Contract Cheating – Dead or Reborn?

Sathiamoorthy Manoharan†, Ulrich Speidel†, Anthony Edward Ward*, and Xinfeng Ye†
*School of Physics, Engineering and Technology
University of York
United Kingdom
†School of Computer Science
University of Auckland
New Zealand

***Abstract*—Online question-answering sites such as Chegg and Bartleby made contract cheating affordable to a wider audience by outsourcing answer services to developing countries such as India and often providing answers within 30 minutes. GPT (Generative Pre-trained Transformer) language models can now give such answers in real-time and for free (or nearly free).**

This paper assesses the quality and promptness of answers provided by Chegg to a range of questions from CS1 and CS2 over a two-year period, comparing them to GPT-generated answers. Results demonstrate that GPT answers are of equal or better quality compared to those provided by human experts, and thus provide an equitable platform for students seeking answers. The paper considers some of the implications from a pedagogical standpoint.

***Index Terms*—academic honesty, positive learning, contract cheating, student assessment, GPT**

I. Introduction

For several years, contract cheating facilitated by online essay mills has been a major issue in academic circles [1]. However, this problem has become increasingly serious with the advent of online question-answering services provided by companies like Chegg and Bartleby, which offer affordable solutions by outsourcing their answering services to developing countries like India. These services are often able to provide answers within 30 minutes, making it easier than ever for students to cheat.

Unfortunately, during the COVID-19 pandemic, the problem has become even more widespread. Chegg, in particular, has become the go-to source for students seeking answers to almost every assessment question in anglophone courses. Although these companies have published honor codes, they are rarely enforced. As a result, questions that are clearly part of current assessed activities are answered without reservation [2], allowing students to breach academic integrity and infringe on copyright at the same time.

The emergence of GPT (Generative Pre-trained Transformer) language models has created a new challenge. These models are capable of solving a wide range of problems in a variety of domains, often with a high degree of accuracy that is comparable to human performance [3]. This raises an important research question: how do the answers generated by GPT language models compare to those provided by online tutoring companies like Chegg?

978-90-386-5690-8/23/$31.00 © 2023 IEEE

This paper seeks to answer this question by conducting a comparison of the quality and speed of answers provided by Chegg and GPT language models. Specifically, we will focus on a wide range of assessment questions at the CS1 and CS2 levels.

The rest of the paper is organized as follows. Section II discusses the background. Section III discusses the question categories and the basis of comparing the answers. Section IV presents the results, showing the quality of answers from both Chegg and GPT. The pedagogical implications of the results are then considered in Section V. The final section concludes this paper with a summary.

II. Background

A. Online Question-Answering Services

Commercial question-answering services such as Chegg and Bartleby provide affordable solutions to student questions. These sites essentially enable contract-cheating, and a large percentage of questions posted in these sites are from current assessed activities from various educational institutions.

Chegg, for example, charges US$20 a month for a premium subscription that includes the question-answering service. The subscriber can seek solutions for up to 20 questions a month, with unlimited access to the solutions in the existing question database. Solutions to additional question cost US$3 each.

A previous study has found that Chegg answers questions even if they contain a clear indication that they are part of a current assessment, enabling student to cheats [2]. In addition, Chegg's hosting of such questions on their site violates copyright in most cases. While Chegg has previously assisted academic institutions in their academic integrity investigations by supplying questioners' metadata such as IP address, e-mail, school name, and time of access, it no longer provides much of this information. Cheating students have also learned not to use e-mail addresses that reveal their identity, or to access from traceable IP addresses such as school networks.

B. Generative Pre-trained Transformer

GPT (Generative Pre-trained Transformer) is a state-of-the-art language generation model developed by OpenAI [4]. It is a deep learning model that uses unsupervised learning to pre-train a large neural network on a massive amount of text data, and then fine-tune the network for various language tasks

such as language translation, text summarisation, and question answering.

The model is based on the transformer architecture, a type of neural network architecture designed for processing sequential data such as text. The transformer architecture uses self-attention mechanisms, allowing the model to selectively focus on different parts of the input when generating text. This allows the model to generate more coherent and fluent text compared to previous models.

GPT-4 is the current version of GPT. It uses a transformer architecture with several billion parameters, making it one of the most powerful language models. It was trained on a diverse range of Internet text, and has the ability to generate human-like text, complete a variety of language tasks such as translation, and perform simple reasoning. It is also multimodal in that it can accept and produce not only text but also images.

ChatGPT[1] is an interactive interface to GPT, which lets users ask questions and give feedback to GPT.

III. Methodology

A. Formulating a Question Set

We formulated a small set of questions drawn from typical CS1 and CS2 courses. These questions are classified into three categories: (1) computer systems – which comprises of computer organization, assembly programming, computer networks, and digital security, (2) algorithms and data structures, and (3) computer programming.

The questions in the computer systems category cover the following topics: binary numbers and arithmetic operations, two's complement, classic ciphers (e.g., Caesar and Vigenère ciphers), encryption, applications of encryption, assembly language programming, assembly macros, loop optimisation, network protocols, HTTP, and HTTPS,

The questions in the algorithms and data structures category cover the following topics: time complexity, sorting, tree traversals, depth-first and breadth-first search, graph algorithms including Dijkstra's shortest path algorithm, and minimum spanning tree using Prim's and Kruskal's algorithms.

The questions in the computer programming category cover the following topics: loops, lists, iterating through list elements, list operations, forming and using dictionaries, using pseudo-random numbers, and basic text encoding (such ROT13 and URL encoding).

The first two categories contain 7 questions each, the computer programming category contains 16 questions, i.e., the total number of questions in the set is 30.

Each question indicates in the prelude that this is part of a graded assessment, indicating the marks awarded for the question. In addition, it also specifies a due date in the future.

B. Posting Questions to Chegg

Each question was posted to Chegg, and the answer measured using two metrics: (1) quality, a percentage mark we awarded to the answer based on a set grading rubric, and (2) promptness, the time between posting of the question and receipt of the answer.

This question set was first posted in 2020 [2]. In 2022, an isometric question set was posted consisting of paraphrased versions of the 2020 questions and (where applicable) a change in some of the data. Isometric questions were necessary to prevent Chegg from automatically linking the new questions to the previously posted ones.

The second trial was important to see if there was any change on Chegg's part in terms of their claim to take academic integrity seriously, which would have meant blocking questions that violate their published honour code.

C. Posing Questions to ChatGPT

We also posed each question from the set to ChatGPT and recorded the answers. Compared to the human "tutors" used by Chegg, the time ChatGPT took for answers was neglible and was recorded as zero. The ChatGPT-provided answers were then marked using the same grading rubric we used for marking Chegg-supplied answers.

IV. Results

A. Chegg's Question-Answering Service

TABLE I: Quality and promptness of the answers in the computer systems category. Questions posted in 2020 to Chegg. Promptness is the number of minutes it took to obtain a solution.

Question	Quality (Marks)	Promptness (Time)	Notes
S1 (Cryptanalysis)	80%	420	
S2 (Crypto essay)	0%	17	Plagiarised
S3 (Assembly A)	100%	60	
S4 (Assembly B)	–	–	No answer
S5 (Assembly macro)	–	–	No answer
S6 (Web security)	20%	120	
S7 (Two's complement)	100%	10	Good explanation

TABLE II: Quality and promptness of the answers in the algorithms and data structures category. Questions posted in 2020 to Chegg. Promptness is the number of minutes it took to obtain a solution.

Question	Quality (Marks)	Promptness (Time)	Notes
A1 (Time complexity)	100%	44	
A2 (Djkstra)	100%	60	
A3 (DFS)	100%	26	
A4 (BFS)	100%	12	
A5 (MST 1)	100%	180	
A6 (MST 2)	100%	45	
A7 (Tree traversals)	100%	3	

Tables I–III indicate the quality and promptness of the answers Chegg provided in 2020. Questions in the algorithms and data structures category as well as the programming category received good solutions. Chegg answered most programming questions quickly, but the answers for the questions in algorithms and data structures took longer. Solutions for

[1]https://chat.openai.com

TABLE III: Quality and promptness of the answers in the computer programming category. Questions posted in 2020 to Chegg. Promptness is the number of minutes it took to obtain a solution.

Question	Quality (Marks)	Promptness (Time)	Notes
P1	100%	14	
P2	100%	8	
P3	100%	8	
P4	100%	7	
P5	100%	7	
P6	100%	8	
P7	100%	6	
P8	100%	50	
P9	100%	6	
P10	100%	60	
P11	100%	8	
P12	70%	8	Some test cases fail
P13	90%	27	Minor mistake.
P14	0%	60	
P15	–	–	No answer
P16	100%	39	Excellent solution

computer systems questions were mediocre and took time: Two of the questions did not attract any answer, and one of the answers was plagiarised.

TABLE IV: Quality and promptness of the answers in the computer systems category. Questions posted in 2022 to Chegg. Promptness is the number of minutes it took to obtain a solution.

Question	Quality (Marks)	Promptness (Time)	Notes
S1 (Cryptanalysis)	0%	22	Wrong; Plagiarised
S2 (Crypto essay)	0%	28	Plagiarised
S3 (Assembly A)	0%	65	Incorrect answer
S4 (Assembly B)	0%	65	Incorrect answer
S5 (Assembly macro)	0%	1145	Incorrect answer
S6 (Web security)	0%	15	Plagiarised
S7 (Two's complement)	100%	96	

TABLE V: Quality and promptness of the answers in the algorithms and data structures category. Questions posted in 2022 to Chegg. Promptness is the number of minutes it took to obtain a solution.

Question	Quality (Marks)	Promptness (Time)	Notes
A1 (Time complexity)	100%	37	
A2 (Djkstra)	100%	141	
A3 (DFS)	100%	49	
A4 (BFS)	100%	49	
A5 (MST 1)	100%	73	
A6 (MST 2)	100%	34	
A7 (Tree traversals)	100%	39	

The questions posted to Chegg in 2020 were paraphrased, the data therein were changed, and the resulting isomorphic questions were posted again to Chegg in 2022 to see if Chegg still aided cheating by not flagging questions that were clearly marked as part of a current assessment. As we expected, all questions were answered with none of them being flagged as violating Chegg's honour code. Tables IV–VI indicate the quality and promptness of the answers Chegg provided in 2022. Questions in the algorithms and data structures category as well as the programming category received good solutions. Only one of the solutions in the computer systems category was correct; in the remaining six, three were incorrect, and three were plagiarised.

TABLE VI: Quality and promptness of the answers in the computer programming category. Questions posted in 2022 to Chegg. Promptness is the number of minutes it took to obtain a solution.

Question	Quality (Marks)	Promptness (Time)	Notes
P1	100%	156	
P2	100%		73
P3	100%	18	
P4	100%	8	
P5	100%	82	Incorrect explanation
P6	100%	29	
P7	100%	53	
P8	100%	302	
P9	100%	163	
P10	75%	59	Correct, but convoluted
P11	100%	19	
P12	100%	24	
P13	100%	50	Two solutions provided
P14	0%	17	Copied from StackOverflow
P15	50%	37	Partially correct
P16	100%	6	

B. *ChatGPT*

The questions posted to Chegg in 2022 were posed to ChatGPT. Tables VII–IX indicate the quality and promptness of the answers ChatGPT provided. The answers were generated in less than a minute, and therefore the promptness metric is noted as 0 for all solutions.

TABLE VII: Quality and promptness of the answers in the computer systems category. Questions posed in 2023 to ChatGPT. Promptness is the number of minutes it took to obtain a solution.

Question	Quality (Marks)	Promptness (Time)	Notes
S1 (Cryptanalysis)	40%	0	Partial answer
S2 (Crypto essay)	100%	0	
S3 (Assembly A)	90%	0	
S4 (Assembly B)	90%	0	
S5 (Assembly macro)	100%	0	
S6 (Web security)	100%	0	
S7 (Two's complement)	100%	0	

All the solutions ChatGPT generated for the programming questions were correct. Of the seven solutions for the questions in algorithms and data structures category, two solutions had incorrect steps, but the other five solutions were correct. For the questions in the computer systems category, ChatGPT generated far superior solutions than Chegg: four solutions were correct, two were almost correct, and one was partially correct.

TABLE VIII: Quality and promptness of the answers in the algorithms and data structures category. Questions posed in 2023 to ChatGPT. Promptness is the number of minutes it took to obtain a solution.

Question	Quality (Marks)	Promptness (Time)	Notes
A1 (Time complexity)	100%	0	
A2 (Djkstra)	100%	0	
A3 (DFS)	100%	0	
A4 (BFS)	100%	0	
A5 (MST 1)	80%	0	Some steps incorrect
A6 (MST 2)	80%	0	Some steps incorrect
A7 (Tree traversals)	100%	0	

TABLE IX: Quality and promptness of the answers in the computer programming category. Questions posed in 2023 to ChatGPT. Promptness is the number of minutes it took to obtain a solution.

Question	Quality (Marks)	Promptness (Time)	Notes
P1	100%	0	
P2	100%	0	
P3	100%	0	
P4	100%	0	
P5	100%	0	
P6	100%	0	
P7	100%	0	
P8	100%	0	
P9	100%	0	
P10	100%	0	
P11	100%	0	
P12	100%	0	
P13	100%	0	
P14	100%	0	
P15	100%	0	
P16	100%	0	

V. Reflections

Even though Chegg has an honour code, our experiments show that Chegg does not enforce it: Questions with clear indications of being part of a current assessed activity still get answered.

The quality of answers provided by Chegg is, on average, well below that of the answers generated by ChatGPT. This is especially true for our questions in the computer systems category. Given that ChatGPT is also more or less free to use and the answers are generated almost instantly, we believe that Chegg[2] is no longer truly relevant in the CS1 and CS2 domains. This has significant pedagogical implications for academics and academic institutions. More students will be inclined to use ChatGPT for generating solutions for assessment questions. Unlike Chegg, where an academic can check if a question was posted online, there is no online record of a student using ChatGPT, and posing the same question to ChatGPT again will normally generate a completely different answer. With Chegg, courses that use individualised questions can trace students who post such questions on Chegg. This is not possible with ChatGPT.

There are attempts to distinguish AI-generated content from human-generated content.

GPTZero[3], for example, uses a measure known as *perplexity score* to detect AI-generated content. ChatGPT explains perplexity score as follows:

> Intuitively, perplexity can be thought of as the number of possible choices a language model is considering when trying to predict the next word in a sequence. A lower perplexity score indicates that the language model is more certain about its predictions and is considering fewer possibilities, while a higher perplexity score indicates that the language model is less certain and is considering more possibilities.

GPTZero assigns this AI-generated content a perplexity score of 22.5 and correctly classifies the content as AI-generated. If we split the second sentence of the AI-generated content into two, deleting the word "while", then GPTZero assigns a much higher perplexity score of 66.7 and classifies the content as human-generated.

While downstream detection of AI-generated content is an interesting approach, it is currently unreliable and not sufficient to penalize cheating students. The most effective way to prevent cheating through generative AI is to use supervised assessments, such as invigilated tests, labs, and exams.

We believe that contract-cheating using services such as Chegg and Bartleby will be short-lived, but cheating using generative AI will be something all academics will have to tackle through awareness, assessment design, and assessment practices.

VI. Summary and Conclusions

Despite Chegg having an honor code, our experiments have shown that the platform fails to enforce it. We have found that questions that are clearly identifiable as part of an ongoing assessed activity still receive answers, despite the platform's supposed commitment to academic integrity.

We conducted a study over a two-year period, evaluating the accuracy and speed of responses provided by Chegg for a range of questions in the fields of CS1 and CS2. We then compared these answers to those generated by ChatGPT. The results indicated the answers generated by ChatGPT were consistently superior or of equal quality compared to Chegg's answers.

Given the availability of ChatGPT as a free and almost instantaneous alternative, we believe that Chegg's relevance in the CS1 and CS2 domains is questionable. Students are likely to be more inclined to use ChatGPT to generate solutions for assessment questions.

We believe that the use of contract-cheating services such as Chegg and Bartleby will be short-lived, but the use of generative AI for cheating will be a persistent issue that academics will need to address through increased awareness, assessment

[2]A recent announcement by Chegg says *CheggMate*, a new service to be offered by Chegg, will use GPT-4, trained on the Chegg's proprietary database of solutions, to auto-generate solutions.

[3]https://gptzero.me

design, and assessment practices. To prevent cheating using generative AI, the only effective solution is to use supervised assessments, such as invigilated tests, labs, and exams.

REFERENCES

[1] R. Clarke and T. Lancaster, "Commercial aspects of contract cheating," in *Proceedings of the* 18^{th} *ACM Conference on Innovation and Technology in Computer Science Education*, ser. ITiCSE '13. Association for Computing Machinery, 2013, pp. 219–224.

[2] S. Manoharan and U. Speidel, "Contract cheating in computer science: A case study," in *2020 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 2020, pp. 91–98.

[3] M. Trim, "Faking it and breaking it: Responsible AI is needed now," *SIGCAS Comput. Soc.*, vol. 51, no. 3, pp. 7–9, feb 2023.

[4] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskeve, "Improving language understanding by generative pre-training," 2018.