ASTR

ORIGINAL ARTICLE

pISSN 2288-6575 • eISSN 2288-6796
https://doi.org/10.4174/astr.2023.104.5.269
Annals of Surgical Treatment and Research

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

Namkee Oh, Gyu-Seong Choi, Woo Yong Lee

Department of Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea

Purpose: This study aimed to assess the performance of ChatGPT, specifically the GPT-3.5 and GPT-4 models, in understanding complex surgical clinical information and its potential implications for surgical education and training.
Methods: The dataset comprised 280 questions from the Korean general surgery board exams conducted between 2020 and 2022. Both GPT-3.5 and GPT-4 models were evaluated, and their performances were compared using McNemar test.
Results: GPT-3.5 achieved an overall accuracy of 46.8%, while GPT-4 demonstrated a significant improvement with an overall accuracy of 76.4%, indicating a notable difference in performance between the models ($P < 0.001$). GPT-4 also exhibited consistent performance across all subspecialties, with accuracy rates ranging from 63.6% to 83.3%.
Conclusion: ChatGPT, particularly GPT-4, demonstrates a remarkable ability to understand complex surgical clinical information, achieving an accuracy rate of 76.4% on the Korean general surgery board exam. However, it is important to recognize the limitations of large language models and ensure that they are used in conjunction with human expertise and judgment.
[Ann Surg Treat Res 2023;104(5):269-273]

Key Words: Artificial intelligence, Continuing medical education, General surgery, Medical education

INTRODUCTION

Significant advancements in large language model (LLM) technology have recently revolutionized the field of artificial intelligence (AI), with ChatGPT released by OpenAI in November 2022 standing out as a prime example [1]. ChatGPT has exhibited exceptional performance in evaluating knowledge related to fields such as medicine, law, and management, which have traditionally been considered to be the domain of experts. Notably, the system achieved high accuracy on the United States Medical Licensing Examination, the Bar exam, and the Wharton MBA final exam, even without fine-tuning the pretrained model [2-5].

Surgical education and training demand a significant investment of time, with the process involving a combination of didactic learning, hands-on training, and supervised clinical experience [6]. During residency, surgical trainees work alongside experienced surgeons, gaining practical experience in patient care, surgery, and clinical decision-making. Additionally, trainees engage in a series of didactic courses and conferences covering the principles of surgery, medical knowledge, and surgical techniques. Due to the comprehensive nature of surgical education and training, it can take more than a decade to become a skilled and competent surgeon. Given the

Received March 23, 2023, **Revised** March 30, 2023,
Accepted March 30, 2023

Corresponding Author: Woo Yong Lee
Department of Surgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, 81 Irwon-ro, Gangnam-gu, Seoul 06351, Korea
Tel: +82-2-3410-0261, **Fax:** +82-2-3410-6980
E-mail: wooyong123.lee@samsung.com
ORCID: https://orcid.org/0000-0002-9558-9019

time-intensive nature of surgical education and training, it is important to explore how emerging technologies, such as AI and LLMs, can augment the learning process [7].

This study aims to employ ChatGPT to evaluate the general surgery board exam in Korea and assess whether LLMs possess expert-level knowledge. Moreover, the study compared the performance of GPT-3.5 and GPT-4. By exploring the potential of LLMs in the context of surgical education and training, this study seeks to provide a foundation for future research on how these advancements can be effectively integrated into clinical education and practice, ultimately benefiting surgical residents, and practicing surgeons.

METHODS

The study did not involve human subjects and did not require Institutional Review Board approval.

General surgery board exam of Korea

The goal of surgical education and training is to develop the ability to actively evaluate the pathological conditions of surgical diseases and to acquire the surgical skills to treat traumatic, congenital, acquired, neoplastic, and infectious surgical diseases. To quantitatively evaluate this knowledge and skill set of surgical residents, a board certification exam is required after completion of their training, in order to become a board-certified general surgeon in Korea. The exam is composed of 2 parts: the first part is a 200-question multiple-choice test, and those who pass the first part are eligible to take the second part. The second part consists of questions based on high-resolution clinical images and surgical video clips. The questions are created and supervised by the Korean Surgical Society and the Korean Academy of Medical Science (KAMS).

Dataset for model testing

The actual board exam questions are held by KAMS, but due to limited access to the usage of these questions, we constructed our dataset by gathering questions recalled by examinees who took the actual exam. As the LLM cannot process visual information such as clinical images, radiology, and graphs, questions that included visual information were excluded from our dataset. All problems were manually inputted in their original Korean text. Finally, our dataset included a total of 280 questions from the first stage of the board exam in 2020, 2021, and 2022 (Fig. 1A).

Large language model and performance evaluation

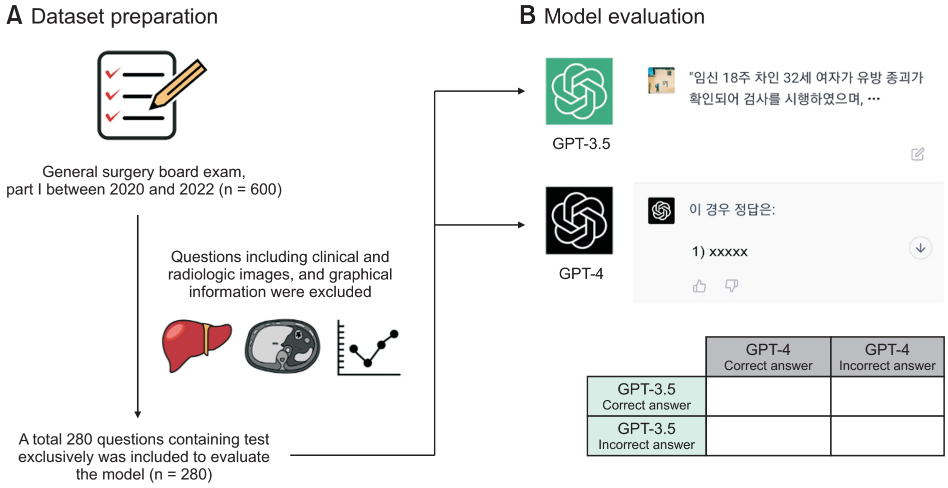
In this study, we utilized the ChatGPT generative pretrained transformer (GPT) language model developed by OpenAI to evaluate its performance on a dataset of questions. We performed model testing using both GPT-3.5 and GPT-4, with the former conducted from March 1 to March 3, 2023, and the latter scheduled for March 15, 2023. To evaluate the model's performance, we manually entered the questions into the ChatGPT website and compared the answers provided by the model to those provided by examinees (Fig. 1B).

Statistical analysis

This study compared the performance of the GPT-3.5 and GPT-4 models with the McNemar test. A P-value less than 0.05 would indicate a statistically significant difference between the performance of the GPT-3.5 and GPT-4 (IBM SPSS Statistics ver. 27; IBM Corp.).

RESULTS

The dataset used for model evaluation consisted of a total of 280 questions, which were classified into subspecialties and listed in order of frequency as follows: endocrine (16.8%), breast


A Dataset preparation
General surgery board exam, part I between 2020 and 2022 (n = 600)
Questions including clinical and radiologic images, and graphical information were excluded
A total 280 questions containing test exclusively was included to evaluate the model (n = 280)
B Model evaluation
GPT-3.5
GPT-4

"임신 18주 차인 32세 여자가 유방 종괴가 확인되어 검사를 시행하였으며, ⋯

이 경우 정답은:

1) xxxxx

	GPT-4 Correct answer	GPT-4 Incorrect answer
GPT-3.5 Correct answer		
GPT-3.5 Incorrect answer		

Fig. 1. (A) Dataset preparation process for model evaluation. (B) How models were evaluated from ChatGPT website.

(16.1%), lower gastrointestinal (LGI, 14.3%), upper gastrointestinal (UGI, 13.2%), general (13.2%), pediatric (6.4%), hepatobiliary and pancreas (HBP, 6.1%), vascular (6.1%), transplantation (4.0%), and trauma and critical care (4.0%) (Fig. 2).

A significant difference in performance was observed between the GPT-3.5 and GPT-4 models ($P < 0.001$). The GPT-3.5 model achieved an overall accuracy of 46.8%, providing correct answers for 131 out of the 280 questions (Table 1). In terms of individual subspecialties, the model's accuracy rates were as follows (sorted from highest to lowest): transplantation (72.7%), breast (62.2%), HBP (52.9%), general (48.6%), UGI (45.9%), trauma and critical care (45.5%), LGI (45.0%), endocrine (36.2%), pediatric (33.3%), and vascular (29.4%). In contrast, the GPT-4 model demonstrated a substantial improvement in overall accuracy, attaining a rate of 76.4% by providing correct answers for 214 out of the 280 questions. The accuracy rates for each subspecialty were as follows: pediatric (83.3%), breast (82.2%), UGI (81.1%), endocrine (78.7%), general (75.7%), transplantation (72.7%), LGI (72.5%), vascular (70.6%), HBP (64.7%), and trauma and critical care (63.6%) (Fig. 3).

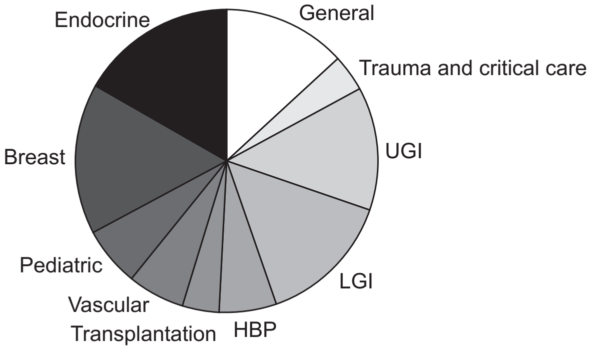


Fig. 2. The dataset was composed of 280 questions, and it is classified into subspecialties in the field of general surgery. HBP, hepatobiliary and pancreas; LGI, lower gastrointestinal; UGI, upper gastrointestinal.

DISCUSSION

The primary objective of this study was to conduct a quantitative assessment of ChatGPT's ability to comprehend complex surgical clinical information and to explore the potential implications of LLM technology for surgical education and training. Specifically, we tested the performance of ChatGPT using questions from the Korean general surgery board exam and observed that the model achieved an accuracy of 76.4% with GPT-4 and 46.8% with GPT-3.5. Remarkably, this accuracy was achieved without fine-tuning the model and by using prompts in the Korean language exclusively, thus highlighting the significance of our findings.

Table 1. Comparison table for the accuracy of GPT-3.5 and GPT-4

Variable	GPT-3.5	GPT-4
Correct answer	131	214
Incorrect answer	149	66
Accuracy	46.8% (131/280)	76.4% (214/66)

GPT, generative pretrained transformer.

Table 2. A 2-by-2 contingency table summarizing performance of GPT-3.5 and GPT-4

Variable	GPT-4 correct answer	GPT-4 incorrect answer	Total
GPT-3.5 correct answer	113	18	131
GPT-3.5 incorrect answer	101	48	149

GPT, generative pretrained transformer.

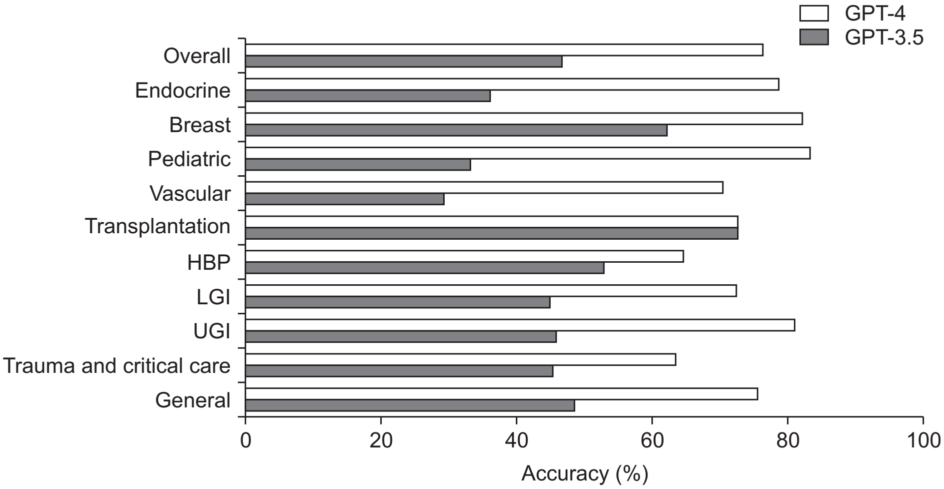


Fig. 3. Comparison of the performance of GPT-4 and GPT-3.5 with overall accuracy and accuracies according to its subspecialties. GPT, generative pretrained transformer; HBP, hepatobiliary and pancreas; LGI, lower gastrointestinal; UGI, upper gastrointestinal

The comparative analysis revealed a notable improvement in GPT-4's performance compared to GPT-3.5 model across all subspecialties. GPT-4 not only exhibited a higher overall accuracy rate but also demonstrated more consistent performance in each subspecialty, with accuracy rates ranging from 63.6% to 83.3%. However, for 18 questions, GPT-3.5 provided the correct answer while GPT-4 did not (Table 2). It is unclear why GPT-4 gave incorrect answers for these questions despite the overall increase in accuracy. Pinpointing the exact reason for this discrepancy is challenging. Differences in training data, model architecture, or other factors could have contributed to the variation in the performance between the 2 versions.

The authors kindly recommend that the surgeon's society proactively adapts and utilizes these technological advancements to enhance patient safety and improve the quality of surgical care. In the context of surgical education, it is crucial to transition from the traditional rote learning approach to a method that emphasizes problem definition in specific clinical situations and the acquisition of relevant clinical information for problem resolution. LLMs serve as generative AI models, providing answers to given problems. Consequently, the quality of the answers relies on the questions posed [8]. Surgeons must conduct thorough history-taking and physical examinations to accurately define the problems they face. By providing LLMs with comprehensive summaries of patients' chief complaints, present illnesses, and physical examinations, the models have the potential to assist in decision-making regarding diagnostic tests and treatment options in certain clinical situations. However, it is essential for medical professionals to remember that LLMs should not replace the fundamentals of patient care, which include maintaining close connections with patients and actively listening to their concerns [9].

Moreover, active surgeons who completed their training over a decade ago may find LLMs helpful for continuous medical education (CME). Accessing new knowledge may be challenging for them due to the time elapsed since their training, potentially leading to outdated management practices. While numerous surgical societies offer CME programs, altering ingrained routines in clinical practice can be difficult. By utilizing an up-to-date LLM as a supplementary resource in their decision-making process, surgeons may have additional means to stay informed and strive for evidence-based care in their patient management [10].

In medicine, decision-making has a profound impact on patient safety, demanding a higher level of accuracy and a conservative approach to change compared to other fields. Although GPT-4 achieved a 76.4% accuracy rate on the Korean surgical board exam, it is important to remember that LLMs are generative models, often referred to as "stochastic parrots" [11]. Instead of providing strictly accurate information, they generate responses based on the probability of the most appropriate words given the data they have been trained on. Consequently, the current level of accuracy is not yet sufficient for immediate clinical application in patient care.

However, it is noteworthy that a service released less than 6 months ago exhibits such remarkable performance, and ChatGPT is only one example of LLMs. Recently, Microsoft released BioGPT, an LLM trained on PubMed literature, and Meta introduced LLaMA, an LLM with an accessible application programming interface (API) for open innovation and fine-tuning [12,13]. Based on these trends, we can anticipate future LLMs to be trained on an even larger and more diverse set of medical information, providing specialized knowledge in the medical field. In addition, the GPT-4 framework itself is capable of processing and analyzing visual information, including images and videos [14]. This capability raises the possibility that, in the future, the performance of GPT-4 could be evaluated on datasets containing clinical photos and surgical videos. Such advancements would further enhance the applicability of GPT-4 in surgical fields, broadening its utility beyond text-based tasks and offering a more comprehensive understanding of complex clinical scenarios assisting professionals in their decision-making processes and contributing to improved patient care.

The limitations of this study include the fact that the dataset was compiled using questions recalled by examinees, which may not accurately represent the full set of actual board exam questions due to restricted access. Another limitation is the exclusion of visual information. Since the models used in the study are unable to process visual information, such as clinical images, radiology, and graphs, questions containing visual components were excluded from the dataset. As a result, we cannot determine whether ChatGPT would pass or fail the board exam based on these limitations. Despite these constraints, this study holds significance as it confirms the ability of LLMs to analyze surgical clinical information and make appropriate clinical decisions.

In conclusion, ChatGPT, particularly GPT-4, demonstrates a remarkable ability to understand complex surgical clinical information, achieving an accuracy rate of 76.4% on the Korean general surgery board exam. However, it is important to recognize the limitations of LLMs and ensure that they are used in conjunction with human expertise and judgment.

ACKNOWLEDGEMENTS

The authors would like to thank Da Hyun Lee, an audiovisual engineer at Samsung Medical Information & Medical Services, for designing Fig. 1 for this work.

Fund/Grant Support

This study was supported by Future Medicine 2030 Project of the Samsung Medical Center (No. SMX1230771).

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

ORCID iD

Namkee Oh: https://orcid.org/0000-0002-6594-8973
Gyu-Seong Choi: https://orcid.org/0000-0003-2545-3105
Woo Yong Lee: https://orcid.org/0000-0002-9558-9019

REFERENCES

1. OpenAI. Introducing ChatGPT [Internet]. OpenAI; c2015-2023 [cited 2023 Feb 10]. Available from: https://openai.com/blog/chatgpt
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:e0000198.
3. 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:e0000205.
4. Bommarito II M, Katz DM. GPT takes the Bar Exam [Preprint]. Posted online 2022 Dec 29. arXiv:2212.14402. Available from: https://doi.org/10.48550/arXiv.2212.14402
5. Choi JH, Hickman KE, Monahan A, Schwarcz DB. Chatgpt goes to law school. Minnesota Legal Studies Research Paper No. 23-03 [Internet]. SSRN; 2023 [cited 2023 Feb 10]. Available from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4335905
6. Debas HT, Bass BL, Brennan MF, Flynn TC, Folse JR, Freischlag JA, et al. American Surgical Association Blue Ribbon Committee Report on Surgical Education: 2004. Ann Surg 2005;241:1-8.
7. Wartman SA, Combs CD. Medical education must move from the information age to the age of artificial intelligence. Acad Med 2018;93:1107-9.
8. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training [Internet]. OpenAI; 2018 [cited 2023 Feb 10]. Available from: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
9. Kapadia MR, Kieran K. Being affable, available, and able is not enough: prioritizing surgeon-patient communication. JAMA Surg 2020;155:277-8.
10. Han ER, Yeo S, Kim MJ, Lee YH, Park KH, Roh H. Medical education trends for future physicians in the era of advanced technology and artificial intelligence: an integrative review. BMC Med Educ 2019;19:460.
11. Bender EM, Gebru T, Mcmillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? In: FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; 2021 Mar 3-10. p. 610-23. Available from: https://doi.org/10.1145/3442188.3445922
12. Luo R, Sun L, Xia Y, Qin T, Zhang S, Poon H, et al. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Brief Bioinform 2022;23:bbac409.
13. Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, et al. Llama: Open and efficient foundation language models. arXiv [Preprint] 2023 Feb 27. https://doi.org/10.48550/arXiv.2302.13971
14. OpenAI. GPT-4 technical report [Internet]. Open AI; 2023 [cited 2023 Feb 10]. Available from: https://cdn.openai.com/papers/gpt-4.pdf