https://doi.org/10.63684/dk.2025.03.01

LÁSZLÓ BOGNÁR *–ANTAL JOÓS **

The Impact of Allowing ChatGPT on Responses to Different Question Types in a Mid-Term Math Exam

Abstract: This study investigates the impact of ChatGPT on student performance in mid-term math exams, focusing on differences in scores across various types of test questions. The findings reveal that students using ChatGPT exhibited significantly lower average scores compared to their non-GPT counterparts, with more erratic performance patterns. In particular, ChatGPT users struggled with complex mathematical operations, such as matrix inverses and vector multiplications. Both ChatGPT and Copilot displayed similar levels of consistency, occasionally providing incorrect or mixed answers, which may have contributed to the lower performance of GPT users. The study suggests that inadequate preparation and unfamiliarity with using GPT during exams could also have played a role in these results. These findings raise important questions about the integration of AI tools in education, particularly in subjects like mathematics, where precision is essential. Future research should explore optimal ways to integrate AI tools like ChatGPT into learning environments to enhance, rather than hinder, academic performance.
Keywords: ChatGPT; mathematics education; chatbot.

Absztrakt: Ez a tanulmány a ChatGPT hatását vizsgálja a diákok félévközi matematikavizsgán nyújtott teljesítményére, a különböző típusú tesztkérdések eredményei közötti különbségekre összpontosítva. Az eredmények azt mutatják, hogy a ChatGPT-t használó diákok jelentősen alacsonyabb átlagos pontszámokat értek el, mint a ChatGPT-t nem használó társaik, és a teljesítményük is kiszámíthatatlanabb volt. A ChatGPT-felhasználók különösen az összetett matematikai műveletekkel, például a mátrix inverzekkel és a vektorok szorzásaival küszködtek. Mind a ChatGPT, mind a Copilot hasonló következetességet mutatott, időnként helytelen vagy vegyes válaszokat adtak, ami hozzájárulhatott a GPT-felhasználók alacsonyabb teljesítményéhez.

* *University of Dunaújváros, Institute of Computer Engineering, Department of Mathematics*
Email: bognarl@uniduna.hu

** *University of Dunaújváros, Institute of Computer Engineering, Department of Mathematics*
Email: joosa@uniduna.hu

[1] Gulwani, S.–Polozov, O.–Singh, R. (2017): Program Synthesis. *Foundations and Trends® in Programming Languages,* 4., (1–2.), pp. 1–119.

[2] Miller, L. A. (1981): Natural Language Programming: Styles, Strategies, and Contrasts. *IBM Systems Journal,* 20., (2.), pp. 184–215.

[3] Dohmke, T. (2022): *GitHub Copilot Is Generally Available to All Developers.* Retrieved from https://github.blog/2022-06-21-github-copilot-is-generally-available-to-all-developers/

[4] OpenAI. (2022): *Introducing ChatGPT.* Retri-eved from https://openai.com/blog/chatgpt

[5] Hu, K. (2023): *ChatGPT Sets Record for Fastest-Growing User Base – Analyst Note. Reuters.* Retrieved from https://www.reuters.com/technology/chatgpt-sets-recordfastest-growing-user-base-analyst-note-2023-02-01/

[6] Mehdi, Y. (2023): *Reinventing Search with a New AI-powered Microsoft Bing and Edge, Your Copilot for the Web.* Retrieved from https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edgeyour-copilot-for-the-web/

[7] Pichai, S. (2023): *An Important next Step on Our AI Journey.* Retrieved from https://blog.google/technology/ai/bard-google-ai-search-updates/

[8] Buolamwini, J.–Gebru, T. (2018): Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: *Conference on Fairness, Accountability and Transparency,* pp. 77–91. PMLR.

[9] Liang, P. P.–Wu, C.–Morency, L.-P.–Salakhutdinov, R. (2021): Towards understanding and mitigating social biases in language models. In: *International Conference on Machine Learning,* pp. 6565–6576. PMLR.

[10] Bender, E. M.–Gebru, T.–McMillan-Major, A.–Shmitchell. S. (2021): On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* pp. 610–623.

A tanulmány szerint a nem megfelelő felkészülés és a GPT vizsgák során történő használatának ismeretlensége is szerepet játszhatott ezekben az eredményekben. Ezek az eredmények fontos kérdéseket vetnek fel a mesterséges intelligencia eszközeinek az oktatásba való integrálásával kapcsolatban, különösen az olyan tantárgyak esetében, mint a matematika, ahol a pontosság elengedhetetlen. A jövőbeni kutatásoknak fel kell tárni, hogyan lehet a ChatGPT-hez hasonló AI-eszközöket optimálisan integrálni a tanulási környezetbe, hogy javítsák, ne pedig akadályozzák a tanulmányi teljesítményt.
Kulcsszavak: ChatGPT; matematikaoktatás; chatbot.

Introduction

Research areas such as neural networks, program synthesis [1], and natural language programming [2] have been evolving for decades. However, it was only in the past year that these technologies became widely available to the general public through prominent commercial launches. In June 2022, GitHub Copilot, an AI-powered code generation tool, was officially released following a year of private beta testing [3]: Shortly after, in November 2022, OpenAI launched the ChatGPT AI chatbot [4], which rapidly attracted an estimated 100 million users within two months, setting a record for the fastest app growth in history [5]: By early 2023, Microsoft and Google had integrated similar conversational AI into their search engines [6; 7]: The quick adoption of AI tools has sparked widespread debate about several concerns, including bias [8; 9], ethics [10],

misinformation [11], data privacy [12], and environmental impact [13]: Among these issues, educators have raised concerns about the role of AI in assisting students with homework and exams in various subjects [14]: In the field of computer science education, AI tools have been shown to be particularly effective for programming tasks [15]: This paper investigates the impact of ChatGPT on student performance in mathematics, with an emphasis on how scores vary across different types of exam questions. We designed an experiment to evaluate whether the use of ChatGPT during a mid-term exam affects student outcomes. Research Question: Is there a significant difference between the average scores of students who use ChatGPT and those who do not on different types of test questions?

Conditions of the experiment

The experiment was conducted with international students enrolled in Mathematics 1 and Engineering Mathematics 1 at the University of Dunaújváros, Hungary. For simplicity, we will refer to both subjects as Mathematics 1 throughout this paper. In Mathematics 1, students are required to take two tests within the Moodle Learning Management System during the semester, and their final grade is based on the combined scores from these two mid-term tests. This study focuses exclusively on the results from the first test.

In this test, each student was presented with 5 questions, each covering a distinct sub-topic. While the structure of the questions was consistent across all students, the parameters within each question were randomized by the Moodle system. A total of 140 students submitted valid tests, of which 22 students self-reported using ChatGPT during the exam. Therefore, across the 140 students, a total of 700 individual questions were answered (140 students × 5 questions), with 110 of those questions (22 students × 5 questions) being solved with the assistance of ChatGPT.

The learning material for this test covered fundamental topics in linear algebra, including an introduction to matrices, matrix operations, calculating determinants and inverses, operations with vectors such as scalar and vector

[11] Kreps, S.–McCain, R. M.–Brundage, M. (2022): All the News That's Fit to Fabricate: AI-generated Text as a Tool of Media Misinformation. *Journal of Experimental Political Science,* 9., (1), pp. 104–117.

[12]Butterick, M. (2022): *GitHub Copilot Inv.estigation.* Joseph Saveri Law Firm & Matthew Butterick. Retrieved from https://githubcopilotinvestigation.com/

[13] Strubell, E.–Ganesh, A.–McCallum, A. (2019): *Energy and Policy Considerations for Deep Learning in NLP.* arXiv preprint arXiv:1906.02243.

[14] Johnson, A. (2023): *ChatGPT In Schools: Here's Where It's Banned–And How It Could Potentially Help Students.* Retrieved from https://www.forbes.com/sites/ariannajohnson/2023/01/18/chatgpt-in-schools-heres-where-its-banned-and-how-it-could-potentially-help-students/

[15] Lau, S.–Guo, J. P. (2023): From "Ban It Till We Understand It" to "Resistance is Futile": How University Programming Instructors Plan to Adapt as More Students Use AI Code Generation and Explanation Tools such as ChatGPT and GitHub Copilot. In: *ACM Conference on International Computing Education Research (ICER).*

multiplication, mixed multiplication, calculating angles between vectors, and operations with complex numbers in algebraic form such as addition, subtraction, multiplication, and division. Specifically, the test evaluated students on these core concepts:

– The first question (Q1) was on addition, subtraction, multiplication, transposition, determinant, adjoint and inverse calculus with matrices.
– The second question (Q2) was on addition and subtraction of vectors, multiplication by scalar, linear independence of vectors, base of vectors, rank of matrix, scalar multiplication of vectors and solution of linear system of equations.
– The third question (Q3) was about mixed product of vectors, scalar product of vectors, equation of a plane, equation of a line, length of a vector, product of vectors and closed angle of vectors.
– The fourth question (Q4) was taken from the same set as the first question.
– The fifth question (Q5) was taken from the topics: real and imaginary parts of complex numbers, sum of complex numbers, difference of complex numbers, multiplication of complex numbers and absolute value of complex numbers.

Students had 45 minutes to complete the test, with any unsubmitted tests automatically submitted at the end of this time. Immediately after submission, students were able to view their test results along with the correct answers. While the teaching was conducted face-to-face, the test itself was administered online. On the morning of the test, students were informed that they were permitted to use ChatGPT during the test. The test included a self-report question asking whether they had used ChatGPT. There were no penalties or incentives tied to the use of ChatGPT, and it had not been integrated into any of the prior mathematics lessons. Additionally, the problem-solving capabilities and limitations of ChatGPT had not been demonstrated in class. Students were allowed to use not only ChatGPT but also other similar AI tools.

Comparison of the goodness of GPTs

The students were free to use any learning aid they wished, but we specifically examined two, ChatGPT 3.5 and Copilot with Bing Chat. For all 110 questions that GPT users received, we asked the GPTs three times for the correct answer and categorized the goodness of the GPTs based on these answers.

If all three answers for a question were the same, then the GPT's answer to that question is called consistent, otherwise it is called inconsistent. If all three answers were the same and good, the GPT answer is called correct. If the answers are different, but there was a good answer among them, it is called mixed. If each answer was wrong (same or different), it is called incorrect.

Figure 1. presents a distribution of the correctness of GPT responses, with ChatGPT accounting for 50% and Bing Chat for the remaining 50%. The graph demonstrates that no statistically significant difference exists in the proportion of correct answers between ChatGPT 3.5 and Copilot with Bing Chat. ChatGPT 3.5 displays a slight edge in accuracy, but the difference is not substantial. The figure also reveals that there is no significant difference in the proportion of incorrect answers and mixed answers between the two programs. ChatGPT produces marginally fewer incorrect responses compared to Copilot with Bing Chat. The difference in mixed answers is negligible. The Mixed column indicates that approximately 13% of the time, the responding program is simply making random guesses.

Figure 1. The correctness of ChatGPT and Bing

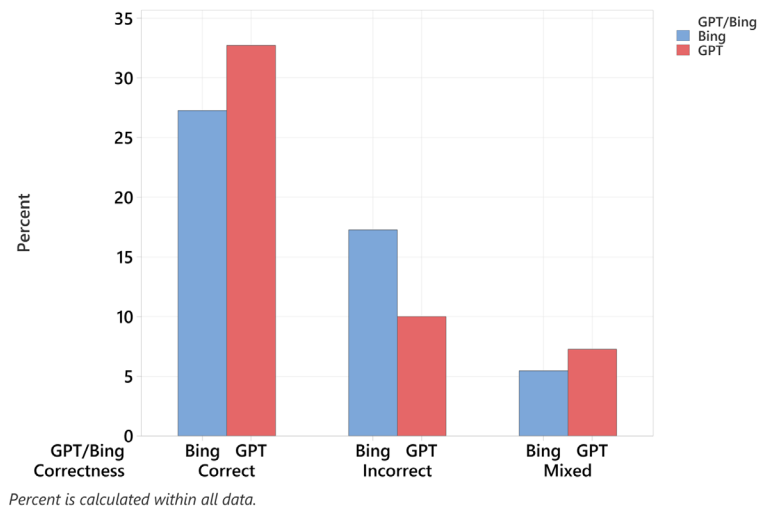


Figure 2. depicts the distribution of consistent and inconsistent answers for both GPT models, with ChatGPT and Bing Chat representing 50-50% of the data. From *Figure 2.*, it's evident that the consistency and inconsistency properties are comparable for the two programs. The rate of inconsistent responses is less than 20% within both ChatGPT and Copilot with Bing Chat categories. The Inconsistent column signifies that nearly 18% of the time, the responding program is resorting to random guesses.

Figure 3. illustrates the percentage of correct, incorrect, and mixed responses in ChatGPT's answers for each question. It indicates that for ChatGPT, question Q2 (involving basic vector operations) was the easiest, but even for this question, the model resorted to guesswork almost 10% of the time. Question Q3 (encompassing various vector multiplications and 3-dimensional coordinate geometry

problems) proved to be the most challenging, with a percentage of correct answers falling below 30%. *Figure 4.* presents the percentage of correct, incorrect, and mixed responses in Copilot with Bing Chat's answers for each question. The results are intriguing, as question Q5 (requiring basic operations on complex numbers) emerged as the easiest, with no mixed answers. Question Q3, once again, presented the most arduous task, with a percentage of correct answers plummeting below 10%.

Figure 2. The consistency of ChatGPT and Bing

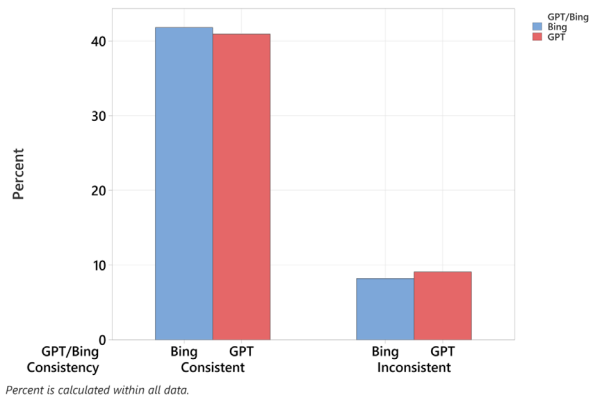


Figure 5. demonstrates the relative proportion of correct responses for each question across ChatGPT and Copilot with Bing Chat. In the case of question Q2, ChatGPT's ratio of correct answers hovered around 91%. This implies that the ratio of incorrect answers for ChatGPT was approximately 9%. Similarly, for Copilot with Bing Chat, the ratio of correct answers neared 72%, while the ratio of incorrect answers approached 28%. Overall, questions Q2 and Q5 emerged as the most accurate, while Q3 (involving various vector multiplications and 3-dimensional coordinate geometry problems) consistently posed the greatest challenge for both models.

Figure 3. The correctness of ChatGPT by question

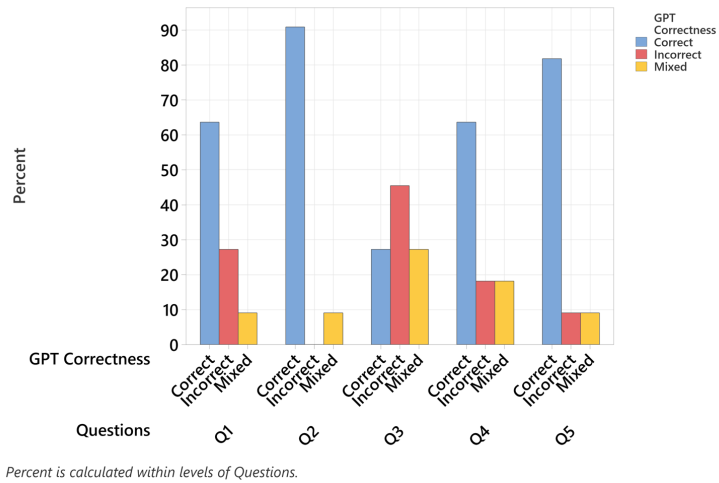


Figure 4. The correctness of Bing by question

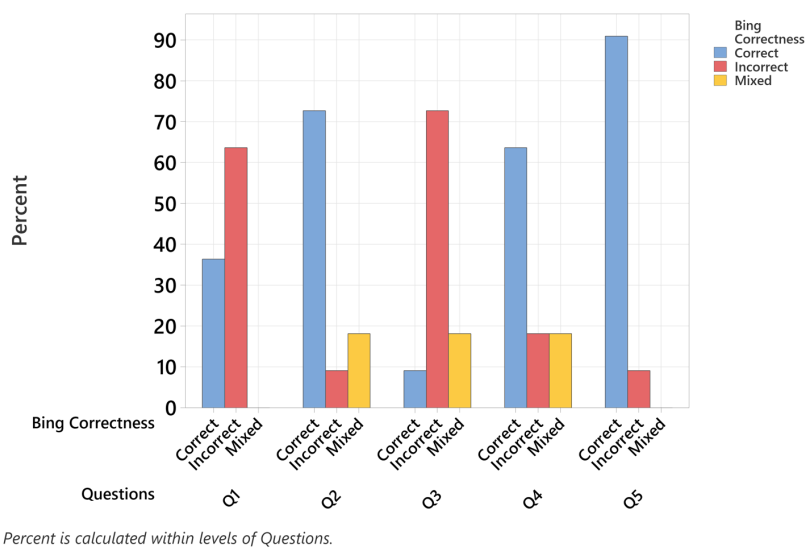
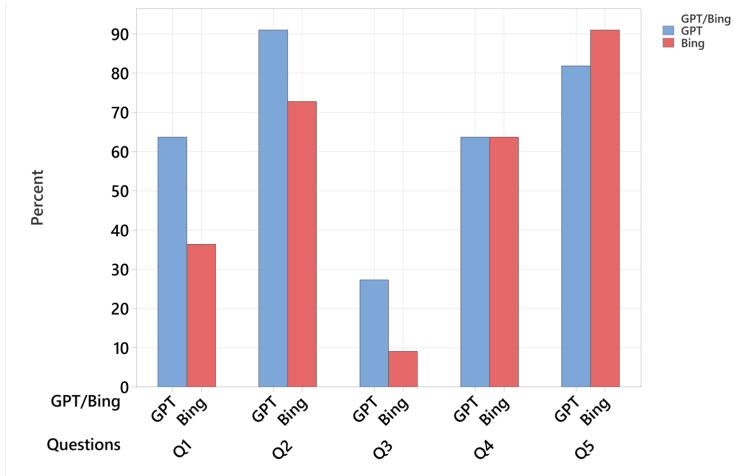


Figure 5. The correctness of ChatGPT and Bing by question



Comparison the goodness of student outcomes with and without a chatbot

We now delve into a comparison of the correctness of the students' responses, considering the accuracy of GPT's responses.

Figure 6. illustrates the proportion of correct responses for ChatGPT, students not using GPT, and students using GPT for each question. The first three columns reveal that 63% of ChatGPT's responses were accurate, 92% of students not using GPT provided correct answers, and 46% of students using GPT correctly answered the first question. This implies that if students using GPT had solely relied on ChatGPT's responses, their performance on the first question would have improved.

Conversely, on the third question, ChatGPT's performance significantly trailed that of students using GPT.

ChatGPT's accuracy on the third question was a mere 28%, while students using GPT achieved a correct response rate of 46%. Hence, there were instances where students identified incorrect answers provided by GPT or at that question they did not use it.

Notably, students not using GPT demonstrated superior performance compared to ChatGPT for questions Q1, Q3, Q4, and Q5.

Figure 6. The correctness of ChatGPT, student without GPT and with GPT

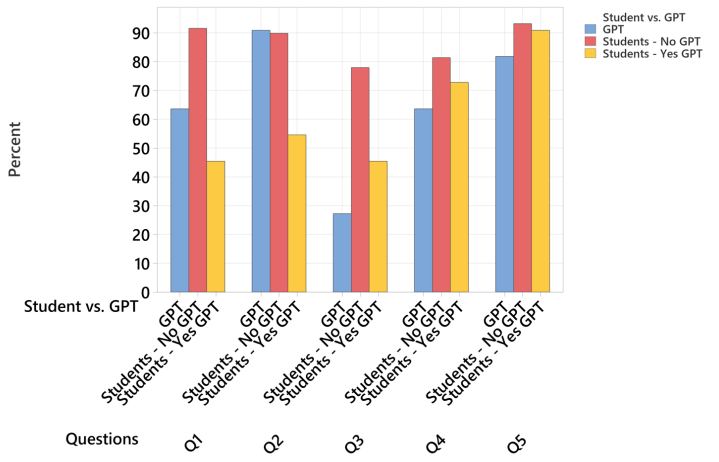


Figure 7. The correctness of Bing, student without Bing and with Bing

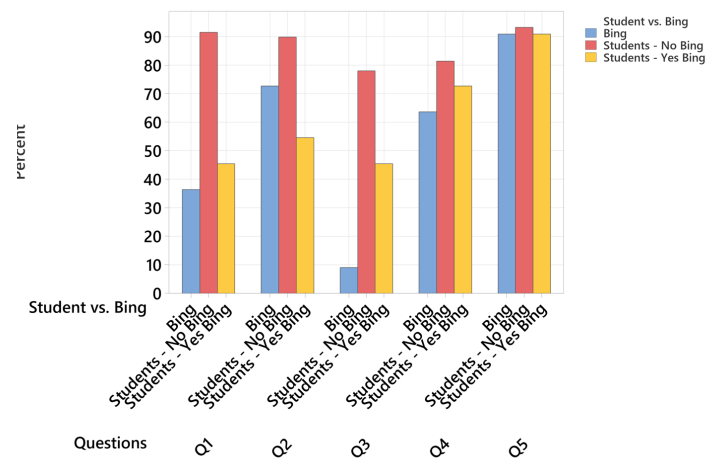


Figure 7. depicts the proportion of correct responses for Copilot with Bing Chat, students not using Bing, and students using it. It's worth noting that the performance of students not using Bing consistently surpassed those of using it.

Conclusions

This study investigating the use of ChatGPT in a mid-term math exam revealed several important findings, highlighting the complexities and challenges involved in integrating AI tools into educational environments.

GPT Consistency:
Both ChatGPT and Copilot with Bing Chat demonstrated comparable levels of correctness and consistency in their responses to the questions posed. The quality of responses from ChatGPT and Copilot did not show a significant difference, with both exhibiting occasional incorrect or mixed answers.

Question-Specific Performance:
Students using ChatGPT displayed more extreme variations in scores across different questions, especially compared to non-GPT users. Notably, questions related to complex mathematical operations, such as matrix inverses and vector multiplications, posed substantial challenges for ChatGPT users.

Implications and Further Questions:
The study prompts further exploration into the reasons behind the lower scores of ChatGPT users. Questions regarding students' study habits, the reliability of GPT-generated answers, and the impact of GPTs on learning outcomes merit deeper investigation.

The circumstances of the current experiment, in particular the fact that students were not prepared to use GPT in class, may have a significant impact on the findings. A further research task could be to compare different experimental set-ups.

The findings underscore the importance of carefully considering the integration of AI tools into educational practices, especially in disciplines like mathematics where precision is crucial.

In conclusion, while the introduction of AI tools like ChatGPT holds promise in educational contexts, this study emphasizes the need for a nuanced understanding of their impact, considering both the potential benefits and challenges associated with their implementation.

Future research should delve into refining AI tools for specific educational contexts and addressing the identified concerns to optimize their effectiveness.