Methodological framework for automated solving of mathematics competition problems in Serbia using LLMs

Marina Svičević[1], Aleksandar Milutinović[1], Nemanja Vučićević[1], Aleksandar Milenković[1], Miloš Pavković[2]

[1]Faculty of Science, University of Kragujevac, Kragujevac, Serbia, marina.svicevic@pmf.kg.ac.rs, 0000-0003-2791-3849, milutinovic.aleksandar9@gmail.com, 0009-0004-1300-3232, nemanja.vucicevic@pmf.kg.ac.rs, 0000-0002-4903-7280, aleksandar.milenkovic@pmf.kg.ac.rs, 0000-0001-6699-8772

[2]Singidunum University, Belgrade, Serbia, mpavkovic@singidunum.ac.rs, 0000-0001-7776-6045

Corresponding author: marina.svicevic@pmf.kg.ac.rs

Abstract

This paper presents a methodological framework for the automated solving of mathematics competition problems for high school students in Serbia using large language models (LLMs). The corpus includes 348 text-based problems from the municipal, regional, and national competition levels, distributed across grades and Categories A and B. The problems and their corresponding official solutions were organized in a structured format with LaTeX notation, defining prompt templates for inference and reference solutions for automatic evaluation. The study compares several models, including both general purpose and mathematically specialized LLMs. A particular contribution of the paper is a multi-aspect automatic evaluation framework based on final answer accuracy, logical flow of reasoning, and explanation quality. The results reveal a clear gap between commercial and locally executable models, as well as the importance of domain specialization for more successful solving of competition level mathematical problems.

Keywords: large language models; mathematics competitions; automated problem solving; automatic evaluation

1. Introduction

Mathematics problems that require multi step reasoning, precise interpretation of conditions, and well-argued conclusions represent a serious challenge for LLMs. As noted in [1], research on LLM-based mathematical reasoning covers diverse tasks, datasets, approaches, and evaluation settings, which makes it difficult to clearly assess the actual capabilities and limitations of these models.

Previous studies have shown that the performance of AI tools and LLMs in solving mathematical problems depends on task format, domain, and language. In [2], dedicated to problems from the Mathematical Kangaroo competition, models achieved better results on text based problems than on image based ones, and performed better in algebra and number theory than in geometry and logic. These findings also highlighted the importance of linguistic context, that is, the need to examine model performance on problems formulated in Serbian Language.

Following research focused on more complex problems from the national mathematics competition in Serbia. In [3, 4], DeepSeek-R1, o1, and o3-mini were evaluated on LaTeX-formatted problems in Serbian, with the generated solutions assessed by a member of the official competition committee using the same criteria applied to students' work. The results showed that, although the models could in some categories achieve rankings comparable to those of successful contestants, they still struggled with less standard formulations, logic problems, and tasks requiring the consistent execution of all solution steps. A similar conclusion is reported in [5], which emphasizes that competition mathematics problems remain a demanding test for LLMs and that final answer accuracy alone is insufficient for reliably assessing reasoning quality.

Based on this, the aim of this paper is to present a methodological framework for the automated solving of mathematics competition problems in Serbia. The main contribution of the paper lies in integrating data preparation, model prompting, and multi aspect automatic evaluation into a unified experimental workflow.

2. Methodology

Figure 1 presents the overall workflow used in this study, from the preparation of competition problems in LaTeX format to the automatic evaluation of generated solutions.

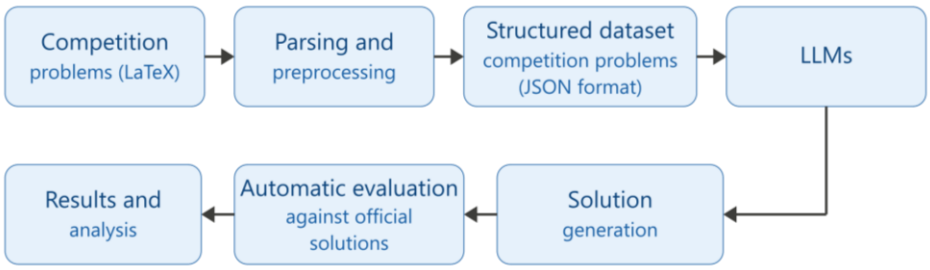


Figure 1. Workflow of the proposed methodology for automated solving of mathematics competition problems in Serbia using LLMs.

The dataset used in this study consists of mathematics competition problems for high school students in Serbia, organized by the Mathematical Society of Serbia. The corpus covers three competition levels: municipal, regional, and national. It also includes all four high school grades in Categories A and B. In the official competition structure, five problems are assigned in all grades and categories, except for Category A at the national level, where four problems are assigned. Based on the collected material, the dataset used in this paper was constructed, and its distribution by competition level and grade is shown in Table 1. For the purposes of this study, 348 problems suitable for text-based processing were selected, forming a corpus for automated solving and evaluation.

Table 1. Distribution of problems by competition level and grade.

Level	Grade 1	Grade 2	Grade 3	Grade 4
Municipal	30	30	30	30
Regional	30	30	30	30
National	27	27	27	27

Each problem was structured as a prompt reference pair: the prompt contains the task metadata (domain, grade, and competition level) together with the problem statement, while the reference represents the full official solution. Both the problem statement and the reference solution preserve LaTeX notation to maintain the precision of mathematical expressions.

For the experiments, models of different sizes and purposes were selected, including general purpose LLMs and models specialized in mathematical reasoning, enabling a comparison of their initial capabilities in solving competition level problems.

Automatic evaluation was based on three independent criteria: Answer Accuracy (A), Logical Coherence (B), and Explanation Quality (C) and was performed by an LLM judge (Claude 3.7 Sonnet). For each of the 348 problems, the generated solution, together with the problem statement and the corresponding reference solution, was submitted to LLM, with a separate system prompt defined for each criterion. Answer Accuracy (A) evaluates the correctness of the final answer and returns a binary score of 0 or 1, based entirely on a comparison with the reference solution. Logical Coherence (B) evaluates the soundness of the reasoning process on a scale from 0 to 1, independently of whether the final answer is correct. Explanation Quality (C) assesses clarity, completeness, and exactness of the presented solution, also on a scale from 0 to 1. The final score is computed as a weighted average of these three criteria as $0.30 \times A + 0.40 \times B + 0.30 \times C$.

3. Results

General results of the experiments are presented in Table 2. The table includes results for three evaluation aspects; final answer accuracy, logical coherence of reasoning, and explanation quality, as well as the weighted average, which represents an aggregated measure of overall model performance.

Table 2. General results of the experiments.

Model	A	B	C	Weighted average
Gemini 3 Flash	54.5%	74.1%	69.8%	66.9%
Claude 3.7 Sonnet	37.6%	39.6%	69.8%	48.1%
Mathstral 7B	15.6%	20.1%	19.9%	18.7%
LLaMA 3 8B	16.8%	14.3%	26.2%	18.6%
Mistral 7B Instruct	12.6%	14.2%	27.9%	17.8%
DeepSeek Math 7B	12.9%	15.1%	23.0%	16.8%
Qwen2.5 Math 7B	13.2%	4.7%	11.6%	9.3%
MAmmoTH 34B	2.6%	6.9%	11.5%	7.0%
Llemma 7B	2.4%	6.1%	10.3%	6.2%

The results reveal a clear gap between commercial models and locally executable models. Gemini 3 Flash achieved the best overall result, with a weighted average of 66.9%, while Claude 3.7 Sonnet reached 48.1%. All remaining models stayed below 20%, with Mathstral 7B achieving the best result among locally executable models, at 18.7%. The results also indicate that model size alone is not a decisive factor of success. In some cases, models specialized in mathematical reasoning outperform larger general-purpose models, highlighting the importance of domain specialization. Among local models, a particularly visible gap appears between the

logical coherence of reasoning and final answer accuracy, suggesting that models often begin solving in the right direction but fail to produce a fully correct final solution.

4. Discussion and Conclusions

This paper presented a methodological framework for the automated solving of mathematics competition problems for high school students in Serbia using LLMs. A particular contribution of the paper is the establishment of a multi aspect automatic evaluation framework, which includes final answer accuracy, logical coherence of reasoning, and explanation quality, thus enabling a more comprehensive assessment of model performance than one based solely on the final result. The obtained results indicate a clear gap between commercial and locally executable models, as well as the importance of domain specialization for better performance on competition level problems. At the same time, the differences between the values obtained in criteria A, B, and C show that model performance cannot be reliably assessed solely based on final answer accuracy, but that the quality of the reasoning process and how the solution is explained must also be considered. Future work will focus on fine-tuning selected models to improve their performance in solving problems in this domain.

Acknowledgements

The authors are supported by the Ministry of Science, Technological Development and Innovation, Republic of Serbia, Contract No. 451-03-34/2026-03/200122.

References

1. Ahn J, Verma R, Lou R, Liu D, Zhang R, Yin W. Large Language Models for Mathematical Reasoning: Progresses and Challenges. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop. 2024:225–237.

2. Svičević M, Milenković A, Vučićević N, Stanković M. Evaluating the Success of AI Tools in Supporting Student Performance in Mathematical Kangaroo Competition. Computer Applications in Engineering Education. 2025;33:e70063.

3. Vučićević N, Svičević M, Milenković A. Challenging DeepSeek-R1 with Serbian High School Math Competition Problems. In: Sinteza 2025 – International Scientific Conference on Information Technology, Computer Science, and Data Science. 2025:274–280.

4. Milenković A, Vučićević N, Svičević M. Evaluating Open AI Tools o1 and o3-mini in Solving High School Problems from Serbian National Mathematics Competition. The Teaching of Mathematics. 2025;28(1):14–29.

5. Mao Y, Kim Y, Zhou Y. CHAMP: A Competition-level Dataset for Fine-Grained Analyses of LLMs' Mathematical Reasoning Capabilities. In: Findings of the Association for Computational Linguistics: ACL 2024. 2024:13256–13274.