Surgery for Obesity and Related Diseases 20 (2024) 609–613

SURGERY FOR OBESITY AND RELATED DISEASES

ASMBS Guidelines/Statements

Performance of artificial intelligence in bariatric surgery: comparative analysis of ChatGPT-4, Bing, and Bard in the American Society for Metabolic and Bariatric Surgery textbook of bariatric surgery questions

Yung Lee, M.D., M.P.H.[a,b], Léa Tessier, M.D.[a], Karanbir Brar, M.D.[a], Sarah Malone, B.H.Sc.[a], David Jin, B.H.Sc.[a], Tyler McKechnie, M.D., M.Sc.[a], James J. Jung, M.D., Ph.D.[c], Matthew Kroh, M.D.[d], Jerry T. Dang, M.D., Ph.D.[d,*], ASMBS Artificial Intelligence and Digital Surgery Taskforce

[a] *Division of General Surgery, McMaster University, Hamilton, Ontario, Canada*
[b] *Harvard T.H. Chan School of Public Health, Harvard University, Boston, Massachusetts*
[c] *Division of General Surgery, University of Toronto, Toronto, Ontario, Canada*
[d] *Digestive Disease Institute, Cleveland Clinic, Cleveland, Ohio*

Received 12 April 2024; accepted 14 April 2024

Abstract

Background: The American Society for Metabolic and Bariatric Surgery (ASMBS) textbook serves as a comprehensive resource for bariatric surgery, covering recent advancements and clinical questions. Testing artificial intelligence (AI) engines using this authoritative source ensures accurate and up-to-date information and provides insight in its potential implications for surgical education and training.

Objectives: To determine the quality and to compare different large language models' (LLMs) ability to respond to textbook questions relating to bariatric surgery.

Setting: Remote.

Methods: Prompts to be entered into the LLMs were multiple-choice questions found in "The ASMBS Textbook of Bariatric Surgery, second Edition. The prompts were queried into 3 LLMs: OpenAI's ChatGPT-4, Microsoft's Bing, and Google's Bard. The generated responses were assessed based on overall accuracy, the number of correct answers according to subject matter, and the number of correct answers based on question type. Statistical analysis was performed to determine the number of responses per LLMs per category that were correct.

Results: Two hundred questions were used to query the AI models. There was an overall significant difference in the accuracy of answers, with an accuracy of 83.0% for ChatGPT-4, followed by Bard (76.0%) and Bing (65.0%). Subgroup analysis revealed a significant difference between the models' performance in question categories, with ChatGPT-4's demonstrating the highest proportion of correct answers in questions related to treatment and surgical procedures (83.1%) and complications (91.7%). There was also a significant difference between the performance in different question types, with ChatGPT-4 showing superior performance in inclusionary questions. Bard and Bing were unable to answer certain questions whereas ChatGPT-4 left no questions unanswered.

Conclusions: LLMs, particularly ChatGPT-4, demonstrated promising accuracy when answering clinical questions related to bariatric surgery. Continued AI advancements and research is required

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

*Correspondence: Jerry T. Dang, M.D., Ph.D., Digestive Disease Institute, Cleveland Clinic, Cleveland, Ohio.

E-mail address: yung.lee@medportal.ca or leey27@mcmaster.ca (J.T. Dang).

https://doi.org/10.1016/j.soard.2024.04.014

to elucidate the potential applications of LLMs in training and education. (Surg Obes Relat Dis 2024;20:609–613.)

Keywords: Large language models; Chatbot; Artificial intelligence; Bariatric surgery textbook; ASMBS textbook

Artificial intelligence (AI)–based chat models, also known as large language models (LLMs), such as ChatGPT have recently garnered substantial attention, showing promise in a wide range of settings including education, disease diagnosis, triage and screening, and risk analysis [1]. Furthermore, previous studies have investigated the potential applications of these LLMs in medical and surgical education and training. Following its success with the United States Medical Licensing Exam (USMLE), the ability of LLMs to answer resident-level board exams has been investigated in multiple studies [2]. Recently, ChatGPT has shown accuracy in answering board exams from the American Board of Thoracic Surgery (ABTS), European Board of Pediatric Surgery (EBPS), the Canadian Otolaryngology–Head and Neck Surgery Royal College board examination, the Ophthalmic Knowledge Assessment (OKAP) exam, the American Board of Neurological Surgery (ABNS) exam, and more [3–9].

To our knowledge, the performance of LLMs in answering clinical questions related to bariatric surgery has not yet been evaluated. To test the use of AI for clinical questions, this study utilized the American Society for Metabolic and Bariatric Surgery (ASMBS) textbook, which serves as a comprehensive guide on bariatrics, encompassing recent advancements, and the management of bariatric complications. Furthermore, while previous studies have evaluated the ability of ChatGPT in answering clinically relevant questions, there is a paucity of research comparing different AI models in this context. Using the clinical questions presented in this textbook, this study aims to determine the quality and reliability of different AI chat models when responding to clinical questions relating to bariatric surgery and provide insight into the strengths and limitations of the application of these models.

Methods

Selection of prompts and use of AI-based large language models

The prompts selected for use in the LLMs were multiple-choice questions found at the end of all chapters (1–56) of “The ASMBS Textbook of Bariatric Surgery: Second Edition” [10]. A total of 200 questions were included. These questions were queried into 3 AI-based chat models: Open-AI’s ChatGPT-4 (OpenAI), Bing (Microsoft), and Bard (Google), with a new chat conversation being created for each question from May 25th, 2023 to May 30th, 2023. The LLMs were not trained specific to the ASMBS Textbook contents and their ability to answer the questions or generate responses are based on their default settings. All questions were inputted following the prompt: “Answer this multiple-choice question:” (Appendices 1–3). If the LLMs did not provide one of the answer choices listed, it was prompted to choose a single answer for a maximum of 3 attempts by the follow-up statement “please pick an answer.” An Excel Spreadsheet was used to organize the questions and responses. Textbook questions were organized into columns pertaining to chapter and question number, question text, multiple-choice options, and the correct multiple-choice option. Responses were organized into 3 columns based on the type of LLM.

Outcomes

The responses generated by the AI chat models were compared to the correct responses provided in “The ASMBS Textbook of Bariatric Surgery, second Edition”. The accuracy of the generated answers, the number of correct answers based on subject matter, and the number of

Table 1
Accuracy of provided answers by different large language models

AI model	Correct n (%)	*P*	Incorrect n (%)	*P*	Randomly chosen n (%)	Unable to answer n (%)
Chat-GPT4	166 (83.0%)	<.001	34 (17.0%)	<.001	5 (2.5%)	0
Bard	152 (76.0%)		48 (24.0%)		0	5 (2.5%)
Bing	131 (65.5%)		69 (34.5%)		0	1 (.5%)

AI = artificial intelligence.

Table 2
Accuracy of provided answers by question subject types

AI model	Case scenario n (%) n = 20	Studies/policy n (%) n = 37	Treatment and surgical procedures n (%) n = 77	Complications/ adverse events n (%) n = 24	Biochemistry/ pharmaceutical n (%) n = 13	Diagnosis/ evaluation n (%) n = 11	Definitions n (%) n = 8	Epidemiologic/ socioeconomic n (%) n = 10
Chat-GPT	14 (70.0%)	29 (78.3)	64 (83.1%)	22 (91.7%)	11 (84.6%)	9 (81.8%)	8 (100.0%)	9 (90.0%)
Bard	12 (60.0%) 2 unable to answer	27 (73.0%) 1 unable to answer	60 (77.9%) 2 unable to answer	21 (87.5%)	11 (84.6%) 1 unable to answer	5 (45.4%)	8 (100.0%)	8 (80.0%)
Bing	10 (50.0%)	25 (67.6%)	48 (63.3%) 1 unable to answer	15 (62.5%)	12 (92.3%) 1 unable to answer	7 (63.6%)	6 (75.0%)	8 (80.0%)
P value	.610	.495	.012	.022	.795	.385	.545	.725

AI = artificial intelligence.

correct answers based on question type were assessed. The accuracy of generated answers was separated into the following subgroups: correct, incorrect, randomly chosen (as stated by LLM), and unable to answer. A separate subgroup analysis was also conducted based on the content of the question. These subgroups include case scenario (e.g., questions asking how to proceed after presenting a scenario of a patient and their condition); studies/policy (e.g., questions regarding guidance provided by organizations or the government, notable studies surrounding bariatric surgery, as well as questions regarding regulations or programs); treatment and surgical procedures (e.g., questions involving outcomes following different treatments as well as specific surgical techniques); complications/adverse events; biochemistry/pharmaceutical (e.g., questions regarding hormones/biochemistry pathways and medications); diagnosis/evaluation (e.g., questions about when to use a diagnostic test or how to follow-up on a patient with a certain condition); definitions; and epidemiologic/socioeconomic (e.g., questions that give a statistical value related to public health or asked about race/ethnicity) (Appendix 4). Further, the questions were categorized based on the question types, such as inclusionary (i.e., identifying the correct option from multiple choice), exclusionary (i.e., identifying the wrong option from multiple choice), and true or false.

Statistical analysis

All statistical analyses were performed using STATA (StataCorp version 17). Data are shown as proportion of correct answers per LLMs and per question category in percentages (%). Dichotomous variables were compared using the 2 × 3 chi-square test. All statistical tests were 2-sided with the threshold for significance set at $P < .05$ and 95% confidence intervals (CIs) were provided where applicable.

Results

From a total of 200 questions, there were 20 (10%) case scenario, 37 (18.5%) studies/policy, 77 (38.5%) treatment and surgical procedures, 24 (12%) complications/adverse events, 13 (6.5%) biochemistry/pharmaceutical, 11 (5.5%) diagnosis/evaluation, 8 (4%) definitions/evaluation, and 10 (2%) epidemiologic/socioeconomic questions. The questions were also categorized into 162 (81%) inclusionary, 31 (15.5%) exclusionary, and 7 (3.5%) true or false question types.

ChatGPT-4 correctly answered 166 (83.0%), whereas Bard correctly answered 152 (76.0%) and Bing correctly answered 131 (65.5%) questions ($P < .001$). When the chatbots could not provide an initial answer and were forced to randomly select one with the follow-up prompt “please pick an answer,” only ChatGPT-4 was able to do so, with a proportion of 5 (2.5%) randomly selected answers (Table 1).

Within the subgroup analysis based on different question types, ChatGPT-4 had the highest proportion of correct

Table 3
Accuracy of provided answers by question types

AI model	Inclusionary, n (%) n = 162	Exclusionary, n (%) n = 31	True or false, n (%) n = 7
Chat-GPT4	134 (82.7%)	27 (87.1%)	5 (71.4%)
Bard	123 (75.9%) 6 unable to answer	22 (71.0%)	7 (100.0%)
Bing	106 (65.4%) 2 unable to answer	19 (61.3%)	5 (71.4%)
P value	.002	.069	.291

AI = artificial intelligence.

responses for treatment and surgical procedures (64/77, 83.1%) (*P* = .012), as well as complications/adverse events (22/24, 91.7%) (*P* = .022). In contrast, Bing presented the lowest proportion of correct answers for the aforementioned subject categories.

Both ChatGPT-4 and Bard had the highest proportion of correct responses for definitions/evaluations and answered all of these questions correctly. In addition, Bard was unable to answer the following questions: 2 case scenarios, 1 studies/policy, 2 treatment and surgical procedures, and 1 biochemistry/pharmaceutical. Bing was unable to answer 1 treatment and surgical procedures question and 1 biochemistry/pharmaceutical question (Table 2).

Regarding question type, ChatGPT-4 ranked the highest for inclusionary questions, correctly answering 134 of 162 (82.7%) questions (*P* = .002). Within this question type category, Bing continued to display the lowest proportion of correct answers (Table 3).

Discussion

The current study is the first study evaluating the quality and reliability of 3 AI-based chat models in responding to textbook questions related to bariatric surgery. We found a statistically significant difference in the overall accuracy when comparing the 3 models, with ChatGPT-4 demonstrating the highest proportion of correct answers, followed by Bard, then Bing. While ChatGPT-4 provided answers to all questions, Bard and Bing could not answer certain questions. In terms of question categories, ChatGPT-4 demonstrated the highest number of correct responses in questions related to treatment/surgical procedures and complications/adverse events, as well as inclusionary-type questions.

These findings indicate that LLMs, particularly ChatGPT-4, exhibit promising performance in answering clinically important questions related to bariatric surgery. Despite its infancy, the application of LLMs in healthcare has quickly gained attention. Based on our findings, a potential application of LLMs could include marking and assessing surgical board exams. By training or refining the AI with numerous credible sources such as reputable articles and textbooks, these models can efficiently evaluate short-answer exam questions in a streamlined manner. In addition, these models could be employed to evaluate the clarity and fairness of exam questions. For instance, if a model trained with the same resources as surgical residents consistently struggled with a particular question, it may indicate an inherent fault within the question. Furthermore, these AI models may also assist in the generation of exam questions. Its ability to accurately answer a diverse range of questions suggests its potential in augmenting the learning process in the evaluation of surgical residents. Finally, LLMs can be used as a learning tool for residents studying for their board exams, by generating simulated clinical scenarios and practice questions followed by feedback.

The effectiveness and accuracy of LLMs in these applications will heavily rely on the quality and diversity of the training data [11]. Ongoing training and fine-tuning with a large volume of high-quality data on bariatric surgery will promote ongoing improvement in accuracy [11]. It is also worth mentioning that while AI models can process text-based questions, their limitations in contextual understanding may impede their ability to understand nuances in complex exam questions. Physician oversight will be required in all these processes to ensure standardization and fairness. Overall, our research demonstrates the accuracy of LLMs and their ability to answer questions across various subject matters and question types, thus highlighting their potential use as a reliable and efficient tool in enhancing surgical training and evaluation.

While the results are encouraging, it is essential to acknowledge certain limitations of the present study. Firstly, the evaluation was specifically focused on bariatric surgery and relied on a single textbook as the reference. Generalizability to other medical domains and resources may vary. Secondly, the study only evaluated 3 LLMs; other models or variations could yield different results. For instance, Bing utilizes GPT-4 as its base model. Next, as the LLMs models are currently unable to process visual information, no questions containing photos or videos were used. As such, conclusions could not be made on the ability of these models to accurately answer questions containing clinical images such as radiology and graphs. However, with their continued development future versions of LLMs may possess these abilities, which subsequent studies could examine. Third, we determined the "correct" answer based on the answer key provided by the ASMBS textbook and the relative accuracy of the answers to some questions could be debated depending on the specific details of the clinical scenario or future research evidence in bariatric and metabolic surgery. Finally, the results of AI analyses are limited by the quality and types of data available. Supervised learning relies on labeled data, which can be expensive to gather, and if done poorly can result in inaccurate data [12,13]. This highlights the importance of ongoing training and fine-tuning, leveraging a growing quantity of high-quality data and supervision by a diverse team of experts to ensure continued enhancement in accuracy [11].

Conclusion

In conclusion, this study demonstrates the potential of LLMs, particularly ChatGPT-4, in providing accurate and reliable responses to textbook questions in the field of bariatric surgery. Further advancements in AI technology and continued evaluation will contribute to the continued improvement and refinement of LLMs, paving the way for their wider integration into practice and education.

Disclosure

The authors have no commercial associations that might be a conflict of interest in relation to this article.

Supplementary data

Supplementary material associated with this article can be found, in the online version, at https://doi.org/10.1016/j.soard.2024.04.014.

References

[1] He J, Baxter SL, Xu J, et al. The practical implementation of artificial intelligence technologies in medicine. Nat Med 2019;25:30–6.

[2] Kung TH, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health 2023;2:e0000198.

[3] ChatGPT takes on thoracic surgery: a comparative analysis of AI performance on board exams - Khalpey AI Lab | Khalpey AI Lab. Available from: https://khalpey-ai.com/chatgpt-takes-on-thoracic-surgery-a-comparative-analysis-of-ai-performance-on-board-exams/. Accessed August 9, 2023.

[4] Azizoğlu M, Hani Okur M. How does ChatGPT perform on the European Board of Pediatric Surgery examination? A randomized comparative study. Research Square 2023. https://doi.org/10.21203/rs.3.rs-3018641/v1.

[5] Long C, Lowe K, dos Santos A, et al. Evaluating ChatGPT-4 in otolaryngology–head and neck surgery board examination using the CVSA model. medRxiv 2023. https://doi.org/10.1101/2023.05.30.23290758.

[6] Antaki F, Touma S, Milad D, et al. Evaluating the performance of ChatGPT in ophthalmology: an analysis of its successes and shortcomings. Ophthalmol Sci 2023;3:100324.

[7] Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. medRxiv 2023. https://doi.org/10.1101/2023.03.25.23287743.

[8] Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. medRxiv 2023. https://doi.org/10.1101/2023.04.06.23288265.

[9] Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res 2023;104:269–73.

[10] Nguyen NT, Brethauer SA, Morton JM, et al. The ASMBS Textbook of Bariatric Surgery. 2nd Ed, Springer, New York City, NY; 2020.

[11] Fine-tuning - OpenAI API. Available from: https://platform.openai.com/docs/guides/fine-tuning. Accessed August 9, 2023.

[12] Exploring the ChestXray14 dataset: problems – Lauren Oakden-Rayner. Available from: https://laurenoakdenrayner.com/2017/12/18/the-chestxray14-dataset-problems/. Accessed July 16, 2023.

[13] Hashimoto DA, Rosman G, Rus D, et al. Artificial intelligence in surgery: promises and perils. Ann Surg 2018;268:70–6.