Medical Science Educator
https://doi.org/10.1007/s40670-025-02293-z

ORIGINAL RESEARCH

Can ChatGPT-4o Really Pass Medical Science Exams? A Pragmatic Analysis Using Novel Questions

Philip M. Newton[1] · **Christopher J. Summers**[1] · **Uzman Zaheer**[1] · **Maira Xiromeriti**[1] · **Jemima R. Stokes**[1] · **Jaskaran Singh Bhangu**[1] · **Elis G. Roome**[1] · **Alanna Roberts-Phillips**[1] · **Darius Mazaheri-Asadi**[1] · **Cameron D. Jones**[1] · **Stuart Hughes**[1] · **Dominic Gilbert**[1] · **Ewan Jones**[1] · **Keioni Essex**[1] · **Emily C. Ellis**[1] · **Ross Davey**[1] · **Adrienne A. Cox**[1] · **Jessica A. Bassett**[1]

Accepted: 17 January 2025

Abstract

ChatGPT apparently shows excellent performance on high-level professional exams such as those involved in medical assessment and licensing. This has raised concerns that ChatGPT could be used for academic misconduct, especially in unproctored online exams. However, ChatGPT has previously shown weaker performance on questions with pictures, and there have been concerns that ChatGPT's performance may be artificially inflated by the public nature of the sample questions tested, meaning they likely formed part of the training materials for ChatGPT. This led to suggestions that cheating could be mitigated by using novel questions for every sitting of an exam and making extensive use of picture-based questions. These approaches remain untested. Here, we tested the performance of ChatGPT-4o on existing medical licensing exams in the UK and USA, and on novel questions based on those exams. ChatGPT-4o scored 94% on the United Kingdom Medical Licensing Exam Applied Knowledge Test and 89.9% on the United States Medical Licensing Exam Step 1. Performance was not diminished when the questions were rewritten into novel versions, or on completely novel questions which were not based on any existing questions. ChatGPT did show reduced performance on questions containing images when the answer options were added to an image as text labels. These data demonstrate that the performance of ChatGPT continues to improve and that secure testing environments are required for the valid assessment of both foundational and higher order learning.

Keywords Assessment validity · Academic integrity · Cheating · Evidence-based education · MCQs · Pragmatism

Introduction

New generative artificial intelligence (GenAI) tools such as ChatGPT have attracted enormous attention, in part for their apparent ability to pass high-level professional exams. These bots are based on underlying large language models (LLMs) which differ widely in their performance. The version of ChatGPT, running the GPT-4 LLM, scored 86% on the United States Medical Licensing Exam (USMLE) Step 1 [1] and 76.4% on the United Kingdom Medical Licensing Exam Applied Knowledge Test (UK MLA AKT) [2]. These MCQ-based exams test high-level problem-solving, requiring the application of core knowledge to clinical scenarios [3], and so have raised questions about the security of online exams.

However, there have been a number of responses to, and criticisms of, the claim that ChatGPT is genuinely solving the problems presented in these questions. Instead, critics propose tools like ChatGPT are more likely 'regurgitating' content which has been in their training materials [4], and many studies use sample papers which are in the public domain and have been for some time. This regurgitation is proposed to be a paraphrasing of prior training materials in a way that resembles a student who is plagiarising a piece of text by changing key words but without understanding the meaning, and so occasionally getting things (very) wrong [5]. Thus, the argument goes to counter the apparent threat of ChatGPT to exam security and integrity; educators could use novel questions for each sitting of the exam [6]. In addition, there have been efforts to identify features of exam questions which ChatGPT might struggle with, for example

---

✉ Philip M. Newton
p.newton@swansea.ac.uk

[1] Swansea University Medical School, Swansea, Wales SA2 8PP, UK

an increase in the number of answer items, increasing language complexity or having multiple correct answers. However, none of these appears to have any effect on ChatGPT performance [7]. Many early papers which tested the performance of ChatGPT on sample exams deliberately excluded questions containing images, on the basis that older versions of ChatGPT, even GPT-4, could not process these images. Thus, the reported performance of ChatGPT may be an over-estimation, since the percentage scored by ChatGPT uses a lower denominator once image-based questions are excluded (e.g. [8]). This also led to proposals that educators could author 'ChatGPT-proof' questions by including images, along with mathematical calculations and reasoning tests, which it is proposed that ChatGPT does not perform well at [9].

These issues are important in part because of wider questions about the security, but also the inclusivity and cost, of examinations. Online examinations are cheaper and more flexible than their in-person equivalents, but they potentially increase the risk of cheating. During the COVID-19 pandemic, the percentage of students who admitted to cheating in online exams appeared to double, and more students reported cheating than not [10]. One apparent solution to this problem is to increase the use of online proctoring/invigilation systems to monitor student behaviour. However, these then drive back up the cost of the online exams, and the student experience of remote proctoring is poor, with concerns about privacy, fairness, inclusivity, and cost [11, 12]. An alternative is to avoid the use of proctoring altogether. A high-profile 2023 publication analysed exam performance data from the COVID lockdown and concluded that unproctored online exams are a 'valid and meaningful' way of measuring student learning [13], although this analysis has been challenged [14] and does not include a consideration of ChatGPT. Thus, it is important to understand whether ChatGPT truly can pass exams, including novel questions with images, as part of a consideration about how best to deploy exams, online or in-person, proctored or not.

Pragmatism is the research paradigm adopted here. It prioritises the asking of questions whose answers will be useful, rather than perhaps asking more academic or basic questions [15]. If ChatGPT truly can pass high-level STEM exams, even with novel questions containing images, then from a pragmatic standpoint, this is important because it essentially settles any debate about whether these examinations can be conducted in an online, unproctored format. From the pragmatic perspective, it does not matter *how* ChatGPT is answering these questions, either by truly solving problems or through some sophisticated paraphrasing. There is a related pragmatic issue, which is that for most STEM subjects there is a core curriculum: a basic set of knowledge and skills which graduates must be able to demonstrate in order to graduate, and also to be able to apply knowledge to practice. This cumulative view of learning has a long history and remains prevalent today through the use of instruments such as Bloom's taxonomy [16]. In essence, we cannot expect students to undertake learning and practice at the higher levels of Bloom's taxonomy unless they have the core foundational knowledge to be applied to those higher levels. Thus, educators need to assess that foundational knowledge first, before it is applied, particularly where there are safety concerns, e.g. for patients. However, from the pragmatic perspective, it seems reasonable to propose that there are only so many ways that one can phrase exam questions which assess these core principles. This then creates a risk that if educators strive to write completely novel questions on every core topic for every exam sitting, just to thwart ChatGPT, then this will rapidly become impossible. These issues also have relevance for the proposed positive benefits of ChatGPT. It offers great promise as a tutoring tool for students who are preparing for exams [17] but educators and learners both need to be confident that the answers given are logical and reasonable [18].

Some of the controversy and discourse about the apparent ability of ChatGPT to pass and perform well (or not) on exams likely come from the frequent updating of ChatGPT over a short timescale. A review of ChatGPT's performance on exams from multiple disciplines found that the then-subscription version of ChatGPT, running the LLM GPT-4, outperformed the then-free version running GPT-3 or 3.5, with the average difference being 25 percentage points [19]. On May 13, 2024, OpenAI, the creators of ChatGPT, released another update, entitled ChatGPT-4o, showing enhanced performance compared to GPT-4, particularly on the integration of text, visual, and audio information [20]. The performance of ChatGPT-4o on medical licensing exams has not yet been examined, and determining this performance also has a pragmatic value—efforts to thwart ChatGPT based on the performance of GPT-4 may be redundant if GPT-4o has an even better performance.

Here then we address the following research questions. It is important to be clear that the specific medical licensing-type exams tested here are used, for pragmatic purposes, as a model for STEM exams generally, given that they are written to a high standard and are aimed at problem-solving and the application of knowledge [3, 21].

1. How well does ChatGPT-4o perform on sample medical licensing exams in the USA and UK?
2. Is the performance of ChatGPT-4o affected when these sample questions are rewritten into novel formats, but assessing the same core curricular concepts?
3. How well does ChatGPT-4o perform on completely novel medical-licensing type questions?

These are pragmatic research questions since their answers should be useful for educators, regardless of what those answers are. For example, if ChatGPT-4o shows impaired performance on novel questions, then the security of online exams might be increased by ensuring that questions are completely novel, whereas if performance is unchanged on novel vs sample questions, then the converse is true.

Methods

The following question sources were tested.

1. (Pilot) Wikiversity Fundamentals of Neuroscience Exam [22]
2. Sample paper 1, UK Medical Licensing Assessment Applied Knowledge Test [23]
3. USMLE Step 1 Sample paper [24]
4. Rewritten questions from 2 + 3
5. Completely Novel USMLE-style questions.

Rewriting of Existing Questions in the Public Domain

Each question from sources 1–3 was rewritten by a member of the research team. Each question was rewritten three times with each rewrite undertaken by a different team member. Instructions were to 'change as much as possible about the question without changing the underlying learning. Change all the text where possible'. Suggestions of specific items to change included demographic details in the scenarios, answer options, and answer order. Each team member was also provided with a summary of common issues found when writing USMLE-style questions [3] and asked to avoid any of the identified writing flaws. All rewritten items were checked for accuracy and originality by registered doctors (CJS, RD) or a subject matter expert (PMN) and adjusted where necessary, for example if the revised question could be made even more different to the original question.

An initial pilot was undertaken using five questions on neuroscience from the 'Wikiversity' website. These were considered 'lower order' questions, assessing basic factual knowledge of neurological disease. The questions have been in the public domain since 2013. Each question was rewritten into three different forms by a member of the research team, who then discussed the process and feasibility of scaling the methodology to a larger exam. All four versions of each question were then pilot tested using GPT-4 on 23/04/24 and 24/04/24.

Analysis of Existing Medical Licensing Exams and Rewrites

Each question was tested using a single-shot method, the most likely approach taken by a student who was seeking to cheat on an MCQ exam, where the text was highlighted in the pdf (original questions) or word document (rewrites), copied and then pasted directly into ChatGPT-4o with no attempt to format the text. Where the question included a picture, this was copied using screen clipping, saved, and uploaded as an image (.png) file with only the country and the question number as the file name (e.g. 'UK32'). No additional prompts were given apart from the content of the question. Each question was asked in a new chat and no memory functions were activated. For the USMLE questions, a 'temporary chat' was activated for each question. No responses were given to ChatGPT. ChatGPT's first response was recorded each time as correct/incorrect. ChatGPT-4o tests were undertaken May 14–24 2024.

Creation and Analysis of Novel Questions

Three sets of completely novel questions were generated, totalling 90 questions in all. A first set of forty novel questions were created in the style of questions for the UK MLA AKT and USMLE, by an author who is experienced in the creation of these assessment items (CS), according to guidance from the United States National Board of Medical Examiners [3]. Ten of these questions included novel images that were either created for this study or were images from the private collection of one of the authors (CS). None of these images is available in the public domain. All images were obtained with appropriate consent and anonymised prior to use in keeping with paragraph 10 of the General Medical Council's professional standards on making and using visual recordings of patients [25]. These questions were mapped to curricula items from the MLA content map [26] and were of a comparative style and difficulty to the MLA. A second set of questions was written by an author (PMN) using guidance for the creation of multiple-choice questions which assess higher order learning in STEM. These guidelines include identifying assumed knowledge, creating problem-solving scenarios, and the use of actions as answer options [21]. Some of these questions included images sourced from Wikimedia Commons. During this process, the authors observed a trend that ChatGPT appeared to struggle with anatomical images that had novel text labels, e.g. a brain section with the labels A-H added, with arrows to specific brain regions that corresponded to question answers. To probe this further, the third set of questions comprised 14 pairs which assessed the same learning but either using a labelled image, or text equivalent. Finally, ChatGPT was then asked simply to identify the labels on the images from these questions where possible. Each question was asked in a new 'temporary chat'. ChatGPT-4o tests were undertaken May 24–Jun 18, 2024.

Results

Summary

We asked three research questions: first, how well does ChatGPT-4o perform on sample medical licensing exams in the USA and UK, then whether this performance is affected when these sample questions are rewritten into novel formats, but assessing the same core curricular concepts, and finally how well does ChatGPT perform on completely novel medical-licensing type questions? Our findings demonstrate that ChatGPT-4o performs extremely well, outperforming previous versions, and that this performance is maintained on novel questions of either type. We did however find reduced performance on a specific format of image-based question where the answer options were given as text labels added to the image. An example of this format and the text equivalent is given in Fig. 1. All results are summarized in Table 1, with the details given below.

We tested a total of 705 assessment items, of which ChatGPT answered 635 (90%) correctly. Of the 219 original sample questions, 28 of these questions contained images, of which ChatGPT answered 20 (71.4%) correctly. A breakdown of these items is below.

Wikiversity Pilot

GPT-4 correctly answered all versions of all questions, both the originals and the rewritten versions.

United Kingdom Medical Licensing Assessment, Applied Knowledge Test

ChatGPT-4o answered 94 of 100 questions on the original sample paper. Five of the questions included pictures. ChatGPT answered four of these correctly. ChatGPT then scored 93%, 91%, and 95% on the three collections of rewrites. One question, on herpes zoster ophthalmicus, was answered incorrectly on all four occasions. In all other cases, there was no consistent pattern. Some questions that ChatGPT had answered incorrectly on the original sample paper were answered correctly once rewritten, but the converse was also true for other questions. A majority (85%) of questions were answered correctly in all four versions (original and all three rewrites). The full dataset and question ID is in Supplementary Data S1.

United States Medical Licensing Exam Step 1

ChatGPT-4o scored 89.9% (107/119) of the original questions correctly. Of the original 119, there were images in 23 of them, of which 16 (69.6%) were answered correctly. This suggested that ChatGPT might struggle more with the picture questions in this particular exam. Given that ChatGPT-4o had already demonstrated no impairment of performance when rewriting text questions from the UK MLA AKT into a novel format, we decided to rewrite only a sample of 27 of the USMLE questions, but to probe further this possible diminished performance on questions containing pictures by including 13 picture questions, of which 5 had been answered incorrectly from the original paper. Of the sample of 27, ChatGPT scored 74.1% (20/27) on the original versions, and then 85.2 (23/27), 70.4% (19/27), and 85.2% (23/27) on the rewrites. Only one question was answered incorrectly in all four versions. This was a picture question based on a graph, while the other four picture questions which ChatGPT had answered incorrectly were then answered correctly at least once during the rewrites. A majority (55.6%, 15/27) of questions were answered correctly on all four occasions. The full dataset and question ID is in Supplementary Data S1.

Novel USMLE-Style Questions

A total of 62 novel questions were generated, of which ChatGPT answered 60 (98.8%) correctly.

Questions with Labels on an Image

Twenty-eight questions were generated in pairs (2 × 14) which assessed the same learning in each pair. One version of the question contained a labelled image where the labels were simple letters (A, B, C etc.) and these were the answer options; for example, the image was a picture of the brain with different regions labelled A-H. The paired question contained answer options in text form; for example, the brain regions were listed as text. An example of this format is in Fig. 1. ChatGPT answered 13/14 of the text version of these questions, but only 2/14 of the labelled image questions. A summary of the analysis is in Supplementary Data S1. The novel questions may be shared upon request but are not published here due to the images contained within.

Identification of Labels on Images

Ten of the labelled images were structured in a way that it was reasonable to upload them to ChatGPT-4o with the prompt 'Can you identify all the labels (A–X) on the uploaded image?' where 'X' was either E, F, G, or H depending on the number of labels. Of a total of 66 labels

Common scenario *An elderly gentleman is rushed to hospital after being found on the floor at home. He appears to be able to breathe and his heartrate is elevated but stable. However he appears to be completely paralysed and does not respond when asked questions. His pupils are pinpoints. He does not blink when something goes near his eyes, but when a light is shone into his eyes, they move horizontally to follow the light.*

Text question *Damage to which structure in the brain is most likely to result in the above presentation?*

A. Primary Motor Cortex
B. Hippocampus
C. Cerebellum
D. Nucelus Accumbens
E. Globus Pallidus
F. Substantia Nigra
G. Pons
H. Medulla

Image Question *Which of the structures in the image is most likely to be damaged?*

A
B
C
D
E
F
G
H

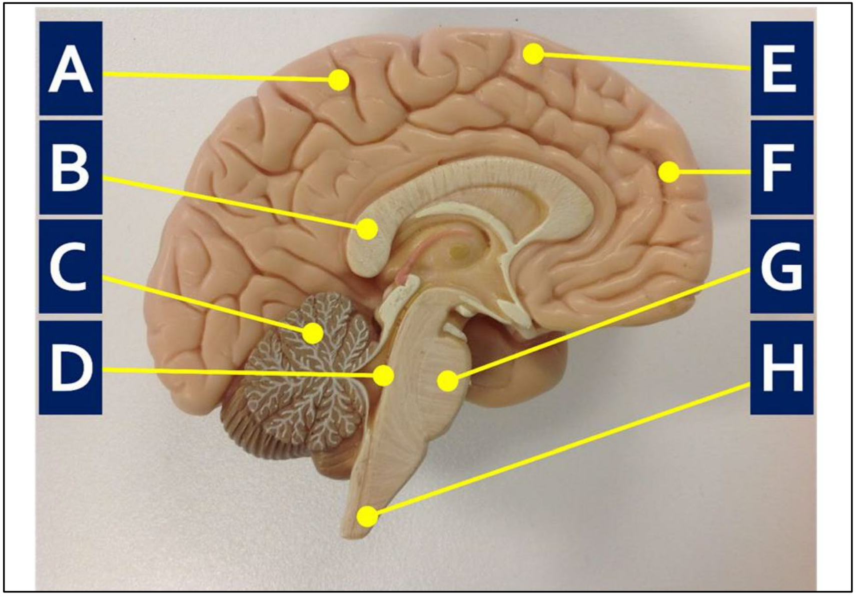

Fig. 1 An example of a novel higher order MCQ written using established guidelines [21], with text options as answers (which ChatGPT answers correctly), or a labelled image (which ChatGPT answers incorrectly). Note that the answer options do not correspond exactly

across the 10 images, ChatGPT correctly labelled 25 items. For all 10 images, ChatGPT correctly identified the main structure in the image (e.g. brain, kidney) but not the labelled subregions.

Discussion

ChatGPT-4o showed a very high level of performance on the papers tested, even when the questions were rewritten so that they assessed the same learning but with different wording. This level of performance was also found on completely novel questions written in the style of professional licensing exams. Our analysis included many questions based on images, and almost all questions were designed to assess higher-order problem-solving [3, 21].

A repeated finding from the research on academic misconduct demonstrates that one of the strongest factors contributing to an increased likelihood in the occurrence of academic dishonesty is the ease with which it can be committed [10, 27]. Cheating in online exams was already high before the emergence of ChatGPT [10] and our findings

Table 1 Summary performance of ChatGPT-4o on the questions tested here. The results for the 'images' format show only results for questions containing images, whereas those from 'All' contain those same image questions along with every other format. See "Methods" for acronyms and question sources

Question source/type	Number of questions	Version	Format	*N* correct	% correct
UK MLA AKT	100	Original	All	93	93
		Rewrite #1	All	93	93
		Rewrite #2	All	91	91
		Rewrite #3	All	95	95
	5	Original	Images	4	80
USMLE Step 1	119	Original	All	107	89.9
	23	Original	Images	16	69.6
	27	Original	All[a]	20	74.1
		Rewrite #1	All[a]	23	85.2
		Rewrite #2	All[a]	19	70.4
		Rewrite #3	All[a]	23	85.2
Novel USMLE	62	-	All	60	96.8
	10	-	Images	9	90
Novel USMLE with labelled image	14	Text only		13	92.9
	14	Text only		2	14.3

[a] Contained 13/27 picture questions

demonstrate that any student using ChatGPT would likely receive an excellent mark even if they had no prior knowledge whatsoever, further increasing any temptation to cheat. Thus, it seems reasonable to propose that our findings mean online unproctored summative exams are difficult to justify as a valid form of assessment, a conclusion which is in contrast to findings published following an analysis of exam performance during the COVID pandemic, but before the emergence of ChatGPT [13].

The high-performance levels of ChatGPT may also increase the temptation to cheat using ChatGPT even in proctored exams, particularly if they are taken online; data suggest that proctoring considerably reduces cheating in online exams but does not eliminate it completely [10]. We are not aware of any current data on the extent to which students are using ChatGPT to cheat in online exams, proctored or unproctored, although this is the subject of ongoing work. A study conducted in Vietnam in May 2023 showed that 23.7% of undergraduates cheated using ChatGPT, although the assessment formats were not specified [28]. A study conducted at around the same time in US high schools found similar numbers in one school, though lower in two others [29]. These figures seem likely to increase as ChatGPT becomes better known and more widely available, along with similar tools such as Claude.AI.

One intuitive response to these challenges is to design questions which ChatGPT finds harder to answer. This 'arms race' approach is partly the genesis of the current paper, based on earlier studies which observed that ChatGPT could not process image-based questions at all, and other studies suggesting that ChatGPT is a 'copy and paste' machine whose impact can be minimized by using novel questions for each sitting of an exam [6]. Our findings do suggest that ChatGPT-4o struggled more on questions with images overall, though the sample was quite small. However, we saw a strong indication of poor performance a very specific type of MCQ, where the answer items were single-letter labels and arrows on images. There is more than one possible explanation for this apparent weakness. These questions are designed to require 'assumed knowledge' and so to be harder to answer than factual recall questions [21]. For example, the picture item shown in Fig. 1 requires the test taker to know that the scenario represents the clinical condition locked-in syndrome, and then to know that this condition is associated with damage to the part of the brain called the pons, and then to be able to identify the anatomical location of the pons on a picture of a model. ChatGPT consistently struggled with these specific formats of image questions and so one interpretation is that it is the 'multi-step' nature of these questions which trips up ChatGPT. However, ChatGPT was consistently correct on the text versions of these questions and would give detailed descriptions of the answer option. ChatGPT was also clearly able to identify, in text form, where the pons is located (for example). But when simply asked to identify the labels on these images, ChatGPT struggled, indicating that it is the processing of these specific types of text-labelled images which ChatGPT struggles with, rather than the solving of multi-step problems.

One potential conclusion from these findings with images is that such questions could be used to thwart ChatGPT and so deter cheating in online exams. However, from our pragmatic perspective, we caution against this interpretation. Writing an entire exam based on these types of questions seems implausible and unlikely to be valid. This limitation likely applies to other methods identified as a way of 'defeating' ChatGPT. For example, an older study, using an unidentified version of ChatGPT, showed that ChatGPT overselects answer options 'all of the above' or 'none of the above', meaning that when these answer options are present but are incorrect, ChatGPT shows a much lower performance compared to when these answer options are absent or when they are present but are the correct answer. However, designing questions which incorporate this flaw also seems likely to be a short-term measure that may well result in poorer quality questions and weaker curriculum coverage. These types of answer options are also advised against when writing high-quality assessment items [21]. It may be that there are other patterns of question format that ChatGPT is unable to answer, but the very high accuracy shown here by ChatGPT means that only a small number of questions were answered incorrectly, making it difficult to identify any other consistent patterns.

Any reduction in the use of online unproctored exams will clearly not eradicate academic misconduct. There are a wide range of dishonest behaviours undertaken by medical and other students [30], and the performance of ChatGPT on assessment formats such as essays is also very strong [31]. Essays are, by design, asynchronous and unmonitored, meaning that it would be almost impossible to prevent a student from using ChatGPT to complete assignments in these formats. Detection tools have been developed and these appear to show good accuracy for raw text generated by tools such as ChatGPT [32] but they can be easily circumvented [33] and even a very small rate of false-positives is problematic since there is no independent source to match a student assignment to, unlike with 'conventional' plagiarism, meaning that problematic, adversarial situations can quickly arise when students are accused of cheating on essays using ChatGPT [34].

The performance of ChatGPT-4o demonstrated here shows a modest improvement when compared to that seen using GPT-4, which itself shows a much improved performance compared to GPT-3 and GPT-3.5 [19], although many prior papers excluded image-based questions from their analyses whereas they are included here. This trend of improving performance seems likely to continue; at the time of writing (July 2024), OpenAI are rolling out enhanced visual recognition features in GPT-4o to their subscribers, meaning that users will be able to simply point their camera at an exam question and it will scan and 'read' the text before generating an answer [20].

The high performance of ChatGPT-4o on the exams tested here and elsewhere leads naturally to a question of whether these tools might also be able to *write* such exams. A review on some of the older versions of these tools concluded that question generation is possible although with some limitations, and proposed further testing [35]. It is now possible to upload considerable volumes of data to ChatGPT and to build custom GPTs which have specific instructions tailored to certain tasks, as designed by the creator. This approach has already shown promise for the creation of USMLE-style assessment items and may even be able to generate an entire exam and blueprint it to a curriculum, saving considerable time and cost for educators and universities [36]. This possibility arose during the conduct of the study here wherein some questions that were initially answered incorrectly by ChatGPT revealed either strong distractors or potential ambiguities in the question stem or associated image, suggesting weaknesses in the question itself. No questions tested here were eliminated from analysis for being actually incorrect or of poor quality, but this analysis suggested that such issues might be easily identified by using ChatGPT as an adjunct to exam creation and standard setting.

Similar benefits could also be obtained for students. The research team here noted the accuracy and value of the explanations provided by ChatGPT when answering the questions, and these naturally suggest the potential of ChatGPT, and the aforementioned custom GPTs, as study tools for students. Such an approach has been successfully used in ophthalmology [37] and anatomy learning [38].

Conclusion

ChatGPT-4o shows very high levels of performance on MCQ-based applied knowledge tests, including questions with images. These data echo but improve further upon findings from earlier versions of ChatGPT [39] and suggest that educators will find it extremely difficult to write valid questions which are 'ChatGPT-proof', even if they are completely novel and image-based. Given the expanding use and continued improvement of LLM-based chatbots, our data suggest that it is difficult to defend the use unproctored online exams for summative assessment, even when assessing higher order learning. These assessments, and lower-level MCQs-based exams testing core foundational knowledge, should only be conducted under secure conditions.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40670-025-02293-z.

Data Availability All raw data is available in the supplementary material.

Declarations

Conflict of Interest The authors declare no competing interests.

References

1. Garabet R, Mackey BP, Cross J, Weingarten M. ChatGPT-4 performance on USMLE Step 1 style questions and its implications for medical education: a comparative study across systems and disciplines. Med Sci Educ. 2024;34(1):145–52.
2. Lai UH, Wu KS, Hsu TY, Kan JKC. Evaluating the performance of ChatGPT-4 on the United Kingdom medical licensing assessment. Front Med. 2023;19(10):1240915.
3. Billings M, DeRuchie K, Hussie K, Kulesher A, Merrell J, Morales A, et al. Constructing written test questions for the health sciences [Internet]. National Board of Medical Examiners; 2020 [cited 2022 Apr 7]. Available from: https://www.nbme.org/sites/default/files/2020-11/NBME_Item%20Writing%20Guide_2020.pdf.
4. Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023;29(3):721–32.
5. Marcus G. Partial regurgitation and how LLMs really… [Internet]. Marcus on AI. 2024 [cited 2024 Jun 3]. Available from: https://garymarcus.substack.com/p/partial-regurgitation-and-how-llms/comments.
6. Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. Educ Sci. 2023;13(4):410.
7. Ram S, Qian C. A study on the vulnerability of test questions against ChatGPT-based cheating. In: 2023 International Conference on Machine Learning and Applications (ICMLA) [Internet]. 2023 [cited 2024 Jun 17]. p. 1710–5. Available from: https://ieeexplore.ieee.org/abstract/document/10460039.
8. Abbas A, Rehman MS, Rehman SS. Comparing the performance of popular large language models on the national board of medical examiners sample questions. Cureus. 2024;16(3):e55991.
9. Arkoudas K. GPT-4 can't reason [Internet]. arXiv: 2308.03762 [Preprint]. 2023 [cited 2024 Feb 18]. Available from: http://arxiv.org/abs/2308.03762.
10. Newton PM, Essex K. How common is cheating in online exams and did it increase during the COVID-19 pandemic? A systematic review. J Acad Ethics [Internet]. 2023 Aug 4 [cited 2023 Aug 7]; Available from: https://doi.org/10.1007/s10805-023-09485-5.
11. Marano E, Newton PM, Birch Z, Croombs M, Gilbert C, Draper MJ. What is the student experience of remote proctoring? A pragmatic scoping review. High Educ Q. 2024;78(3):1031–47.
12. Meulmeester FL, Dubois EA, Krommenhoek-van Es C (Tineke), de Jong PGM, Langers AMJ. Medical students' perspectives on online proctoring during remote digital progress test. Med Sci Educ. 2021; 31(6):1773–7.
13. Chan JCK, Ahn D. Unproctored online exams provide meaningful assessment of student learning. Proc Natl Acad Sci. 2023;120(31):e2302020120.
14. Newton PM. The validity of unproctored online exams is undermined by cheating. Proc Natl Acad Sci. 2023;120(41):e2312978120.
15. Newton PM, Da Silva A, Berry S. The case for pragmatic evidence-based higher education: a useful way forward? Front Educ [Internet]. 2020 [cited 2021 May 8]; 5. Available from: https://www.frontiersin.org/articles/https://doi.org/10.3389/feduc.2020.583157/full.
16. Newton PM, Da Silva A, Peters LG. A pragmatic master list of action verbs for Bloom's taxonomy. Front Educ [Internet]. 2020 [cited 2020 Jul 14]; 5. Available from: https://www.frontiersin.org/articles/https://doi.org/10.3389/feduc.2020.00107/full.
17. Koga S. The potential of ChatGPT in medical education: focusing on USMLE preparation. Ann Biomed Eng. 2023;51(10):2123–4.
18. Daungsupawong H, Wiwanitkit V. ChatGPT-4 performance on USMLE Step 1 style questions and its implications for medical education: correspondence. Med Sci Educ [Internet]. 2024 Apr 5 [cited 2024 Jun 3]; Available from: https://doi.org/10.1007/s40670-024-02033-9.
19. Newton P, Xiromeriti M. ChatGPT performance on multiple choice question examinations in higher education. A pragmatic scoping review. Assess Eval High Educ. 2024;0(0):1–18.
20. OpenAI. Hello GPT-4o [Internet]. [cited 2024 Jun 3]. Available from: https://openai.com/index/hello-gpt-4o/.
21. Newton PM. Guidelines for creating online MCQ-based exams to evaluate higher order learning and reduce academic misconduct. In: Eaton SE, editor. Handbook of academic integrity [Internet]. Singapore: Springer Nature; 2023 [cited 2023 Jul 13]. p. 1–17. Available from: https://doi.org/10.1007/978-981-287-079-7_93-1.
22. Wikiversity. Fundamentals of neuroscience/exams - Wikiversity [Internet]. 2013 [cited 2024 Feb 10]. Available from: https://en.wikiversity.org/wiki/Fundamentals_of_Neuroscience/Exams.
23. Medical Schools Council. Practice exam for the MS AKT | Medical Schools Council [Internet]. 2023 [cited 2024 Mar 10]. Available from: https://www.medschools.ac.uk/medical-licensing-assessment/preparing-for-the-ms-akt/practice-exam-for-the-ms-akt.
24. United States Medical Licensing Examination. Step 1 sample test questions | USMLE [Internet]. 2021 [cited 2024 Jun 10]. Available from: https://www.usmle.org/prepare-your-exam/step-1-materials/step-1-sample-test-questions.
25. GMC. Making and using visual and audio recordings of patients (summary) [Internet]. General Medical Council; 2011 [cited 2023 Jun 15]. Available from: https://www.gmc-uk.org/professional-standards/professional-standards-for-doctors/making-and-using-visual-and-audio-recordings-of-patients.
26. GMC. MLA content map [Internet]. 2021 [cited 2024 Jun 15]. Available from: https://www.gmc-uk.org/education/medical-licensing-assessment/mla-content-map.
27. Bretag T, Harper R, Burton M, Ellis C, Newton P, Rozenberg P, et al. Contract cheating: a survey of Australian university students. Stud High Educ. 2019;44(11):1837–56.
28. Nguyen HM, Goto D. Unmasking academic cheating behavior in the artificial intelligence era: evidence from Vietnamese undergraduates. Educ Inf Technol [Internet]. 2024 Feb 5 [cited 2024 Feb 18]; Available from: https://doi.org/10.1007/s10639-024-12495-4.
29. Lee VR, Pope D, Miles S, Zárate RC. Cheating in the age of generative AI: a high school survey study of cheating behaviors before and after the release of ChatGPT. Comput Educ Artif Intell. 2024;1(7):100253.
30. Henning MA, Chen Y, Ram S, Malpas P. Describing the attributional nature of academic dishonesty. Med Sci Educ. 2019;29(2):577–81.
31. Herbold S, Hautli-Janisz A, Heuer U, Kikteva Z, Trautsch A. AI, write an essay for me: a large-scale comparison of human-written versus ChatGPT-generated essays [Internet]. arXiv: 2304.14276 [Preprint] 2023 [cited 2023 May 8]. Available from: http://arxiv.org/abs/2304.14276.
32. Weber-Wulff D, Anohina-Naumeca A, Bjelobaba S, Foltýnek T, Guerrero-Dib J, Popoola O, et al. Testing of detection tools for AI-generated text [Internet]. arXiv: 2306.15666 [Preprint] 2023 [cited 2023 Aug 7]. Available from: http://arxiv.org/abs/2306.15666.

33. Perkins M, Roe J, Vu BH, Postma D, Hickerson D, McGaughran J, et al. arXiv: 2403.19148v1 [Preprint] [cited 2024 Jun 11]. GenAI detection tools, adversarial techniques and implications for inclusivity in higher education. Available from: https://arxiv.org/abs/2403.19148v1.
34. Gorichanaz T. Accused: how students respond to allegations of using ChatGPT on assessments. Learn Res Pract [Internet]. 2023 Jul 3 [cited 2024 May 3]; Available from: https://www.tandfonline.com/doi/abs/https://doi.org/10.1080/23735082.2023.2254787.
35. Artsi Y, Sorin V, Konen E, Glicksberg BS, Nadkarni G, Klang E. Large language models for generating medical examinations: systematic review. BMC Med Educ. 2024;24(1):354.
36. Kıyak YS, Kononowicz AA. Case-based MCQ generator: a custom ChatGPT based on published prompts in the literature for automatic item generation. Med Teach [Internet]. 2024 Feb 6 [cited 2024 Jun 11]; Available from: https://www.tandfonline.com/doi/abs/https://doi.org/10.1080/0142159X.2024.2314723.
37. Sevgi M, Antaki F, Keane PA. Medical education with large language models in ophthalmology: custom instructions and enhanced retrieval capabilities. Br J Ophthalmol [Internet]. 2024 May 7 [cited 2024 Jun 11]; Available from: https://bjo.bmj.com/content/early/2024/05/07/bjo-2023-325046.
38. Collins BR, Black EW, Rarey KE. Introducing AnatomyGPT: a customized artificial intelligence application for anatomical sciences education. Clin Anat [Internet]. [cited 2024 Jun 11];n/a(n/a). Available from: https://onlinelibrary.wiley.com/doi/abs/https://doi.org/10.1002/ca.24178.
39. Sood A, Mansoor N, Memmi C, Lynch M, Lynch J. Generative pretrained transformer-4, an artificial intelligence text predictive model, has a high capability for passing novel written radiology exam questions. Int J Comput Assist Radiol Surg. 2024;19(4):645–53.