Human Interaction and Emerging Technologies (IHIET 2023), Vol. 70, 2023, 311–318
https://doi.org/10.54941/ahfe1002957

Can ChatGPT Help College Instructors Generate High-Quality Quiz Questions?

Kai Lu

University of Michigan, Ann Arbor, MI, 48107, USA

ABSTRACT

ChatGPT is getting increasing attention in both academic and professional settings. Since its release, there has been a discussion on how services such as ChatGPT may change education. Many teachers have shared that they use ChatGPT to help them generate assignment prompts, questions, and lesson plans in various subject areas. Mixed opinions have been shared with regard to the quality of materials created by ChatGPT. While some teachers believe that the materials are of reasonable quality, others worry that ChatGPT may not always generate accurate or reliable information and may reproduce biases and stereotypes that exist in the data it was trained on. In this study, I explore the research question of whether ChatGPT can really replace teachers in generating high quality assessment questions. Specifically, I compare ChatGPT-generated questions with instructor-written questions that have been used in two classes at a public research University in the US. The preliminary results show that although ChatGPT can produce logically sensible questions, the quality is not always comparable to instructor-written ones. The ChatGPT-generated questions are not specific to student misconceptions and do not align with the learning objectives instructors have in mind, which often lead to such questions being relatively obvious and easy to answer. I further discuss the capabilities and limitations of ChatGPT in generating high quality assessment questions. This study provides insights into how we may leverage advanced AI tools such as ChatGPT to support education.

Keywords: ChatGPT, Education, Question generation, Human-AI interaction

INTRODUCTION

ChatGPT is getting increasing attention in both academic and professional settings (Southern, 2023). Soon after its release, it gained over a million users in just a few days. ChatGPT uses Artificial Intelligence to engage in conversation with users. It is trained on huge amounts of data and designed to respond to any natural language prompts the user gives it. As examples, users can ask ChatGPT for the definitions of a word, directions of a task (e.g., recipes), summaries of a paragraph, write an essay on any topic, etc. Since its release, there has been a discussion on whether services such as ChatGPT can change education. Specifically, many teachers have shared over blog posts and social media that they have been exploring using ChatGPT to help them generate assignment prompts, questions, and lesson plans in various subject areas (Cohen, 2023; Ramin, 2022; Stacey, 2023). Mixed opinions have been shared regarding the quality of materials created by ChatGPT. While some teachers believe that the materials are of reasonable quality (Stacey, 2023), others worry that ChatGPT may not always generate accurate or reliable information and may reproduce biases and stereotypes that exist in the data it was trained on (Metzler, 2022). Teachers can then use these assessment tools to quickly evaluate student progress on a variety of topics. However, can ChatGPT really replace teachers in generating high quality assessment questions? In this study, we compare ChatGPT-generated questions with instructors-written questions that have been used in three classes at a University in the US. We discuss the capabilities and limitations of ChatGPT in generating high quality assessment questions. This study provides insights into how we may leverage advanced AI tools such as ChatGPT to support education.

RELATED WORK

Recent articles have showcased different ways teachers may use ChatGPT to support their instruction. Here are some examples of how teachers may leverage ChatGPT to support their teaching. For example, for a writing class, a teacher could choose a piece of student writing, enter it into ChatGPT, and have the tool generate a customized guide about writing skills the student could work on (Cohen, 2023). Additionally, teachers can use ChatGPT to generate writing prompts. For example, Chat GPT can help teachers generate engaging writing prompts for students to respond to. A teacher could ask ChatGPT to generate a story starter or a creative writing prompt, and then have students use the prompt as the basis for their own writing (Ramin, 2022). Moreover, people have also suggested using ChatGPT as a reading comprehension tool. For example, teachers can ask ChatGPT to generate questions about a certain passage and have students answer questions about it. This can be a helpful way to assess students' understanding of the material (Ramin, 2022). Some teachers admitted that ChatGPT can help them with certain teaching tasks and since teachers are stretched over-thin especially during the pandemic, they find such AI support to be very beneficial (Stacey, 2023). As an example, a science teacher used ChatGPT to create three lesson plans to explain how volcanoes are formed. Although such lesson plans are not perfect as admitted by the teacher, they considered the ChatGPT response to be a reasonable start.

Although ChatGPT has drawn wide attention and demonstrated initial success in many cases in education, it is still unclear how ChatGPT-generated materials compare to human-written ones. In this project, I aim to explore when asking ChatGPT to generate educational assessments, how these assessments compare to instructor-written assessments.

METHODS

I compared the ChatGPT-generated questions with instructor-written questions in two classes that I have taken at the University of Michigan. These two classes include a Linear Algebra class and an Introduction to Astronomy class.

The Linear Algebra class focuses on the fundamental concepts such as finding kernel image, row operations, eigenvalues and other fundamental concepts. I chose this class because it is a good representation of a math class which is very common for college students. Linear algebra is also a prerequisite for computer science, data science, and other tech related majors which is very popular.

The Astronomy class is an entry level introduction course on Astronomy and teaches basic concepts such as structures of the universe, the planets of the solar system, etc. This is a good representation of a science course that students take for science distribution credits and it also provides weekly quizzes for students which makes it a good resource for the comparison task.

For the Linear Algebra class, I selected exam questions from our midterm and final exams. We usually have two types of questions in the Linear Algebra classes, including calculation questions or short answer questions. For the astronomy class, I selected review questions after each chapter in the textbook, which are primarily open-ended long answer questions regarding concepts discussed within the corresponding chapters.

I used different prompts in ChatGPT to generate questions for both classes. The goal is to compare the quality of the questions produced in ChatGPT with the actual questions we had in my exams or in the textbook. For the Astronomy class, I copied and pasted an entire chapter from the textbook and asked ChatGPT to generate chapter review questions based on the text. For the Linear Algebra class, I summarized the concepts tested in the examples, including QR factorization, calculations of determinants, etc. I then asked ChatGPT to generate multiple-choice style questions testing those specific concepts.

I then made qualitative comparisons between the instructor-written questions in my exams and the ChatGPT-generated questions. One of the main metrics I used was difficulty rating. For the Linear Algebra class, I rated questions in my exam and the ChatGPT-generated questions on a difficulty scale of 1-10. I also used other metrics including the length of the questions, the number of concepts tested in each question, and the fluency of the questions in the comparison. For the Astronomy class, I compared the quality of the ChatGPT-generated questions with the questions in the textbook by checking the relevance of questions to the text, how easy it is to find the answer to the questions through Google, and the length of the answer that is required (e.g., whether the question can be answered with just a few words or it requires a long paragraph as the answer).

Findings

In this section, I summarize the main takeaways from this study. For the Linear Algebra class, I found that ChatGPT was able to generate Linear algebra questions testing a specific concept, e.g., calculating determinants. However, there are certain limitations of ChatGPT when compared to the questions made by instructors.

First, instructors are better at making questions targeted at certain concepts and use strategies such as simple numbers or providing constraints, hints, and scaffolds (such as a formula or partial answer), which makes sure that students do not spend too much time on the non-targeted aspects of the question (e.g., complex calculations). Instructor-written questions thus do a better job at drawing students' attention to a specific learning objective when they are solving the problem, and not wasting students' time working on unnecessary complicated math calculations that are not the intended target of the question. However, on the contrary, ChatGPT often uses random numbers such as 1,2,3,4,5, which can make certain questions really easy to solve while making others very difficult.

Second, if instructors plan to use ChatGPT to generate questions, I consider it difficult for them to come up with accurate prompts to generate a specific type of questions. For example, instructors may write questions that ask students to calculate, however, using a reasonably specific prompt, ChatGPT generates a short-answer question that asks students to explain a concept. When users use simple prompts, they may not get the questions they wanted. It requires some thoughts for the users to use accurate prompts to describe their desires.

Third, we do observe that for straightforward domains, ChatGPT can generate a similar question as the instructor.

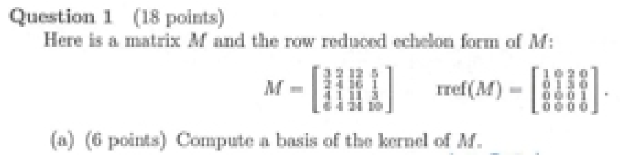
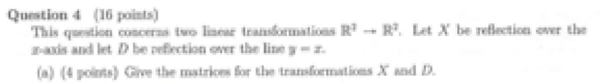
For the Astronomy class, I observed that there are similarities between the ChatGPT-generated questions with the question in the textbook. For example, the textbook asks this question "Describe the major levels of the universe's structure", whereas ChatGPT generates a question "Describe the relationship between the solar system and the galaxy". I consider both questions to test students' understanding of the structure of our universe and are almost interchangeable.

A major difference is that ChaptGPT almost always starts the question stem with "what" or "how", which makes the question easily answerable with a short response. The questions in the textbook, on the other hand, are more sophisticated and diverse, which asks students to explain certain concepts. Moreover, since the ChatGPT questions are relatively more straightforward, it could be easy for students to get the answer through a keyword search without actually reading the chapter. The instructor-written questions often involve explaining and applying key ideas, which exercise higher order thinking. The instructor-written questions may require students to actually read and fully understand the chapter.

CONCLUSION

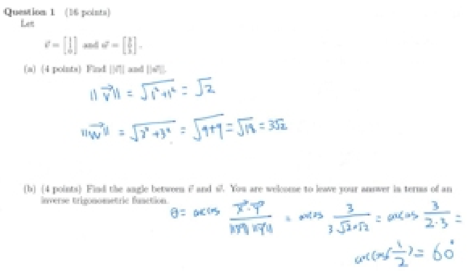
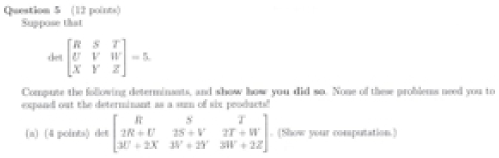
In conclusion, the research shows that ChatGPT is capable of generating quiz questions that are relevant. However, these questions are not sufficient to replace instructor-written questions that are used in practice, at least for the two classes I investigated in this study. Here I summarize the main weaknesses of ChatGPT identified in the study in terms of generating educational assessment questions that are of high educational value. First, for math calculation questions, an instructor's choice of numbers if the questions are usually very deliberate and they do so to adjust the difficulty of the question

Table 1. Comparison of questions generated by ChatGPT and instructor-written questions. The first row shows the prompt I provided to ChatGPT. The difficulty rating of the questions are in parentheses.

Questions 1	
Prompt used in ChatGPT	Generate linear algebra questions to compute basis of kernel with values
Instructor-written question (difficulty)	 **Question 1** (18 points) Here is a matrix M and the row reduced echelon form of M: $$M = \begin{bmatrix} 3 & 2 & 12 & 5 \\ 2 & 4 & 16 & 1 \\ 4 & 1 & 11 & 3 \\ 6 & 4 & 24 & 10 \end{bmatrix} \qquad \mathrm{rref}(M) = \begin{bmatrix} 1 & 0 & 2 & 0 \\ 0 & 1 & 3 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$ (a) (6 points) Compute a basis of the kernel of M. (6/10)
ChatGPT-generated question (difficulty)	Given the matrix A = [[1, 2, 3], [4, 5, 6], [7, 8, 9]] and the vector b = [1, 2, 3], what is the basis for the kernel of the linear transformation represented by the matrix A? (4/10)
Quality difference	The instructor - written question involves a more complex construction of the matrix which makes it harder to construct equations to derive the kernel. ChatGPT does not do a good job because it provides vector b which is irrelevant and that the matrix is made up of three linearly independent, so the kernel would just be 0 and no further calculation is required.
Question 2	
Prompt used in ChatGPT	Generate linear algebra questions to compute reflection over x axis
Instructor-written question (difficulty)	 **Question 4** (16 points) This question concerns two linear transformations $\mathbb{R}^2 \to \mathbb{R}^2$. Let X be reflection over the x-axis and let D be reflection over the line $y = x$. (a) (4 points) Give the matrices for the transformations X and D. (5/10)
ChatGPT-generated question (difficulty)	How do we use matrix multiplication to find the reflection of a vector [x, y] over the x-axis? (5/10)
Quality difference	ChatGPT generated a question requiring a short answer although I specifically requested a computation-style question.
Question 3	
Prompt used in ChatGPT	Generate linear algebra questions to calculate the length and angle between two vectors with values

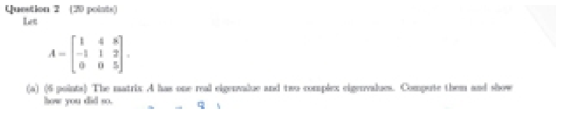
(continued)

Table 1. Continued

Question 3	
Instructor-written question (difficulty)	 Question 1 (16 points) Let $\vec{v} = \begin{bmatrix}1\\1\\0\end{bmatrix}$ and $\vec{w} = \begin{bmatrix}3\\0\\3\end{bmatrix}$. (a) (4 points) Find $\|\vec{v}\|$ and $\|\vec{w}\|$. $\|\vec{v}\| = \sqrt{1^2+1^2} = \sqrt{2}$ $\|\vec{w}\| = \sqrt{3^2+3^2} = \sqrt{9+9} = \sqrt{18} = 3\sqrt{2}$ (b) (4 points) Find the angle between $\vec{v}$ and $\vec{w}$. You are welcome to leave your answer in terms of an inverse trigonometric function. $\theta = \arccos \frac{\vec{v}\cdot\vec{w}}{\|\vec{v}\|\|\vec{w}\|} = \arccos \frac{3}{3\sqrt{2}\cdot\sqrt{2}} = \arccos \frac{3}{2\cdot 3} = \arccos(\frac{1}{2}) = 60^\circ$ (6/10)
ChatGPT-generated question (difficulty)	Given vectors A = [2, 4, 6] and B = [3, 5, 7], find the length of vector A. Given vectors A = [5, 7, 9] and B = [−2, 3, 4], find the angle (in degrees) between the two vectors. (5/10)
Quality difference	These two questions are pretty much the same. ChatGPT specifically asked for the answer in degrees while the instructor allowed an answer in the form of a trigonometric equation. Instructor asked for the length of both vectors while ChatGPT asked for one.
Question 4	
Prompt used in ChatGPT	Generate linear algebra questions to compute determinants with variables as values
Instructor-written question (difficulty)	 Question 5 (12 points) Suppose that $\det \begin{bmatrix} R & S & T \\ U & V & W \\ X & Y & Z \end{bmatrix} = 5.$ Compute the following determinants, and **show how you did so**. None of these problems need you to expand out the determinant as a sum of six products! (a) (4 points) $\det \begin{bmatrix} R & S & T \\ 2R+U & 2S+V & 2T+W \\ 3U+2X & 3V+2Y & 3W+2Z \end{bmatrix}$. (Show your computation.) (6/10)
ChatGPT-generated question (difficulty)	Given the matrix B = [x y z; p q r; s t u], what is the determinant when x, y, z, p, q, r, s, t, and u are variables? (4/10)
Quality difference	A big difference is that the ChatGPT question only requires the student to know the formula for calculating determinants and simply using that on the variables. For the instructor-written question, the student doesn't necessarily need to know the rules of the determinant but they need to know how the manipulation of rows in a matrix affects the determinant.

(continued)

Table 1. Continued

Question 5	
Prompt used in ChatGPT	Generate linear algebra questions to compute real and complex eigenvalues with a 3 × 3 matrix with values
Instructor-written question (difficulty)	 Question 2 (20 points) Let $A = \begin{bmatrix} 1 & 4 & 8 \\ -1 & 1 & 2 \\ 0 & 0 & 5 \end{bmatrix}.$ (a) (6 points) The matrix A has one real eigenvalue and two complex eigenvalues. Compute them and show how you did so. (7/10)
ChatGPT-generated question (difficulty)	Given the matrix A = [1 2 3; 4 5 6; 7 8 9], find the real eigenvalues of A. Given the matrix B = [2 −1 2; 1 2 −1; −1 −1 2], find the complex eigenvalues of B. (6/10)
Quality difference	Both questions asked to compute real and complex eigenvalues. The instructor-written question has a matrix with two zeros to make the computation easier. The ChatGPT question is difficult in the sense that the computation is hard since the matrix has no zeros. However, this is not the intended target of this question.

aligned with their teaching objectives. However, ChatGPT uses random numbers in such calculation questions, which oversimplifies the process most of the time. Second, when using ChatGPT to generate short answering or essay questions, it usually generates questions that start with “what” or “how” and can be answered with just a few words. It also makes it easier for students to find the correct answer without the need to fully understand the content. Third, it is difficult for users to find good prompts to generate the questions they want. For example, it may be difficult for a user to come up with good prompts to generate multi-step questions in Linear Algebra.

REFERENCES

Katie Metzler (2022) How ChatGPT Could Transform Higher Education. https://www.socialsciencespace.com/2022/12/how-chatgpt-could-transform-higher-education/

Marcee Harris (2022). ChatGPT- The Game-Changing App Every Teacher Should Know About. https://www.learnersedge.com/blog/chatgpt-the-game-changing-app-every-teacher-should-know-about

Matt G. Southern (2023) ChatGPT’s Popularity Boosts OpenAI’s Value To $29 Billion. https://www.searchenginejournal.com/chatgpts-popularity-boosts-openais-value-to-29-billion/475762/#close

Shana Ramin (2022). 3 Ways Teachers Can Use ChatGPT in the Classroom. https://www.helloteacherlady.com/blog/2022/12/3-ways-teachers-can-use-chatgpt-in-the-classroom-according-to-chatgpt

Stephanie Stacey (2023). These Teachers Think ChatGPT Can Help Them Spend Less Time on Writing Reports- And More Time with Their Students. https://www.businessinsider.com/teachers-think-chatgpt-let-them-spend-more-time-actually-teaching-2023-1

Zak Cohen (2023). Leveraging ChatGPT: Practical Ideas for Educators. https://www.ascd.org/blogs/leveraging-chatgpt-practical-ideas-for-educators