SPECIAL COMMUNICATION
Volume 16 Issue 1 2024

DOI: 10.21315/eimj2024.16.1.13

ARTICLE INFO
Received: 09-05-2023
Accepted: 04-08-2023
Online: 29-03-2024

From Chatbots to Mind-bots: Navigating Assessments in the Advent of ChatGPT

Mohd Salami Ibrahim, Nurulhuda Mat Hassan

Department of Family Medicine, Faculty of Medicine, Universiti Sultan Zainal Abidin, Terengganu, MALAYSIA

To cite this article: Ibrahim MS, Mat Hassan N. From chatbots to mind-bots: navigating assessments in the advent of ChatGPT. Education in Medicine Journal. 2024;16(1):205–211. https://doi.org/10.21315/eimj2024.16.1.13

To link to this article: https://doi.org/10.21315/eimj2024.16.1.13

ABSTRACT

Recently, the authors had a situation in a mobile chat group (WhatsApp) when a medical lecturer was asking what value a human papillomavirus (HPV) test adds to the pap smear for cervical cancer screening. The field experts in the group replied with detailed, erudite explanations. Soon after, another lecturer posted an answer from GPT-3 based chatbot. Despite needing more depth of an expert's reply, the chatbot gave concise answers, reframing complex medical jargon in plain English without losing crucial medical information, and more. They were easier to understand. All these with the leisure of a human-like engagement. This narration is one of countless news related to ChatGPT, which have been making headlines, academic journals included, to illustrate how the large language model technology may have disrupted conventional educational practice. One discriminatory element distinguishes this technology from all its predecessors; it is not trying to mimic a human response but responding like a human. In this writing, we navigate discussion based on the most fundamental aspect of assessment, its purpose. We revisit the concept of fidelity from the field of simulation to explain how the technology may have rejuvenated the purpose of assessment for learning (formative assessment). Then, we articulate several associated challenges in the conduct of the high-stake assessment of learning (summative assessment). We conclude with an emphasis on the purposes as the guiding principles that remain the same despite the changes in the landscape on the conduct of the assessment.

Keywords: *ChatGPT, Assessment, Formative assessment, Summative assessment, Medical education*

CORRESPONDING AUTHOR

Nurulhuda Mat Hassan, Department of Family Medicine, Faculty of Medicine, Universiti Sultan Zainal Abidin, Jalan Sultan Mahmud, 20400 Kuala Terengganu, Terengganu, Malaysia

Email: nurulhudamh@unisza.edu.my

INTRODUCTION

A few months ago, a colleague of ours, who is a medical lecturer, was enquiring in a WhatsApp (Meta Inc., CA, US) social group about the rationale behind adding human papillomavirus (HPV) to pap smear for cervical cancer screening. A senior gynaecologist responded aptly with a detailed explanation. Soon after, another lecturer shared the following conversation with a GPT-3 based chatbot mobile app (ChatGod version 1.0.11 on the Android 13 platform, accessed on 21 February 2023).

© Malaysian Association of Education in Medicine and Health Sciences and Penerbit Universiti Sains Malaysia. 2024
This work is licensed under the terms of the Creative Commons Attribution (CC BY)
(http://creativecommons.org/licenses/by/4.0/).

> **Question:** How HPV testing change the way we usually do screening for cervical cancer with pap smear?
>
> **Chatbot:** HPV testing has been shown to be more accurate than Pap testing alone in detecting cervical cancer. HPV testing is also more sensitive in detecting precancerous lesions as it looks for the presence of HPV, the virus known to cause cervical cancer. HPV testing in combination with Pap smear screening is now recommended for women over the age of 30, as it is more effective in detecting cervical cancer at an earlier stage. However, it does have some limitations. HPV testing can only detect the presence of the virus, which does not always lead to cervical cancer. Additionally, it does not always accurately predict which women will go on to develop cervical cancer. Furthermore, it has higher false-positive results than other screening methods such as Pap smears.
>
> **Question:** What then are the positive predictive value and the negative predictive value of an HPV test?
>
> **Chatbot:** The positive predictive value of HPV testing to screen for cervical cancer is approximately 14%–39%, and the negative predictive value is approximately 94%–97%.

The replies lacked the intimate knowledge of an expert who linked the complex pathophysiological process of insults to healthy cells to architect meaningful pathways to guide clinical decision-making for cervical cancer screening. However, members of the WhatsApp group appreciated the easy-to-understand answers for the use of plain language, concise and, more importantly, it did feel like talking to a human. Consequently, even though the natural language processing (NLP) technology itself is old, the excitement of what it can do now is new.

The excitement rewinds to November 2022 when ChatGPT, a GPT-3.5 NLP model, was unveiled to the public (1). The revelation has metamorphosed the previously saturated phenomenon in machine learning into a new global sensation. When ChatGPT scored beyond the passing threshold for all Step 1, Step 2CK, and Step 3 of the United States Medical Licensing Exam (USMLE) (2, 3), it raises a key question; has the debut initiated a butterfly effect that disrupts the conventional understanding and practice of assessment?

In this special communication, we discuss how medical and higher education assessment may have sailed into uncharted territory due to this technology's redefined use and potential misuse. To ensure that the boat is still navigating in the right direction, we will engage from the most fundamental aspect of assessment; its purpose of driving learning (formative assessment) and substantiating learning (summative assessment).

FORMATIVE ASSESSMENT

The Concept of Simulation Fidelity

We believe understanding simulation fidelity is one of the vital educational philosophies to dissect this issue. Simulation fidelity is commonly misunderstood as a measure of how advanced the technology is being used to create realistic simulations. A more robust approach perhaps can begin by considering fidelity from an educational context rather than from the view of simulation per se. The term denotes the approximation of the transfer of learning that is largely dependent on the "degree of faithfulness" between simulation and real practice (4). Illustrating this concept, an old but interesting experiment by De Groot (5) showed that a grandmaster could recall their location better than average players whenever chess pieces were set in a game-play setting. However, this finding did not occur when the chess pieces were set randomly. This experiment demonstrated that the human ability to make associations is a function of exposure to specific patterns of information rather than any perceptual ability (4). Furthermore, when Proteau et al. (6) conducted an arm-movement experiment, participants who were trained with visual input needed the same acuity to reproduce their skills with fewer errors. In contrast, their counterparts who were trained in a dark room could achieve similar accuracy without any visual cue (6). Therefore, fidelity can be understood as the notion of the "principle of information processing specificity" where the transfer of learning is proportionately enhanced by the degree of overlap of learning context between simulation and real practice (4).

This understanding is crucial because the volume of meaningful learning from simulation training does not depend on how real the simulation is but rather the degree of overlap between what needs to be learned and how simulation fulfils this purpose. For example, for urological stones extraction training, a randomised controlled trial by Matsumoto et al. (7) showed there was no statistically significant difference between medical students who were trained using a model made of a plastic cup and drinking straws, which cost CAD20 compared with those who were trained with an advanced manikin which carried a CAD3,700 price tag. Similarly, a mannequin head can be regarded as a low-fidelity simulation for endotracheal intubation, but it is high-fidelity for practising nasogastric tube insertion (8). Likewise, early laparoscopic simulators confer high-fidelity simulation of the laparoscopic visual field. However, its fidelity is poor for suturing skills, which rely more on haptic feedback to "feel" the needle against the tissue (9). Understanding the concept of fidelity is fundamental because the concept underpins the design for mastery of learning via simulation.

ChatGPT as a Screen-based Simulation

Consequently, it is useful to appreciate ChatGPT as a screen-based simulation to understand how it may invigorate formative assessment for the mastery of learning. Unlike the typical search engine or any other chatbots before it, the example given in the introduction above illustrates how the intuitively engaging, appealing, and meaningful conversation managed to consolidate learning because of a simple but meaningful reply. The simulation-based learning has never been more accelerated at this scale, especially considering the new height of fidelity is coupled with the brainpower of a chatbot trained by big data of the internet. ChatGPT has written abstracts for academic journals to deceive even scientists in the field. An expert in the study of technology and regulation was quoted (10):

> …we're now in a situation where the experts are not able to determine what's true or not.

As a result, any students in the world with internet access, regardless of their levels and fields of study, are now having access to a simulation technology that manifests as a pseudo-teacher who can explain almost anything, attach emotion and insights to its reply, share light humour to cheer the mood up, and will never get tired from addressing any enquiries.

Implementing Formative Assessment: The Value of Knowing the Right Questions

Paradoxically, the heightened fidelity and wealth of information may have mandated more emphasis on the roles of teachers to embrace formative assessment. The roles of teachers for ChatGPT-based learning via formative assessment may summon two key crucial advantages.

Firstly, students will know the right questions for their learning. Having access to vast information and high-fidelity teaching simulation does not immediately translate into meaningful learning. With formative assessments, teachers' roles are crucial via framing questions as the instructional tool to challenge assumptions, expose contradictions, and lead to new knowledge and wisdom (11). These questions equip the teachers with tools to guide students from being overwhelmed by excessive information due to the extensive abilities of the chatbot, which may be redundant, unnecessary, or distracting to the trajectory of students' progress. Thus, guiding students with the right questions via formative assessments is a sine-qua-non for structured, systematic, relevant, and progressive efforts toward the mastery of learning.

Secondly, since the chatbot may provide high accuracy and human-like responses at any time and any place, students may learn much faster via deliberative practice. The theory of deliberative practice posits that expert performance is a monotonic function of the accumulation of practice that leads to the improvement or maintenance of skill (12). The centre of this theory is persistent activities and assessments that expose gaps in knowledge and skills, so quality feedback can be optimised to correct errors and improve performance (12). We do not subscribe to the idea that relies on ChatGPT as the main and sole source of feedback because of potential issues with the validity of the feedback. However, we propose that the use of ChatGPT in this context is primarily useful as the initial retrieval of answers and feedback so students may then be enriched with some preliminary knowledge, ideas, and recommendations, which can subsequently be substantiated with credible resources. Thus, the latter signified the crucial roles of teachers as the trustworthy source of information who may deliver not only direct feedback but also values and professionalism via role modelling.

IS CHATGPT A THREAT TO SUMMATIVE ASSESSMENT?

The question of credible information and reliable learning subsequently leads to the fundamental question of ChatGPT and its impacts on summative assessments. Since ChatGPT has demonstrated competent responses to the challenges in clinical practices and beyond (13), there are concerns about the possibility that students may unethically capitalise on the technology in their favour (14). Will this be a problem for medical schools to validate students' learning assessments?

Some conducts of summative assessments may be at high-risk more than others. High-stake examinations, which are mostly conducted face-to-face and directly invigilated, may secure a degree of reassurance. However, credit-bearing continuous examinations such as essays may face serious challenges. These challenges are not new. Previously, options did exist for unethical conduct on high-stakes assessments. For example, a student can use paper mills such as Killer Papers, a company which has earned a seven-figure monthly revenue for its essay-writing services (15). Meanwhile, cheaper alternatives include utilising search engines such as Google to find samples of essays related to the topic and using artificial intelligence (AI)-based editing software to escape issues related to plagiarism (16, 17). The advent of ChatGPT adds a new option, albeit a better one, as it is currently free, articulating high-fidelity responses that can even deceive the experts in the field and being better to escape issues related to plagiarism (18, 19). Abstracts generated by ChatGPT even have a median originality score of 100% (20). Therefore, concerns related to ChatGPT are more directed at the conducts rather than the summative assessment itself, with the risks appearing to be heavily weighed on digital-based rather than conventional-based conduct of summative assessments.

As a result, the arrival of ChatGPT has catapulted the priority on increasing the trustworthiness and self-credibility of the students by educators. Awareness of the importance and benefits of undergoing the process of writing the assignments themselves should be instilled in the students, as well as the values of integrity and honesty. The process of thinking, analysing, and effectively expressing own self while executing the assignment should serve as the foundation of professional education. The students need to be aware that copying the answers of their assignments from AI, such as ChatGPT, is a form of academic dishonesty that would mask the catastrophic outcomes in the longer run. Instilling these values is crucial to be done as early as possible during medical education so that even in the face of the most advanced AI, the students can choose to do the hard work rather than take the easy way out. Hence, although ChatGPT may have accelerated concerns related to the integrity of certain forms of summative assessments, we believe it ultimately highlights the need for a stronger emphasis on developing critical thinking and academic integrity among students, which will benefit them in their future professional endeavours.

Consequently, instead of shying away from ChatGPT, educators may adopt a contrasting strategy to “befriending” and getting to know ChatGPT and other AIs to equip themselves with operations of the language AI chatbots. These skills may prove useful for educators to assume the roles of competent examiners. In a paper describing ChatGPT passing the law school examinations, the authors noted that despite passing the examinations, two out of three examiners managed to identify the papers written by ChatGPT (21). The papers written by the chatbot were said to be repetitive and had uncannily perfect grammatical structures. Furthermore, some have revealed that ChatGPT was unable to match critical thinking skills questions to satisfaction (22). Others have also noted that maths questions were the weakness of ChatGPT (21). Therefore, educators must be more creative in designing critical examinations and questions to reflect desired outcomes.

As evidenced by the conversation earlier in the paper, the answers of ChatGPT contained the facts, but not critical-thinking-based answers whereby linking the basic knowledge of the pathophysiology of cervical cancer leads to the decision on the mode of screening. Suppose medical schools were to only test on the factual knowledge of diseases. In that case, it invariably denies the nature of medicine, in which contextual application of knowledge is the crucial process of decision-making. Thus, as Mbakwe et al. (13) highlighted in their paper, the success of ChatGPT in the USMLE should prompt teachers of medicine to rethink whether we are assessing future medical graduates on the actual skills needed to become good doctors. The authors also highlighted the frightening possibility of ChatGPT passing based on the biased information available to it from the internet. Given the seemingly original output of ChatGPT and other AI results from autonomous learning of the vast pool of online information, this is a given. It also points out the possibility of overreliance on factual-based questions in assessing medical students in USMLE. Thus, this evidence signifies the mandate of redirection in assessing medical students and towards the more complex cognitive skills such as problem-solving and critical thinking.

CONCLUSION

To conclude, the emergence of new large language model technology such as ChatGPT has the potential to revolutionise the way assessments are conducted, with a focus on enhancing learning and substantiating knowledge. Through the lens of simulation and deliberative practice, we have emphasised the importance of providing students with appropriate guidance, questions, and feedback in conjunction with the benefits of ChatGPT's high-fidelity simulations for effective formative assessments. Furthermore, we have highlighted the challenges educators face in developing the necessary skills to harness this technology for secure and reliable summative assessments. Although this technology has transformed the landscape of assessment, the purpose remains the same: to facilitate learning and measure knowledge effectively.

ACKNOWLEDGEMENTS

The authors would like to thank all the lecturers from the Faculty of Medicine, Universiti Sultan Zainal Abidin for their support and initiating the discussion on ChatGPT.

REFERENCES

1. Gordijn B, Have HT. ChatGPT: evolution or revolution? Med Health Care Philos. 2023;26(1):1–2. https://doi.org/10.1007/s11019-023-10136-0
2. 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 Digit Health. 2023;2(2):e0000198. https://doi.org/10.1371/journal.pdig.0000198
3. Biswas S. Passing is great: can ChatGPT conduct USMLE exams? Ann Biomed Eng. 2023; 51(9):1885–6. https://doi.org/10.1007/s10439-023-03224-y
4. Grierson LEM. Information processing, specificity of practice, and the transfer of learning: considerations for reconsidering fidelity. Adv in Health Sci Educ. 2014;19(2):281–9. https://doi.org/10.1007/s10459-014-9504-x
5. De Groot AD. Thought and choice in chess. 2nd ed. The Hague: Mouton Publishers; 1978.
6. Proteau L, Marteniuk RG, Girouard Y, Dugas C. On the type of information used to control and learn an aiming movement after moderate and extensive training. Hum Mov Sci. 1987;6(2):181–99. https://doi.org/10.1016/0167-9457(87)90011-X
7. Matsumoto ED, Hamstra SJ, Radomski SB, Cusimano MD. The effect of bench model fidelity on endourological skills: a randomised controlled study. J Urol. 2002;167(3):1243–7. https://doi.org/10.1016/S0022-5347(05)65274-3
8. Hamstra SJ, Brydges R, Hatala R, Zendejas B, Cook DA. Reconsidering fidelity in simulation-based training. Acad Med. 2014;89(3):387–92. https://doi.org/10.1097/ACM.0000000000000130
9. Norman G, Dore K, Grierson L. The minimal relationship between simulation fidelity and transfer of learning. Med Educ. 2012;46(7):636–47. https://doi.org/10.1111/j.1365-2923.2012.04243.x
10. Else H. Abstracts written by ChatGPT fool scientists. Nature. 2023;613:423. https://doi.org/10.1038/d41586-023-00056-7

11. Corley MA, Rauscher WC. TEAL Fact Sheet No. 12: deeper learning through questioning. Arlington, VA: American Institutes for Research (AIR); 2013.
12. Campitelli G, Gobet F. Deliberate practice: necessary but not sufficient. Curr Dir Psychol Sci. 2011;20(5):280–5. https://doi.org/10.1177/0963721411421922
13. Mbakwe AB, Lourentzou I, Celi LA, Mechanic OJ, Dagan A. ChatGPT passing USMLE shines a spotlight on the flaws of medical education. PLOS Digit Health. 2023;2(2):e0000205. https://doi.org/10.1371/journal.pdig.0000205
14. Cotton D, Cotton P, Shipway JR. Chatting and cheating: ensuring academic integrity in the era of ChatGPT. Innov Educ Teach Int. 2023. https://doi.org/10.1080/14703297.2023.2190148
15. Stokel-Walker C. Inside a highly lucrative, ethically questionable essay-writing service; 2022 [cited 2023 Mar 26]. Available from: https://www.inverse.com/input/culture/killer-papers-essay-mill-academics-writing-service
16. Naznean A. Cheating during online examinations – literature review. J Pedagogy. 2021;2:7–21. https://doi.org/10.26755/RevPed/2021.2/7
17. Masic I, Jankovic SM, Kurjak A, Donev DM, Zildzic M, Sinanovic O, et al. Guidelines for editing biomedical journals: recommended by Academy of Medical Sciences of Bosnia and Herzegovina. Acta Inform Med. 2020;28(4):232–6. https://doi.org/10.5455/aim.2020.28.232-236
18. Dwivedi YK, Kshetri N, Hughes L, Slade EL, Jeyaraj A, Kar AK, et al. “So what if ChatGPT wrote it?” multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. Int J Inf Manag. 2023;71:102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642
19. Khalil M, Er E. Will ChatGPT get you caught? rethinking of plagiarism detection; 2023 [cited 2023 Mar 17]. https://doi.org/10.35542/osf.io/fnh48
20. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. NPJ Digit Med. 2023;6(1):75 https://doi.org/10.1038/s41746-023-00819-6
21. Choi JH, Hickman KE, Monahan A, Schwarcz, D. ChatGPT goes to law school. Minnesota legal studies research paper no 23-03; 2023 [cited 2023 Mar 26]. https://doi.org/10.2139/ssrn.4335905
22. Stokel-Walker C. AI bot ChatGPT writes smart essays — should professors worry? Nature. 2022. https://doi.org/10.1038/d41586-022-04397-7