CHALLENGING DEEPSEEK-R1 WITH SERBIAN HIGH SCHOOL MATH COMPETITION PROBLEMS

Nemanja Vučićević,
[0000-0002-4903-7280]
Marina Svičević*,
[0000-0003-2791-3849]
Aleksandar Milenković
[0000-0001-6699-8772]

University od Kragujevac,
Faculty of Science,
Kragujevac, Srebia

Abstract:

This paper investigates the application of the large language model DeepSeek-R1 in solving problems from the Serbian National Mathematics Competition for high school students (2023/2024), with the aim of examining its ability to understand, reason, and generate accurate mathematical solutions. The analysis covers all grade levels and both competition categories (A and B), comprising a total of 36 problems across various mathematical domains, including algebra, geometry, combinatorics, number theory, and logic. All problems were presented to the model in their original textual form, using LaTeX syntax to ensure accurate representation of mathematical expressions.

The results obtained by the model were compared to official grading criteria and to the average scores achieved by student participants. In several cases, DeepSeek-R1 reached scores that would have qualified for official awards, especially in the higher grades of category B. The model showed stronger performance on algebraic problems and those with a more formal structure, while it encountered difficulties with logic-based problems and less standard formulations. Most errors were due to occasional misinterpretation of the problem statements or the omission of key reasoning steps.

This research provides a realistic assessment of the capabilities of a contemporary large language model in solving complex mathematical problems. It also highlights possible directions for its integration as a support tool in the teaching and learning of mathematics.

Keywords:

Non-standard problem solving, DeepSeek-R1, Large Language Models, Math competitions.

INTRODUCTION

Mathematics competitions have a long-standing tradition in the Serbian education system and represent one of the most important forms of nurturing giftedness, developing logical thinking, and encouraging creative problem solving among students. From the lower grades of primary school to the end of secondary education, students have the opportunity to compete in solving demanding problems that often go far beyond the standard curriculum. These competitions not only recognize and affirm talent but also prepare students for future academic and professional achievements in fields such as mathematics, computer science, physics, and engineering.

Correspondence:
Marina Svičević

e-mail:
marina.svicevic@pmf.kg.ac.rs

CC BY

DOI: 10.15308/Sinteza-2025-274-280

One of the most prestigious events in this system is the National Mathematics Competition for High School Students, which represents the final stage following the municipal and district competitions. Participation at this level is limited to the most successful students, which gives the event particular weight in both educational and societal contexts. The problems featured at this stage are known for their difficulty—not merely testing mastered mathematical techniques, but encouraging students to independently recognize connections, construct solution strategies, and apply non-standard ideas. This demand for a creative approach to problem solving makes the problems especially challenging and valuable for analysis. The competition is organized in accordance with the Official Rulebook on Mathematics Competitions for Secondary School Students [1], which defines participation criteria, problem structure, and ranking methodology. The competition includes four grade levels (I–IV) and two categories: Category A, for students from the Mathematical Grammar School, and Category B, for students from all other grammar schools. In both categories, the exam lasts four hours, but the format differs: Category A has four problems worth up to 25 points each, while Category B has five problems worth up to 20 points, both totaling 100 points. Problems are adapted to the students' grade level and category, and cover areas defined by the Competition Curriculum [2], including logic, algebra, geometry, combinatorics, and number theory.

The increasing capability of large language models (LLM) to process and solve tasks involving formal reasoning raises important questions about their applicability in educational contexts and their potential use as tools for supporting learning or assessment. In this light, it is particularly interesting to examine how a modern model such as DeepSeek-R1 [3], which is not exclusively specialized in mathematics but demonstrates strong performance on logical reasoning tasks, performs when faced with problems of high complexity.

The main objective of this paper is to examine how the DeepSeek-R1 model would perform on problems from the Serbian National Mathematics Competition for high school students in the 2023/2024 school year, across both categories—A and B. By analyzing the solutions generated by the model, the study aims to assess its ability to understand problem statements, plan solutions, and produce precise mathematical reasoning in the context of problems designed for the most successful high school students in Serbia.

Within this objective, the paper addresses the following research questions:

- How many points would DeepSeek-R1 score if evaluated according to the official competition criteria by a human grading committee?
- Where would the model rank compared to real student participants in the same year?
- What are the characteristics of the problems where the solutions are incorrect, in terms of the mathematical domains they belong to, the way problems are formulated, and how the data provided in the problem statements are interpreted?

In addition to providing an empirical insight into the capabilities of an advanced artificial intelligence model, this study also contributes to the broader discussion of how LLMs can be integrated into educational practice—as tools for practice, diagnosis, or inspiration in the teaching of mathematics.

2. BACKGROUND AND RELATED WORK

In recent years, we have witnessed significant progress in the development of LLMs, which are increasingly capable of solving tasks well beyond the scope of natural language processing. One area that has attracted particular attention in recent literature is the application of these models in mathematics, where solving textual and symbolic problems—requiring reasoning, formal precision, and solution planning—has become a realistic possibility.

One of the first notable models focused on mathematical reasoning was Minerva, developed by Google and introduced in 2022 [4]. Trained on technical and scientific literature, including large corpora from ArXiv, Minerva achieved impressive results on the MATH dataset, with its 62-billion-parameter version reaching the average score on the Polish national high school graduation exam. The model employs a chain-of-thought technique, which enables gradual solution development through step-by-step explanations. GPT-4 [5] currently represents one of the most powerful available solutions, capable of solving problems at the Olympiad level. According to OpenAI's technical report, GPT-4 scores within the top 10% on standardized tests such as the SAT, GRE, and other math-related exams. A key strength of GPT-4 is its ability to generate logically coherent solutions with detailed reasoning steps, which significantly improves accuracy compared to previous model generations. A more recent contribution to this

domain is DeepSeekMath [6], which continued training on a corpus of over 120 billion tokens of mathematical content. Although relatively small (with 7 billion parameters), the model achieved 51.7% accuracy on the MATH dataset, and the use of self-consistency techniques increased its performance to 60.9%. Additionally, it incorporates Group Relative Policy Optimization (GRPO) to improve reasoning efficiency with limited resources. A notable advancement in prompt design is the MathPrompter technique [7], which combines multiple independent solution strategies for self-verification. This approach demonstrated that the accuracy of GPT-3.5 and GPT-4 on the MultiArith dataset of elementary arithmetic problems can be improved from 78.7% to 92.5% by employing parallel methods such as algebraic solutions, Python code, and verbal explanations.

In addition to individual performance analyses, there is a growing body of research focused on direct comparisons of different LLMs in solving mathematical problems of varying difficulty. Such comparisons allow for a more precise assessment of each model's strengths and limitations in terms of accuracy, robustness, sensitivity to formulation changes, and domain-specific reasoning. Recent evaluations have placed special emphasis on the DeepSeek-R1 [8] [9], OpenAI o1 [10], and the most recent, improved version o3-mini [11], all of which represent state-of-the-art LLMs optimized for reasoning tasks. DeepSeek-R1, based on a Mixture-of-Experts architecture and trained using reinforcement learning with a focus on mathematical reasoning, achieves outstanding results on the MATH-500 benchmark, reaching an accuracy of 97.3%, thereby outperforming the OpenAI o1 model, which scores 96.4% on the same dataset. These results were obtained using the self-consistency technique, which has become a standard in evaluating mathematical reasoning in modern LLMs [8]. Beyond its high accuracy, DeepSeek-R1 also demonstrates greater robustness to lexical and structural reformulations of problem statements compared to the o1 model, whose performance tends to degrade under such variations. Analyses show that this robustness is particularly evident in areas such as algebra and number theory, while more complex spatial problems, such as geometry, remain a challenge for all models. A comparison on the AIME 2024 benchmark confirms this close performance: DeepSeek-R1 solves 79.8% of the problems, while OpenAI o1 solves 79.2%, with the smaller o1-mini model falling significantly behind. In more recent evaluations, the o3-mini model, released in early 2025, shows even better performance on categories such as Olympiad-level and scientific problems, surpassing both previous architectures on the GPQA-Diamond dataset and the most difficult AIME tasks [11]. On the GSM8K dataset, which consists of elementary-level arithmetic word problems, GPT-4 achieves an accuracy of 92.0% [12], while DeepSeek-R1, using a self-consistent configuration, reaches 96.3% [13]. These results lead to a high level of precision in basic arithmetic tasks. In summary, comparisons among the most advanced LLMs show that models such as DeepSeek-R1 and OpenAI o3-mini have reached a level of mathematical reasoning comparable to that of top-performing human competitors across various levels. The differences between them are increasingly reflected in their robustness, domain-specific precision, and adaptability to reformulated problems, which represent key directions for future development in this field.

The standard evaluation of LLMs is mostly based on automatically comparing the generated answer with the expected solution, where accuracy is measured as the percentage of correctly solved problems. Although the problems are typically presented in textual form, solving them requires multi-step logical reasoning, which has led to the adoption of more advanced evaluation methods such as self-consistency, partial credit scoring, and the assessment of intermediate steps. Chen et al. [14] demonstrated that GPT-4o [15] is capable of assigning partial credit to student-generated solutions with a high level of agreement with human graders (70–80%), using multiple evaluation passes and a detailed scoring rubric.

While modern models have achieved remarkable results on existing problem sets, there remains a need for further evaluation in new contexts—especially on problems that were not part of their training data. Such cases—like problems from national mathematics competitions—offer valuable insight into the true capabilities of these models and the possibilities for their application in educational and competitive settings, while also presenting an added challenge in evaluating the quality of generated solutions.

3. METHODOLOGY

This section of the paper outlines the steps taken to examine the ability of the large language model DeepSeek-R1 to solve mathematics problems from the Serbian National Mathematics Competition for high school students in the 2023/2024 school year. The choice of both the model and the problem set is based on two key criteria: on the one hand, DeepSeek-R1 is one of the most advanced open-source models specifically trained for mathematical reasoning; on the other hand, the competition problems represent an authentic and challenging dataset that spans a wide range of mathematical topics, formulation styles, and diverse solution approaches.

This setup makes it possible to test the model not only on well-known and frequently used datasets but also on real-world problems that were not part of its training data, offering valuable insight into its ability to generalize and apply reasoning in an educational context.

3.1. DATASET: SERBIAN NATIONAL MATHEMATICS COMPETITION 2023/2024

The dataset used in this study consists of original problems from the Serbian National Mathematics Competition for high school students, held during the 2023/2024 school year. The competition is organized by the Mathematical Society of Serbia, and the problems span a wide range of high school mathematics topics, including logic, algebra, geometry, combinatorics, number theory, and elements of analysis.

The competition is divided into four grade levels and two categories:

- Category A is intended for students from mathematical grammar schools,
- Category B is intended for students from all other grammar schools.

For each grade, Category A includes four problems, while Category B includes five. Problems in Category A are worth 25 points each, and those in Category B are worth 20 points each, giving a maximum of 100 points in both cases. In the 2024 competition, all problems were presented in textual form, without diagrams, illustrations, or graphical aids. The formulations typically require multiple steps and abstract thinking, making them well-suited for testing LLMs in the domain of mathematical reasoning. Since modern models support LaTeX syntax, the problem statements in this study were formatted accordingly, allowing for accurate representation of mathematical expressions and symbols. Each problem was presented to the model as an independent prompt in Serbian, without additional context, in order to simulate realistic solving conditions. The evaluation of the responses was performed manually by a current member of the official competition committee, based on the official solutions and the same criteria used during the actual competition. Special attention was given to the accuracy of the final answer, the logical coherence of the reasoning steps, and the identification of typical errors.

3.2. DEEPSEEK-R1

The DeepSeek-R1 model [16] is a next-generation large language architecture developed by DeepSeek AI with the goal of enhancing models' abilities to perform complex reasoning, particularly in the domains of mathematics and programming. Released as an open-source model in early 2025, DeepSeek-R1 is based on a Mixture-of-Experts (MoE) architecture, with a total capacity of 671 billion parameters, of which approximately 37 billion are active per query. This approach enables a better balance between response quality and execution efficiency.

Unlike some earlier LLMs that relied primarily on supervised learning, DeepSeek-R1 was trained using reinforcement learning, where the model was optimized directly based on the quality of its logical reasoning. Specifically, the training employed a “reasoning-in-the-loop” technique, allowing the model to learn from chains of thought and feedback on the correctness of its conclusions. The training process included a strong emphasis on the mathematical domain, further reinforced through fine-tuning on a large collection of problems, including datasets such as MATH, GSM8K, AIME, and OlympiadBench.

DeepSeek-R1 supports an input context of up to 128,000 tokens, making it well suited for handling long mathematical formulations and complex problem statements. In addition to the full model, the team has released several distilled versions, including R1-Zero, a variant without additional fine-tuning, which contributes to transparency in research settings.

The choice of DeepSeek-R1 for this study is based on several reasons. First, it is open source, which allows for direct access and reproducibility of results. Second, prior experiments have shown that DeepSeek-R1 performs on par with, and often surpasses, closed models such as

OpenAI o1, particularly on mathematical benchmarks. Third, its training is explicitly focused on reasoning, making it highly relevant for the types of tasks featured in mathematics competitions, where modeling multi-step solutions is more critical than simply generating a correct final answer.

4. RESULTS AND DISCUSSION

This section presents the results of applying the DeepSeek-R1 model to the problems from the Serbian National Mathematics Competition 2023/2024. The analysis covers a total of 36 problems, four in Category A and five in Category B for each of the four grade levels. In each case, the model was given only the textual formulation of the problem, presented in LaTeX format, without additional instructions, examples, or context, and consistent with the original tasks, which did not include any illustrations. The answers generated by the model were manually reviewed according to the official scoring criteria. The results obtained were compared with the average scores achieved by human participants by grade and category. Additionally, an analysis was conducted to determine which types of problems the model solved more or less successfully. Special attention was given to the qualitative analysis of errors and to the model's potential applications in educational contexts.

The first step in the evaluation involved comparing the total number of points that DeepSeek-R1 achieved per grade with the average number of points earned by students in the same categories at the 2023/2024 National Mathematics Competition.

The results are shown separately for Category A, which includes students from mathematical grammar schools, and Category B, which includes students from all other grammar schools (Figure 1). In both categories, the model scored below average in the first grade but generally outperformed the average results of students in the higher grades. Although there is no clear upward trend across all grades, the model shows consistent advantages in the second, third, and fourth grades, especially in the second grade of both categories. It is important to note that neither problem difficulty nor grade level alone necessarily explain the model's performance. According to the grading committee, problems for the first grade often include non-standard formulations, with an increased presence of combinatorial and logic-oriented tasks, which may have contributed to the model's lower performance at this level. Models like DeepSeek-R1 tend to be more effective at solving problems with a more formal structure, which aligns with the stronger results observed in grades dominated by algebraic and analytical problems. Overall, the diagram confirms that the model performs significantly better in Category B, where the problems are generally less complex.

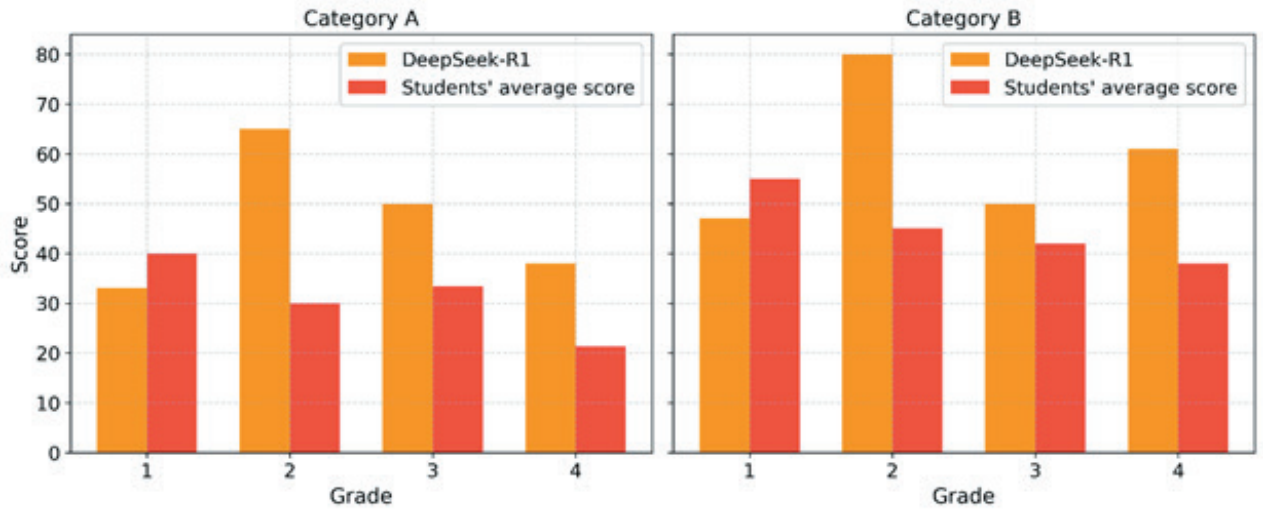

Figure 1. Comparison of DeepSeek-R1's results with the students' average scores achieved by students at the National Mathematics Competition 2023/2024, across both A and B categories

In addition to comparing DeepSeek-R1's total score with average values, it is also insightful to examine the model's placement on the official competition ranking lists by grade and category, including visual indicators of awarded prizes (Figure 2). In the first grade, the model did not manage to reach an award-winning position in either category — it ranked 33rd out of 50 in Category A and 31st out of 42 in Category B. However, from the second grade onward, the model achieved significantly better results. In Category A, it placed 4th out of 41 in the second grade, earning second prize, and maintained a top-ten placement in the third and fourth grades, earning third prizes. Even better results were achieved in Category B, where the model placed 3rd out of 67 in the second grade, earning first prize, and secured 13th and 7th place in the third and fourth grades, respectively, both corresponding to second prizes. These results indicate that the model is most successful in Category B from the second grade onward, which aligns with the relatively lower complexity of problems in that category. At the same time, the model's consistent placement in the top third of Category A in the higher grades demonstrates its capacity to handle more challenging problems as well.

The results show that the DeepSeek-R1 model, when evaluated according to official competition criteria, can achieve high rankings and even win awards in certain categories, especially in higher grades and with less demanding problems. Its errors most frequently occur in tasks with ambiguous formulations, logic-based reasoning, or geometry problems that require constructive or proof-based solutions, indicating limitations in the model's ability to comprehend abstract problems and carry out multi-step reasoning. These insights confirm that LLMs have the potential to serve as valuable educational tools in mathematics instruction — for independent practice, reasoning diagnostics, and encouraging diverse problem-solving approaches.

5. CONCLUSION

This study has shown that the large language model DeepSeek-R1 is capable of successfully solving high-complexity mathematical problems, such as those featured in the National Mathematics Competition for Serbian high school students. Through systematic evaluation, it was found that the model not only achieves results comparable to top-performing contestants in certain categories but even reaches rankings that would earn it official awards in specific grades and categories. Its highest effectiveness was observed in algebraic and analytical problems, while weaker performance was noted in logic-based tasks and less clearly formulated problems.

These findings suggest that LLMs like DeepSeek-R1 have already reached a level that makes them relevant in educational contexts—not as replacements for students or teachers, but as complementary tools for practice, automated assessment, and the encouragement of creative problem-solving. The unique value of such analyses lies

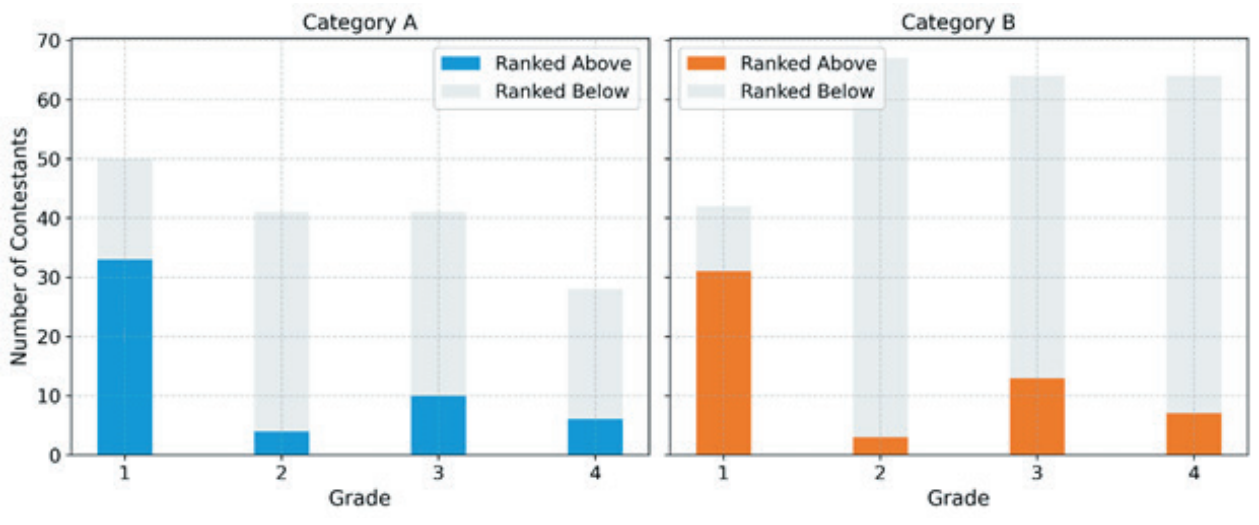

Figure 2. DeepSeek-R1's relative ranking among contestants at the National Mathematics Competition 2023/2024, across both Category A and Category B. Each bar shows the number of students ranked above the model (colored) and those ranked below (light gray)

in the fact that competition problems represent challenging, real-world tasks that are not part of the models' training datasets, thus offering a more objective view of their capabilities and limitations.

Future research could involve comparative evaluations of multiple models on the same problem set, as well as an exploration of how LLMs might be integrated into mathematics education through interactive platforms and tailored feedback mechanisms.

6. ACKNOWLEDGEMENTS

This work was supported by the Ministry of Science, Technological Development and Innovation of the Republic of Serbia through the Agreement No. 451-03-65/2024-03/200122.

REFERENCES

[1] Društvo matematičara Srbije, "Pravilnik o takmičenjima," [Online]. Available: https://dms.rs/wp-content/uploads/2021/12/Pravilnik_o_takmicenjima_SS_matematika.pdf.

[2] Društvo matematičara Srbije, "Program takmičenja," [Online]. Available: https://dms.rs/wp-content/uploads/2016/12/program_2016.pdf.

[3] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang and X. Bi, "Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[4] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag and T. Gutman-Solo, "Solving quantitative reasoning problems with language models," *Advances in Neural Information Processing Systems*, vol. 35, p. 3843–3857, 2022.

[5] OpenAI, "GPT-4," [Online]. Available: https://openai.com/index/gpt-4-research/. [Accessed 10 March 2025].

[6] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li and Y. Wu, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[7] S. Imani, L. Du and H. Shrivastava, "Mathprompter: Mathematical reasoning using large language models," *arXiv preprint arXiv:2303.05398*, 2023.

[8] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang and X. Bi, "Deepseek-r1: Incentivizing reasoning capability in LLMs via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[9] DeepSeek. [Online]. Available: https://www.deepseek.com/. [Accessed 10 March 2025].

[10] OpenAI, "o1," [Online]. Available: https://openai.com/index/openai-o1-system-card/. [Accessed 10 March 2025].

[11] OpenAI, "o3-mini," [Online]. Available: https://openai.com/index/openai-o3-mini/. [Accessed 10 March 2025].

[12] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman and S. Anadkat, "GPT-4 Technical Report," *arXiv preprint arXiv:2303.08774*, 2023.

[13] A. Eassa, A. Shah, H. Mao, H. Lu, E. Ho, J. Xin and O. Almog, "NVIDIA," [Online]. Available: https://developer.nvidia.com/blog/nvidia-blackwell-delivers-world-record-deepseek-r1-inference-performance/. [Accessed 20 March 2025].

[14] Z. Chen and T. Wan, "Using Large Language Models to Assign Partial Credits to Students' Problem-Solving Process: Grade at Human Level Accuracy with Grading Confidence Index and Personalized Student-facing Feedback," *arXiv preprint arXiv:2412.06910*, 2024.

[15] OpenAI, "GPT-4o," 13 May 2024. [Online]. Available: https://openai.com/index/gpt-4o-system-card/. [Accessed 10 March 2025].

[16] DeepSeek-R1, 2025. [Online]. Available: https://huggingface.co/deepseek-ai/DeepSeek-R1.