PLOS ONE

RESEARCH ARTICLE

ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom)

Billy Ho Hung Cheung[1], Gary Kui Kai Lau[1], Gordon Tin Chun Wong[1], Elaine Yuen Phin Lee[1], Dhananjay Kulkarni[2], Choon Sheong Seow[3], Ruby Wong[4], Michael Tiong-Hong Co[1]*

1 L.K.S. Faculty of Medicine, University of Hong Kong, Hong Kong, Hong Kong S.A.R, **2** Department of Surgery, University of Edinburgh, Edinburgh, United Kingdom, **3** Department of Surgery, National University Cancer Institute Singapore, Singapore, Singapore, **4** Department of Surgery, University of Galway, Galway, Ireland

* mcth@hku.hk

OPEN ACCESS

Citation: Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, et al. (2023) ChatGPT versus human in generating medical graduate exam multiple choice questions—A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS ONE 18(8): e0290691. https://doi.org/10.1371/journal.pone.0290691

Editor: Jie Wang, Education University of Hong Kong, HONG KONG

Received: June 5, 2023

Accepted: August 15, 2023

Published: August 29, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0290691

Abstract

Introduction

Large language models, in particular ChatGPT, have showcased remarkable language processing capabilities. Given the substantial workload of university medical staff, this study aims to assess the quality of multiple-choice questions (MCQs) produced by ChatGPT for use in graduate medical examinations, compared to questions written by university professoriate staffs based on standard medical textbooks.

Methods

50 MCQs were generated by ChatGPT with reference to two standard undergraduate medical textbooks (Harrison's, and Bailey & Love's). Another 50 MCQs were drafted by two university professoriate staff using the same medical textbooks. All 100 MCQ were individually numbered, randomized and sent to five independent international assessors for MCQ quality assessment using a standardized assessment score on five assessment domains, namely, appropriateness of the question, clarity and specificity, relevance, discriminative power of alternatives, and suitability for medical graduate examination.

Results

The total time required for ChatGPT to create the 50 questions was 20 minutes 25 seconds, while it took two human examiners a total of 211 minutes 33 seconds to draft the 50 questions. When a comparison of the mean score was made between the questions constructed by A.I. with those drafted by humans, only in the relevance domain that the A.I. was inferior to humans (A.I.: 7.56 +/- 0.94 vs human: 7.88 +/- 0.52; p = 0.04). There was no significant

Data Availability Statement: All relevant data are within the paper and its Supporting information files.

Funding: No.

Competing interests: No.

difference in question quality between questions drafted by A.I. versus humans, in the total assessment score as well as in other domains. Questions generated by A.I. yielded a wider range of scores, while those created by humans were consistent and within a narrower range.

Conclusion

ChatGPT has the potential to generate comparable-quality MCQs for medical graduate examinations within a significantly shorter time.

Introduction

The workload of university medical staff is a pressing issue that requires attention [1], Medical staff are often tasked with multiple responsibilities that can include, but are not limited to, patient care, teaching, student assessment, research, and administrative work [2]. This heavy workload can be especially challenging for medical academic staff, who are also required to uphold the standard of undergraduate medical exams. The demand for exam quality has increased in recent years, as students and other stakeholders expect assessments to be fair, accurate, unbiased and aligned with the predefined learning objectives [3, 4].

In this context, the development of artificial intelligence (A.I.), machine learning, and language models offers a promising solution. A.I. is a rapidly developing field that has the potential to transform many industries, including education [5]. A.I. is a broad field that encompasses many different technologies, such as machine learning, natural language processing, and computer vision [6]. Machine learning is a subset of A.I. that involves training algorithms to make predictions or decisions based on the database [7]. This technology has been used in a variety of applications, including speech recognition, image classification, and language generation.

A large language model is a type of A.I. that has been trained on a massive amount of data and can generate human-like text with impressive accuracy [8]. These models are called "large" because they have a huge number of parameters, which are the weights in the model's mathematical equations that determine its behaviour. The more parameters a model has, the more information it can store and the more complex tasks it can perform.

ChatGPT is a specific type of large language model developed by OpenAI [9]. It is a state-of-the-art model that has been trained on a diverse range of texts, including news articles, books, and websites. This makes it highly versatile and capable of generating text on a wide range of topics with remarkable coherence and consistency. It is a pre-trained language model with knowledge up to 2021. However, it is possible to feed in relevant reference text by the operator, such that updated text or desired text outputs can be generated based on the operator's command and preference.

The potential implications and possibilities of using ChatGPT for assessment in medical education are significant. A recent publication confirmed that the knowledge provided by ChatGPT is adequate to pass the United States Medical Licensing Exam [10]. It is also believed that ChatGPT could be used to generate high-quality exam questions, provide personalized feedback to students, and automate the grading process, hence reducing the workload of medical staff and improving the quality of assessments [11]. This technology has the potential to be a game-changer in the field of medical education, providing new and innovative ways to assess student learning and evaluate exam quality.

Meanwhile, multiple-choice questions (MCQs) have been used as a form of knowledge-based assessment since the early twentieth century [12]. MCQ is an important and integral component in both undergraduate and post-graduate exams, due to its standardization, equitability, objectiveness, cost-effectiveness, and reliability [13]. When compared to essay-type of questions or short answer questions, MCQs also allow assessment of a broader range of content—each exam paper can include large numbers of MCQs [14]. This makes the MCQ format particularly suitable for summative final examinations. The major drawback to the MCQ format, however, is that high-quality questions are difficult and time-consuming to draft.

We hypothesize that advanced large language models, such as ChatGPT, can reliably generate high-quality MCQs that are comparable to those of an experienced exam writer. Here, with an updated large language model A.I. available, we aim to evaluate the quality of exam questions generated by ChatGPT versus those drafted by university professoriate staff based on international gold-standard medical reference textbooks.

Methods

This is a prospective study to compare the quality of MCQs generated by ChatGPT versus those drafted by experienced university professoriate staff for medical exams (Fig 1). The study was conducted in February 2023. To allow a fair comparison, certain criteria are set for question developments.

1. The questions were designed to meet the standard for a medical graduate exam.
2. Only four choices were allowed for each question.
3. The questions were limited to knowledge-based questions only.
4. The questions were text-based only.
5. Topics regarding the exam context were set by an independent researcher before the design of the questions.
6. Both the professorial staff who designed the questions and the research operating ChatGPT were not allowed to view questions from the other side.

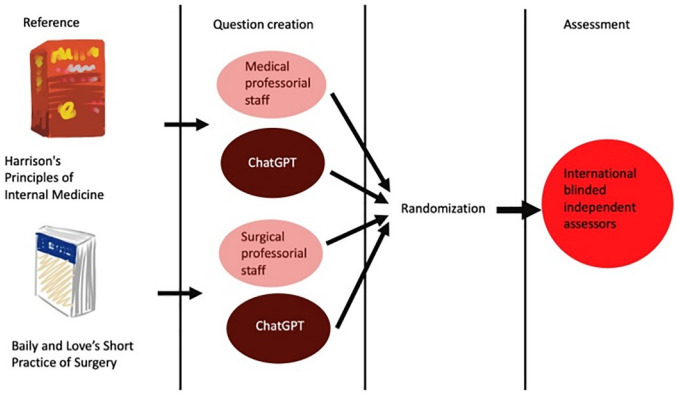

Fig 1. Schematic diagram of the study design.

https://doi.org/10.1371/journal.pone.0290691.g001

7. Standardized, internationally-used textbooks, with Harrison's Principles of Internal Medicine 21th edition for medicine [15], and Baily and Love's Short Practice of Surgery 27th Edition for surgery [16], were used as the reference for question generation.
8. Distractors were allowed to refer from sources other than the original reference provided.
9. No explanation was required for the question.

ChatGPT plus [17] was used for question creation. This version is a small update to the standard ChatGPT version on the 1st of February 2023. According to the official webpage, ChatGPT Plus carries the ability to answer follow-up questions, and challenge incorrect assumptions. This also provides better access during peak hours, faster response time and priority to updates.

Question construction

Fifty multiple choice questions (MCQs) with four options covering Internal Medicine and Surgery were generated by the ChatGPT with reference articles selected from the two reference medical textbooks.

ChatGPT was assessed in Hong Kong S.A.R. via a virtual personal network to overcome its geographical limitation [18]. The commander using the ChatGPT will only copy and paste the instruction "Can you write a multiple-choice question based on the following criteria, with the reference I am providing for you and your medical knowledge?" (also known as the prompt), the criteria and the reference text (selected from the reference textbook and input into ChatGPT interface as pure text input), to yield a question (S1 File). After sending the request to ChatGPT, a response will be automatically generated. Interaction is only allowed for clarification but not any modification. Questions provided by ChatGPT will be directly copied and used as the output for assessment by the independent quality assessment team. Moreover, a new chat session was started in ChatGPT for each entry to reduce memory retention bias.

The duration of work by A.I. was defined by the response time that ChatGPT needed to generate the questions once the prompt and the material were given, and the time for the operator to copy and paste the questions was not included. The duration of work by ChatGPT and humans was documented and compared.

Another 50 MCQs with four options were drafted by two experienced university professoriate staff (with more than 15 years of clinical practice and academic experience in Internal Medicine and Surgery, respectively) based on the same textbook reference materials.

They were given the exact instructions as the criteria mentioned above. Additionally, while they were allowed to refer back to additional textbooks or online resources, they were not allowed to use any A.I. tool to assist their question writing.

Blinded assessment by a multinational expert panel

All 100 MCQs were individually numbered and randomized using a computer-generated sequence.

A multiple-choice item consists of the stem, the options, and any auxiliary information [12]. The stem contains context, content, and/or the question the student is required to answer. The options include a set of alternative answers with one correct option and one or more incorrect options or distractors. Auxiliary information includes any additional content, in either the stem or option, required to generate an item. The incorrect options to an MCQ are known as distractors; they serve the purpose of diverting non-competent candidates away

from the correct answer, which they serve as an important hallmark of a high-quality question [19].

The question set was then sent to five independent international blinded assessors (From the United Kingdom, Ireland, Singapore and Hong Kong S.A.R.) for assessment of the question quality (who were also the authors of this study as well). Members of the international panel of assessors are experienced clinicians with heavy involvement in medical education in their locality.

An assessment tool with five domains is specifically designed after literature review of relevant metrics for MCQ quality assessment in this study [20–23]. Specific instructions were given to the assessors for the meaning of each domain. These include (I) Appropriateness of the question, defined as if the question is correct, appropriately constructed with appropriate length and well-formed; (II) Clarity and specificity, defined as if the question is clear and specific without ambiguity, its answerability and without being under- or over-informative; (III) Relevance, defined as the relevance to clinical context; (IV) Quality of the alternatives & discriminative power for the assessment of the alternative choices provided; and (V) Suitability for graduate medical school exam focused on the level of challenge including if the question higher order of learning outcomes such as application, analysis and evaluation, as elaborated by Bloom and his collaborators [24]. The quality of MCQs was objectively assessed by a numeric scale of 0–10. "0" is defined as extremely poor, while "10" means that it is at the quality of a gold-standard question. The total score of this section ranges from 0 to 50.

In addition, the assessors were also asked to determine if the questions were constructed by A.I. or by humans, in which they were blinded about the total number of questions created by each arm. In addition, G.P.T. -2 Output Detector, an A.I. output detector [25] was also used to predict as if the question was written by A.I. or a human. This detector assigns a score between 0.02 and 99.98% to each question, with a higher score indicating a greater likelihood that the question was constructed by A.I.

Statistical analysis

All data were prospectively collected by a research assistant and computerized into a database. All statistical analyses were performed with the Statistical Product and Service Solution (SPSS) version 29. A comparison was made between questions created by A.I. and by humans. Student T test or Mann-Whitney U test was used for the comparison of continuous variables for the five domains individually and combined. P value of less than 0.05 were considered statistically significant. Chi-squared test or Fisher's exact test were used to compare discrete variables, namely the perception of the assessor whether a question was produced by A.I. or human.

A paired t-test was performed to assess for systematic differences between the mean measures of each rater (including G.P.T. -2 Output Detector).

Results

Question construction

The question writing was performed by ChatGPT on two separate dates, 11th and 17th February. The work was carried out with stable internet via Wifi at a minimum of 15.40 Mbps for downloads and 11.96 Mbps for update speed.

The total time required for ChatGPT to create the 50 questions was 20 minutes 25 seconds, while it took two human examiners a total of 211 minutes 33 seconds (84 minutes 56 seconds for the surgical examiner and 126 minutes 37 seconds for the medical examiner).

Table 1. Mean score of each domain.

	Mean (± SD)	Range
Appropriateness of the question	7.78 ± 0.74	5.40–9.80
Clarity and specificity	7.63 ± 0.69	5.60–9.20
Relevance	7.72 ± 0.77	5.60–9.20
Quality of the alternatives & discriminative power	7.31 ± 0.65	5.60–7.31
Suitability for graduate medical school exam	7.32 ± 0.84	5.00–9.20
Total score	37.76 ± 3.34	27.20–46.20

https://doi.org/10.1371/journal.pone.0290691.t001

Assessment of question quality

The results of the assessment by the independent blinded assessors were summarized in Table 1. We can see that the overall score was satisfactory with a mean score of each domain above 7.

When a comparison of the mean score was made between the questions constructed by A.I. with those constructed by humans, only in the relevance category that the A.I. was inferior to humans (A.I.: 7.56 ± 0.94 vs human: 7.88 ± 0.52; p = 0.04, Table 2). There was no significant difference in other domains of question quality assessment. The same applies to the total scores between A.I. and humans (Table 1).

Questions generated by A.I. yielded a wider range of scores, while those created by humans were consistent and within a narrower range (Fig 2). A similar distribution was also observed across all five domains (Fig 3).

A.I. vs human

The average scores of the questions based on the same reading material were compared between that generated by A.I. vs by human writers. A "win" denotes a higher average score. Questions generated by humans generally received a higher mark when compared to the A.I. counterpart (Table 3). However, AI-drafted questions outperformed human ones in 36% to 44% of cases across five assessment domains, including the total score.

In addition, when the questions generated by ChatGPT were reviewed, we observed few negative features, including minimal use of negative stem (only 14% (7/50), compared to 12% (6/50) by human examiners) with a lack of "except", "All/none of the above".

Blinded guess of the question writer by a panel of assessors

Assessors were asked to deduce if the question was written by A.I. or humans, and the results are shown in Table 4. The results of the assessment by G.P.T. -2 Output Detector were also

Table 2. Comparison of mean scores between questions generated by AI and human.

	AI (± SD)	Human (± SD)	*P*
Appropriateness of the question	7.72 ± 0.83	7.84 ± 0.65	0.45
Clarity and specificity	7.56 ± 0.81	7.69 ± 0.55	0.34
Relevance	7.56 ± 0.94	7.88 ± 0.52	0.04
Quality of the alternatives & discriminative power	7.26 ± 0.68	7.36 ± 0.61	0.46
Suitability for graduate medical school exam	7.25 ± 0.94	7.40 ± 0.72	0.39
Total score	37.36 ± 3.92	38.16 ± 2.62	0.23

SD = standard deviation

https://doi.org/10.1371/journal.pone.0290691.t002

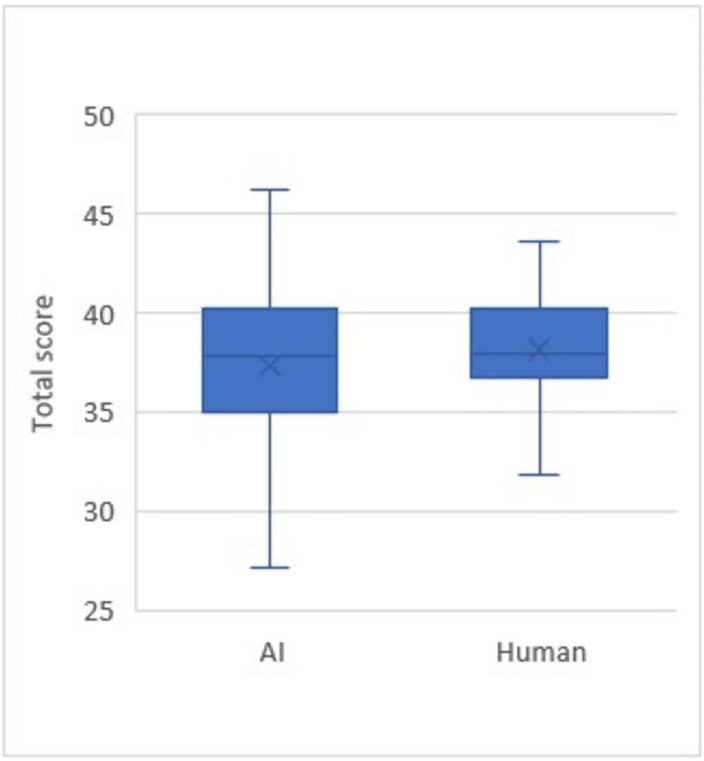

Fig 2. Assessment scores of MCQ quality between A.I. and human.

https://doi.org/10.1371/journal.pone.0290691.g002

shown. The percentage of correct guesses was consistently low. None of the five assessors could achieve a correct "guess" rate of 90% or above. Only G.P.T. -2 Output Detector could achieve a higher correct "guess" rate of 90% for human exam writers. In addition, our results showed that there is no correlation between the guess made by the assessor and the actual writer of the question.

Discussion

This is the first evidence in the literature showing that a commercially available, pre-trained A. I. can prepare exam material with a compatible quality with experienced human examiners.

MCQ is a crucial tool for assessment in education because they permit the direct measurement of many knowledge, skills, and competencies across a broad range of disciplines with the ability to test their concepts and principles, make judgments, drawing inferences, reasoning, interpretation of data, and information application [12, 14]. MCQs are also efficient to administer, easy to score objectively, and provide statistical information regarding the class performance on a particular question and assess if the question was appropriate to the context that was presented [21, 26]. A standard MCQ consists of the stem, the options, and occasionally auxiliary information [27]. The stem contains context, and content, and sets the question. The options include a set of alternative answers with one correct option and other incorrect options known as distractors [22]. Distractors are required to divert non-competent candidates away from the right answer, which serves as an essential hallmark of a high-quality question [19]. However, the major drawback to the MCQ format is that high-quality questions are difficult, time-consuming, and costly to write [28]. From our results, it is evident that even

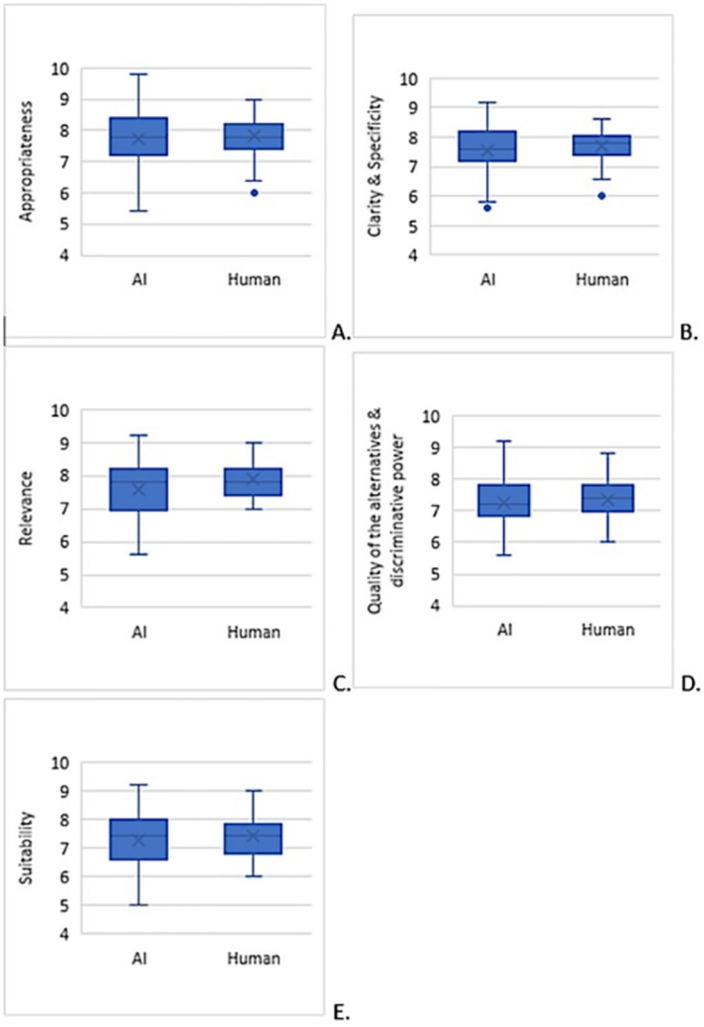

Fig 3. Assessment scores across all five assessment domains.

https://doi.org/10.1371/journal.pone.0290691.g003

experienced examiners required more than ten minutes to prepare one question on average. With encouraging results demonstrated by the current study, we believe that A.I. could have the potential to generate quality MCQs for medical education. Artificial Intelligence (A.I.) has been used in education to improve learning and teaching outcomes for several years. The history of A.I. in education can be traced back to the 1950s when scientists and mathematicians

Table 3. Comparison between AI vs human with the same reference.

	AI wins	Human wins	Equal	Mean difference (AI–human) (± SD)
Appropriateness of the question	18 (36%)	27 (54%)	5 (10%)	- 0.11 ± 1.05
Clarity and specificity	18 (36%)	26 (52%)	6 (12%)	- 0.13 ± 1.08
Relevance	18 (36%)	27 (54%)	5 (10%)	- 0.32 ± 1.04
Quality of the alternatives & discriminative power	21 (42%)	26 (52%)	3 (6%)	- 0.10 ± 0.94
Suitability for graduate medical school exam	22 (44%)	28 (56%)	2 (4%)	- 0.14 ± 1.12
Total score	20 (40%)	30 (60%)	0 (0%)	- 0.80 ± 4.82

https://doi.org/10.1371/journal.pone.0290691.t003

Table 4. Blinded guess of question writer (i.e. AI vs human).

	AI (total = 50) (Correct guess, %)	Human (total = 50) (Correct guess, %)	Correlation	p
Assessor A	24, 48%	23, 46%	- 0.14–0.26	0.55
Assessor B	14, 28%	41, 82%	- 0.38–0.10	0.24
Assessor C	33, 66%	24, 48%	- 0.35–0.06	0.16
Assessor D	27, 53%	26, 52%	- 0.26–0.14	0.55
Assessor E	26, 52%	32, 64%	- 0.36–0.04	0.11
GPT-2 Output Detector	7, 14%	45, 90%	- 0.40–0.21	0.54

https://doi.org/10.1371/journal.pone.0290691.t004

explored the mathematical possibility of A.I. [29]. In recent years, A.I. has been used to analyze student learning abilities and history, allowing teachers to create the best learning program for all students [5].

The introduction of ChatGPT, an AI-powered chatbot, has also transformed the landscape of A.I. in education. ChatGPT was trained on a large dataset of real human conversations and online data, leading to its capability of song or poem writing, storytelling, list creation, and even passing exams [10]. In our study, no additional training is required for the user or extra tuning of ChatGPT to yield similar results except for the need to follow the commands. As demonstrated in our study, a reasonable MCQ can be written by ChatGPT using simple commands with a text reference provided. The questions written by humans were rated superior to those written by A.I. when we compare the questions with the same reference head-to-head (Table 3). However, the difference in our raters' average scores was insignificant except in the relevance domain (Table 2). This shows that despite the apparent superiority of the MCQ written by humans, the score difference is actually narrow and mostly insignificant. This indicates a huge potential to explore the use of such tools as assistance in other educational scenarios.

However, ChatGPT and other AI-powered tools in education have also raised concerns about their negative impacts on student learning [11]. Including ChatGPT, most A.I. models are trained by the vast content available on the internet, and their reliability and credibility are questionable. Moreover, many A.I.s were found to have significant bias due to their training data [30]. Another major potential setback related to natural language generator A.I. is called hallucination [31]. Like hallucination described in humans, this condition refers to a phenomenon where the A.I. generates nonsensical, or unfaithful to the provided source input. This has led to immediate recall even for some initially promising A.I. from some of the largest internet companies, such as Galactica from Meta Inc [32]. Hence, our team proposed the use of the A.I. by educators, which demonstrate a feasible way of utilization for educational purpose. To minimize bias and hallucination, our proposed methodology consists of providing a reliable reference for the A.I. to generate questions instead of complete dependence on its database [33].

Compatible with the guidance from the Division of Education, American College of Surgeons, there were minimal negative features in both the MCQs written by ChatGPT and humans (14% (7/50) vs 12% (6/50)), and there was also no use of "except" or "All/none of the above", which could create additional confusion to exam candidates [20]. However, we also acknowledge that there were intrinsic limitations with ChatGPT in generating MCQs for medical graduate examinations. First, as it is a pure language-based model, it cannot create any text and correlate it with clinical photos or radiological images. This is also an essential area in the exam, assessing candidates' interpretation skills. In addition, our pre-study tests have found that ChatGPT performed poorly when it was instructed to generate a clinical scenario, possibly due to the high complexity of knowledge and experience required to create a relevant scenario,

limiting its use to assess candidates' ability of application of their knowledge. However, with the continuous effort from the OpenAI team and the rapid evolution of A.I. technology, these barriers might be solved in the near future [34].

Limitations

The first limitation is that the reference material used was obtained directly from a textbook, and the length of the text was limited by the A.I. platform, which is currently at 510 tokens, potentially leading to selection bias by the operator. In contrast, human exam writers can recall information associated with the text, resulting in higher quality questions. Another limitation is the limited number of people involved and the number of questions generated, which could limit the applicability of the results. Only two professoriate staff from Medicine and Surgery departments, but not experts in all fields, developed the MCQs in this study. This differs from the real-life scenario where graduate exam questions were generated by a large pool of question writers from various clinical departments, which were then reviewed and vetted by a panel of professoriate staff. Besides, only 50 questions were generated by humans and another 50 by A. I., and only five assessors were involved. Real-world performance, in particular, the differentiate index, was impossible to assess due to the lack of actual students doing the test. Hence, efforts were made to improve the generalizability, including comprehensive coverage of all areas in both medicine and surgery and the participation of a multinational panel for assessment. Hence, efforts were made to improve the generalizability, including comprehensive coverage of all areas in both medicine and surgery and the participation of a multinational panel for assessment. The third limitation concerns the absence of human interference in the question-writing process. A.I. generated the questions, and the first-available question was captured without any polishing, while ChatGPT is also known to be sensitive to small changes in the prompt and can provide various answers even with the same prompt. And with extra fine-tuning in the form of further conversation with ChatGPT, the output quality from ChatGPT can be significantly enhanced. In addition, this study only evaluated the use of A.I. in generating MCQs. The full application of A.I. in developing the entire medical exam questions is yet to be thoroughly assessed. Lastly, with the improvements in the newer generation ChatGPT AI platform and other adjuncts, A.I. may perform better than what was observed in this current study. Nonetheless, this study provided solid evidence of the ability of ChatGPT and its strong potential in assisting medical exam MCQ preparation.

Conclusion

This is the first study showing that ChatGPT, a widely available large language model, can be utilized as an exam question writer for graduate medical exams with comparable performance to experienced human examiners. Our study supports the continuous exploration of how large language model A.I. can assist academia in improving their efficiency while maintaining a consistently high standard. Further studies are required to explore additional applications and other limitations of the booming A.I. platforms to enhance reliability with minimal bias.

Supporting information

S1 Checklist. PLOS ONE clinical studies checklist.
(DOCX)

S1 File. ChatGPT webpage.
(DOCX)

Author Contributions

Conceptualization: Billy Ho Hung Cheung, Michael Tiong-Hong Co.

Data curation: Billy Ho Hung Cheung, Gary Kui Kai Lau, Gordon Tin Chun Wong, Elaine Yuen Phin Lee.

Formal analysis: Billy Ho Hung Cheung, Michael Tiong-Hong Co.

Investigation: Billy Ho Hung Cheung, Gary Kui Kai Lau, Gordon Tin Chun Wong, Elaine Yuen Phin Lee, Dhananjay Kulkarni, Choon Sheong Seow, Ruby Wong.

Methodology: Billy Ho Hung Cheung, Gary Kui Kai Lau, Choon Sheong Seow, Michael Tiong-Hong Co.

Project administration: Billy Ho Hung Cheung, Gary Kui Kai Lau, Gordon Tin Chun Wong, Elaine Yuen Phin Lee, Dhananjay Kulkarni, Ruby Wong, Michael Tiong-Hong Co.

Resources: Billy Ho Hung Cheung, Gordon Tin Chun Wong, Elaine Yuen Phin Lee, Dhananjay Kulkarni, Choon Sheong Seow.

Software: Billy Ho Hung Cheung.

Supervision: Gordon Tin Chun Wong, Michael Tiong-Hong Co.

Validation: Billy Ho Hung Cheung, Michael Tiong-Hong Co.

Visualization: Billy Ho Hung Cheung.

Writing – original draft: Billy Ho Hung Cheung.

Writing – review & editing: Billy Ho Hung Cheung, Michael Tiong-Hong Co.

References

1. Nassar AK, Reid S, Kahnamoui K, Tuma F, Waheed A, McConnell M. Burnout among Academic Clinicians as It Correlates with Workload and Demographic Variables. Behavioral Sciences. 2020; 10(6):94. https://doi.org/10.3390/bs10060094 PMID: 32471265
2. Rao SK, Kimball AB, Lehrhoff SR, Hidrue MK, Colton DG, Ferris TG, et al. The Impact of Administrative Burden on Academic Physicians: Results of a Hospital-Wide Physician Survey. Academic Medicine. 2017; 92(2):237–43. https://doi.org/10.1097/ACM.0000000000001461 PMID: 28121687
3. Yeoh KG. The future of medical education. Singapore Med J. 2019; 60(1):3–8. https://doi.org/10.11622/smedj.2019003 PMID: 30840994
4. Wong BM, Levinson W, Shojania KG. Quality improvement in medical education: current state and future directions. Med Educ. 2012; 46(1):107–19. https://doi.org/10.1111/j.1365-2923.2011.04154.x PMID: 22150202
5. Chen L, Chen P, Lin Z. Artificial Intelligence in Education: A Review. IEEE Access. 2020; 8:75264–78.
6. Scotti V. Artificial intelligence. IEEE Instrumentation & Measurement Magazine. 2020; 23(3):27–31.
7. Jordan MI, Mitchell TM. Machine learning: Trends, perspectives, and prospects. Science. 2015; 349 (6245):255–60. https://doi.org/10.1126/science.aaa8415 PMID: 26185243
8. Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent abilities of large language models. arXiv preprint arXiv:220607682. 2022.
9. AI O. ChatGPT: Optimizing Language Models for Dialogue San Francisco, Canada2023. https://openai.com/blog/chatgpt/.
10. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. PLOS Digital Health. 2023; 2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198 PMID: 36812645
11. O'Connor S, ChatGpt. Open artificial intelligence platforms in nursing education: Tools for academic progress or abuse? Nurse Education in Practice. 2023; 66:103537. https://doi.org/10.1016/j.nepr.2022.103537 PMID: 36549229

12. Haladyna TM, Downing SM, Rodriguez MC. A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. Applied Measurement in Education. 2002; 15(3):309–33.
13. Kilgour JM, Tayyaba S. An investigation into the optimal number of distractors in single-best answer exams. Adv Health Sci Educ Theory Pract. 2016; 21(3):571–85. https://doi.org/10.1007/s10459-015-9652-7 PMID: 26597452
14. Dion V, St-Onge C, Bartman I, Touchie C, Pugh D. Written-Based Progress Testing: A Scoping Review. Academic Medicine. 2022; 97(5):747–57. https://doi.org/10.1097/ACM.0000000000004507 PMID: 34753858
15. Loscalzo J, Fauci AS, Kasper DL, Hauser S, Longo D, Jameson JL. Harrison's Principles of Internal Medicine, Twenty-First Edition (Vol.1 & Vol.2): McGraw Hill LLC; 2022.
16. Williams NS, O'Connell PR, McCaskie AW. Bailey & Love's Short Practice of Surgery: Taylor & Francis Group; 2018.
17. OpenAI. Introducing ChatGPT Plus 2023. https://openai.com/blog/chatgpt-plus/.
18. OpenAI. Supported countries and territories 2023. https://platform.openai.com/docs/supported-countries.
19. Kumar D, Jaipurkar R, Shekhar A, Sikri G, Srinivas V. Item analysis of multiple choice questions: A quality assurance test for an assessment tool. Medical Journal Armed Forces India. 2021; 77:S85–S9. https://doi.org/10.1016/j.mjafi.2020.11.007 PMID: 33612937
20. Brame CJ. Writing good multiple choice test questions 2013. https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/.
21. Iñarrairaegui M, Fernández-Ros N, Lucena F, Landecho MF, García N, Quiroga J, et al. Evaluation of the quality of multiple-choice questions according to the students' academic level. BMC Med Educ. 2022; 22(1):779. https://doi.org/10.1186/s12909-022-03844-3 PMID: 36369070
22. Gierl MJ, Bulut O, Guo Q, Zhang X. Developing, Analyzing, and Using Distractors for Multiple-Choice Tests in Education: A Comprehensive Review. Review of Educational Research. 2017; 87(6):1082–116.
23. Shin J, Guo Q, Gierl MJ. Multiple-Choice Item Distractor Development Using Topic Modeling Approaches. Front Psychol. 2019; 10:825. https://doi.org/10.3389/fpsyg.2019.00825 PMID: 31133911
24. Adams NE. Bloom's taxonomy of cognitive learning objectives. Journal of the Medical Library Association: JMLA. 2015; 103(3):152. https://doi.org/10.3163/1536-5050.103.3.010 PMID: 26213509
25. OpenAI. GPT-2 Output Detector 2022. https://huggingface.co/openai-detector.
26. Haladyna TM. Developing and validating multiple-choice test items: Routledge; 2004.
27. Vegada B, Shukla A, Khilnani A, Charan J, Desai C. Comparison between three option, four option and five option multiple choice question tests for quality parameters: A randomized study. Indian J Pharmacol. 2016; 48(5):571–5. https://doi.org/10.4103/0253-7613.190757 PMID: 27721545
28. Epstein RM. Assessment in Medical Education. New England Journal of Medicine. 2007; 356(4):387–96. https://doi.org/10.1056/NEJMra054784 PMID: 17251535
29. Haenlein M, Kaplan A. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. California management review. 2019; 61(4):5–14.
30. Howard FM, Dolezal J, Kochanny S, Schulte J, Chen H, Heij L, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. Nat Commun. 2021; 12(1):4423. https://doi.org/10.1038/s41467-021-24698-1 PMID: 34285218
31. Maynez J, Narayan S, Bohnet B, McDonald R. On faithfulness and factuality in abstractive summarization. arXiv preprint arXiv:200500661. 2020.
32. Heaven WD. Why Meta's latest large language model survived only three days online US: MIT Technology Review 2023. https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/.
33. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. Cureus. 2023; 15(2):e35179. https://doi.org/10.7759/cureus.35179 PMID: 36811129
34. OpenAI. GPT-4 2023. https://openai.com/research/gpt-4.