Online Exams in the Era of ChatGPT

Erik Buchmann[12] and Andreas Thor[3]

Abstract: Recent versions of ChatGPT demonstrate an amazing ability to answer difficult questions in natural languages on a wide range of topics. This puts homeworks or online exams at risk, where a student can simply forward a question to the chatbot and copy its answers. We have tested ChatGPT with three of our exams, to find out which kinds of exam questions are still difficult for a generative AI. Therefore, we categorized exam questions according to a knowledge taxonomy, and we analyze the wrong answers in each category. To our surprise, ChatGPT even performed well with procedural knowledge, and it earned a grade of 2.7 (B-) in the IT Security exam. However, we also observed five options to formulate questions that ChatGPT struggles with.

Keywords: Online Exams, ChatGPT

1 Introduction

In Nov. 2022, OpenAI rolled out ChatGPT[4], a chatbot based on Generative Pretrained Transformer Models. It answers question from simple ("How do i clean my cat?") to challenging ("What are the security properties of homomorphic encryption?"). It generates texts ("Write an invitation letter for a DAAD travel grant"), and it masters different writing styles. Alternatives such as ChatSonic[5] even access Google search. This puts exams in danger, where the student is without supervision, e.g., bachelor theses or online exams.

In this work, we let ChatGPT answer three exams, and we grade it as we grade our students. We analyze the way ChatGPT answers, particularly the questions ChatGPT was not able to answer correctly, in the four dimensions factual, conceptual, procedural and metacognitive knowledge. Furthermore, we let a plagiarism tool scan the answers. Finally, we compile five options to formulate questions that ChatGPT cannot answer easily. We were surprised to see ChatGPT earning a grade of 2.7 (B-) in our IT Security exam, one point short of a 2.3 (B). ChatGPT correctly spotted the pseudo-code that caused a buffer overflow, and generated firewall rules from a textual description, i.e., it applied procedural knowledge. On the other hand, ChatGPT just passed the other exams. For example, it gives broad answers to specific questions, and struggles with numbers and calculations. In Section 2, we describe our experiments with ChatGPT. In Section 3, we discuss our findings. In Section 4, we compare our findings with related work. Section 5 concludes.

[1] Dept. of Computer Science, Leipzig University, Germany. buchmann@informatik.uni-leipzig.de
[2] Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) Dresden/Leipzig, Germany
[3] Leipzig University of Applied Sciences, Germany. andreas.thor@htwk-leipzig.de
[4] https://chat.openai.com/chat
[5] https://writesonic.com

doi: 10.18420/delfi2023-15

2 Examining ChatGPT

We tested ChatGPT, release Jan 30, 2023, with three different exams that had to be answered in 90 minutes. The *IT Security exam* for international master students of computer science and business informatics consists of 18 questions in the topics "Access Control", "Encryption", "Secure Programming", "Network Security", and "IT-Grundschutz". Figure 1 shows a subset of the exam questions.

```
I1) Explain the difference between vulnerability and threat by using an example.
I2) Why is the private key of the RSA crypto system a prime number?
I3) Explain how the One-time Pad approach works.
I4) Is the One-time Pad approach more secure than a public-key crypto system?
I5) Describe the seven steps of the IT Security Process (BSI-Standard 200-2).
I6) Which line in the pseudo-code is vulnerable against a buffer overflow, and why?
        bool verifyWallet() {
                char wallet[512]; // declare buffer of 512 Bytes
                readFile(wallet, "C:\wallet.cry"); // read wallet file
                return (verify(wallet)); // validate the wallet
        }
```

Fig. 1: Questions from the IT Security exam

According to Bloom's revisited taxonomy [Kr02], the questions test factual (I3, I5), conceptual (I1, I2), procedural (I6) and, to some extent, metacognitive knowledge (I4). To exclude arbitrary answers, several questions build on each other (I3, I4). Figure 2 exemplarily shows the answer to question I2.

```
The private key in the RSA cryptography system is a prime number because of the
mathematical properties of prime numbers. The RSA encryption system is based on the
mathematical properties of prime factorization, where a large composite number can be
factored into two prime numbers. The security of RSA encryption is based on the
difficulty of finding the prime factors of a large composite number, which is used as
the private key. If someone were to obtain the private key, they would be able to
decode any message encrypted with that key, so it's important to keep the private key
secure.
```

Fig. 2: ChatGPT's answer to I2.

Table 1 shows the allocation of the points of the IT Security exam to the dimensions of knowledge. Numbers in brackets denote the points ChatGPT earned.

Exam	Factual Knowledge	Conceptual Knowledge	Procedural Knowledge	Metacognitive Knowledge
IT Security	12 (9)	51 (33)	24 (20)	3 (2)
Data Privacy	28 (11)	41 (19)	15 (11)	6 (5)
Foundations of CS	3 (0)	20 (9)	26 (17)	2 (0)

Tab. 1: Allocation of points to knowledge dimensions

The German *Data Privacy exam* has the same structure and target group as the IT Security exam. It tests the areas "Societal Foundations", "Anonymity Measures", "Internet Privacy", and "Internet of Things Privacy". As Table 1 shows, this exam emphasizes

conceptual knowledge, and has a stronger focus on factual knowledge than the IT Security exam.

The exam *Foundations of Computer Science* targets at German Bachelor students of computer science in the first semester. This exam focuses on conceptual and procedural knowledge, as shown in Table 1: 46 of 51 points can be earned by answering questions like "What is the Hamming distance of the code words for A and D?" or "Transfer a pseudo-code into a flowchart". To avoid guessing, the students had to explain their answers. The exam addresses the areas "Data Structures", "Encoding Theory", "Computing Architectures", "Fundamentals of Programming" and "Logic". We omitted two image-based questions, because the version of ChatGPT we experimented with did not understand images.

Experiment Procedure: At first, we fed ChatGPT with our exam questions, one by one, for all three exams. If ChatGPT obviously did not understand, we rephrased the question as a student would do in an exam. We graded the answers with the same scheme as for the real exam. We also applied a commercial plagiarism tool[6]. We used a qualitative approach to analyze all answers that were not entirely correct, and we compiled hypotheses why ChatGPT might have given a wrong answer. We consolidated the hypotheses, and we tried to verify those hypotheses with different and rephrased questions. Finally, we derived common characteristics of questions that are challenging for ChatGPT (see Section 3).

Evaluation of the Answers: For each exam, Table 1 lists ChatGPT's score in brackets for each dimension of knowledge. In the *IT Security exam*, ChatGPT scored 64 from 90 points. This is sufficient for a 2.7 (B-), one point short of a 2.3 (B). As expected, a knowledge model is good with of factual knowledge. ChatGPT also demonstrated procedural knowledge in many cases, e.g., to assess pseudo-code (I6) and to generate firewall rules. However, it struggled to apply concepts to very specific cases, or to compare two concepts. For example, ChatGPT could answer I3, but it did not answer I4 correctly. Sometimes, ChatGPT answered inconsistently, e.g., Figure 2: The first sentence of the answer says that the private key is a prime number (wrong), followed by the explanation that the private key is the composite of two prime numbers (true). For I5, ChatGPT invented ("hallucinated") some steps.

In the Data Privacy exam, ChatGPT obtained 46 of 90 points, which means a grade of 4.0 (D). In this exam, ChatGPT missed many points for factual and conceptual knowledge. A closer inspection revealed, that this exam frequently asks for facts and concepts in relation to very specific use cases, e.g., "What are the differences between the interaction model of the Internet of Things and the interaction model of traditional PC applications?" We already know from the IT Security exam that this kind of question is difficult for ChatGPT. In the exam Foundations of Computer Science, ChatGPT earned 25 out of 51 points, which corresponds to a grade of 4.0 (D). Again, we observed that ChatGPT struggled with some calculations and questions that required a specific answer instead of a broad one.

[6] https://www.plagaware.com/

Identifying ChatGPT's Answers: For our *IT Security exam*, the plagiarism tool found that only 79 out of 2434 words in four sentences might have been paraphrased. The results for *Data Privacy* and *Foundations of Computer Science* were also not sufficient to prove plagiarism. Others tested AI tools [AJ23] for detecting AI-written text, with little success.

However, ChatGPT's answers have a unique pattern: At first, ChatGPT repeats the most important keyword(s) from the question. Then it relates the question to a generalized concept. ChatGPT always finishes with a summarizing sentence, even if it is repetitive or contradicting to the preceding text. ChatGPT often provides unnecessary details. For example, compare I2 with its answer (Figure 2). The answer starts by repeating the keywords "RSA cryptography" and "prime numbers". It describes the concept of RSA, and it concludes by saying that the private key must be kept secure, although this was not part of the question. It may indicate the use of ChatGPT, if this pattern can be found multiple times in an exam.

3 Exam Questions That are Difficult for ChatGPT

In this section, we list options for questions that are difficult for ChatGPT. We observed them in answers across all three exams, and we verified them with follow-up questions.

O1: Suggestive numbers of steps. ChatGPT tends to respond to suggestive numbers of steps, phases, cycles, stages, etc. Consider I5: ChatGPT invents any number of steps the questioner asks for, at least for reasonable numbers.

O2: Specific answers. Some question call for very specific cases. ChatGPT gives a perfect answer to I1, because this asks for a generic concept. However, ChatGPT still produces a generic answer, if I1 is turned into a specific question like "Explain the difference between vulnerability and threat for a UDP connection between a cloud server and an IoT device."

O3: Contradictions and repetitions. Control questions consider a subject from multiple perspectives, e.g., "(a) Does the code contain a buffer overflow?" and "(b) In which line is the buffer overflow?". If (a) is answered with "No", a consistent answer to (b) must be "There is no such line". We observed ChatGPT giving contradicting or repetitive answers, even within the same response (cf. Figure 2).

O4: Charts and figures. The version of ChatGPT we experimented with did not understand images. Thus, questions such as "If the certificate authority in a figure is exposed, which certificates are invalid?" could not be answered. However, in the meantime OpenAI offers a subscription model of GPT-v4 that also interprets images.

O5: Math beyond elementary school. The math knowledge of ChatGPT is limited. Questions like "Multiply two dual numbers 1101 and 101 in the dual system and explain your path to the solution." were answered incorrectly. For some math-related questions, ChatGPT produced an explanation that did not match the results it calculated.

4 Related Work

Related work for a *hot topic* like ChatGPT is practically always incomplete, as new field reports, analyses and handouts are published almost daily. Furthermore, since ChatGPT has only been available to the general public for a few months, there are hardly any peer-reviewed papers. The discussion on how ChatGPT can be used in university teaching, especially for exams, and whether this puts exam integrity in danger [Su22] is in full swing[7]. On one side are the "preventers" who want to ban and prevent the use of AI-based systems through rules, such as a return to pen and paper exams [Ca23], and software tools[8] to automatically detect whether text submitted by students has been auto-generated. On the other side are the "proponents" [Ru23; Sp23], who not only allow but also encourage the use of such tools. They emphasize good scientific practice (naming all tools and sources) and media literacy (assessing the possibilities and limitations of tools and sources).

Similar to our approach, other researchers have had ChatGPT answer exam questions from their fields of expertise, e.g., business administration [Te23] and law [Ch23]. Their experiences are similar to ours. In all cases ChatGPT would have passed the exam with grades between 2 (B) and 3 (C). ChatGPT is surprisingly good at describing facts and topics, which is especially evident in good performance on essay writing tasks. In contrast, the results for multiple choice questions were worse, but still significantly better than random guessing, as clueless students would do.

Regardless of the task format, ChatGPT surprisingly makes errors in simple mathematical calculations. ChatGPT's ability to change its answer through further hints allows ChatGPT to correct itself or to handle the task correctly.

Furthermore, ChatGPT can be effectively used in teaching and for the creation of exams [Mo23; Pr23]). For example, distractors for multiple choice questions or suggestions for exam questions in general can be auto-generated by ChatGPT. To avoid that such questions can be answered too easily by ChatGPT, teachers should (re-)formulate questions that challenge a computer more than a human [Su22]. These include multi-hop questions that combine multiple facts, and questions that require logical reasoning to answer.

This shows that students and teachers have to acquire the competence of *prompt engineering* [Wh23] in the long run, in order to be able to use ChatGPT efficiently for their studies also outside of exams. A prompt is a set of instructions provided to ChatGPT to customize the dialog. Examples for prompt engineering are question refinement, if ChatGPT apparently did not understand the question, or setting specific contexts.

[7] https://hochschulforumdigitalisierung.de/de/dossiers/generative-ki
[8] https://gptzero.me

5 Conclusion

Generative Pretrained Transformer Models such as ChatGPT have reached a degree of maturity, that puts any exam at risk, where a student is without supervision. The generated answers are seemingly convincing, and cannot be reliably identified by tools. To find out if there are questions an AI cannot easily answer, we tested ChatGPT with three different exams. ChatGPT demonstrated factual, conceptual and procedural knowledge, but it struggled with some calculations, questions aiming for specific (instead of broad) answers, and it invents procedure descriptions. Future approaches might overcome these issues.

Bibliography

[AJ23] Alimardani, A.; Jane, E. A.: We pitted ChatGPT against tools for detecting AI-written text, and the results are troubling, https://theconversation.com/we-pitted-chatgpt-against-tools-for-detecting-ai-written-text-and-the-results-are-troubling-199774, retrieved: March 2023, 2023.

[Ca23] Cassidy, C.: Australian universities to return to 'pen and paper' exams after students caught using AI to write essays. In: The Guardian. 2023.

[Ch23] Choi, J. H.; Hickman, K. E.; Monahan, A.; Schwarcz, D. B.: ChatGPT Goes to Law School. In: Minnesota Legal Studies Research Paper No. 23-03. 2023.

[Kr02] Krathwohl, D. R.: A Revision of Bloom's Taxonomy: An Overview. Theory Into Practice 41/4, pp. 212–218, 2002.

[Mo23] Mohr, G. et al.: Übersicht zu ChatGPT im Kontext Hochschullehre, https://www.hul.uni-hamburg.de/selbstlernmaterialien/dokumente/hul-chatgpt-im-kontext-lehre-2023-01-20.pdf, retrieved: March 2023, 2023.

[Pr23] ProLehre TUM: Einsatz von ChatGPT in der Lehre, https://www.prolehre.tum.de/fileadmin/w00btq/www/Angebote_Broschueren_Handreichungen/prolehre-handreichung-chatgpt-v2.2.pdf, retrieved: March 2023, 2023.

[Ru23] Rudolph, J. et al.: ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? Journal of Applied Learning & Teaching 6(1)/, 2023.

[Sp23] Spannagel, C.: Rules for Tools, https://csp.uber.space/phhd/rulesfortools.pdf, retrieved: March 2023, 2023.

[Su22] Susnjak, T.: ChatGPT: The End of Online Exam Integrity?, 2022, url: https://arxiv.org/abs/2212.09292.

[Te23] Terwiesch, C.: Would Chat GPT3 Get a Wharton MBA? A Prediction Based on Its Performance in the Operations Management Course, Mack Institute for Innovation Management, University of Pennsylvania, 2023.

[Wh23] White, J.; Fu, Q.; Hays, S.; Sandborn, M.; Olea, C.; Gilbert, H.; Elnashar, A.; Spencer-Smith, J.; Schmidt, D. C.: A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT, 2023, url: https://arxiv.org/abs/2302.11382.