EDUCATION AND ARTIFICIAL INTELLIGENCE

International Scientific Conference

November 29/30, 2024

BOOK OF PROCEEDINGS

ПЕДАГОШКИ ФАКУЛТЕТ
ВРАЊЕ

EDUCATION AND ARFTIFICIAL INTELLIGENCE

(EDAI 2024)

ПЕДАГОШКИ ФАКУЛТЕТ · ВРАЊЕ ·

1st International Scientific Conference

Education and Artificial Intelligence (EDAI 2024)

Publisher

Pedagogical Faculty in Vranje, University of Niš, Serbia

For the publisher

Prof. Dragana Stanojević, PhD

Editors

Aleksandar Spasić, PhD

Darko Stojanović, PhD

Organizer

Pedagogical Faculty in Vranje, University of Niš, Serbia

Co-organizers

Faculty of Pedagogy, South-West University "Neofit Rilski", Blagoevgrad, Republic of Bulgaria

Faculty of Education, University "St. Kliment Ohridski", Bitola, Republic of North Macedonia

Faculty of Educational Sciences, University "Goce Delčev", Štip, Republic of North Macedonia

Faculty of Philosophy, University of Banja Luka, Republic of Srpska, Bosnia and Herzegovina

Proofreader

Biljana Savić

Technical editor

Darko Stojanović, PhD

Cover design

MSc Lidija Tasić

The publication of the Book of Proceedings has been funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

Република Србија

Министарство просвете,
науке и технолошког развоја

1st International Scientific Conference

Education and Artificial Intelligence (EDAI 2024)

BOOK OF PROCEEDINGS

Vranje, November 29–30, 2024

International Programme Committee

Asya Berberyan, PhD, Psychology Department at Russian-Armenian (Slavonic) University, Armenia

Danče Sivakova Neškovski, PhD, Faculty of Pedagogy in Bitola, University “St. Kliment Ohridski”, Republic of North Macedonia

Emilija Petrova Gjorgjeva, PhD, Faculty of Educational Sciences, University “Goce Delčev”, Štip, Republic of North Macedonia

Mariana Neagu, PhD, Faculty of Letters - "Dunărea de Jos" University of Galați, Romania

Danijela Zdravković, PhD, Pedagogical Faculty in Vranje, University of Niš, Serbia

Miljana Mladenović, PhD, Pedagogical Faculty in Vranje, University of Niš, Serbia

Tatjana Milosavljević Đukić, PhD, Pedagogical Faculty in Vranje, University of Niš, Serbia

Biljana Novković Cvetković, PhD, Pedagogical Faculty in Vranje, University of Niš, Serbia

Aleksandar Stojadinović, PhD, Pedagogical Faculty in Vranje, University of Niš, Serbia

Ana Spasić Stošić, PhD, Pedagogical Faculty in Vranje, University of Niš, Serbia

Ivana Tasić Mitić, PhD, Pedagogical Faculty in Vranje, University of Niš, Serbia

Lazar Stošić, PhD, Union - Nikola Tesla University, Belgrade, Faculty of Informatics and Computer Science, Serbia / Don State Technical University, Rostov-on-Don, Russian Federation

Dragan Janković, PhD, Faculty of Electronic Engineering, University of Niš, Serbia

Bratislav Predić, PhD, Faculty of Electronic Engineering, University of Niš, Serbia

Marija Jovanović, PhD, Faculty of Philosophy, University of Niš, Serbia

Dragana Jovanović, PhD, Faculty of Philosophy, University of Niš, Serbia

Organizing committee

Aleksandra Milanović, PhD, **Chairmen**, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Milica Aleksić, PhD, **Co-Chairmen**, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Darko Stojanović, PhD, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Marko Stanković, PhD, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Dragana Stanković, PhD, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Vladislav Krstić, PhD, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Milica Ristić, PhD, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Biljana Prodović Milojković, PhD, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Vesna Zdravković, PhD, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Sanja Anđelković, PhD, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Katarina Stanković, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Jelena Jovanović Kostić, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Marija Tasić, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Mirjana Đokić, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Jelena Krstić, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Lidija Tasić, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Marija Dejković, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Marija Nešić, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Jovana Arsić, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Anđela Protić, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Jovana Stošić, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Milan Krstić, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Bratislav Nikolić, MSc, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

Biljana Savić, *Pedagogical Faculty in Vranje, University of Niš, Serbia*

PROCEEDINGS OF THE 1ST INTERNATIONAL SCIENTIFIC CONFERENCE
Education and Artificial Intelligence (EDAI 2024)
https://doi.org/10.46793/EDAI24.165S pp. 165 - 174

Original research paper

PERFORMANCE OF AN AI TOOL IN SOLVING NON-STANDARD MATHEMATICS COMPETITION PROBLEMS

UDC 372.27::51:004.85

Marko Stanković[1], Aleksandar Milenković[2], Marina Svičević[2], Nemanja Vučićević[2]

[1]Pedagogical Faculty in Vranje, University of Niš, Serbia
[2]University of Kragujevac, Faculty of Science, Serbia

Abstract*. For some time now, researchers around the world have been examining the effects of using AI in mathematics education to provide additional support and assistance to students. One line of research focuses on helping students who wish to participate in 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 competitions aimed at popularizing mathematics and developing logical thinking in students. One such competition is the international Kangaroo competition. In this paper, we analyze the performance of the AI Math Solver on the Interactive Mathematics platform in solving problems from the 2024 Kangaroo competition for students in the 3rd and 4th grades of elementary school, as well as the 7th and 8th grades of elementary school, and the 3rd and 4th grades of high school. The tasks were uploaded in the form of images (screenshots), both in Serbian and English, because in the formulation of the tasks and/or provided answers for the Kangaroo competition, images often appear. Out of a total of 84 tasks, both in Serbian and in English, it correctly solved 24, which is just under 30% success in both cases. Furthermore, some tasks solved in Serbian were not solved in English, and vice versa. Additionally, differences were found in the distribution of correct answers among tasks of different difficulty levels.*

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

1. 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 shown both potential and challenges for use in education. 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 contexts (cf. Lo, 2023). For example, it can be effectively integrated into education to automate routine tasks and enhance the learning experience for students (Elbanna and Armstrong, 2023), as well as to assist teachers in lesson preparation (Spasić and Janković, 2023). However, using ChatGPT in

Corresponding author: Marko Stanković
University of Niš, Pedagogical Faculty in Vranje, Partizanska 14, 17500 Vranje, Serbia
Phone: + 381 17 422 962 • E-mail: markos@pfvr.ni.ac.rs

education raises ethical concerns, including plagiarism, reduced learning engagement, and user privacy, etc. (Memarian and Doleck, 2023).

Although studies have shown possible use in various school subjects, ChatGPT cannot be adequately used for certain subjects such as mathematics. A detailed assessment of the capabilities of ChatGPT indicates significant limitations in solving complex mathematical tasks, especially at the level of postgraduate studies (Frieder et al., 2023). However, ChatGPT can be a valuable assistant in math fact checking and information retrieval. Recent research by Wei (2024) shows that advanced AI models like ChatGPT-4 and ChatGPT-4o generally surpass U.S. students' performance across all grades, content areas, item types, and difficulty levels. However, they still struggle with specific areas such as geometry and high-difficulty mathematical tasks. It is therefore not surprising that there is a growing range of AI tools specifically designed to solve mathematical problems, which are adapted for use in formal education and research work.

Particularly interesting are tools that can solve problems from International Mathematical Olympiad (IMO) at the level of human performance (DeepMind, 2024). For example, AlphaGeometry is a neuro-symbolic AI system designed to solve complex Euclidean geometry problems without requiring human-annotated training data. It synthesizes millions of theorems and proofs, allowing it to solve Olympiad-level problems with human-like performance, producing readable proofs, and even outperforming prior methods on key benchmarks (Trinh et al, 2024). It was soon upgraded to AlphaGeometry 2, with even better performance. AlphaProof is an AI system designed to solve formal mathematical problems by proving or disproving statements using a formal language called Lean. It combines reinforcement learning with a language model to generate and verify solutions. This system solved one of the hardest problems at the IMO 2024, achieving a silver-medal equivalent score (Castelvecchi, 2024).

In addition to IMO, there are competitions designed to popularize mathematics and enhance students' logical thinking skills. A prominent example of one such 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 aims to explore.

The challenges associated with AI systems solving tasks that demand broad reasoning skills have been previously explored in the context of the SMART-101 dataset (Cherian et al., 2023). SMART-101 evaluates visuo-linguistic puzzles designed for children aged 6–8, requiring skills such as arithmetic, algebra, and spatial reasoning. This dataset comprises 101 puzzles derived from nearly 10 years of Math Kangaroo USA competitions. While large models show promising reasoning abilities in this domain, their solutions often fall short of accuracy. This shows the limitations of current AI systems in generalization and abstraction, particularly when the problem-solving context requires multimodal reasoning and the integration of diverse skills. In contrast to SMART-101, which targets children aged 6–8, this study extends the investigation to tasks suited for older students from the Mathematical Kangaroo competition. These tasks are more diverse in complexity, often involving higher-order reasoning, and provide a broader perspective on the performance of AI tools in educational settings. Additionally, we employ a specialized AI tool for mathematics and test the tasks in both English and Serbian (using Cyrillic script).

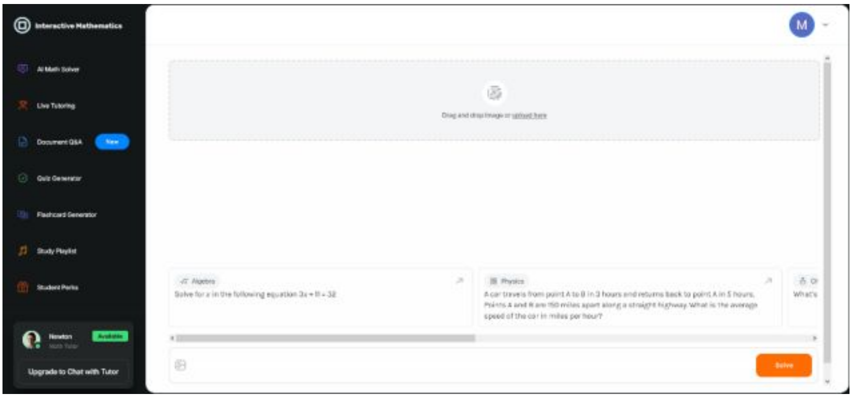
In the following sections, we provide an overview of the Interactive Mathematics platform, and the AI tool used in this study, AI Math Solver. We then describe the Mathematical Kangaroo competition, highlighting its unique features. The study proceeds to present the results of applying the AI tool to competition tasks, comparing its performance on problems presented in Serbian and in English. Many popular large language models (e.g., LLaMA) are trained primarily on English-centric data, which hampers their performance in languages other than English (Zhao et al., 2024). For this reason, we decided to test the AI Math Solver to solve tasks both in Serbian and in English. Finally, we offer a discussion of the results and provide concluding remarks.

2. Interactive Mathematics and AI Math Solver

The research was conducted using the Interactive Mathematics platform[1], a math education platform combining human educators and AI computing. This platform has many features, like an AI Math Solver[2], live tutoring, quiz generator, and flashcards to help with learning and problem solving. The platform is designed with a strong educational focus to really help students practice and improve their math performance. Although the platform includes numerous features, it maintains an intuitive and user-friendly interface. Some features on the platform are free to use, while others are available for a reasonable fee, making them affordable for students. It is also important to mention that Interactive Mathematics is trusted by prestigious institutions such as MIT, Yale, and Harvard.

AI Math Solver is a tool on the platform Interactive Mathematics that uses a mathe-matical computation engine that excels at solving mathematical formulas with the power of AI 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 Fig. 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, 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. Additionally, for tasks involving functions, the platform sometimes presents a "Generate graph" button, whenever the tool can generate a suitable graphical representation for the given problem type, enabling users to visualize the function's graph, which is especially helpful for understanding complex problems.



Solve for x in the following equation 3x + 11 = 32

A car travels from point A to B in 3 hours and returns back to point A in 5 hours. Points A and B are 150 miles apart along a straight highway. What is the average speed of the car in miles per hour?

What's

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

[1] https://www.intmath.com/

[2] https://app.intmath.com/

3. MATHEMATICAL KANGAROO

Mathematical Kangaroo is an international mathematics competition held in over 100 countries, coordinated globally by the Association Kangourou Sans Frontières (AKSF). The main goal of the Mathematical Kangaroo competition is the popularization of mathematics, i.e., to increase interest in mathematics and natural sciences, as well as the level of logical and combinatorial thinking, understanding of texts and application of acquired mathematical knowledge.

In order to achieve the goals of the competition and show the beauty of mathematics, the tasks are carefully chosen at the annual AKSF meeting. Mathematicians and educators from all over the world select tasks from a database of pre-submitted and rated questions. "The questions are not standard textbook problems and come from a large variety of topics. Besides inspiring ideas, perseverance and creativity, they require imagination, basic computational skills, logical thinking and other problem-solving strategies" (Akveld et al., 2020).

Mathematical Kangaroo is a test competition. Each task has 5 possible answers, of which only one is correct. The test contains 30 tasks with three degrees of difficulty (10 tasks each), which students from the 5th grade of elementary school to the 4th grade of high school do for 90 minutes. Students in the 3rd and 4th grades of elementary school must answer 24 questions in 75 minutes and for them the tasks are divided according to the difficulty of the group of 8 tasks. Also, in Serbia, students in the 1st and 2nd grades of elementary school must answer 18 questions (divided into three groups of six problems based on the difficulty degree) in 60 minutes.

4. RESEARCH QUESTION

The aim of this paper is to determine the performance of the AI Math Solver in solving three categories of non-standard tasks from the Kangaroo competition, specifically for 3rd-4th grade elementary school, 7th-8th grade elementary school, and 3rd-4th grade high school. Additionally, we aim to compare these results with those achieved by students in Serbia in 2024, in the same categories. We decided to test performance in these three categories because we wanted to examine the impact of the age group (and the complexity of the tasks in terms of the content students need to master to solve them) for which the tasks are intended. Therefore, we deliberately did not choose consecutive categories.

To address this, we presented the tasks as screenshots, as many tasks (45 out of 84) include an image in their formulation or provide answer options in image form. Additionally, considering previous findings on the impact of language when presenting tasks to the AI tool, we uploaded the tasks in both Serbian and English to determine whether there are significant differences in the AI Math Solver's performance depending on the language.

The research was conducted during September and October 2024.

5. RESULTS

First, we analyzed whether there were differences in the number of tasks that the AI Math Solver solved successfully versus number of the tasks that it solved unsuccessfully or it didn't produce the solution. It was expected that, as task complexity increased with higher grade levels, the number of correct answers provided by the AI Math Solver would decrease. However, based on the results, this conclusion cannot be drawn.

Moreover, regarding task-solving in Serbian language, it can be observed that the highest percentage of solved tasks was in the 7th-8th grade category (33.33%). In the other two categories, the percentage was somewhat lower – 25% of the tasks were solved in the 3rd-4th grade category for elementary school, while a nearly identical percentage of correct answers (26.67%) was found in the 3rd-4th grade category for high school (Fig. 2).

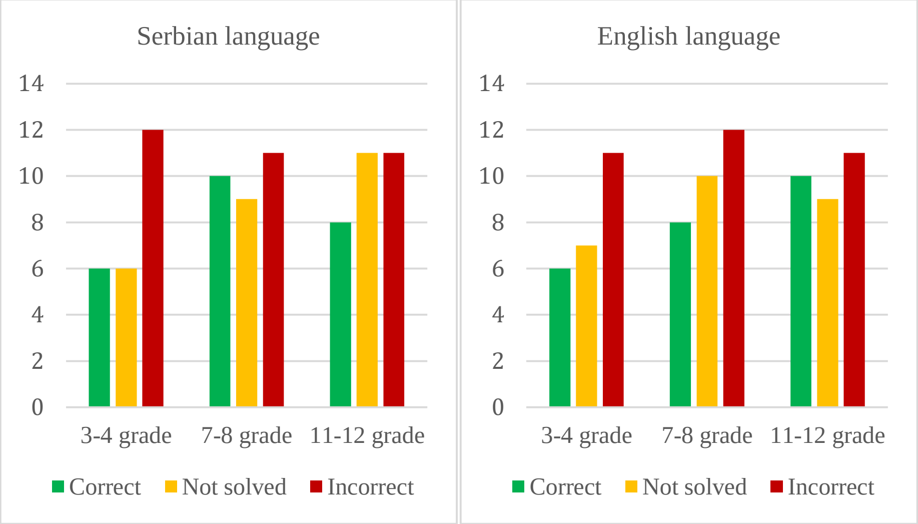


Fig. 2 Success in Solving Problems in Serbian language (left) and in English language (right) by category

Interestingly, when it comes to accuracy in problems formulated in English language, there is a slight increase in success rates as the grade level for which the tasks are intended rises. Specifically, for 3rd-4th grade, 25% of the tasks were solved correctly; in 7th-8th grade, the rate was 26.26%, and in high school (3rd-4th grade), it reached 33.33%. Given the very small differences in percentages and the limited sample size, we cannot conclude that this AI tool's success rate in solving non-standard math competition tasks significantly increases or decreases with the grade level for which the tasks are designed.

After that, we aimed to examine whether there is a difference in the success of the AI Math Solver tool in solving tasks of varying difficulty levels (Fig. 3). Out of a total of 28 tasks across all three difficulty levels, the highest number of correct answers in Serbian language was observed in the easiest category, with tasks worth 3 points (42.86%), while the proportion of correct solutions was much lower for tasks worth 4 points (25%) and those worth 5 points (17.86%). Interestingly, in the English version, the AI tool solved one fewer task worth 3 points (39.29%) but managed to solve a higher percentage of the most difficult tasks (those worth 5 points, 28.57%) compared to medium-difficulty tasks (only 17.86%).

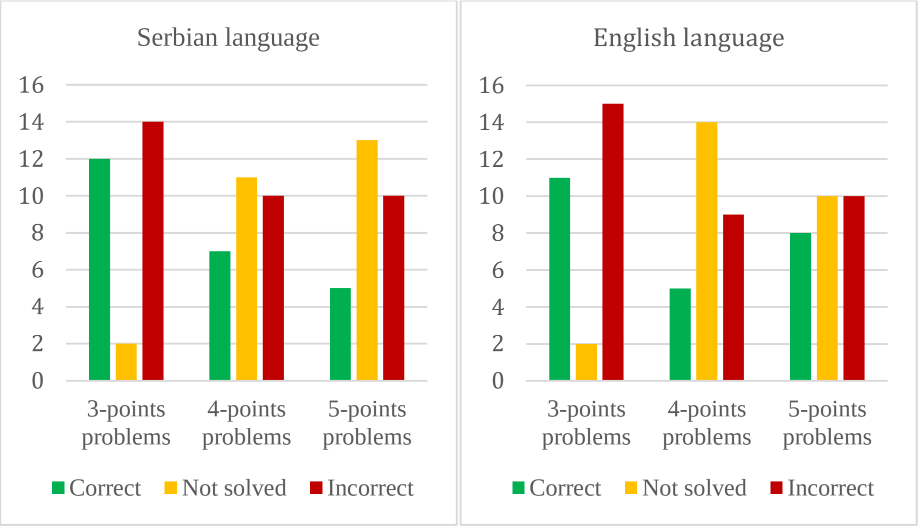


Fig. 3 Success in Solving Problems in Serbian language (left) and in English language (right) by degree of difficulty

Regarding the success rate in solving tasks based on whether the problem formulation or answer choices included images requiring interpretation of certain details, out of a total of 84 tasks, 45 contained some graphical representations. Among these, for tasks in Serbian, the AI Math Solver successfully solved 13 tasks (28.88%), while out of the remaining 39 tasks without images, it solved exactly 11 (28.20%), which shows an almost identical success rate.

When it comes to success in solving tasks in English, the identical percentage of solved tasks was observed in tasks that included a picture in the formulation and/or the provided answers (28.88%), as well as in tasks that did not include a picture (28.20%).

From Fig. 4, it can also be observed that there is a significantly higher number of tasks for which this AI tool failed to provide a solution when solving the problem required understanding the image and abstracting data from it. This highlights the limitations of AI Math Solver's capabilities, particularly in solving geometric problems.

The results obtained are below expectations. With only 24 out of a total of 84 tasks successfully solved in both Serbian and English, the performance cannot be considered successful. For illustration, if a 3rd or 4th-grade student gave the same answers as the AI Math Solver (in Serbian language), they would score 35.5 points out of a maximum of 120. This would place a 4th-grade student at approximately 2825th out of 3377 competitors, or a 3rd-grade student at about 2778th out of 3953 competitors who participated in the Kangaroo competition in March 2024. In summary, for both 3rd and 4th grades, the AI Math Solver's performance would position it roughly in the bottom third of all participants. A 7th or 8th-grade student with identical answers to the AI Math Solver in their category would score 56 points out of a maximum of 150, ranking them at around 462nd place among 1197 7th-grade students or 510th among 964 8th-grade students. In summary, for both 7th and 8th grades, the AI Math Solver's performance would place it roughly in the middle of the rankings.

As for the AI Math Solver's performance on tasks for 3rd and 4th-year high school students, its score in Serbian was 49.5 points out of a maximum of 150. Precise comparisons with student results would depend on the specific high school track or vocational program, so due to this complexity, we omit detailed analysis. Nevertheless, the AI Math Solver's performance is far below that of the top-performing students. For instance, one 3rd-year student achieved an impressive 145 points. Moreover, in the 4th-year category, two students attained the maximum possible score of 150 points.

We do not compare the results of the AI Math Solver achieved in English due to the small differences in the percentages of successfully solved tasks between the Serbian and English languages. However, in any case, the AI's performance is far from that of the top competitors in the Kangaroo competition in Serbia.

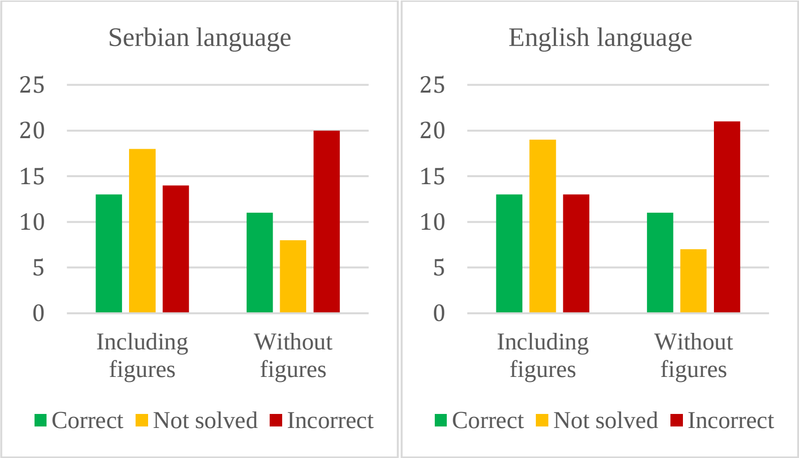


Fig. 4 Success in solving tasks in Serbian language (left) and in English language (right) depending on whether the formulation includes an image

In addition to these results, when testing tasks in AI Math Solver, several specific errors were observed, indicating challenges in linguistic and technical processing. For instance, the tool provided an English solution for a task written in Serbian, a phenomenon that can be linked to language confusion in LLMs (see, e.g., Marchisio et al., 2024). Additionally, some tasks showed a mix of Cyrillic and Latin alphabets, along with the insertion of words from other Slavic languages. For tasks containing table-like data, the tool frequently displayed an error message: *Unknown environment 'tabular'*. Furthermore, certain formulas were rendered in LaTeX source code instead of a standard readable format, etc.

6. Discussion

As previously mentioned, AI systems for solving mathematical tasks exhibit significant limitations in generalization and abstraction, particularly when problem-solving contexts demand multimodal reasoning and the integration of diverse skills (Cherian et al., 2023). The same limitations are evident in the performance of the AI Math Solver. Below, we analyze key reasoning modalities that contributed to its poor results.

One of the main reasons is the lack of adequate training datasets for non-standard mathematical tasks, which often feature unusual formulations or infrequent graphical representations, limiting AI models' ability to generalize and effectively solve such problems. For instance, Zhang et al. (2024) demonstrate that pre-trained models consistently outperform their non-pre-trained counterparts in text generation tasks within mathematical context, highlighting the importance of training data in achieving improved results.

When it comes to the concerning results in the performance of the AI Math Solver tool in solving simpler tasks that don't require advanced mathematical knowledge, the causes can be found in language specifics and contextual understanding. For example, when tasks are formulated for younger students, they are often designed to be engaging and motivating. While students, parents, and teachers understand the essence and requirements of the formulation, this can pose a challenge for the AI tool, which is trained to solve tasks with clearly structured text and requests. For an in-depth survey of challenges in understanding and solving mathematical word problems, see Sundaram et al. (2024). In addition, AI has a reduced ability to adapt to the new context, further reducing its performance. Unlike humans, who rely on intuition and past experiences when faced with new problems, artificial intelligence systems can perform poorly when faced with unconventional tasks. This lack of adaptability makes the AI less capable in non-standard scenarios.

For a comprehensive discussion of mathematical problems, their corresponding datasets, and the factors influencing LLMs in mathematical problem-solving, see Ahn et al. (2024).

Although images in tasks may make it easier for students, this may not be the case when the task is being solved by an AI tool. Graphical elements like color schemes, labels, and image quality can hinder AI systems in accurately interpreting math problems, with vibrant, playful images in lower-grade tasks and more complex visuals in higher grades adding to the challenge. These visual complexities make it difficult for AI to discern relationships between these geometric components. These observations align with findings from Yiu et al. (2024), which examine visual analogical reasoning in large multimodal models (LMMs). Their results reveal that while models like GPT-4V can identify simple visual attributes, they struggle with abstract reasoning, such as quantifying relationships or extrapolating rules to new contexts. These limitations stem from training data that primarily includes 2D images and text, reducing the model's ability to handle complex visual transformations or scenarios requiring a deeper understanding of the 3D physical world.

Non-standard tasks often feature formulations that are not very common, reducing the effectiveness of AI tools typically trained on tasks with conventional symbols and labels. These variations can pose additional challenges for AI tools in their operation. Furthermore, as mentioned, in Kangaroo competition tasks, some data is presented in text, while other data is provided in images (not necessarily images of geometric objects). This requires students to synthesize information presented in different forms and then combine it to solve the task. Such task formats demand a combination of language and image processing and, therefore, multimodal reasoning, which presents an additional difficulty for AI tools.

7. CONCLUSION

Considering that tasks in the Kangaroo competition are not considered as quite difficult, as students are expected to solve a large number of tasks within a relatively short time and the solutions do not involve complex or lengthy procedures, it was expected that the AI Math Solver would perform much better on these tasks. However, based on the results obtained, this cannot be confirmed; in fact, the results are significantly below expectations. Of course, one reason for these poor results could be that all tasks were uploaded as screenshots. It is possible that text-based tasks entered by copying the text and provided answers would yield better results, but this needs to be verified. The main conclusion could be that, due to the likely lack of training of the AI tool to solve non-standard mathematical tasks like those in the Kangaroo competition, the model requires additional training with tasks solved by students in previous years. Future research plans include analyzing the performance on tasks presented to the AI Math Solver via copied text and answer options, as well as testing a larger number of tasks to generalize the results. Additionally, we plan to test the performance of other widely used tools, such as ChatGPT, to compare the effectiveness of different tools and determine which one performs better in this context.

Acknowledgement: Authors are supported by the Ministry of Science, Technological Development and Innovation, Republic of Serbia. The first author acknowledges support under Contract No. 451-03-66/2024-03/200139, while the second, third and fourth author are supported under Contract No. 451-03-65/2024-03/200122.

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 Lamoneda, 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.-C, 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

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

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

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/feduc.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 ALATA U REŠAVANJU NESTANDARDNIH ZADATAKA SA MATEMATIČKIH TAKMIČENJA

Apstrakt. *Već neko vreme, istraživači širom sveta ispituju efekte korišćenja veštačke inteligencije u matematičkom obrazovanju kako bi učenicima pružili dodatnu podršku i pomoć. 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 takvo jeste i međunarodno takmičenje Kengur bez granica. U ovom radu analiziramo uspešnost AI Math Solver-a na platformi Interactive Mathematics u rešavanju zadataka sa takmičenja Kengur bez granica iz 2024. godine za učenike 3. i 4. razreda osnovne škole, kao i 7. i 8. razreda osnovne škole, i 3. i 4. razreda srednje škole. Zadaci su postavljeni u formi slika (screenshot-ova), na srpskom i engleskom jeziku, jer se u formulaciji zadataka i/ili ponuđenih odgovora na takmičenju Kengur bez granica često pojavljuju slike. 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*

University of Niš

Pedagogical Faculty in Vranje

1st International Scientific Conference

Education and Artificial Intelligence

BOOK OF PROCEEDINGS

Computer design

Darko Stojanović

Printed by

Plutos doo, Vranje

Printed in 70 copies

CIP - Каталогизација у публикацији
Народна библиотека Србије, Београд

37.091:004.8(082)(0.034.2)

INTERNATIONAL Scientific Conference Education and Artificial Intelligence (1 ; 2024 ; Vranje)
Book of proceedings / 1st International Scientific Conference Education and Artificial Intelligence (EDAI 2024), Vranje, November 29-30, 2024 ; [organizer Pedagogical Faculty in Vranje, University of Niš, Serbia] ; [co-organizers Faculty of Pedagogy, South-West University "Neofit Rilski", Blagoevgrad, Republic of Bulgaria ... [et al.]] ; [editors Aleksandar Spasić, Darko Stojanović]. - Vranje : University of Niš, Pedagogical Faculty, 2025 (Vranje : Plutos). - elektronski optički disk (CD-ROM) ; 12 cm

Sistemski zahtevi: Nisu navedeni. - Nasl. sa naslovne strane dokumenta. - Tiraž 70. - Bibliografija uz svaki rad. - Apstrakti.

ISBN 978-86-6301-060-4

а) Образовање -- Вештачка интелигенција -- Зборници

COBISS.SR-ID 168080393

ПЕДАГОШКИ ФАКУЛТЕТ · ВРАЊЕ ·