Computer Science and Information Systems 00(0):0000–0000 https://doi.org/10.2298/CSIS251025050V

Paper Type: Research Paper

Generative AI in Financial Mathematics: Evaluating LLMs on a Bond-Focused Benchmark Dataset

Nemanja Vučićević[1], Marina Svičević[1], Aleksandar Milenković[1], Miloš Pavković[2], Lazar Stošić[3,4], and Jelica Protić[5]

[1] Faculty of Science, University of Kragujevac, Radoja Domanovića 12
34000 Kragujevac, Serbia
marina.svicevic@pmf.kg.ac.rs (corresponding author),
{nemanja.vucicevic, marina.svicevic, aleksandar.milenkovic}@pmf.kg.ac.rs
[2] Singidunum University, Danijelova 32
11000 Belgrade, Serbia
mpavkovic@singidunum.ac.rs
[3] Faculty of Informatics and Computer Science, University Union—Nikola Tesla, Belgrade, Serbia
[4] Don State Technical University, Rostov-on-Don, Russian Federation
lstosic@unt.edu.rs
[5] School of Electrical Engineering, University of Belgrade, Bulevar kralja Aleksandra 73
11000 Belgrade, Serbia
jelica.protic@etf.bg.ac.rs

Abstract. This paper presents a comprehensive evaluation of the performance of large language models (LLMs) on a newly created set of 50 tasks in financial mathematics, with a focus on bond valuation. Models o3, Claude Opus 4, DeepSeek-R1, and Gemini were tested on three types of tasks (practical, real-life, and theoretical) and three difficulty levels (easy, medium, and hard). Model o3 achieved the highest average accuracy (80.9%), while DeepSeek-R1 showed the weakest results (70.6%). The detailed analysis included error classification (methodological, numerical, conceptual), measuring solution time as well as the application of ANOVA and post-hoc tests, which confirmed statistically significant differences between the models. The results indicate that more complex and theoretical tasks remain challenging for all models. The proposed methodology and the set of created tasks provide a basis for further research and for the development of specialized LLM systems for financial analysis and educational purposes.

Keywords: Large Language Models, Financial Mathematics, Bond Valuation, Dataset Creation, Model Evaluation.

1. Introduction

The use of deep learning-based artificial intelligence models, especially large language models, is increasingly recognized as a transformative tool across a variety of application domains, including mathematics education [1], [2], financial modeling, and the solving of problem-oriented tasks. Trained on vast amounts of textual and numerical data, these

models have the potential to generate structured responses, carry out symbolic reasoning, and automate tasks that have traditionally been reserved for experts in different fields.

One challenge in modern research is understanding the limits and capabilities of these models on tasks that require the use of explicit mathematical formulas, the combination of multiple problem solving approaches, the interpretation of specialized terminology, and the derivation of proofs and conclusions through multiple steps. This challenge arises from a fundamental difference: LLMs are probabilistic sequence models that generate text by predicting likely next tokens, whereas formal mathematics, including financial mathematics, imposes deterministic requirements in which each step must be exactly correct and verifiable. A generated solution is therefore a statistically likely output rather than a formally guaranteed derivation, and its correctness cannot be assumed but must be established empirically. Although modern reasoning models narrow this gap through reinforcement-learning-trained chains of thought and tool-assisted verification, they still provide no formal guarantees. Systematic, domain-specific evaluation is consequently the primary means of determining where, and to what extent, these systems can be relied upon, which is precisely the aim of this study. Of particular relevance is the discipline of financial mathematics that integrates theoretical knowledge of economics and mathematics with demanding analytical and computational techniques [3]. Financial mathematics is characterized by high precision and procedural structures, especially for bond modeling, discounting future cash flows, yield to maturity estimation, and risk assessment [4], [5]. Unlike classical mathematical problems, tasks in this area often combine textual context, numerical tables, and complex formulas, making them particularly suitable for evaluating LLM performance.

Although LLMs have advanced rapidly, systematic studies in narrowly defined areas of applied mathematics such as bonds within financial mathematics remain limited. Most of the work targets general tasks from algebra, analysis, geometry, or combinatorics [6], [7], [8], [9], [10], while other studies focus on nonstandard competition tasks [11], [12], [13], [14]. Domains like financial mathematics, which are rich in formulas and procedures and typically taught at the university level, are still underexplored.

To address this gap, a dataset of financial mathematics tasks focused on bonds is created. The tasks were carefully selected to cover theoretical principles, practical calculations, and problem scenarios from real-world applications, and were classified into three levels of difficulty. Using standardized prompts, the ability of several widely used LLMs to understand, interpret, and correctly solve these tasks is tested, conducting both quantitative and qualitative analyses.

The goal of this research is a multilayered evaluation of LLM performance on a carefully designed financial mathematics set, emphasizing:

- accuracy of solution procedures as well as the types of errors;
- quantifying LLM performance and compare models on problem-solving tasks, across task types and difficulty levels;
- documenting characteristic failure modes and performance patterns across task types and difficulty levels, providing diagnostic signals for model developers and for integrating LLMs into AI/SE evaluation and regression-testing workflows;
- examining the models' potential for educational and professional finance use.

The contribution of this paper lies in its systematic evaluation approach within a well-defined mathematical finance domain, and in its purpose-built task set and methodology.

These components support future studies, benchmarking, and enhancement of artificial intelligence models in the financial domain through detailed error analytics and statistical analysis of differences.

2. Background and Related Work

2.1. Large Language Models and Generative AI

Large language models, developed on the transformer architecture, represent one of the most important technological innovations in modern artificial intelligence. Their ability to generate text, perform multistep reasoning, and respond to requests across diverse knowledge areas has enabled wide adoption in education, engineering, finance, medicine, and other domains. These models are trained on extensive datasets that include textual corpora, source code, mathematical expressions, and tables.

Among current LLMs, representative models include o3 (OpenAI), Claude Opus 4 (Anthropic), DeepSeek-R1, and Gemini 2.5 Pro (Google DeepMind), which will be referred to as Gemini in this paper. The o3 model, part of OpenAI's o-series of reasoning models trained with large-scale reinforcement learning, is characterized by advanced reasoning and the handling of formal expressions in complex problem settings [15]. Claude Opus 4, Anthropic's flagship model, demonstrates robust performance on tasks that require precise reasoning, argumentation, and the solution of technical problems [16]. DeepSeek-R1, an open model of the new generation, shows particular efficiency in mathematical problem solving and program coding, based on a combination of LLM principles and reinforcement learning [17]. Gemini, a multimodal model, is notable for integrating text with charts, tables, and images, extending beyond the classical boundaries of text-only modeling and finding use in applications such as medical diagnostics and financial analysis [18].

LLMs' core abilities include language understanding, information synthesis, symbolic and numerical reasoning, the generation of structured solutions, and the simulation of expert knowledge. Educational applications are becoming increasingly important, particularly for personalized learning, automated assessment, and support for problem solving. Studies show that LLMs can contribute to educational processes by generating content, providing explanations, and structuring responses, while also highlighting the need for careful integration in educational settings given the limitations in reliability, transparency, and pedagogical suitability [19].

These models are increasingly integrated into different expert domains and are not limited only to educational contexts. In software engineering, LLMs have become tools for code development and maintenance. They can automatically generate code segments based on description, saving developers time on repetitive, boilerplate tasks. For example, LLMs can write standard code and implement common algorithms, significantly accelerating development [20]. At the same time, they assist in debugging more complex software components by analyzing logs and error messages. They can also indicate possible bug causes and potential fixes. In practice, LLM tools increase developer productivity by allowing a greater focus on creative aspects of design and architecture rather than routine work. Related evaluations in algorithmic reasoning and complexity analysis further indicate that LLMs can assist in structured technical tasks, while requiring careful interpretation of their intermediate reasoning steps [21].

Beyond software engineering, LLMs are increasingly evaluated in other expert domains, including clinical decision support in medicine [22], [23], qualitative data analysis in the social sciences [24], and legal reasoning [25], [26]. A common finding across these domains is that strong general-purpose performance does not automatically translate into reliable expert-level behavior: domain-specific evaluation, validation, and expert oversight remain necessary. This observation directly motivates the present study, since financial mathematics is precisely such a specialized, formula-rich domain in which general benchmark results cannot be assumed to transfer.

In addition to these domains, a growing body of work focuses specifically on the application of large language models in finance. Recent survey studies provide systematic overviews of financial use cases, including market analysis, risk assessment, pricing, and decision support, while also highlighting domain-specific challenges related to numerical precision, regulatory constraints, and interpretability [27,28,29].

2.2. Previous evaluations of LLMs on mathematical tasks

LLMs have been extensively tested on mathematical tasks to assess their step-by-step reasoning and ability to solve complex problems. In recent years, standardized benchmarks have been developed that range from elementary word problems to high-level competition problems, with modern LLMs demonstrating steadily improving performance. In the following, the findings for ChatGPT, Claude Opus, DeepSeek-R1, and Gemini are highlighted, along with specialized mathematical models.

OpenAI's ChatGPT-4 demonstrates strong performance in standard arithmetic tasks such as GSM8K, reaching 85%-90% accuracy, indicating the ability to perform multi-step reasoning on basic text-defined problems. However, its performance drops below 50% on more complex high-school and Olympiad-level problems in the MATH dataset. Code-based self-verification boosts performance by up to 84%, approaching expert-level proficiency. [30]. A notable contribution to understanding the mathematical capabilities of GPT models, comes from the GHOSTS and miniGHOSTS datasets, which target graduate-level challenges. Analysis by more than 1,600 experts indicates that models such as GPT-4 are helpful assistants for retrieving mathematical facts and basic inference, but still lag in tasks requiring deeper conceptual understanding and advanced proof techniques [1]. Similarly, a recent study using interactive evaluation shows that GPT-4 can be highly useful for discussion and idea checking, although caution is required due to occasional mismatches between formal correctness and perceived usefulness [31]. From an educational perspective, analyzing GPT-4's errors is a powerful learning method that improves students' authentic problem-solving skills. A quasi-experimental study confirmed this, showing that primary students using this approach achieved superior results and greater mathematical self-efficacy compared to peers receiving traditional instruction [32]. Further insights into generative AI appear in evaluations on competition-level tasks, which demand deeper mathematical understanding and multi-step reasoning.

A 2025 study evaluated o1 and o3-mini on tasks from Serbia's National Mathematics Competition for high-school students on Category A and Category B (highly challenging and moderately challenging tasks respectively)[11]. Although perfect solutions were rare, the models' performance was typically strong enough to secure a top-three prize in the competition. The authors emphasize that these tools can benefit advanced students who compare their solutions with alternative approaches, provided that each step of the

model is interpreted critically and cautiously. The open DeepSeek-R1 model achieved comparable success, demonstrating strong performance on 36 different tasks from the same competition [14]. The model performed exceptionally well on tasks that required symbolic manipulation and algebraic reasoning, while logic and atypically formulated problems presented greater challenges. The model's scores in upper-grade Category B matched those of prize-winning students. To ensure accuracy, all problems were provided to the models in their original LaTeX format to preserve the integrity of the mathematical notation during the evaluation.

In a related study, several generative tools were evaluated using tasks from the Kangaroo competition [13]. The models tested included ChatGPT (free and paid), AI Math Solver, Math Mentor, and o1-preview. While the first three processed both text and images, the latter two were restricted to text-only input. Performance was strong in problems involving numbers and algebra, but struggled with geometry and logic. The models performance declined when required to interpret visual or abstract content, particularly diagrams and drawings. The o1-preview model stood out with excellent accuracy on text-only tasks. The study also highlights the need for models that more effectively process math tasks in languages other than English. Related evaluations have recently been conducted in other languages with similarly sized speaker populations: mathematical reasoning of LLMs, including reasoning-oriented models such as DeepSeek-R1, has been benchmarked in Catalan [33], while a multilingual benchmark based on Kangaroo tests has been proposed for evaluating visual mathematics in multimodal LLMs [34].

Analyses comparing leading commercial models, including Claude Opus 3, ChatGPT-4, and Gemini, reveal significant performance differences across the LLMs. In an analysis of first-year high-school textbook problems, GPT-4 emerged as the top performing model, achieving 85% accuracy and excelling at "compute" and "prove" tasks [35]. Claude Opus 3 followed with 67% accuracy, while Gemini lagged at 42%. Notably, all models struggled with tasks requiring explanation or visual interpretation, such as graphs and Venn diagrams. These findings align with Jonathan Vidal's study [36], where Gemini ranked last (67% accuracy) compared to Claude Sonnet 3 (80%) and ChatGPT-4 (93%) on textual word problems for high-school students. In particular, both Claude and ChatGPT-4 achieved perfect accuracy in counting tasks, while Gemini lagged significantly.

The wide spectrum of difficulty in benchmarks like GSM8K and MATH highlights the need for specialized LLMs trained on targeted mathematical corpora [1]. This has led to the emergence of specialized math models that enhance reasoning and precision by training on domain-specific data, enabling better multi-step logic, symbolic manipulation, and formal proofs.

Developed at Google Research, Minerva [7] represents one of the first systematic efforts to solve quantitative tasks using language models. By training on mathematical and scientific texts, it achieved high accuracy on STEM tasks like MATH, establishing a foundation for future work.

WizardMath [37] continues this direction of research, using a Reinforced Evol-Instruct approach to optimize mathematical reasoning without need for external code execution. Although a version based on Mistral-7B achieves high accuracy on standard benchmarks such as GSM8K and MATH, larger variants such as WizardMath-70B is better than some proprietary commercial models.

Similarly, Li et al. [38] found that a 7B model such as LLaMA-2 7B can achive high accuracy (up to 97.7% on GSM8K) through multi-sample generation with best-of selection. The primary challenge remains first-attempt reliability, or stability in generating a correct solution.

MathGenie introduces synthetic task generation using back-translation from solutions starting from a limited seed set. Models trained with this method (MathGenieLM) achieve state-of-the-art results on all five math benchmarks, including 87.7% on GSM8K and 55.7% on MATH [39].

In the multimodal domain, the MathVista benchmark assesses the ability to solve mathematical problems presented in visual contexts. Although GPT-4V leads as the best performer, it remains more than 10% behind human accuracy, highlighting the difficulty of integrating visual understanding with mathematical reasoning [40].

In the specific domain of financial mathematics, ChatGPT has shown solid performance on standard tasks but faces difficulties with complex analytical problems [41,42]. Another study focused on its ability to solve text-based tasks, concluding that ChatGPT is a useful tool for visualizing deductive steps and creating instructional graphics [43,44].

Beyond educational and theoretical tasks, there is growing interest in applying LLMs directly within financial mathematics, including recent evaluations of ChatGPT 5.2 on university-level tasks in this domain [45]. In this context, BloombergGPT, trained on a mixture of general and specialized financial text (about 700 billion tokens), is the first large language model focused on finance. While it is not the best performer on abstract math benchmarks, BloombergGPT performs exceptionally well on tasks such as risk assessment, market signal analysis, and understanding financial documents [46]. Complementing such domain-specific models, XFinBench benchmarks LLMs on complex financial problem-solving and reasoning tasks, further illustrating that strong performance on generic math datasets does not necessarily guarantee reliable applied financial reasoning [47].

Despite impressive LLM performance on math benchmarks, most of the mentioned research focused on general education or pure theoretical mathematics. A clear gap exists in studies targeting specific applied domains, such as financial mathematics and analysis, which require a combination of theoretical knowledge with practical calculation. Furthermore, few evaluations assess LLMs in non-English technical contexts. Our study addresses these gaps by providing the first evaluation of multiple LLMs on a curated set of financial mathematics problems focused on bonds.

3. Research Methodology

This study proceeds in two phases: (i) building a targeted dataset and (ii) systematically evaluating selected large language models (LLMs) on that dataset. The goal is to provide a comprehensive and reproducible assessment of model capabilities to solve financial mathematics tasks with a focus on bonds.

In the first phase, a diverse and structurally complex set of financial mathematics tasks is developed that cover the practical, real-life, and theoretical aspects of bonds. The dataset was carefully designed to include different types of financial instruments and real-world scenarios, as well as abstract theoretical tasks that test deeper mathematical reasoning in large language models.

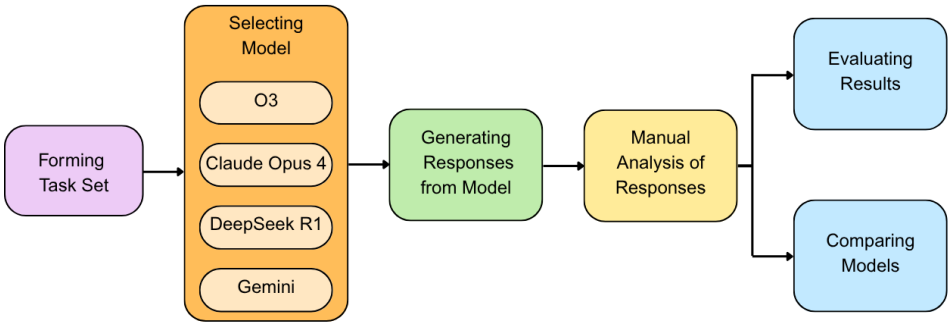


Fig. 1. Workflow of the evaluation process, from task creation and model selection to response generation and final analysis.

In the second phase, selected LLMs were evaluated on the constructed dataset using a standardized testing procedure consisting of the following steps (Fig. 1):

1. a task is selected from the dataset;
2. the task is provided as input to the selected AI model. Each task is presented in its original, unchanged format, in the Serbian language, including LaTeX syntax for mathematical notation, without any supplementary explanations or instructions;
3. the complete response generated by the model is recorded;
4. the recorded response is analyzed and scored according to predefined evaluation metrics.

All evaluations were conducted during June and July 2025 through the models' official public chat interfaces (ChatGPT for o3, claude.ai for Claude Opus 4, the DeepSeek chat application with the DeepThink (R1) mode for DeepSeek-R1, and the Gemini application for Gemini 2.5 Pro). This setting was chosen deliberately, as it reflects the way students and practitioners typically interact with these tools. All models were used with the default interface settings; decoding parameters such as temperature or top-p are not exposed in these interfaces and were therefore not modified, and no tools or extensions were manually enabled or disabled beyond the default chat configuration. The evaluation was carried out in a single chat session per model, initialized with the following standardized prompt (translated from Serbian): "In what follows, I will give you tasks from the field of financial mathematics, one by one. Give me the detailed steps of the procedure for solving each task." After this instruction, the tasks were submitted one at a time in their original form, and for each task only the model's first complete response was recorded and evaluated, without any follow-up interaction, clarification, or retries.

This method ensures an objective analysis of each model's essential capabilities, free from external factors or input adjustments. Such an approach not only achieves a high level of research validity, but also opens pathways for identifying the models' potential applications in educational and professional settings.

3.1. Construction and Analysis of the Financial Mathematics Task Dataset

The test set for this study features 50 various financial mathematics tasks, with a strong focus on bond-related tasks including practical, theoretical, and real-life scenarios. The selection process was designed to cover a broad spectrum of problem types and difficulty levels. To ensure relevance, the tasks are primarily based on common financial instruments like fixed-coupon bonds, zero-coupon bonds, T-Bills and T-Bonds, as well as certain types of inflation-indexed bonds (TIPS), complemented by abstract problems that generalize key concepts.

While constructing the dataset, a broad range of the latest literature in finance, financial modeling, and financial mathematics was used, with the goal of covering the widest possible variety of problem types, including both standard models and real-life problems. The tasks, concepts and ideas for designing the new tasks were inspired by the literature [48],[49], [50],[51],[52],[53], as well as the collections of tasks used for obtaining CFA and FRM certifications in finance. Instead of using these sources directly, they served as inspiration for creating new problems in bond valuation, interest-rate risk sensitivity analysis, portfolio construction, and other related topics. The dataset was also supplemented with elementary examples from lecture materials for the Financial Mathematics course at the Faculty of Science, University of Kragujevac. Theoretical problems were developed in collaboration with instructors from the Financial Mathematics course. All tasks were formulated and presented in Serbian, consistent with the language of instruction in the educational setting from which the benchmark was developed. Retaining the original language preserves the authenticity of the educational context and avoids introducing variation caused by translation. At the same time, presenting the tasks exclusively in Serbian may influence model performance and is therefore treated as a limitation in Section 4.4. This approach aimed to cover a wide variety of fundamental concepts related to the properties and analysis of bonds.

For a more detailed presentation, the dataset is categorized into the following groups based on the complexity, difficulty, and variety of the problems:

- **Practical bond valuation models**: This group includes tasks of varying difficulty focused on calculating core bond metrics such as bond price, current yield, yield to maturity (YTM), and determining transaction dates based on the given face value, payment frequency, and other parameters. These tasks test the correct application of the valuation of financial models, computational precision, and the overall problem-solving methodology. The problems range from direct formula application to complex scenarios that require advanced mathematical analysis and the use of sophisticated tools to solve implicitly defined functions.
- **Real-life bond models**: This category features tasks that simulate real financial situations using actual market data (e.g., T-Notes, fixed coupons). The tasks focus on portfolio immunization; constructing portfolios to meet future liabilities while minimizing interest rate risk and bond value sensitivity analysis. They cover topics such as cash-flow immunization for pension funds, price sensitivity, loss estimation, and transactions involving instruments like Treasury futures. Solving these tasks often requires a range of methods, from linear algebra to various optimization techniques, with difficulty determined by the scenario's complexity.
- **Theoretical bond models**: This category comprises formal tasks requiring proofs of general bond properties. Key topics include continuous compounding, the convexity

of the price-yield function (including under constraints), the asymptotic behavior of duration, and the structure of spot and forward rates. The primary motivation for including such tasks is to assess the models' capacity for deeper reasoning and abstraction. Solving these types of problem typically requires calculus (derivatives, limits), algebraic transformations, and interpretation of the obtained results.

Figure 2 provides a clearer illustration of the task distribution within the dataset using two pie charts. The left chart displays the distribution by difficulty, showing that medium-complexity tasks are most prevalent (44%), followed by easy (34%) and hard (22%) tasks. The right chart details the distribution by type: of the 50 tasks, 22 are practical, 21 are real-life, and 7 are theoretical. Although the theoretical group is the smallest, it covers diverse analytical requirements (see Fig. 2). This balance facilitates a multi-layered evaluation of the models across both practical and conceptual challenges.

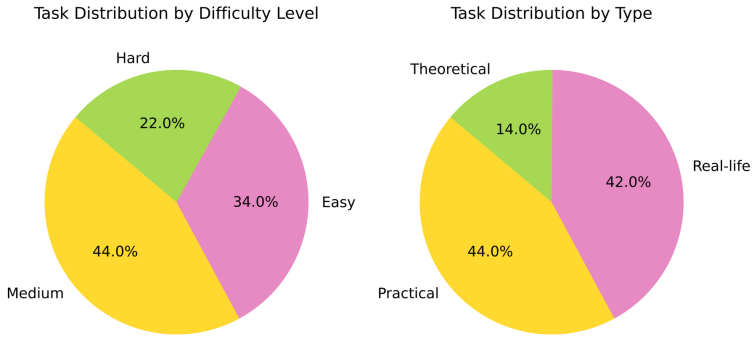


Fig. 2. Task distribution within the dataset, categorized by difficulty level (left) and problem type (right)

3.2. Model Selection and Technical Characteristics

For the purposes of this study, four modern large language models were selected; OpenAI's o3, Anthropic's Claude Opus 4, DeepSeek-R1, and Google DeepMind's Gemini, as they represent the most predominant and advanced approaches in generative AI today. The selection was made to cover different development paradigms, from closed commercial models to open-source solutions, and to include models with proven success on mathematical tasks. This strategic choice enables a comprehensive evaluation of the ability of generative systems to solve problems in financial mathematics by ensuring diversity in model design and prior performance.

OpenAI's o3 is a representative of the latest generation of proprietary reasoning models. It belongs to OpenAI's o-series, trained with large-scale reinforcement learning to produce extended chains of thought before answering, and it achieves high reported accuracy across a broad range of mathematical benchmarks [15]. During its reasoning process,

o3 can invoke tools such as a Python execution environment, which allows it to verify intermediate computations in quantitative tasks [15], [54].

Claude Opus 4 is Anthropic's flagship hybrid reasoning model, which offers both near-instant responses and an extended thinking mode for longer multi-step reasoning [16]. According to the official model documentation, it is optimized for sustained performance on long-running, multi-step tasks, a characteristic directly relevant for problems requiring extended mathematical derivations [16].

Gemini 2.5 Pro is Google DeepMind's multimodal reasoning model, which processes text and visual inputs and generates explicit reasoning steps before producing an answer [18]. Google reports strong results on mathematical benchmarks, including competition-level mathematics such as AIME 2025 [55]. At the same time, reported evaluations indicate limitations on tasks requiring strictly formal, step-by-step derivations [56]. Since all tasks in our benchmark are text-based, only the model's textual reasoning capabilities are exercised in this study.

Among open models, DeepSeek-R1 represents a new generation of LLMs aimed at improving mathematical reasoning through reinforcement learning. It is built on the DeepSeek-V3 Mixture-of-Experts (MoE) architecture, with approximately 671B total parameters of which about 37B are activated per token, and was trained primarily using large-scale reinforcement learning on chains of thought, without relying on extensive supervised fine-tuning [17]. According to its authors, its reasoning performance is comparable to OpenAI's o1 on complex reasoning tasks. The model weights are openly released, which facilitates reproducibility and domain-specific adaptation by the community [17].

The selected models thus differ in design characteristics that are directly relevant for the experiment: o3 and DeepSeek-R1 rely on reinforcement-learning-trained long chains of thought, with o3 additionally able to invoke tools during reasoning; Claude Opus 4 combines fast responses with an extended thinking mode; Gemini 2.5 Pro adds multimodal input processing, although only its textual capabilities are used here; and DeepSeek-R1 is the only open-weight model in the set, enabling full reproducibility.

This strategic selection of models enables not only a quantitative comparison on benchmark metrics but also a qualitative assessment of their problem-solving approaches on a well-defined set of tasks in financial mathematics.

3.3. Evaluation Metrics

Evaluating the performance of large language models in financial mathematics requires an approach that goes beyond assessing the numerical accuracy of final solutions. It requires a deeper analysis of the reasoning process, the appropriateness of the methods applied, and the validity of the resulting interpretations. Given the complexity of bond-related tasks, we developed a multi-layered framework to enable both quantitative and qualitative comparisons across models.

To ensure grading precision, we introduced a percentage-based scoring system, assigning each LLM response a value from 0 to 100 points based on its correctness and completeness. Each response was graded independently by three experienced evaluators, all of whom are instructors of the Financial Mathematics course and authors of the tasks. The final score of a response was defined as the arithmetic mean of the three individual grades. Inter-rater reliability was assessed using the intraclass correlation coefficient (two way random effects model, absolute agreement). Across all 200 graded responses,

the reliability of a single rater was ICC$(A, 1) = 0.97$ (95% CI $[0.96, 0.98]$), while the reliability of the mean of the three raters, which is the score used in all analyses, was ICC$(A, 3) = 0.99$, indicating excellent inter-rater agreement [57]. The three raters assigned identical scores for 118 of the 200 responses, with a mean within-response score range of 5.3 points. The evaluation is not based solely on the final numerical result and the conclusion; the model's ability to correctly interpret the given information is also considered. We also assess the selection of an appropriate theoretical framework, the execution of mathematical operations, and the provision of a meaningful interpretation of the results. Particular attention is paid to the chain of reasoning, as the tasks are often multi-step and require a combination of mathematical and economic knowledge. The evaluators followed a common scoring rubric, summarized in Table 1. The guiding principle is that the deduction reflects both the nature of the error and its impact on the final result: numerical inaccuracies within an otherwise valid procedure incur the mildest deductions, whereas an inappropriate choice of method or a misunderstanding of the underlying financial concepts leads to substantially lower scores. Consistent with this rubric, in our data the mean score of responses containing numerical errors was 79.6 (median 90), compared to 51.7 for methodological and 53.3 for conceptual errors.

Table 1. Scoring rubric used by the evaluators

Score	Description of the solution
100	Complete and correct procedure, accurate final result, and meaningful interpretation.
85–95	Correct method and valid reasoning; minor numerical inaccuracies (rounding, arithmetic slips) with limited impact on the result.
50–80	Generally correct approach, but with errors that materially affect the final result or its interpretation.
20–45	Substantial methodological or conceptual flaws; only some steps or partial insights are correct.
0–15	Fundamentally incorrect approach with negligible correct content.

For the quantitative evaluation, the model's overall mean accuracy is defined as the arithmetic mean of the individual accuracy scores:

$$AverageAccuracy = \frac{1}{N} \sum_{i=1}^{N} Accuracy(P_i) \tag{1}$$

where N is the total number of tasks and $Accuracy(P_i)$ is the number of points the model achieved on the task P_i.

In addition to quantitative scoring, we classify the types of errors that models make when solving bond-related tasks. By analyzing the solutions, we identify patterns that indicate conceptual misunderstandings, incorrect formulation of mathematical expressions, numerical inaccuracies in calculations, or inadequate interpretation of results. This approach enables not only quantitative measurement of success but also an understanding of the nature of individual model weaknesses.

Errors are further categorized into three basic types:

- methodological errors;
- conceptual errors;
- numerical errors.

Methodological errors refer to the wrong choice of the theoretical domain or the inadequate application of the mathematical framework. Conceptual errors are taken into account when the model shows partial understanding of the problem but fails to connect key financial concepts and relationships correctly. Numerical errors are technical inaccuracies in computations in which the reasoning is valid but the resulting numerical value is incorrect.

For the analysis of the model behavior, we evaluate the performance across two key dimensions: the type of problem (practical, real-life, or theoretical) and its difficulty (easy, medium, or hard). This two-dimensional framework allows us to assess how each model handles challenges ranging from straightforward computation to the high-level abstraction required for theoretical analysis.

For each of the two dimensions, we compute the mean accuracy using the following:

$$CategoryAccuracy = \frac{1}{N_c} \sum_{j=1}^{N_c} Accuracy(P_j) \tag{2}$$

where N_c is the number of tasks in the observed category and $Accuracy(P_j)$ is the score for the j-th task in that category.

The overall mean accuracy of each model was calculated for the full dataset and for each category. This approach provides a reliable foundation for identifying the strengths and limitations of different LLM systems in solving financial mathematics problems in the bond domain.

In addition to solution accuracy, we measured the response generation time for each task. This was recorded as the interval from the moment of task submission until the model produced its complete and final solution. The time denoted $T(P_i)$ for task P_i, is used to assess the average solving speed of different models.

The average solving time for each model is computed as:

$$AverageTime = \frac{1}{N} \sum_{i=1}^{N} T(P_i) \tag{3}$$

where N is the total number of tasks and $T(P_i)$ is the time required to solve the i-th task. This indicator adds an additional dimension to model evaluation, particularly relevant in time-sensitive scenarios.

4. Results and Discussions

4.1. Evaluation of LLM-generated Solutions

The results presented in this section provide a multi-layered view of the ability of selected large language models to solve problems in financial mathematics, with a focus on the bond domain. Unlike previous studies that predominantly examined general mathematical

problems, this study evaluates model performance in a precisely defined and applied area, combining quantitative and qualitative evaluation.

We analyze three aspects: (i) overall solution accuracy, (ii) performance as a function of task type and complexity, and (iii) time solving efficiency. In addition, a detailed error type analysis provides more insight into weaknesses and potential risks when applying the models in real-world scenarios.

The first part of our analysis considers the overall mean accuracy of each LLM across the entire task set. Figure 3 summarizes these results, showing the arithmetic mean of the scores for each of the four models, expressed as a percentage as defined in Equation (1). The o3 model achieved the highest accuracy, with an average of 80.9%, followed by Gemini 78.9% and Claude Opus 4 74.9%, while the lowest result was recorded by DeepSeek-R1, with an average accuracy of 70.6%. These results indicate that all models can solve complex bond-related tasks with a high level of success while also revealing clear differences in precision and reliability. Although the differences between models may be considered moderate, it is noticeable that proprietary models slightly outperform the open-source model DeepSeek-R1.

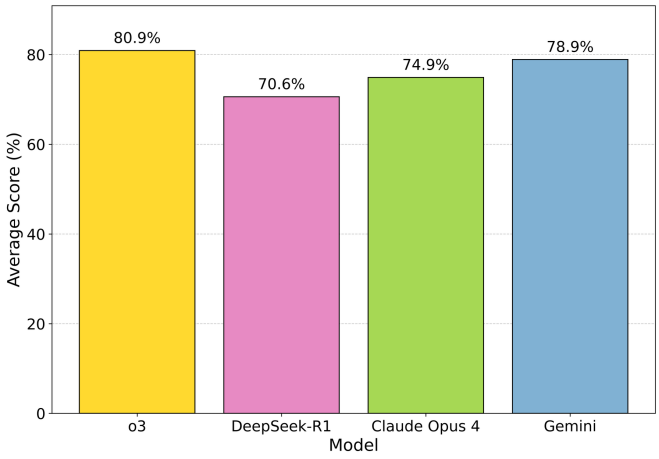


Fig. 3. Average accuracy scores of the evaluated LLM models on the complete set of financial mathematics tasks

Figure 4 presents a breakdown of each model's mean accuracy by task difficulty (easy, medium, and hard), with scores calculated according to Equation (2). This approach provides a deeper insight into how different LLMs respond to variations in the complexity of problems in financial mathematics.

The results show that all models achieved exceptionally high scores on easy tasks, with average accuracy above 90%, where o3 model achieved the highest accuracy of 95.3%, and DeepSeek-R1 of 92.6% which performs the worst. Although all models experience a slight performance decline on medium-difficulty tasks, the results are still robust. o3 maintains its lead with a score of 83.9%, while DeepSeek-R1 performs the weakest in this category at 65.7%. Hard tasks present the greatest challenge, with all models showing

a substantial drop in accuracy. Gemini achieved the highest accuracy in this category at 53.6%, followed by o3 at 52.7%, while Claude Opus 4 and DeepSeek-R1 showed lower accuracies of 47.3% and 46.4%, respectively. This analysis clearly indicates a downward trend in model performance as task complexity increases, confirming that higher levels of abstract reasoning and complex mathematical analysis remain a significant challenge for current AI models.

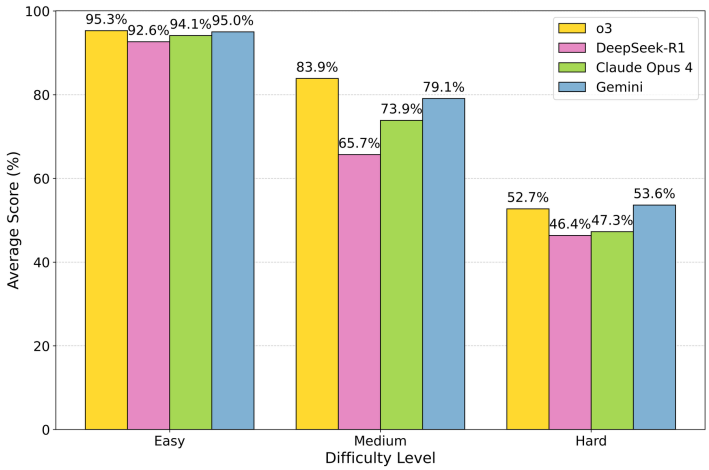


Fig. 4. Average accuracy of the models across different task difficulty levels (Easy, Medium, Hard)

Breaking performance down by task category, Fig. 5 shows the average accuracy achieved on theoretical, practical, and real-life problem tasks based on Equation (2). This structured analysis enables a more precise view of each model's capabilities with respect to the nature of the tasks being solved. On theoretical tasks, which require deeper mathematical understanding and formal inference, the average accuracies ranged between 71.4% and 76.4%, with DeepSeek-R1 achieving the best result. In this category, all models exhibited similar performance without major variation. The highest accuracy levels were achieved on practical bond valuation tasks, which mainly involve applying financial formulas and performing numerical calculations. The o3 model was the top performer at 90.0%, followed closely by Gemini at 88.9%, while the remaining models followed with a more significant performance gap. Likewise, real-life tasks involving applied bond models, which require portfolio construction, sensitivity analysis, or solving complex financial scenarios, also revealed larger performance disparities and an overall drop in accuracy. The o3 model achieved the best result at 73.8%, while DeepSeek-R1 recorded the lowest accuracy at 62.4%. These results indicate that the models perform best on tasks involving well-defined numerical calculations and familiar patterns. In contrast, they struggle with more complex, real-life tasks that require multi-step reasoning and a combination of different approaches. Overall, the analysis shows that model performance is not uniform across categories, demonstrating that the type of problem is an important factor in each

model's success. Table 2 summarizes the corresponding means and standard deviations for each problem category, quantitatively supporting the performance patterns illustrated in Figure 5.

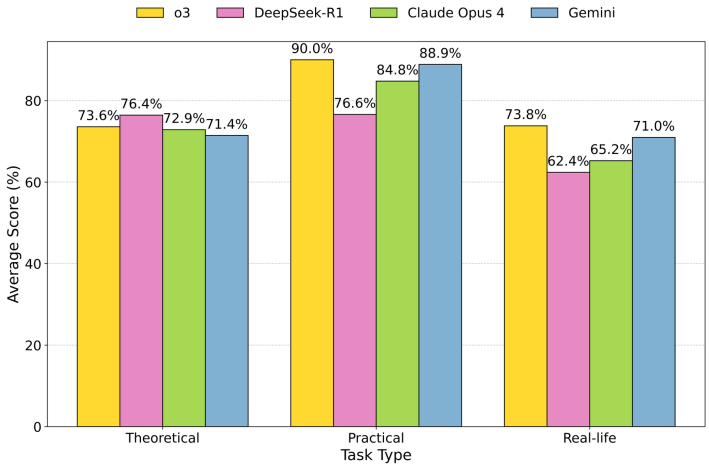


Fig. 5. Average accuracy of the four LLM models across different task categories: Theoretical, Practical, and Real-life

Table 2. Performance of the four models by problem type

Task type	Statistics	o3	DeepSeek-R1	Claude Opus 4	Gemini
Theoretical	M	73.6	76.4	72.9	71.4
	SD	35.4	35.0	35.6	38.5
Practical	M	90.0	76.6	84.8	88.9
	SD	11.9	24.7	16.2	13.1
Real-life	M	73.8	62.4	65.2	71.0
	SD	28.0	32.7	29.8	30.6

In addition to accuracy, we analyzed the average time each model required to solve a task using Equation (3), with the results shown in Fig. 6. This analysis offers insight into model efficiency in terms of response generation speed, a key factor in real-time applications. Claude Opus 4 recorded the shortest average time at $11.0s$, indicating a high processing speed. The o3 model averaged $44.1s$, while Gemini was slightly slower with an average time of $53.5s$.

In contrast, DeepSeek-R1 required substantially more time, averaging $289.4s$ per task. This slower performance is largely attributed to DeepSeek-R1's design as a reasoning model that produces detailed step-by-step solutions, generating many intermediate tokens and computations that can slow down inference [58]. Additionally, the public DeepSeek-

R1 service often runs at capacity, which means user requests are queued until server resources are freed during busy periods [59]. These factors together, explain why DeepSeek-R1's completion times were substantially higher compared to other LLMs.
Overall, the results show that large differences in execution time do not necessarily track differences in accuracy, confirming that inference speed is not a reliable predictor of accurate solutions in bond-focused financial mathematics. Table 3 provides the corresponding 95% confidence intervals for both the average completion times and the average scores, further confirming the substantial performance gap between DeepSeek-R1 and the remaining models.

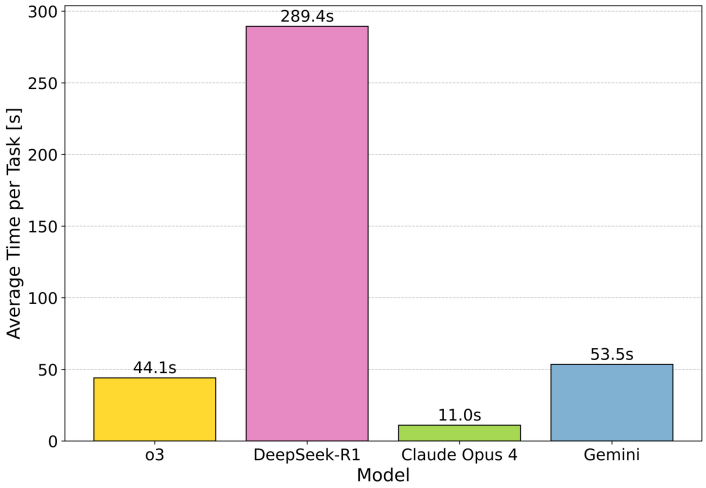


Fig. 6. Average time required per task (in seconds) across the four evaluated LLM models

Table 3. 95% Confidence intervals for average times and average scores for all four models

	o3	DeepSeek-R1	Claude Opus 4	Gemini
Time	[33.42; 54.78]	[225.79; 353.05]	[7.56; 14.44]	[47.77; 59.15]
Score	[73.93; 87.87]	[62.09; 79.11]	[67.31; 82.49]	[71.29; 86.51]

Figure 7 presents a detailed analysis of incorrect solutions, classifying the errors for each model into three categories: methodological, conceptual, and numerical. This part of the analysis focuses on tasks where the maximum score was not achieved, providing deeper insight into the failure mechanisms of each LLM system. The total count of tasks with errors varied by model, reflecting differences in their ability to consistently produce accurate solutions. DeepSeek-R1 recorded the highest number of errors (36), while o3, Claude Opus 4, and Gemini had 31, 34, and 32 errors, respectively.

An analysis of error types shows that numerical errors were the most prevalent for all models, indicating difficulties with precise calculation even when the logical approach was correct. These errors are particularly significant in financial mathematics, where minor inaccuracies in number rounding, imprecise methods, or arithmetic flaws can lead to substantial downstream consequences. Furthermore, many problems require correct interpretation of numerical results, such as calculating the number of bonds, so even a small error of 10^{-4} can significantly affect the final conclusion. These inaccuracies sometimes caused the models to misinterpret the results, highlighting their critical importance.
Methodological errors, which result from the selection of an inappropriate specialized domain of financial mathematics, were also notable across all models, with Claude Opus 4, DeepSeek-R1, and Gemini recording the same number of such errors. Conceptual errors, indicating a deeper misunderstanding of financial principles, were the least frequent overall. DeepSeek-R1 exhibited more of these errors than the other models, while Claude Opus 4 and Gemini had the fewest, with three each. This detailed error analysis not only pinpoints specific weaknesses, but also reveals characteristic failure patterns for each model, a crucial step for improving their performance in specialized domains like financial mathematics.

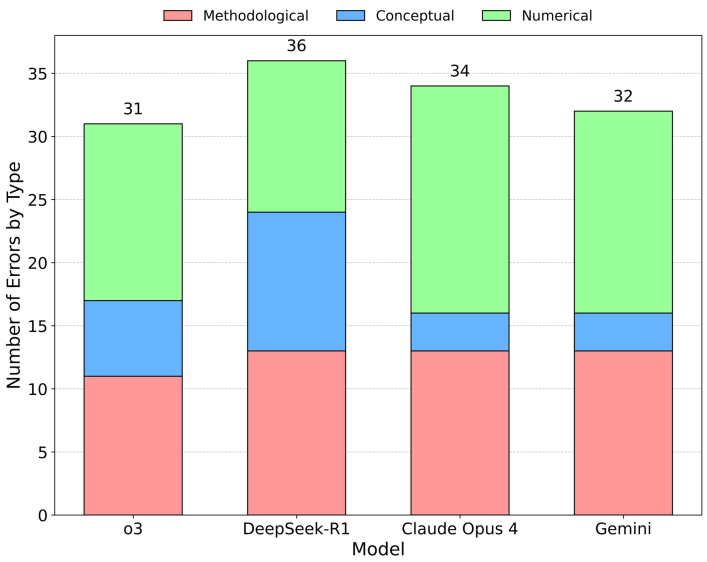


Fig. 7. Distribution of error types by model for incorrectly solved tasks

4.2. Statistical Significance of Results

First, we examine whether there are statistically significant differences in the average time each of the four models required to solve the 50 tasks. Since the assumptions of normality and homogeneity of distribution for the observed variables were not violated,

we employed an analysis of variance (ANOVA) test [60]. An ANOVA confirmed that the differences are statistically significant ($F = 79.247, df = 3, p < 0.0005$). The value of Partial Eta Squared was 0.847, indicating an exceptionally large effect size, confirming the major differences in solution generation times. The descriptive statistics for task completion times are presented in Table 4.

Table 4. Descriptive statistics for task completion times (in seconds) by AI model

AI tool	o3	DeepSeek-R1	Claude Opus 4	Gemini
Mean	44.10	289.42	11.00	53.46
SD	37.58	223.88	3.07	20.02
Minimum	3	20	2	21
Maximum	202	943	80	131

The Bonferroni post-hoc comparisons [61], detailed in Table 5, reveal that the time differences were statistically significant in five out of the six pairs. The only non-significant difference was between the o3 and Gemini models. Regarding model stability, Claude Opus 4 not only had the lowest average time but also the smallest standard deviation, while DeepSeek-R1's was substantially larger than the others.

Table 5. p-values for the Bonferroni post-hoc test for the time AI tools needed for solving bond tasks

AI tool	o3	DeepSeek-R1	Claude Opus 4	Gemini
o3				
DeepSeek-R1	$<$ **0.0005**			
Claude Opus 4	$<$ **0.0005**	$<$ **0.0005**		
Gemini	0.443	$<$ **0.0005**	$<$ **0.0005**	

The primary results of the study concern the differences in the average scores achieved by each model, with each problem scored on a scale from a minimum of 0 to a maximum of 100 points. The descriptive statistics for these scores are presented in Table 6. An ANOVA test confirmed that the differences in the mean scores observed in Table 6 are statistically significant ($F = 4.379, df = 3, p = 0.008$). The Partial Eta Squared value of 0.218, although considerably smaller than the result for response times, still indicates a large effect size (according to Cohen), suggesting the choice of AI model has a substantial impact on the average score.

Out of six total pairwise comparisons, statistically significant differences were found in two pairs of AI models. Specifically, based on the p-values in Table 7 and the mean scores in Table 6, we conclude that o3 and Gemini performed significantly better than DeepSeek-R1. The performance difference between o3 and Gemini was not significant. Similarly, the results of Claude Opus 4 did not differ significantly from any of the other three models. In this analysis, the standard deviations varied less than the solution times,

Table 6. Descriptive statistics for the points AI models achieved for solving bond tasks

AI tool	o3	DeepSeek-R1	Claude Opus 4	Gemini
Mean	80.90	70.60	74.90	78.90
SD	24.51	29.96	26.72	26.79
Minimum	15	10	10	10
Maximum	100	100	100	100

but it is notable that the model with the highest mean score, o3, also had the smallest standard deviation.

Table 7. p-values for the Bonferroni post-hoc test for the points AI models achieved for solving bond tasks

AI tool	o3	DeepSeek-R1	Claude Opus 4	Gemini
o3				
DeepSeek-R1	**0.004**			
Claude Opus 4	0.206	0.565		
Gemini	0.998	**0.034**	0.764	

The difficulty of the task had a statistically significant impact on the performance of all four models: o3 ($F = 17.217, p < 0.0005, df = (2, 47), \eta^2 = 0.423$), DeepSeek-R1 ($F = 12.482, p < 0.0005, df = (2, 47), \eta^2 = 0.347$), Claude Opus 4 ($F = 17.035, p < 0.0005, df = (2, 47), \eta^2 = 0.420$), and Gemini ($F = 11.314, p < 0.0005, df = (2, 47), \eta^2 = 0.325$). The effect size was very large for all four models, with the strongest effects observed for o3 and Claude Opus 4.

For the o3 model, performance differed significantly between easy and hard tasks ($p < 0.0005$) and between medium and hard tasks ($p < 0.0005$), however, the difference between easy and medium tasks was not statistically significant ($p = 0.069$). In contrast, the other three models showed significant performance differences across all pairs of difficulty levels: easy vs. medium ($p = 0.001$ for DeepSeek-R1, $p = 0.004$ for Claude Opus 4, $p = 0.033$ for Gemini), medium vs. hard ($p = 0.004$ for DeepSeek-R1, $p = 0.001$ for Claude Opus 4, $p = 0.04$ for Gemini), and easy vs. hard ($p < 0.0005$ for all three). This finding underscores the substantial impact of task complexity on model performance and highlights the need for specialized training, particularly for more complex problems.

Finally, we analyzed performance differences between three task types: practical calculations, theoretical tasks, and real-life applications. Our analysis involved two stages: comparing the models against each other within each category and examining each model's individual performance within each of the three groups. Within the first group, practical calculations that involve the application of bond-related formulas, statistically significant differences were found between the four AI models ($F = 3.294, df = 3, p = 0.043, \eta^2 = 0.34$). In this category, o3 achieved the highest score ($M = 90.00, SD = 11.85$), while DeepSeek-R1 recorded the lowest ($M = 76.59, SD = 24.71$).

For theoretical tasks, which require understanding relationships and interpreting special cases with formal mathematical methods, the performance differences were not sta-

tistically significant ($F = 1.00, df = 3, p = 0.479, \eta^2 = 0.43$). This group notably contains the fewest tasks (only 7). The mean score difference between the best performer, DeepSeek-R1 ($M = 76.43, SD = 34.96$), and the lowest performer, Gemini ($M = 71.43, SD = 38.48$), was exactly 5 points.

Similarly, a comparison of mean scores on real-life, context-rich problems revealed no statistically significant differences ($F = 1.81, df = 3, p = 0.182, \eta^2 = 0.23$). DeepSeek-R1 again had the lowest score in this group ($M = 62.38, SD = 32.69$), while o3 performed the best ($M = 73.81, SD = 28.01$).

Thus, differences between the models reached statistical significance only in the practical calculations category; the absence of significant differences in the remaining two categories may be attributable to the smaller number of tasks within each group. However, when the results were aggregated across all categories, the overall differences in model performance were statistically significant.

The analysis of each tool's performance across the task categories revealed that there were no statistically significant internal differences for: o3 ($F = 3.366, df = 2, p = 0.065, \eta^2 = 0.33$); DeepSeek-R1 ($F = 1.384, df = 2, p = 0.261, \eta^2 = 0.05$); Claude Opus 4 ($F = 3.513, df = 2, p = 0.057, \eta^2 = 0.33$); and Gemini ($F = 3.391, df = 2, p = 0.064, \eta^2 = 0.33$). Although the eta squared values indicate relatively large effect sizes for several tools (all except DeepSeek-R1), the observed differences did not reach statistical significance, which may again be attributed to the small sample size.

4.3. Case Study: Limitations of LLMs on theoretical financial mathematics tasks

Despite generally strong overall results, there were specific problems on which most, or even all, models performed poorly.

The following example illustrates this type of task:

The future value of the selling price of a bond at the maturity date of a liability that occurs strictly before the bond's own maturity is a convex function of the yield to maturity.

From a theoretical perspective, statements of this type are closely related to the analysis of interest-rate risk and the concept of portfolio immunization, so it is particularly important that they are established by a rigorous mathematical proof rather than by intuition alone.

The DeepSeek-R1 model identified the main idea behind the problem, but its analysis was entirely restricted to the special case of a zero-coupon bond. Although it hinted at the right proof strategy (expressing the bond price as a function of the yield and examining its second derivative), it did not extend the reasoning to the general coupon case, nor did it explicitly incorporate the future value of the selling price at the earlier liability date. The resulting arguments were not carried through to completion: no clear expression for the second derivative was given, nor was its non-negativity established in the general setting. Consequently, DeepSeek-R1 did not produce a full, rigorous proof of the statement.

For Claude Opus 4 and Gemini, the initial line of reasoning was considerably closer to what is mathematically required. Both systems correctly recognized that, in order to prove convexity, one needs to: (i) explicitly write down the future value of the selling price as a function of the yield to maturity, taking into account all coupon payments up to the selling date, and (ii) compute the first and second derivatives of this function with respect to the yield and analyze the sign of the second derivative. However, despite this largely correct proof blueprint, the resulting mathematical expressions were incomplete

or incomprehensible, several steps were skipped, and the final conclusion that the function is indeed convex relied more on intuitive arguments (e.g., references to the typical shape of price to yield curves) than on a fully workedout computation. Thus, in both cases, the proof remained conceptually promising but formally insufficient.

The o3 model performed the worst on this test. Instead of starting from the standard theory of bond pricing and its dependence on the yield, it attempted to construct counterexamples and to disprove the statement. In doing so, it misinterpreted key aspects of the problem and its parameters, in particular, the structure of cash flows and the timing of the sale, so that the proposed counterexamples were based on inconsistent or incomplete models. As a result, o3 produced the erroneous suggestion where the statement may fail in general, even though, under the properly specified model, it is in fact true. This outcome clearly illustrates the risk of relying uncritically on generative models in situations where subtle mathematical structure is crucial.

In summary, in this particular problem none of the analyzed AI tools produced a fully correct and rigorous mathematical proof of the convexity of the future value of the bond selling price as a function of the yield to maturity in the general setting of a coupon bond sold before maturity. DeepSeek-R1 provided a useful partial insight through the special case of a zero-coupon bond, while Claude Opus 4 and Gemini outlined a reasonably sound conceptual framework without fully carrying out all formal derivations. The o3 model, on the other hand, explored a line of reasoning aimed at constructing counterexamples and thus suggested that the statement might fail in general, which in this setting reflects a misinterpretation of some of the problem assumptions. Overall, this case study indicates that current generative AI systems can offer valuable support for brainstorming proof strategies and recalling key ideas, but that, in specialized areas such as financial mathematics, they should still be used with caution when fully rigorous and complete proofs are required.

4.4. Discussion

The results of this study indicate that modern large language models can solve problems in financial mathematics, particularly within the bond domain. However, this analysis shows that these capabilities vary by task type, difficulty, and model, which is consistent with previous findings that mathematical reasoning remains a challenge for generative AI systems [1], [7], [31].

The o3 model achieved the highest average accuracy, aligning with earlier work on the dominance of GPT-4 architectures for moderately complex mathematics tasks [30]. Gemini demonstrated comparable performance, especially on practical and real-life tasks, supporting the findings that multimodal approaches assist problem-oriented tasks [36]. In contrast, DeepSeek-R1, despite its optimization for mathematics, showed lower accuracy, suggesting that proprietary models still outperform open-source systems on complex quantitative problems [17]. However, in our experiments, DeepSeek-R1 achieved the strongest performance on theoretical tasks. Our results also align with broader studies in financial mathematics, such as [41], which find that models perform well on standard tasks, but struggle with those that are analytically more demanding.

Performance differences among models became more pronounced as the task difficulty increased. All models achieved high accuracy on simpler tasks, consistent with results on datasets like GSM8K [7]. In contrast, performance declined on harder tasks,

confirming that multi-step reasoning and abstract inference remain difficult [1]. This suggests that current LLMs do not yet fully replicate the expert knowledge required for precise mathematical interpretation in financial contexts.

An analysis by task type revealed that all models performed best on practical tasks, likely due to their strength in applying known formulas. Theoretical tasks, which require abstract modeling and general conclusions, proved to be more demanding, while the weakest results of the tools were observed on real-life tasks.

The error analysis identified similar failure patterns across all models, with a prevalence of numerical errors and a notable number of methodological mistakes. This points to the need to improve internal reasoning and verification. Conceptual errors, though less common, show that a model's understanding of a problem's core principles can be limited even with correct technical execution, a finding consistent with [31].

The timing analysis revealed statistically significant differences in efficiency. DeepSeek-R1 had the longest solution time by a wide margin, whereas Claude Opus 4 was the fastest and most stable. The only non-significant time difference was between o3 and Gemini. These results indicate that the response speed is a separate evaluation aspect that is not directly correlated with the accuracy.

This study has limitations, including a relatively small number of tasks per category, which can affect statistical power. In particular, the distribution of tasks across categories is uneven, with only 7 theoretical tasks compared to 22 practical and 21 real life tasks. This imbalance is partly a consequence of a deliberate design decision: the theoretical corpus related to bonds is inherently narrow, covering a limited number of fundamental properties (e.g., convexity of the price–yield relation, asymptotic behavior of duration, the structure of spot and forward rates, and continuous compounding). Adding further theoretical tasks within this scope would largely result in restatements or minor variations of the same underlying propositions, rather than genuinely new problems. Nevertheless, the small sample size reduces the statistical power of the sub-category analyses, so the results reported for the theoretical category should be interpreted as exploratory rather than confirmatory. In a follow-up version of the benchmark, we plan to release a larger and more balanced dataset in which the theoretical component is expanded by broadening its scope (e.g., to other fixed-income instruments, term-structure theory, and immunization theorems), thereby avoiding redundancy while enabling more robust parametric inference at the category level.

Furthermore, all tasks were presented in Serbian, which may affect the performance of models trained predominantly on English-language data. An earlier study compared generative AI tools on identical mathematical tasks presented in Serbian and English and reported broadly comparable performance across the two languages in several cases; for o1-preview, accuracy on text-based tasks was 86% in Serbian and 84% in English [13]. This relatively small difference suggests that the language effect may be limited for advanced reasoning capable models. However, the earlier study concerned competition mathematics and evaluated a different set of tools, while the present study did not include a matched English language control. Consequently, the effect of language cannot be isolated from other sources of error, nor can the proportion of errors attributable to the Serbian language be quantified. A controlled bilingual evaluation of the present benchmark is therefore an important direction for future work.

Two further limitations concern the scope of the evaluation. First, the study covers four flagship models, three of which are proprietary. Recent work evaluated QLoRA fine-tuning of Mistral-7B-Instruct-v0.3 on Serbian high-school mathematics competition tasks [62]. Although the fine-tuned model did not outperform the base model and the study concerned competition mathematics rather than financial mathematics, it represents an initial step toward determining how smaller open-weight models can be adapted for specialized mathematical reasoning in Serbian. This research direction is ongoing and will be extended to the direct evaluation of smaller open-weight models and their distilled variants on the bond-focused benchmark introduced here. Second, while the risk that models merely reproduce memorized solutions is mitigated by the design of the benchmark, since the tasks were newly constructed for this study, were not publicly available prior to it, and are formulated in Serbian, our evaluation does not formally separate genuine reasoning from retrieval of memorized patterns, perturbation-based analyses that could make this distinction explicit are left for future work.

The results should also be interpreted in light of the inherent mismatch between the probabilistic nature of LLM architectures and the deterministic requirements of rigorous financial mathematics. Since these models generate statistically plausible token sequences rather than formally verified derivations, occasional errors are not incidental defects but a structural property of the approach. This is most visible in our case study, where a model confidently pursued counterexamples to a statement that is in fact true. Rather than diminishing the contribution of this research, this mismatch defines it: precisely because no formal guarantees exist, empirically mapping the reliability frontier of these systems in a well-defined domain is a necessary step. Our results delineate this frontier concretely. On formula driven practical computations, current models achieve accuracy high enough to serve as effective assistants, provided their outputs are verified; on complex real-life scenarios and formal proofs, they should be treated as sources of ideas and draft strategies rather than as authorities. The error taxonomy further specifies what kind of failures a user should anticipate in each setting, from benign numerical slips to methodological misdirection. In educational and professional practice, this positions LLMs as tools whose value depends on the user's ability to critically evaluate their output, and our benchmark provides an empirical basis for calibrating exactly that critical stance.

In conclusion, while general-purpose LLMs can solve many financial mathematics tasks, their weaknesses on complex tasks highlight the need for further development and potentially for specialized models. This aligns with the trend towards finance-focused systems like *BloombergGPT* [46], which suggests that steering AI toward narrower domains is a necessary step for future progress.

5. Conclusion

This paper presents a systematic evaluation of the capabilities of modern large language models in solving financial mathematics problems, with a specific focus on the bond domain. To this end, we developed a dedicated set of 50 tasks spanning practical, theoretical, and real-world applications to test four models: o3, Claude Opus 4, DeepSeek-R1, and Gemini. The models were assessed using a multi-layered framework that combines solution accuracy, error analysis, and response time, allowing us to study both the quality and efficiency of model behaviour.

The first aspect of our evaluation focused on accuracy and error types. Confirming our statistical analysis, the o3 model achieved the highest mean accuracy (80.9%), while DeepSeek-R1 had the lowest (70.6%). The error analysis revealed that numerical errors were the most common across all models, while theoretical tasks tended to expose more conceptual weaknesses.

The second aspect of our analysis addressed comparative performance by type and difficulty of the task. As verified, all models performed well on easy tasks (over 90% accuracy), but their performance dropped below 55% on hard tasks. Practical tasks resulted in the highest scores, while theoretical and real-life problems proved more challenging. Our statistical analysis (ANOVA) confirmed significant differences among the models in both accuracy and time required to generate the solution. As shown in the results, Claude Opus 4 was the fastest model (11.0 seconds), while DeepSeek-R1 was the slowest (289.4 seconds).

Beyond these empirical findings, an important contribution of this work lies in the evaluation and benchmarking perspective. We introduced a bond-focused benchmark specification that includes a carefully curated task set, explicit categorization by type (practical, real-life, theoretical) and difficulty (easy, medium, hard), and a detailed scoring section with an explicit taxonomy of methodological, numerical, and conceptual errors. Together with the measurement of solution times and the use of ANOVA and Bonferroni post-hoc tests, this yields a standardized, model-agnostic evaluation pipeline that can be reused for subsequent generations of LLMs in quantitative finance and related domains.

The findings indicate that while current general-purpose LLMs can solve many financial mathematics tasks, more complex analytical and theoretical tasks involve further improvement or the use of specialized models. Our constructed dataset and evaluation methodology, including detailed statistical analysis, provide a valuable contribution to future research on the development and application of LLMs in both educational, financial practice, and professional use. To support reproducibility and future benchmarking, the complete benchmark, including all 50 tasks in their original form, the standardized prompt, the full responses generated by each of the four models, and the detailed per-task scores, error classifications, and solution times, is publicly available at `https://github.com/marinasvicevic/bond-llm-benchmark`.

Acknowledgement. This work was supported by the Serbian Ministry of Education, Science and Technological Development (Agreement No. 451-03-34/2026-03/200122).

References

1. Frieder, S., Pinchetti, L., Griffiths, R.R., Salvatori, T., Lukasiewicz, T., Petersen, P., Berner, J.: Mathematical capabilities of chatgpt. Advances in neural information processing systems 36, 27699–27744 (2023)
2. Rane, N.: Enhancing mathematical capabilities through chatgpt and similar generative artificial intelligence: Roles and challenges in solving mathematical problems. Available at SSRN 4603237 (2023)
3. Broverman, S.A.: Mathematics of investment and credit. Actex Publications (2010)
4. Fabozzi, F.J.: Duration, convexity, and other bond risk measures. John Wiley & Sons (1999)

5. Mondello, E.: Duration and convexity. In: Applied Fundamentals in Finance: Portfolio Management and Investments, pp. 413–447. Springer (2023)
6. Ahn, J., Verma, R., Lou, R., Liu, D., Zhang, R., Yin, W.: Large language models for mathematical reasoning: Progresses and challenges. arXiv preprint arXiv:2402.00157 (2024)
7. Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., et al.: Solving quantitative reasoning problems with language models. Advances in Neural Information Processing Systems 35, 3843–3857 (2022)
8. Moon, I.H.: Performance comparison of large language models on advanced calculus problems. arXiv preprint arXiv:2503.03960 (2025)
9. Wei, X.: Evaluating chatgpt-4 and chatgpt-4o: performance insights from naep mathematics problem solving. In: Frontiers in Education. vol. 9, p. 1452570. Frontiers Media SA (2024)
10. Zhang, J., Li, Z., Zhang, M., Yin, F., Liu, C., Moshfeghi, Y.: Geoeval: benchmark for evaluating llms and multi-modal models on geometry problem-solving. arXiv preprint arXiv:2402.10104 (2024)
11. Milenković, A., Vučićević, N., Svičević, M.: Evaluating open ai tools o1 and o3-mini in solving high school problems from serbian national mathematics competition. The Teaching of Mathematics 28(1), 14–29 (2025)
12. Stankovic, M., Milenkovic, A., Svicevic, M., Vucicevic, N.: Success of ai math solver tool in solving non-standard mathematics competition problems. Facta Universitatis, Series: Teaching, Learning and Teacher Education (2025)
13. Svičević, M., Milenković, A., Vučićević, N., Stanković, M.: Evaluating the success of ai tools in supporting student performance in mathematical kangaroo competition. Computer Applications in Engineering Education 33(4), e70063 (2025)
14. Vučićević, N., Svičević, M., Milenković, A.: Challenging deepseek-r1 with serbian high school math competition problems. In: Sinteza 2025 - International Scientific Conference on Information Technology, Computer Science, and Data Science. pp. 274–280 (2025)
15. OpenAI: Introducing o3 and o4-mini (2025), https://openai.com/index/introducing-o3-and-o4-mini/ (Accessed: 2025-06-28)
16. Anthropic: Introducing claude 4: Claude opus 4 and claude sonnet 4. https://www.anthropic.com/news/claude-4 (May 2025), accessed: 2025-06-28
17. Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., et al.: DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
18. DeepMind, G.: Introducing gemini 2.5 pro: Our most intelligent ai model. https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/ (Mar 2025), accessed: 2025-06-28
19. Łodzikowski, K., Foltz, P.W., Behrens, J.T.: Generative ai and its educational implications. In: Trust and Inclusion in AI-Mediated Education: Where Human Learning Meets Learning Machines, pp. 35–57. Springer (2024)
20. Haque, M.A.: Llms: A game-changer for software engineers? BenchCouncil Transactions on Benchmarks, Standards and Evaluations p. 100204 (2025)
21. Pešić, D., Janičić, M.V., Mišić, M., Protić, J.: Assessing chatgpt for algorithm time complexity education. In: World Congress in Computer Science, Computer Engineering & Applied Computing. pp. 182–191. Springer (2024)
22. Saab, K., Tu, T., Weng, W.H., Tanno, R., Stutz, D., Wulczyn, E., Zhang, F., Strother, T., Park, C., Vedadi, E., et al.: Capabilities of gemini models in medicine. arXiv preprint arXiv:2404.18416 (2024)
23. Yang, L., Xu, S., Sellergren, A., Kohlberger, T., Zhou, Y., Ktena, I., Kiraly, A., Ahmed, F., Hormozdiari, F., Jaroensri, T., et al.: Advancing multimodal medical capabilities of gemini. arXiv preprint arXiv:2405.03162 (2024)

24. Bennis, I., Mouwafaq, S.: Advancing ai-driven thematic analysis in qualitative research: a comparative study of nine generative models on cutaneous leishmaniasis data. BMC Medical Informatics and Decision Making 25(1), 1–14 (2025)
25. Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D.M., Aletras, N.: LexGLUE: A benchmark dataset for legal language understanding in English. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 4310–4330 (2022)
26. Guha, N., Nyarko, J., Ho, D.E., Ré, C., Chilton, A., Narayana, A., Chohlas-Wood, A., Peters, A., Waldon, B., Rockmore, D.N., et al.: LegalBench: A collaboratively built benchmark for measuring legal reasoning in large language models. Advances in Neural Information Processing Systems 36, 44123–44279 (2023), arXiv:2308.11462
27. Lee, J., Stevens, N., Han, S.C.: Large language models in finance (finllms). Neural Computing and Applications pp. 1–15 (2025)
28. Li, Y., Wang, S., Ding, H., Chen, H.: Large language models in finance: A survey. In: Proceedings of the fourth ACM international conference on AI in finance. pp. 374–382 (2023)
29. Nie, Y., Kong, Y., Dong, X., Mulvey, J.M., Poor, H.V., Wen, Q., Zohren, S.: A survey of large language models for financial applications: Progress, prospects and challenges. arXiv preprint arXiv:2406.11903 (2024)
30. Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., et al.: Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712 (2023)
31. Collins, K.M., Jiang, A.Q., Frieder, S., Wong, L., Zilka, M., Bhatt, U., Lukasiewicz, T., Wu, Y., Tenenbaum, J.B., Hart, W., et al.: Evaluating language models for mathematics through interactions. Proceedings of the National Academy of Sciences 121(24), e2318124121 (2024)
32. Lin, Y.F., Fu-Yu Yang, E., Wu, J.S., Yeh, C.Y., Liao, C.Y., Chan, T.W.: Enhancing students' authentic mathematical problem-solving skills and confidence through error analysis of gpt-4 solutions. Research & Practice in Technology Enhanced Learning 20 (2025)
33. Rhomrasi, L., Ahsini, Y., Igualde-Sáez, A., Vinuesa, R., Hoyas, S., García-Sabater, J.P., Fullana-i Alfonso, M.J., Conejero, J.A.: LLM performance on mathematical reasoning in Catalan language. Results in Engineering 25, 104366 (2025)
34. Igualde-Sáez, A., Rhomrasi, L., Ahsini, Y., et al.: Evaluating visual mathematics in multimodal LLMs: a multilingual benchmark based on the Kangaroo tests. The Journal of Supercomputing 82, 159 (2026)
35. Oh, S., Yoon, J., Chung, Y., Cho, Y., Shim, H., Kwon, O.N.: Analysis of generative ai's mathematical problem-solving performance: Focusing on chatgpt 4, claude 3 opus, and gemini advanced. The Mathematical Education 63(3), 549–571 (2024)
36. Vidal, J.: Evaluation of the performance of state-of-the-art large language models (llms) in solving math word problems. Available at SSRN 4902960 (2024)
37. Luo, H., Sun, Q., Xu, C., Zhao, P., Lou, J., Tao, C., Geng, X., Lin, Q., Chen, S., Zhang, D.: Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct. arXiv preprint arXiv:2308.09583 (2023)
38. Li, C., Wang, W., Hu, J., Wei, Y., Zheng, N., Hu, H., Zhang, Z., Peng, H.: Common 7b language models already possess strong math capabilities. arXiv preprint arXiv:2403.04706 (2024)
39. Lu, Z., Zhou, A., Ren, H., Wang, K., Shi, W., Pan, J., Zhan, M., Li, H.: Mathgenie: Generating synthetic data with question back-translation for enhancing mathematical reasoning of llms. arXiv preprint arXiv:2402.16352 (2024)
40. Lu, P., Bansal, H., Xia, T., Liu, J., Li, C., Hajishirzi, H., Cheng, H., Chang, K.W., Galley, M., Gao, J.: Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. arXiv preprint arXiv:2310.02255 (2023)
41. Liu, L.X., Sun, Z., Xu, K., Chen, C.: Ai-driven financial analysis: Exploring chatgpt's capabilities and challenges. International Journal of Financial Studies 12(3), 60 (2024)

42. Zhao, Y., Liu, H., Long, Y., Zhang, R., Zhao, C., Cohan, A.: Financemath: Knowledge-intensive math reasoning in finance domains. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 12841–12858 (2024)
43. Shirota, Y., Kuwana, A., Chakraborty, B.: Generative ai in financial mathematics problem-solving: Integration of linguistic and symbolic processing functions. In: International Conference on Artificial Intelligence in Education Technology. pp. 341–355. Springer (2024)
44. Srivastava, P., Malik, M., Gupta, V., Ganu, T., Roth, D.: Evaluating llms' mathematical reasoning in financial document question answering. arXiv preprint arXiv:2402.11194 (2024)
45. Vučićević, N., Svičević, M., Milenković, A., Pavković, M., Protić, J.: Chatgpt 5.2 i finansijska matematika: analiza tačnosti i obrazovne vrednosti generisanih rešenja. In: 32. skup TRENDOVI RAZVOJA: TRANSFORMACIJA VISOKOG OBRAZOVANJA U NOVIM TEHNOLOŠKIM OKVIRIMA: Kvalitet, održivost i institucionalni razvoj. pp. 43–46. Fakultet tehničkih nauka Univerziteta u Novom Sadu (2026), http://www.trend.uns.ac.rs/TREND2026-ZBORNIK-RADOVA.pdf
46. Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., Kambadur, P., Rosenberg, D., Mann, G.: Bloomberggpt: A large language model for finance. arXiv preprint arXiv:2303.17564 (2023)
47. Zhang, Z., Cao, Y., Liao, L.: Xfinbench: Benchmarking llms in complex financial problem solving and reasoning. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 8715–8758 (2025)
48. Campolieti, G., Makarov, R.N.: Financial mathematics: a comprehensive treatment. Chapman and Hall/CRC (2018)
49. Fabozzi, F.J., Fabozzi, F.A.: Bond markets, analysis, and strategies. MIT Press (2021)
50. Hastings, K.J.: Introduction to financial mathematics. Chapman and Hall/CRC (2024)
51. Jarrow, R.: Modeling fixed income securities and interest rate options. Chapman and Hall/CRC (2019)
52. Wilders, R.J.: Financial mathematics for actuarial science: the theory of interest. CRC press (2020)
53. Wu, L.: Interest rate modeling: Theory and practice. Chapman and Hall/CRC (2024)
54. OpenAI: OpenAI o3 and o4-mini system card (2025), https://openai.com/index/o3-o4-mini-system-card, accessed: July 2025
55. DeepMind, G.: Introducing gemini (2025), https://deepmind.google/models/gemini/ (Accessed: 2025-06-28)
56. Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
57. Koo, T.K., Li, M.Y.: A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine 15(2), 155–163 (2016)
58. Tejankar, A., Van Eaton, W.: Self-distilling DeepSeek-R1 with turbo speculation - 2x inference (Feb 2025), https://predibase.com/blog/deepseek-r1-self-distillation-turbo-speculation, accessed: 2025-06-28
59. MimicPC: Why deepseek server is busy? 14 effective ways to fix the error (2 2025), https://www.mimicpc.com/learn/why-deepseek-show-server-busy-how-to-fix
60. Montgomery, D.C.: Design and analysis of experiments. John Wiley & Sons, 9th edn. (2017)
61. Dunn, O.J.: Multiple comparisons among means. Journal of the American Statistical Association 56(293), 52–64 (Mar 1961)
62. Pavković, M., Svičević, M., Milutinović, A., Vučićević, N., Milenković, A.: QLoRA fine-tuning of Mistral-7B for serbian high school mathematics competition tasks. In: Sinteza 2026 - International Scientific Conference on Information Technology, Computer Science, and Data Science. pp. 129–136 (2026)

Nemanja Vučićević, PhD, is a Teaching Assistant at the Department of Mathematics and Informatics, Faculty of Science, University of Kragujevac, Serbia. His research interests include stochastic optimization, operations research, mathematical modeling, generalized convexity, integral inequalities, machine learning, and artificial intelligence. His scientific work focuses on the development and theoretical analysis of optimization algorithms and their application to practical problems. He has published in international scientific journals and presented his work at national and international conferences, with a particular interest in the application of artificial intelligence to mathematics education.

Marina Svičević is an Associate Professor at the Faculty of Science, University of Kragujevac, Serbia. Her research interests include computer modeling and simulation, machine learning, and data analysis, as well as the methodology of teaching informatics. She applies machine learning techniques in education, medical physics, big data processing, image analysis, etc. Marina has published numerous scientific papers and actively participates in national and international conferences, contributing to advancements in computer science and its applications.

Aleksandar Milenković is an Assistant Professor at the Faculty of Science, University of Kragujevac, Serbia. His research interests include mathematics education at all three levels (elementary, high school, and university), mathematics teacher education, gifted students in mathematics, mathematics competitions, and the use of AI in mathematics education. He has published numerous scientific papers and has presented his findings to the scientific and professional community at various international and national conferences.

Miloš Pavković is an Assistant Professor at Singidunum University, Serbia. His research interests include applied artificial intelligence, machine learning, deep neural networks, and AI system integration. He teaches courses related to applied AI and machine learning, with a strong focus on real-world applications. With over a decade of hands-on industry experience, he actively collaborates on projects involving AI deployment, data engineering, and model optimization. Miloš regularly participates in academic and professional initiatives that bridge the gap between research and industry needs.

Lazar Stošić, PhD in Computer Science, is a Professor and Dean of the Faculty of Informatics and Computer Science, University Union—Nikola Tesla, Belgrade, Serbia, as well as a Leading Researcher at Don State Technical University in Rostov-on-Don, Russian Federation. He is also the President of the Association for the Development of Science, Technology, and Education in Vranje, Serbia, and a Senior Researcher at the Institute for Management and Knowledge in Skopje, North Macedonia. Additionally, he serves as the Editor-in-Chief of the *International Journal of Cognitive Research in Science, Engineering, and Education*.

Jelica Protić received the PhD degree in electrical engineering from the University of Belgrade in 1999. She is currently a Full Professor and the Head of the Department of Computer Engineering and Informatics with the University of Belgrade, the School of Electrical Engineering. With Milo Tomasevic and Veljko Milutinovic, she co-authored *Distributed Shared Memory: Concepts and Systems* (IEEE CS Press, 1997) and presented numerous pre-conference tutorials on this subject. She has long term experience in teaching a diversity of courses in programming languages, as well as the development of various educational software tools. Her research interests include distributed systems, consis-

tency models, complex networks, and all aspects of computer-based quantitative performance analysis and modeling.

Received: March 30, 2026; Accepted: September 15, 2026.