Evaluating GPT-4's Proficiency in Addressing Cryptography Examinations

Vasily Mikhalev, Nils Kopal, and Bernhard Esslinger

University of Siegen, Germany
vasily.mikhalev@uni-siegen.de

Abstract. In the rapidly advancing domain of artificial intelligence, ChatGPT, powered by the GPT-4 model, has emerged as a state-of-the-art interactive agent, exhibiting substantial capabilities across various domains. This paper aims to assess the efficacy of GPT-4 in addressing and solving problems found within cryptographic examinations. We devised a multi-faceted methodology, presenting the model with a series of cryptographic questions of varying complexities derived from real academic examinations. Our evaluation encompasses both classical and modern cryptographic challenges, focusing on the model's ability to understand, interpret, and generate correct solutions while discerning its limitations. The model was challenged with a spectrum of cryptographic tasks, earning 201 out of 208 points by solving fundamental queries inspired by an oral exam, 80.5 out of 90 points on a written Crypto 1 exam, and 287 out of 385 points on advanced exercises from the Crypto 2 course. The results demonstrate that while GPT-4 shows significant promise in grasping fundamental cryptographic concepts and techniques, certain intricate problems necessitate domain-specific knowledge that may sometimes lie beyond the model's general training. Insights from this study can provide educators, researchers, and examiners with a deeper understanding of how cutting-edge AI models can be both an asset and a potential concern in academic settings related to cryptology. To enhance the clarity and coherence of our work, we utilized ChatGPT-4 to help us in formulating sentences in this paper.

1 Introduction

Artificial Intelligence (AI) stands as a revolutionary technological advancement, profoundly impacting a myriad of domains by emulating and augmenting human intelligence. The field of AI has demonstrated substantial growth, offering unprecedented capabilities, notably in the creation of conversational agents such as ChatGPT which was launched on November 30, 2022, by OpenAI.

ChatGPT is a large language-model-based chatbot that provides users with the capability to refine and steer conversations, ensuring they adhere to a desired length, format, style, level of detail, and language. ChatGPT is built on generative pre-trained transformer (GPT) models. As of the time this paper was written, the GPT-3.5 and GPT-4 versions are available. From this point onward, when referencing the model name alone, it is to be understood that we are indicating ChatGPT, which utilizes the corresponding model.

Despite the expansive growth and applications of AI, there has been a gap in the assessment of its capabilities within the field of cryptography. Cryptography is pivotal in securing information, and evaluating potential and limitations of AI in this domain is crucial for understanding its application and implications. To this end, this work serves as a pioneering exploration into the proficiency of AI, particularly GPT-4, in addressing and solving cryptographic problems, marking the first comprehensive attempt to assess AI's impact on cryptographic examinations.

This study integrates multiple scenarios to evaluate the efficacy of GPT-4, focusing on various cryptographic problems derived from academic examinations of different complexities. These scenarios are crucial, granting higher precedence to comprehend the model's adaptability and understanding within cryptographic contexts, allowing a comprehensive investigation into its ability to interpret, solve, and generate correct solutions to a range of cryptographic challenges.

Our results indicate that GPT-4 has shown significant promise, managing to grasp fundamental cryptographic concepts and solve a multitude of problems, demonstrating a proficiency level considered excellent for a student. However, the model does encounter limitations when dealing with specific intricate problems that require an in-depth, domain-specific understanding.

This paper is structured as follows: In Section 2, we discuss the applications and proficiency of GPT-4 by using exams from various disciplines. Section 3 offers a detailed overview of our applied methodology.

Subsequently, Section 4 showcases the results. In Section 5, we articulate our observations on the capabilities and constraints of GPT-4 in solving cryptology exams. Finally, Section 6 concludes the paper.

2 Related work

In this section, we provide an overview of related work that have already assessed the performance of GPT-3.5 and GPT-4 in various fields and their abilities to tackle exams. To the best of our knowledge, none of these models have been assessed so far in terms of their knowledge of cryptography. We note that several authors refer to GPT-3.5 as ChatGPT, while they distinguish the latest model as GPT-4. It is our understanding that when ChatGPT was initially released, its underlying model GPT-3.5 was also labeled as ChatGPT, leading to the source of confusion. The summaries of the different articles and papers were written with the help of ChatGPT.

We considered five different areas where GPT was used:

Mathematics: The authors of [2] tested GPT-3.5 and GPT-4 on challenging mathematical tasks using the newly introduced GHOSTS and miniGHOSTS datasets. While GPT-3.5 proved adept at assisting with factual queries and mathematical searches, it struggled with advanced mathematics. GPT-4 showed promise in handling undergraduate-level math but fell short on graduate-level problems. Despite their inconsistencies, both models exhibited moments of surprisingly high-quality answers. The authors emphasize the value of ChatGPT as a universal tool for mathematicians while calling for further contributions to the GHOSTS dataset to establish a comprehensive benchmark for evaluating language models' mathematical capabilities.

[10] delves into the perspectives of stakeholders regarding GPT-3.5 role in teaching mathematics. Initial findings indicate it's potential to enhance educational success by imparting basic math knowledge, particularly in geometry. However, caution is advised due to limitations, such as its lack of deep understanding in certain areas. The study suggests the need for guidelines and teacher-student proficiency with chatbot technology. It emphasizes the importance of a new educational philosophy and responsible chatbots, offering both theoretical insights and practical implications for the safe integration of AI, like ChatGPT, into mathematics education.

Medicine: [3] assesses the performance GPT-3.5 in answering questions from the United States Medical Licensing Examination Step 1 and Step 2 exams, comparing it with other language models. It achieved accuracies ranging from 42% to 64.4% across different datasets, surpassing InstructGPT by 8.15% on average. Logical justification was present in all outputs for some datasets, and internal information to the question was found in the majority of cases. The study demonstrates GPT-3.5 potential as a valuable tool for medical education, achieving passing scores for third-year medical students and providing contextual and logical answers to medical queries.

In [4] the authors evaluate GPT-3.5 performance on the United States Medical Licensing Exam (USMLE) Step 1, Step 2CK, and Step 3 without specialized training. ChatGPT achieved accuracy levels exceeding the pass thresholds, indicating its potential to assist with medical education and clinical decision-making. The study highlights ChatGPT's increasing accuracy over previous models, surpassing 60% accuracy, and suggests its proficiency in handling broader clinical content. ChatGPT's performance varied across exam steps, mirroring the perceived difficulty, and the study discusses its potential to enhance medical education and question-explanation writing processes. It also emphasizes the need for standardized research methods to assess human-AI interactions in medical education.

[5] evaluates the performance of GPT-4 on medical competency examinations and benchmark datasets, including the United States Medical Licensing Examination (USMLE) and the MultiMedQA suite. GPT-4 surpasses the passing scores for USMLE exams by more than 20 points and outperforms both earlier general-purpose models (GPT-3.5) and models specifically fine-tuned for medical knowledge. The model also demonstrates improved calibration of answer probabilities. The study explores GPT-4's ability to explain medical reasoning, personalize explanations, and create counterfactual scenarios. It suggests potential uses of GPT-4 in medical education and clinical practice while emphasizing the need for caution, error consideration, and risk mitigation in real-world applications.

Law: [1] examines GPT-3.5 autonomous performance in answering law school exam questions at the University of Minnesota Law School. ChatGPT achieved an average grade equivalent to a C+ student across four courses, including multiple-choice and essay questions, without human assistance. The findings suggest potential applications of ChatGPT in legal education and writing. The paper also offers guidance on prompt engineering for legal writing, addressing tone, word limits, citations, and essay length. While GPT-3.5 showed promise, the study emphasizes the need for prudent use, considering possible errors and challenges in real-world scenarios.

MBA: [8] evaluates performance of GPT-3.5 on an MBA Operations Management final exam. GPT-3.5 excels in basic operations management and process analysis questions, with correct answers and explanations. However, it makes significant errors in simple calculations and struggles with advanced process analysis questions. Notably, it can adapt its responses based on human hints, correcting itself when initially wrong. Overall, its performance is akin to a B to B- grade, raising important considerations for business school education and the role of AI in exams and collaborative learning.

Surveys discussing various scientific fields: Paper [11] reviews the evolution of large language models (LLMs) and their impact on AI. The survey covers four aspects: pre-training, adaptation tuning, utilization, and capacity evaluation. It highlights the rapid advancements and discusses available resources and future challenges in the field of LLMs.

Article [9] discusses the academic achievements of AI models, GPT-3.5 and GPT-4, across various exams, including the bar exam, SAT, GRE, USA Biology Olympiad, AP exams, AMC exams, and sommelier examinations. It highlights GPT-4's superior performance compared to GPT-3.5 in multiple exams. The article also mentions ChatGPT's performance in law school exams, medical licensing exams, essays, and a clinical reasoning final at Stanford Medical School. These results raise implications for education and professional assessment, as AI models continue to demonstrate their capabilities in different domains.

In the technical report [6] GPT-4 achieves human-level or better performance on professional and academic benchmarks, notably excelling in a simulated bar exam, ranking in the top 10%. The report highlights scalable infrastructure and predictive capabilities. Despite its advancements, GPT-4 introduces new risks, including bias, disinformation, over-reliance, privacy concerns, cybersecurity threats, and issues related to proliferation. These risks require ongoing safety measures, marking a substantial step toward versatile and secure AI systems.

[7] provides an extensive review of ChatGPT's impact on scientific research, highlighting its applications, technological development, and potential challenges. It underscores ChatGPT's role in enhancing research efficiency, collaboration, and innovation, along with its proficiency in contextual understanding, language generation, task adaptability, and multilingual capabilities. However, ethical concerns and biases are acknowledged as challenges that need to be addressed. Despite these issues, ChatGPT holds promise for shaping the future of AI-driven research, provided ethical considerations are thoroughly managed, enabling responsible AI integration into scientific endeavors.

3 Methods

In our study, we aimed to evaluate the capabilities of GPT-4 in answering cryptographic exams across three distinct scenarios.

1. The first scenario encompassed fundamental cryptographic queries, covering the core terminology and overarching principles. The compiled list of questions was inspired by oral exams held by Prof. Arno Wacker, formerly professor of the "Applied Information Security" work group of the University of Kassel, and one of the authors of this paper, when he worked as a scientist at that work group. The lecture was called "Grundlagen der angewandten Kryptologie" (Engl. "Fundamentals of applied cryptology") and was a basic lecture for students with no previous knowledge about cryptography.
2. The second scenario involved evaluating GPT-4 on a real written exam that was administered to the students of the University of Mannheim in 2019 for the course "Crypto 1".
3. For our third scenario, we assessed the capabilities of GPT-4 in tackling the more complex exercises from the "Crypto 2" course. These exercises, curated by Prof. Frederik Armknecht, were used at the

University of Mannheim to impart advanced knowledge in cryptography. The exercises were consistently employed from 2010 through to 2020 for students seeking a deeper understanding of the subject. These questions were not offered to the students at the exams but rather were discussed in the exercise sessions. The students were given these tasks a week prior to the session, allowing them ample time to strategize solutions using various resources such as literature, internet access, and peer discussions.

We questioned ourselves about the easiest way to provide a formal mathematical description to GPT-4 so that it comprehends the task. After considering a couple of methods, we concluded that providing the LaTeX code of the mathematical description was the easiest and most effective approach. In the majority of cases, this enabled GPT-4 to precisely understand the problem. We employed this method in both the second and third scenarios.

To implement this, we utilized the chat.openai.com application with the Plus subscription to gain access to GPT-4's capabilities, the version that was available on August 07, 2023. Questions were extracted directly from our LaTeX files and then copy-pasted into the chat interface. Significantly, we ensured that all questions corresponding to a single scenario were presented to ChatGPT in the same session.

When evaluating the results, we tried to follow the same criteria as we would use when evaluating the answers given by real students.

3.1 Fundamental cryptographic queries

In our first test set, we compiled a large list of basic questions about cryptography. The basic questions (a total of 104 questions) given to GPT-4 encompassed:

- **Basics of cryptology (4 questions)**: Differentiating between cryptology, cryptography, and cryptanalysis.
- **Classical ciphers (11 questions)**: A deep dive into various classical ciphers, their categorization, and specific characteristics.
- **Statistical measures in cryptography (9 questions)**: Discussion on key statistical metrics such as the Index of Coincidence, entropy, and the concept of keyspace.
- **Cryptanalysis of classical ciphers (8 questions)**: Techniques and tests to decipher or break classical ciphers, including frequency analysis and various attack methodologies.
- **Machine ciphers (8 questions)**: Comprehensive exploration of the Enigma machine, its components, operation, vulnerabilities, and methods to break its code.
- **Security objectives in modern cryptography (8 questions)**: Delving into the primary security goals, understanding perfect security, and the intricacies of the one-time-pad.
- **Symmetric modern ciphers (18 questions)**: An analysis of symmetric cryptography, its types, standard protocols, and specific components such as the S-box and P-box.
- **Cryptographic hash functions (10 questions)**: Insights into cryptographic hash functions, their properties, and associated attacks.
- **Asymmetric modern ciphers (Focus on RSA – 7 questions)**: A deep dive into RSA, hybrid cryptography, and the significance of certificates and certificate authorities.
- **Key exchange (6 questions)**: Addressing the key exchange problem and understanding protocols like Diffie-Hellman.
- **Cryptographic protocols (11 questions)**: A study of various cryptographic protocols, their importance, encryption methods, and recommendation on key sizes for ciphers.
- **Quantum cryptography (4 questions)**: Introduction to quantum cryptography, post-quantum cryptography, the BB84 protocol, and lattice-based cryptography.

3.2 Written exam from Crypto 1 course at University of Mannheim (2019)

The exam, conducted in German, presented a comprehensive test of both theoretical knowledge and practical application. The core topics covered in the examination were:

1. **Grundlegendes (fundamentals):** This section tested the foundational concepts of cryptography. Participants were expected to understand and explain basic cryptographic terms, principles, and classifications.

2. **Entropie (entropy):** Examinees were required to demonstrate their grasp over the concept of entropy, particularly its significance in cryptography. Tasks involved calculating entropy values and understanding their implications.
3. **LFSRs und Schlüsselstromgeneratoren (LFSRs and keystream generators):** This topic assessed the understanding of linear feedback shift registers (LFSRs) and how they function within keystream generators. Practical tasks included analyzing simple keystream generators using LFSRs and explaining their weaknesses.
4. **Blockchiffren (block ciphers):** Participants needed to showcase their proficiency in block cipher algorithms. This involved explaining their operations, strengths, and weaknesses. The practical task included explanation of a meet-in-the-middle attack against 2-DES, evaluation of the time complexity of the attack, and analyzing the suggested improvement of the scheme.
5. **RSA:** This section focused on the RSA public-key cryptographic algorithm. Tasks ranged from explanation of the scheme, encryption and decryption using RSA to discussing its security properties.
6. **Hashfunktionen (hash functions):** Examinees were tested on their understanding of hash functions. This included their properties, use-cases, and potential vulnerabilities. Practical task included explanation of how the birthday paradox could help to reduce the time of finding 2 colliding inputs.

3.3 Advanced exercises from the crypto 2 Course at University of Mannheim (2010-2020)

The exercises provided in the third scenario address advanced cryptographic concepts and their applications, with a focus on computational security, cryptanalysis, and formal cryptographic properties.

1. **Crypto mathematics:**
   - Investigate and provide proofs regarding the negligibility of specific functions.
2. **Stream ciphers:**
   - Conduct a guess-and-determine attack targeting the Geffe generator.
   - Analyze and determine the probability of biases present in the output.
   - Execute a key recovery attack on the modified Geffe generator utilizing the simplest algebraic approach.
   - Undertake a more intricate algebraic attack, recognizing and leveraging specific properties.
   - Implement an algebraic attack on a more advanced but yet straightforward keystream generator (KSG).
   - Compute the output derived from LFSR and determine its period.
   - Strategize attacks on various simple KSGs.
   - Explain an attack on the Geffe generator using detailed parameters.
   - Design and explain a challenging attack strategy for a more advanced KSG.
3. **Block ciphers:**
   - Estimate the time requirement to breach AES 128 considering specified hardware conditions.
   - Evaluate the security robustness of a simplified AES, where certain operations are excluded.
4. **Feistel networks:**
   - Validate the pseudorandomness of Feistel networks across various number of rounds, constructed using a pseudorandom round function.
   - Prove a specific property of Feistel networks.
5. **Differential cryptanalysis:** Formulate differentials for a designated S-box.
6. **Linear cryptanalysis:** Ascertain and prove a linear property associated with an S-box.
7. **Hash functions:**
   - Calculate the likelihood of output alteration when modifying several input elements across diverse functions.
   - Scrutinize a given function to determine if it qualifies as a hash function, ensuring its preimage resistance, second preimage resistance, and/or collision resistance.
   - Interpret the interdependencies between various hash function attributes and discern their implications.
8. **Provable security:**
   - Demonstrate that a specific probabilistic RSA scheme is vulnerable under the IND-CPA model.
   - Establish and prove particular properties of the ElGamal encryption scheme.

- Determine the conditions necessary for a scheme to achieve perfect security.
- Evaluate the absolute secrecy of various schemes.

9. **Public-key encryption schemes:**
   - Apply the baby-step/giant-step algorithm.
   - Implement the ElGamal encryption scheme using predefined parameters.
   - Utilize the RSA encryption scheme with set parameters.
   - Strategize and implement an attack on the standard RSA in real-world conditions.
   - Verify the decryption functionality of a proposed randomized RSA scheme and propose a known-plaintext attack.
   - Expose the vulnerabilities of RSA by employing the Chinese remainder theorem.

4 Results

This section provides a comprehensive overview of GPT-4 performance across the three cryptographic assessments.

4.1 Fundamental queries

Table 1 summarizes the results of GPT-4 working on the fundamental cryptographic queries. The AI (or student) received 2 points for a completely correct answer, 1 point for a partially correct answer, and 0 points if they could not answer correctly at all.

Task field	Max points	Points earned
Basics of cryptology	8	8
Classical ciphers	22	20
Statistical measures in cryptography	18	16
Cryptanalysis of classical ciphers	16	15
Machine ciphers	16	15
Security objectives in modern cryptography	16	16
Symmetric modern ciphers	36	35
Cryptographic hash functions	20	20
Asymmetric modern ciphers	14	14
Key exchange	12	12
Cryptographic protocols	22	22
Quantum cryptography	8	8
Total	208	201 (96.63%)

Table 1. Results in solving fundamental queries (oral exam)

In total GPT-4 received 201 points out of 208, which corresponds to 96.63% correctness. In the real oral exam this was the best possible grade of 1.0.

4.2 Written exam on Crypto 1

We summarized the results of GPT-4 working on this exam in the Table 2.

Task field	Max points	Points earned
Basics	10	9.5
Entropy	12	9
LFSRs and keystream generators	20	20
Block ciphers	16	12.5
RSA	13	10.5
Hash functions	19	19
Total	90	80.5 (89.44%)

Table 2. Results in solving a real Crypto 1 exam

In total GPT-4 received 80.5 points out of 90, which corresponds to 89.44% correctness. In the real written exam this was the best possible grade of 1.0. In fact the best student earned 76 points and received the same grade.

4.3 Crypto 2 advanced exercises

The summary of solving the advanced exercises from the course Crypto 2 are given in Table 3.

#	Task	Pts	Solved	Got	Comments
Crypto mathematics					
1.1	Prove ineligibility of functions	9	100%	9	Tasks likely used in training.
Stream ciphers					
2.1.a	Guess and determine on Geffe	10	95%	9.5	Good but simpler approach exists.
2.1.b	Evaluate output biases	10	100%	10	
2.1.c	Attacking modified Geffe generator	10	100%	10	
2.1.d	Notice certain algebraic property	15	20%	3	Mistake at the beginning.
2.2	Algebraic attack on KSG	12	20%	2.4	Mistake at the beginning.
2.3	Compute LFSR output and period	7	60%	4.2	Correct output, but period not found.
2.4	Attack other KSGs	10	95%	9.5	Valid but inefficient.
2.5	Geffe generator attack	10	60%	6	Good algorithm, wrong computations.
2.6	Harder KSG attack	12	20%	2.4	Valid but off-task.
Block ciphers					
3.1	Time to break AES 128	4	100%	4	
3.2	Security of simplified AES	28	100%	28	
Feistel networks					
4.1	Pseudorandomness of Feistel networks	27	70%	18.9	False assumption for 1 out of 3 tasks.
4.2	Prove Feistel properties	10	100%	10	
Differential cryptanalysis					
5.1	Construct S-box differentials	12	80%	9.6	Correct start, later comp. errors.
Linear cryptanalysis					
6.1	Prove linear S-box property	12	100%	12	
Hash functions					
7.2	Probability of flipping output by flipping inputs	9	60%	5.4	4/9 probabilities correct; issue with boolean function evaluations.
7.3	Evaluate hash function properties	24	66%	15.84	8/12 answers correct.
7.4	Hash function attribute implications	12	80%	9.6	False preimage resistance implication.
Provable security					
8.1	Unsecure RSA in IND-CPA	15	30%	4.5	Correct approach, misinterpreted task.
8.2	Properties of ElGamal scheme	18	100%	18	
8.3	Perfect security conditions	24	66%	15.84	Misunderstood task.
8.4	Evaluate perfect secrecy schemes	15	67%	10.05	2/3 schemes evaluated correctly.
Public-key encryption schemes					
9.1	Apply baby-step/giant-step algorithm	12	80%	9.6	Correct algorithm, incomplete computations.
9.2	Use ElGamal with parameters	9	90%	8.1	Arithmetic error.
9.3	Use RSA with parameters	9	70%	6.3	Correct formula, wrong computations.
9.4	Attack textbook RSA in practice	10	100%	10	
9.5	Validate RSA decryption	18	100%	18	
9.6	RSA vs Chinese remainder theorem	12	60%	7.2	Correct formulas, wrong computations.
In total		**385**	**74.53%**	**286.93**	

Table 3. Results in solving advanced crypto exercises

In total, GPT-4 was able to obtain 287 out of 385 points which is about 75 %. Given the complexity of the tasks, a student who would show the same performance would be considered excellent.

5 Discussion

We took a close look at how GPT-4 handles various cryptographic problems and drew some key conclusions from the different outcomes that we observed:

- **Mathematical proficiency**: GPT-4 showcased an adept understanding of LaTeX math formatting and many cryptographic algorithms. For simpler arithmetic tasks like computations in GF2, multiplication, and division, the model was consistent and accurate. However, its accuracy waned with more complex tasks. For instance, modular operations were often calculated incorrectly, as were evaluations of boolean functions with multiple arguments.

- **Solution consistency**: There were instances where the model displayed knowledge of how to solve a problem, evident from its previous responses. However, it occasionally refrained from computing the specific example provided, offering comments like "finding exact value requires deeper analysis".
- **Error propagation**: Mistakes in initial computations often led to cascading errors. For instance, while the first entries of a difference distribution table might be accurate, subsequent errors often resulted in further inaccuracies.
- **Solution generalization**: In instances where GPT-4 lacked a direct solution, it attempted to generalize from similar tasks encountered during training. This could result in both correct and incorrect answers. Nevertheless, this capability positions the model as a potential brainstorming tool or sparring partner.
- **Precision in questioning**: Ambiguities in question formulation led GPT-4 to provide technically correct answers which might differ from expected human responses. This highlights the importance of precise problem statements.
- **Unwarranted assumptions**: There were occasions when the model presented assumptions not provided in the question, failing to justify its reasoning. This could lead to misleading or incorrect solutions.
- **Answering strategy**: The model sometimes responded to a variant of the posed question or overlooked parts of it. This was especially evident in multi-part questions, where some sections might be addressed thoroughly, while others were completely ignored.
- **Complex exercises**: For more challenging exercises, particularly in the provable security field, GPT-4 displayed inconsistencies, often attempting related but distinct problems.
- **Fundamental mathematical tasks**: The model exhibited proficiency in basic mathematical operations and evaluations, potentially as a result of its training regimen.
- **Acknowledgment of uncertainty**: Notably, GPT-4 rarely, if ever, explicitly acknowledged when it did not know the answer to a given problem.
- **Invention of false answers**: In a few instances, GPT-4 came up with answers even when it didn't know the correct ones. For instance, when asked about a specific historical cipher, GPT-4 confidently explained how it worked, but the information was entirely incorrect. This happened only a few times for a handful of questions. If a student uses GPT-4 for learning, they might end up picking up the wrong answers.

In conclusion, while GPT-4 demonstrates significant prowess in many areas of cryptography, it has limitations, especially when considering its application to real-world cryptographic tasks.

6 Conclusion

The exploration of the application of GPT-4 in the cryptographic domain reveals its substantial potential as an auxiliary educational tool. It can significantly facilitate the learning process for students, elucidating complex cryptographic concepts and providing insights that can be pivotal for understanding. The interactive and dynamic learning environment enabled by GPT-4 augments traditional learning approaches and offers students a novel way to engage with the subject matter.

However, this technological advancement does not come without challenges and risks. The possibility of students using GPT-4 as a substitute for independent problem-solving and critical thinking poses a significant concern. Depending on GPT-4 too heavily could weaken the learning experience and slow the growth of crucial analytical abilities. Furthermore, since the model can sometimes provide inaccurate answers, students without a solid foundation might have difficulty determining the correctness of the responses. This highlights the importance of students developing their comprehension and fact-checking abilities. But almost every pupil and student in his/her life has also experienced human teachers who have not quite mastered their subject matter.

Given the difficulty in ascertaining whether students have solved exercises independently, a reevaluation of assessment methods is crucial. We recommend a diminished emphasis on homework-based assessments and advocate for a more substantial focus on written or oral examinations, where the use of tools like GPT-4 is inherently precluded. This approach can help to maintain academic honesty and truly help students grow their knowledge and abilities in cryptography.

It is also pivotal for students to maintain a skeptical approach when using AI tools, corroborating the provided solutions independently to confirm their correctness. The utility of GPT-4 should be perceived as supplementary, encouraging students to refine their knowledge and verification skills concurrently.

Future Work: For subsequent studies, a profound analysis of how GPT-4 and similar AI models can augment teaching and research in cryptology is essential. This includes exploring innovative methodologies for integrating such tools ethically and effectively into the learning environment and researching detection methods to discern the usage of GPT-4 and similar tools. These future endeavors can provide deeper insights and facilitate the optimized and ethical integration of advanced AI in educational settings, enhancing learning experiences while maintaining the sanctity of educational values.

Acknowledgments

We express our sincere gratitude to Prof. Dr. Frederik Armknecht and Prof. Dr. Arno Wacker for their generosity in allowing us to utilize their teaching materials for this study. Their willingness to share knowledge and resources has been instrumental in the execution and success of this research.

This work has been supported by the Swedish Research Council, grant 2018-06074, DECRYPT – Decryption of Historical Manuscripts.

References

1. J. H. Choi, K. E. Hickman, A. Monahan, and D. Schwarcz. ChatGPT Goes to Law School. *Available at SSRN*, 2023.
2. S. Frieder, L. Pinchetti, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, A. Chevalier, and J. Berner. Mathematical Capabilities of ChatGPT. *arXiv preprint arXiv:2301.13867*, 2023.
3. A. Gilson, C. W. Safranek, T. Huang, V. Socrates, L. Chi, R. A. Taylor, D. Chartash, et al. How Does ChatGPT Perform on the United States Medical Licensing Examination? The Implications of Large Language Models for Medical Education and Knowledge Assessment. *JMIR Medical Education*, 9(1):e45312, 2023.
4. T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted Medical Education using Large Language Models. *PLoS digital health*, 2(2):e0000198, 2023.
5. H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv preprint arXiv:2303.13375*, 2023.
6. GPT-4 Technical Report, 2023. `https://cdn.openai.com/papers/gpt-4.pdf`.
7. P. P. Ray. ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope. *Internet of Things and Cyber-Physical Systems*, 2023.
8. C. Terwiesch. Would Chat GPT3 Get a Wharton MBA. *A prediction based on its performance in the operations management course. Wharton: Mack Institute for Innovation Management/University of Pennsylvania/School Wharton*, 2023.
9. L. Varanasi. AI Models like ChatGPT and GPT-4 are Acing Everything from the Bar Exam to AP Biology. *Here's a List of Difficult Exams both AI Versions have Passed. Business Insider*, 2023. `https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1`.
10. Y. Wardat, M. A. Tashtoush, R. AlAli, and A. M. Jarrah. ChatGPT: A Revolutionary Tool for Teaching and Learning Mathematics. *Eurasia Journal of Mathematics, Science and Technology Education*, 19(7):em2286, 2023.
11. W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al. A Survey of Large Language Models. *arXiv preprint arXiv:2303.18223*, 2023.