THE IMPACT OF LLM-BASED CHATBOTS ON SECONDARY COMPUTING EDUCATION

Milić Vukojičić[1]*,
[0009-0002-1218-5893]
Ivana Korica[1],
[0009-0009-1918-347X]
Mladen Veinović[2]
[0000-0001-6136-1895]

[1]British International School Belgrade,
Belgrade, Serbia

[2]Singidunum University,
Belgrade, Serbia

Abstract:

This study aims to show the difference in teaching and learning approaches in secondary education. The purpose is to compare the output of a Computing project of three different groups of students with different resources for the same task. The main resource for the first group was the Computing book and teacher presentations, for the second group the main resource were the search engines Google and Bing, and the third group's main resource was the Large Language Model. By comparing the outcomes and performance of the three groups, the effectiveness of using different resources in a controlled classroom environment is assessed, showing the difference in overall performance of each group - project completion times vary widely, as do the definitions of key terms and answers to specific questions. The study also shows that resources such as books can be very useful as the limited experience of students means that the standards are very clear. Using a search engine expands the reach of information to students but also leads to too many choices. The problem that arises here is that of data accuracy, for example: whether the data obtained by LLM is accurate and sufficient for defining key terms and completing tasks.

Keywords:

Large Language Model, Artificial Intelligence in education, AI Chatbots, Computing education, Tutoring systems.

INTRODUCTION

In recent years education has become one of the main fields that have experienced major changes and the was faced with the need for improvement. The difference between secondary students now and those 10 years ago is dramatic, particularly due to their attention span, their understanding of what they have read, etc. The main difference that we can see in the past two years has been made by the increased use of Large Language Models (LLM) like ChatGPT (Generative Pre-trained Transformer). From the discussions of LLM replacing teachers in the classrooms [1] to the differences in the results when students are using LLMs [2], but also the research in the domain of essay writing [3] and digitised education [4,5] can give us the idea that ChatGPT will and is already affecting all forms of education. In addition to the impact observed on teachers and students within the educational framework, it is noticeable that both parties are subject to influences exerted by Large Language Models (LLMs) [6], thereby shaping and altering their roles, interactions, and experiences within the educational environment.

Correspondence:
Milić Vukojičić

e-mail:
vukojicic.milic@gmail.com

DOI: 10.15308/Sinteza-2024-31-37

Interest and research attention in the field of education regarding Large Language Models (LLMs) are experiencing a notable increase. Recent scholarly activities reflect a global trend wherein educators across diverse contexts are actively assessing the integration of LLMs within their instructional practices. This growing interest is particularly evident in evaluations centred around language teaching [7], as well as computer science and programming education [2], highlighting predominant concerns and inquiries within these domains.

ChatGPT embodies multiple roles within educational settings, functioning as an interlocutor, content provider, teaching assistant, and evaluator. Equally, teachers undertake complex roles that encompass orchestrating resources with pedagogical decisions, fostering student agency as active investigators, and instilling awareness of AI ethics [7].

Moreover, ChatGPT's utility extends to academia as evidenced by its role as a writing assistant in scholarly pursuits, as demonstrated in the paper titled "Analysing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. Contemporary Educational Technology" [8].

In various educational domains, ChatGPT – as an example of a Large Language Model – offers invaluable support to both students and teachers, facilitating Automated Grading and Feedback, Customized Learning experiences, Language Translation and Vocabulary Assistance, Personalized Educational Resources, Efficient Lesson Design, and Time Savings for Educators [9]. This likeness highlights the potential for ChatGPT to similarly benefit students across these areas.

Researchers are actively exploring avenues for integrating ChatGPT into educational practices, accompanied by guidance for its responsible implementation [10]. This concentrated effort reflects a growing recognition of ChatGPT's potential to enhance teaching and learning experiences while ensuring ethical considerations are prioritized. This paper aims to explain the impact and difference between students' results in the domain of Computing education in Secondary school education. The experiment was planned and conducted on a group of 39 students who were divided into three separate groups. Each group was equipped with varying resources to facilitate learning and task completion. Specifically, the resources provided to the students included a Computing coursebook accompanied by teacher presentations, access to conventional search engines such as Google and Bing, and exposure to an innovative LLM model, the above mentioned ChatGPT (Generative Pretrained Transformer). This comprehensive approach aimed to examine the efficacy and comparative advantages of different learning resources within the educational context, thereby offering insights into their respective impacts on student performance and outcomes.

2. EXPERIMENT METHODOLOGY

The experiment of the study is based upon the participation of three separate groups, totaling 39 Secondary school students. These groups were differentiated as follows: the first comprised of 12 students, the second included 15 students, and the third encompassed 12 students. Each group was equipped with unique resources to tackle the same Computing task. Specifically, the first group relied primarily on Computing books and teacher presentations, while the second group utilized the search engines Google and Bing. In contrast, the third group's main resource was the LLM ChatGPT, as depicted in Figure 1. This deliberate allocation of resources across groups enabled a comparative analysis of their respective impacts on task completion and learning outcomes within the context of Computing education.

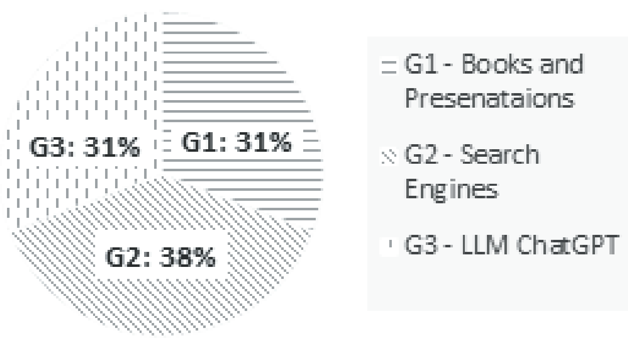


Figure 1. Student Groups and Resources assigned.

The experiment involved each student completing four relatively different tasks within two 40-minute lessons, totalling 80 minutes. Although group discussions on subject topics were present, all the tasks were written tasks and the discussions were not considered in regards to timing nor as a resource and the task-time-point distribution can be seen in Figure 2.

For the first task, students were tasked with defining keywords. They needed to define ten keywords, with each correct definition earning them 1 point. Students had 15 minutes to complete this assignment. Moving on to the second task, students participated in a class discussion followed by answering short questions. After discussing as groups for 20 minutes, they had an additional 15 minutes for further discussion and 5 minutes to answer 3 questions. Each correct answer was worth 10 points. The third task involved completing a short test comprising 10 brief questions. This test carried a maximum of 27 points, and students were allotted 15 minutes for completion. Lastly, the fourth task centred on creating a presentation. Students were provided with a template and specific instructions for each slide. They were required to create 8 to 10 slides, including a Title slide and a "Thank you" slide, with the remaining slides containing topic-related information. Each slide was valued at 10 points, and students had 30 minutes to finalize their presentations.

This structured approach ensured that students engaged in diverse activities within the designated time frame, covering aspects such as keyword definition, group discussion, individual assessment, and presentation development. The clear description of tasks and time allocation facilitated efficient completion while allowing for comprehensive evaluation of student performance across different skill sets.

The values displayed in Table 1, Table 2 and Table 3, and Figure 3, Figure 4 and Figure 5 offer a comprehensive overview of the outcomes obtained across Task 1 through Task 4 for each of the three groups under study. These tables serve as essential sources of data, summarising the performance metrics and achievements of the students throughout the experimental process. In Task 1 through Task 4, encompassing activities ranging from keyword definition to presentation development, the values presented in the tables are expressed in percentages. This percentage representation offers a standardized means of comparison, enabling a nuanced understanding of the relative performance levels achieved across the different tasks and groups.

Furthermore, the time allocated for completing each task is recorded in seconds, providing insight into the efficiency and pace at which students engaged with the assigned activities. This time-based measurement adds detail to the analysis, enabling an overview of patterns related to time management and task completion rates among the groups.

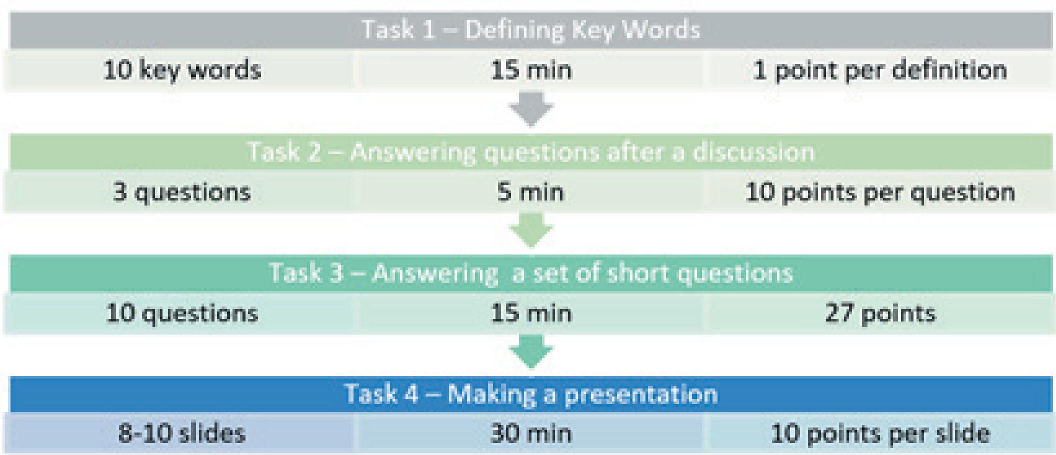
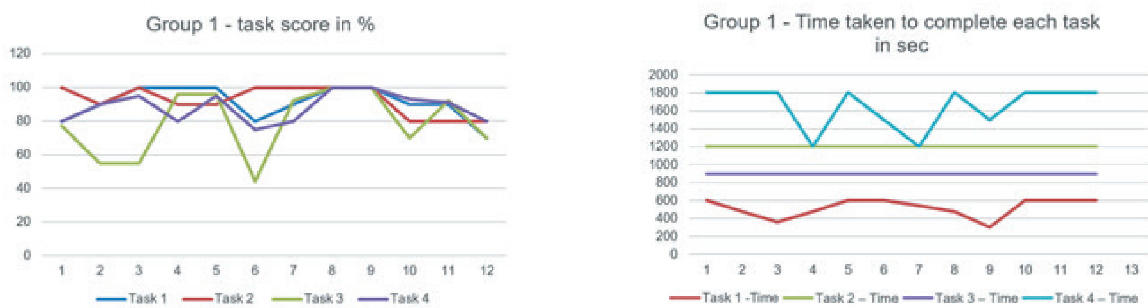


Figure 2. Task details.

Table 1. Results of group 1, Computing book and teacher presentations as a main resource.

Student Number	Task 1 (%)	Task 1 (sec)	Task 2 (%)	Task 2 (sec)	Task 3 (%)	Task 3 (sec)	Task 4 (%)	Task 4 (sec)
1	80	600	100	1200	77	900	80	1800
2	90	480	90	1200	55	900	90	1800
3	100	360	100	1200	55	900	95	1800
4	100	480	90	1200	96	900	80	1200
5	100	600	90	1200	96	900	95	1800
6	80	600	100	1200	44	900	75	1500
7	90	540	100	1200	92	900	80	1200
8	100	480	100	1200	100	900	100	1800
9	100	300	100	1200	100	900	100	1500
10	90	600	80	1200	70	900	93	1800
11	90	600	80	1200	92	900	91	1800
12	70	600	80	1200	70	900	80	1800



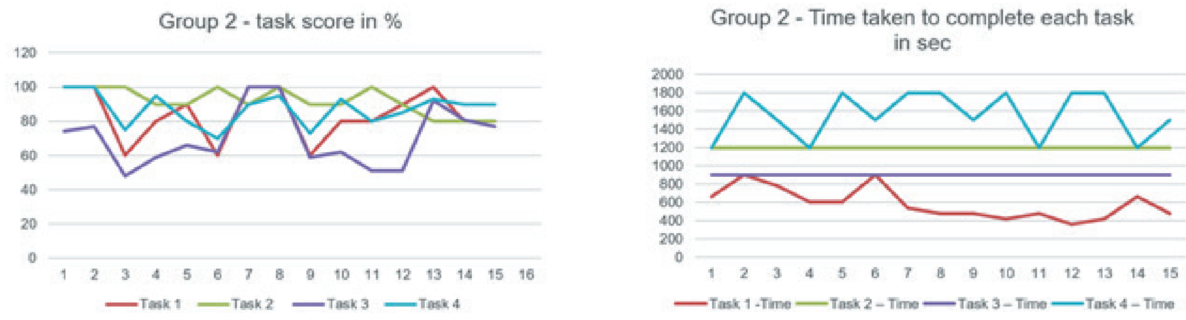
(a)

(b)

Figure 3. Group 1 - task score in percentages (a) and Group 1 - Time taken to complete each task in seconds (b).

Table 2. Results of group 2, Search engines Google and Bing as a min resource.

Student Number	Task 1 (%)	Task 1 (sec)	Task 2 (%)	Task 2 (sec)	Task 3 (%)	Task 3 (sec)	Task 4 (%)	Task 4 (sec)
1	100	660	100	1200	74	900	100	1200
2	100	900	100	1200	77	900	100	1800
3	60	780	100	1200	48	900	75	1500
4	80	600	90	1200	59	900	95	1200
5	90	600	90	1200	66	900	80	1800
6	60	900	100	1200	62	900	70	1500
7	100	540	90	1200	100	900	90	1800
8	100	480	100	1200	100	900	95	1800
9	60	480	90	1200	59	900	73	1500
10	80	420	90	1200	62	900	93	1800
11	80	480	100	1200	51	900	80	1200
12	90	360	90	1200	51	900	85	1800
13	100	420	80	1200	92	900	93	1800
14	80	660	80	1200	81	900	90	1200
15	80	480	80	1200	77	900	90	1500

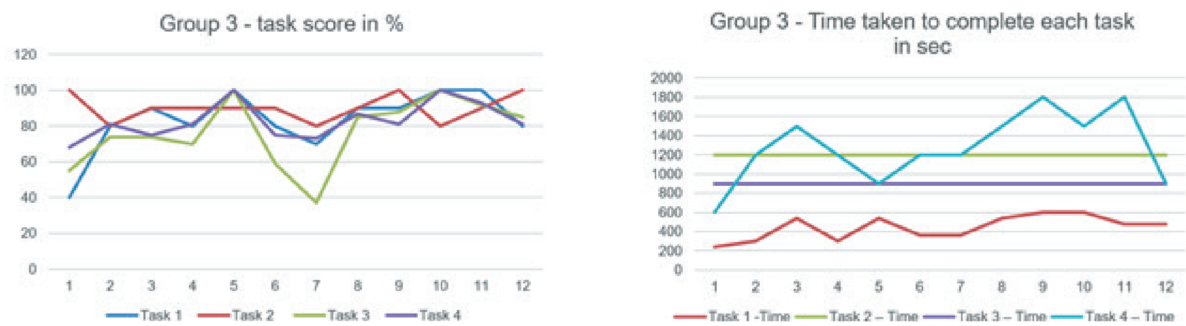


(a) (b)

Figure 4. Group 2 - task score in percentages (a) and Group 2 - Time taken to complete each task in seconds (b).

Table 3. Results of group 3, Large Language Model (ChatGPT) as a main resource.

Student Number	Task 1 (%)	Task 1 (sec)	Task 2 (%)	Task 2 (sec)	Task 3 (%)	Task 3 (sec)	Task 4 (%)	Task 4 (sec)
1	40	240	100	1200	55	900	68	600
2	80	300	80	1200	74	900	81	1200
3	90	540	90	1200	74	900	75	1500
4	80	300	90	1200	70	900	81	1200
5	100	540	90	1200	100	900	100	900
6	80	360	90	1200	59	900	75	1200
7	70	360	80	1200	37	900	73	1200
8	90	540	90	1200	85	900	87	1500
9	90	600	100	1200	88	900	81	1800
10	100	600	80	1200	100	900	100	1500
11	100	480	90	1200	92	900	93	1800
12	80	480	100	1200	85	900	81	900



(a) (b)

Figure 5. Group 3 - task score in percentages (a) and Group 3 - Time taken to complete each task in seconds (b)

3. RESULTS AND DISCUSSION

This paper presents the outcomes obtained by students upon completing all four tasks (Tasks 1-4) as well as the corresponding time required for task completion. These results are compiled and presented in Table 4, offering a comparative analysis across Group 1 (G1), Group 2 (G2), and Group 3 (G3). The data in Table 4 is structured to showcase key performance metrics, including the Minimum (Min), Maximum (Max), Average, and Median percentages attained by each group following evaluation. Additionally, the time taken for task completion is presented in seconds. This facilitates a clear and concise examination of the performance outcomes and time-based aspects associated with each group's engagement in the assigned tasks. Through this comprehensive presentation of results, valuable insights can be gained into the effectiveness of different instructional strategies and resource allocations organised within the educational context.

Upon closer examination of the data, it becomes evident that Group 1 emerged with the highest test scores, attributed to their use of Computing books and teacher presentations as primary resources. Equally, Group 3, which relied on the Large Language Model ChatGPT, demonstrated the fastest response times among the groups.

Another crucial aspect to consider is the discrepancy between the minimum and maximum values across the groups. Notably, Group 3 exhibited the most significant variance, with the widest range observed between the lowest and highest scores attained by students across tasks. Specifically, for Task 1, the score range spanned from 40 to 100, while for Task 3, it ranged from 37 to 100. Interestingly, Task 3 exhibited consistent differences in score ranges across all three groups.

These findings underscore the nuanced interplay between resource allocation, task performance, and time management within the educational setting. While certain groups may excel in specific tasks owing to their chosen resources, variations in individual performance levels underscore the need for tailored instructional approaches. Furthermore, the consistent patterns observed in Task 3 across all groups permits further investigation into potential underlying factors influencing student outcomes. Overall, this comprehensive analysis provides a robust foundation for refining educational strategies and optimizing resource allocations to enhance student learning experiences.

4. CONCLUSION

In recent years we have seen more papers on the topic of how LLM models can be used in teaching and learning to solve tasks, perform tests, make presentations, as well as help students in a variety of tasks. The criticism of LLM models [11,12] and the difference between human and LLM output [13] was discussed in several papers.

In this paper, we noted the difference between the three groups of students which relied on different sources – coursebooks, presentations, search engines, and LLM models – in the domain of completing school work tasks. The results show that the best-performing

Table 4. Comparison between results of (Group 1 - G1, Group 2 - G2, and Group 3 - G3), represented in Min, Max, Average, and Median percentage received after evaluation and time represented in seconds.

	Task 1 (%)	Task 1 (sec)	Task 2 (%)	Task 2 (sec)	Task 3 (%)	Task 3 (sec)	Task 4 (%)	Task 4 (sec)
Min (G1)	70	300	80	1200	44	900	75	1200
Max (G1)	100	600	100	1200	100	900	100	1800
Average (G1)	**90.83333333**	520	**92.5**	1200	**78.91666667**	900	**88.25**	1650
Median (G1)	**90**	570	**95**	1200	**84.5**	900	**90.5**	1800
Min (G2)	60	360	80	1200	48	900	70	1200
Max (G2)	100	900	100	1200	100	900	100	1800
Average (G2)	84	584	92	1200	70.6	900	87.26666667	1560
Median (G2)	80	540	90	1200	66	900	90	1500
Min (G3)	40	240	80	1200	37	900	68	600
Max (G3)	100	600	100	1200	100	900	100	1800
Average (G3)	83.33333333	**445**	90	1200	76.58333333	900	82.91666667	**1275**
Median (G3)	85	**480**	90	1200	79.5	900	81	**1200**

students are students from the first group, which had only coursebooks and presentations as a resource. We can also observe that they needed the longest amount of time to finish tasks as the information available to them about task topics was limited. The second group, which used search engines, took the most amount of time to finish tasks, as they had unlimited resources on the internet to choose from for their tasks. Here it was very hard to differentiate which information was crucial to them and which was not. Group 3 was the fastest, they used ChatGPT as the main resource, but they also had the lowest score on the task results. The biggest difference between the lowest-scoring student and highest highest-scoring student we observed in Group 3.

While students can be confused when they have unlimited choices for resources like search engines, the performance of the students decreases when they have limited resources like ChatGPT. LLM-based models can improve the speed of completing tasks, however they rely on the student's ability to ask questions. Future work in the domain of students' performance in secondary school education, based on LLM, can be done by using different LLM-based chatbots like Gemma and Llama 2.

5. REFERENCES

[1] A. M. A. Ausat, B. Massang, M. Efendi, N. Nofirman, and Y. Riady, “Can Chat GPT replace the role of the teacher in the classroom: A fundamental analysis,” Journal on Education, vol. 5, no. 4, pp. 16100-16106, 20232.

[2] M. Vukojičić and J. Krstić, “ChatGPT in programming education: ChatGPT as a programming assistant,” InspirED Teachers’ Voice, vol. 2023, no. 1, pp. 7-13, 2023

[3] S. Sharma and R. Yadav, “Chat GPT – A technological remedy or challenge for the education system,” Global Journal of Enterprise Information System, vol. 14, no. 4, pp. 46-51, 2022

[4] A. G. R. Castillo, H. V. H. Rivera, R. M. V. Teves, H. R. P. Lopez, G. Y. Reyes, M. A. M. Rodriguez, et al., “Effect of Chat GPT on the digitized learning process of university students,” Journal of Namibian Studies: History Politics Culture, vol. 33, pp. 1-15, 2023

[5] K. Aleksić-Maslać, F. Borović, and Z. Biočina, “Perception and Usage of Chat Gpt in the Education System,,” in INTED2024 Proceedings, pp. 1842-1848, IATED, 2024

[6] J. Whalen and C. Mouza, “ChatGPT: Challenges, opportunities, and implications for teacher education,” Contemporary Issues in Technology and Teacher Education, vol. 23, no. 1, pp. 1-23, 2023

[7] J. Jeon and S. Lee, “Large language models in education: A focus on the complementary relationship between human teachers and ChatGPT,” Education and Information Technologies, vol. 28, no. 12, pp. 15873-15892, 2023.

[8] M. Imran and N. Almusharraf, “Analysing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature,” Contemporary Educational Technology, vol. 15, no. 4, pp. ep464, 2023

[9] M. Javaid, A. Haleem, R. P. Singh, S. Khan, and I. H. Khan, “Unlocking the opportunities through ChatGPT Tool towards ameliorating the education system,” BenchCouncil Transactions on Benchmarks, Standards and Evaluations, vol. 3, no. 2, p. 100115, 2023

[10] M. Halaweh, “ChatGPT in education: Strategies for responsible implementation,” Contemporary Educational Technology, vol. 15, no. 2, p. ep421, 2023

[11] E. M. Bender and A. Koller, “Climbing towards NLU: On meaning, form, and understanding in the age of data,” in Proceedings of the 58th annual meeting of the Association for Computational Linguistics, July 2020, pp. 5185-5198

[12] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “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

[13] M. Vukojičić, J. Krstić, and M. Veinović, "Imitation Drawing: Can We Spot the Difference Between Ai and Human Generated Drawing?," Jan. 2023. [Online]. Available: https://doi.org/10.15308/sinteza-2023-23-27