Jurnal Penelitian Sains Teknologi

NuminaMath 7B: Revolutionizing Math Solving with Integrated Reasoning Advanced Generative AI Tools and Python REPL

Adi Jufriansah[1], Irwan Akib[2], Naufal Ishartono[3], Azmi Khusnani[4], Tanti Diyah Rahmawati[5], Edwin Ariesto Umbu Malahina[6], Osniman Paulina Maure[7], Nova Tri Romadloni[8]

[1,4]Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia
[2]Faculty of Teacher Training and Education, Universitas Muhammadiyah Makassar, Indonesia
[3]Faculty of Teacher Training and Education, Universitas Muhammadiyah Surakarta, Indonesia
[5]Department of Ship Machinery Engineering Technology, Politeknik Pelayaran Surabaya, Indonesia
[6]Department of Informatics, STIKOM Uyelindo Kupang, Indonesia
[7]Faculty of Mathematics and Natural Sciences, Universitas San Pedro, Indonesia
[8]Faculty of Science, Technology and Animal Husbandry, Universitas Muhammadiyah Karanganyar, Indonesia

doi: 10.23917/saintek.v2i1.15728
Received: December 2nd, 2025 | Revised: January 30th, 2026 | Accepted: February 10th, 2026
Available Online: February 11th, 2026 | Published Regularly: March, 2026

Abstract

The efficacy of NuminaMath 7B, an AI model that was created to address mathematical challenges, is assessed in this investigation. We evaluated the model's accuracy and efficiency against conventional methods through experiments that produced quantitative data. Qualitative data were collected through surveys and interviews with users to gain insight into their experiences and pinpoint areas for improvement. The survey results indicated that users found NuminaMath 7B to be pertinent, effective, and user-friendly, as evidenced by the exceptionally high average scores in user experience (95), perception of features and interface (90), and additional feedback (85). NuminaMath 7B was able to offer mathematical solutions with logical and detailed explanations as a result of the model's development through two phases of adjustments, which were conducted using the Chain of Thought (CoT) methodology and inspiration from the Tool-Integrated Reasoning Agent (ToRA) framework. Testing demonstrated that the model achieved a score of 29 out of 50 in the AI Math Olympiad competition, despite encountering difficulties in resolving more intricate problems. This study underscores the significance and urgency of AI technology, particularly in the field of mathematics, as well as the significant potential of AI models to facilitate a more comprehensive comprehension of mathematical concepts.

Keywords: NuminaMath 7b, large language model, problem solving, chain of thought, AI math olympiad

✉**Corresponding Author:**
Adi Jufriansah, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Indonesia
Email: adijufriansah@gmail.com

Introduction

In the digital era that is becoming more advanced, the Large Language Model (LLM) has emerged as a revolutionary instrument that alters the manner in which information and knowledge interact [1], [2]. This is demonstrated by Parra et al. [3] and Almeida et al. [4], who have determined that LLM's capacity to resolve mathematical problems is of the utmost importance in a variety of disciplines due to its

high accuracy and speed. Furthermore, this decision is also substantiated by numerous studies, including Mazzullo et al. [5] and Hu et al. [6], which assert the same regarding the potential of LLM to enhance mathematical abilities and its function as a universal language in STEM. After being associated with discovery is intriguing in light of the growing demand for mathematical abilities [7], [8], [9]. The primary challenge is how to enhance the accuracy and efficiency of mathematical problem solving, which is exceedingly intricate. In numerous disciplines, including algorithm development, extensive data analysis, and other scientific applications, small calculation errors or the use of inappropriate approaches can be fatal [10], [11]. According to Drori et al. [12], the manual solution of mathematical problems is time-consuming and susceptible to human error. A model is required to effectively resolve the challenges posed by the growing volume of data and the complexity of problems [13], [14], [15], [16], [17].

The AI-MO/NuminaMath-7B-TIR model is expected to deliver substantial improvements in the efficiency and effectiveness of mathematical content explanation through the use of LLM [18]. Therefore, this context will be consistent with the research conducted by Chen at al. [19] on the challenges of maximizing the potential of LLM to support more dynamic and connected mathematics learning in accordance with the requirements of the present day. NuminaMath 7B is an innovative solution that integrates Python REPL (Read-Eval-Print Loop) and artificial intelligence (AI) to establish a platform that can enhance the efficiency and accuracy of mathematical problem-solving. The objective of this platform is to automate the identification and application of the most suitable problem-solving methods by utilizing AI technology to facilitate the process of resolving intricate mathematical problems. In the interim, Python REPL offers users the ability to construct and customize solutions to meet their unique requirements [20]. This combination allows NuminaMath 7B to offer solutions that are quick, accurate, and flexible, thereby reducing the manual workload and enhancing user productivity.

NuminaMath 7B can be employed as an effective teaching instrument in the educational context, facilitating the comprehension of intricate mathematical concepts by students. This platform has the capacity to offer students a more interactive and immersive learning experience by offering step-by-step explanations and rapid and accurate solutions. In addition, NuminaMath 7B can be employed by data analysts, engineers, and scientists to enhance the accuracy and productivity of their work in a professional setting. Consequently, NuminaMath 7B not only provides technical solutions but also establishes the foundation for a variety of future innovations. The primary goal of this research, as indicated by the background description, is to create technology that enhances the precision and efficacy of mathematical problem-solving. This research also demonstrates how the incorporation of Python REPL and AI can deliver solutions that are more precise and expeditious than those obtained through conventional methods. This study also looks at how well NuminaMath 7B works in a number of different situations, such as basic math and more complicated problems like optimization and data analysis. The goal is to make important contributions to the fields of artificial intelligence and mathematical problem-solving.

Method

This study employs both quantitative and qualitative methodologies to assess the efficacy of NuminaMath 7B (Figure 1). Quantitative data will be gathered through trials aimed at assessing the correctness and speed of solving mathematical problems with NuminaMath 7B in comparison with conventional approaches.

Figure 1. Research Methodology

1. Quantitative Methodology

Figure 2 illustrates the two intricate tuning phases that were implemented during the development of NuminaMath 7B TIR. The base model deepseek-math-7b was tailored to a diverse array of datasets that encompassed natural language mathematical problems and solutions during the initial stage [21]. This phase is essential for the development of a comprehensive understanding of a variety of mathematical concepts and their respective solutions [22], [23], [24], [25]. The Chain of Thought (CoT) methodology was implemented to enhance each solution, thereby facilitating logical and systematic reasoning. Further refining was implemented in the second stage to improve the model's capacity to address more intricate mathematical problems. The NuminaMath 7B TIR is capable of solving mathematical problems with high accuracy and providing solutions that are accompanied by detailed logical explanations, thereby enhancing user understanding and confidence in the results through the use of this two-stage approach.

The subsequent tuning stage was more specialized and concentrated, with a particular emphasis on synthetic datasets that emphasize integrated reasoning with assistive tools. During this phase, each mathematical problem is deconstructed into a sequence of logical steps, a Python program, and the resulting output. This methodology is motivated by Microsoft's ToRA (Tool-Integrated Reasoning Agent) framework, which employs GPT-4 to produce executable Python code that generates comprehensive solutions [26], [27]. The outcome is a model that is capable of solving mathematical problems in an efficient manner by integrating computational tools and natural language-based reasoning. This model provides solutions that are not only accurate but also directly implementable. In addition, the system architecture design encompasses the integration of a Python REPL for the direct execution of Python code in the solution of mathematical problems, as well as the development of an interactive user interface that is connected to the backend via an API service.

NuminaMath 7B TIR training is a meticulously organized process that employs specific hyperparameters to enhance performance. The learning rate is established at 2e-05, with a training batch size of 4 and an

evaluation batch size of 8. In order to guarantee reproducibility, the seed value is 42. A total training batch size of 32 and a total evaluation batch size of 64 were achieved by utilizing a multi-GPU configuration with 8 devices for the training. In order to guarantee stability during training, the Adam optimizer was implemented with beta values of 0.9 and 0.999 and an epsilon value of 1e-08. A preheat ratio of 0.1 was employed for four epochs, and a cosine learning rate scheduler was employed. In order to guarantee optimal compatibility and performance, the model was trained with cutting-edge frameworks. Transformers 4.40.1, PyTorch 2.3.1, Datasets 2.18.0, and Tokenizers 0.19.1 were the versions employed. The infrastructure required for the successful development and deployment of models is provided by these tools. The model was subjected to functional and performance testing to guarantee that all features operate in accordance with the specifications and operate proficiently. Collins et al. [28] suggest that the system's efficacy and satisfaction in enhancing mathematical comprehension can be evaluated by testing it with actual users. Additionally, the system results were compared to conventional mathematics teaching methods in order to assess the enhancements that this LLM offered.

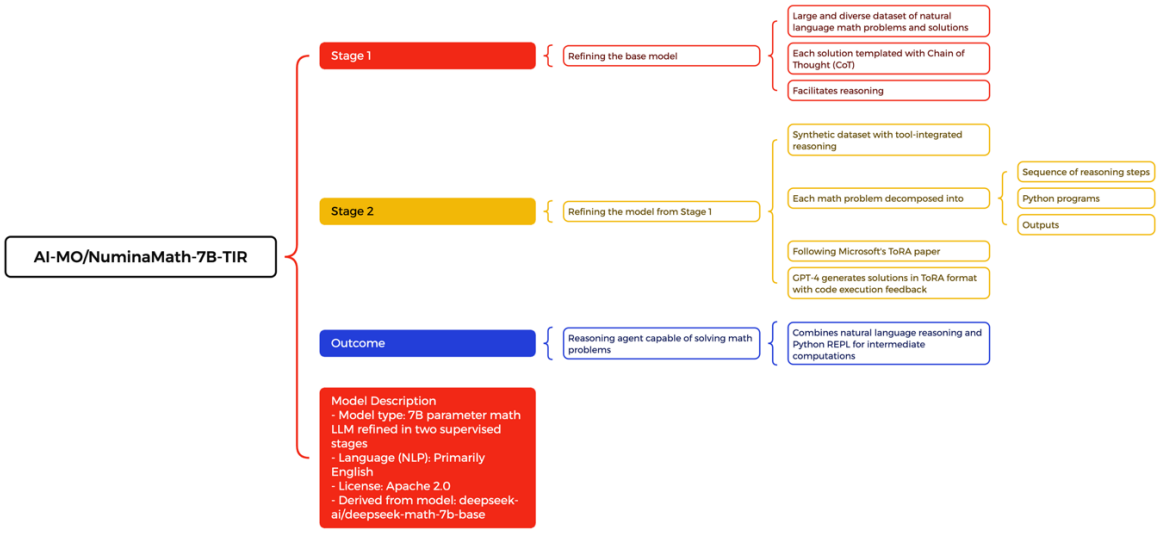

Figure 2. System Development Flow

2. Qualitative Approach

Qualitative data was collected through a survey with users to gain insight into the user experience using the platform, as well as to identify areas for improvement. The insight questionnaire is presented in Table 1.

Table 1. Insight Questionnaire on User Experience Using the Platform

Indicator	Statement
Experience with Usage	1. NuminaMath 7B is relevant for solving math problems. 2. NuminaMath 7B helps me understand math concepts better. 3. NuminaMath 7B improves my efficiency in solving math problems.
Perception and Feedback	1. The features in NuminaMath 7B are of good quality.

Indicator	Statement
	2. The interface of NuminaMath 7B is easy to use.
Additional Feedback	1. I am satisfied with the use of NuminaMath 7B overall.
	2. I recommend NuminaMath 7B to others.

The score for each questionnaire was obtained using the Likert scale formula according to equation 1 [29].

$$p_i = \frac{n_i}{N} \times 100\% \quad (1)$$

Where *p*i is the percentage, *n*i is the score obtained, and N is the total score. The scores are then converted into interval categories [30] as shown in Table 2.

Table 2. Likert Scale Index

Percentage	Definition
80-100	Very Good
60-79	Good
40-59	Fair
20-39	Bad
0-19	Very Bad

Result and Discussion

1. Development and Refinement Process

A Natural Language Processing (NLP) pipeline is presented in the form of an appealing paragraph using Python code that incorporates the `re` and `torch` modules, as well as the `transformers` library from Hugging Face. The `transformers` library from Hugging Face, which has transformed the way we work with complex language models, is utilized in the Python code provided to capitalize on the processing power of NLP [31], [32]. First, the `re` (regular expressions) module is imported, which enables the manipulation of text in a sophisticated and flexible manner by matching patterns. Next, PyTorch is imported to provide a deep learning framework that is both potent and efficient. This framework supports tensor operations and differentiation automation, which are essential for the training and use of language models.

The utilization of the `pipeline` from `transformers` is the most intriguing aspect of this code. This pipeline is a versatile utility that streamlines the process of establishing and executing a variety of NLP tasks, including sentiment analysis, text processing, and question answering [33], [34]. With this pipeline, we can utilize state-of-the-art pre-trained models for a diverse array of NLP applications with the help of a few lines of code. With this advancement, researchers and application developers alike can effortlessly incorporate state-of-the-art NLP technology into their projects. This code demonstrates the ease with which it is possible to develop robust AI-based applications using contemporary tools, thereby revolutionizing the manner in which we interact with text data in the digital age of today. One of the most sophisticated features of the `transformers` library is the establishment of a text-generation pipeline, which is the starting point of the code. This model has been optimized for mathematical problem-solving tasks, as indicated by the `model="AI-MO/NuminaMath-7B-TIR"` parameter. The `torch_dtype=torch.bfloat16` setting guarantees that the tensor data type is `bfloat16`, which enhances computational efficiency and memory usage without compromising accuracy. In addition, the `device_map="auto"` parameter enables this pipeline to autonomously identify and utilize available hardware, such as GPUs, to expedite the text generation process. The model and

token request execution procedure is illustrated in Figure 3, which provides the complete presentation.

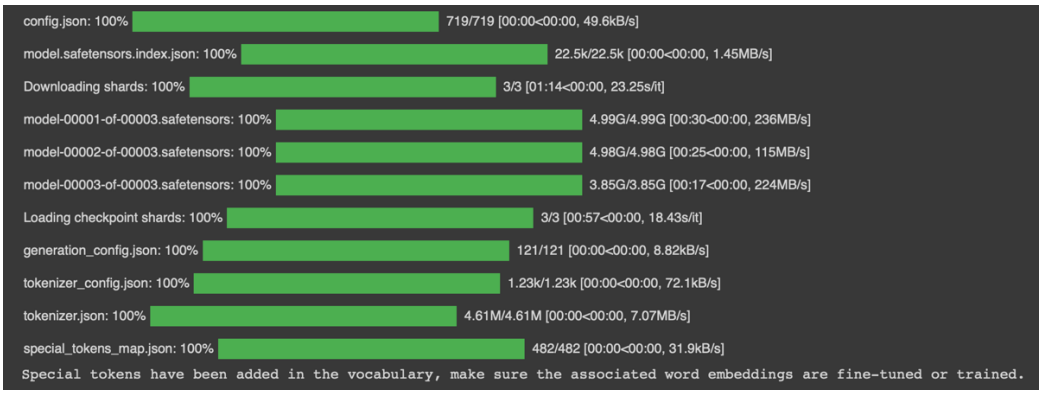

Figure 3. Model and Token Request Process

A math problem is defined in chat format as follows: "For what number of values of the constant k will the polynomial $x^{2}+kx+36$ have two distinct integer roots?" The apply_chat_template method of the tokenizer provided by the pipeline is then used to apply this message to the chat template. This method formats the message in accordance with the input structure anticipated by the model, without explicitly tokenizing it. Additionally, it includes a generation prompt to prompt the model to generate text. Subsequently, the text generation configuration (gen_config) is established. Several critical parameters are established in this configuration to regulate the text generation process. The utmost number of new tokens that the model can generate is defined by the max_new_tokens parameter, which is set to 1024. The model will generate deterministic text without random sampling, ensuring consistency of the results, as the do_sample parameter is set to False. The model is instructed to cease text generation upon reaching the conclusion of a Python code block by setting the stop_strings parameter to ["output"]. This ensures that the entire mathematical solution is generated in its entirety. Finally, the tokenizer parameter is set to the tokenizer from the pipeline, which guarantees consistency in the tokenization process with the subsequent code.

```
messages = [
  {"role": "user", "content": "For how many values of the constant $k$ will the polynomial $x^{2}+kx+36$
have two distinct integer roots?"},
]
prompt = pipe.tokenizer.apply_chat_template(messages, tokenize=False, add_generation_prompt=True)

gen_config = {
  "max_new_tokens": 1024,
  "do_sample": False,
  "stop_strings": ["```output"], # Generate until Python code block is complete
  "tokenizer": pipe.tokenizer,
}
```

Figure 4. Text Generation Configuration and Inference Setup in NuminaMath 7B TIR

2. Performance and Achievement

The capabilities of NuminaMath 7B TIR have been thoroughly evaluated and validated through a variety of tests. The AI Math Olympiad (AIMO) was attended by the model, which achieved a First Advancement award with a remarkable score of 29 out of 50 on both public and private test sets. This accomplishment underscores the model's proficiency in addressing mathematics problems of a competitive nature. Nevertheless, NuminaMath 7B TIR exhibited its capabilities in resolving problems up to the American Mathematics Competitions (AMC) 12 level. Nevertheless, it encountered difficulties with more intricate problems at the AIME and Math Olympiad levels, particularly in geometry. This accomplishment is indicative of the model's significant potential and pinpoints areas that require additional improvement.

3. Technical Specifications and Limitations

NuminaMath 7B TIR training entails the optimization of model performance through the use of several critical hyperparameters. The learning rate is established at 2e-05, with a training batch size of 4 and an evaluation batch size of 8. The training process is conducted in a distributed multi-GPU environment, which results in a total training batch size of 32 and an evaluation batch size of 64. Adam is the optimizer employed, and it is configured with specific beta parameters and epsilon values to ensure stability throughout the training process. Utilizing a cosine learning rate scheduler with a warmup ratio of 0.1, the training process spans four epochs to ensure optimal optimization. NuminaMath 7B TIR is subject to numerous constraints, despite its potent training program. The model has been explicitly developed for the restricted field of competitive mathematics and is not appropriate for general chat applications. Further, its performance may be inconsistent when confronted with more intricate problems and geometry as a result of its capacity constraints and the absence of multimodal capabilities, such as vision. These limitations suggest that, despite the fact that NuminaMath 7B TIR is exceptional in numerous respects, there is still space for additional development to address more intricate challenges.

4. Implementation and Usage

NuminaMath 7B TIR is accessible via an Inference Endpoint, which enables users to submit mathematical problems that are subsequently resolved by the model through the execution of Python code and natural language

processing. The model is a highly beneficial instrument in educational and competitive mathematics environments due to the numerous logical steps required to arrive at the final solution when applied in a real-world scenario. The NuminaMath 7B prototype interface is depicted in Figure 5.

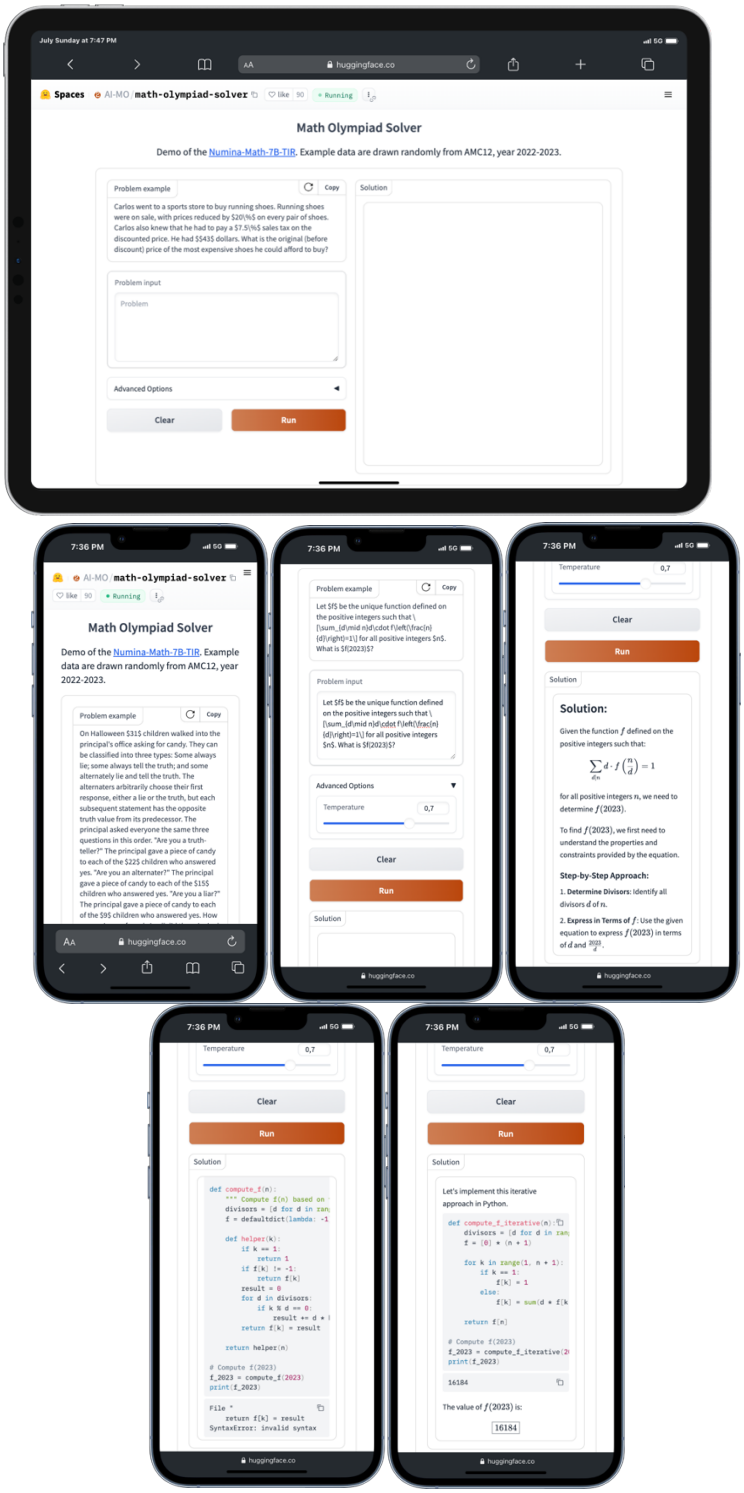

Figure 5. NuminaMath 7B Interface

Furthermore, users may capitalize on the public APIs that are accessible. The procedure for efficiently accessing functions in Gradio-based applications using a Python client is delineated in the subsequent steps. Initially, it is imperative to install the Python Gradio client if it has not already been done so. The command `pip install gradio_client` can be used to initiate the installation procedure. The subsequent step is to determine the API endpoint that corresponds to the desired function in the application after the implementation is complete. The application's documentation or user interface typically contain information regarding this endpoint. After locating the appropriate endpoint, replicate the provided code snippet and replace the placeholder values with the pertinent input data. In the event that the application is located in a private Space, authentication may be necessary through a Hugging Face token. The procedure for this can be reviewed in the official documentation. Alternatively, it is advisable to employ API Recorder, a tool that facilitates the automated recording of API requests. This feature simplifies the process of automatically generating API requests and minimizes the likelihood of errors in manual code writing. The integration with Gradio functions is made more structured and efficient by adhering to this methodology, which enables the maximum exploitation of the capabilities of Gradio-based applications. This implementation not only enhances the process's reliability but also guarantees the optimal utilization of Gradio technology, which in turn affects the efficiency and effectiveness of research that employs this instrument with the following code.

```
from gradio_client import Client
client = Client("AI-MO/math-olympiad-solver")
result = client.predict(
  text="Hello!!",
  api_name="/lambda"
)
print(result)

from gradio_client import Client
client = Client("AI-MO/math-olympiad-solver")
result = client.predict(
  example="The sum of three numbers is $96.$
The first number is $6$ times the third number,
and the third number is $40$ less than the second
number. What is the absolute value of the
difference between the first and second
numbers?",
  api_name="/lambda_1"
)
print(result)

from gradio_client import Client
client = Client("AI-MO/math-olympiad-solver")
result = client.predict(
  api_name="/clear"
)
print(result)

from gradio_client import Client
client = Client("AI-MO/math-olympiad-solver")
result = client.predict(
  api_name="/get_init_problem_input"
)
print(result)

from gradio_client import Client
client = Client("AI-MO/math-olympiad-solver")
result = client.predict(
  inp_text="Hello!!",
  temperature=0.1,
  api_name="/solve_problem_wrapper"
)
print(result)

from gradio_client import Client

client = Client("AI-MO/math-olympiad-solver")
result = client.predict(
  api_name="/get_running_btns"
```

```
)
print(result)

from gradio_client import Client

client = Client("AI-MO/math-olympiad-solver")
result = client.predict(
   api_name="/get_run_after_problem_input"
)
print(result)

from gradio_client import Client
```

```
client = Client("AI-MO/math-olympiad-solver")
result = client.predict(
   api_name="/update_example_problem"
)
print(result)

from gradio_client import Client

client = Client("AI-MO/math-olympiad-solver")
result = client.predict(
   api_name="/update_example_problem_1"
)
print(result)
```

Figure 6. Python-Based API Integration Using Gradio Client for NuminaMath 7B TIR

After conducting the survey, the NuminaMath 7B platform (Figure 6) received an average score of 95 out of 100 for the user experience. This indicates that the majority of users believe that NuminaMath 7B is highly relevant and effective in assisting them in the resolution of mathematical problems, the enhancement of conceptual understanding, and the efficiency of learning or instructing. Conversely, NuminaMath 7B's interface and features were also evaluated favorably, with an average score of 90, as indicated by the feedback and perceptions. The interface was user-friendly, and users found the features to be of high quality, which ultimately contributed to a positive experience. The platform received a score of 85 for the additional feedback aspect. However, this score still suggests a high level of satisfaction, despite being marginally lower than the former two categories. Users offered constructive suggestions for future development and positive recommendations, suggesting that the platform has the potential to be improved and adapted to meet the requirements of users in the future. In general, this data demonstrates that NuminaMath 7B is well-received by its users and has a positive impact on mathematics learning. Additionally, it offers valuable insights for ongoing development.

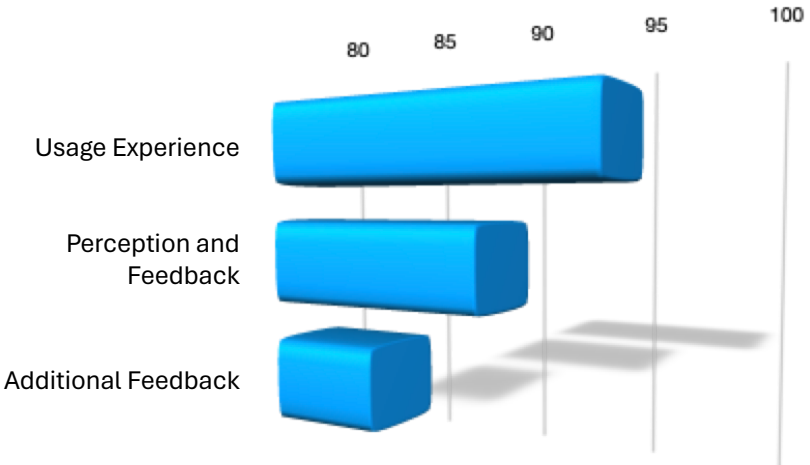

Figure 7. User Survey Results

The process of education and learning has been revolutionized by advancements in artificial intelligence (AI) and natural language processing in the era of digital technology [5], [19], [35]. The approach and comprehension of intricate disciplines, such as mathematics, have been revolutionized by these technologies [36], [37]. NuminaMath 7B is a noteworthy innovation that has had a substantial influence on mathematics across a variety of domains, distinguishing it from the numerous applications of AI. NuminaMath 7B enables students to pose questions and identify solutions to each step of a mathematical problem. NuminaMath 7B's personalized model enables interactive exploration of equations and functions, such as those found in high-level algebra problems, thereby facilitating a more profound comprehension. Furthermore, NuminaMath 7B improves conceptual clarity by presenting intricate mathematical solutions and offering visual representations. NuminaMath 7B facilitates students' comprehension of abstract concepts by providing textual descriptions and diagrams in subjects such as geometry, which elucidate three-dimensional shapes and spatial relationships.

NuminaMath 7B analyzes the learning patterns of each individual and adjusts the educational content accordingly using AI algorithms. In areas such as calculus, where proficiency levels differ, this adaptive approach guarantees that students receive personalized lessons and practice problems that enable them to advance at their own pace. NuminaMath 7B also provides assistance to students and professionals in the resolution of intricate mathematical problems, providing guidance in fields such as physics, engineering, and computer science. NuminaMath 7B facilitates the comprehension and application of sophisticated statistical methods, thereby guaranteeing the accuracy of data analysis, in the fields of probability theory and statistics. NuminaMath 7B illustrates the integration of concepts from various disciplines, thereby bridging the gap between mathematics and a wide range of fields. For instance, NuminaMath 7B displays how concepts from linear algebra and calculus are integrated to simulate real-world phenomena in applied mathematics, offering an interdisciplinary viewpoint. By facilitating students' comprehension of mathematical reasoning, theorems, and proofs, NuminaMath 7B also contributes to the advancement of higher education and research.

NuminaMath 7B is an essential instrument for the academic community, as it elucidates complex mathematical concepts and recommends methodologies for problem-solving in research. NuminaMath 7B fosters critical thinking and mathematical creativity by presenting open-ended challenges and dilemmas. For instance, NuminaMath 7B introduces unsolved problems in number theory, which motivates students to engage in discussions about theoretical aspects of mathematics and to investigate innovative solutions. NuminaMath 7B's language translation capabilities facilitate cross-cultural learning by facilitating the exchange of mathematical concepts across languages. NuminaMath 7B fosters interdisciplinary learning by offering insights into mathematical concepts associated with language in fields such as mathematical linguistics. NuminaMath 7B addresses the global challenge of educational accessibility by offering inclusive mathematics education. It is compatible with

assistive technology due to its text-based interface, which enables individuals with disabilities to access mathematical content. In addition, NuminaMath 7B provides alternative problem-solving methods and simplified explanations to accommodate a diverse range of learning requirements. NuminaMath 7B revolutionizes mathematics education by incorporating adaptive techniques, conceptual clarity, interactive learning, problem-solving assistance, the integration of mathematical fields, support for higher education and research, the encouragement of creativity and critical thinking, language translation, and increased accessibility. The function of NuminaMath 7B in mathematics education will continue to expand as technology advances, thereby cultivating a more profound understanding of the beauty and significance of mathematics worldwide. In order to establish an inclusive and empowering future for mathematics education, it is imperative to capitalize on this innovation. Table 3 provides a comprehensive overview of the role and challenges of NuminaMath 7B in the resolution of mathematical problems.

Table 3. Role and Challenges of Numinamath 7B in Solving Mathematical Problems

No.	Aspect	Role	Challenges	Technology	Solution
1	Problem-Solving Support	Providing solutions, guidance, and explanations for math problems	Accuracy, complexity, avoiding over-reliance	ML algorithms, NLP techniques	Rigorous testing, multiple solution paths, user education
2	Conceptual Understanding	Explaining concepts, clarifying doubts, offering real-life applications	Clarity, diverse learning styles, abstract concepts	Interactive simulations	Customized simulations, context-based examples
3	Learning Assistance	Customizing learning paths, recommending advanced math materials, interactive quizzes	Personalization, active engagement, maintaining interest	ML algorithms, gamification platforms	Adaptive learning, gamified content, interactive challenges
4	Language Translation	Translating math content into various languages	Accuracy, math notation, consistency	Neural translation engines	Human editing, handling specialized notations
5	Collaborative Learning	Facilitating collaborative problem-solving, ensuring a conducive environment	Security, misinformation, moderation	Secure platforms, AI moderation tools	Encryption, AI-based content filters
6	Accessibility	Making math accessible to individuals with special needs	User-friendly assistive technology, compatibility	Adaptive interfaces	Collaboration with experts, compatibility testing

No.	Aspect	Role	Challenges	Technology	Solution
7	Ethical Considerations	Ensuring privacy, preventing misuse, addressing bias	Data security, content moderation, unbiased responses	Bias detection algorithms	Secure data, continuous moderation, bias detection, and mitigation

NuminaMath 7B is capable of participating in natural language dialogue, a feature that is reminiscent of the renowned ChatGPT. With its intricate capabilities and ability to communicate mathematically with precision in symbolic representation and precise logical reasoning. NuminaMath 7B is capable of precisely interpreting mathematical questions, thereby preventing misunderstandings that may erode user confidence in the system. Another issue arises from the necessity of a profound contextual comprehension, as mathematical expressions can differ in meaning depending on the context. The manipulation of complex and non-standard mathematical symbols, as well as the resolution of ambiguities and misunderstandings, also present significant challenges. NuminaMath 7B consistently updates and incorporates specialized knowledge in a variety of mathematical domains, including algebra, calculus, geometry, and statistics, to ensure its adaptability and relevance. In order to ensure that users have a seamless and informative interaction with the system, it is crucial to maintain an up-to-date knowledge base. Table 4 illustrates the comprehensive role and obstacles of NuminaMath 7B in a variety of mathematical disciplines.

Table 4. Roles and Challenges of NuminaMath 7B in Various Fields of Mathematics

No.	Mathematical Field	Role of NuminaMath 7B	Challenges in Improvement
1	Algebra	NuminaMath 7B can assist in solving algebraic equations, explaining concepts, and offering step-by-step solutions	Interpreting various problem formats and accommodating diverse user input styles
2	Calculus	NuminaMath 7B can help solve calculus problems, perform differentiation, integration, and explain complex calculus concepts	Managing complex symbolic expressions and producing accurate, user-friendly mathematical explanations
3	Geometry	NuminaMath 7B can assist with solving geometry problems, explaining theorems, and visualizing geometric shapes and their properties	Interpreting and generating geometric diagrams and providing highly accurate logical proofs
4	Statistics	NuminaMath 7B can support statistical analysis, explain statistical concepts, and interpret datasets to aid in data-driven decision-making	Handling large datasets, understanding context-specific statistical methods, and ensuring accurate interpretations tailored to user needs
5	Trigonometry	NuminaMath 7B can assist in solving trigonometric problems, explaining trigonometric functions, and solving various trigonometric equations	Managing different trigonometric identities and equations, ensuring precise and context-appropriate solutions
6	Number Theory	NuminaMath 7B can assist with number theory concepts, prime factorization, and solving problems and theorems related to numbers	Handling large numbers, providing accurate prime factorizations, and

No.	Mathematical Field	Role of NuminaMath 7B	Challenges in Improvement
			comprehending complex number theory theorems
7	Linear Algebra	NuminaMath 7B can help solve systems of linear equations, perform matrix operations, and explain abstract linear transformations	Handling diverse matrix dimensions, ensuring accurate matrix computations, and explaining complex linear algebra concepts in an understandable manner
8	Differential Equations	NuminaMath 7B can help solve ordinary differential equations, explain solution methods, and visualize solutions in various scenarios	Handling different types of differential equations, understanding boundary conditions, and ensuring the accuracy and applicability of the generated solutions

NuminaMath 7B encounters numerous substantial obstacles in its pursuit of enhanced mathematical capabilities. One of these capabilities is the capacity to deconstruct intricate mathematical problems into user-friendly, structured steps. In order to effectively guide users through the solution process and provide a clear explanation of each step, it is necessary to possess not only profound mathematical expertise but also effective pedagogical skills. Furthermore, NuminaMath 7B must demonstrate proficiency in quantitative analysis and data interpretation, comprehend data visualization techniques, and effectively communicate the insights acquired to users through a diverse array of communication formats. Other challenges include effectively managing user feedback to identify areas for development and to continuously enhance the model's capabilities. The mathematical responses must be fair and accurate, and the system must be capable of learning from user interactions and refining its responses over time. NuminaMath 7B must be meticulously trained to avoid bias in mathematical responses and to comply with the principles of responsible use of AI and data privacy, which are also significant ethical concerns. NuminaMath 7B is anticipated to offer substantial advantages in enhancing the mathematical acumen of users from a variety of backgrounds, as it is dedicated to the development of high-quality AI technology and adheres to rigorous ethical standards.

The significance of developing and refining machine learning models that can facilitate the solution of mathematical problems with high levels of accuracy and efficiency is underscored by this study. NuminaMath 7B demonstrates that AI technology can be adapted to comprehend and resolve intricate mathematical problems, offering detailed and logical solutions with the appropriate approach. This model's success in the AI Math Olympiad and its overall performance are indicative of the significant potential of AI in mathematics education and competitions. This has also been the subject of discussion in studies [38], [39], with the results demonstrating the significance of developing and evaluating methods to evaluate the capacity of AI to solve International Mathematical Olympiad (IMO) problems.

The primary challenge is to apply mathematical creativity and thinking abilities to AI, as the assessment methods used for students

can be applied. The success of AI in mathematics competitions will be contingent upon the implementation of appropriate assessments. The AI Math Olympiad's rigorous testing of NuminaMath 7B's capabilities demonstrated that this model is capable of competing at a high level. Despite the fact that there are still obstacles to resolving more intricate problems, particularly those involving geometry, the model's accomplishments are already quite impressive. This implies that this model has the potential to be a highly beneficial instrument in academic competitions and mathematics education with additional refinement and development. The development of this model was conducted with meticulous attention to the technical details necessary to achieve optimal performance, as evidenced by the use of appropriate hyperparameters and a distributed multi-GPU training environment in terms of technical specifications. The Adam optimizer is recommended in the studies to ensure that the model is trained in an efficient and stable fashion by maintaining stability during the training process with specific beta parameters and epsilon values [40], [41], [42], [43]. The training process, which lasted for four epochs and utilized a cosine learning rate scheduler, also demonstrates that every aspect of the training was tailored to enhance the model's performance. Naturally, NuminaMath 7B has limitations, as do all technologies. General chat applications may not be appropriate for this model, as it is expressly intended for the restricted domain of competitive mathematics. Furthermore, the model's performance may be inconsistent when confronted with more intricate problems and geometric fields, which suggests its multimodal incapacity and limitations, including vision.

The significance of additional development to enhance the model's capacity to address a wide range of mathematical issues is underscored by these constraints. Underestimating the urgency and significance of this research is impossible. The utilization of AI technology in education is essential in the current digital era. This study offers compelling evidence that artificial intelligence (AI) can be employed to enhance comprehension and problem-solving in mathematics, a subject that presents a significant obstacle in the field of education. NuminaMath 7B not only assists in the resolution of mathematical issues but also instructs users on the methodology that underpins the solution. This is achieved through the provision of comprehensive and logical solutions. This has a significant effect on the comprehension of mathematical concepts and the efficacy of learning for both students and teachers. Additionally, the research demonstrates that AI models can be customized to suit a diverse array of specific applications with the appropriate modifications. In the realm of mathematics, the integration of AI technology with computational tools like Python demonstrates that it is capable of producing solutions that are not only precise but also directly implementable. This presents significant opportunities for further development in other disciplines that necessitate the resolution of intricate and logical problems.

The user survey results indicate that NuminaMath 7B has been well received by its users, as evidenced by the exceptionally high average scores in user experience, perception, and additional feedback. This suggests that users find the platform to be highly relevant and effective in assisting them in the resolution of

mathematical problems, the enhancement of conceptual understanding, and the efficiency of learning or instructing. The significance of continuing to develop and refine this model to accommodate the changing requirements of users is emphasized by its favorable reception. In conclusion, this investigation underscores the substantial potential of AI technology in the field of education, particularly in mathematics. The creation of more effective and efficient instruments to aid in the learning and solving of math problems can be achieved by continuing to develop and refine models such as NuminaMath 7B. In addition, this investigation demonstrates that AI technology can be adjusted to suit a diverse array of specific applications with the appropriate methodology, thereby creating significant opportunities for innovation in numerous other sectors.

The urgency and significance of this research are derived from its capacity to enhance the quality of education and assist in the resolution of intricate mathematical challenges, thereby generating substantial positive effects for students, instructors, and the education sector as a whole. This research has far-reaching implications in the long term. The utilization of AI technologies, such as NuminaMath 7B, has the potential to revolutionize the way we approach education, particularly in subjects that necessitate a profound comprehension and logical reasoning, such as mathematics. By continuing to further develop and refine this technology, we can develop more effective instruments to educate future generations, thereby facilitating their comprehension of complex concepts in a more efficient and effective manner. Additionally, this research presents opportunities for the development of AI applications in other domains that necessitate complex problem-solving, demonstrating the vast potential of this technology that has not yet been fully investigated. This research demonstrates that effective and efficient solutions can be provided to complex educational challenges through the use of AI technology when the appropriate approach is taken. By continuing to develop and refine this technology, we can create more effective instruments to aid learning and problem-solving, thereby having a substantial positive impact on society and education.

Conclusion

NuminaMath 7B is a prospective innovation in the field of mathematical problem solving. The platform provides rapid, accurate, and flexible solutions to a diverse array of mathematical problems by integrating the Python REPL and artificial intelligence. The introduction of NuminaMath 7B TIR, which boasts a structured approach to problem solving and sophisticated capabilities, provides a valuable resource for individuals who are confronted with more complex mathematical challenges. NuminaMath 7B TIR demonstrates the significant potential of AI to transform the way we solve mathematical problems, despite the fact that there are some areas for development, particularly in the integration of multimodal data and the handling of more complex problems.

Reference

[1] M. A. K. Raiaan *et al.*, "A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges," *IEEE Access*, vol. 12, pp. 26839–26874, 2024, doi: 10.1109/ACCESS.2024.3365742.

[2] S. Minaee *et al.*, "Large Language Models: A Survey," Feb. 2024.

[3] V. Parra, P. Sureda, A. Corica, S. Schiaffino, and D. Godoy, "Can Generative AI Solve Geometry Problems? Strengths and Weaknesses of LLMs for Geometric Reasoning in Spanish," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 8, no. 5, p. 65, 2024, doi: 10.9781/ijimai.2024.02.009.

[4] G. F. C. F. Almeida, J. L. Nunes, N. Engelmann, A. Wiegmann, and M. de Araújo, "Exploring the psychology of LLMs' moral and legal reasoning," *Artif Intell*, vol. 333, p. 104145, Aug. 2024, doi: 10.1016/j.artint.2024.104145.

[5] E. Mazzullo, O. Bulut, T. Wongvorachan, and B. Tan, "Learning Analytics in the Era of Large Language Models," *Analytics*, vol. 2, no. 4, pp. 877–898, Nov. 2023, doi: 10.3390/analytics2040046.

[6] B. Hu, L. Zheng, J. Zhu, L. Ding, Y. Wang, and X. Gu, "Teaching Plan Generation and Evaluation With GPT-4: Unleashing the Potential of LLM in Instructional Design," *IEEE Transactions on Learning Technologies*, vol. 17, pp. 1471–1485, 2024, doi: 10.1109/TLT.2024.3384765.

[7] C. Zhou, "Integration of modern technologies in higher education on the example of artificial intelligence use," *Educ Inf Technol (Dordr)*, vol. 28, no. 4, pp. 3893–3910, Apr. 2023, doi: 10.1007/s10639-022-11309-9.

[8] Y. Jiang and B. Li, "Exploration on the Teaching Reform Measure for Machine Learning Course System of Artificial Intelligence Specialty," *Sci Program*, vol. 2021, pp. 1–9, Nov. 2021, doi: 10.1155/2021/8971588.

[9] Y. Qiu, J. Pan, and N. A. Ishak, "Effectiveness of Artificial Intelligence (AI) in Improving Pupils' Deep Learning in Primary School Mathematics Teaching in Fujian Province," *Comput Intell Neurosci*, vol. 2022, pp. 1–10, Sep. 2022, doi: 10.1155/2022/1362996.

[10] L. Jiang *et al.*, "Opportunities and challenges of artificial intelligence in the medical field: current application, emerging problems, and problem-solving strategies," *Journal of International Medical Research*, vol. 49, no. 3, p. 030006052110001, Mar. 2021, doi: 10.1177/03000605211000157.

[11] F. Agterberg, "Geomathematics," 2023, pp. 512–519. doi: 10.1007/978-3-030-85040-1_12.

[12] I. Drori *et al.*, "A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level," *Proceedings of the National Academy of Sciences*, vol. 119, no. 32, Aug. 2022, doi: 10.1073/pnas.2123433119.

[13] Y. Wang, K. Chen, H. Tan, and K. Guo, "Tabi: An Efficient Multi-Level Inference System for Large Language Models," in *Proceedings of the Eighteenth European Conference on Computer Systems*, New York, NY, USA: ACM, May 2023, pp. 233–248. doi: 10.1145/3552326.3587438.

[14] R. K. Kodali, Y. Prasad Upreti, and L. Boppana, "Large Language Models in AWS," in *2024 1st International Conference on Robotics, Engineering, Science, and Technology (RESTCON)*, IEEE, Feb. 2024, pp. 112–117. doi: 10.1109/RESTCON60981.2024.10463557.

[15] S. Prasad, H. Gupta, and A. Ghosh, "Leveraging the Potential of Large Language Models," *Informatica*, vol. 48, no. 8, May 2024, doi: 10.31449/inf.v48i8.5635.

[16] J. Ji *et al.*, "GenRec: Large Language Model for Generative Recommendation," 2024, pp. 494–502. doi: 10.1007/978-3-031-56063-7_42.

[17] S. Kukreja, T. Kumar, A. Purohit, A. Dasgupta, and D. Guha, "A Literature Survey on Open Source Large Language Models," in *Proceedings of the 2024 7th International Conference on Computers in Management and Business*, New York, NY, USA: ACM, Jan. 2024, pp. 133–143. doi: 10.1145/3647782.3647803.

[18] E. Beeching *et al.*, "NuminaMath 7B TIR," *Hugging Face repository*, 2024, Accessed: Jul. 16, 2024. [Online]. Available: https://huggingface.co/AI-MO/NuminaMath-7B-TIR

[19] J. Chen *et al.*, "When large language models meet personalization: perspectives of challenges and opportunities," *World Wide Web*, vol. 27, no. 4, p. 42, Jul. 2024, doi: 10.1007/s11280-024-01276-1.

[20] L. T. van Binsbergen, M. Verano Merino, P. Jeanjean, T. van der Storm, B. Combemale, and O. Barais, "A principled approach to REPL interpreters," in *Proceedings of the 2020 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, New York, NY, USA: ACM, Nov. 2020, pp. 84–100. doi: 10.1145/3426428.3426917.

[21] Z. Shao *et al.*, "DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models," Feb. 2024, [Online]. Available: http://arxiv.org/abs/2402.03300

[22] B. Masikisiki, V. Marivate, and Y. Hlophe, "Investigating the Efficacy of Large Language Models in Reflective Assessment Methods through Chain of Thought Prompting," in *Proceedings of the 4th African Human Computer Interaction Conference*, New York, NY, USA: ACM, Nov. 2023, pp. 44–49. doi: 10.1145/3628096.3628747.

[23] Z. Zhang, A. Zhang, M. Li, and A. Smola, "Automatic Chain of Thought Prompting in Large Language Models," Oct. 2022.

[24] X. Wang *et al.*, "Self-Consistency Improves Chain of Thought Reasoning in Language Models," Mar. 2022.

[25] M. Suzgun *et al.*, "Challenging BIG-Bench Tasks and Whether Chain-of-Thought Can Solve Them," Oct. 2022.

[26] S. Schulhoff *et al.*, "The Prompt Report: A Systematic Survey of Prompting Techniques," Jun. 2024.

[27] F. Xu, Q. Sun, K. Cheng, J. Liu, Y. Qiao, and Z. Wu, "Interactive Evolution: A Neural-Symbolic Self-Training Framework For Large Language Models," Jun. 2024.

[28] K. M. Collins *et al.*, "Evaluating language models for mathematics through interactions," *Proceedings of the National Academy of Sciences*, vol. 121, no. 24, Jun. 2024, doi: 10.1073/pnas.2318124121.

[29] D. Sulisworo *et al.*, "Enhancing the science teacher skills on integration of augmented reality based media and learning strategy," 2023, p. 020045. doi: 10.1063/5.0154257.

[30] D. Sulisworo *et al.*, "The Science Teachers' Optimism Response to the Use of Marker-Based Augmented Reality in the Global Warming Issue," *Educ Res Int*, vol. 2021, pp. 1–9, Dec. 2021, doi: 10.1155/2021/7264230.

[31] A. Kolides *et al.*, "Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts," *Simul Model Pract Theory*, vol. 126, p. 102754, Jul. 2023, doi: 10.1016/j.simpat.2023.102754.

[32] S. M. Jain, *Introduction to Transformers for NLP*. Berkeley, CA: Apress, 2022. doi: 10.1007/978-1-4842-8844-3.

[33] I. Lipovac and M. B. Babac, "Developing a data pipeline solution for big data processing," *International Journal of Data Mining, Modelling and Management*, vol. 16, no. 1, pp. 1–22,

2024, doi: 10.1504/IJDMMM.2024.136221.
[34] S. Pais, J. Cordeiro, and M. L. Jamil, "NLP-based platform as a service: a brief review," *J Big Data*, vol. 9, no. 1, p. 54, Dec. 2022, doi: 10.1186/s40537-022-00603-5.
[35] L. Chen, P. Chen, and Z. Lin, "Artificial Intelligence in Education: A Review," *IEEE Access*, vol. 8, pp. 75264–75278, 2020, doi: 10.1109/ACCESS.2020.2988510.
[36] V. D. Kirova, C. S. Ku, J. R. Laracy, and T. J. Marlowe, "Software Engineering Education Must Adapt and Evolve for an LLM Environment," in *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, New York, NY, USA: ACM, Mar. 2024, pp. 666–672. doi: 10.1145/3626252.3630927.
[37] R. Nakamoto, B. Flanagan, T. Yamauchi, Y. Dai, K. Takami, and H. Ogata, "Enhancing Automated Scoring of Math Self-Explanation Quality Using LLM-Generated Datasets: A Semi-Supervised Approach," *Computers*, vol. 12, no. 11, p. 217, Oct. 2023, doi: 10.3390/computers12110217.
[38] S. Yang, "Mathematical Analysis: How Would AI Tackle Math Olympiad Problems?," *International Journal of High School Research*, vol. 3, no. 3, pp. 61–65, Jun. 2021, doi: 10.36838/v3i3.13.
[39] T. H. Trinh, Y. Wu, Q. V. Le, H. He, and T. Luong, "Solving olympiad geometry without human demonstrations," *Nature*, vol. 625, no. 7995, pp. 476–482, Jan. 2024, doi: 10.1038/s41586-023-06747-5.
[40] H. Kabiri, Y. Ghanou, H. Khalifi, and G. Casalino, "AMAdam: adaptive modifier of Adam method," *Knowl Inf Syst*, vol. 66, no. 6, pp. 3427–3458, Jun. 2024, doi: 10.1007/s10115-023-02052-9.
[41] M. Bhandari, P. Parajuli, P. Chapagain, and L. Gaur, "Evaluating Performance of Adam Optimization by Proposing Energy Index," 2022, pp. 156–168. doi: 10.1007/978-3-031-07005-1_15.
[42] F. Mehmood, S. Ahmad, and T. K. Whangbo, "An Efficient Optimization Technique for Training Deep Neural Networks," *Mathematics*, vol. 11, no. 6, p. 1360, Mar. 2023, doi: 10.3390/math11061360.
[43] M. Reyad, A. M. Sarhan, and M. Arafa, "A modified Adam algorithm for deep neural network optimization," *Neural Comput Appl*, vol. 35, no. 23, pp. 17095–17112, Aug. 2023, doi: 10.1007/s00521-023-08568-z.