Analysis of the Possibilities of Using LLM Chatbots for Solving Course and Exam Tasks

Svetlana Stefanova, Yordan Kalmukov
Department of Computer Systems and Technologies, University of Ruse, Ruse, Bulgaria

***Abstract*—With the widespread introduction of new technologies and, in particular, AI in various areas of life, students are increasingly using large language models (LLMs) such as ChatGPT and other similar tools to help them with their academic tasks. By using them, they can improve their productivity, improve their understanding of complex topics, and support their academic work. LLMs are used both in research, information gathering and preparation for exams and tests, as well as for generating ideas, creating code, and more. This study explores the possibility of using ChatGPT, Claude and DeepSeek for solving course and exam tasks. The results of the analysis could serve as a warning signal and motivation for future transformation of student testing and assessment methods. The ability to use AI systems to search, analyze, and summarize large volumes of information should shift the focus of assessment from classical fact-finding and practical performance of elementary tasks to creativity, combinability, and skills for adapting and applying the already gained knowledge.**

Keywords*—Large language models (LLM); artificial intelligence (AI); ChatGPT; Claude AI; DeepSeek; AI in education; AI for solving exams

I. Introduction

The use of large language models (LLMs) and chatbot platforms in educational contexts, especially for solving course and exam tasks, is growing and offers various possibilities, but also raises important ethical and pedagogical questions.

One of the directions of use is for teaching assistance, as learners can ask questions and receive information and explanations. In some of the tasks, the solution can be given step by step, rather than just providing the final results. In this way, students can be supported in the process itself and encouraged to think critically by asking them questions step by step instead of providing them with final answers. Unfortunately, some students rely too much on LLMs and do not even check the information they generate, which prevents them from mastering key skills in solving problems. An additional potential risk for them is the sometimes inaccurate or downright false information generated.

Another direction of use is when working on projects. In this case, logical frameworks would be helpful before starting the writing process itself, detecting and correcting stylistic and grammatical inaccuracies and errors, as well as generating quotes. Greater attention should be paid when such tools are used to solve practical tests. If this opportunity is used for exam simulation and self-assessment, then the usefulness is once again a fact. Cases of concern, however, are when LLMs are used to directly solve exams or online tests, which violates academic ethics. The use of real-time artificial intelligence tools during exams is a violation of academic education policies.

Types of AI-based platforms

The platforms and tools used by students can be divided into general and those that are focused on education.

Common platforms include:

- ChatGPT [1] – used for information gathering and problem solving;
- Claude [2] – suitable for longer documents and reasoning tasks;
- Gemini [3];
- Copilot [4] – already built into Word, Excel and other tools of the Office suite;
- DeepSeek by the Chinese hedge fund High-Flyer [5].

The tools focused on education are Khanmigo by Khan Academy [6]; Google Socratic [7] as a mobile application for explaining test problems; Quillionz [8] for generating test questions, and Elicit [9] for searching and finding academic articles and other publications.

In order to prevent risk, it is advisable to set clear rules for when the use of artificial intelligence tools is allowed, giving priority to originality. It is good for students to understand the strengths and weaknesses of the LLM so that they can be aware of the role of analytical and critical thinking.

This study performs a series of experimental analyses aiming to test whether the LLM chatbots like ChatGPT, Claude and DeekSeek could solve exams in real time without any further interaction, rather than just providing exam materials as an image taken from a smartphone or smart watch. Results revealing the short answer “yes” should serve as a warning signal and motivation for future transformation of student testing and assessment methods.

The study is structured as follows: Section II reviews related work done by other researchers. Section III presents the experimental setup and the examination materials used to test the LLM chatbots. Section IV discusses the experimental results, their alignment with the related work and their impact on the methods of examining and evaluating students. Finally, Section V ends the study with a conclusion, outlining the need for future transformation in education in the age of AI.

II. Related Work

Puthumanaillam et al. [10] tested if LLMs (ChatGPT) can successfully complete an entire bachelor course in Aerospace Control Systems. Not just a single final exam, but an entire course with 115 course deliverables, ranging from multiple-choice questions to complex Python programming tasks and long-form analytical writing. ChatGPT successfully completed the course and earned a B grade (82.24%). Its strongest results are in structured assignments and greatest limitations in open-ended projects [10].

Rytilahti and Kaila [11] investigated the capabilities of LLM ChatGPT (GPT3.5 and GPT-4) tools to solve coding exercises in an Introductory programming course. Their results show that the LLM can indeed be quite effective in solving the coding exercises. Depending on the version of the tool, the selected approach (amateur or already experienced student), and the prompt used, ChatGPT was able to achieve between 63.4% and 86.2% of the course's total number of points. If programming exercises were only considered, then ChatGPT answered 100% correctly to 107 (75.9%) and to 139 (98.6%) of the course's 141 programming exercises [11]. This means that even students with no previous experience in programming can successfully complete programming courses by utilizing freely available tools [11].

VarastehNezhad et al. [12] evaluated the performance of popular LLMs (GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Flash, Llama-3.1-70B, Mistral Large 2, DeepSeek-V2, and Gemma-2-27B) in answering questions in data structure and algorithm design from the Iranian university entrance exams for master's programs in computer science. They analyzed the accuracy and the length of responses to the exams in 2022, 2023 and 2024. The questions were given to the LLMs in both Persian and English so that the authors could compare the LLMs' performance in both languages. Results indicate that GPT-4o achieved the highest average accuracy (75.0%), followed by Claude 3.5 Sonnet (67.2%) and Mistral Large 2 (64.1%) [12]. As expected, evaluated LLMs perform better in English than in Persian, with GPT-4o having the largest performance gap (81.3% in English vs 68.8% in Persian).

Felicia Burlacu [13] analyzes the patterns in the performance of LLMs, i.e. what factors influence the accuracy of the LLMs' responses. She identified that all matters - the format of the question (multiple-choice or short answer), the length of the question and its type (factual or analytical). According to the results, LLMs perform best (71.42%) on multiple-choice questions of medium length (50 to 100 words) and factual type. Question length is important since a short question may not provide enough details to the LLM, while a long question could confuse it. Expectedly, factual questions get a higher average accuracy (64.28%) than analytical questions (42.85%).

Knowing that LLM chatbots could be used for cheating during exams, Simon Kaare Larsen [14] proposes guidelines and strategies for creating LLM-resistant exams, including content moderation, deliberate inaccuracies, real-world scenarios beyond the model's knowledge base, effective distractor options, evaluating soft skills, and incorporating non-textual information.

III. Experimental Setup

Since we teach in subjects related to Computer Science, Web Programming and Artificial Intelligence, we are interested in how well the LLM chatbots could solve the exam materials in these specific subjects.

The most commonly used exam formats for IT courses are "theoretical" with many questions that expect short open answers or multiple-choice answers, and practical exams, where students are required to write programming code and develop working applications. The first one is used to check students' overall knowledge or common sense in a specific area of science, while the practical exam tries to evaluate students' ability to develop real-life working applications on a computer. During practical exams, students are usually allowed to use any third-party educational resources on the Internet, except AI chatbots.

The exam in Distributed Web Applications is a "theoretical" one containing 18 questions that require short open-ended answers. Students have 1 hour to answer all questions, but they usually do it faster. All questions are printed on A4-sized paper, and students are required to answer with 1 to 3 sentences directly on that sheet of paper. Here are some example questions: "Specify two advantages of orchestration over choreography", or "If you need to guarantee a strictly defined order of execution of web services, will you choose synchronous or asynchronous communication between them?"

The exam in Information Retrieval is similar (see Fig. 1). It consists of 17 questions that require short open-ended answers between one word and 2 to 3 sentences. Students receive all 17 questions printed on A4-sized paper and should write their answers on the same sheet as well. Here are some example questions: "What are the main differences between the Latent Semantic Analysis and the Vector Space Model?", "Which similarity measures could be used to calculate similarity between sets of keywords?" or "What is the difference between flat clustering and hierarchical clustering?"

In contrast to the previous two, the exam in Web Programming is practical and is conducted on a computer (see Fig. 2). Every student receives an individual assignment to develop a working web application from scratch. Examples, templates and partly implemented code from the exercises are allowed to be used as reference. The assignment consists of four sub-assignments that upgrade one another and increase the final grade from "Sufficient D" to "Excellent A". An example assignment is shown in Fig. 2.

For the purpose of our experiments, the blank exam papers or assignments, as received by students, are photographed with an ordinary smartphone and sent to the three chatbots – ChatGPT, Claude and DeepSeek. The answers provided by the AIs are then checked for correctness and evaluated by the authors of this study. We use the same grading system that we apply to our students as well. It is based on the percentage of correct answers as follows:

< 40%, Fail F	70 – 84%, Very good B
40 – 54%, Sufficient D	85%+, Excellent A
55 – 69%, Good C	

Изпит по **„Търсене и извличане на информация"**, ОКС Магистър, 15.04.2025 г.

1. Кои са основните разлики между векторния модел за анализ на текст и латентния семантичен анализ?

2. Чрез кои мерки за сходство може да се изчисли семантична близост (степен на подобие) между два възела в таксономия?

3. Чрез кои показатели (метрики) се оценява точността/адекватността на търсенето?

4. Как може да се прецени дали подредбата на върнатите резултати е правилна?

5. По какъв начин (чрез кои мерки за сходство) може да се изчисли коефициент на подобие между два документа/обекта, описани чрез неподредени множества от ключови думи?

6. Защо описанието на документите чрез таксономия от ключови думи е по-подходящо, отколкото описанието чрез неподредено множество?

7. Как се представят документите (във вид на какво) при векторния модел за анализ на текст?

8. Защо е необходимо изчислените коефициенти на подобие да се нормализират спрямо дължината на векторите?

9. Защо при практическата реализация на векторния модел за анализ на текст всъщност се използва обърнат (инвертиран) индекс, а подобията не се изчисляват чрез векторите на документите?

10. Какви са основните разлики между плоските и йерархичните методи за клъстеризация?

11. Кои са двете основни характеристики, чрез които се изчисляват теглата на думите в документите/заявката при векторния модел за анализ на текст?

12. Защо е желателно да се премахнат семантично незначимите думи от текста преди прилагането на векторния модел или който и да е друг модел?

13. Как се изчислява семантичната близост между два документа при векторния модел за анализ на текст?

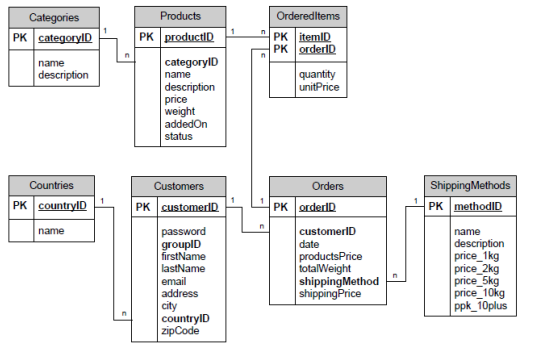
14. Какво представлява операцията стемиране и защо е препоръчително да се извършва преди векторния модел за анализ на текст?

15. Кои фактори влияят върху точността на коефициентите на подобие, изчислени чрез латентния семантичен анализ?

16. Избройте няколко начина, по които може да се изчисли разстоянието между два клъстера?

17. Кой метод позволява откриването и разпознаването на синоними като свързани думи – векторният модел за анализ на текст или латентният семантичен анализ?

Fig. 1. An example assignment for a theoretical exam in information retrieval.



За незнаещите английски:
Categories – категории
Products – продукти
Orders – поръчки
Ordereditems – поръчани артикули
Customers – клиенти
Countries – държави
ShippingMethods – методи за доставка

name – наименование
description – описание
price – цена
weight – тегло
addedOn – дата на добавяне
status – статус
zipCode – пощенски код
unitPrice – единична цена
quantity - количество

firstName – собствено име
lastName – фамилия
city – град
date – дата на поръчване
orderID – номер на поръчката
productsPrice – цена на продуктите
totalWeight – общо тегло на поръчката
shippingPrice – цена за доставка

Навсякъде, където пише "...ID", предпоследният символ е I (голямо i).

За достъп до БД:
host: **127.0.0.1** username: **pti_shop** password: **qwerty** db name: **pti_shop**

Да се реализира уеб приложение, което

а) (за 3) извежда *номерата на поръчките, датата на която са направени, имената на клиентите, които са ги направили и името на метода за доставка, с който са изпратени* за всички клиенти от България.

б) (за 4) Резултатите от подточка а) да се представят в таблица с видима рамка между клетките. Първият ред на таблицата трябва да съдържа заглавията на колоните.

в) (за 5) Да се надгради приложението така, че държавата да не е твърдо заложена от в sql заявката, а да се избира от потребителя от падащо меню.

г) (за 6) Падащото меню и таблицата с резултати да се разположат на един екран (страница). След извеждане на резултатите, избраната държава трябва да остава предварително маркирана в падащото меню.

Fig. 2. An example assignment for a practical exam in web programming.

IV. Results and Discussion

Results from the theoretical exams in Distributed Web Applications and Information Retrieval are shown in Table I and Table II, respectively.

TABLE I Evaluation of the Answers Provided by ChatGPT, Claude and DeepSeek to the Questions from the Exam in Distributed Web Applications

LLM Model	ChatGPT (GPT-3.5)	Claude (Sonnet v3.7)	DeepSeek (V3)
Result, % correct answers	**91 %**	**92 %**	**82 %**
Grade	Excellent A	Excellent A	Excellent A-

Obviously, all three LLM chatbots will get an Excellent A in distributed web applications. All models provide correct answers only, but as seen in Table I, they achieve different percentages. How is that possible? Since there are questions that require a subset of the correct answers, like "Specify two advantages of orchestration over choreography", that makes it possible. There are 4 to 5 advantages of orchestration over choreography, but they have different importance. Some of them are more important than others. DeepSeek, for example, provided two which are correct, but less important, so it does not get full points for this question.

TABLE II Evaluation of the Answers Provided by ChatGPT, Claude and DeepSeek to the Questions from the Exam in Information Retrieval

LLM Model	ChatGPT (GPT-3.5)	Claude (Sonnet v3.7)	DeepSeek (V3)
Result, % correct answers	**95 %**	**99.4 %**	**82 %**
Grade	Excellent A	Excellent A	Excellent A-

Similarly, when solving the exam in Information Retrieval, DeepSeek gives the lowest percentage of correct answers again, while Claude achieves 99.4%. In respect to language clarity, Claude does an excellent job generating beautiful sentences that sound like written by a real human. It should be mentioned here, that all exam materials are in Bulgarian, thus the AI answers are in Bulgarian as well.

The next challenge for the three LLM chatbots is to solve practical exams and write or generate real programming code. It is known that they are good in providing working code fragments or entire basic applications, but it is curious if they can correctly understand our specific assignments and generate working and efficient code. Again, the assignment is photographed with an ordinary smartphone and sent to them for execution. Results are summarized in Table III.

In terms of programming code, all models generate completely working code that satisfies all four sub-assignments, so the chatbots will get an Excellent A grade. ChatGPT and Claude use two additional and unnecessary arrays to cache data from the database, and thus two unnecessary cycles to create these arrays. When told that these two arrays are not necessary,

ChatGPT argues and motivates its decision to cache the data, while Claude agrees they could be omitted.

TABLE III EVALUATION OF THE PROGRAMMING CODE PROVIDED BY CHATGPT, CLAUDE AND DEEPSEEK AS A SOLUTION TO THE PRACTICAL EXAM IN WEB PROGRAMMING

LLM Model	ChatGPT (GPT-3.5)	Claude (Sonnet v3.7)	DeepSeek (V3)
Working code?	**Yes**	**Yes**	**Yes**
Optimal code?	Uses 2 unnecessary arrays and thus 2 unnecessary cycles to create them (but could motivate its decision why)	Uses 2 unnecessary arrays and thus 2 unnecessary cycles to create them	Yes
Grade	Excellent A	Excellent A	Excellent A

The results of our experiments with both theoretical and practical IT exams are similar and confirm those achieved by Puthumanaillam et al. [10] and Rytilahti and Kaila [11] in other subjects.

Since students are illegally trying to use artificial intelligence (AI) during their exams, it is interesting to know how the AI chatbots themselves would evaluate a student who, during a practical exam, has to create an application, has the right to use all Internet resources, without AI, but he or she cannot do anything alone. However, if he or she is allowed to use artificial intelligence, then he or she gets an excellent solution from the AI, but does not understand it.

Claude replies that, in its opinion, the student deserves "Fail F", because the essence of both education and assessment is to measure students' knowledge and skills, not their ability to find someone or something to do the work for them.

ChatGPT states that the grade should be "Sufficient D" because we should not only evaluate knowledge, but also the ability to solve problems. If the student can solve problems with the help of AI, then he or she has at least some ability to find a solution.

According to DeepSeek, the fair grade is "Sufficient D" since the problem is "solved", but the lack of understanding is a critical flaw in the educational context. The chatbot adds that, if the student can demonstrate how he or she used the AI, this shows some metacognitive skills and would even justify a "Good C" grade.

Although LLM chatbots seem to easily solve exams, there are other tasks they cannot solve, even if they generate perfectly working programming code for solving them. For example, the assignment problem [15] or other optimization tasks. When asked why they are able to generate a programming code to solve the task but cannot provide the solution directly, LLM chatbots reply that they generate the code as text based on their training, but they do not have an access to servers to run it. That is why they can provide just the code, but not the overall final solution. This means that the programming code they generate for the practical exams has not been tested before giving it to students, and thus the LLM chatbots cannot actually guarantee that it is really working.

V. CONCLUSION

A series of experiments have been conducted, aiming to test whether large language models (LLM)-based chatbots could solve exam tasks in real-time with no or minimum interaction. Results could be summarized as follows:

1) All the three AI chatbots (ChatGPT, Claude and DeepSeek) do an excellent job in solving both theoretical and practical exam tasks in the specified subjects, and would have earned an Excellent A grade.

2) Students are not required to have any special skills in working with AI, nor skills in how to ask questions in order to get an accurate and correct answer. They simply take a picture of the assignment and send it to the chatbot. This could happen illegally during the exam, even without the teacher noticing that.

3) When generating short open-ended answers in text format, Claude does the best job, answering most clearly and purposefully.

4) In general, when generating text in Bulgarian, Claude performs best by providing short, beautiful and human-like sentences.

5) DeepSeek answers the leanest and often tends to omit basic and well-known facts in its answer.

Our experimental results completely align with those of other researchers, showing that LLM chatbots are quite good at solving both theoretical and practical IT exams. Especially when it comes to factual (fact-finding) questions or programming code generation, LLMs will not just pass the exam, but will get an excellent A or B grade.

It seems that it is time to change the educational system again and shift the focus of assessment a little bit from the classical fact-finding and practical performance of elementary tasks to creativity, combinability, and skills for adapting and applying already gained knowledge.

ACKNOWLEDGMENT

This study is supported by the Scientific Research Fund of the "Angel Kanchev" University of Ruse, Bulgaria.

REFERENCES

[1] OPENAI. Introducing ChatGPT, 2022, https://openai.com/index/chatgpt/ (Accessed April 2025).

[2] ANTHROPIC PBC. Meet Claude – the AI for all of us, 2025, https://www.anthropic.com/claude (Accessed April 2025).

[3] GOOGLE DEEPMIND. Gemini – Google's most intelligent AI models, 2025, https://deepmind.google/technologies/gemini/ (Accessed April 2025).

[4] MICROSOFT. Copilot: Your AI companion, 2025, https://copilot.microsoft.com (Accessed April 2025).

[5] DEEPSEEK, 2025. https://www.deepseek.com/en (Accessed April 2025).

[6] KHAN ACADEMY. Meet Khanmigo: Khan Academy's AI-powered teaching assistant & tutor, 2025, https://www.khanmigo.ai/ (Accessed April 2025).

[7] GOOGLE. Socratic: Get unstuck. Learn better, 2025, https://socratic.org (Accessed April 2025).

[8] QUILLIONZ. World's First AI-Powered Question Generator, 2025, https://www.quillionz.com/ (Accessed April 2025).

[9] ELICIT. The AI Research Assistant, 2025, https://elicit.com/ (Accessed April 2025).

[10] G. Puthumanaillam, T. Bretl, and M. Ornik. (2025). The Lazy Student's Dream: ChatGPT Passing an Engineering Course on Its Own. arXiv preprint arXiv:2503.05760.

[11] J. Rytilahti, and E. Kaila. (2024). HOW EASY IS IT TO CHEAT?-SOLVING PROGRAMMING EXERCISES AUTOMATICALLY WITH AI. In Proceedings of the 20th International CDIO Conference. Proceedings of the 20th International CDIO Conference.

[12] A. VarastehNezhad, R. Tavasoli, M. Masumi and F. Taghiyareh, "LLM Performance Assessment in Computer Science Graduate Entrance Exams," 2024 11th International Symposium on Telecommunications (IST), Tehran, Iran, Islamic Republic of, 2024, pp. 232-237, doi: 10.1109/IST64061.2024.10843484.

[13] F. Burlacu. (2024). Patterns of success and failure: Analysing Large Language Models in Question Answering in Exam Contexts (Bachelor's thesis, University of Twente).

[14] S. K. Larsen. (2023). Creating Large Language Model Resistant Exams: Guidelines and Strategies. arXiv preprint arXiv:2304.12203.

[15] Y. Kalmukov, “An algorithm for automatic assignment of reviewers to papers”, Scientometrics, 2020, No 124 (3), pp. 1811–1850, https://doi.org/10.1007/s11192-020-03519-0.