Received: 1 March 2023 | Revised: 8 July 2023 | Accepted: 24 July 2023

DOI: 10.1111/eje.12937

ORIGINAL ARTICLE

WILEY

ChatGPT—A double-edged sword for healthcare education? Implications for assessments of dental students

Kamran Ali | Noha Barhom | Faleh Tamimi | Monty Duggal

College of Dental Medicine QU Health, Qatar University, Doha, Qatar

Correspondence
Kamran Ali, College of Dental Medicine QU Health, Qatar University, Doha, Qatar.
Email: ali.kamran@qu.edu.qa

Funding information
Open access funding was provided by the Qatar National Library

Abstract

Introduction: Open-source generative artificial intelligence (AI) applications are fast-transforming access to information and allow students to prepare assignments and offer quite accurate responses to a wide range of exam questions which are routinely used in assessments of students across the board including undergraduate dental students. This study aims to evaluate the performance of Chat Generative Pre-trained Transformer (ChatGPT), a generative AI-based application, on a wide range of assessments used in contemporary healthcare education and discusses the implications for undergraduate dental education.

Materials and Methods: This was an exploratory study investigating the accuracy of ChatGPT to attempt a range of recognised assessments in healthcare education curricula. A total of 50 independent items encompassing 50 different learning outcomes ($n=10$ per item) were developed by the research team. These included 10 separate items based on each of the five commonly used question formats including multiple-choice questions (MCQs); short-answer questions (SAQs); short essay questions (SEQs); single true/false questions; and fill in the blanks items. Chat GPT was used to attempt each of these 50 questions. In addition, ChatGPT was used to generate reflective reports based on multisource feedback; research methodology; and critical appraisal of the literature.

Results: ChatGPT application provided accurate responses to majority of knowledge-based assessments based on MCQs, SAQs, SEQs, true/false and fill in the blanks items. However, it was only able to answer text-based questions and did not allow processing of questions based on images. Responses generated to written assignments were also satisfactory apart from those for critical appraisal of literature. Word count was the key limitation observed in outputs generated by the free version of ChatGPT.

Conclusion: Notwithstanding their current limitations, generative AI-based applications have the potential to revolutionise virtual learning. Instead of treating it as a threat, healthcare educators need to adapt teaching and assessments in medical and dental education to the benefits of the learners while mitigating against dishonest use of AI-based technology.

KEYWORDS
artificial intelligence, ChatGPT, dental education, education technology, machine learning

1 | INTRODUCTION

Artificial intelligence (AI) is a transdisciplinary field, which involves the use of computer algorithms model intelligent behaviour with minimal human intervention and is informed by logic, statistics, cognitive psychology, linguistics, decision theory, neuroscience, cybernetics and computer engineering.[1] AI applications are primarily based on machine learning and use information retrieval, image and speech recognition, sensor technologies, robotic devices and cognitive decision support systems. AI is already creating a global impact and is fast transforming all spheres of modern life including industry, social media, healthcare, space technology, as well as a wide range of functions at the level of governments.[2–4] The ultimate aim of AI is to create machines which are capable to perform intellectual tasks like humans.[5]

Chat Generative Pre-trained Transformer (ChatGPT) is an open-source AI-based application freely available on the Internet (https://chat.openai.com/). Presently, it is one of the most advanced natural language processing model based on 175 billion parameters and is trained on Azure's AI supercomputer.[6] ChatGPT is a generative AI which is capable of creating new content during real-time conversations.[7] ChatGPT uses various AI models trained on massive amount of text data to respond to user queries.[8] It offers conversational responses to user queries. It has the ability to remember the user input into the conversation thread and its own response and builds on its previous outputs with subsequent queries.

Since its launch in November 2022 by open AI, ChatGPT has generated incredible excitement globally and continues to dominate tech media headlines due to its remarkable abilities.[9] Powered with a conversational interface, ChatGPT allows users to perform numerous text-based tasks such as answering questions on an unprecedented scale, generating codes, translations and generating bespoke content. Like all sectors, inevitably, ChatGPT will affect academia in multiple ways and healthcare education is no exception. One of the main concerns about ChatGPT in healthcare education is related to its ability to generate content and answer questions, which may potentially encourage dishonesty in academic work and assessments.

The aim of this study was to investigate how ChatGPT performs on common assessments used in contemporary healthcare education curricula.

2 | METHODS

2.1 | Study design

This was a web-based exploratory study investigating the accuracy of ChatGPT to attempt a range of recognised assessments in healthcare education curricula. The assessment items were contextualised to dental curricula. ChatGPT Feb 9 free version was used in this study.

2.2 | Development of assessment items

For knowledge-based assessments, a total of 50 independent items on 50 encompassing 50 different learning outcomes ($n = 10$ per item) using five common formats used in the assessment of students in healthcare curricula were developed by the research term. The format and number of questions in each format were as follows:

- Single best answer multiple-choice questions (MCQs) 10 items ($n = 10$)
- Short-answer questions (SAQs) 10 items ($n = 10$)
- Short essay questions (SEQs) 10 items ($n = 10$)
- Single true/false questions 10 items ($n = 10$)
- Fill in the blanks 10 items ($n = 10$).

Most items (over 80%) were based on clinical vignettes with a focus on the application of knowledge rather than mere factual recall. The items used for knowledge-based assessments were based on core clinical topic areas in restorative dentistry, periodontics, fixed prosthodontics, removable prosthodontics, endodontics, pedodontics, orthodontics, preventive dentistry, oral surgery and oral medicine. One assessment item was created for each of 10 aforementioned clinical subjects, using five different formats.

Assessments on reflective reports based on multisource feedback (MSF) portfolio assignments, research methodology and critical appraisal of the literature were also prepared.

2.3 | Quality assurance

The knowledge-based assessment items ($n = 50$) were reviewed by a panel of five experienced clinical academics and checked for accuracy, clarity of language, relevance and agreement on correct answers. No differences among assessors were observed for MCQs, SEQs, fill in the blanks and single true/false items. All assessment items were blueprinted against the learning outcomes for undergraduate dental students, as identified by the General Dental Council (UK) and benchmarked against the knowledge level expected from newly qualified dental graduates to ensure the items were within the scope of undergraduate dental education.[10]

For written assessments (SAQs, MSF reflective reports, research methodology and critical appraisal of literature), assessments rubrics and model answer keys were developed and reviewed by two blinded external reviewers to evaluate their face validity, accuracy and suitability for undergraduate dental students. Minor differences in model answers were encountered for SAQs which were resolved through deliberations among the assessors.

2.4 | Administration of assessments

A user account was created on ChatGPT website and response to each assessment item was generated in turn. A log of every query regarding individual assessment items and response(s) generated by ChatGPT was automatically recorded on the account dashboard. All items and responses by ChatGPT were exported from the website for assessment.

2.5 | Scoring of assessments

Assessment of ChatGPT responses to MCQs, SAQs, single true/false and fill in the blank items was done using dichotomous scoring supported by answer keys for individual items. Correct responses were awarded one mark while incorrect response received a zero score.

Each of the five SEQs and written assignments, that is, MSF reflective reports, research methodology and critical appraisal of literature were marked by two assessors independently using a grading rubric. The following grade boundaries were used for assessing SEQs and written reports:

- Unsatisfactory 0%–49%
- Borderline 50%–59%
- Satisfactory 60%–69%
- Excellent: 70% or more

Inter-rater reliability was evaluated with intraclass correlation coefficient (ICC) in Microsoft Excel [Version 17]. Redmond, WA: Microsoft Corporation. ICC values above 0.61 were considered to be indicative of good agreement between assessors. The ICC values were good for all written assessments including SEQs (0.78); MSF reflective reports (0.81); research methodology (0.71); and critical appraisal of literature (0.70) assessments. Differences in assessor scores were moderated further in the post assessment meeting. Average scores were awarded for scripts with residual differences in scores by individual assessors.

3 | RESULTS

ChatGPT performed above satisfactory level on all types of dental assessments included in this study except critical appraisal of literature. Exemplars of each type of assessment items and responses by ChatGPT are included in the Appendix S1. The mean score of ChatGPT was in the range of 70%–100%. The highest score was recorded for single true/false items (100%) while the lowest mean score of 70% was observed for SEQs. The performance of ChatGPT on other assessment items, that is, MCQs, SEQs and fill in the blank items yielded a mean score of 90%. The performance of ChatGPT on knowledge-based assessments including the number of items and the scoring system used for each type of assessment are summarised in Table 1.

The main limitation observed for ChatGPT performance on these items was that is only able to answer text-based questions and did not allow processing of questions based on images. The main limitation in responses to SEQs was limited details of clinical interventions and follow-up visits which resulted in low scores in comparison to other question formats. Nevertheless, no factually inaccurate information was identified in the responses to any of the SEQ items.

Performance of ChatGPT on MSF assignments, research methodology and critical appraisal of literature is summarised in Table 2 and exemplars are provided in the Appendix S1.

MSF reflective reports generated by ChatGPT and all five reports received an excellent grade. Each of the five research methodology reports generated by ChatGPT received a satisfactory grade. However, deficiencies were noted in sample size calculations, and assessment of outcome measures was noted consistently for all five attempts. Finally, the lowest grade of performance was observed for critical appraisal of literature and all reports generated by ChatGPT received a borderline grade. The key limitations observed consistently for all attempts by the ChatGPT included a low word count (upper limit of 650 words). As explained in the discussion section, this limitation can be addressed by the latest version of ChatGPT. Moreover, the references initially cited by ChatGPT were more than 5 years old, and it missed some key studies based on RCTs. However, further improvements in the quality of critical appraisal and references could be achieved through a conversation with the ChatGPT as shown in the Appendix S1.

TABLE 1 Chat Generative Pre-trained Transformer (ChatGPT) performance on knowledge-based assessments.

Assessment	Number of assessment items (*N*)	Scoring system	Mean percentage score ChatGPT (%)
1. Multiple-choice questions (MCQs)	10	Dichotomous	90
2. Fill in the blanks items	10	Dichotomous	90
3. Short-answer questions (SAQs)	10	Dichotomous	90
4. Single true/false questions	10	Dichotomous	100
5. Short essay questions (SEQs)	10	Grading rubric	70

TABLE 2 Chat Generative Pre-trained Transformer (ChatGPT) performance on written assignments.

Assignments	Number of assessment items (*N*)	Scoring system	Average ChatGPT grade
1. Reflective portfolio report	5	Grading rubric	Excellent (70% or more)
2. Research methodology	5	Grading rubric	Satisfactory (60%–69%)
3. Critical appraisal of literature	5	Grading rubric	Borderline (50%–59%)

4 | DISCUSSION

This study is a first in investigating the impact of generative AI represented by ChatGPT on commonly used assessments in dental education. Our results demonstrate the capabilities of ChatGPT to attempt dental assessments and achieve acceptable grades. Our results corroborate with few recent studies which showed that ChatGPT was able to perform at or near the passing threshold of all three parts of on *United States Medical Licensing Examination®* without any additional training or reinforcement.[11] The findings underscore the need for dental education providers to recognise the impact of rapid technological advances on dental education. It warrants strategies to mitigate against inappropriate use of technology while ensuring that students and faculty are able to benefit from the technology. This is not the first time that education providers are confronted by the challenges posed by innovative technologies. The current generation of academics has already witnessed the internet revolution. Access to information has been transformed by powerful search engines such as Google, and web-based applications like YouTube, as well as use of digital flashcards.[12] Unlike the pre-millennium era, teachers and textbooks are no longer the exclusive source of information for students.

One of the main strengths of ChatGPT over traditional web search engines is that ChatGPT offers a conversation style interactive platform for the users and provides a direct response to queries instead of signposting the user to multiple websites. Also, the users are able to engage with ChatGPT to dissect the information and question its authenticity, and sources. The utility of ChatGPT is akin to a virtual tutor, which may be accessed round the clock free of cost and it is likely there will be increased use of this tool in higher education including medical and science education.

ChatGPT offers an intelligent learning platform that enables students' learning to be scaffolded, offering the capability to adapt and personalise learning content according to individual needs. Use of ChatGPT as a learning platform appears to be a suitable option and does not raise any concerns. Inaccurate information is always a risk with any web search and is also applicable to ChatGPT warranting the user to cross check information when in doubt.

ChatGPT and similar platforms which may be rolled out in the future are likely to evolve further with further human input from technical experts as well as users. Following its initial launch, ChatGPT is undergoing rapid developments, and recently, Open AI announced the launch of ChatGPT4, the latest version in its line of AI language models with enhanced capabilities such as image processing and processing of up to 25000 words of text.[13] The new model requires a monthly subscription (20 US dollars per month at present) and is available to the general public via ChatGPT Plus. These developments reflect the remarkable pace at which AI language models are developing and will ultimately transform into multi-modal systems that integrate text, images, audios and videos.

Assessments are a critical part of dental education and inform decisions regarding student progression during successive stages of a dental programme.[14] Quality assurance of assessments is essential for institutional reputation and public confidence. Education providers need to ensure that the assessment content is maintained securely and student submissions represent original work. The main limitation of the reflective portfolio reports generated by ChatGPT was that there was little reference to specific learning activities or events. However, it may be argued that if a student is prepared to use ChatGPT to prepare their assignment works, the output can be tweaked to address such limitations. Similarly, ChatGPT may be able to provide a solid foundation for other assignments such as research methodology and critical appraisal and students may be able to refine their assignments to address any limitations with or without further help from ChatGPT.

For academic assignments submitted by the students, an increasing number of institutions conduct plagiarism checks using appropriate software applications such as iThenticate (Turnitin LLC).[15] However, plagiarism check software is primarily aimed at quantifying the similarity index with published works available online and identifying the source. Given that, ChatGPT is capable of generating new text, routine software applications used for checking plagiarism may not be dependable for identifying outputs by ChatGPT.

Detecting text generated by AI is an active research area in the field of AI. One approach involves using machine learning models to differentiate between AI-written and human-written text.[16] An increasingly number of AI-based tools are being developed to address this problem such as AI Text Classifier by OpenAI, such as DetectGPT and GPTZero.[17] However, these tools are not always accurate and misclassification can happen. For instance, the outputs generated by ChatGPT were subjected to scrutiny using these tools, but the results were inconclusive. Outputs generated by ChatGPT were subjected to scrutiny using ChatGPT detector software but the results were inconclusive. Moreover, with further expansion and sophistication of language processing, such detection may become even for difficult in the future.[18] Therefore, further research is required to enhance the precision of these tools.

Knowledge-based assessments are a core component of undergraduate dental curricula and provide serve to demonstrate that dental graduates have the underpinning scientific knowledge to inform their clinical practice.[19–21] Historically, knowledge-based assessments in dental education have been delivered face to face in university settings using pen and paper. Although this mode of assessment delivery remains the most common, web-based digital assessment platforms are gaining popularity and allow assessments to be administered online as well as offline. Digital assessment platforms allow secure storage of assessment content in addition to designing, blueprinting, audit and psychometric analyses of assessments. Following the COVID-19 pandemic, remote proctoring has been used to deliver online assessments.[12] Although face-to-face assessments have returned, remote assessments offer some advantages to dental institutions especially if there are resource constraints to provide space and IT equipment for large groups of students sitting an assessment.

ChatGPT does not pose any threats to knowledge-based assessments delivered face to face on university campus under direct invigilation. During COVID-19, remote online delivery of assessments were undertaken on a large scale by dental institutions.[12,22] However, only few institutions had the resources to use commercially available platforms designed to proctor students appropriately.[23] Most dental institutions, particularly in developing countries, use open-source online platforms such as Zoom and Webex to deliver assessments remotely which allowed students to be observed on camera during the assessments. However, these platforms did not permit restricting internet access to assessment content for candidates during assessments. This was mitigated, in part, by creating 'non-searchable' questions so that students could not get a quick answer by searching the internet. However, with the availability of bots like ChatGPT, this approach may not be appropriate if remote assessments are used without proctoring.

The educational value of ChatGPT is promising and there its use in dental education can provide a personalised learning experience to support the varying learning needs of dental students. Face-to-face invigilated assessments are unlikely to be impacted directly by ChatGPT and do not warrant any modification at this stage. However, there is risk of dishonesty in academic assignments which are completed by students off-campus and dental educators need to develop appropriate policies to mitigate against such risks.

A few limitations of this study need to be acknowledged. First, the total number of knowledge-based items was 50, with 10 items each belonging to five different question formats. Our aim was to evaluate ChatGPT on new questions rather than using questions which might be accessible to ChatGPT from existing online resources. Moreover, we did not use questions from our institutional bank as it would have made them openly accessible and rendered them unusable for future assessments. Developing quality questions is time-consuming and it was not practical to produce additional items. It is recognised that additional questions would have been helpful to explore the capabilities of ChatGPT. Second, assessment of the difficulty level of items requires separate measures such as Angoff/Hofstee methods or more accurately using one of item response theory models such as Rasch analysis.[24,25] While this would have been informative, quantification of ChatGPT capabilities in the context of item difficulty was not undertaken and would merit an additional study. In any case, our aim was to demonstrate that ChatGPT is able to answer questions based on a range of formats which are used widely in the assessment of undergraduate dental students. Finally, we only used the open version of ChatGPT and it would be interesting to explore the capabilities of the paid version to process questions based on dental images.

Open access to ChatGPT is a recent phenomenon and represents a learning curve for dental educators. A quick fix approach is unlikely to be the correct way forward and dental educators need to deliberate, identify and implement measures to address challenges posed by ChatGPT. Direct invigilation of written assignments by moving remote assessment to campus appears to be a possible solution. However, it may not be appropriate to complete these assignments within a timed session. Another option may be to change the format of written assignments to oral presentations followed by questions from assessors.[26] However, it would be vital to consider resources and faculty time required to assess substantial number of students. Inevitably, dental education will adapt modern technology as it has done in the past and it is important to avoid knee jerk reactions. Initially, students may be asked to sign a declaration to confirm that all works submitted for summative assessments reflect student's original work and were not generated using AI. Engagement with all stakeholders in dental education including the students to find sustainable solutions appears to be an appropriate strategy. It is imperative that dental educators share their institutional approaches and experiences with colleagues to identify best practices which can be adopted more widely.

5 | CONCLUSION

ChatGPT is a double-edged sword and while it can be helpful for both students and teachers alike, it can be used to generate assignments and answer assessment questions, which raises concerns regarding potential cheating and dishonesty in academic works. Notwithstanding their current limitations, generative AI applications have the potential to revolutionise virtual learning. Instead of treating it as a threat, dental educators need to adapt teaching and assessments in dental education to the benefits of the learners while mitigating against dishonest use of generative AI applications.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the ChatGPT for generating responses to a wide range of assessment items. The author queries and responses by ChatGPT are included in the Appendix S1.

CONFLICT OF INTEREST STATEMENT

The authors have no conflicts of interest to declare.

DATA AVAILABILITY STATEMENT

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

1. Howard J. Artificial intelligence: implications for the future of work. *Am J Ind Med*. 2019;62:917-926.
2. King MR. The future of AI in medicine: a perspective from a Chatbot. *Ann Biomed Eng*. 2023;51:291-295.
3. Girimonte D, Izzo D. Artificial intelligence for space applications. In: Schuster AJ, ed. *Intelligent Computing Everywhere*. Springer; 2007:235-253.
4. Sharma GD, Yadav A, Chopra R. Artificial intelligence and effective governance: a review, critique and research agenda. *Sustain Futures*. 2020;2:100004.
5. Grace K, Salvatier J, Dafoe A, Zhang B, Evans O. Viewpoint: when will ai exceed human performance? Evidence from ai experts. *J Artif Intell Res*. 2018;62:729-754.
6. *Microsoft teams up with OpenAI to exclusively license GPT-3 language model—the official microsoft blog*. 2023. https://blogs.microsoft.com/blog/2020/09/22/microsoft-teams-up-with-openai-to-exclusively-license-gpt-3-language-model/
7. Qadir J. *Engineering education in the era of ChatGPT: promise and pitfalls of generative AI for education*. 2022. https://www.techrxiv.org/articles/preprint/Engineering_Education_in_the_Era_of_ChatGPT_Promise_and_Pitfalls_of_Generative_AI_for_Education/21789434
8. *Model index for researchers—OpenAI API*. 2023. https://platform.openai.com/docs/model-index-for-researchers
9. *OpenAI: everything to know about the company behind ChatGPT*. 2023. https://www.augustman.com/sg/gear/tech/openai-what-to-know-about-the-company-behind-chatgpt/
10. Innes N, Hurst D. GDC learning outcomes for the undergraduate dental curriculum. *Evid Based Dent*. 2012;13:2-3.
11. Aidan G, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States medical licensing examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312.
12. Ali K, Alhaija ESA, Raja M, et al. Blended learning in undergraduate dental education: a global pilot study. *Med Educ Online*. 2023;28(1):2171700. doi:10.1080/10872981.2023.2171700
13. *GPT-4*. 2023. https://openai.com/product/gpt-4
14. Patel US, Tonni I, Gadbury-Amyot C, van der Vleuten CPM, Escudier M. Assessment in a global context: an international perspective on dental education. *Eur J Dent Educ*. 2018;22:21-27.
15. Lapeña JFF. A dozen years, a dozen roses. *PJOHNS*. 2018;33(2):4-5.
16. Mitchell E, Lee Y, Khazatsky A, Manning CD, Finn C. *DetectGPT: zero-shot machine-generated text detection using probability curvature*. 2023. https://arxiv.org/abs/2301.11305
17. *GPTZero*. 2023. https://gptzero.me/
18. *How to detect AI-generated text, according to researchers—WIRED*. 2023. https://www.wired.com/story/how-to-spot-generative-ai-text-chatgpt/
19. Ali K, Jerreat M, Zahra D, Tredwin C. Correlations between final-year dental students' performance on knowledge-based and clinical examinations. *J Dent Educ*. 2017;81(12):1444-1450.
20. Zahra D, Bennett J, Belfield L, et al. Effect of constant versus variable small-group facilitators on students' basic science knowledge in an enquiry-based dental curriculum. *Eur J Dent Educ*. 2019;23(4):448-454.
21. Ali K, Cockerill J, Bennett JH, Belfield L, Tredwin C. Transfer of basic science knowledge in a problem-based learning curriculum. *Eur J Dent Educ*. 2020;24(3):542-547.
22. Jaap A, Dewar A, Duncan C, Fairhurst K, Hope D, Kluth D. Effect of remote online exam delivery on student experience and performance in applied knowledge tests. *BMC Med Educ*. 2021;21(1):86.
23. Ali K, Barhom N, Duggal MS. Online assessment platforms: what is on offer?. *Eur J Dent Educ*. 2023;27(2):320-324. doi:10.1111/eje.12807
24. Ali K, Slade A, Kay EJ, Zahra D, Chatterjee A, Tredwin C. Application of Rasch analysis in the development and psychometric evaluation of dental undergraduates preparedness assessment scale. *Eur J Dent Educ*. 2017;21(4):e135-e141.
25. Ali K, Coombes L, Kay E, et al. Progress testing in undergraduate dental education: the peninsula experience and future opportunities. *Eur J Dent Educ*. 2016;20(3):129-134.
26. *ChatGPT raises uncomfortable questions about teaching and classroom learning—the straits times*. 2023. https://www.straitstimes.com/opinion/need-to-review-literacy-assessment-in-the-age-of-chatgpt

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Ali K, Barhom N, Tamimi F, Duggal M. ChatGPT—A double-edged sword for healthcare education? Implications for assessments of dental students. *Eur J Dent Educ*. 2023;00:1-6. doi:10.1111/eje.12937