DOI: https://doi.org/10.4038/seajme.v18i2.584 **Short Communication**

Comparing the Effectiveness of ChatGPT3.5 and Bing GPT4 as Supplementary Resources to Enhance the Teaching, Learning, and Assessment of Human Anatomy

Wickramarathna, A.M., Kumara, S.S., Buddhasinghe R.U., Wijayasekara, A.E.

Abstract

Introduction: Artificial intelligence (AI) has shown promise in revolutionizing healthcare education by providing efficient and beneficial learning opportunities for students and reducing the workload for educators.

Materials and methods: This study was conducted from September 2023 to October 2023 assessing the accuracy, relevance, comprehensiveness, and user-friendliness of responses given by ChatGPT3.5 and Bing GPT4 on human anatomy.

Results: Both AI tools were able to provide detailed descriptions of gross anatomy, human embryology, and histology. While both showed a need for improvement in generating sample questions, Traditional methods still offer superior accuracy and relevance compared to AI Chatbots. But Chatbots provide personalized and convenient responses.

Conclusion: As AI technology continues to evolve, caution should be exercised in using AI tools for anatomy education, as their capabilities and performance may change over time.

Keywords: Artificial intelligence, Medical education, Anatomy

Introduction

Artificial intelligence (AI); the modeling of intelligent behavior by a computer with no or minimal human involvement has the potential to revolutionize healthcare education by providing efficient and beneficial learning opportunities for students and reducing the workload for educators (Hamet and Tremblay, 2018; Kulkarni et al., 2020; Grunhut et al., 2021;Loeckx, 2016; Wartman and Combs, 2019 Fenwick, 2018; Anu and Ansah, 2023; Abdellatif et al., 2023).

Faculty of Medicine, Wayamba University of Sri Lanka

Corresponding author: Dr. Amila Madhubhashani Wickramarathna
Email: amilaw@wyb.ac.lk

AI chatbots, such as ChatGPT and Bing GPT4, also known as conversational artificial intelligence (CAI), are being increasingly popular in education to engage and motivate students while providing a powerful personalized self-learning experience (Bubaš et al., 2023; Corral, 2021;Jiang et al., 2017;Sedaghat, 2023;Rudolph et al., 2023).

However, the accuracy of AI Chatbots depends on the data they are trained on (Sallam, 2023), leading to biases and errors in responses (Jiao et al., 2023). Bing GPT4 and ChatGPT are two popular CAI has the potential to enhance student learning by promoting knowledge dissemination and critical evaluation skills (Wu et al., 2023; Mollick and Mollick, 2022; Mollick and Mollick, 2023; Rudolph et al., 2023). As further research and evaluation are needed to ensure the best integration of AI resources into medical education, (Lee, 2023; Sedaghat,

2023) this study assessed the accuracy, relevance, comprehensiveness, and user-friendliness of ChatGPT3.5 and Bing GPT4s responses on human anatomy.

Materials and methods

The study evaluated the performance of ChatGPT3.5 and Bing GPT4 in providing fundamental information, create questions on anatomy, and give anatomical explanations to clinical conditions and aid in research on anatomical variations. Twelve questions were prepared in categories regarding gross anatomy, histology, embryology, imaging, clinical anatomy and anatomical variations spanning the module based medical undergraduate anatomy curriculum. Topics were randomly selected from the each module. Model answers for each question was prepared based on the recommended reading materials for the medical undergraduate anatomy curriculum. Two investigators tasked the each Chabot independently and the responses were collected separately to ensure unbiased followed assessment. Responses were then reviewed for accuracy, relevance, comprehensiveness, and user-friendliness using the model answers as a guide by another investigator who was a subject expert and blind to the tasking phase. Each response was then graded as 'Excellent', 'Satisfactory', or 'Inadequate' subjectively based on the quality of the responses.

Results

Both ChatGPT3.5 and Bing GPT4 version were able to provide detailed descriptions of gross anatomy, human embryology, and histology. In general ChatGPT3.5 was superior to Bing GPT4 in the provision of information in gross anatomy, embryology, and histology (Table 1). However, we captured one instance where ChatGPT3.5 provided the wrong information. Therefore, we still cannot guarantee that all the information provided in ChatGPT3.5 is accurate and adequate compared to standard text books. Though both AIs were incapable of providing illustrations and images, interestingly, both suggested links from which those could be downloaded.

Both were unsuccessful in translating common anatomy terms into Sinhala and Tamil, the most commonly used native languages in Sri Lanka. When compared with traditional methods like referring text books when searching for information, illustrations and translation, their accuracy remains superior to both Chatbots with no doubt. However, the relevance of the responses, comprehensiveness of the responses may vary with the situation and individual abilities as text books does not provide customized or personalized responses like summary note. So in terms of efficacy and user friendliness Chatbots would be more advance.

Both AI tools further showed a need for improvement in generating sample questions (Table 1). Their ability to accurately and effectively assess knowledge in this aspect was lacking. The single-best-answer questions and the true and false-type questions generated by both AIs were below average. Essay questions were too broad and asked to cover a wide area. Both AIs were incapable of providing illustrated spot questions. In terms of applied anatomy, both tools were capable of providing explanations regarding the clinical significance and anatomical basis of clinical scenarios to a certain extent. However, we found that Bing GPT4 was providing confusing information on two occasions. Among the two AIs, ChatGPT3.5 was more accurate, detailed, and user-friendly. Furthermore, only the ChatGPT3.5 performed well in providing proper information on anatomical variations (Table 1).While traditional methods like referring recommended text books in generation of questions still provide superior accuracy and relevance compared to Chatbots, the ability to understand those questions can vary depending on user’s individual abilities and level of preparedness. In terms of applied anatomy aspect textbooks and research articles always do not offer personalized or customized responses like interactive Chatbots, making the latter more advanced in terms of comprehensiveness and user-friendliness.

Table 1: Critical analysis of the ability of ChatGPT and Bing Chat to provide knowledge, aid in assessment and explain applied anatomy and anatomical variations

Question	Critical analysis of the answers given by the chatbots	
	ChatGPT	**Bing**
Gross Anatomy "Give me a summary note on brachial plexus"	**Summary of results of the request** Inaccurate information was provided (i.e., brachial plexus consists of five primary nerve roots). Parts rami/roots, trunks, divisions, cords, and terminal branches were described. The relation to scalene muscles was not mentioned. Nerves arising from each part were not mentioned (only the names of some important nerves were mentioned). Clinical significances of the brachial plexus were not mentioned.	**Summary of results of the request** Information on its origin "originates from the anterior rami of spinal nerves C5-T1" was provided Did not describe the parts Did not mention the relation to scalene muscles Did not describe any nerves, not even the names the nerves arising from each part Did not mention clinical significance
Reviewer's comments		
Accuracy	**Inadequate**	**Satisfactory**
Relevance	**Excellent**	**Excellent**
Comprehensiveness	**Satisfactory**	**Satisfactory**
User-friendliness	**Excellent**	**Satisfactory**
Embryology "Describe the process of neurulation"	Stated that "Neurulation begins shortly after fertilization" Described the fusion of the neural tube and bidirectional (cranial and caudal) proceeding Did not mention about the role of the notocord Did not mention about primary neurulation and secondary neurulation Mentions the consequences of failure of the neural tube closure giving an example	Stated that "Neurulation begins shortly after fertilization" Described the fusion of neural tube and bidirectional (cranial and caudal) proceeding Mentioned the role of the notocord mentioned primary neurulation and secondary neurulation Mentioned the consequences of failure of the neural tube closure without giving an example
Reviewer's comments		
Accuracy	**Satisfactory**	**Excellent**
Relevance	**Excellent**	**Excellent**
Comprehensiveness	**Satisfactory**	**Satisfactory**
User-friendliness	**Excellent**	**Satisfactory**
Histology "Provide me a description on human urinary bladder epithelium"	Mentioned that it is a transitional epithelium and transitional in nature Described basal cells, intermediate cells, and surface (umbrella) cells Described specialized barrier function, tight junctions and glycocalyx	Mentioned that it is a transitional epithelium and transitional in nature Does not mention about basal cells, intermediate cells, and surface (umbrella) cells Does not mention about specialized barrier function, tight junctions and glycocalyx

Reviewer's comments		
Accuracy	**Satisfactory**	**Inadequate**
Relevance	**Excellent**	**Excellent**
Comprehensiveness	**Satisfactory**	**Satisfactory**
User-friendliness	**Excellent**	**Satisfactory**
Illustrations "Provide illustration of brain, thyroid and heart valves"	Cannot provide illustration Provided links	Cannot provide illustration Provided links
Reviewer's comments		
Accuracy	**Satisfactory**	**Satisfactory**
Relevance	**Inadequate**	**Inadequate**
Comprehensiveness	**Inadequate**	**Inadequate**
User-friendliness	**Inadequate**	**Inadequate**
Imaging "Provide me following; *x-ray couple bones *IV urogram *MRI brain *Echocardiogram of heart * CT chest * Ultrasound scan of abdomen"	Cannot provide illustration Provided links	Cannot provide illustration Provided links
Reviewer's comments		
Accuracy	**Satisfactory**	**Satisfactory**
Relevance	**Inadequate**	**Inadequate**
Comprehensiveness	**Inadequate**	**Inadequate**
User-friendliness	**Inadequate**	**Inadequate**
Translation from English to Sinhala/Tamil (The two common native languages of Sri Lanka) "Translate the following terms into Sinhala and Tamil • Anterior • Inferior • Sagittal • Abduction • Abdomen"	All translations were incorrect	All translations were incorrect. except the Sinhala translation of "Inferior" "පහළ"
Reviewer's comments		
Accuracy	**Inadequate**	**Inadequate**
Relevance	**Inadequate**	**Inadequate**
Comprehensiveness	**Inadequate**	**Inadequate**
User-friendliness	**Inadequate**	**Inadequate**

The ability of ChatGPT and Bing Chat to generate assessment items in Human anatomy		
"Generate a single best response type question on anatomy of the stomach"	Created single answer not single best Assess only the knowledge level Gave the correct answer	Created single answer not single best Assess only the knowledge level Do not give the answer
Reviewer's comments		
Accuracy	**Inadequate**	**Inadequate**
Relevance	**Inadequate**	**Inadequate**
Comprehensiveness	**Inadequate**	**Inadequate**
User-friendliness	**Satisfactory**	**Inadequate**
"Generate a quiz that ask to match different lower limb muscles to the corresponding motor nerves in lower limb"	Gave some good quality questions	Stated that it is beyond its capability. Provided some links with quizzes
Reviewer's comments		
Accuracy	**Satisfactory**	**Inadequate**
Relevance	**Satisfactory**	**Inadequate**
Comprehensiveness	**Satisfactory**	**Inadequate**
User-friendliness	**Satisfactory**	**Inadequate**
The ability of ChatGPT and Bing Chat to respond regarding applied anatomy and anatomical variations		
"Explain that anatomical basis of intracapsular fractures of hip joint being mole Likely to cause avascular necrosis of femur head in comparison to extra capsular fractures"	Describes that it is due to a difference in in the blood supply to the femoral head. Do not mention about trochanteric and cruciate anastomosis except that fact that blood supply to the femoral head primarily comes from branches of the medial and lateral circumflex femoral arteries, which are located outside the joint capsule Do not mention about the obturator artery	Mentions that blood supply to the femoral head travels in a retrograde direction via the capsule. Do not mention about trochanteric and cruciate anastomosis Do not mention about the obturator artery Gives confusing information "The effects of traumatic dislocation on femoral and acetabular articular cartilage can also lead to arthrosis. Damage to the hip capsule and hip musculature may cause periarticular fibrosis and heterotopic ossification, which can produce functional limitations".
Reviewer's comments		
Accuracy	**Inadequate**	**Inadequate**
Relevance	**Inadequate**	**Inadequate**
Comprehensiveness	**Inadequate**	**Inadequate**
User-friendliness	**Inadequate**	**Inadequate**

"Explain the embryological reason for congenital diaphragmatic hernia"	states that it is due to an abnormality occurs during fetal development Describes some embryological structure that forms that diaphragm Mentions "developmental timing, herniation of abdominal contents, effects on lung development, factors contributing to CDH"	states that it is believed to be related to the failure of the pleuroperitoneal folds to fuse properly" Do not mention any embryological structure that forms that diaphragm Do not mention "developmental tim herniation of abdominal contents, effects on lung development, factors contributing to CDH" except the fact "lungs may not develop fully"
Reviewer's comments		
Accuracy	**Satisfactory**	**Inadequate**
Relevance	**Satisfactory**	**Inadequate**
Comprehensiveness	**Satisfactory**	**Inadequate**
User-friendliness	**Satisfactory**	**Inadequate**
"Explain the embryological reason for congenital brachydactyly"	Gives correct general idea of the condition as "abnormal shortening or underdevelopment of one or more fingers or toes". Gives correct causes (genetic mutations or variations that disrupt the normal processes of digital ray development) Describes types of brachydactyly Describes digital ray development and how it is related to brachydactyly	Gives correct general idea of the condition as "that causes fingers and toes to appear shorter than usual in proportion to other parts of the body". Gives confusing information by mentioning it as a genetic condition in one place and the effect of non-genetic causes like anticonvulsant medication and poor blood flow in intrauterine life in another place. Describes types of brachydactyly Do not describe digital ray development and how it is related to brachydactly
Reviewer's comments		
Accuracy	**Excellent**	**Inadequate**
Relevance	**Excellent**	**Inadequate**
Comprehensiveness	**Excellent**	**Inadequate**
User-friendliness	**Excellent**	**Inadequate**
"Describe the Anatomical variations of drainage of drainage of dural venous sinuses"	Give a good account on anatomical variations including variations in confluence of sinuses, variations in the number or size of dural sinuses, absence (agenesis) or underdevelopment (hypoplasia) of certain dural sinuses or their tributaries, presence of accessory sinuses, and hemispheric dominance	Do not give adequate information Only just a few examples of the normal dural venous sinuses and their drainage pathways.
Accuracy	**Satisfactory**	**Inadequate**

Relevance	**Satisfactory**	**Inadequate**
Comprehensiveness	**Satisfactory**	**Inadequate**
User-friendliness	**Satisfactory**	**Inadequate**

Discussion

Studies are being conducted to compare different AI tools used for learning, assessment, and research in human anatomy (Mollick and Mollick, 2022). ChatGPT3.5 and Bing GPT4 were found to provide accurate descriptions of gross anatomy and human embryology, although Bing GPT4 performed poorly in histology. These findings are compatible with the findings of the evaluation of the general use of these Chatbots in education to a certain extent, where the investigators have found that the Chatbots are not performing as well in assignment questions that are not difficult to write (Rudolph et al., 2023; Totlis et al., 2023). Both tools used similar sources of information but ChatGPT3.5 had an edge in providing academic references and links to related sites.

Our study found that AI tools need significant improvement in aiding assessment of human anatomy as none of the AI tools tested produced accurate, high-quality sample questions. However, ChatGPT-3.5 and Google Bard showed promise in creating and answering questions in anatomy for medical education (Ilgaz and Çelik 2023). While both tools were able to provide explanations of clinical significance in applied anatomy, ChatGPT3.5 was found to be more user-friendly and capable of providing concise summaries. In a study on pharmacology, Bing GPT4 AI outperformed Google Bard, ChatGPT3.5, and ChatGPT4 in terms of accuracy and specificity (Al-Ashwal et al., 2023). The ability of AI tools to provide information on anatomical variations varied depending on the area of concern, with ChatGPT performing better in this aspect compared to Bing GPT4.

Regarding the applied anatomy aspect, ChatGPT3.5 was found to be superior in providing concise summaries and user-friendly information. However, there was a study in pharmacology where Bing GPT4 AI had higher accuracy than ChatGPT. ChatGPT was also better at providing information on anatomical variations compared to Bing GPT4 (Al-Ashwal et al., 2023). However, there are conflicting findings on the adequacy of information on anatomical variations in different studies, suggesting the ability to provide such information may vary depending on the specific anatomical area (Totlis et al., 2023).

Based on the findings we recommend using ChatGPT3.5 and Bing GPT4 cautiously, in anatomy education acknowledging that the representation of the bots and their potential to aid learning anatomy may change as AI technology evolves. Traditional methods still offer accurate information, while Chatbots provide convenience and user-friendliness and personalized responses. However, a single snapshot of the performance during the rapid advancement of AI technology is inadequate to provide recommendations for students to choose between the Chatbots and old methods for learning anatomy. Further, the limitations of this study includes analysis of only two AI tools and focusing mainly on subjective content analysis. Therefore, we suggest a wider assessment of anatomy curriculum covering more content using a blueprint to ensure systematic sampling of topics. Involvement of more subject experts in the review process with calculated inter-rater reliability and development of objective review process in future research is suggested.

Ethics committee approval: Since there is no human or animal involvement, specific ethical clearance was not received.

References

Abdellatif, H., Al Mushaiqri, M., Albalushi, H., Al-Zaabi, A.A., Roychoudhury, S. and Das, S., 2022. Teaching, learning and assessing anatomy with artificial intelligence: the road to a better future. International Journal of Environmental Research and Public Health, 19(21), p.14209.

Al-Ashwal, F.Y., Zawiah, M., Gharaibeh, L., Abu-Farha, R. and Bitar, A.N., 2023. Evaluating the sensitivity, specificity, and Accuracy of ChatGPT-3.5, ChatGPT-4, Bing AI, and bard against conventional drug-drug interactions clinical tools. Drug, Healthcare and Patient Safety, pp.137-147.

Baidoo-Anu, D. and Ansah, L.O., 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. Journal of AI, 7(1), pp.52-62.

Bubaš, G., Čižmešija, A. and Kovačić, A., 2023. Development of an Assessment Scale for Measurement of Usability and User Experience Characteristics of Bing Chat Conversational AI. Future Internet, 16(1), p.4.

Corral, J., 2021. Artificially intelligent chatbots for health professions education. In Digital Innovations in Healthcare Education and Training (pp. 127-135). Academic Press.

Fenwick, T., 2018. Pondering purposes, propelling forwards. Studies in Continuing Education, 40(3), pp.367-380.

Grunhut, J., Wyatt, A.T. and Marques, O., 2021. Educating future physicians in artificial intelligence (AI): an integrative review and proposed changes. Journal of Medical Education and Curricular Development, 8, p.23821205211036836.

Hamet, P. and Tremblay, J., 2017. Artificial intelligence in medicine. Metabolism, 69, pp.S36-S40.

Ilgaz, H.B. and Çelik, Z., 2023. The significance of artificial intelligence platforms in anatomy education: an experience with ChatGPT and google bard. Cureus, 15(9).

Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., Wang, Y., Dong, Q., Shen, H. and Wang, Y., 2017. Artificial intelligence in healthcare: past, present and future. Stroke and vascular neurology, 2(4).

Jiao, W., Wang, W., Huang, J.T., Wang, X., Shi, S. and Tu, Z., 2023. Is ChatGPT a good translator? Yes with GPT-4 as the engine. arXiv preprint arXiv:2301.08745.

Kulkarni, S., Seneviratne, N., Baig, M.S. and Khan, A.H.A., 2020. Artificial intelligence in medicine: where are we now?. Academic radiology, 27(1), pp.62-70.

Lee, H., 2023. The rise of ChatGPT: Exploring its potential in medical education. Anatomical sciences education.

Loeckx, J., 2016. Blurring boundaries in education: Context and impact of MOOCs. International Review of Research in Open and Distributed Learning, 17(3), pp.92-121.

Mollick, E.R. and Mollick, L., 2022. New modes of learning enabled by ai chatbots: Three methods and assignments. Available at SSRN 4300783.

Mollick, E.R. and Mollick, L., 2023. Using AI to implement effective teaching strategies in classrooms: Five strategies, including prompts. Including Prompts (March 17, 2023).

Rudolph, J., Tan, S. and Tan, S., 2023. War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. Journal of Applied Learning and Teaching, 6(1).

Sallam, M., 2023, March. ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In Healthcare (Vol. 11, No. 6, p. 887). MDPI.

Sallam, M., 2023. The utility of ChatGPT as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations. MedRxiv, pp.2023-02.

Sedaghat, S., 2023. Early applications of ChatGPT in medical practice, education and research. Clinical Medicine, 23(3), pp.278-279.

Sedaghat, S., 2023. Early applications of ChatGPT in medical practice, education and research. Clinical Medicine, 23(3), pp.278-279.

Totlis, T., Natsis, K., Filos, D., Ediaroglou, V., Mantzou, N., Duparc, F. and Piagkou, M., 2023. The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. Surgical and Radiologic Anatomy, 45(10), pp.1321-1329.

Wartman, S.A. and Combs, C.D., 2019. Reimagining medical education in the age of AI. AMA journal of ethics, 21(2), pp.146-152.