

SUCCESS OF AI MATH SOLVER TOOL IN SOLVING NON-STANDARD MATHEMATICS COMPETITION PROBLEMS*

UDC 51::37.016::004.8
371.314.6:51(079.1)
51-057.875
004.8:51-057.875:373.3+373.5

**Marko Stanković¹, Aleksandar Milenković²,
Marina Svičević², Nemanja Vučičević²**

¹University of Niš, Pedagogical Faculty in Vranje, Serbia

²University of Kragujevac, Faculty of Science, Serbia

ORCID iDs: MarkoStanković
Aleksandar Milenković
Marina Svičević
Nemanja Vučičević

 <https://orcid.org/0000-0003-0563-4267>
 <https://orcid.org/0000-0001-6699-8772>
 <https://orcid.org/0000-0003-2791-3849>
 <https://orcid.org/0000-0002-4903-7280>

Abstract. Artificial intelligence is increasingly transforming how students learn, including their approach to mathematics and problem-solving, by offering additional support and assistance—a trend that continues to attract research interest. One line of research focuses on helping students prepare for math competitions by solving more complex mathematical problems. In addition to regular national math competitions, which allow students to progress to international mathematical Olympiads, there are also competitions aimed at popularizing mathematics and developing logical thinking in students. One such competition is the international Kangaroo competition. In this paper, we examine the performance of the AI Math Solver, available on the Interactive Mathematics platform, in solving tasks from the 2024 Kangaroo competition. The selected tasks targeted three student categories: 3rd and 4th grade elementary, 7th and 8th grade elementary, and 3rd and 4th grade high school students. The problems were uploaded as images (screenshots) in both Serbian and English, since visual elements frequently appear in the problem formulations and answer choices in the Kangaroo competition. The results are presented in two sections: a qualitative analysis of selected problems that illustrate common patterns and errors, and a quantitative analysis that

Received April 29, 2025 / Accepted June 30, 2025

Corresponding author: Marko Stanković

Pedagogical Faculty in Vranje, University of Niš, Partizanska 14, 17500 Vranje, Serbia

E-mail: markos@pfvr.ni.ac.rs

* *Note.* This article is an extended and substantially revised version of the paper entitled "Performance of an AI Tool in Solving Non-Standard Mathematics Competition Problems" presented at the 1st International Scientific Conference *Education and Artificial Intelligence (EDAI 2024)*, held at the Pedagogical Faculty in Vranje, University of Niš. The original version of the paper appeared in the conference proceedings (Stanković et al., 2025).

summarizes the tool's overall performance. Out of a total of 84 tasks, in both Serbian and English, the solver correctly answered 24, corresponding to a success rate of just under 30% in both languages. Furthermore, some tasks solved in Serbian were not solved in English, and vice versa. Additionally, differences were observed in the distribution of correct answers across tasks of varying difficulty levels.

Key words: *AI tools, Kangaroo competition, math education, non-standard tasks*

INTRODUCTION

In recent years, the integration of artificial intelligence (AI) into educational environments has attracted significant attention. Various forms of Generative Artificial Intelligence (GenAI) have demonstrated both potential and challenges in educational contexts. Many GenAI tools are either free or affordable and easy to use, making them attractive options for a wide range of educational purposes. Among these tools, ChatGPT (Chat Generative Pre-trained Transformer), a publicly accessible chatbot, stands out in terms of popularity. Numerous studies have examined the use of ChatGPT in educational settings (cf. Lo, 2023). For example, it can be effectively integrated into education to automate routine tasks and enhance students' learning experience (Elbanna and Armstrong, 2023), as well as to assist teachers with lesson preparation (Spasić and Janković, 2023). However, the use of ChatGPT in education also raises ethical concerns, including plagiarism, reduced learning engagement, and user privacy (Memarian and Doleck, 2023).

Although many studies have highlighted ChatGPT's potential across various school subjects, its application in mathematics remains limited. A thorough evaluation of its capabilities reveals serious shortcomings when it comes to solving complex mathematical problems, particularly at the postgraduate level (Frieder et al., 2023). Nevertheless, ChatGPT can serve as a helpful assistant for checking mathematical facts and retrieving information. Recent research by Wei (2024) demonstrates that advanced AI models such as ChatGPT-4 and ChatGPT-4o generally outperform U.S. students across all grade levels, content domains, item types, and difficulty ranges. Yet, they continue to struggle with geometry and higher-level problems, prompting the development of specialized AI tools tailored for mathematical reasoning. Particularly notable are tools capable of solving problems from the International Mathematical Olympiad (IMO) at a human level (DeepMind, 2024), such as AlphaGeometry (Trinh et al., 2024), later improved into AlphaGeometry 2 and AlphaProof (Castelvecchi, 2024).

To engage a wider range of students, researchers have emphasized the importance of using interesting and contextually relevant problems in teaching mathematics (e.g., Simjanović et al., 2021). Therefore, in addition to the IMO, there are competitions designed to popularize mathematics and enhance students' logical thinking skills. A prominent example of such a competition is the Mathematical Kangaroo. Solving problems in this competition requires creativity, imagination, logical reasoning, and the application of diverse problem-solving strategies. This raises the question of how AI tools will perform on such tasks—a question this study seeks to investigate.

The limitations of AI systems in solving tasks that require broad reasoning skills have been previously explored using datasets such as SMART-101 and SMART-840 (Cherian et al., 2023, 2025). SMART-101 evaluates visuo-linguistic puzzles designed for children

aged 6–8, focusing on skills like arithmetic, algebra, and spatial reasoning, and is based on nearly 10 years of Math Kangaroo USA competitions. SMART-840 extends this evaluation across all age groups (grades 1–12) using 840 puzzles from recent Mathematical Kangaroo competitions. Although large models demonstrate promising reasoning abilities, their solutions often lack accuracy, especially on tasks intended for younger children, revealing persistent limitations in generalization and abstraction, particularly when multimodal reasoning and the integration of diverse skills are required. Building on these insights, this study examines Mathematical Kangaroo tasks by selecting non-consecutive age groups to better capture performance variations. We also test a specialized AI math tool in both English and Serbian (Cyrillic script), offering a broader perspective on language and script-related challenges.

In the following sections, we present an overview of the Interactive Mathematics platform and the AI tool employed in this study—AI Math Solver. We also provide a brief description of the Mathematical Kangaroo competition, highlighting its unique features. To assess the performance of the AI tool, we applied it to a selection of competition tasks in both Serbian and English. Given that many large language models (e.g., LLaMA) are predominantly trained on English-language data, their performance often degrades when applied to tasks in other languages (Zhao et al., 2024). Therefore, a bilingual testing approach was adopted. The results are organized into two subsections: a *qualitative analysis of selected problems* that illustrates the tool’s reasoning patterns and typical errors, and a *quantitative analysis of results* that summarizes its overall performance across the dataset. We conclude with a discussion of key findings and their implications for the use of AI in mathematics education.

2. INTERACTIVE MATHEMATICS AND AI MATH SOLVER

The research was conducted using the Interactive Mathematics platform¹, a math education platform that combines human educators and AI computing. The platform offers features such as an AI Math Solver², live tutoring, a quiz generator, and flashcards—tools designed to enhance students’ problem-solving skills and overall math performance. Despite its range of functionalities, the platform remains intuitive and user-friendly. Some tools are freely available, while others require a modest fee, making the platform accessible to a broad audience. It is also worth noting that Interactive Mathematics is trusted by leading institutions such as MIT, Yale, and Harvard.

AI Math Solver is a tool on the Interactive Mathematics platform that uses a mathematical computation engine and leverages the power of large language models to generate natural language answers. The AI Math Solver addresses a wide range of scientific fields, including Basic Math, Math Word Problems, Pre-Algebra, Algebra, Geometry, Graphing, Trigonometry, Precalculus, Calculus, Statistics, Finite Math, Linear Algebra, Chemistry, and Physics, allowing users to tackle diverse mathematical and scientific challenges. The user interface is very intuitive, with a clean, minimalistic design that allows students to focus on problem-solving (see Figure 1). Users can either manually input problems or upload documents for automatic analysis, which is particularly useful for non-standard problems that include images or diagrams. The tool provides real-time solutions,

¹<https://www.intmath.com/>

²<https://app.intmath.com/>

showing both step-by-step explanations and final results. After solving a problem, an "Ask follow-up question" button allows users to easily seek clarification or further assistance.

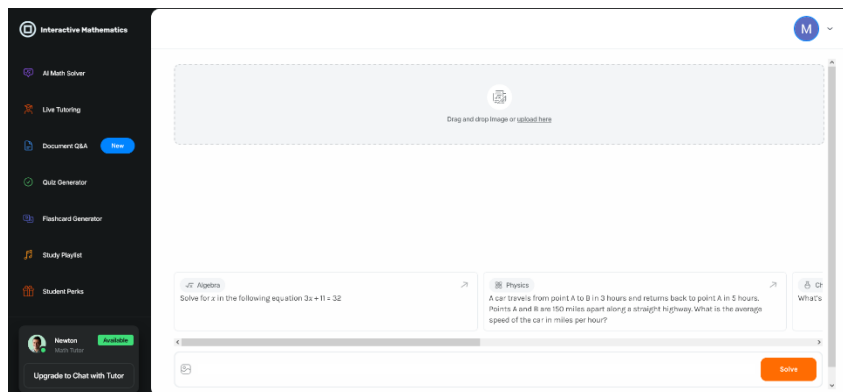


Fig. 1 Interface of the AI Math Solver after login.

3. MATHEMATICAL KANGAROO

The Mathematical Kangaroo is an international mathematics competition held in over 100 countries since last year, organized globally by the Association Kangourou Sans Frontières (AKSF). The primary mission of the Mathematical Kangaroo competition is to popularize mathematics, with the goal of increasing interest in mathematics and the natural sciences, as well as fostering combinatorial and logical thinking, text comprehension, and the application of acquired mathematical knowledge.

Each year, the competition tasks are carefully selected at the annual AKSF meeting, where university professors, mathematics teachers, and educators from around the world choose problems from a database of tasks proposed by representatives of different countries. Special attention is given to ensuring that the selected problems are not standard textbook exercises. In addition to requiring inspiring ideas, perseverance, and creativity, solving these tasks also demands imagination, basic computational skills, logical reasoning, and a variety of problem-solving strategies (Akveld et al., 2020).

Each problem in the competition has five answer choices, with only one correct answer. For students in 5th grade and above, the test consists of 30 problems categorized into three difficulty levels (10 problems per level), to be solved within 90 minutes. Students in grades 3 and 4 must answer 24 questions within 75 minutes, with problems divided into three groups of eight based on difficulty level. In Serbia, students in grades 1 and 2 solve 18 problems, divided into three groups of six, within 60 minutes.

4. METHODOLOGY

The aim of this study is to evaluate the performance of the AI tool AI Math Solver in solving problems from various categories of non-standard competition tasks, specifically those from the Mathematical Kangaroo competition. To better assess the impact of task complexity on the success rate of AI Math Solver's, we chose to analyze its performance

on solving problems from non-consecutive student categories. Therefore, we selected the test sets used in the Kangaroo competition by students in grades 3 and 4 of elementary school, grades 7 and 8 of elementary school, and grades 3 and 4 of high school.

Additionally, this study aims to compare the success of AI Math Solver with that of students from Serbia who participated in the competition in the aforementioned categories.

Given that more than half of the tasks (45 out of 84) include images either in the problem statements and/or within the answer choices, the tasks were presented to the AI tool in the form of screenshots. Furthermore, considering previous research on the impact of language on AI performance, we provided the problems in both Serbian and English in order to determine whether significant differences exist in AI Math Solver's performance depending on the language.

The research was conducted during September and October 2024.

5. RESULTS

5.1 Qualitative Analysis of Selected Problems

In this subsection, we present three problems in English to illustrate the performance of the AI Math Solver.

1. (3-4 grade problem, 3 points) Lizzy pays 7 dollars for 3 items. The cost of each item is different and is a whole number. How much is the most expensive item?

(A) 2 dollars (B) 3 dollars (C) 4 dollars (D) 5 dollars (E) 6 dollars

Solution provided by AI Math Solver:

To find the cost of the most expensive item, we need to consider the total cost and the number of items. Let's break down the problem step by step.

1. Total cost: Lizzy pays 7 dollars for 3 items.
2. The cost of each item is different and is a whole number.

Since the total cost is 7 dollars and there are 3 items, we can start by finding the maximum possible cost of the most expensive item.

Let's assume the cost of the most expensive item is x dollars. Then, the cost of the remaining two items must be $(7-x)$ dollars in total.

Since the cost of each item is a whole number, the cost of the remaining two items must be a sum of two whole numbers that add up to $(7-x)$ dollars.

We can try different values of x to see which one satisfies the condition.

If $x=6$, then $(7-x)=1$, which means the remaining two items must cost 1 dollar in total. However, this is not possible since the cost of each item is a whole number and must be at least 1 dollar.

If $x=5$, then $(7-x)=2$, which means the remaining two items must cost 2 dollars in total. This is possible if one item costs 1 dollar and the other item costs 1 dollar.

If $x=4$, then $(7-x)=3$, which means the remaining two items must cost 3 dollars in total. This is possible if one item costs 1 dollar and the other item costs 2 dollars.

If $x=3$, then $(7-x)=4$, which means the remaining two items must cost 4 dollars in total. This is possible if one item costs 1 dollar and the other item costs 3 dollars.

However, if $x=3$, then the cost of the most expensive item is not unique, since the other two items can also have a cost of 3 dollars.

Therefore, the maximum possible cost of the most expensive item is $x=5$ dollars, which satisfies the condition that the cost of each item is different and is a whole number.

5 dollars

AI Math Solver produced an incorrect solution in this case. The correct answer is (C) 4 dollars. Although the solver employed a case analysis method, it failed to account for the requirement that all three item prices must be distinct whole numbers summing to 7. Interestingly, the solver applied a different approach when solving the same problem in Serbian, yet arrived at the same incorrect conclusion. Furthermore, in both versions, the variables were incorrectly formatted as plain text rather than being displayed in proper mathematical notation (e.g., x instead of x).

2. (7-8 grade problem, 5 points) The digits 0-9 can be drawn with horizontal and vertical segments, as shown.



Greg chooses three different digits. In total, his digits have 5 horizontal segments and 10 vertical segments. What is the sum of his three digits?

(A) 9

(B) 10

(C) 14

(D) 18

(E) 19

Solution provided by AI Math Solver:

To solve this problem, we need to analyze the given information and devise a plan. Here's the step-by-step approach:

****Step1: Understand the problem and the given information****

We are given that Greg chooses three different digits, and in total, his digits have 5 horizontal segments and 10 vertical segments.

****Step 2: Analyze the number of segments in each digit****

Let's analyze the number of horizontal and vertical segments in each digit from 0 to 9:

-0: 1 horizontal, 2 vertical
 -1: 0 horizontal, 2 vertical
 -2: 1 horizontal, 3 vertical
 -3: 1 horizontal, 3 vertical
 -4: 2 horizontal, 2 vertical
 -5: 1 horizontal, 3 vertical
 -6: 1 horizontal, 4 vertical
 -7: 1 horizontal, 2 vertical
 -8: 2 horizontal, 4 vertical
 -9: 1 horizontal, 4 vertical

****Step3: Determine the possible combinations of digits****

We need to find combinations of three digits that add up to 5 horizontal segments and 10 vertical segments.

****Step 4: Look for combinations that satisfy the given conditions****

After analyzing the possible combinations, we can see that one possible combination is 1, 4, and 9 (or any permutation of these digits). This combination adds up to 5 horizontal segments ($0+2+3$) and 10 vertical segments ($2+2+6$). However, we need to verify if this combination is unique.

****Step 5: Verify the uniqueness of the combination****

After re-examining the possible combinations, we can see that the combination 1, 4, and 9 is indeed unique.

****Step 6: Calculate the sum of the digits****

Now, we can calculate the sum of the digits:

$$1+4+9=14$$

Therefore, the sum of Greg's three digits is 14.

The solution provided by AI Math Solver is incorrect, as the correct answer is (A) 9, corresponding to the digits 0, 1, and 8. The model misidentified the number of horizontal and vertical segments per digit, likely due to difficulties interpreting the visual representation. Nevertheless, its reasoning process follows a coherent and structured approach, which can still aid students in understanding problem-solving strategies. Interestingly, when the problem was presented in Serbian, the model produced different segment counts, leading to its failure to reach the correct answer and generating numerous solutions not included among the provided options.

3.(11-12 grade problem, 5 points) The values of x , y and z satisfy $2^x = 3$, $2^y = 7$ and $6^z = 7$. Which of the following gives the relationship between x , y and z ?

(A) $z = \frac{y}{1+x}$ (B) $z = \frac{x}{y} + 1$ (C) $z = \frac{y}{x} - 1$ (D) $z = \frac{x}{y-1}$ (E) $z = y - \frac{1}{x}$

Solution provided by AI Math Solver:

To find the relation between x , y and z , let's start by analyzing the given equations:

1. $2^x = 3$
2. $2^y = 7$
3. $6^z = 7$

We can rewrite equation (3) using the fact that $6 = 2 \times 3$:

$$(2 \times 3)^z = 7$$

Using the property of exponents that $(ab)^c = a^c b^c$, we can rewrite the equation as:

$$2^z \times 3^z = 7$$

Now, we can substitute $2^x = 3$ into this equation to get:

$$2^z \times (2^x)^z = 7$$

Using the property of exponents that $(a^b)^c = a^{bc}$, we can rewrite the equation as:

$$2^z \times 2^{xz} = 7$$

Using the property of exponents that $a^b \times a^c = a^{b+c}$, we can rewrite the equation as:

$$2^{z+xz} = 7$$

Now, we can equate this equation with equation (2) $2^y = 7$:

$$2^{z+xz} = 2^y$$

Since the bases are the same, we can equate the exponents:

$$z + xz = y$$

Now, we can factor out z from the left-hand side:

$$z(1 + x) = y$$

Dividing both sides by $(1 + x)$, we get:

$$z = \frac{y}{1+x}$$

Therefore, the relationship between x , y and z is given by:

$$z = \frac{y}{1+x}$$

This solution is correct and clearly demonstrates the steps leading to the right answer. The solver effectively applies exponent rules and logical substitutions, resulting in a precise derivation. The step-by-step reasoning adds significant pedagogical value, making the solution useful for educational purposes. Interestingly, when presented with the same problem in Serbian, the AI failed to reach the correct conclusion despite starting with a similar logical approach.

5.2. Quantitative Analysis of Results

We will now analyze the success of the AI Math Solver in relation to task category, difficulty, and the presence of images in the problem statement or answer choices.

First, we compared the proportion of tasks the AI Math Solver successfully solved, those for which it produced incorrect solutions, and those for which it failed to generate any solution, relative to the total number of tasks across the three selected categories. It was expected that as the mathematical complexity of the tasks increased, the number of correct answers provided by the AI Math Solver would decrease, accompanied by a rise in incorrect solutions or unsolved tasks. However, the results do not support this conclusion.

Specifically, when analyzing tasks presented in Serbian, the highest percentage of correctly solved tasks was recorded in the 7th and 8th grade elementary school category, where exactly one-third of the tasks were solved correctly. In the other two categories, the success rate was lower - 6 out of 24 tasks were correctly solved in the 3rd and 4th grade elementary school category, while in the 3rd and 4th grade high school category, the percentage of correct answers was slightly higher (26.67%) (Figure 2).

When it comes to accuracy in solving tasks formulated in English, the results differ somewhat. There is a slight increase in the AI Math Solver's success rate as the target students' age increases. In the 3rd and 4th grade elementary school category, 25% of the tasks were solved correctly, while in the 7th and 8th grade category, this percentage rose slightly to 26.26%. In the oldest category, intended for 3rd and 4th year high school students, the AI tool solved one in every three tasks successfully (33.33%).

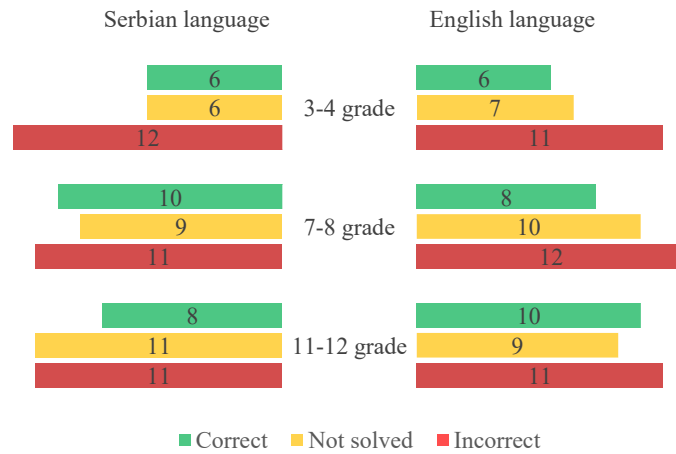


Fig. 2 Success in Solving Problems in Serbian language (left) and in English language (right) by category

Given the very small differences in percentages and the limited sample size, we cannot draw a clear and unambiguous conclusion about the relationship between the success of the AI Math Solver tool and the group for which the tasks were intended.

Furthermore, since the tasks within each category were divided into three difficulty levels, we examined whether the AI Math Solver's success varied across these difficulty levels (Figure 3). Of the 28 tasks worth 3, 4, and 5 points respectively, the highest number of correct answers in Serbian was observed for tasks worth 3 points (12 tasks), while the number of correct solutions was significantly lower for tasks worth 4 points (7 tasks) and even lower for those worth 5 points (5 tasks).

For the same tasks formulated in English, the AI Math Solver correctly solved 11 tasks in the 3-point category. Interestingly, in English, the AI tool solved a higher percentage of the most difficult tasks (those worth 5 points – 8 tasks) compared to the tasks of medium difficulty (only 5 out of 28 tasks).

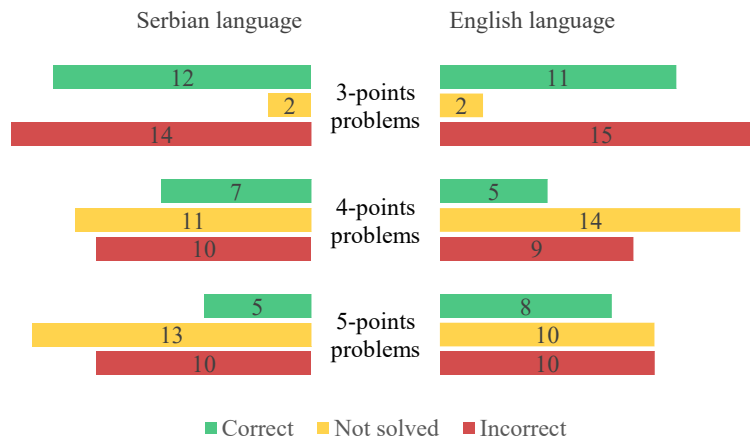


Fig. 3 Success in Solving Problems in Serbian language (left) and in English language (right) by degree of difficulty

As mentioned, a large number of tasks contain images in their formulation. Therefore, we sought to examine whether the presence of an image makes the problem-solving process more difficult or possibly easier for the AI Math Solver. Out of a total of 84 tasks, 45 included some graphical representations, both in the problem statements and in the provided answer choices. Of these 45 tasks containing images, the AI Math Solver successfully solved 13 tasks (28.88%) in Serbian, while it solved exactly 11 out of the remaining 39 tasks without images (28.20%). Thus, the success rate is nearly identical. When the tasks were formulated in English, the same percentages of successfully solved tasks was observed for those that with images (28.88%) and those without images (28.20%). This indicates that the presence of an image does not affect the overall success rate of the AI Math Solver on Kangaroo competition tasks.

However, Figures 4 reveal another interesting finding. For a significantly larger number of tasks, where solving the problem required understanding the image and abstracting data from it, the AI tool was unable to provide a solution compared to tasks without images. Specifically, the percentage of tasks with no solution provided in the image – containing group was 40% in Serbian and 42.22% in English, whereas the percentage of unsolved tasks without images was 20.51% in Serbian and 17.95% in English. This highlights the limitations of the AI Math Solver's capabilities, particularly when solving geometric problems.

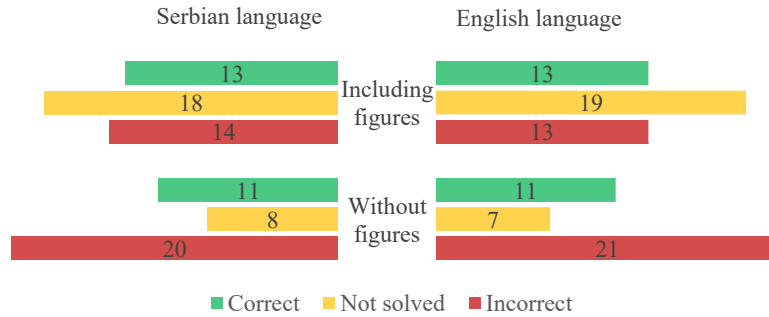


Fig. 4 Success in solving tasks in Serbian language (left) and in English language (right) depending on whether the formulation includes an image

As mentioned earlier, for some problems, the AI Math Solver provided different answers depending on whether they were formulated in Serbian or English. Moreover, in a few cases, it generated an answer in one language but failed to produce a solution in the other. Therefore, we aimed to determine the extent to which the answers produced by this AI tool matched when the problems were given in both Serbian and English. Out of a total of 24 problems for 3rd and 4th grades of elementary school, the AI Math Solver produced the same answer in both languages for only four problems (16.67%). Additionally, for three more problems, it failed to generate an answer in either Serbian or English. A similar percentage of identical answers was in the set of 30 problems for 7th and 8th grades of elementary school, where 7 problems (23.33%) had matching answers in both languages. Furthermore, for five problems, the AI MathSolver failed to produce a solution in both languages.

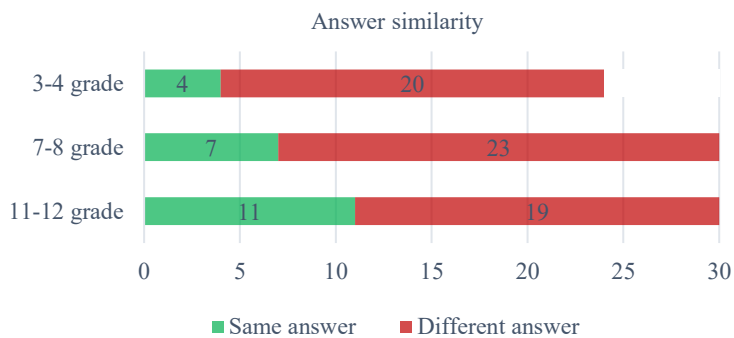


Fig. 5 Answer similarity based on answers given in Serbian and English language

The highest percentage of identical answers occurred in the group of problems for students in the 3rd and 4th years of high school. Out of 30 problems in this category, the AI tool produced

the same answers in both English and Serbian in 11 cases (36.67%). This category included the most challenging problems, and the solver failed to generate a solution in both languages for as many as 20% of the tasks. Overall, out of a total of 84 problems, the AI tool produced identical solutions for only 22 problems (26.19%), in both English and Serbian which is still a relatively low percentage. One possible reason for this low level of answer similarity is that some problems were slightly modified during translation to make them more accessible to students, often placing them in a more realistic (quasi-realistic) context.

The success of AI Math Solver fell below expectations. With only 24 correctly solved tasks out of 84 in both Serbian and English, the tool's overall performance cannot be considered successful. Specifically, if a 3rd or 4th grade elementary school student had provided the same responses (correct, incorrect, or unanswered) as AI Math Solver when solving tasks in Serbian, they would have scored 35.5 points out of a maximum of 120. This result would place AI Math Solver in the lowest third of 3rd and 4th grade students who participated in the 2024 competition (Table 1). Regarding the 7th and 8th grade students, the tool's performance would position it approximately in the middle of the ranking. Similarly, in the competition among 3rd and 4th year high school students, the obtained result would place AI Math Solver around the median of the ranking.

Based on these findings, we conclude that the performance of the analyzed AI tool remains significantly below that of the top competitors in the 2024 Kangaroo competition in Serbia.

Table 1 AI Math solver results compared to Serbian students' results

	AI Math Solver points	Maximum number of points	Number of competitors in Serbia	The ranking of AI Math Solver among students in Serbia
3 rd grade of elementary school	35.5	120	3953	2778 th
4 th grade of elementary school			3377	2825 th
7 th grade of elementary school	56	150	1197	462 nd
8 th grade of elementary school			964	510 th
3 rd grade of high school	49.5	150	671	350 th
4 th grade of high school			407	214 th

In addition to these results, several specific errors were observed when testing tasks with AI Math Solver, highlighting challenges in linguistic and technical processing. For example, the tool sometimes provided a solution in English for a problem originally written in Serbian, which may be attributed to language confusion in large language models (see, e.g., Marchisio et al., 2024). In some solutions, there were alternating uses of Cyrillic and Latin scripts, as well as the appearance of words that do not belong to Serbian but to other Slavic languages. Additionally, certain mathematical formulas were rendered in LaTeX source code rather than a standard readable format, etc.

6. DISCUSSION

This section discusses key factors that influenced the performance of the AI Math Solver. As highlighted in previous research, AI systems designed for mathematical problem-solving often face substantial limitations in generalization and abstraction, especially when dealing with tasks that require multimodal reasoning and the integration of diverse skills (Cherian et al., 2023). These challenges are clearly reflected in the results of our evaluation. Below, we examine the main reasoning patterns that contributed to the tool’s underperformance.

One of the primary reasons is the lack of adequate training datasets for non-standard mathematical tasks, which frequently feature unusual formulations or uncommon graphical representations. This limitation restricts AI models’ ability to generalize and effectively solve such problems. For example, Zhang et al. (2024) demonstrate that pre-trained models consistently outperform their non-pre-trained counterparts in mathematical text generation tasks, highlighting the crucial role of training data in achieving better results. This limitation is well illustrated by the first problem discussed above. Although the AI Math Solver applied a seemingly valid case analysis, it failed to properly model a key constraint—that the item prices must be distinct whole numbers summing to seven. This suggests a deeper issue in problem modeling rather than in the specific solution approach.

Regarding the concerning results in the performance of the AI Math Solver on simpler tasks that do not require advanced mathematical knowledge, the causes can largely be attributed to language specifics and contextual understanding. For example, tasks designed for younger students are often formulated to be engaging and motivating. While students, parents, and teachers understand the essence and requirements of such formulations intuitively, this can pose challenges for the AI tool, which is trained to solve problems with clearly structured text and explicit instructions. For an in-depth survey of challenges in understanding and solving mathematical word problems, see Sundaram et al. (2024). Another limitation is evident in the third task, where the AI Math Solver provided a correct solution in English but produced an incorrect answer when the same task was translated into Serbian. Although the underlying reasoning process was similar, this discrepancy highlights a weakness in the model’s multilingual mathematical reasoning—small differences in phrasing or grammatical structure between languages can lead to divergent outcomes and misalignment between natural language and mathematical interpretation.

Additionally, AI systems exhibit a reduced ability to adapt to new contexts, which further reduce their performance. Unlike humans, who rely on intuition and past experiences when facing unfamiliar problems, AI systems tend to perform poorly when confronted with unconventional tasks. This lack of adaptability limits the AI’s effectiveness in non-standard scenarios. For a comprehensive discussion of mathematical problems, related datasets, and factors influencing LLMs in mathematical problem-solving, see Ahn et al. (2024).

While images can support student understanding, they often complicate tasks for AI tools. Factors such as colour, labeling and image quality, especially in playful visuals aimed at younger students or more intricate graphics for higher grades, can hinder accurate interpretation. These challenges align with findings by Yiu et al. (2024), who demonstrate that although large multimodal models like GPT-4V handle basic visual features well, they struggle with abstract reasoning and complex visual transformations, largely because their training focuses primarily on 2D data and text.

A similar limitation is evident in the second task, which involved counting line segments used to write digits. Here, the AI followed a structured and pedagogically clear line of reasoning

but misinterpreted the visual representation, resulting in an incorrect answer. While the reasoning process itself was coherent, the failure to accurately extract visual features highlights a broader challenge in AI's ability to process and integrate image-based information.

Non-standard tasks often employ uncommon formulations, reducing the effectiveness of AI tools that are typically trained on conventional symbols and labels. In Kangaroo competition tasks, information is sometimes divided between text and images (not always geometric), requiring students to synthesize data from multiple sources. Such formats demand multimodal reasoning—combining language and image processing, which poses an additional challenge for AI systems.

7. CONCLUSION

Considering that tasks in the Kangaroo competition are designed to be relatively straightforward, requiring students to solve many problems quickly and without resorting to complex procedures, stronger performance from the AI Math Solver was reasonable expected. However, the obtained results fall significantly below these expectations. One contributing factor might be the mode of task input: all problems were uploaded as screenshots, and it remains unclear whether providing tasks as copied text along with answer choices would yield better outcomes. Overall, the findings suggest that the AI Math Solver would benefit from additional fine-tuning using data from previous years' problems, likely due to insufficient training on non-standard mathematical tasks such as those featured in Kangaroo competitions.

Acknowledgement: Authors are supported by the Ministry of Science, Technological Development and Innovation, Republic of Serbia. The first author acknowledges support under Contract No. 451-03-136/2025-03/200139, while the second, third and fourth author are supported under Contract No. 451-03-137/2025-03/ 200122.

REFERENCES

- Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., & Yin, W. (2024). Large Language Models for Mathematical Reasoning: Progresses and Challenges. *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 225–237. Association for Computational Linguistics. <https://aclanthology.org/2024.eacl-srw.17/>
- Akveld, M., Caceres-Duque, L. F., Nieto Said, J. H., & Sánchez Lamóneda, R. (2020). The Math Kangaroo Competition. *Espacio Matemático* 1(2), 74–91. <https://doi.org/10.3929/ETHZ-B-000456237>
- Castelvecchi, D. (2024). DeepMind hits milestone in solving maths problems — AI's Next Grand Challenge. *Nature*, 632(8024), 236–237. <https://doi.org/10.1038/d41586-024-02441-2>
- Cherian, A., Peng, K., Lohit, S., Smith, K.A., & Tenenbaum, J.B. (2023). Are Deep Neural Networks SMARTer Than Second Graders? *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10834–10844. <https://doi.org/10.1109/cvpr52729.2023.01043>
- Cherian, A., Peng, K.-C., Lohit, S., Matthiesen, J., Smith, K., & Tenenbaum, J.B. (2025). Evaluating Large Vision-and-Language Models on Children's Mathematical Olympiads. *NIPS '24: Proceedings of the 38th International Conference on Neural Information Processing Systems*, 15779–15800 <https://dl.acm.org/doi/10.5555/3737916.3738420>
- DeepMind. (2024). AI achieves silver-medal standard solving International Mathematical Olympiad problems. *DeepMind Blog*. Retrieved December 2, 2024, from <https://deepmind.google/discover/blog/ai-solves-imo-problems-at-silver-medal-level/>
- Elbanna, S., & Armstrong, L. (2023). Exploring the integration of ChatGPT in education: adapting for the future. *In Management & Sustainability: An Arab Review* 3(1), 16–29. <https://doi.org/10.1108/msar-03-2023-0016>

- Frieder, S., Pinchetti, L., Chevalier, A., Griffiths, R.R., Salvatori, T., Lukasiewicz, T., Petersen, P., & Berner, J. (2024). Mathematical Capabilities of ChatGPT. *Proceedings of the 37th International Conference on Neural Information Processing Systems*, 27699–27744. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/58168e8a92994655d6da3939e7cc0918-Paper-Datasets_and_Benchmarks.pdf
- Koncel-Kedziorski, R., Roy, S., Amini, A., Kushman, N., & Hajishirzi, H. (2016). MAWPS: A Math Word Problem Repository. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1152–1157. Association for Computational Linguistics. <https://doi.org/10.18653/v1/n16-1136>
- Lo, C. K. (2023). What Is the Impact of ChatGPT on Education? A Rapid Review of the Literature. *Education Sciences* 13(4), 410. MDPI. <https://doi.org/10.3390/educsci13040410>
- Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.-W., Galley, M., & Gao, J. (2024). MathVista: Evaluating Mathematical Reasoning of Foundation Models in Visual Contexts. *Proceedings of ICLR*. <https://openreview.net/attachment?id=KUNzEQMWU7&name=pdf>
- Marchisio, K., Ko, W., Bérard, A., Dehaze, T., & Ruder, S. (2024). Understanding and mitigating language confusion in LLMs. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 6653–6677. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.380>
- Memarian, B., & Doleck, T. (2023). ChatGPT in education: Methods, potentials, and limitations. *Computers in Human Behavior: Artificial Humans* 1(2), 100022. Elsevier BV. <https://doi.org/10.1016/j.chbah.2023.100022>
- Simjanović, D., Randjelović, B., Vesić, N., & Penjišević, A. (2022). Examples of mathematical problems in primary and secondary education that include the actual calendar year. *Facta Universitatis, Series: Teaching, Learning and Teacher Education*, 5(2), 191–200. <https://doi.org/10.22190/futlte210617015s>
- Spasić, A. J., & Janković, D. S. (2023). Using ChatGPT Standard Prompt Engineering Techniques in Lesson Preparation: Role, Instructions and Seed-Word Prompts. *2023 58th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)*, 47–50. <https://doi.org/10.1109/icest58410.2023.10187269>
- Stanković, M., Milenković, A., Svičević, M., & Vučićević, N. (2025). Performance of an AI Tool in Solving Non-Standard Mathematics Competition Problems. *1st International Scientific Conference Education and Artificial Intelligence (EDAI 2024)*, 165–174. <https://doi.org/10.46793/EDAI24.165S>
- Sundaram, S. S., Gurajada, S., Padmanabhan, D., Abraham, S. S., & Fisichella, M. (2024). Does a language model “understand” high school math? A survey of deep learning based word problem solvers. *Wiley Interdisciplinary Reviews. Data Mining and Knowledge Discovery* 14(4). <https://doi.org/10.1002/widm.1534>
- Trinh, T. H., Wu, Y., Le, Q. V., He, H., & Luong, T. (2024). Solving olympiad geometry without human demonstrations. In *Nature*, 625(7995), 476–482. <https://doi.org/10.1038/s41586-023-06747-5>
- Wei, X. (2024). Evaluating chatGPT-4 and chatGPT-4o: performance insights from NAEP mathematics problem solving. *Frontiers in Education*, 9, Article 1452570. <https://doi.org/10.3389/educ.2024.1452570>
- Yiu, E., Qraitem, M., Wong, C., Majhi, A. N., Bai, Y., Ginosar, S., Gopnik, A., & Saenko, K. (2024). KiVA: Kid-inspired Visual Analogies for Testing Large Multimodal Models (Version 1). *arXiv*. <https://doi.org/10.48550/ARXIV.2407.17773>
- Zhang, F., Li, C., Henkel, O., Xing, W., Baral, S., Heffernan, N., & Li, H. (2024). Math-LLMs: AI Cyberinfrastructure with Pre-trained Transformers for Math Education. *International Journal of Artificial Intelligence in Education*. <https://doi.org/10.1007/s40593-024-00416-y>
- Zhao, J., Zhang, Z., Zhang, Q., Gui, T., & Huang, X. (2024). LLaMA Beyond English: An Empirical Study on Language Capability Transfer. *ArXiv*. <https://doi.org/10.48550/arXiv.2401.01055>

USPEŠNOST AI MATH SOLVER ALATA U REŠAVANJU NESTANDARDNIH ZADATAKA SA MATEMATIČKIH TAKMIČENJA

Veštačka inteligencija sve više menja način na koji se učenici obrazuju, pa tako i pristup učenju matematike i rešavanju zadataka, pružajući dodatnu podršku i pomoć, što predstavlja sve češćipredmet naučnih istraživanja. Jedan pravac istraživanja fokusira se na pomoć učenicima koji žele da učestvuju na matematičkim takmičenjima u rešavanju složenijih matematičkih problema. Pored redovnih nacionalnih matematičkih takmičenja, koja učenicima omogućavaju napredovanje do međunarodnih matematičkih olimpijada, postoje takmičenja usmerena na popularizaciju

matematike i razvoj logičkog mišljenja kod učenika. Jedno od takvih jeste međunarodno takmičenje Kengur bez granica. U ovom radu ispitujemo uspešnost alata AI Math Solver, dostupnog na platformi Interactive Mathematics, u rešavanju zadataka sa takmičenja Kengur bez granica iz 2024. godine. Istraživanje obuhvata zadatke namenjene za tri uzrasne grupe: 3. i 4. razreda osnovne škole, 7. i 8. razreda osnovne škole i 3. i 4. razreda srednje škole. Zadaci su postavljeni u vidu slika (screenshot-ova), na srpskom i engleskom jeziku, zbog česte prisutnosti vizuelnih elemenata u formulacijama i ponuđenim odgovorima. Rezultati su predstavljeni kroz dve celine: kvalitativnu analizu odabranih zadataka i kvantitativnu analizu dobijenih rezultata. Od ukupno 84 zadatka, kako na srpskom tako i na engleskom jeziku, tačno su rešena 24 zadatka, što je nešto manje od 30% uspešnosti u oba slučaja. Dalje, neki zadaci rešeni na srpskom nisu rešeni na engleskom jeziku, i obrnuto. Pored toga, uočene su razlike u raspodeli tačnih odgovora među zadacima različitih nivoa težine.

Ključne reči: *AI alati, Kengur bez granica, matematičko obrazovanje, nestandardni zadaci*