# QLORA FINE-TUNING OF MISTRAL-7B FOR SERBIAN HIGH SCHOOL MATHEMATICS COMPETITION TASKS

Miloš Pavković[1]*,
[0000-0001-7776-6045]

Marina Svičević[2],
[0000-0003-2791-3849]

Aleksandar Milutinović[2],
[0009-0004-1300-3232]

Nemanja Vučićević[2],
[0000-0002-4903-7280]

Aleksandar Milenković[2]
[0000-0001-6699-8772]

[1]Singidunum University,
Belgrade, Serbia

[2]University of Kragujevac,
Faculty of Science,
Kragujevac, Serbia

Abstract:

This paper examines the extent to which QLoRA fine-tuning can improve the performance of the large language model Mistral-7B-Instruct-v0.3 on Serbian high school mathematics competition tasks. Based on a dataset of tasks in Serbian, a fine-tuned model, Math-SRB-Mistral-7B, was developed and compared with the base model. The responses were evaluated using Claude 3.7 Sonnet as a judge, according to multiple criteria, including final answer accuracy, logical coherence, explanation quality, and an aggregate score. The results suggest that the applied fine-tuning did not lead to improved performance; instead, the fine-tuned model achieved slightly lower scores across all evaluated dimensions. This finding suggests that parameter-efficient adaptation of general-purpose LLMs on small and challenging mathematical datasets does not necessarily result in better generalization to new tasks. At the same time, the results highlight the importance of multi-criteria evaluation in the analysis of mathematical reasoning generated by LLMs.

Keywords:

Large Language Models, Qlora Fine-Tuning, Mathematics Competitions, LLM-As-A-Judge, Multi-Criteria Evaluation.

## INTRODUCTION

Mathematics competitions occupy an important place in the Serbian educational system and provide a valuable context for the development of logical thinking, creative problem-solving, and mathematical precision. Unlike standard school tasks, competition tasks often require multi-step reasoning, careful interpretation of conditions, the integration of different concepts, and clearly articulated explanations. For this reason, they provide a suitable framework for examining the actual capabilities of modern large language models (LLMs) in the domain of mathematical reasoning ( [1], [2]).

In recent years, LLMs have shown substantial progress on tasks that require more complex forms of reasoning, including mathematical problems. However, previous studies have shown that their performance depends on several factors, such as the type of task, the mathematical area, the formulation of the problem, the language in which the task is presented, and the way generated solutions are evaluated [1], [3].

Correspondence:
Miloš Pavković

e-mail:
mpavkovic@singidunum.ac.rs



DOI: 10.15308/Sinteza-2026-129-136

Results obtained on standard benchmark datasets do not always provide a sufficiently reliable picture of model behavior when solving tasks that were not part of the usual training sets and that require not only a correct final answer, but also a consistent, understandable, and mathematically justified explanation.

Additional motivation for this line of research is provided by studies that have examined the use of AI tools and LLMs on mathematics competition tasks in the Serbian-speaking context. An earlier study devoted to the Mathematical Kangaroo Competition showed that model performance depends both on the task format and on the mathematical area, with text-based tasks and certain areas, such as algebra and number theory, being more favourable for models than geometry and logic [2]. Similarly, a study focused on tasks from the national mathematics competition in Serbia showed that even advanced models can achieve notable results, but still encounter difficulties with less standard formulations, logic-based tasks, and the consistent execution of all steps in a solution [4], [5]. These findings indicate that tasks from national mathematics competitions represent a demanding and relevant test domain for further examining the capabilities of LLM models.

Beyond model performance itself, the way their responses are evaluated is also an important issue. In previous studies, solutions were often assessed manually, following the approach used in grading students' work according to the official competition criteria [4], [5]. Although such an approach has high expert value, it is time-consuming and difficult to scale when comparing a larger number of experiments, model variants, or different configurations. At the same time, more recent studies have shown that LLMs can also be used to support the evaluation of open-ended student responses and partial scoring, with a satisfactory level of agreement with human evaluators [6]. This creates room for the development of structured automatic frameworks for assessing the quality of mathematical solutions, in which not only final accuracy is considered, but also the logical coherence of the solution process and the quality of the explanation.

In this context, a particularly important question is whether a general-purpose model that was not originally specialized for mathematical reasoning can be successfully adapted to the narrower domain of competition tasks. Parameter-efficient approaches, such as LoRA and QLoRA, make such adaptation possible while requiring significantly fewer computational resources than full fine-tuning, which makes them particularly suitable for academic and research settings [7], [8]. Starting from this, the aim of this paper is to examine whether and to what extent QLoRA fine-tuning can improve the performance of the Mistral-7B-Instruct-v0.3 model [9] on tasks from competitions organized by the Serbian Mathematical Society, using a multi-criteria automatic evaluation of generated solutions. In the remainder of the paper, the fine-tuned variant of this model is referred to as **Math-SRB-Mistral-7B**.

In line with this objective, the paper addresses the following research questions:

1. To what extent does QLoRA fine-tuning improve the performance of the Mistral-7B-Instruct-v0.3 model on tasks from competitions organized by the Serbian Mathematical Society?
2. Does the observed effect appear equally across different evaluation aspects, such as final answer accuracy, logical coherence, and explanation quality?

## 2. BACKGROUND AND RELATED WORK

Fine-tuning of LLMs for mathematical reasoning has become an active research area in which several important approaches can be identified. One line of research focuses on the construction and augmentation of training datasets. Yu et al. [10] propose MetaMath, a model fine-tuned on the MetaMathQA dataset, obtained by reformulating questions from the GSM8K and MATH datasets from multiple perspectives. The authors show that this approach leads to a substantial improvement over earlier open-source models of the same size, with MetaMath-7B achieving 66.5% on GSM8K and 19.8% on MATH, which highlights the importance of data quality and diversity for mathematical fine-tuning. A similar direction is followed by Luo et al. [11] with the WizardMath model, which applies Reinforcement Learning from Evol-Instruct Feedback (RLEIF) to improve mathematical chain-of-thought reasoning. In this setting, WizardMath-Mistral-7B achieves 90.7% on GSM8K and 55.4% on MATH, showing that the combination of evolved instructions and a feedback-driven approach can significantly improve the mathematical capabilities of such models.

Another line of research focuses on specialized pretraining on mathematically relevant corpora. Shao et al. [12] introduce DeepSeekMath-7B, a model further pretrained on approximately 120 billion mathematically relevant tokens, achieving 51.7% on the MATH dataset and 60.9% when combined with a self-consistency approach.

Azerbayev et al. [13] propose Llemma, a model obtained through continued pretraining of Code Llama on the Proof-Pile-2 corpus, which includes scientific papers, web content containing mathematical notation, and formal mathematical and programming records. In contrast, MAmmoTH [14] shows that strong results can also be achieved through an instruction-tuning approach that combines natural language explanations with program-oriented rationales. A common characteristic of these studies is that they rely on large, predominantly English-language datasets, such as GSM8K and MATH, or on very large mathematical corpora. In this sense, the present study considers a considerably more demanding practical scenario: QLoRA fine-tuning on a small, domain-specific dataset of mathematics competition tasks from Serbia.

When it comes to evaluation, most of the studies discussed above rely primarily on the automatic comparison of the final answer with a reference solution, which does not provide a complete view of the quality of the solution process itself. In the context of the CHAMP dataset, Mao et al. [15] emphasize the need for a fine-grained analysis of mathematical reasoning on competition-level tasks, precisely because final accuracy alone is insufficient for understanding the actual capabilities of LLMs. Stephan et al. [16] show that LLMs acting as judges on mathematical reasoning tasks can reliably distinguish the stronger model, but also that their decisions may be influenced by stylistic characteristics of the responses. Chen and Wan [6] further show that LLMs can successfully support partial scoring and the evaluation of solution explanations, with agreement rates of approximately 70–80% relative to human evaluators. Building on these findings, the evaluation framework used in this paper does not reduce assessment to final answer accuracy alone, but separates it into three criteria – Answer Accuracy, Logical Coherence, and Explanation Quality – thus providing a more detailed view of the effects of fine-tuning.

# 3. METHODOLOGY

## 3.1. DATASET AND EXPERIMENTAL SPLIT

The dataset used in this study consists of 348 tasks from mathematics competitions organized by the Serbian Mathematical Society for high school students, collected from the 2022, 2023, and 2024 competition years. The tasks come from three competition levels – municipal, regional, and national – and cover all four years of high school, in categories A and B.

According to their mathematical content, the tasks are classified into four areas: algebra, geometry, number theory, and combinatorics.

The original tasks and their official solutions were available in the form of LaTeX documents. For the purposes of this study, a specialized parser was developed to extract the task statements and corresponding solutions from the LaTeX sources, remove visual and formatting commands without semantic significance, and produce structured JSON records. Each record contains a unique identifier, the relevant metadata (mathematical area, grade, competition level, and category), the task statement, and the reference solution. Tasks containing graphical elements or images were excluded from the dataset, as they require multimodal models, which are beyond the scope of this study.

The dataset was split chronologically into training, validation, and test sets in order to ensure that the model had no access during training to tasks from the year used for final evaluation. The training set contains 192 tasks, the validation set 40, and the test set 116. The entire set of tasks from 2024 was reserved as a hold-out test set, while the validation set was formed from regional-level tasks from 2023, so that validation would be based on a competition level whose difficulty and structure naturally lie between the municipal and national levels. The distribution of tasks across the three sets is shown in Table 1.

The base model selected in this study was Mistral-7B-Instruct-v0.3, a general-purpose LLM with 7 billion parameters. Since this model was not originally specialized for mathematical reasoning, its use makes it possible to directly assess the effect of domain-specific adaptation

**Table 1.** Dataset split

| Set | N | Source |
|---|---|---|
| Training | 192 | Entire 2022 set (116) + 2023 excluding the regional level (76) |
| Validation | 40 | Regional competition 2023 |
| Test | 116 | Entire 2024 set (hold-out) |

on mathematics competition tasks. In this way, any observed improvement in performance can largely be attributed to the applied fine-tuning procedure rather than to prior mathematical specialization of the model.

To adapt the base model, the QLoRA method was applied, combining 4-bit NF4 quantization with low-rank LoRA adapters. This approach enables parameter-efficient fine-tuning with substantially lower memory and computational requirements than full fine-tuning, while still maintaining competitive model performance. In the experimental setup, the following hyperparameters were used: LoRA rank r = 16, LoRA alpha α = 32, 4-bit NF4 quantization, 5 epochs, an initial learning rate of $2 \times 10^{-4}$, a cosine scheduler, a per-device batch size of 2, and gradient accumulation of 16, resulting in an effective batch size of 32. The selected configuration was established through preliminary testing and reflects settings commonly used for small, domain-specific datasets; a systematic ablation study is identified as a direction for future work. The configuration of the hyperparameters used is shown in Table 2. An early stopping mechanism with a patience of two consecutive evaluations without improvement in validation loss was also applied. The LoRA adapters are stored separately from the base model, which allows a direct comparison between the base and fine-tuned variants on the same test set.

### 3.2. EXPERIMENTAL PIPELINE

The entire experimental process was organized as a modular pipeline that includes data preparation, model fine-tuning, solution generation, and automatic evaluation. The architecture of the experimental pipeline is shown in Figure 1. After parsing and preprocessing the original LaTeX documents, a structured dataset in JSON format is created, which serves as the basis both for QLoRA fine-tuning and for subsequent inference and evaluation.

QLoRA fine-tuning of the base model Mistral-7B-Instruct-v0.3 is performed on the training set of 192 tasks, resulting in the fine-tuned variant Math-SRB-Mistral-7B. After that, solutions for all 116 tasks in the test set are generated using both the base model and the fine-tuned model under identical inference conditions. Both sets of generated responses are stored in a structured format for subsequent processing and evaluation.

Automatic evaluation is performed using the LLM judge model Claude 3.7 Sonnet, based on three independent criteria: Answer Accuracy (A), Logical Coherence (B), and Explanation Quality (C). The first criterion evaluates the correctness of the final answer, the second assesses the logical validity and consistency of the solution process, and the third evaluates the clarity, completeness, and overall quality of the explanation.

**Table 2.** Configuration of QLoRA fine-tuning

| Parameter | Value |
|---|---|
| LoRA rank (r) | 16 |
| LoRA alpha (α) | 32 |
| Quantization | 4-bit NF4 |
| Number of epochs | 5 |
| Learning rate | $2 \times 10^{-4}$ |
| LR scheduler | Cosine |
| Batch size (per device) | 2 |
| Gradient accumulation steps | 16 |
| Effective batch size | 32 |
| Early stopping patience | 2 |

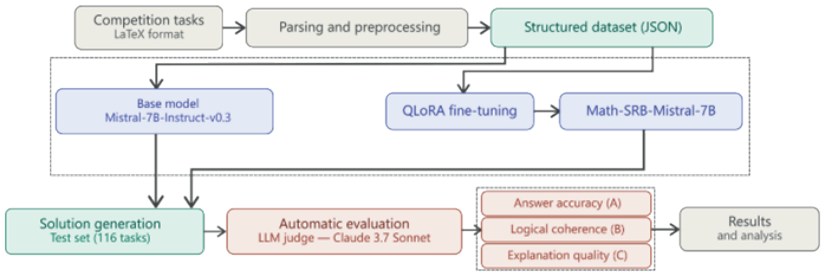


**Figure 1.** Experimental pipeline for fine-tuning and evaluating the Math-SRB-Mistral-7B model on Serbian mathematics competition tasks

For each task, the generated solution, together with the task statement and the reference solution, is submitted to the judge separately for each of these criteria, yielding a multi-criteria assessment suitable for further analysis of the performance of the base and fine-tuned models. Claude 3.7 Sonnet was selected as the evaluation judge because of its multilingual capabilities, including Serbian, and its reported reliability in assessing mathematical explanations and reasoning tasks [6], [16]. Since all tasks were intentionally kept in their original Serbian form, the study focuses on evaluating model performance in the native task language rather than through a translation-based baseline. The use of a single LLM judge enabled a consistent automatic evaluation framework, while future work will further strengthen reliability through human adjudication, cross-judge validation, and comparisons with math-specialized models and translation-based approaches. The aggregate score is computed as a weighted average according to the following formula

$$Score = 0.40 \times A + 0.30 \times B + 0.30 \times C$$

**Equation 1**. Weighted aggregate score used for model evaluation.

where the higher weight assigned to Answer Accuracy reflects the fact that, in the competition context, the correctness of the final answer remains a key component of the overall evaluation, alongside logical coherence and explanation quality.

## 4. RESULTS AND DISCUSSION

The evaluation results on the hold-out test set, which comprises 116 tasks from 2024, show that the applied QLoRA fine-tuning did not improve the performance of the base model, Mistral-7B-Instruct-v0.3. Instead, the fine-tuned variant, Math-SRB-Mistral-7B, achieved lower scores across all evaluated criteria. The base model obtained 8.5% on final answer accuracy, 21.6% on logical coherence, 28.1% on explanation quality, and 18.3% on the aggregate score, whereas the fine-tuned model achieved 6.5%, 19.3%, 26.6%, and 16.4%, respectively (Figure 2). This finding suggests that domain adaptation did not enhance model performance in this setting, but instead led to a slight yet consistent decline in overall performance.

With regard to the first research question, QLoRA fine-tuning did not improve the performance of the base model in this setting. Parameter-efficient fine-tuning of a general-purpose LLM on a small and heterogeneous set of tasks does not necessarily lead to better generalization – a finding that highlights the limitations of such an approach in the context of complex mathematical tasks. The stratified results in Table 3 confirm that the performance decline is consistent across all mathematical areas; breakdown by grade, competition level, and deterministic matching for numeric answers are identified as further directions.

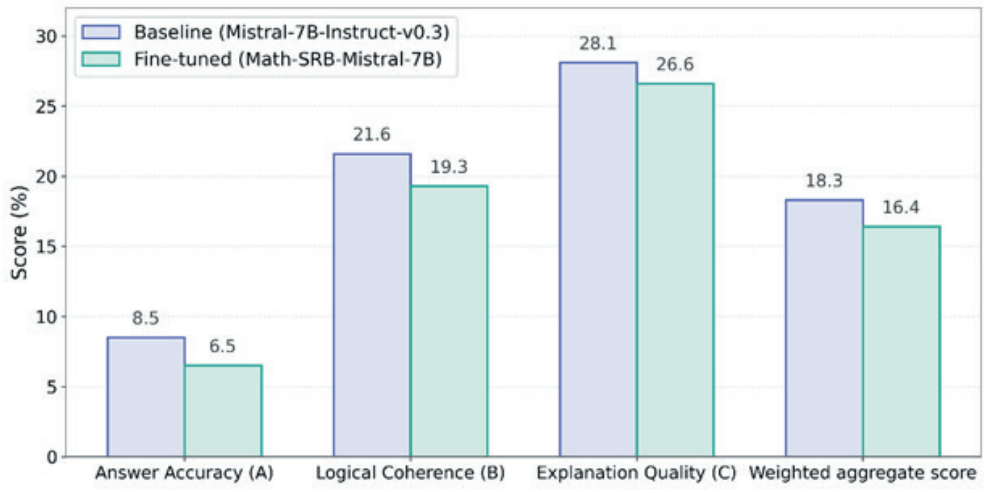


**Figure 2.** Comparison of baseline and fine-tuned model performance across the evaluation criteria

**Table 3.** Weighted aggregate score by mathematical area (test set, N=116)

| Model | Algebra | Geometry | Combinatorics | Number Theory | Average |
|---|---|---|---|---|---|
| Mistral-7B-Instruct | 18.2% | 19.8% | 15.8% | 19.4% | 18.3% |
| Math-SRB-Mistral-7B | 16.5% | 18.4% | 14.4% | 16.3% | 16.4% |

It should be noted that the observed differences are small in absolute terms; no statistical significance testing was performed, as the study is exploratory in nature. For broader context, a companion study comparing multiple locally executable models – including Mathstral 7B, DeepSeek Math 7B, and Llemma 7B – on the same Serbian competition corpus shows that all remain below 20% on the weighted aggregate score [17], confirming that the difficulty of the domain is not specific to Mistral-7B-Instruct.

The answer to the second research question is also clear. The performance decline is consistent across all evaluation aspects: the fine-tuned model achieved lower scores in logical coherence and explanation quality as well, indicating that additional training did not improve any of the key dimensions of mathematical reasoning examined here. For both models, the highest scores were recorded for explanation quality, followed by logical coherence, while final answer accuracy remained the lowest, suggesting that models are more likely to produce well-structured responses than mathematically correct ones.

Additional insight into the fine-tuning process is provided in Figure 3. The training loss decreases steadily, while the evaluation loss initially declines to around 0.940 by epoch 2.5, after which it increases again. Similarly, evaluation mean token accuracy rises to approximately 0.771 and then decreases, while final training accuracy reaches 0.804. This dynamic suggests that the model captured patterns in the training and validation sets but failed to generalize to the test set – a pattern consistent with the small, domain-specific, and highly challenging nature of the dataset.

One possible explanation for these results lies in the structure of the experiment itself. The training set comprised 192 tasks, the validation set included 40 tasks, while the entire 2024 competition year was reserved as the final test set. This chronological split is methodologically justified, as it ensures a strict separation between training and final evaluation, but it is also highly demanding. The limited amount of training data may not have been sufficient for fine-tuning to produce robust gains and may instead have led to partial adaptation to a narrow set of examples without broader improvement in the model's mathematical reasoning ability.

The nature of competition tasks may also have contributed to this outcome. Unlike standard benchmark datasets, mathematics competition tasks often require careful reading of the problem statement, a more creative choice of solution strategy, multi-step reasoning, and mathematically rigorous justification of each step. For this reason, additional training on a limited number of examples may not have been sufficient to improve precisely those abilities that are crucial for this type of task. The results further suggest that domain adaptation success depends on training set size, diversity, and domain characteristics, and that the model's limited exposure to Serbian-language input may have been an additional contributing factor.

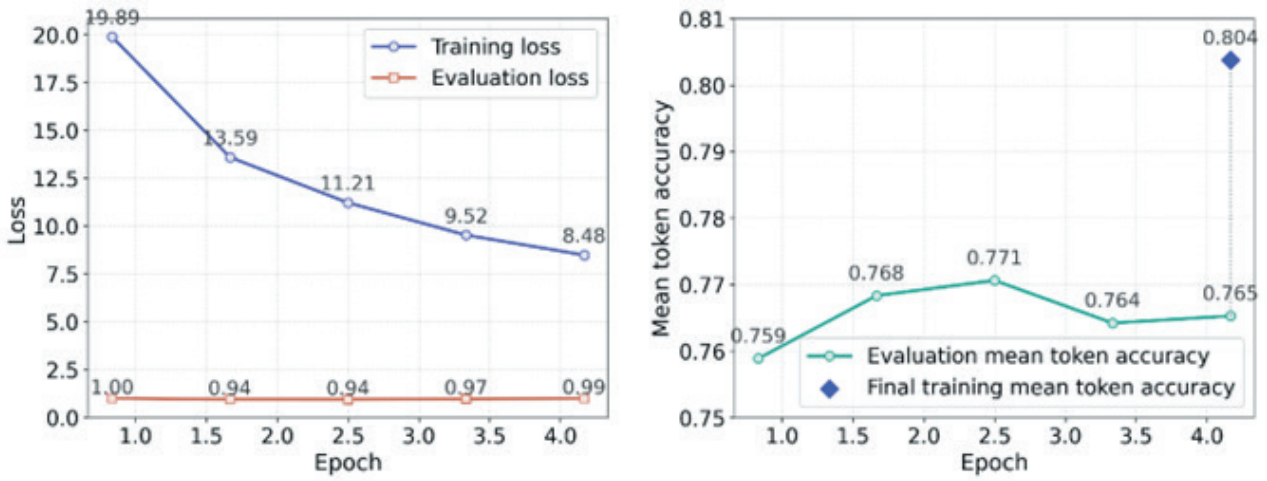


**Figure 3.** Training dynamics of the fine-tuned model in terms of loss and mean token accuracy

## 5. CONCLUSION

The results of this study show that QLoRA fine-tuning of the Mistral-7B-Instruct-v0.3 model on a domain-specific dataset of Serbian high school mathematics competition tasks did not lead to improved performance on the held-out test set. Instead of the expected improvement, the fine-tuned model, Math-SRB-Mistral-7B, achieved slightly lower scores across all evaluated dimensions, including final answer accuracy, logical coherence, explanation quality, and the aggregate score. Although the anticipated improvement was not confirmed, this finding is still important, as it indicates that parameter-efficient adaptation of general-purpose LLMs to small and challenging mathematical datasets does not necessarily result in better generalization to unseen tasks. In addition, the results confirm that multi-criteria evaluation provides an important framework for analyzing LLM-generated responses, as it enables the simultaneous consideration of mathematical accuracy, logical soundness, and explanation quality. Future research may therefore focus on expanding and diversifying the training set, as well as on fine-tuning models that are already more strongly specialized for mathematical reasoning and better adapted to Serbian-language input. Systematic ablation studies over prompt formats, adapter rank, dropout, and learning rate sensitivity are also identified as directions for future work.

## 6. ACKNOWLEDGEMENTS

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

## REFERENCES

[1] J. Ahn, R. Verma, R. Lou, D. Liu, R. Zhang and W. Yin, "Large Language Models for Mathematical Reasoning: Progresses and Challenges," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, St. Julian's, Malta, 2024. doi:10.18653/v1/2024.eacl-srw.17.

[2] M. Svičević, A. Milenković, N. Vučićević and M. Stanković, "Evaluating the Success of AI Tools in Supporting Student Performance in Mathematical Kangaroo Competition," *Computer Applications in Engineering Education*, vol. 33, no. e70063, 2025. doi:10.1002/cae.70063.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le and D. Zhou, "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in neural information processing systems*, vol. 35, pp. 24824-24837, 2022.

[4] N. Vučićević, M. Svičević and A. Milenković, "Challenging Deepseek-R1 with Serbian High School Math Competition Problems," in *Sinteza 2025 - International Scientific Conference on Information Technology, Computer Science, and Data Science*, Belgrade, Singidunum University, Serbia, 2025, pp. 274-280. doi:10.15308/Sinteza-2025-274-280.

[5] A. Milenković, N. Vučićević and M. Svičević, "Evaluating open ai tools o1 and o3-mini in solving high school problems from serbian national mathematics competition," *Teaching of Mathematics*, vol. 28, no. 1, 2025. doi:10.57016/TM-EYRQ1676.

[6] Z. Chen and T. Wan, "Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy," *Physical Review Physics Education Research*, vol. 21, no. 1, p. 010126, 2025. doi:10.1103/PhysRevPhysEducRes.21.010126.

[7] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in *International Conference on Learning Representations*, 2022.

[8] T. Dettmers, A. Pagnoni, A. Holtzman and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," *Advances in neural information processing systems*, vol. 36, pp. 10088-10115, 2023.

[9] Mistral AI, "Mistral-7B-Instruct-v0.3," 2024. [Online]. Available: https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3. [Accessed 1 April 2026].

[10] L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, J. T. Kwok, Z. Li, A. Weller and W. Liu, "MetaMath: Bootstrap your own mathematical questions for large language models," in *The Twelfth International Conference on Learning Representations*, 2024.

[11] H. Luo, Q. Sun, C. Xu, P. Zhao, J. Lou, C. Tao, X. Geng, Q. Lin, S. Chen, Y. Tang and D. Zhang, "WizardMath: Empowering mathematical reasoning for large language models via reinforced evol-instruct," in *The Thirteenth International Conference on Learning Representations*, 2025.

[12] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. K. Li, Y. Wu and D. Guo, "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models," *arXiv*:2402.03300, 2024.

[13] Z. Azerbayev, H. Schoelkopf, K. Paster, M. Dos Santos, S. McAleer, A. Q. Jiang, J. Deng, S. Biderman and S. Welleck, "Llemma: An open language model for mathematics," *arXiv*:2310.10631, 2023.

[14] X. Yue, X. Qu, G. Zhang, Y. Fu, W. Huang, H. Sun, Y. Su and W. Chen, "MAmmoTH: Building math generalist models through hybrid instruction tuning," *arXiv*:2309.05653, 2023.

[15] Y. Mao, Y. Kim and Y. Zhou, "CHAMP: A competition-level dataset for fine-grained analyses of llms' mathematical reasoning capabilities," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 13256-13274. doi:10.18653/v1/2024.findings-acl.785.

[16] A. Stephan, D. Zhu, M. Aßenmacher, X. Shen and B. Roth, "From calculation to adjudication: Examining llm judges on mathematical reasoning tasks," in *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, 2025

[17] M. Svičević, A. Milutinović, N. Vučićević, A. Milenković and M. Pavković, "Methodological framework for automated solving of mathematics competition problems in Serbia using LLMs," in *DeepTech Connect 2026*, 2026. (in press)