ГОДИШНИК НА СОФИЙСКИЯ УНИВЕРСИТЕТ „СВ. КЛИМЕНТ ОХРИДСКИ“
ФАКУЛТЕТ ПО МАТЕМАТИКА И ИНФОРМАТИКА
Том 112, 2025

ANNUAL OF SOFIA UNIVERSITY “ST. KLIMENT OHRIDSKI”
FACULTY OF MATHEMATICS AND INFORMATICS
Volume 112, 2025

LARGE LANGUAGE MODELS: A TOOL FOR SOLVING MATHEMATICAL PROBLEMS IN HIGH SCHOOL

Ralitza Stamenkova

Mathematics learning resources have evolved from static textbooks to collaborative online forums and conversational artificial intelligence (AI) tools. This evolution reflects students' ongoing demand for clarity, adaptability, and accessible support. While traditional textbooks offered privacy, they limited flexibility. Online forums, such as Yahoo! Answers, Math Stack Exchange, and Mathematika.bg, enabled collaborative problem-solving, but they required public participation. Large language models (LLMs), including ChatGPT, now offer private, adaptive “comfort mode” interactions, combining the autonomy of self-study with the responsiveness of a personal tutor.

The potential of LLMs to support mathematics education in Bulgaria is examined through a dual approach in this study: (1) eleven models solving problems from the Bulgarian National External Assessment (NEA) for 10th grade are empirically evaluated, and (2) practicing mathematics teachers enrolled in a Master program are qualitatively observed. Model performance was evaluated based on accuracy, methodological alignment with the national curriculum, and linguistic appropriateness.

The findings indicate that, although several models, such as Mistral (22B) and DeepSeek R1, achieved perfect accuracy, they often used solution strategies that deviated from national standards. Locally fine-tuned models (e.g., BgGPT) demonstrated stronger curriculum alignment and the use of precise Bulgarian mathematical terminology. Teacher feedback revealed recognition of AI's potential for personalized student support, as well as caution toward integration, reflecting a preference to retain creative and methodological control. The study concludes that three conditions are necessary for successful AI integration in mathematics education: mathematical accuracy, adherence to curriculum-specific methods, and linguistically precise explanations. Large language models (LLMs) can complement, but not replace, the teacher's role when deliberately embedded in the continuum of educational resources, from books to forums to conversational AI.

Keywords: mathematics education, large language models, ChatGPT, AI in education, curriculum alignment, teacher perceptions, personalised learning

2020 Mathematics Subject Classification: 97D40

DOI: 10.60063/GSU.FMI.112.165-183

1. Introduction

The evolution of digital technologies and the emergence of AI have profoundly transformed the landscape of mathematics education, leading to "a paradigm shift in how education is delivered ... from static content to dynamic, personalized learning environments" [3,9]. From the limited print-based resources of the late 20th century to the vast digital ecosystem of today, students and educators have gained access to a wide array of tools, ranging from educational videos and interactive platforms to AI-powered assistants. However, this transformation has also introduced challenges, including inconsistencies in content quality, unequal access, and shifting expectations regarding the roles of teachers and technology in the learning process. As noted by Ali et al. [2], while ChatGPT offers real-time, adaptive support, its use also raises concerns about overreliance, academic dishonesty, and uneven student autonomy.

According to Dey [7], AI is "reshaping educational experiences, enhancing learning outcomes, and addressing long-standing challenges in the education sector", offering significant potential for personalization and innovation across disciplines. In this context, the present study explores the intersection of AI and mathematics education, with a particular focus on the perceptions and behaviours of practicing teachers enrolled in the master's program "Innovation and Multidisciplinarity in Teaching Mathematics, Computer Modelling, and Information Technologies" during the winter semester of 2024/2025. The cohort under observation is currently enrolled in secondary education programs and is being observed as part of two graduate-level courses: "Educational Methods and Tools and Innovative Approaches" and "Project-Based Learning." The objective of this study is to investigate how these experienced educators interact with emerging technologies, particularly LLM, in the design of learning tasks and pedagogical strategies.

The study also considers the evolving needs of school-aged learners, who increasingly seek diverse, personalized, and on-demand forms of academic support. According to Ayeni et al. [3], AI enables "individualized learning trajectories that dynamically adapt to student needs," making it a powerful tool for inclusivity and motivation. The way students seek assistance in mathematics has evolved over time. Initially, students predominantly relied on textbook-based study methods. Subsequently, there has been a shift toward online peer collaboration, which has emerged as an increasingly prevalent approach. More recently, there has been a notable increase in the use of AI-powered platforms as a tool for seeking academic assistance. This progression is exemplified by the transition from early web forums, such as Yahoo! Answers, to structured question-and-answer communities, including Math Stack Exchange and `Matematika.bg`. The current era is characterized by the emergence of platforms such as ChatGPT, Le Chat (Mistral AI), and DeepSeek, which offer immediate solutions and, in some cases, curriculum-aligned explanations. These developments have diversified the forms of academic support available to learners, offering both new opportunities and new challenges for integration into formal education.

In light of these considerations, the present study aims to address three interconnected objectives:

(1) To examine how practicing teachers engage with AI tools in their own pedagogical design.
(2) To explore how students' historical and present-day needs shape their interactions with online mathematical resources.
(3) To analyse the potential and limitations of chatbot-based AI systems for delivering curriculum-aligned mathematical explanations and solutions.

To achieve these aims, the present study evaluates the capacity of contemporary language models to solve mathematics problems from the Bulgarian National External Assessment (NEA) for 10th grade. The evaluation focuses on three factors: accuracy, methodological alignment with the national curriculum, and linguistic appropriateness. The empirical results are complemented by qualitative insights from practicing teachers, offering a nuanced understanding of both the opportunities and constraints of integrating AI into mathematics teaching and learning. The central question guiding this research is how these technologies can be meaningfully integrated into the curriculum to meet authentic educational needs while preserving pedagogical integrity.

2. Observations

2.1. Master of education student

The observation focuses on students from the Master's program "Innovation and multidisciplinarity in teaching Mathematics, Computer Modelling and Information Technologies" in the winter semester of 2024/2025. A review of their work on ongoing assignments in the courses "Educational methods and tools and innovative approaches" and "Project-based learning" has been conducted. It is noteworthy that the participants in the study were master's-level students who also practice as mathematics teachers in well-regarded secondary schools, primarily based in the capital. Most of these individuals possess extensive experience in teaching and demonstrate a robust comprehension of the subject matter.

One of the assignments given in their coursework posed the following question: "Is artificial intelligence considered an educational technology?" A total of nine written responses were submitted, with most of these responses limited to a single printed page. One student explicitly stated that they initiated the process by consulting ChatGPT's response to the prompt. However, it is challenging to ascertain the extent to which the other submissions were independently authored or supported by AI tools.

The prevailing sentiment among the student cohort was that AI has the potential to function as a substantial support instrument for educators, particularly in the domains of data processing, test creation, and the automation of administrative tasks. The participants underscored the merits of intelligent educational platforms, which possess the capacity to adapt content in accordance with students' proficiency levels. However, it was acknowledged that such platforms are not widely used in the Bulgarian educational context, rendering many of the arguments more theoretical

than experiential. Many students have expressed ethical and social concerns regarding the implementation of artificial intelligence in educational settings. There was unanimous consensus that AI cannot substitute for the teacher as the primary figure in the educational process. However, AI can enhance and support the teacher's work. This mirrors findings suggesting that while educators see potential in AI, they remain sceptical and cautious in adopting it widely due to unclear guidelines and insufficient training [7]. This is further supported by findings from Pepin et al. [19], who emphasize that while teachers acknowledge the exploratory potential of tools like ChatGPT, they often remain hesitant to integrate them meaningfully into their pedagogy without clearer didactic frameworks or institutional guidance.

In the course entitled "Educational Methods and Tools and Innovative Approaches," students were assigned the following task: "Formulate a mathematics task that could be used as the basis for a mathematical essay, presentation, group activity, or practice-based individual project. In the case of employing a large language model (LLM) during task development, it is necessary to specify the model's name and version, the original prompt, and the generated result. The objective is to ingeniously modify the AI-generated content to create a final version of the task." The course was conducted online, enabling continuous feedback and iterative improvement of student work, which fostered both consultation and reflection. A study of the submitted assignments revealed that only one of the ten was explicitly inspired or co-authored with the help of AI. The remaining submissions appear to be authentic and original works, as evidenced by their style and depth. In subsequent discussions, it became clear that the students viewed the creation of assignments independently as a matter of professional pride. This attitude is in partial conflict with the course's underlying objective, which is to cultivate proficiency in the utilization of digital technologies and their pedagogical applications. In a similar vein, the course titled "Project-Based Learning" incorporated two assignments that advocated for the utilization of AI. However, the same cohort of students predominantly demonstrated a reluctance to engage with AI tools.

From a research perspective, this offers an insightful glimpse into teachers' attitudes toward creative collaboration with AI. Despite the modest sample size, the findings offer a promising line of inquiry for future studies delving into teachers' perceptions and inclination to incorporate AI in educational practice.

In another assignment, participants were obliged to select a problem from their own teaching practice:

"Choose a problem from your practice and write it down.

Look at the solution, using one or more of the different tools, like PhotoMath, ChatGPT, WolframAlpha, or feedback in a forum about a specific topic (these are just some examples, and they are not mandatory).

Pay attention to the steps of the generated solutions and answer the following questions:

(1) How similar is the generated solution to the one you would show your students, and in what ways are they different? In your analysis, clearly indicate which tool was used to generate the solution you are discussing.

(2) How did you choose the tool? Was it based on experience, skills, the specifics of the task, the students' knowledge, or other factors?
(3) What ideas does the solution give you for changing the task or making a new one?"

The responses indicated that the university students were aware of AI tools commonly used by school learners. It was observed that PhotoMath, a tool that was once prevalent among students, appears to be losing its relevance. Conversely, WolframAlpha is regarded as a favoured instrument among advanced students, university learners, and educators. Despite that the students demonstrated familiarity with the listed tools, their contributions were characterized by a lack of initiative and innovation in expanding the list of technologies. This finding suggests a potential area for curricular enhancement. According to Ravšelj et al. [23], teachers recognize the growing role of tools like ChatGPT in education but emphasize the need for clear institutional support and ethical guidelines for integration, noting a lack of consistent policy and training as barriers to meaningful adoption. This is true not only for educators themselves but also because educators serve as role models and conduits through which these competencies are transmitted to students. As emphasized by Pepin et al. [19], educators have a significant impact on classroom practices, shaping not only the pedagogical approach but also modelling digital engagement and the judicious use of tools such as ChatGPT. This role, as highlighted by the researchers, is crucial in determining the acceptance or restriction of these technological tools within educational settings.

2.2. The online materials vs. school students needs

The expansion of digital technologies and the ubiquity of internet access has led to a substantial augmentation in the array of mathematics resources accessible to students. In the contemporary educational landscape, learners have access to a variety of educational materials, including instructional videos created by individual enthusiasts and educational organizations, as well as collections of solved examination problems offered by private tutoring institutions and online portals. While this abundance provides unprecedented opportunities for independent learning, it also poses a significant challenge: the quality, clarity, and pedagogical consistency of these materials vary considerably.

In order to accurately assess the impact of AI in education, it is essential to understand the specific needs of students in the context of evolving educational resources.

In the late 20th century, students had limited access to resources for independent study. Mathematics textbooks were frequently minimalistic in nature, with answers to exercises being included only sporadically. Problem books, which are designed to address the most challenging aspects of mathematics education, seldom offer solutions to these problems. However, these solutions are often accompanied by expressions that may be perceived as discouraging, such as "it follows obviously from here," which may not provide adequate support to learners who are still grappling with fundamental concepts. In that context, students relied heavily on peer

collaboration and the goodwill of their teachers as the primary sources of additional guidance. The phenomenon of self-directed learning was largely constrained by a lack of accessible, comprehensive, and pedagogically appropriate materials.

The rise of the Internet, however, has had a profound impact on the educational landscape. Online forums have emerged as valuable spaces where students can seek clarification, engage in discourse, and disseminate solutions. These forums evolved into informal learning communities that reflected the diversity of students' learning styles. As Raban and Harper observe, "in general, question asking communities are formed thanks to two complementary human instincts: the need to ask questions and the inclination of answer people to contribute their knowledge to others" [22]. For some students, a fully written solution was sufficient; for others, the opportunity to pose follow-up questions and receive clarifying explanations was essential. This shift signalled a global trend toward more interactive and learner-centred support environments.

During this period, platforms such as Yahoo! Answers (active from 2005 to 2021) gained international popularity as public knowledge-sharing hubs. Yahoo! Answers, launched in 2005 [11], was among the first large-scale community-driven platforms that enabled users to ask and answer questions across a wide array of domains. The mathematics section of the website rapidly became a popular resource for students, educators, and enthusiasts seeking assistance with problem-solving, ranging from basic arithmetic to advanced calculus. Yahoo! Answers was a prominent example of an online knowledge-sharing community, characterized by its substantial size and diversity. This platform facilitated a wide range of interactions, encompassing both technical and social domains. Notably, it featured a robust Mathematics section that was specifically designed to cater to high school learners [1]. Yahoo! Answers and similar platforms have served not only as information sources but as informal learning communities, where the interplay of social cognition and perceived knowledge value drives both asking and answering behaviours [22].

Despite its initial success, Yahoo! Answers experienced a steady decline after 2010 due to diminished content quality, reduced moderation, and increased spam. Consequently, its educational value has diminished. In April 2021, the platform was officially deactivated [4]. Despite the absence of a systematic curation of its contents, the impact of this collection remains substantial. Subsequently, a considerable number of users and contributors migrated to more specialized platforms such as Math Stack Exchange, Reddit, and Quora. These platforms continue to address the evolving needs of students in the digital age.

The inception of Math Stack Exchange (MSE) in 2010 coincided with the broader expansion of the Stack Exchange network. The establishment of MSE was driven by an increasing demand for specialized, peer-reviewed mathematical support. In contrast to earlier platforms such as Yahoo! Answers, MSE implemented a reputation system, stringent moderation policies, and LaTeX integration, thereby facilitating engagement with mathematical content across diverse levels of proficiency, ranging from secondary education to advanced research. As Mansouri et al. observe, "the presence of Community Question Answering sites such as Math Stack

Exchange and Math Overflow suggests that there is a great public interest in finding answers to mathematical questions posed in natural language, using both text and mathematical notation" [14].

It is evident that, over time, Math Stack Exchange evolved into one of the most active and trusted online mathematics communities. The platform's emphasis on systematic topic categorization, the collaborative oversight of content by its members, and the establishment of trust based on reputation have collectively fostered a remarkably coherent and sustainable environment for knowledge sharing. Preliminary research on Stack Exchange communities indicates that "active communities have higher local cohesiveness and develop stable and more strongly connected cores." Social trust, as determined by reputation, plays a pivotal role in the long-term sustainability of these communities [27].

Concurrently, `Matematika.bg` was established by Yordan Petrov and remains under his maintenance [20]. Since its inception in 2005, the platform has been instrumental in fostering the development of online mathematics resources, with a particular emphasis on interactive content such as national assessment tests (e.g., for Bulgarian students in Grades 4 and 7). Notably, the platform has been led by Petrov in its development and authorship. The site offers a distinctive integration of several components. The materials have been meticulously aligned with the curriculum and include the official NEA tests for various grades. Petrov's engagement in the robust forum system is characterized by a wide range of participation in discussions, encompassing a diverse array of subjects, from high school-level topics to advanced mathematics, including differential equations. Under Petrov's leadership, `Matematika.bg` has evolved into a distinctive Bulgarian educational platform that integrates teacher-led resource development, learner interaction, and an ongoing vision for adaptive, personalized learning through exercises and structured guidance.

These platforms have addressed a significant gap in educational resources by providing students with mechanisms to seek clarification and receive explanations tailored to their cognitive needs and mathematical maturity levels. Rather than merely serving as answer banks, these communities foster social learning environments that reinforce mathematical learning through peer dialogue and explanation. According to Padayachee and Campbell [18], online mathematics forums can extend the benefits of face-to-face discussions to digital spaces. These digital spaces help students build a "community of inquiry" that promotes deep, interactive engagement with complex mathematical ideas.

The rapid dissemination of learning resources accelerated further due to the COVID-19 pandemic [26]. Notably, the number of educational videos created by teachers to methodically guide students through problem-solving processes increased. These videos promote conceptual understanding and demonstrate mathematical thinking in action. Meanwhile, private tutoring and after-school academies remained popular alternatives for customized educational support. Instructional videos, especially those with heuristic worked examples and interactive features, have proven to be valuable resources for fostering metacognitive engagement and supporting diverse learning needs. As Wirth and Greefrath [30] observe, such videos allow learners to

adapt the content to their own pace and provide step-by-step visual and verbal guidance, which students consider essential for grasping complex mathematical modelling processes.

2.3. From foundations to frontiers

The theoretical underpinnings of AI were established in the mid-20th century with foundational work in symbolic reasoning, machine learning, and computational models of cognition [24]. However, a substantial paradigm shift occurred at the end of the century with the emergence of deep learning, a subset of machine learning that uses multi-layered neural networks to model complex patterns in large datasets [8, 12]. The implementation of deep learning techniques has made training large-scale models, including LLMs, possible using extensive textual data sets [5, 25]. These techniques have enabled the generation of human-like language and the successful execution of a wide range of tasks with increasing precision [6].

A significant turning point in the evolution of LLMs and their public perception occurred in 2022 with the release of ChatGPT by OpenAI. Initially based on the GPT-3.5 architecture, ChatGPT later transitioned to the GPT-4 architecture [6, 17]. This release substantially increased access to conversational AI, showcasing the potential of LLMs in various domains, such as education, programming, creative writing, and research assistance [13, 17]. In 2025, the landscape evolved further with the introduction of DeepSeek [15], a next-generation model that expanded the capabilities of multilingual reasoning and domain specialization.

All these marked a pivotal shift in the ecosystem's evolution, introducing a new dimension. The reactions ranged from enthusiasm to scepticism. However, the rapid adoption of LLMs by students, particularly for mathematical tasks, underscores the ongoing demand for personalized, context-sensitive academic support [31].

It has become evident that students have consistently sought out learning platforms that align with their preferred learning styles [2], whether these styles are visual, verbal, exploratory, or procedural. The emergence of forums, videos, and now AI-powered assistants does not signify a fundamental shift in the needs of students. Rather, it indicates a shift in the methods through which these needs can be effectively and flexibly addressed. While some students continue to engage with conventional methods, such as individual teacher consultations or peer interaction within online forums, chatbot-based AI systems are increasingly being adopted as a widely accepted and accessible alternative. This is further supported by recent findings: "ChatGPT has a large positive impact on improving learning performance ($g = 0.867$) and a moderately positive impact on enhancing learning perception ($g = 0.456$) and fostering higher-order thinking ($g = 0.457$)" [28]. This evolution necessitates a critical reassessment of the design, integration, and evaluation of educational technologies. This re-evaluation should not be limited to the novelty of the technologies but should also consider their responsiveness to longstanding and authentic learner demands [28].

A comprehensive understanding of the expectations that contemporary students hold toward tools like ChatGPT necessitates a multifaceted examination of several

factors, including their motivational profiles, learning styles, and prior experiences with academic support systems. While some students continue to exhibit high levels of engagement through conventional methods such as individual teacher consultations or peer interaction within online forums, there has been an increasing adoption of chatbot-based AI systems as a widely accepted and accessible alternative [23].

What kind of support do students typically receive from AI Chatbots? Most of the time, it is a thorough, step-by-step plan with the necessary reasons and explanations [6, 29]. Often, these solutions follow a chain-of-thought format that breaks down complex problems into smaller, logically connected steps [23]. This type of reasoning has been shown to significantly improve the interpretability and effectiveness of AI responses in educational contexts. A valuable feature of this interaction is that students can ask clarifying questions about any stage of the solution process, creating a dynamic, adaptive learning experience.

One of the key advantages of using AI tools like ChatGPT is the psychologically safe environment they create. Students can engage with academic content without facing judgment or embarrassment, which are common in traditional classroom settings. The AI does not evaluate performance, assign grades, or offer opinions about knowledge. Instead, it functions as an impartial, always-available assistant whose sole purpose is to provide support and clarification. This neutrality makes ChatGPT especially appealing to learners who lack confidence or hesitate to participate actively in class. Research also shows that students value ChatGPT for its nonjudgmental tone and accessibility, especially when they are struggling with subjects such as mathematics [23].

3. Empirical evaluation

3.1. Dataset

The empirical study employed eight mathematics problems from the Bulgarian NEA for 10th grade, administered at the conclusion of the 2023/2024 academic year. The selected items were identified as Problems 1, 2, 4, 6, 7, 8, 9, and 10 [16].

The original examination contained multiple-choice items, with four options and one correct answer. For the purposes of this study, these multiple-choice items were adapted into a short-answer format. This transformation ensured that evaluation would account for the accuracy and format of the final answer in the context of the given problem. According to the official scoring rubric, each of these tasks is worth 4 points. The preparation of the problem involved the transcription of all mathematical formulae, expressions, and symbols into LaTeX. This was done to ensure precise automated processing.

3.2. Methodology

An automated evaluation was conducted using eleven popular large language models (LLMs) representing different architectures and parameter scales, all current

as of March 2025, and obtained from Ollama, a public repository of pre-trained generative models.

The following models were analyzed:

- BigGPT (9B and 27B)
- LLaMA 3 (8B and 70B; labeled `llama_31_8b` and `llama_33_70b`)
- DeepSeek R1 (7B)
- Mathstral (7B)
- Mixtral (8×22B and 8×7B)
- Mistral (7B, 22B, and 123B).

In addition to the automated batch evaluation, a manual assessment was performed on Le Chat (Mistral), DeepSeek, and ChatGPT. In these sessions, the same problems were posed interactively. In certain instances, supplementary constraints, guidelines, or requirements are delineated. The constraints are predominantly contingent upon the implementation of a particular mathematical approach, with the objective of adhering to the outlined curriculum.

4. Synopsis of the results

4.1. Automated evaluation

As illustrated in Figure 1, the success rate (i.e., the percentage of correctly solved problems) varies across models. The validation of a solution was contingent upon its capacity to yield an exact match with the anticipated outcome, presented in the prescribed format. In addition to raw success rates, the evaluation incorporated four qualitative scoring dimensions derived from the rubric.

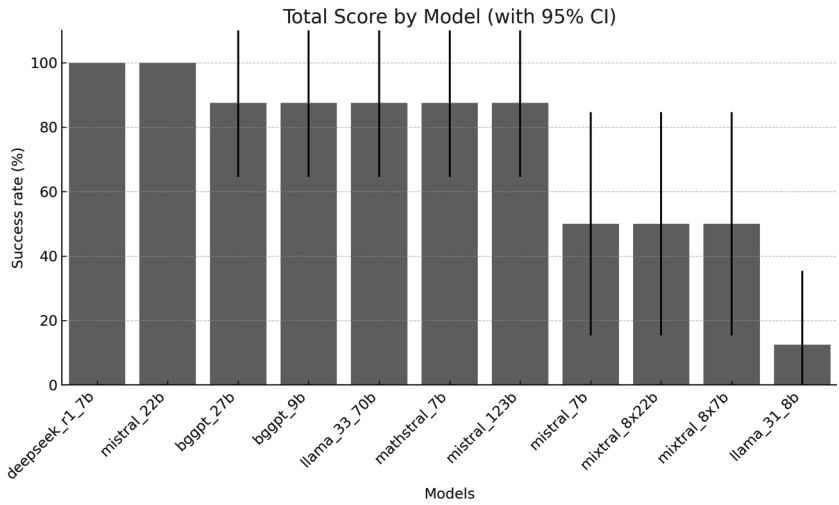

Figure 1. Total score by model

- The *Structured solution* (0–2) is the extent to which the model's response is characterized by a coherent, stepwise derivation leading to the conclusion;
- *Pedagogical quality* (0–2) – the alignment of the reasoning and methods with the official Bulgarian educational plan;
- *Critical error*, otherwise referred to as a penalty, is characterized by deductive or conceptual flaws, ranging from minor miscalculations that result in a score of −1 to incorrect underlying concepts that yield a score of −2;
- *Language clarity* (0–2) – the degree to which the language used is correct and appropriate, including grammar, orthography, and domain-specific terminology.

As illustrated in Table 1, there is a strong correlation between performance in raw accuracy and structured presentation and pedagogical quality, though the relationship is not perfect. For instance:

Table 1

Models by rubrics

Model	**Success rate (%)**	**Mean quality (%)**	**Structured solution (0–2)**	**Pedagogical quality (0–2)**	**Critical error (pen.)**	**Language clarity (0–2)**
DeepSeek R1 (7B)	100.0	54.17	1.50	1.50	0.00	0.25
Mistral (22B)	100.0	81.25	1.75	1.75	0.00	1.38
BgGPT (27B)	87.50	79.17	1.75	1.50	−0.38	1.88
BgGPT (9B)	87.50	81.25	1.75	1.50	−0.25	1.88
LLaMA 3.3 (70B)	87.50	56.25	1.38	1.00	−0.13	1.13
Mathstral (7B)	87.50	62.50	1.88	1.63	0.00	0.25
Mistral (123B)	87.50	85.42	1.88	1.88	−0.25	1.63
Mistral (7B)	50.00	41.67	1.13	0.75	−1.00	1.63
Mixtral (8×22B)	50.00	56.25	1.43	1.29	−0.50	1.50
Mixtral (8×7B)	50.00	33.33	1.00	0.88	−0.75	0.88
LLaMA 3.1 (8B)	12.50	12.50	0.75	0.38	−1.13	0.75

The DeepSeek R1 (7B) and Mistral (22B) models both demonstrated 100% success rates; however, a divergence in their mean quality scores was observed (54.17% vs. 81.25%). DeepSeek's lower score is attributable to more frequent minor deviations from the anticipated Bulgarian curriculum style, despite the presence of mathematically correct answers.

Mathstral (7B) attained the highest structured solution score (1.88/2), yet exhibited a lower mean quality (62.5%), frequently attributable to minor methodological discrepancies with the Bulgarian syllabus, such as the utilization of internationally prevalent solution techniques in conjunction with those dictated by the national curriculum.

Statistical significance

A one-way analysis of variance (ANOVA) was conducted on the success rates across the eleven models to determine whether the differences were statistically significant. The results of the study indicate the following:

The F-statistic $(10, 77)$ was determined to be 15.42, and the p-value was found to be less than 0.001. This indicates that there is a statistically significant difference in the success rates across the models.

Subsequent post-hoc Tukey HSD comparisons revealed that:

- The performance of DeepSeek R1 (7B) and Mistral (22B) was found to be significantly superior to that of LLaMA 3.1 (8B) ($p < 0.001$).
- The BgGPT variants (9B & 27B) and Mistral (123B) formed a statistically indistinguishable high-performance cluster, with success rates ranging from 87.5% to 100%.
- The lowest-performing group comprised of Mixtral (8×7B) and LLaMA 3.1 (8B).

Error analysis

A qualitative review of the findings indicated that linguistic and methodological issues predominated over purely mathematical errors.

The linguistic issues encompassed solutions that were entirely in English despite Bulgarian inputs, or grammatically awkward phrasing. In the majority of these cases, mathematical correctness remained unaltered, suggesting a sufficient understanding of the Bulgarian problem statement. Terminology frequently reflected literal translation rather than the locally accepted mathematical vocabulary.

The methodological deviations that were identified included correct but inefficient approaches. For example, solving for the individual roots and then multiplying, instead of directly applying Viète's formulas; or tackling higher-degree inequalities by case analysis rather than interval methods prescribed in the Bulgarian curriculum. These deviations appear to reflect global training bias rather than failure of reasoning.

Structured solution

It has been observed that models in the top performance group – namely, Mistral (123B), Mathstral, and BgGPT variants – typically exhibited clear multi-step solutions accompanied by formula citations prior to utilization. In contrast, LLaMA 3.1 (8B) exhibited the lowest structured score (0.75/2), frequently producing only minimal reasoning before yielding a definitive answer.

Pedagogical quality

The highest pedagogical alignment was observed in Mistral (123B) and Mistral (22B) (1.88 and 1.75/2, respectively). These models demonstrated a strong adherence to the Bulgarian curriculum methods, as evidenced by their utilization of the interval method for inequalities, which superseded the more ad hoc and case-specific approaches. Lower-scoring models frequently resorted to alternative yet valid methods derived from global training data.

Critical errors

The most proficient group exhibited a low incidence of critical errors, with DeepSeek R1 (7B), Mathstral (7B), and Mistral (22B) demonstrating no penalties. In contrast, LLaMA 3.1 (8B) accumulated the largest average penalty (-1.13), indicative of multiple conceptual misunderstandings rather than isolated arithmetic slips.

Language clarity

The BgGPT variants demonstrated a high level of linguistic appropriateness, with an average score of 1.88 out of 2. These variants exhibited a capacity to produce grammatically correct Bulgarian, incorporating locally relevant terminology, thereby exhibiting a strong alignment with the subject's linguistic nuances. DeepSeek R1 (7B) demonstrated near-flawless accuracy, yet its clarity score ($0.25/2$) was notably lower due to frequent mixing of Bulgarian with Russian or English, a combination that may present challenges in classroom adoption.

Observed patterns and implications

A thorough examination of rubric data reveals a discernible trade-off between accuracy and educational alignment in select models. Global models such as DeepSeek R1 and LLaMA 3.3 (70B) have been shown to achieve high levels of accuracy; however, these models have been observed to occasionally overlook region-specific pedagogical norms. Locally trained or fine-tuned models (BgGPT) demonstrate enhanced alignment with the national curriculum and superior language utilization, despite their inability to attain absolute precision. From an instructional perspective, structured reasoning and adherence to curricular conventions are as important as raw correctness, particularly for formative assessment and classroom demonstration. Consequently, while models such as DeepSeek R1 can function as problem-solving backends, their outputs may necessitate post-processing or fine-tuning to align with the pedagogical and linguistic standards for Bulgarian education.

The findings indicate that, despite noteworthy advancements in mathematical reasoning, prevailing LLMs lack the reliability necessary for autonomous tool use in this category of mathematics assessment problems, particularly in scenarios where pedagogical alignment and linguistic precision are paramount.

4.2. Manual assessment

In order to evaluate the applicability of contemporary AI chatbots in supporting Bulgarian high school mathematics education, three widely discussed models, DeepSeek, Le Chat, and ChatGPT, were examined with respect to their problem-solving capabilities, alignment with national curricular standards, and suitability for both teacher and student use.

DeepSeek

A notable constraint of the DeepSeek chatbot is its restricted capacity for image recognition. Consequently, the necessity arose to assess the efficacy of the LaTeX version of the mathematical problems. This option is not realistic for the average high school student. As with automated testing scenarios, the model produced accurate solutions. Notably, the production of grammatically correct and contextually appropriate Bulgarian in the formulation of solutions was identified as a significant advantage. Once more, discrepancies in problem-solving methodologies became evident. For instance, when determining the maximum of a quadratic function, the model's initial solution incorporated mathematical analysis. After the implementation of constraints that confined the methods to those pertinent to the curriculum, the vertex coordinates were obtained through the utilization of the method of completing the square. When prompted to use a standard formula, the model produced the desired solution. However, iterative refinement and methodological adjustment of this sort would exceed the capabilities of a typical tenth-grade student seeking assistance. Consequently, it is not appropriate to regard DeepSeek as a substitute for direct consultation with a teacher or participation in a peer discussion forum.

Le Chat (by Mistral)

Conversely, the Le Chat model has been shown to effectively implement image recognition, thereby enabling the direct processing of problem statements presented in image format. In contrast to the automated testing scenarios involving other Mistral-based models, all issues encountered in the interactive chat sessions were resolved correctly. However, disparities in the implemented solution strategies – relative to those with which Bulgarian students are more familiar – became evident, primarily due to the specificity of national educational standards. While Le Chat may serve as a convenient tool for teachers, enabling efficient recording and formatting of solutions and the generation of similar problems, for students, these methodological discrepancies limit its value as an autonomous learning aid.

ChatGPT

ChatGPT is historically the first model to overcome the language barrier, enabling seamless communication in Bulgarian. It continues to be the most prevalent model among students and educators in Bulgaria. Furthermore, product integrations have been developed; for example, SmarTest incorporates ChatGPT, thereby enabling automatically generated test questions to be copied and used in test creation [21]. However, the discrepancy between ChatGPT's problem-solving methodologies and the Bulgarian educational curriculum gives rise to compatibility issues. While many solutions are mathematically correct, they may not align with the methods or presentation formats expected within the national educational system, rendering them inapplicable in certain classroom or examination contexts.

The comparative analysis of DeepSeek, Le Chat, and ChatGPT in the context of mathematical problem-solving for Bulgarian high school students reveals distinct strengths and limitations.

5. Discussion and conclusions

The trajectory of mathematics learning resources is indicative of an ongoing process of adaptation to the evolving demands of students, who increasingly seek clarity in the material, accessibility in the form of support services, and a degree of personalization in their academic experiences. For many years, Bulgarian students, like their counterparts around the world, primarily relied on printed textbooks and problem books. These texts frequently provided only rudimentary explanations and omitted comprehensive solutions, impeding independent study. The additional guidance provided was largely derived from teacher consultations or peer interactions, thereby constraining the flexibility and breadth of self-directed learning.

The development of the Internet signified a substantial paradigm shift. The advent of online forums such as Yahoo! Answers, Math Stack Exchange, and the Bulgarian platform `Matematika.bg` has engendered novel opportunities for learners to pose questions, access explanations, and collaborate asynchronously. These communities proffered a plethora of problem-solving perspectives and social engagement, yet also introduced challenges, including inconsistent quality, varying depth, and delayed responses. While forums fostered a sense of community and allowed students to learn from multiple perspectives, they also required public participation, which some learners found uncomfortable.

Conversational AI tools, in contrast, facilitate the creation of a private, adaptive space where students can explore questions without peer visibility, receive immediate feedback, and adjust the depth or pace of explanations to suit their individual preferences. This development signifies a transition from socially mediated, frequently asynchronous learning models to an on-demand, personalized interaction paradigm. This paradigm emulates the psychological comfort of working one-on-one with a tutor. LLMs, such as ChatGPT, exemplify the latest stage in this progression, as they represent a shift from delayed, public exchanges to instant, private "comfort mode" interactions. It has been demonstrated that students could receive step-by-step solutions on demand, to inquire further without the concern of being judged, and to adapt conversations to their own pace and style. This evolution mirrors broader educational trends towards customized, adaptive systems that can cater to individual learning preferences [3, 28]. The psychological safety and flexibility of AI-powered chats have been shown to be especially appealing for learners who are hesitant to participate actively in class or in public online forums.

However, the empirical findings demonstrate that while LLMs frequently attain high levels of accuracy – at times reaching 100% on NEA mathematics problems – their methodologies do not invariably align with the Bulgarian curriculum. While these divergences are mathematically valid, they can pose challenges in high-stakes assessments, where strict adherence to established methods is mandatory [19, 23]. This observation aligns with a previously identified forum-era reality: the correctness of a resource does not inherently guarantee its alignment with the needs of its intended audience.

Linguistic clarity remains a further challenge. Locally fine-tuned models such as BgGPT score highly for using precise Bulgarian mathematical terminology, whereas some global models mix Bulgarian with English or Russian terms. In mathematics learning, as in forum exchanges, language precision and familiarity strongly influence comprehension and student confidence [30].

Insights from the parallel qualitative study reinforce these findings. Master level mathematics teachers recognized AI's potential for providing tailored student support but were cautious about integrating it into their own practice. Many preferred to create materials independently, viewing this as part of their professional identity. This cautious stance echoes earlier hesitations in adopting forums for formal teaching, reflecting a desire to maintain control over content and methodology. Yet, as students increasingly gravitate toward comfort-mode learning environments, the absence of guided integration risks widening the gap between formal instruction and self-directed study [2, 10, 13].

From static books to peer-driven forums, to interactive AI chats, each stage in this evolution reflects a shift not only in technology but also in the social dynamics of learning. Books offered privacy but little adaptability; forums introduced collaboration and multiple perspectives but required public participation; conversational AI combines the privacy of independent study with the adaptability and responsiveness of a personal tutor. Today's LLMs are uniquely positioned to meet these evolving needs by offering private, adaptive, and judgment-free interactions, while retaining the ability to guide students step-by-step.

However, for successful integration into formal education to occur, three specific criteria must be met.

- Mathematical accuracy is critical for maintaining the reliability that fosters student trust.
- Methodological alignment signifies the process of ensuring that solutions are congruent with local curricular standards, particularly within the context of assessment.
- Linguistic appropriateness is defined as the employment of precise, culturally and pedagogically relevant terminology.

The findings of this study suggest that in the absence of teacher mediation, the outputs of mathematically correct AI systems may potentially mislead students if they employ solution methods or language with which the students are unfamiliar. This evidence aligns with broader research indicating that effective AI adoption is contingent upon technical accuracy, institutional support, teacher training, and ethical safeguards [2, 3, 10, 13, 28]. In the domain of mathematics education, these considerations are particularly salient, as the presentation of solutions is as important as the correctness of the solutions themselves [19, 23, 30].

Moving forward, educational institutions should:

- Provide clear guidelines for aligning AI-generated content with curricular standards.
- Offer professional development enabling teachers to adapt AI outputs for local contexts.

- Ensure equitable access and address ethical considerations in AI-supported learning.

When incorporated deliberately into this continuum, ranging from books to forums to comfort-mode AI, LLMs have the potential to augment rather than substitute for the role of the teacher. These resources have the potential to combine the richness of peer-based learning with the safety and adaptability of private study, thereby ensuring that mathematics education remains both rigorous and responsive to students' changing expectations.

Acknowledgements

The research and preparation of this article were funded and supported by the National Program "Young Scientists and Postdoctoral Fellows – 2," administered by the Ministry of Education and Science of the Republic of Bulgaria.

References

[1] L. A. Adamic, J. Zhang, E. Bakshy and M. S. Ackerman, Knowledge sharing and Yahoo Answers: everyone knows something, in: Proc. 17th Int. Conf. on World Wide Web, WWW, Beijing, 2008.

[2] D. Ali, Y. Fatemi, E. Boskabadi, M. Nikfar, J. Ugwuoke and H. Ali, ChatGPT in teaching and learning: a systematic review, Education Sciences 14(6) (2024) 643.

[3] O. O. Ayeni, N. M. A. Hamad, O. N. Chisom, B. Osawaru and O. E. Adewusi, AI in education: a review of personalized learning and educational technology, GSC Advanced Research and Reviews 18(2) (2024) 261–271.

[4] I. Bonifacic, Yahoo!News, Yahoo, 06 04 2021. Available at https://www.yahoo.com/news/yahoo-answers-shutdown-may-4th-210240460.html. Accessed August 2025.

[5] T. Brown, B. Mann, N. Ryder et al., Language models are few-shot learners, in: 34th Conference on Neural Information Processing Systems, Vancouver, 2020.

[6] S. Bubeck, V. Chandrasekaran, R. Eldan et al., Sparks of artificial general intelligence: early experiments with GPT-4, arXiv preprint arXiv:2303.12712, 2023.

[7] D. N. C. Dey, Enhancing educational tools through artificial intelligence in perspective of need of AI, SSRN, 2024.

[8] I. Goodfellow, Y. Bengio and A. Courville, Deep learning, MIT Press, 2016.

[9] P. Halachev, Integration of ChatGPT in e-Learning systems: comprehensive review, Periodicals of Engineering and Natural Sciences 12(1) (2024) 169–182.

[10] G.-J. Hwang and Y.-F. Tu, Roles and research trends of artificial intelligence in mathematics education: a bibliometric mapping analysis and systematic review, Mathematics 9 (2021) 584.

[11] U. P. International, Yahoo! launches Web answering site, 08.12.2005. Available at https://phys.org/news/2005-12-yahoo-web-site.html. Accessed August 2025.

[12] M. I. Jordan and T. M. Mitchell, Machine learning: trends, perspectives, and prospects, Science 349(6245) (2015) 255–260.

[13] E. Kasneci, K. Sessler, S. Küchemann et al., ChatGPT for good? On opportunities and challenges of large language models for education, Learning and Individual Differences 103 (2023) 102274.

[14] B. Mansouri, A. Agarwal, D. Oard and R. Zanibbi, Finding old answers to new math questions: the ARQMath lab at CLEF 2020, in: Advances in Information Retrieval, 42nd European Conference on IR Research, Lisbon, 2020.
[15] A. Mathew, Deep seek vs. ChatGPT: a deep dive into AI language mastery, International Journal for Multidisciplinary Research 7(1) (2025) 5.
[16] MES, NEA, 06/2024. Available at https://www.mon.bg/nfs/2024/06/nvo_10kl_math_12062024.pdf. Accessed August 2025.
[17] OpenAI, GPT-4 Technical Report, 2023. Available at https://cdn.openai.com/papers/gpt-4.pdf. Accessed August 2025.
[18] P. Padayachee and A. L. Campbell, Supporting a mathematics community of inquiry through online discussion forums: towards design principles, Int. J. Math. Educ. Sci. Technol. 53(1) (2021) 35–63.
[19] B. Pepin, N. Buchholtz and U. Salinas-Hernández, Mathematics education in the era of ChatGPT: investigating its meaning and use for school and university education, Digital Experiences in Mathematics Education 11 (2025) 1–8.
[20] Y. Petrov, Matematika.bg, Available at https://www.matematika.bg/. Accessed August 2025.
[21] A. Popov, Artificial intelligence to assist teachers. Available at https://departments.unwe.bg/Uploads/Department/padmin_5c4a3_UNWE-Teacher-Lecture.pdf. Accessed August 2025 (in Bulgarian).
[22] D. R. Raban and F. M. Harper, Motivations for answering questions online, Burda Center Publishing, 2008.
[23] D. Ravšelj, D. Keržič, N. Tomaževič et al. Higher education students' perceptions of ChatGPT: A global study of early reactions, PLoS ONE 20(2) (2025) e0315011, https://doi.org/10.1371/journal.pone.0315011.
[24] S. Russell and P. Norvig, Artificial intelligence: a modern approach, Pearson, 2020.
[25] A. Vaswani, N. Shazeer, N. Parmar et al., Attention is all you need, in: 31st Conference on Neural Information Processing Systems, Long Beach, CA, 2017.
[26] R. Videla, S. Rossel, C. Muñoz and C. Aguayo, Online mathematics education during the COVID-19 pandemic: didactic strategies, educational resources, and educational contexts, Education Sciences 12(7) (2022) 492.
[27] A. Vranić, A. Tomašević, A. Alorić and M. M. Dankulov, Sustainability of Stack Exchange Q&A communities: the role of trust, EPJ Data Science 12 (2023) 4.
[28] J. Wang and W. Fan, The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: insights from a meta-analysis, Humanities and Social Sciences Communications 12 (2025) 621.
[29] J. Wei, X. Wang, D. Schuurmans et al., Chain of thought prompting elicits reasoning in Large Language Models. Available at https://doi.org/10.48550/arXiv.2201.11903. Accessed August 2025.
[30] L. Wirth and G. Greefrath, Working with an instructional video on mathematical modeling: uppersecondary students' perceived advantages and challenges, ZDM – Mathematics Education 56 (2024) 573–587.
[31] K. Žáková, D. Urbano, R. Cruz-Correia et al., Exploring student and teacher perspectives on ChatGPT's impact in higher education, Education and Information Technologies 30 (2024) 649–692.

Received on August 14, 2025
Accepted on November 1, 2025

Ralitza Stamenkova

Faculty of Mathematics and Informatics
Sofia University "St. Kliment Ohridski"
5, James Bourchier Blvd.
1164 Sofia
BULGARIA

E-mail: r.stamenkova@fmi.uni-sofia.bg