XII Triennial International Scientific Conference

Engineering TODAY

2026

CONFERENCE PROCEEDINGS

25 - 27 JUNE 2026, VRNJAČKA BANJA, SERBIA

FACULTY OF MECHANICAL AND CIVIL ENGINEERING IN KRALJEVO
UNIVERSITY OF KRAGUJEVAC

XII INTERNATIONAL TRIENNIAL SCIENTIFIC CONFERENCE

ENGINEERING TODAY
ET 2026

Conference Proceedings

25 – 27 June 2026, Vrnjačka Banja, Serbia

Supported by:

Republic of Serbia
Ministry of Science, Technological Development and Innovation

PUBLISHER: Faculty of Mechanical and Civil Engineering in Kraljevo, University of Kragujevac

EDITOR: Prof. Dr Goran Marković

TECHNICAL EDITOR AND COVER DESIGN:
Prof. Dr Mišo Bjelić

REVIEWS: All papers have been reviewed by the members of the Scientific Committee

PRINT: SaTCIP d.o.o., Vrnjačka Banja

PLACE AND YEAR OF PUBLICATION:
Kraljevo, 2026

CIRCULATION: 60 copies

ISBN 978-86-82434-15-3

CIP - Каталогизација у публикацији Народна библиотека Србије, Београд

621(082)
629.3/.4(082)
621.86/.87(082)
681.5(082)
004(048)

TRIENNIAL International Scientific Conference Engineering Today (12 ; 2026 ; Vrnjačka Banja)

Conference Proceedings / XII International Triennial Scientific Conference Engineering Today 2026, ET 2026, 25 – 27 June 2026, Vrnjačka Banja, Serbia ; [editor Goran Marković]. - Kraljevo : University of Kragujevac, Faculty of Mechanical and Civil Engineering, 2026 (Vrnjačka Banja : SaTCIP). - 1 knj. (razl. pag.) : ilustr. ; 30 cm

Na nasl. str.: Republic of Serbia, Ministry of Science, Technological Development and Innovation. - Tiraž 60. - Str. 7: Preface / Goran Marković. - Bibliografija uz svaki rad. - Registar.

ISBN 978-86-82434-15-3

a) Производно машинство -- Зборници
b) Транспортна средства -- Зборници
v) Аутоматско управљање -- Зборници
g) Информациона технологија -- Зборници

COBISS.SR-ID 201872905

Prof. Dr A. Milašinović, FME Banja Luka, BiH
Prof. Dr I. Milićević, FTS Čačak, Serbia
Prof. Dr V. Milićević, FMCE Kraljevo, Serbia
Prof. Dr D. Milković, FME Belgrade, Serbia
Prof. Dr Z. Miljković, FME Belgrade, Serbia
Prof. Dr B. Milošević, FMCE Kraljevo, Serbia
Prof. Dr V. Milovanović, FE Kragujevac, Serbia
Prof. Dr G. Minak, UNIBO, Italy
Prof. Dr D. Minić, FME Kosovska Mitrovica, Serbia
Prof. Dr A. Nikolić, FMCE Kraljevo, Serbia
Prof. Dr E. Nikolov, TU Sofia, Bulgaria
Prof. Dr V. Nikolić, FME Niš, Serbia
Prof. Dr V. Nikolov, VTU Sofia, Bulgaria
Prof. Dr M. Ognjanović, FME Belgrade, Serbia
Prof. Dr J. Peterka, FMS&T Trnava, Slovakia
Prof. Dr D. Petrović, FMCE Kraljevo, Serbia
Prof. Dr J. Polajnar, BCU Prince George, Canada
Prof. Dr M. Popović, FTS Čačak, Serbia
Prof. Dr D. Pršić, FMCE Kraljevo, Serbia
Prof. Dr Francu Catalin, UTCB, Romania
Prof. Dr N. Radić, FME East Sarajevo, BiH
Prof. Dr B. Radičević, FMCE Kraljevo, Serbia
Prof. Dr V. Radonjanin, FTS Novi Sad, Serbia
Prof. Dr M. Savković, FMCE Kraljevo, Serbia
Prof. Dr D. Sever, FCE Maribor, Slovenia
Prof. Dr B. Sredojević, FMCE Kraljevo, Serbia
Prof. Dr V. Stojanović, FMCE Kraljevo, Serbia
Prof. Dr I. S. Surovcev, VGSU Voronezh, Russia
Prof. Dr J. Tanasković, FME Belgrade, Serbia
Prof. Dr Lj. Tanović, FME Belgrade, Serbia
Prof. Dr D. Todorova, VTU Sofia, Bulgaria
Prof. Dr K. Weinert, TU Dortmund, Germany
Prof. Dr N. Zdravković, FMCE Kraljevo, Serbia
Prof. Dr D. Živanić, FTS Novi Sad, Serbia
Prof. Dr N. Zrnić, FME Belgrade, Serbia
Prof. Dr R. Vujadinovic, FME Podgorica, Montengro
Prof. Dr M. Jordović Pavlović, FMCE Kraljevo, Serbia
Prof. Dr Z. Đinović, ACMIT Wiener Neustadt, Austria
Prof. Dr Limonka Koceva Lazarova, UGD Štip, N. Maced.

ORGANIZING COMMITTEE

Chairman

Prof. Dr Vladimir Milićević, FMCE Kraljevo, Serbia

Vice-Chairman

Dr Nataša Pavlović, FMCE Kraljevo, Serbia

MEMBERS OF THE ORGANIZING COMMITTEE

Dr A. Petrović, FMCE Kraljevo, Serbia
Dr V. Grković, FMCE Kraljevo, Serbia
Dr M. Bošković, FMCE Kraljevo, Serbia
Dr M. Nikolić, FMCE Kraljevo, Serbia
Dr N. Stojić, FMCE Kraljevo, Serbia
Dr I. Franc, FMCE Kraljevo, Serbia
Dr M. Jovanović, FMCE Kraljevo, Serbia
Dr M. Nikolić, FMCE Kraljevo, Serbia
Dr V. Đorđević, FMCE Kraljevo, Serbia
Dr N. Petrović, FMCE Kraljevo, Serbia
Dr J. Bojković, FMCE Kraljevo, Serbia
Dr G. Bošković, FMCE Kraljevo, Serbia

TECHNICAL COMMITTEE

Chairman

Dr Vladimir Mandić, FMCE Kraljevo, Serbia

Vice-Chairman

Bojan Beloica, FMCE Kraljevo, Serbia

MEMBERS OF THE TECHNICAL COMMITTEE

M. Ivanović, FMCE Kraljevo, Serbia
M. Janićijević Milačak, FMCE Kraljevo, Serbia
S. Mihajlović, FMCE Kraljevo, Serbia
V. Sinđelić, FMCE Kraljevo, Serbia
M. Todorović, FMCE Kraljevo, Serbia
S. Pajović, FMCE Kraljevo, Serbia
P. Mladenović, FMCE Kraljevo, Serbia
A. Jovanović, FMCE Kraljevo, Serbia
M. Adamović, FMCE Kraljevo, Serbia
M. Rasinac, FMCE Kraljevo, Serbia
J. Perić, FMCE Kraljevo, Serbia

SESSION D: AUTOMATIC CONTROL, IT AND ARTIFICIAL INTELLIGENCE

Comparing QLoRA fine-tuning of a general-purpose and a mathematically specialized LLM on Serbian mathematics competition tasks

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

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

[2] Singidunum University, Belgrade, Serbia

ARTICLE INFO

* Correspondence: marina.svicevic@pmf.kg.ac.rs

DOI: https://doi.org/10.46793/ET26.D12S **PAGES:** D97 – D103

ABSTRACT

This paper examines the effect of QLoRA fine-tuning of locally executable large language models on mathematics competition tasks for high school students in Serbia. The main focus is on comparing the general-purpose model Mistral-7B-Instruct and the mathematically specialized model Mathstral-7B, in order to determine whether prior specialization for mathematical reasoning affects the success of fine-tuning on a small, domain-specific corpus of tasks in Serbian. For the purposes of the study, an automatic evaluation framework was applied, in which generated solutions were assessed according to several criteria, including final answer accuracy, logical coherence of the solution, and explanation quality. The results indicate different effects of fine-tuning for the observed models: the general-purpose model showed a slight performance decline, while the mathematically specialized model showed an improvement. This outcome suggests that models previously adapted to mathematical reasoning may benefit more from domain-specific fine-tuning, while also confirming that Serbian mathematics competition tasks remain a highly demanding test for locally executable large language models.

KEYWORDS

Large language models, QLoRA fine-tuning, Mistral-7B-Instruct, Mathstral-7B, Mathematics competitions, Automatic evaluation, Mathematical reasoning

1. INTRODUCTION

Large language models (LLMs) have made significant progress in solving mathematical problems, particularly on standard datasets commonly used for their evaluation. However, strong performance on such datasets does not necessarily imply comparable success in solving non-standard problems. Mathematics competition tasks represent a particularly demanding test, as solving them often requires more than the application of an appropriate formula or algorithm. It is necessary to interpret the problem conditions correctly, identify a suitable strategy, connect multiple reasoning steps, and reach a clearly justified conclusion [1], [2].

Assessing the quality of LLM-generated solutions is further complicated by the fact that a mathematically correct answer and a convincingly written explanation do not always match. A model may provide a clearly structured solution that contains an error, while in other cases it may generate a correct answer despite relying on incomplete or insufficiently justified reasoning. Therefore, the evaluation of such solutions should not be limited solely to checking the final result. The correctness of the reasoning process and the quality of the explanation should also be considered. Previous studies conducted on mathematics competition tasks in Serbia have shown that the performance of AI tools depends on the task type, mathematical domain, input format, and characteristics of the evaluated model [3], [4], [5].

Although commercial models often achieve better results, locally executable models are particularly relevant for research and practical applications. Their use provides greater control over data processing, improves experimental reproducibility, and enables model adaptation to specific needs without relying on external services. However, their performance on complex mathematical tasks remains limited. One possible approach to improving their capabilities is parameter-efficient

fine-tuning. Methods such as LoRA and QLoRA enable the adaptation of large language models with substantially lower memory and computational requirements than full fine-tuning of all model parameters [6], [7].

An important question is whether the effectiveness of such adaptation depends on the initial characteristics of the selected model. A general-purpose model and a model already specialized for mathematical reasoning may not benefit equally from additional training on a small, domain-specific corpus. In the first case, fine-tuning aims to adapt a broadly trained model to a narrower type of task. In the second case, additional training further adapts an already specialized model to a more specific linguistic and problem-solving context. This distinction is particularly relevant when the dataset consists of complex competition tasks in Serbian rather than a large collection of standard problems in English.

This paper compares the effects of QLoRA fine-tuning on two locally executable models of similar size: Mistral-7B-Instruct-v0.3, a general-purpose instruction model, and its variant Mathstral-7B, a model specialized for mathematical and scientific tasks. For both models, performance is analyzed before and after fine-tuning on mathematics competition tasks for high school students in Serbia. The generated solutions are assessed using an automatic multi-criteria evaluation framework that considers final answer accuracy, logical coherence, and explanation quality.

The aim of this study is to examine whether prior specialization for mathematical reasoning affects the outcome of domain-specific fine-tuning. Accordingly, the following research questions are addressed:

- How does QLoRA fine-tuning affect the performance of the general-purpose Mistral-7B-Instruct-v0.3 model and the mathematically specialized Mathstral-7B model on Serbian mathematics competition tasks for high school students?
- Does prior mathematical specialization impact the effectiveness of fine-tuning on a small, domain-specific corpus of tasks?

2. RELATED WORK

Research on improving the ability of large language models to solve mathematical problems has developed in several directions. One approach focuses on preparing higher-quality training datasets and expanding them with different variants of existing tasks. MetaMath, for example, was developed through fine-tuning on the MetaMathQA dataset, which was created by reformulating questions from standard mathematical datasets in multiple ways [8]. This approach shows that the diversity of training examples can play an important role in improving mathematical reasoning. Another direction involves continued pretraining on large mathematical corpora. DeepSeekMath-7B was developed by continuing the pretraining of DeepSeek-Coder-Base-v1.5 7B on a large collection of mathematically relevant content while retaining natural language and code data [9]. Similarly, Llemma was 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 records [10]. Mathstral-7B belongs to the same broader group of mathematically oriented models. It was built on Mistral-7B and designed for mathematical and scientific tasks [11]. Unlike general-purpose models, such models start from representations that have already been shaped toward mathematical reasoning. In addition to the choice of training data and the initial model, the method used to adapt the model to a specific domain is also important. Full fine-tuning of large models requires substantial computational resources, which has motivated the development of methods that update only a limited number of specific parameters. LoRA keeps the weights of the base model unchanged and introduces low-rank matrices that are trained for a specific task [6]. QLoRA further reduces memory requirements by using a quantized base model while computing gradients through LoRA adapters [7]. These methods are particularly relevant in research settings where locally executable models are used under limited hardware resources.

Although the mathematical capabilities of LLMs are often evaluated on standard datasets, such results do not provide a complete picture of their performance on non-standard tasks. Mathematics competition tasks require the identification of less obvious relationships, the selection of an appropriate strategy, and the consistent execution of multi-step reasoning. The CHAMP dataset was designed specifically for a more detailed analysis of these abilities using high school competition-level tasks. In addition to the tasks themselves, the dataset includes concept and hint annotations, enabling an analysis of how models use additional information during problem solving [2]. The authors also point out that a correct final answer does not necessarily result from valid reasoning, which is why evaluation should not be reduced solely to checking the final result. The performance of AI tools on mathematics competition tasks has also been examined in the Serbian-language context. A study conducted on tasks from the Mathematical Kangaroo competition showed that results depend on the input format, the language of the task statement, the mathematical domain, and the selected tool [3]. Later studies analyzed solutions to more demanding tasks from the Serbian national mathematics competition for high school students, generated using DeepSeek-R1, as well as o1 and o3-mini [4], [5]. The findings confirmed that advanced models can solve some non-standard tasks successfully, but difficulties remain with more complex formulations, logically demanding problems, and the consistent justification of all solution steps. In the context of mathematics competition tasks in Serbia, a framework for automated solving and multi-criteria evaluation of generated solutions was proposed and applied to commercial and locally executable models with different purposes [12]. The results revealed a clear difference between these two groups of models, while also indicating the potential of mathematically specialized models among locally executable solutions. A subsequent analysis of QLoRA fine-tuning of Mistral-7B-Instruct-v0.3 showed that adapting a general-purpose model to a small dataset of competition tasks does not necessarily improve performance on a held-out test set [13].

Although numerous studies have examined mathematically specialized models and methods for adapting them, considerably less attention has been paid to whether prior specialization affects the outcome of additional fine-tuning on a small, language-specific dataset. Therefore, this paper addresses a gap in the existing literature by directly comparing the effects of QLoRA fine-tuning on the general-purpose Mistral-7B-Instruct-v0.3 model and the mathematically specialized Mathstral-7B model using the same dataset and a unified evaluation framework.

3. EXPERIMENTAL DESIGN

3.1. Dataset and Experimental Workflow

The experimental dataset consists of 348 mathematics competition tasks for high school students in Serbia, organized by the Mathematical Society of Serbia. The dataset includes tasks from the municipal, regional, and national competition levels from 2022, 2023, and 2024, covering all four high school grades, as well as Categories A and B at the national level. According to mathematical domain, the tasks belong to algebra, geometry, number theory, or combinatorics. Only text-based tasks were included in the study, while tasks containing images or graphical elements were excluded. This made it possible to analyze the generated textual solutions without additional processing of visual content.

The original materials were available in LaTeX format and contained the task statements together with the corresponding official solutions. After preprocessing, each task was stored in a structured format along with relevant metadata, including the year, competition level, grade, category, and mathematical domain. Mathematical expressions were kept in LaTeX format to preserve their structural and semantic accuracy. To prevent overlap between the examples used for fine-tuning and the tasks intended for final evaluation, the dataset was split chronologically. The training set contains 192 tasks, the validation set 40 tasks, and the test set all 116 tasks from 2024. The final evaluation is therefore based on tasks that were not available during training. The dataset split is presented in Table 1.

Table 1: Chronological partitioning of the dataset.

Set	Number of tasks	Source
Training	192	Entire 2022 set and selected tasks from 2023
Validation	40	Regional competition tasks from 2023
Test	116	Entire 2024 set

The overall experimental workflow is shown in Figure 1. After data preparation and splitting, the baseline and fine-tuned variants of the two locally executable models were analyzed. For each of the four configurations, solutions were generated for the tasks in the test set, followed by automatic multi-criteria evaluation and a comparative analysis of the results.

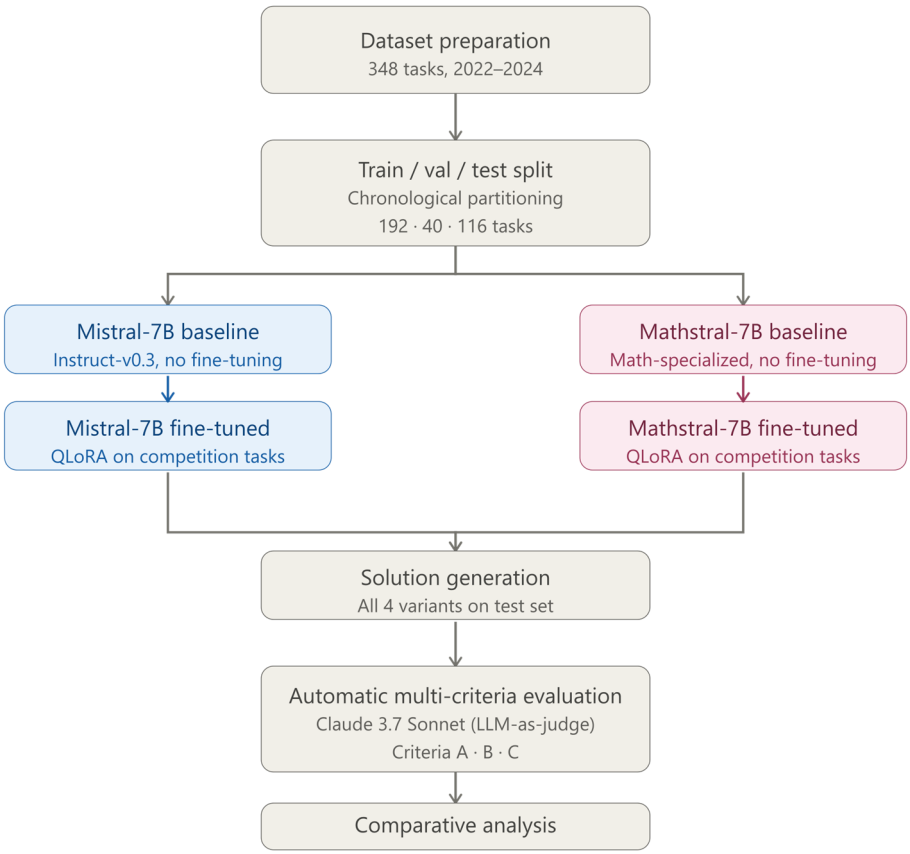

Figure 1: Experimental pipeline for comparing baseline and QLoRA fine-tuned variants of Mistral-7B-Instruct and Mathstral-7B.

3.2. Compared Models and Fine-Tuning Setup

The experiment considered two locally executable models with seven billion parameters, but different initial purposes. Mistral-7B-Instruct-v0.3 is a general-purpose instruction model, whereas Mathstral-7B is specialized for mathematical and scientific tasks. Selecting models of comparable size enables a meaningful comparison of their performance and the effects of additional domain-specific fine-tuning.

Two variants of each model were analyzed: the baseline version, without additional training on tasks from the dataset, and the version obtained through QLoRA fine-tuning. This resulted in four experimental configurations: Mistral-7B Baseline, Mistral-7B Fine-Tuned, Mathstral-7B Baseline, and Mathstral-7B Fine-Tuned. All variants were run locally and evaluated on the same held-out test set.

QLoRA was applied for the domain-specific fine-tuning of both models. The hyperparameters were not fixed identically in advance for the two models but were selected based on the results obtained during the experimental process. For each model, the configuration that achieved the most favorable results on the validation set was retained. The key fine-tuning parameters are presented in Table 2.

Table 2: Main QLoRA fine-tuning parameters.

Parameter	Mistral-7B	Mathstral-7B
LoRA rank (r)	16	16
LoRA alpha (α)	32	32
Number of epochs	5	5
Batch size	2	2
Gradient accumulation steps	16	16
Effective batch size	32	32
Learning rate	2×10^{-4}	5×10^{-5}
LR scheduler	Cosine	Cosine
Early stopping patience	2	2

For fine-tuning, the models were additionally configured with a LoRA dropout of 0.10, a warmup ratio of 0.10, a weight decay of 0.05, and a maximum sequence length of 4000 tokens. The LoRA adapters were applied to the projection modules q_proj, v_proj, k_proj, o_proj, gate_proj, up_proj, and down_proj. During solution generation, the maximum number of new tokens was set to 3400, with a temperature of 0.1, a top_p value of 0.95, and a repetition penalty of 1.1.

3.3. Automatic Multi-Criteria Evaluation

An automatic evaluation procedure based on the LLM-as-a-judge approach was applied to assess the generated solutions. For each task, the evaluation model Claude 3.7 Sonnet received the task statement, the corresponding official solution, and the answer generated by the analysed model. The same procedure was applied to all four experimental configurations, enabling their direct comparison on the same test set.

The choice of an automatic evaluation approach was motivated by the nature of the analysed tasks. Mathematics competition tasks require not only a correct final answer, but also a logically sound solution process and a clear explanation. Manually assessing a large number of generated solutions would be time consuming, whereas the LLM-as-a-judge approach allows the same criteria to be applied consistently and at scale. At the same time, it enables different aspects of a solution to be analysed separately, which is important because a correct answer does not necessarily result from valid reasoning.

The evaluation was conducted according to three criteria. Answer Accuracy (A) refers to the correctness of the final answer and takes a value from 0 to 1. Logical Coherence (B) assesses the logical consistency and mathematical justification of the solution process on a scale from 0 to 1. Explanation Quality (C) evaluates the clarity, completeness, and precision of the provided explanation, also on a scale from 0 to 1.

A separate evaluation prompt was used for each criterion so that answer accuracy, reasoning quality, and explanation quality could be assessed independently. The aggregate score was then calculated as follows:

$$S = 0.40A + 0.30B + 0.30C.$$

A higher weight was assigned to Answer Accuracy, as the correctness of the final result is particularly important in the context of mathematics competitions. Including Logical Coherence and Explanation Quality provides a more detailed view of how the model arrives at a solution. Although the same evaluation procedure was applied to all model variants, the obtained scores should be interpreted as automatic comparative indicators rather than as a substitute for expert mathematical assessment.

4. RESULTS AND DISCUSSION

The evaluation results for the four analyzed configurations are presented in Table 3. In addition to final answer accuracy, the table includes the logical coherence of the generated solution process, the quality of the provided explanation, and the aggregate score, which represents an overall measure of model performance.

Table 3: Performance comparison of baseline and fine-tuned model variants.

Model	Answer Accuracy (A)	Logical Coherence (B)	Explanation Quality (C)	Aggregate Score
Mistral-7B Baseline	8.5%	21.6%	28.1%	18.3%
Mistral-7B Fine-Tuned	6.5%	19.3%	26.6%	16.4%
Mathstral-7B Baseline	22.3%	23.7%	41.2%	28.4%
Mathstral-7B Fine-Tuned	24.1%	28.0%	42.9%	30.9%

A comparison of the baseline variants already reveals a clear difference between the two models. Mathstral-7B Baseline outperforms Mistral-7B Baseline according to all three individual criteria, as well as the aggregate score. The most pronounced differences are observed in final answer accuracy and explanation quality. This result indicates that prior mathematical specialization is relevant even when the model is applied to non-standard tasks in Serbian.

The effects of fine-tuning differ between the two analyzed models. No improvement was observed for Mistral-7B. After additional training, the values of all three criteria decreased, and the aggregate score was therefore lower. This decline suggests that adapting a general-purpose model to a small dataset of complex competition tasks does not necessarily lead to better generalization to unseen examples.

A different pattern is observed for Mathstral-7B. The fine-tuned variant achieves better results according to all the criteria evaluated. The largest improvement is recorded for the logical coherence of the solution process, while final answer accuracy and explanation quality also increased. The aggregate score rises from 28.4% to 30.9%, making Mathstral-7B Fine-Tuned the best-performing configuration among the four analyzed variants.

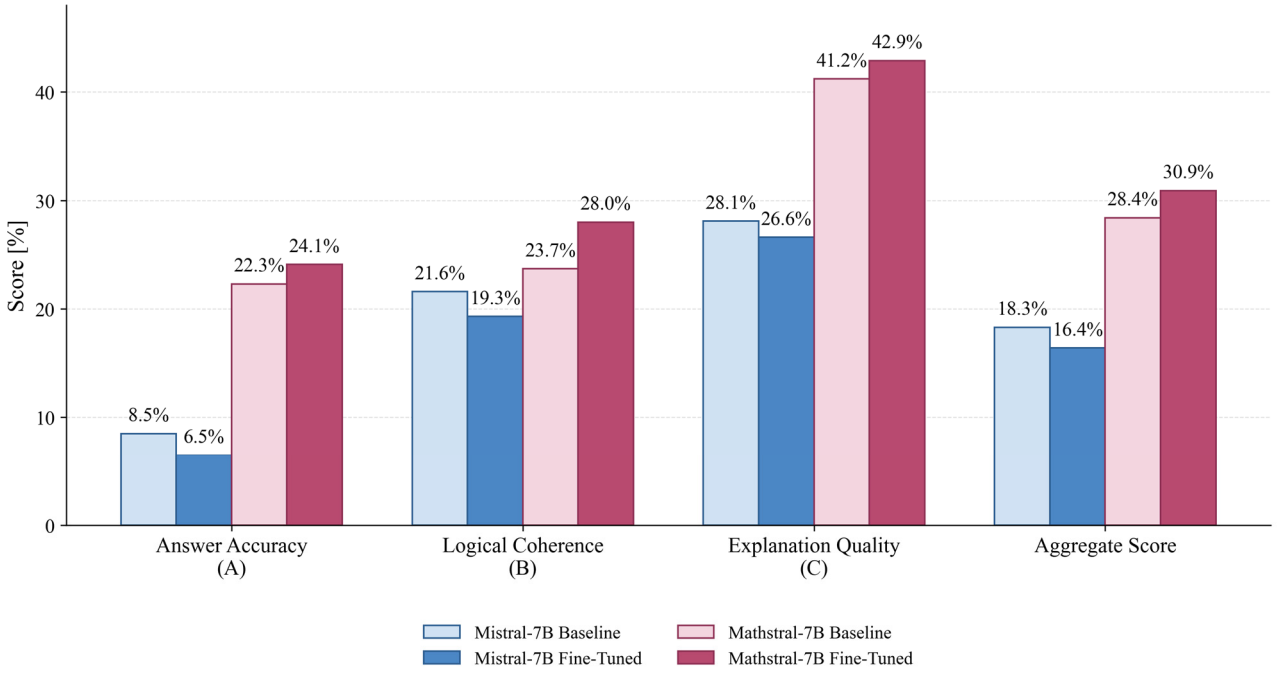

Figure 2: Performance comparison of baseline and QLoRA fine-tuned variants of Mistral-7B and Mathstral-7B across the evaluation criteria on the held-out test set of 116 tasks.

The results indicate that the effectiveness of domain-specific fine-tuning also depends on the initial characteristics of the selected model. Additional training on a relatively small corpus was not sufficient to improve the performance of the general-purpose Mistral-7B model. In contrast, Mathstral-7B benefited more from the same type of corpus, suggesting that fine-tuning may be more effective when it builds on previously developed mathematical reasoning capabilities.

The training dynamics are consistent with this finding. As shown in Figure 3, Mathstral-7B achieves a substantially lower evaluation loss throughout training, with a final value of 0.620 compared with 0.991 for Mistral-7B. At the same time, its peak evaluation mean token accuracy reaches 0.850, whereas Mistral-7B plateaus around 0.771. Notably, Mistral-7B shows a larger gap between its final training accuracy (0.804) and evaluation accuracy (0.765), suggesting a degree of overfitting, whereas Mathstral-7B displays almost no such divergence (0.851 vs. 0.849). These patterns indicate that Mathstral-7B adapted more effectively during fine-tuning, which is consistent with its stronger performance on the held-out test set.

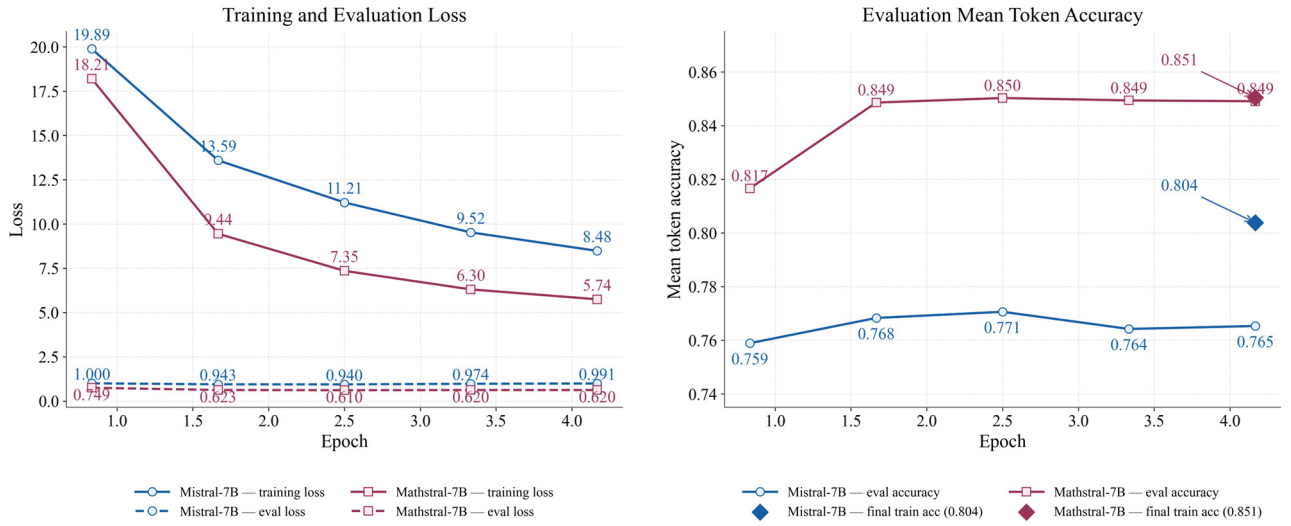

Figure 3: Training and evaluation loss (left) and evaluation mean token accuracy (right) across training epochs for Mistral-7B-Instruct-v0.3 and Mathstral-7B.

At the same time, the results should be interpreted with a degree of caution. The study considers two models of comparable size but different initial purposes, while the hyperparameters were selected based on the best results obtained during the experimental process for each model. Therefore, the observed differences cannot be attributed to a single factor alone. Nevertheless, the opposite directions of performance change after fine-tuning represent an important finding and justify further investigation involving a broader range of models and larger datasets.

The results also confirm the importance of multi-criteria evaluation. Across all configurations, Explanation Quality scores are higher than Answer Accuracy scores. This shows that models can generate responses that appear clear and well structured without necessarily producing a mathematically correct final result. Therefore, an evaluation based solely on the final answer does not provide a complete perspective of the quality of generated solutions. On the other hand, these results point to an additional important consideration: individuals evaluating these responses should possess sufficient experience in interpreting mathematical reasoning, as explanation quality extends beyond answer accuracy. A response may appear convincing due to LLMs presentation and structure, even when the underlying reasoning is incomplete or flawed. Consequently, evaluators with limited mathematical expertise may overestimate the quality of a solution based on the manner in which it is presented rather than on the validity of the mathematical argument itself.

5. CONCLUSION

This paper analyzed the effects of QLoRA fine-tuning on two locally executable models of comparable size but different initial purposes: the general-purpose Mistral-7B-Instruct-v0.3 model and the mathematically specialized Mathstral-7B model. The models were evaluated on mathematics competition tasks for high school students in Serbia by comparing their baseline and fine-tuned variants using an automatic multi-criteria evaluation framework.

The results show that domain-specific fine-tuning does not affect models with different initial characteristics in the same way. Mistral-7B exhibited a decline in performance after additional training, whereas Mathstral-7B improved across all evaluated criteria. The most pronounced improvement was observed in the logical coherence of the generated solutions. The difference between the models is also reflected in the training process, as Mathstral-7B achieved more favorable results on the validation set.

These findings indicate that the choice of the initial model is an important factor when planning domain-specific fine-tuning. A small dataset of complex competition tasks was not sufficient to improve the performance of the general-purpose model, but it enabled further adaptation of a model already specialized for mathematical reasoning. At the same time, the results confirm that Serbian mathematics competition tasks remain a demanding test for locally executable LLMs.

The study is limited to two models with seven billion parameters and a relatively small dataset. Future work may include expanding the corpus, incorporating additional mathematically specialized models, conducting a more detailed analysis of hyperparameter effects, and validating the automatically generated scores through expert assessment. It would also be useful to examine whether the observed pattern persists across other types of mathematical tasks and larger datasets. Future analyses should also examine performance by mathematical domain, competition level, and task type, in order to determine where fine-tuning produces the most significant gains.

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", Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, pp. 225–237, (2024)

[2] Y. Mao, Y. Kim, and Y. Zhou, "CHAMP: A competition-level dataset for fine-grained analyses of LLMs' mathematical reasoning capabilities", Findings of the Association for Computational Linguistics: ACL 2024, pp. 13256–13274, (2024)

[3] 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(4), p. e70063, https://doi.org/10.1002/cae.70063, (2025)

[4] N. Vučićević, M. Svičević, and A. Milenković, "Challenging DeepSeek-R1 with Serbian High School Math Competition Problems", Sinteza 2025-International Scientific Conference on Information Technology, Computer Science, and Data Science, pp. 274–280, (2025)

[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(1), pp. 14–29, https://doi.org/10.57016/TM-EYRQ1676, (2025)

[6] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, et al., "LoRA: Low-rank adaptation of large language models", International Conference on Learning Representations, (2022)

[7] 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)

[8] L. Yu, W. Jiang, H. Shi, J. Yu, Z. Liu, Y. Zhang, et al., "MetaMath: Bootstrap your own mathematical questions for large language models", International Conference on Learning Representations, (2024)

[9] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, et al., "DeepSeekMath: Pushing the limits of mathematical reasoning in open language models", arXiv preprint arXiv:2402.03300, (2024)

[10] Z. Azerbayev, H. Schoelkopf, K. Paster, M. D. Santos, S. McAleer, A. Q. Jiang, et al., "Llemma: An open language model for mathematics", arXiv preprint arXiv:2310.10631, (2023)

[11] Mistral AI, "Model Card for Mathstral-7B-v0.1", Hugging Face, https://huggingface.co/mistralai/Mathstral-7B-v0.1, (2024)

[12] M. Svičević, A. Milutinović, N. Vučićević, A. Milenković, and M. Pavković, "Methodological Framework for Automated Solving of Mathematics Competition Problems in Serbia Using LLMs", DeepTech Connect 2026, in press, https://a-side-project.eu/deeptech-connect-conference-2026/, (2026)

[13] M. Pavković, M. Svičević, A. Milutinović, N. Vučićević, and A. Milenković, "QLoRA Fine-Tuning of Mistral-7B for Serbian High School Mathematics Competition Tasks", Sinteza 2026, in press, (2026)