (score: 4 points) The solution is complete and justified. All values of the parameter for which the system has exactly two solutions are found correctly. All cases are considered, and boundary points are analyzed correctly.

Example 2 (score: 3 points) All stages are present in the solution. Through correct reasoning, a set of parameter values is obtained that differs from the required set only by the exclusion of boundary points or the inclusion of points not belonging to the answer. For example, the answer obtained is $(-2,2)$ instead of $[-2,2]$ or $[-2,-1)\cup(-1,1)\cup(1,2]$ instead of $[-2,-1)\cup(-1,1)\cup[1,2]$.

Example 3 (score: 2 points) The solution is logical, all steps are present, but a computational error was made when solving the inequality, which led to an incorrect answer. Or, through correct reasoning, an interval of the set of parameter values was obtained, possibly including boundary points, but not all possible cases were considered.

Example 4 (score: 1 point) The roots of the equation are found in the solution, and the problem is correctly reduced to the investigation of these roots under the condition $x^2+ax+1\geq 0$, but the further investigation is incorrect or absent.

Example 5 (score: 0 points) The solution contains fundamental errors in understanding the problem or in mathematical reasoning. For example, the roots of the original equation are found incorrectly, or the condition of the problem is misinterpreted.

IMPORTANT: Instructions for working with multiple images

If the solution (correct or student's) is presented in several images:

1. Consider all images as a single solution.
2. Analyze them in sequential order.
3. Take into account information from all images when assessing.

Response Format

Structure your response as follows:

Analysis of the solution [Detailed analysis indicating correct steps and errors]

Comparison with the correct solution [Comparison of the student's approach with the reference solution, indicating key differences]

Final score [Score: X points]

IMPORTANT: The "Final score" section **MUST** be formatted **EXACTLY** in this format: "Final score" - heading. "[Score: X points]" - where X is a number 0, 1, 2, 3, or 4. **DO NOT CHANGE** this format and do not add any other text to it!

Justification of the score [Detailed justification in accordance with the criteria]

Recommendations (if applicable) [Recommendations for correcting errors]

B.3.2 Response from OpenAI o4-mini (Score: 2 - Correct)

Analysis of the solution

1. The student wrote down an equivalent system:
    - a) the equation after squaring;
    - b) the condition that the expression under the root is non-negative.
2. Found the roots of the original equation: $x = 0$, $x = 1\text{-}a$, $x = \text{-}1\text{-}a$.
3. Analyzed the condition $x^2 + ax + 1 \geq 0$ for these values of x:
    - for $x = 0$, the condition is always met;
    - for $x = 1\text{-}a$, from the inequality, they obtained $a \leq 2$;
    - for $x = \text{-}1\text{-}a$, they obtained $a \geq \text{-}2$.
4. Concluded that for three distinct roots, it is necessary that $\text{-}2 \leq a \leq 2$ and $a \neq 1$, $a \neq \text{-}1$.
5. However, in the final answer, the student wrote $a \in (\text{-}1; 1) \cup (1; 2]$, thus completely omitting the interval $[\text{-}2; \text{-}1)$.

Error: when combining the resulting intervals and excluding the boundary points, the student discarded the left part $[\text{-}2; \text{-}1)$ for a purely computational/descriptive reason. All key steps (finding roots, conditions on a, accounting for exclusions) were performed correctly, but the answer contains an inaccuracy.

Comparison with the correct solution

Reference answer: $a \in [\text{-}2; \text{-}1) \cup (\text{-}1; 1) \cup (1; 2]$.

The student obtained the same intermediate boundaries -2 and 2, and correctly excluded $a = \pm 1$, but missed the part $[\text{-}2; \text{-}1)$ in the final answer. Therefore, the method is correct, but a computational (descriptive) slip in combining the results led to an incomplete answer.

Final score

[Score: 2 points]

Justification of the score

According to the criteria:

- All key steps of the solution (finding roots, domain of permissible values, conditions on the parameter, exclusion of boundary values) were performed correctly.
- The final answer is incorrect due to a computational/descriptive error when combining the intervals.
- The approach is mathematically correct, but the result is inaccurate.

This corresponds to the 2-point criterion: "An interval of the set of parameter values is obtained through correct reasoning, possibly including/excluding boundary points, but the overall answer is incorrect due to a computational error."

Recommendations

- Be more careful when combining intervals and check that no sections have been dropped.
- After obtaining the conditions on the parameter, explicitly write down all three 'pieces' and only then combine them, excluding the single points.

B.3.3 Response from Qwen 2.5 VL 32B (Score: 4 - Overestimated)

Analysis of the solution

Step 1: Forming the system of equations The student correctly understands that for the initial equation $\sqrt{3x^2 + 2ax + 1} = x^2 + ax + 1$, two conditions must be met:

1. $x^2 + ax + 1 \geq 0$ (since the square root is defined only for non-negative values).
2. $3x^2 + 2ax + 1 = (x^2 + ax + 1)^2$ (the main equation).

This approach is correct, and the student correctly transforms the equation:

$$3x^2 + 2ax + 1 = x^4 + 2ax^3 + (a^2 + 2)x^2 + 2ax + 1.$$

Then, they combine like terms and obtain:

$$x^4 + 2ax^3 + (a^2 - 1)x^2 = 0.$$

Next, the student correctly factorizes this equation:

$$x^2(x^2 + 2ax + (a^2 - 1)) = 0.$$

From this, they find the roots:

$$x = 0 \quad \text{and} \quad x^2 + 2ax + (a^2 - 1) = 0.$$

For the quadratic equation $x^2 + 2ax + (a^2 - 1) = 0$, the student finds the roots:

$$x = -a + 1 \quad \text{and} \quad x = -a - 1.$$

Thus, the student correctly finds three roots: $x = 0$, $x = -a + 1$, and $x = -a - 1$.

Step 2: Analysis of the conditions for distinct roots The student recognizes that for the equation to have three distinct roots, these roots must be different. They check the cases where the roots coincide:

1. $x = 0$ and $x = -a + 1$: This is possible if $-a + 1 = 0 \Rightarrow a = 1$.
2. $x = 0$ and $x = -a - 1$: This is possible if $-a - 1 = 0 \Rightarrow a = -1$.
3. $x = -a + 1$ and $x = -a - 1$: This is possible if $-a + 1 = -a - 1$, but this is impossible as it leads to a contradiction.

Step 3: Verifying the condition $x^2 + ax + 1 \geq 0$ The student checks the condition $x^2 + ax + 1 \geq 0$ for each root:

1. For $x = 0$: $x^2 + ax + 1 = 1 \geq 0$ is always met.
2. For $x = -a + 1$: $x^2 + ax + 1 = (-a + 1)^2 + a(-a + 1) + 1 = a^2 - 2a + 1 - a^2 + a + 1 = -a + 2$. This is non-negative if $-a + 2 \geq 0 \Rightarrow a \leq 2$.
3. For $x = -a - 1$: $x^2 + ax + 1 = (-a - 1)^2 + a(-a - 1) + 1 = a^2 + 2a + 1 - a^2 - a + 1 = a + 2$. This is non-negative if $a + 2 \geq 0 \Rightarrow a \geq -2$.

Step 4: Combining the conditions The student correctly combines all the obtained conditions:

- $a \neq 1$ and $a \neq -1$ (for distinct roots).
- $a \leq 2$ and $a \geq -2$ (to satisfy the condition $x^2 + ax + 1 \geq 0$).

Thus, the student obtains:

$$a \in [-2, -1) \cup (-1, 1) \cup (1, 2].$$

Comparison with the correct solution

The correct solution and the student's solution completely coincide in the answer:

$$a \in [-2, -1) \cup (-1, 1) \cup (1, 2].$$

The student correctly performed all the key steps:

1. Transforming the original equation.
2. Finding the roots.
3. Analyzing the conditions for distinct roots.
4. Verifying the condition $x^2 + ax + 1 \geq 0$.
5. Combining all the obtained conditions.

Final score

[Score: 4 points]

Justification of the score

The student justifiably obtained the correct answer by correctly performing all the key steps of the solution. All intermediate results are correct, the boundary points are handled correctly, and all possible cases are considered. The solution method fully complies with the mathematical requirements of the problem, and the answer completely matches the correct solution.

Recommendations

The student’s solution contains no errors and requires no corrections. However, the student could improve the presentation by making it more detailed and structured, so that readers can better follow their reasoning. For example, they could explain in more detail why the roots coincide when $a = 1$ or $a = -1$, and why these values are excluded from the answer.