ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Volume 10 Issue 12 December 2023

Evaluating ChatGPT Performance in Answering Questions Related to Abdominal Anatomy

[1] Olena Bolgova, [2] Volodymyr Mavrych *

[1] Department of Anatomy and Genetics, College of Medicine, Alfaisal University, Riyadh, Kingdom of Saudi Arabia
Corresponding Author Email: [1] obolgova@alfaisal.edu, [2] vmavrych@alfaisal.edu *

Abstract*— In construction industry, builders and engineers are more interested in application cold-formed steel (CFS) sections in place of conventional materials not only as the non-structural components but also the structural members of the commercial and residential buildings. As the structural members, the literatures endorsed that built-up CFS sections have higher strength than single detached sections. The objective of this paper is to analyze the buckling behaviors of CFS built-up box slender columns by means of FE software ANSYS 2020 R1 grounded on the recent experimental models and compares the outputs for design optimization. Face-to-face built-up box slender columns were connected with fillet weld joint spacing of 500 mm and analyzed for Eigenvalue buckling loads. Numerical results by ANSYS 2020 R1 were within the range of allowable compressive loads and global buckling governs for slender columns.

Keywords: Anatomy, Artificial intelligence, ChatGPT, Medical Education.

I. INTRODUCTION

During The quickly developing realm of artificial intelligence (AI) has great promise for transforming medical education. AI can help with student assessment, offer individualized learning opportunities, and facilitate the integration of pre-clinical and clinical courses. More research on the application of AI in undergraduate medical education is needed despite the possible advantages. This study compares artificial intelligence to existing teaching and assessment techniques in order to examine the role of AI in undergraduate medical curricula globally [1]

The term artificial intelligence refers to the development of computer systems that are able to do operations like sensing, reasoning, and decision-making that generally call for human intelligence. AI is utilized in the healthcare industry to analyze vast volumes of patient data, including imaging scans, laboratory results, and medical records, to enhance clinical decision-making and patient outcomes. Machine Learning (ML) is a branch of artificial intelligence that creates models and algorithms that can learn from data without explicit programming. ML algorithms can be trained on big datasets in the healthcare industry to find trends, forecast results, and diagnose patients. This increases the precision of medical professionals' diagnoses and aids their decision-making. [2].

A lifetime of learning is required for medical education, which includes postgraduate work, specialty training, undergraduate study, and more. It also holds true for doctors, nurses, and other allied health professionals. In light of this, we must recognize the enormous contribution that artificial intelligence makes to medical education in this day of rapidly advancing technology [3]. It is essential to address the challenges and constraints associated with ChatGPT, including ethical considerations and the potential for adverse effects. Medical educators must remain attuned to the rapidly evolving technological landscape and its implications for curriculum design, assessment methodologies, and teaching strategies. Continuous research and assessment are essential for effectively integrating AI-based learning tools into medical education [4, 5].

An interesting investigation compared the quality of multiple-choice questions (MCQs) produced by ChatGPT to those made by university professors for medical graduate exams. While human examiners needed 211 minutes to complete 50 multiple-choice questions, ChatGPT completed the task in roughly 20 minutes [6]. Independent experts evaluated ChatGPT's questions and concluded that, except for the relevancy category, where it received a slightly lower score, the quality of the questions was on par with those written by humans. On the other hand, human-produced questions displayed a greater scoring consistency than those created by ChatGPT. ChatGPT can provide excellent multiple-choice questions for medical graduate exams and assist with item creation. It emerges as a potent next-generation method for medical evaluation in the future, guaranteeing several high-quality goods in a timely and cost-effective manner [7].

Numerous studies looked into ChatGPT's ability to help medical students with their research and instruction on anatomy. Inquiries were made to ChatGPT to assess its precision, applicability, and thoroughness. ChatGPT offered accurate anatomical and structural information along with clinical significance. It also provided help with terminology and summaries. Nevertheless, systematic classification was required to improve its responses to anatomical variances [8].

According to some recent publications, ChatGPT performs well when answering MCQs, which may affect the educational system. Regarding accuracy and consistency, ChatGPT-3.5 outperformed Google Bard in answering lung cancer prevention and screening questions. ChatGPT-3.5 answered 70.8% of the questions correctly, whereas Google

ISSN (Online) 2456 -1304

International Journal of Science, Engineering and Management (IJSEM)

Volume 10 Issue 12 December 2023

Bard answered 51.7%. [9]. ChatGPT offered pertinent responses to frequently asked patient inquiries concerning total hip replacement and optic disc drusen. However, some of the answers should be more precise, especially those about prognosis and treatment, which can be detrimental in some situations [10, 11].

The need to overhaul medical education is growing along with the evolution of health care. The use of data to enhance clinical decision-making will increase as medicine moves into the era of artificial intelligence (AI), increasing the demand for proficient medicine-machine interaction. Technologies like artificial intelligence (AI) are required to allow medical practitioners to efficiently employ the growing body of medical knowledge to practice medicine. Medical personnel must receive sufficient training on this new technology, including its benefits for enhancing access, affordability, and quality of care, as well as its drawbacks, including liability and transparency. AI must be smoothly incorporated into all facets of the curriculum [12].

There are numerous papers on this subject despite ChatGPT's recent introduction. However, only some of them offer statistical data. Our study's primary goals were to evaluate ChatGPT-3.5's performance on Abdominal material MCQs and its ability to improve the results based on feedback from the evaluator.

II. MATERIALS AND METHODS

This study's research focused on an in-depth evaluation of ChatGPT-3.5's ability to respond to 50 multiple-choice questions (MCQs) written in the USMLE style format. These questions were randomly chosen from the 2020 Gross Anatomy course exam database for medical students. There were varying degrees of difficulty among the questions, and none contained pictures. By doing this, we were able to get around the restriction on the lack of real-time information: ChatGPT-3.5 has no access to real-time information on events that occur after September 2021, as its understanding is dependent on text data up to that date.

The results of 10 successive attempts by ChatGPT to answer this set of questions were evaluated based on accuracy, relevance, and comprehensiveness. The first five successive attempts (Series 1) were made without giving feedback to the GPT-3.5. To check the ability of ChatGPT to be trained, the following five successive attempts (Series 2) were made with the feedback from the evaluator to the system after each attempt, depending on the results compared to the previous attempt (positive feedback if the result was better or negative if the result was worst).

Each ChatGPT attempt's data was compared with previous attempts, finding the percentage of repeated answers and correct answers among them.

To compare the ChatGPT results with random guessing, we generated five sets of random answers to the same test using Excel's RAND () function. Microsoft Excel (Microsoft®365) was utilized for the basic statistical analysis. submission.

III. RESULTS

According to our data, ChatGPT provided accurate answers to 42.4±4.1% of the selected questions across ten successive attempts, much superior to random guessing – 18.4±2.2% (Fig. 1).

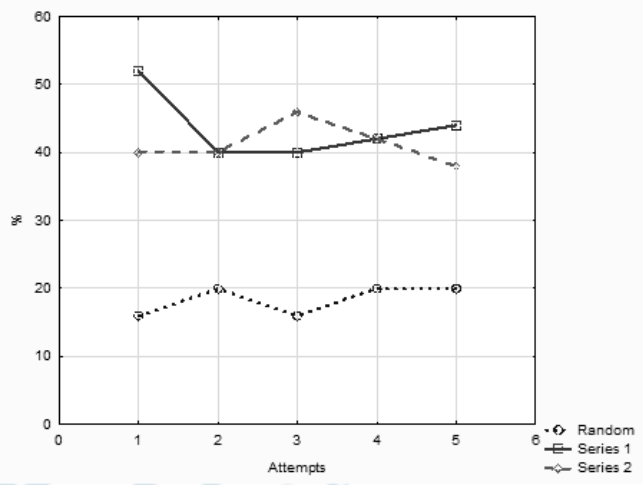

Fig. 1. Percentile of correct answers of ChatGPT on 50 MCQs in Series 1 and Series 2 compared with random answers.

Five attempts from Series 1 without feedback from the evaluator to the Chatbot had 43.6±4.9% correct answers. Each attempt Chatbot generated different list of the answers. The first attempt was the most successful, with 52% correct answers. The results of the following four attempts fluctuated between 40% - 44%. The coincidence of answers with the previous generations was 46% - 60%, and among them, the coincidence of correct answers was 26% - 38% (Tab. 1).

Table 1. % of Correct Answers, Coincidence with a Previous Attempt, and Coincidence of Correct Answers with a Previous Attempt for 5 Attempts From Series 1

Attempt	1	2	3	4	5
Correct answers	52	40	40	42	44
Coincidence with 1		60	54	48	60
Coincidence corrects with 1		38	32	34	38
Coincidence with 2			64	48	56
Coincidence corrects with 2			32	30	32
Coincidence with 3				54	46
Coincidence corrects with 3				28	26
Coincidence with 4					46
Coincidence corrects with 4					26

In Series 2, the following five attempts with positive or negative feedback to the Chatbot had 41.2±3.0% correct

ISSN (Online) 2456 -1304
International Journal of Science, Engineering and Management (IJSEM)
Volume 10 Issue 12 December 2023

answers, paradoxically less than Series 1 without feedback. The coincidence of answers with the previous generations was 50% - 68%, which is higher compared with the previous Series 1 answers and may indicate that the system has fewer potential variations to play with. Among them, the coincidence of correct answers was 26% - 32%, almost the same as in Series 1 (Tab. 2).

Table 2. % of Correct Answers, Coincidence with a Previous Attempt, and Coincidence of Correct Answers with a Previous Attempt for 5 Attempts From Series 2

Attempt	1	2	3	4	5
Correct answers	40	40	46	42	38
Coincidence with 1		68	56	52	54
Coincidence corrects with 1		30	32	32	26
Coincidence with 2			56	50	54
Coincidence corrects with 2			32	30	28
Coincidence with 3				56	58
Coincidence corrects with 3				30	28
Coincidence with 4					56
Coincidence corrects with 4					28

Only 10 questions (20%) were answered correctly across all 10 attempts. The item analysis indicated that all these 10 MCQs were simple recall questions. When it came to more comprehensive questions, the Chatbot's replies were only sometimes adequate and varied across the attempts.

IV. DISCUSSION

Use The result of our study indicated that ChatGPT is a powerful A.I. language model and can answer MCQs in the Gross Anatomy course for medical students, with an average result of 42.4%. Our data correlate well with another recent study that evaluated its performance in answering medical questions from the U.S. Medical Licensing Exams. ChatGPT achieved 42% to 64.4% accuracy, surpassing other models like Instruct GPT. This research also indicates that ChatGPT's performance decreased with question difficulty [13].

Additionally, it strongly correlates with our data, which showed that just 20% of the questions were successfully answered on all 10 tries. As per the item analysis, each of the ten multiple-choice questions was a straightforward recall question. The Chatbot's responses to more in-depth queries need to be more appropriate and inconsistent throughout the efforts. However, ChatGPT is outperforming other areas like biochemistry, physiology, and head and neck surgery - especially when answering questions that are not multiple-choice (MCQs).

Results of Vaira LA et al. suggest that ChatGPT's effectiveness in solving reasoning questions related to the core concepts in physiology across different modules is good. Answering reasoning-type questions created by instructors, ChatGPT shows approximately 74% of the correct answers [14].

In answering questions and solving clinical scenarios in head and neck surgery, ChatGPT achieved a correct response in 84.7% of closed-ended questions and provided an entirely or nearly correct diagnosis in 81.7% of clinical scenarios. However, the completeness of the proposed procedures and the quality of bibliographic references needed to be improved [15].

The other study aimed to assess ChatGPT's ability to answer higher-order questions related to medical biochemistry. ChatGPT was presented with 200 such questions and received a median score of 80% in its responses. The responses were consistent across different modules in medical biochemistry. Authors concluded that ChatGPT shows potential in effectively addressing higher-order questions in medical biochemistry, but ongoing training and development are necessary for optimal performance in academic medical contexts [16].

Our data indicated that the current version of ChatGPT needs a more effective way to be trained. Feedback to ChatGPT based on the test results did not improve the chat performance: 43.6±4.9% correct answers initially and 41.2±3.0% after the training.

One of our study's limitations was that English is the only language used to communicate with ChatGPT because all the questions were written in that language. Comparing how well it does while responding to queries in other languages will be interesting.

Another study aimed to assess ChatGPT's performance in answering the 2022 Brazilian National Examination for Medical Degree Revalidation and its use in evaluating the exam's quality. Two physicians input all exam questions into the system, and after comparing their responses to the official answers, they categorized the model's answers as adequate, inadequate, or indeterminate. Disagreements were resolved to reach a consensus on ChatGPT's accuracy. The model correctly answered 87.7% of Revalida questions, with no significant differences across medical topics [17].

ChatGPT was assessed for its accuracy and reliability in providing knowledge, management, and emotional support for patients with cirrhosis and hepatocellular carcinoma who often require personalized care. The chat demonstrated good knowledge retention for cirrhosis (79.1% correct) and HCC (74.0% correct), with some responses needing more comprehensiveness. It performed better in basic knowledge, lifestyle, and treatment aspects compared to diagnosis and prevention. While it answered 76.9% of quality measure questions correctly, it fell short in specifying decision-making criteria and treatment durations [18].

The other study assessed ChatGPT's ability to respond to cirrhosis-related questions in Arabic and compared its performance to English. Cirrhosis is a growing health concern in Arab countries. The model provided

comprehensive answers in 24.2% of Arabic responses, with 72.5% correct and 13.2% of responses needing to be corrected. Compared to English, Arabic responses could have been more accurate in 33% of cases [19].

In most U.S. medical colleges and schools, the passing grade in MCQ subject-based exams is 65% - 75% [20]. So, according to our data, ChatGPT-3.5's performance is currently below the passing grade in the Abdomen material. To evaluate ChatGPT's performance more precisely, more studies should be done to cover all the material of the gross anatomy course.

V. CONCLUSION

ChatGPT demonstrates considerable potential as an interactive and enriching educational tool for students delving into the intricacies of anatomy. Its unique ability to cultivate student engagement and spark curiosity through conversational responses to inquiries stands out as a commendable feature.

However, it is crucial to acknowledge that the GPT-3.5 model's performance currently needs to catch up to the threshold required for passing grades in examinations. Recognizing this, we tried an available feedback mechanism on the platform to facilitate improvements in its performance and found that it is inadequate now.

It is imperative to emphasize that ChatGPT should be perceived as something other than a replacement for educators' indispensable role in the learning process. Instead, it should be regarded as a supplementary resource to augment and elevate the educational experience. Educators remain the cornerstone of effective instruction, providing guidance, support, and personalized insights that AI tools cannot replicate.

Moving forward, we advocate for dedicated research efforts to formulate comprehensive guidelines. These guidelines will elucidate the optimal utilization and application of ChatGPT within the realm of anatomy instruction. By establishing clear protocols and best practices, educators can leverage ChatGPT as a powerful ally in the educational journey, enhancing students' understanding and engagement with the subject matter. This collaborative approach, combining the strengths of both educators and AI, holds the potential to redefine and elevate the anatomy learning experience.

ACKNOWLEDGMENT

The authors would like to acknowledge all support from the Department of Anatomy and Genetic, College of Medicine at Alfaisal University.

COMPETING INTERESTS

The authors declare no conflicts of interest, financial or otherwise.

REFERENCES

[1] Varma JR, Fernando S, Ting BY, Aamir S, Sivaprakasam R. The Global Use of Artificial Intelligence in the Undergraduate Medical Curriculum: A Systematic Review. Cureus. 2023 May 30;15(5):e39701. doi: 10.7759/cureus.39701. PMID: 37398823; PMCID: PMC10309075.

[2] Dave M, Patel N. Artificial intelligence in healthcare and education. Br Dent J. 2023 May;234(10):761-764. doi: 10.1038/s41415-023-5845-2. Epub 2023 May 26. PMID: 37237212; PMCID: PMC10219811.

[3] Mir MM, Mir GM, Raina NT, Mir SM, Mir SM, Miskeen E, Alharthi MH, Alamri MMS. Application of Artificial Intelligence in Medical Education: Current Scenario and Future Perspectives. J Adv Med Educ Prof. 2023 Jul;11(3):133-140. doi: 10.30476/JAMP.2023.98655.1803. PMID: 37469385; PMCID: PMC10352669.

[4] Lee H. The rise of ChatGPT: Exploring its potential in medical education. Anat Sci Educ. 2023 Mar 14. doi: 10.1002/ase.2270. Epub ahead of print. PMID: 36916887.

[5] Jamal A, Solaiman M, Alhasan K, Temsah MH, Sayed G. Integrating ChatGPT in Medical Education: Adapting Curricula to Cultivate Competent Physicians for the AI Era. Cureus. 2023 Aug 6;15(8):e43036. doi: 10.7759/cureus.43036. PMID: 37674966; PMCID: PMC10479954.

[6] Cheung BHH, Lau GKK, Wong GTC, Lee EYP, Kulkarni D, Seow CS, Wong R, Co MT. ChatGPT versus human in generating medical graduate exam multiple choice questions-A multinational prospective study (Hong Kong S.A.R., Singapore, Ireland, and the United Kingdom). PLoS One. 2023 Aug 29;18(8):e0290691. doi: 10.1371/journal.pone.0290691. PMID: 37643186; PMCID: PMC10464959.

[7] Falcão F, Costa P, Pêgo JM. Feasibility assurance: a review of automatic item generation in medical assessment. Adv Health Sci Educ Theory Pract. 2022 May;27(2):405-425. doi: 10.1007/s10459-022-10092-z. Epub 2022 Mar 1. PMID: 35230589; PMCID: PMC8886703.

[8] Totlis T, Natsis K, Filos D, Ediaroglou V, Mantzou N, Duparc F, Piagkou M. The potential role of ChatGPT and artificial intelligence in anatomy education: a conversation with ChatGPT. Surg Radiol Anat. 2023 Oct;45(10):1321-1329. doi: 10.1007/s00276-023-03229-1. Epub 2023 Aug 16. PMID: 37584720; PMCID: PMC10533609.

[9] Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard. Radiology. 2023 Jun;307(5):e230922. doi: 10.1148/radiol.230922. PMID: 37310252.

[10] Potapenko I, Malmqvist L, Subhi Y, Hamann S. Artificial Intelligence-Based ChatGPT Responses for Patient Questions on Optic Disc Drusen. Ophthalmol Ther. 2023 Sep 12. doi: 10.1007/s40123-023-00800-2. Epub ahead of print. PMID: 37698823.

[11] Mika AP, Martin JR, Engstrom SM, Polkowski GG,

Wilson JM. Assessing ChatGPT Responses to Common Patient Questions Regarding Total Hip Arthroplasty. J Bone Joint Surg Am. 2023 Oct 4;105(19):1519-1526. doi: 10.2106/JBJS.23.00209. Epub 2023 Jul 17. PMID: 37459402.

[12] Paranjape K, Schinkel M, Nannan Panday R, Car J, Nanayakkara P. Introducing Artificial Intelligence Training in Medical Education. JMIR Med Educ. 2019 Dec 3;5(2):e16048. doi: 10.2196/16048. PMID: 31793895; PMCID: PMC6918207.

[13] Gilson A, Safranek CW, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. 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 Feb 8;9:e45312. doi: 10.2196/45312. PMID: 36753318; PMCID: PMC9947764.

[14] Banerjee A, Ahmad A, Bhalla P, Goyal K. Assessing the Efficacy of ChatGPT in Solving Questions Based on the Core Concepts in Physiology. Cureus. 2023 Aug 10;15(8):e43314. doi: 10.7759/cureus.43314. PMID: 37700949; PMCID: PMC10492920.

[15] Vaira LA, Lechien JR, Abbate V, Allevi F, Audino G, Beltramini GA, Bergonzani M, Bolzoni A, Committeri U, Crimi S, Gabriele G, Lonardi F, Maglitto F, Petrocelli M, Pucci R, Saponaro G, Tel A, Vellone V, Chiesa-Estomba CM, Boscolo-Rizzo P, Salzano G, De Riu G. Accuracy of ChatGPT-Generated Information on Head and Neck and Oromaxillofacial Surgery: A Multicenter Collaborative Analysis. Otolaryngol Head Neck Surg. 2023 Aug 18. doi: 10.1002/ohn.489. Epub ahead of print. PMID: 37595113.

[16] Ghosh A, Bir A. Evaluating ChatGPT's Ability to Solve Higher-Order Questions on the Competency-Based Medical Education Curriculum in Medical Biochemistry. Cureus. 2023 Apr 2;15(4):e37023. doi: 10.7759/cureus.37023. PMID: 37143631; PMCID: PMC10152308.

[17] Gobira M, Nakayama LF, Moreira R, Andrade E, Regatieri CVS, Belfort R Jr. Performance of ChatGPT-4 in answering questions from the Brazilian National Examination for Medical Degree Revalidation. Rev Assoc Med Bras (1992). 2023 Sep 25;69(10):e20230848. doi: 10.1590/1806-9282.20230848. PMID: 37792871; PMCID: PMC10547492.

[18] Yeo YH, Samaan JS, Ng WH, Ting PS, Trivedi H, Vipani A, Ayoub W, Yang JD, Liran O, Spiegel B, Kuo A. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. Clin Mol Hepatol. 2023 Jul;29(3):721-732. doi: 10.3350/cmh.2023.0089. Epub 2023 Mar 22. PMID: 36946005; PMCID: PMC10366809.

[19] Samaan JS, Yeo YH, Ng WH, Ting PS, Trivedi H, Vipani A, Yang JD, Liran O, Spiegel B, Kuo A, Ayoub WS. ChatGPT's ability to comprehend and answer cirrhosis related questions in Arabic. Arab J Gastroenterol. 2023 Aug;24(3):145-148. doi: 10.1016/j.ajg.2023.08.001. Epub 2023 Sep 4. PMID: 37673708.

[20] Chew C, O'Dwyer PJ, Young D, Gracie JA. Radiology teaching improves Anatomy scores for medical students. Br J Radiol 2020; 93: 20200463