Contents lists available at ScienceDirect

Academic Pathology 11/1 (2024) 100099

Academic Pathology

journal homepage: www.journals.elsevier.com/academic-pathology

Regular Article

ChatGPT 3.5 fails to write appropriate multiple choice practice exam questions☆

Alexander Ngo[1], Saumya Gupta[1], Oliver Perrine, Rithik Reddy, Sherry Ershadi, Daniel Remick, MD*

Department of Pathology & Laboratory Medicine, Boston University Chobanian and Avidesian School of Medicine, Boston MA, USA

ABSTRACT

Artificial intelligence (AI) may have a profound impact on traditional teaching in academic settings. Multiple concerns have been raised, especially related to using ChatGPT for creating *de novo* essays. However, AI programs such as ChatGPT may augment teaching techniques. In this article, we used ChatGPT 3.5 to create 60 multiple choice questions. Author written text was uploaded and ChatGPT asked to create multiple choice questions with an explanation for the correct answer and explanations for the incorrect answers. Unfortunately, ChatGPT only generated correct questions and answers with explanations in 32 % of the questions (19 out of 60). In many instances, ChatGPT failed to provide an explanation for the incorrect answers. An additional 25 % of the questions had answers that were either wrong or misleading. A grade of 32 % would be considered failing in most courses. Despite these issues, instructors may still find ChatGPT useful for creating practice exams with explanations—with the caveat that extensive editing may be required.

Keywords: Immunology, Artificial intelligence, Education, Pathology

Introduction

Artificial intelligence (AI) emerged as a disruptive technology in multiple disciplines since the release of ChatGPT in 2022.[1] ChatGPT is a sophisticated AI language model developed by OpenAI. It can understand and generate human-like text, making it capable of engaging in natural and coherent conversations on a wide range of topics *(note these prior two sentences were written by ChatGPT)*. The large language model (LLM) technology used by ChatGPT offers the potential to streamline routine tasks including writing essays for college admissions and term articles for humanities classes.[2] However, AI has also been proposed for teaching in the STEM fields.

Similar to ChatGPT, other types of Intelligent systems have been incorporated into STEM classrooms in higher education to facilitate student learning. A systematic review of the use of AI in higher education showed that AI has been implemented in classrooms to teach, grade, provide feedback to students, in addition to supporting teachers in lecture and classroom designs.[3]

A specific study conducted in 2013 used an AI Inquire Biology textbook system that answers student questions and showed that AI use in the classroom engaged students and improved learning outcomes.[4] Another study showed the use of an intelligent tutoring system in medicine, specifically dermatopathology. This study taught histopathologic diagnosis and reporting skills based on commonly used diagnostic algorithms. The results indicated that the students benefited from the system and the frequency of student errors were reduced.[5]

The use of AI to generate multiple choice questions has also been explored previously. A study in 2006 created a program that could successfully generate questions for Basque language exams.[6] The results showed that this specifically designed system was able to generate proper questions 80 % of the time. Another study in 2022 was able to design a program that could take raw text as input and produce multiple choice questions from the raw text.[7] This was accomplished by first having the program process the text, summarize it, score topics, and then generate multiple choice questions using the selected text. The actual generation of the question involved the creation of a question stem followed by generation of correct answers and incorrect answers, called distractors in the study. Similar to the previously mentioned study, the authors focused on a single topic (in this case computer science), but subsequently were able to generate questions from any raw unedited text. The study was able to show, using a similar approach, that AI can generate questions from any inputted topic.[7] While the programs mentioned in these studies did not provide explanations to the answer choices, they were able to accurately generate acceptable questions from provided material.

☆ Supported by NIH grants R21AI147168 and T32GM098308.

* Corresponding author. Department of Pathology & Laboratory Medicine, Boston University Chobanian and Avidesian School of Medicine, 670 Albany St, Boston MA 02118, USA.

E-mail address: remickd@bu.edu (D. Remick).

[1] Both authors contributed equally to this work.

https://doi.org/10.1016/j.acpath.2023.100099

Received 6 July 2023; Received in revised form 5 September 2023; Accepted 29 October 2023, Available online xxxx

eISSN2374-2895/© 2023 The Authors. Published by Elsevier Inc. on behalf of Association of Pathology Chairs. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Students appreciate practice exams that can be used to study for their actual exams. These practice exams have even greater utility when explanations for answers are included as they reinforce correct thinking patterns and provide support for why information in the incorrect answers was wrong. However, creating these practice exams and answers requires substantial faculty time. We investigated whether ChatGPT could be used to create multiple choice questions along with explanations of correct and incorrect choices for a graduate and medical school immunology class.

Materials and methods

A series of self-learning guides for medical school immunology were created in Microsoft Word. These guides covered several immunology topics listed in Table 1. These self-learning guides were 4–12 pages long and were written by medical students (SE and RR) and the senior author (DR). A free account was established on the website https://beta.openai.com/overview and used to log in (https://chat.openai.com/chat). The study was done between January and February 2023 using the version of ChatGPT available at that time (ChatGPT 3.5).

Once logged in the following steps were used to create multiple choice questions:

1) The text and graphs from a single self-learning guide topic were copied and pasted into the textbox and uploaded.
2) The following sentence was pasted into the textbox, exactly as written: “Write 4 multiple choice questions with 4 answers and explanations for the incorrect and correct answers.”
3) The multiple choice questions were then copied, pasted into word and saved.

A total of 60 questions were generated using this format. The format of the questions were reviewed using the National Board of Medical Examiners guide.[8] These questions were then graded by four reviewers (AN, SG, OP and DR) using with the following criteria: 1 = questions and answers were correct requiring only minor formatting, 2 = questions and answers would require substantial modifications in order to be appropriate, 3 = answers were incorrect or significantly misleading. The results are shown in Table 2. For a question to receive a score of 1 all four reviewers needed to agree that the questions and answers were correct. All questions and scores were reviewed by the senior author (DR). To ensure rigorous evaluation of the questions, the worst score from any of the reviewers was used for the final analysis. For questions requiring substantial modifications in order to be useful (score = 2) were further characterized in Table 3.

ChatGPT wrote the questions and answers in the figures and two sentences in the introduction (noted in italics). The authors wrote the remainder of the text.

Results

Overall ChatGPT failed to produce appropriate questions and answers for an immunology course taught by faculty by the Department of Pathology & Laboratory Medicine. An example of an appropriate question is shown in Fig. 1. For this question, ChatGPT correctly wrote the question stem, identified the correct answer, and provided a succinct yet detailed explanation for the correct answer. Appropriate explanations were also provided for each of the incorrect answers.

Table 1
Immunology topics used to generate multiple choice questions.

- Inflammation
- Innate immunity
- Vaccines
- Antigen capture and recognition
- Tolerance and autoimmunity
- T cell immunity
- Humoral immunity
- Autoimmunity
- Hypersensitivity

Table 2
Reviewer scores. The questions and answers generated by ChatGPT were scored using the defined criteria.

Score	Defined criteria	#	%
1	Questions and answers were correct and may be used with only minor formatting	19/60	32 %
2	Questions and answers would require substantial modifications to be appropriate	26/60	43 %
3	Answers were incorrect or significantly misleading	15/60	25 %

Table 3
Reasons why questions or answers were inappropriate.

Score	Defined criteria	#	%
2.1	No explanations for incorrect answers	19/60	32 %
2.2	Not enough detail in the explanation for the incorrect answers	4/60	7 %
2.3	Question written as a negative question, i.e. which of the following is NOT a function	3/60	5 %

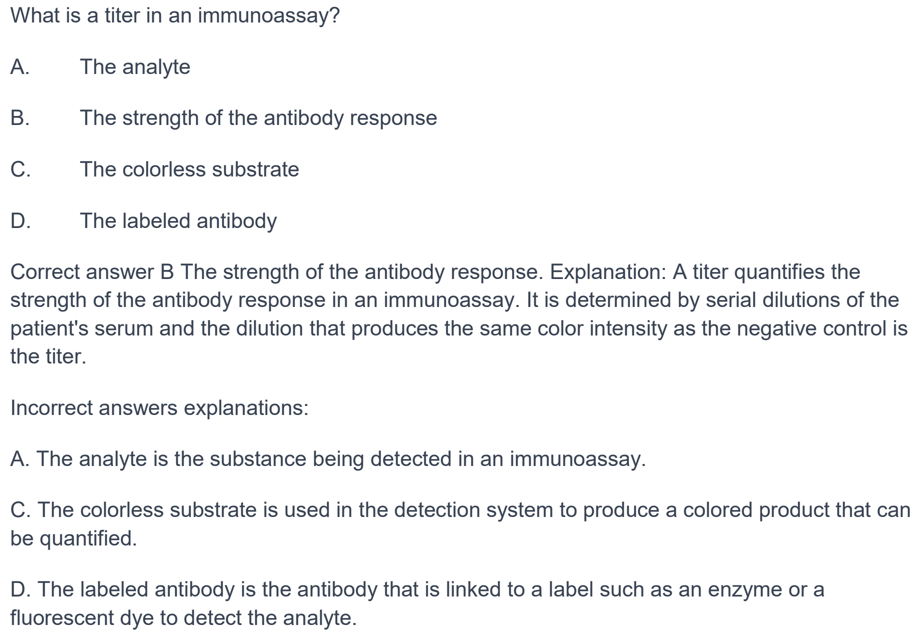
Only 32 % of the questions (19 out of 60) had correct answers and explanations (Table 2). In virtually any class a test score of 32 % would be considered a failing grade.

For 26 of the 60 questions (43 %), substantial modifications would be necessary before the questions would be suitable for a practice exam. The most common reason for a score of two occurred when ChatGPT failed to provide explanations for the incorrect answers (19/60, Table 3). In these cases, the question stem was constructed properly and the correct answer was provided with an appropriate explanation. However, no explanations were given for the incorrect answers. An example of such a question is shown in Fig. 2. While there is no false information in this question, it does not provide an educational opportunity for students to assess their thought process that led them to an answer choice. If a faculty member wants to use this question, they would need to manually write explanations for each of the incorrect answers.

There were four questions where there was insufficient information in the explanation for the wrong answer (Table 3). An additional three questions required a “negative” response such as “Which of the following is NOT a cardinal sign of inflammation” and these types of questions do not comply with the guidelines from the National Board of Medical Examiners.[8]

For those questions where ChatGPT failed to generate explanations for the incorrect answers, attempts were made to give the program a “second chance.” ChatGPT was asked again to write questions and provide answers based on the previously uploaded information. Invariably just asking ChatGPT to generate new questions did not result in appropriate explanations for the incorrect answers. The only approach that worked was to exit the program and start a new session. Even under these circumstances the program would sometimes fail to generate appropriate questions.

About 43 % of the time, ChatGPT produced questions that would require substantial modification in order to be useful for professors and students. Fig. 3 shows an example of such a question, that has nonspecific or vague wording choices in the question stem and thus misleading answer choices. This type of question can lend itself to have multiple right answers without appropriate specificity of certain terms, such as the type of T cells referred to in the question stem. As some of the answer choices can be correct without appropriate context from the question, the



What is a titer in an immunoassay?

A. The analyte

B. The strength of the antibody response

C. The colorless substrate

D. The labeled antibody

Correct answer B The strength of the antibody response. Explanation: A titer quantifies the strength of the antibody response in an immunoassay. It is determined by serial dilutions of the patient's serum and the dilution that produces the same color intensity as the negative control is the titer.

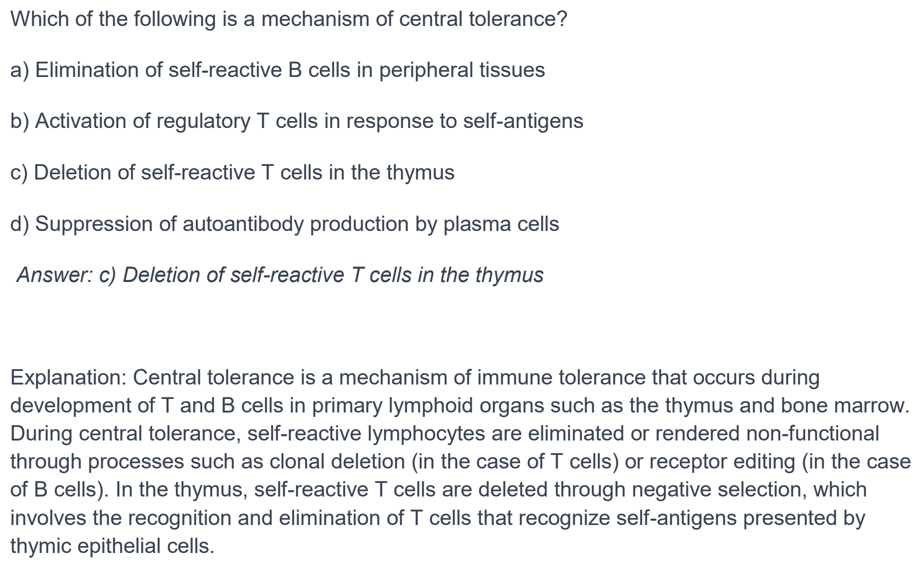
Incorrect answers explanations:

A. The analyte is the substance being detected in an immunoassay.

C. The colorless substrate is used in the detection system to produce a colored product that can be quantified.

D. The labeled antibody is the antibody that is linked to a label such as an enzyme or a fluorescent dye to detect the analyte.

Fig. 1. Example of correct answer and explanations generated by ChatGPT. Text generated by ChatGPT in response to the query "Write 4 multiple choice questions with 4 answers and explanations for the incorrect and correct answers." February 2023. Open AI.



Which of the following is a mechanism of central tolerance?

a) Elimination of self-reactive B cells in peripheral tissues

b) Activation of regulatory T cells in response to self-antigens

c) Deletion of self-reactive T cells in the thymus

d) Suppression of autoantibody production by plasma cells

Answer: c) Deletion of self-reactive T cells in the thymus

Explanation: Central tolerance is a mechanism of immune tolerance that occurs during development of T and B cells in primary lymphoid organs such as the thymus and bone marrow. During central tolerance, self-reactive lymphocytes are eliminated or rendered non-functional through processes such as clonal deletion (in the case of T cells) or receptor editing (in the case of B cells). In the thymus, self-reactive T cells are deleted through negative selection, which involves the recognition and elimination of T cells that recognize self-antigens presented by thymic epithelial cells.

Fig. 2. Example of ChatGPT generated question where no explanations were provided for the wrong answers. Text generated by ChatGPT in response to the query "Write 4 multiple choice questions with 4 answers and explanations for the incorrect and correct answers." February 2023. Open AI.

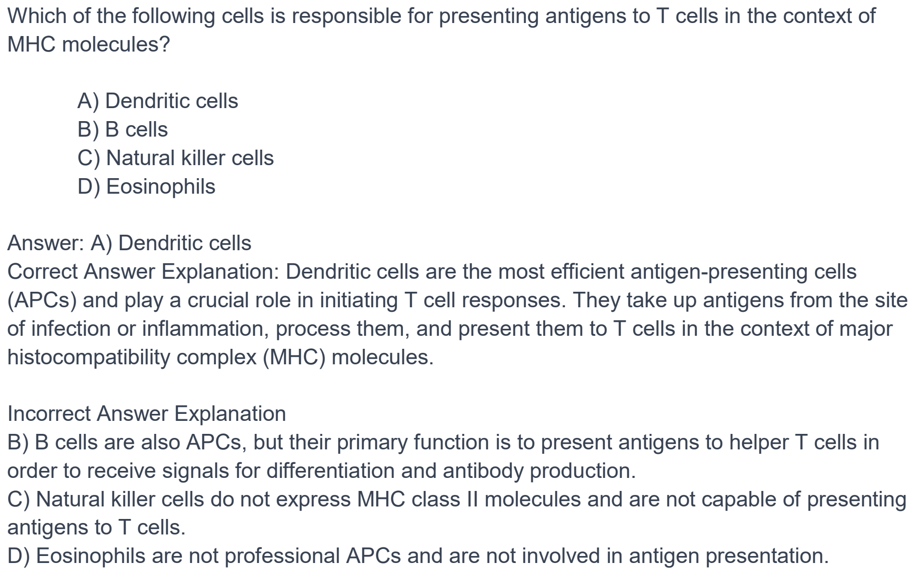
answer explanations provided in the question do not provide sufficient clarification of the incorrect answer choices.

ChatGPT produced questions that would be deemed as completely wrong in 25 % (15/60) of the questions. This could be due to either having incorrect answers labeled as correct, correct answers labeled as incorrect, or incorrect explanations. An example is shown in Fig. 4. In this question, ChatGPT listed multiple correct answers but labeled all but one as incorrect. ChatGPT recognized that all the answers were primary immunodeficiencies but simply chose one as "correct." This type of question would require the complete modification of the answer choices and explanations to be useable.

Discussion

While ChatGPT failed to write appropriate questions, a legitimate issue is whether using ChatGPT saves time. If 32 % of the questions and answers can be used with only minor formatting, would just these questions require less time than having the faculty member write new questions? Without a formal time analysis study this question cannot be answered definitively. It is our impression that even with 32 % correct questions and answers it takes less time to review these for accuracy than writing completely new questions. An alternative to writing these questions would be to have students generate new multiple choice questions. However, a recent article found that this approach was not well received by medical students in a pathology course.[9] A systematic review of the literature to evaluate the educational value of students writing multiple choice questions did not reach definitive conclusions.[10]

While the questions and answers could be used for practice exams, they would also be suitable for the actual exams. Some classes do not



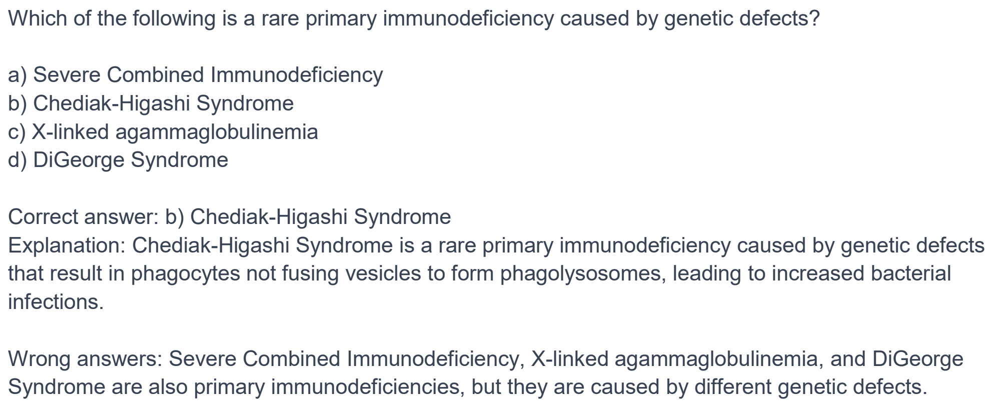
Which of the following cells is responsible for presenting antigens to T cells in the context of MHC molecules?

A) Dendritic cells
B) B cells
C) Natural killer cells
D) Eosinophils

Answer: A) Dendritic cells
Correct Answer Explanation: Dendritic cells are the most efficient antigen-presenting cells (APCs) and play a crucial role in initiating T cell responses. They take up antigens from the site of infection or inflammation, process them, and present them to T cells in the context of major histocompatibility complex (MHC) molecules.

Incorrect Answer Explanation
B) B cells are also APCs, but their primary function is to present antigens to helper T cells in order to receive signals for differentiation and antibody production.
C) Natural killer cells do not express MHC class II molecules and are not capable of presenting antigens to T cells.
D) Eosinophils are not professional APCs and are not involved in antigen presentation.

Fig. 3. Unacceptable question because of non-specific wording or difficult to understand rationale. Text generated by ChatGPT in response to the query "Write 4 multiple choice questions with 4 answers and explanations for the incorrect and correct answers." February 2023. Open AI.



Which of the following is a rare primary immunodeficiency caused by genetic defects?

a) Severe Combined Immunodeficiency
b) Chediak-Higashi Syndrome
c) X-linked agammaglobulinemia
d) DiGeorge Syndrome

Correct answer: b) Chediak-Higashi Syndrome
Explanation: Chediak-Higashi Syndrome is a rare primary immunodeficiency caused by genetic defects that result in phagocytes not fusing vesicles to form phagolysosomes, leading to increased bacterial infections.

Wrong answers: Severe Combined Immunodeficiency, X-linked agammaglobulinemia, and DiGeorge Syndrome are also primary immunodeficiencies, but they are caused by different genetic defects.

Fig. 4. Example of ChatGPT question with incorrectly labeled answers. Text generated by ChatGPT in response to the query "Write 4 multiple choice questions with 4 answers and explanations for the incorrect and correct answers." February 2023. Open AI.

release exam questions to spare faculty the time needed to write appropriate multiple choice questions each year. This could be eliminated if ChatGPT wrote these questions which only needed to be reviewed by the faculty for accuracy.

Aside from it's usefulness as tool for professors to generate exam questions, ChatGPT may also be used by students. Practice questions are a popular study tool among students as they can help gauge a student's understanding of the material while simulating a testing environment.[11] ChatGPT generation of questions with explanations of the incorrect and correct answers offers a great method for students to prepare for exams. Past studies have shown that learning through AI received positive feedback from students who felt that AI increased communication in learning.[12] However, ChatGPT's current accuracy in generating clear questions and accurate answers explanations can be of concern. Since 25 % of the time ChatGPT generated questions with vague or confusing question stems and poor explanations of the answer choices, this study tool may not be entirely viable. These types of misleading questions may create further confusion about the topics, especially since the students have not gained expertise and they may not be able to find errors in the questions. These types of concerns were brought up in a study by Adamson et al., that showed the use of an AI system to facilitate student learning discussions. The study showed that the effectiveness of the AI in aiding the students was dependent upon the difficulty of the material (whether it was new material or review).[13] With newer versions of ChatGPT being released from feedback from users, if the system becomes more accurate in generating the questions and answers, it may be a good study tool for the students.

As with any new tool, there may be concerns about the ethics of using AI in education. One key concern is violation of privacy. In order to be optimized, the AI program may analyze students' personal data to better adapt to their learning techniques.[14] While this may be beneficial for developing the best study plans for students, this may also be considered a violation of personal information. There have been growing concerns about whether or not the usage of ChatGPT by students to write academic essays.[15] Caution is needed when integrating AI programs into the classroom to avoid such ethical conundrums. When creating assessments this should not be a concern since personalized data would not be utilized.

Limitations

The most important limitation of this study is that AI relentlessly improves and newer versions of ChatGPT, or Bard from Google (another AI program), may produce better questions and answers. It should be noted that we did not ask ChatGPT to generate questions from the internet of knowledge. We uploaded the information that should be used to create the questions, rather than asking for questions based on a specific topic, i.e. write four questions about transplant immunology. Better questions and answers may have been generated if ChatGPT were allowed more freedom. Our approach was used because students are responsible for the information presented in class and assigned readings with multiple choice exam questions rather than the internet of knowledge.

Conclusion

ChatGPT may be useful for writing appropriate multiple-choice questions for practice exam. These questions would include explanations for the correct and incorrect answers. The teaching faculty will need to carefully review the generated questions to ensure that the answers and explanations are correct. Future directions: New versions of AI will produce better results. Generating actual exams based on course content may be done more easily.

Declaration of competing interest

The senior author, Daniel Remick, M.D., is a member of the Editorial Board for *Academic Pathology*. No other conflicts of interest exist.

References

1. Wen J, Wang W. The future of ChatGPT in academic research and publishing: a commentary for clinical and translational medicine. *Clin Transl Med*. 2023;13(3), e1207. doi:10.1002/ctm2.1207.
2. Lund BDW, Ting Mannuru, ReddyNie Nishith, Bing Shimray, Wang Somipam, Ziang. ChatGPT and a new academic reality: AI-written research papers and the ethics of the large language models in scholarly publishing. *J Assoc Inf Sci Technol*. 2023;74(5): 570–581. doi:10.1002/asi.24750.
3. Zawacki-Richter O, Marin V, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education - where are the educators? *Intern J Edu Tech in Higher Edu*. 2019;(16):39. doi:10.1186/s41239-019-0171-0.
4. Chaudhri V, Cheng B, Overholtzer A, et al. *Inquire Biology: A Textbook that Answers Questions*. 2013:18.
5. Payne VL, Medvedeva O, Legowski E, et al. Effect of a limited-enforcement intelligent tutoring system in dermatopathology on student errors, goals and solution paths. *Artif Intell Med*. 2009;47(3):175–197. doi:10.1016/j.artmed.2009.07.002.
6. *ArikIturri: An Automatic Question Generator Based on Corpora and NLP Techniques*. Intelligent Tutoring Systems; 2006. presented at.
7. Maheen F, Asif M, Ahmad H, et al. Automatic computer science domain multiple-choice questions generation based on informative sentences. *PeerJ Computer Science*. 2022;8, e1010. doi:10.7717/peerj-cs.1010.
8. Billings MS, DeRuchie K, Hussie K, et al. NBME item-writing guide. Accessed August 25, 2023. https://www.nbme.org/item-writing-guide#request-demo-form; 2023.
9. Grainger R, Dai W, Osborne E, Kenwright D. Medical students create multiple-choice questions for learning in pathology education: a pilot study. *BMC Med Educ*. 2018; 18(1):201. doi:10.1186/s12909-018-1312-1.
10. Touissi Y, Hjiej G, Hajjioui A, Ibrahimi A, Fourtassi M. Does developing multiple-choice questions improve medical students' learning? A systematic review. *Med Educ Online*. 2022;27(1), 2005505. doi:10.1080/10872981.2021.2005505.
11. Wynter L, Burgess A, Kalman E, Heron JE, Bleasel J. Medical students: what educational resources are they using? *BMC Med Educ*. 2019;19(1):36. doi:10.1186/s12909-019-1462-9.
12. Kim Jihyun, Kelly Merrill, Xu Kun, Deanna DS. My teacher is a machine: understanding students' perceptions of AI teaching assistants in online education. *Int J Hum Comput Interact*. 2020;36(20):1902–1911. doi:10.1080/10447318.2020.1801227.
13. Adamson D, Dyke G, Jang H, Rosé CP. Towards an agile approach to adapting dynamic collaboration support to student needs. *Internl J Artificial Intel Edu*. 2014/01/01 2014;24(1):92–124. doi:10.1007/s40593-013-0012-6.
14. Li X. Intelligent agent–supported online education. *Decis Sci J Innovat Educ*. 2007; 5(2):311–331. doi:10.1111/j.1540-4609.2007.00143.x.
15. Anders BA. Is using ChatGPT cheating, plagiarism, both, neither, or forward thinking? *Patterns*. 2023;4(3), 100694. doi:10.1016/j.patter.2023.100694.