

A Comparison Between GPT-3.5, GPT-4, and GPT-4V: Can the Large Language Model (ChatGPT) Pass the Japanese Board of Orthopaedic Surgery Examination?

Review began 02/03/2024

Review ended 02/15/2024

Published 03/18/2024

© Copyright 2024

Nakajima et al. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Nozomu Nakajima¹, Takahito Fujimori², Masayuki Furuya², Yuya Kanie², Hirotatsu Imai², Kosuke Kita², Keisuke Uemura², Seiji Okada²

1. Orthopaedics, Sakai City Medical Center, Sakai, JPN 2. Orthopaedic Surgery, Osaka University, Graduate School of Medicine, Suita, JPN

Corresponding author: Takahito Fujimori, takahito-f@hotmail.co.jp

Abstract

Introduction

Recently, large-scale language models, such as ChatGPT (OpenAI, San Francisco, CA), have evolved. These models are designed to think and act like humans and possess a broad range of specialized knowledge. GPT-3.5 was reported to be at a level of passing the United States Medical Licensing Examination. Its capabilities continue to evolve, and in October 2023, GPT-4V became available as a model capable of image recognition. Therefore, it is important to know the current performance of these models because they will be soon incorporated into medical practice. We aimed to evaluate the performance of ChatGPT in the field of orthopedic surgery.

Methods

We used three years' worth of Japanese Board of Orthopaedic Surgery Examinations (JBOSE) conducted in 2021, 2022, and 2023. Questions and their multiple-choice answers were used in their original Japanese form, as was the official examination rubric. We inputted these questions into three versions of ChatGPT: GPT-3.5, GPT-4, and GPT-4V. For image-based questions, we inputted only textual statements for GPT-3.5 and GPT-4, and both image and textual statements for GPT-4V. As the minimum scoring rate acquired to pass is not officially disclosed, it was calculated using publicly available data.

Results

The estimated minimum scoring rate acquired to pass was calculated as 50.1% (43.7-53.8%). For GPT-4, even when answering all questions, including the image-based ones, the percentage of correct answers was 59% (55-61%) and GPT-4 was able to achieve the passing line. When excluding image-based questions, the score reached 67% (63-73%). For GPT-3.5, the percentage was limited to 30% (28-32%), and this version could not pass the examination. There was a significant difference in the performance between GPT-4 and GPT-3.5 ($p < 0.001$). For image-based questions, the percentage of correct answers was 25% in GPT-3.5, 38% in GPT-4, and 38% in GPT-4V. There was no significant difference in the performance for image-based questions between GPT-4 and GPT-4V.

Conclusions

ChatGPT had enough performance to pass the orthopedic specialist examination. After adding further training data such as images, ChatGPT is expected to be applied to the orthopedics field.

Categories: Orthopedics

Keywords: japanese board of orthopaedic surgery examination, gpt-4v, chatgpt, large language model, artificial intelligence

Introduction

Machines have long been valuable partners to humanity. In our daily lives, we are surrounded by many machines. Modern artificial intelligence (AI), which is attracting significant attention these days, is designed to think and act like humans [1-3]. In the 1970s, attention was drawn to expert systems that aimed to replicate professional advice by teaching machines vast amounts of specialized knowledge [4,5]. A prime example, Mycin, could suggest appropriate antibiotics based on the entered information [6-8]. However, these systems required extensive manual data entry by experts, and due to their rigid rule-based structure, they had limited application [1,9]. Nevertheless, these limitations are being overcome with advancements in machine learning technologies. Introduced by OpenAI (San Francisco, CA) in October 2022, ChatGPT (GPT-3.5) is a prominent large language model. It can instantly generate text based on the vast knowledge

How to cite this article

Nakajima N, Fujimori T, Furuya M, et al. (March 18, 2024) A Comparison Between GPT-3.5, GPT-4, and GPT-4V: Can the Large Language Model (ChatGPT) Pass the Japanese Board of Orthopaedic Surgery Examination?. Cureus 16(3): e56402. DOI 10.7759/cureus.56402

acquired from web data and textbooks. It supports multiple languages and is characterized by its conversational skills, making interactions feel like one is speaking to a human [10].

Since GPT-3.5 was released, it has attracted significant attention worldwide for its usefulness, completeness, and high accuracy. As of October 2023, a search for “Large Language Model” in PubMed has yielded 8182 hits, indicating a high level of attention in the medical field. GPT-3.5 was reported to be at a level of passing the United States Medical Licensing Examination (USMLE) [11]. Furthermore, GPT-4, an upgraded model of GPT-3.5, was released in March 2023. Its accuracy was in the top 10% of the US bar exam (compared with the bottom 10% for the GPT-3.5) [10]. This AI is not only capable of natural conversation but also possesses a broad range of specialized knowledge. It is inevitable that, in the future, it will be utilized in the medical field as an AI that can provide expert advice.

In the field of orthopedics, GPT-4 corresponded to the average performance of post-graduate year (PGY)-5 in the Orthopedic In-Training Examination and exceeded the passing score of the American Board of Orthopedic Surgery Part 1 Examination, while GPT-3.5 only corresponded to a PGY-1 level [12]. However, these evaluations were conducted in English, and their performance inputted in Japanese has not yet been assessed. Furthermore, in October 2023, GPT-4V became available, enabling bimodal input through both text and images. No studies have attempted image-based questions using GPT-4V. This study aimed to evaluate the performance of ChatGPT, one of the representative large-scale language models, on the Japanese Board of Orthopaedic Surgery Examination (JBOSE).

Materials And Methods

The Japanese Board of Orthopaedic Surgery Examination (JBOSE)

This examination is held once a year, consists of 100 questions, and tests a wide range of knowledge related to orthopedics from basic medicine, such as bone metabolism and pathology, to specialized fields, such as conservative therapy, surgical treatment, rehabilitation, and medical insurance. After passing the National Medical Examination in Japan, the applicant must complete two years of initial clinical training, followed by three years and nine months of specialized orthopedic surgery training to be eligible for the examination. The time limit is 120 minutes, and the questions are presented in computer-based testing (CBT) format. Each question consists of a statement and five options. There are three types of questions requiring the selection of one, two, or three correct answers. JBOSE 33rd, 34th, and 35th, conducted in 2021, 2022, and 2023, were used to evaluate the performance of ChatGPT. We focused on the examinations from the past three years because the format changed after the 33rd session, shifting from a combination of written and interview examinations to the CBT format. The questions and correct answers were obtained from the Japanese Orthopaedic Association journal [13-15]. Except for a few questions that were officially announced to be incomplete, all questions were inputted to ChatGPT (Figure 1).

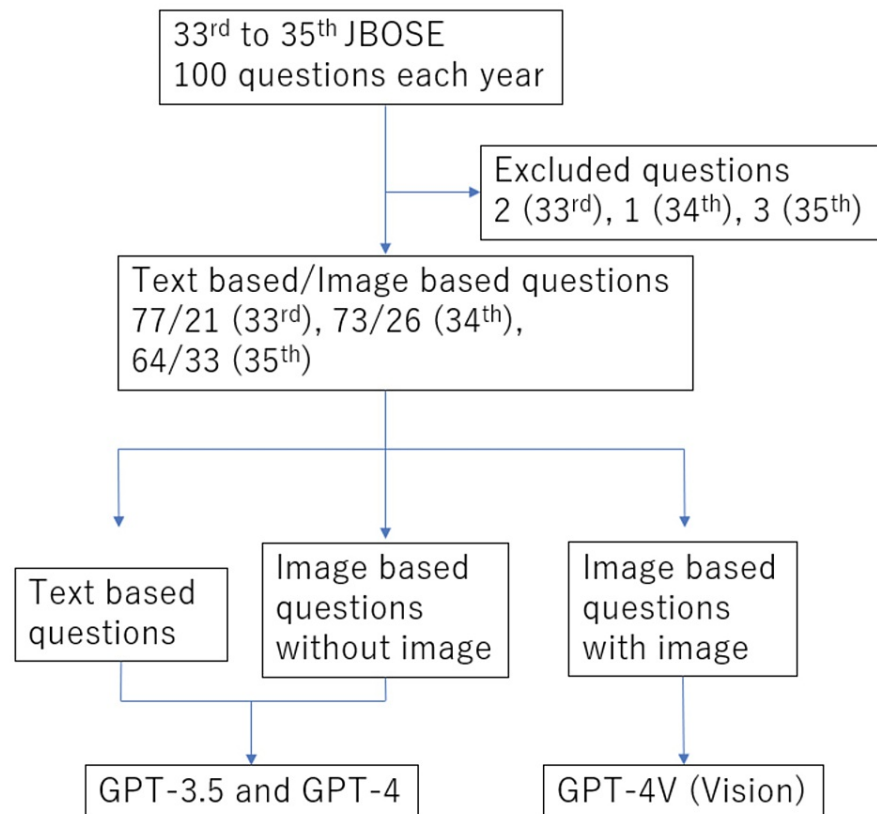


FIGURE 1: Workflow of this study

The Japanese Board of Orthopaedic Surgery Examination (JBOSE) included text-based and image-based questions. A few questions were officially excluded after the examination because of their incompleteness. GPT-4V enables bimodal input through both text and images. The 33rd, 34th, and 35th are ordinal numbers in exams, and they were conducted in 2021, 2022, and 2023, respectively.

JBOSE includes text-based and image-based questions. Image-based questions include both textual statements and images. Although image information is helpful in predicting the correct answers, it is possible to guess the answer from the textual statements alone, even without images. Because GPT-3.5 and GPT-4 only accept textual input, image-based questions were inputted only as textual statements.

Large language model

The models used in this study were GPT-3.5 (standard model), GPT-4 (high-grade model), and GPT-4V (text and image input model), all of which were large language models developed by OpenAI (San Francisco, CA). While GPT-3.5 was available for free, GPT-4 and GPT-4V required a monthly fee of US\$20 as of January 2024. These models generated text based on vast knowledge across various fields and were characterized by their conversational skills, making interactions similar to those of humans. We examined its knowledge in the orthopedic field using JBOSE. The GPT-4V model became available in October 2023. GPT-4V is capable of recognizing images as well as text, therefore we added a verification step to confirm GPT-4V's performance in recognizing medical images.

Data input

Questions and their multiple-choice answers from the JBOSE were used in their original Japanese form, as was the official examination rubric. Instructions for using ChatGPT were also provided in Japanese. Because GPT-3.5 and GPT-4 were not able to recognize images, we input text data only. In other words, for image-based questions, GPT3.5 and GPT4 answered without images, as a reference, only text data for the questions. We employed prompt engineering techniques to ensure consistency in the interaction and emulation of a clinical examination setting. The following statement was instructed, and then each question was entered one by one: "You are an orthopedic specialist. We will present a test, so please answer the following questions. Make sure to answer each question carefully without any mistakes." This approach was designed to prime ChatGPT with a specific role and mindset as an orthopedic specialist. When multiple questions were presented at once, there was a risk of information overload and scattered attention to specific questions. Therefore, in each interaction, we inputted questions individually.

Evaluation

To determine whether GPT-3.5 and GPT-4 could pass the orthopedic specialist examination, the accuracy was calculated separately for the entire set of questions and the text-based questions. Accuracy was defined as the percentage of correct answers. As the minimum scoring rate acquired to pass is not officially disclosed, it was determined using publicly available data (total examinees, number of passers, average score, and standard deviation) using the following formula: passing score = $Z \times \text{standard deviation} + \text{average score}$. Z-score corresponds to the value derived from the standard normal distribution table for “1 – number of passers/total examinees.”

Modified question

To further assess the accuracy of GPT-4, we introduced a “Modified Question” format to the text-based questions. In this approach, the number of correct options was concealed during the question. In other words, the model was instructed to respond without knowing how many correct answers were present among the options. For instance, a question like “Choose the three correct options” would be posed as “Select all the correct options.” The model was assessed on its ability to judge all given options correctly. We presented GPT-4 with modified questions created from the text-based questions of the 33rd to 35th JBOSE and examined its performance and reproducibility. “Reproducibility” was defined as the proportion of questions that were answered correctly in the modified questions out of those that were correctly answered in the conventional questions.

Statistical analysis

Statistical analysis was conducted using Welch’s t-test to compare the performance of GPT-3.5 and GPT-4. All statistical tests were two-tailed, and a p-value of less than 0.05 was considered statistically significant. Statistical analysis was performed using SPSS Statistics version 20 software (IBM Corp., Armonk, NY).

Results

Demographic data

The overview of the examination is shown in Table 1.

JBOSE	33rd	34th	35th
Number of examinees (passers/total)	1021/1048	603/653	530/591
Passing rate among all examinees (%)	97	92	90
Number of questions (total/text-based/image-based)	98/77/21	99/73/26	97/64/33
Average score (%)	61	66	64
Standard deviation (%)	9	8.5	8.9
Estimated minimum scoring rate acquired to pass (%)	43.7	53.8	52.7

TABLE 1: The overview of the 33rd-35th Japanese Board of Orthopaedic Surgery Examination (JBOSE) and the performance of GPT-3.5, GPT-4, and GPT-4V

The estimated minimum scoring rate acquired to pass was calculated as 43.7% for the 33rd, 53.8% for the 34th, and 52.7% for the 35th examination. The pass rate among all examinees is typically around 90% each year. However, the 33rd exam, being the first to adopt the CBT format, saw a higher pass rate of 97%, exceeding the usual average. Subsequently, the pass rates for the 34th and 35th exams were 92% and 90%, respectively. There were 77 (79%) text-based questions and 21 (21%) image-based questions in the 33rd examination, 73 (74%) and 26 (26%) in the 34th examination, and 64 (66%) and 33 (34%) in the 35th examination, respectively.

Performance of each GPT model

For GPT-4, even when answering all questions, including the image-based ones, the accuracy was 60%, 55%, and 61% for each examination. When excluding image-based questions, it was 64%, 63%, and 73%, which was comparable with the average score of examinees. In both cases, the scores achieved the pass line. The accuracy for image-based questions was lower than that for text-based questions. For GPT-4V, the accuracy in image-based questions was 38%, 35%, and 39%, which surpassed that of GPT-4 in the 34th (31%) and 35th (36%) examinations but fell below GPT-4’s performance in the 33rd (48%) examination. For GPT-3.5, the

accuracy was limited to 28%, 32%, and 30%, which was unable to achieve a pass in any of the years (Table 2).

	JBOSE	33rd	34th	35th
		Number of correct answers/total questions, accuracy (%)		
GPT-4	All questions	59/98, (60)	54/99, (55)	59/97, (61)
	Text-based questions	49/77, (64)	46/73, (63)	47/64, (73)
	Image-based questions without image	10/21, (48)	8/26, (31)	12/33, (36)
GPT-4V	Image-based questions with image	8/21, (38)	9/26, (35)	13/33, (39)
GPT-3.5	All questions	27/98, (28)	32/99, (32)	29/97, (30)
	Text-based questions	20/77, (26)	25/73, (34)	23/64, (36)
	Image-based questions without image	7/21, (33)	7/26, (27)	6/33, (18)

TABLE 2: The accuracy of GPT-3.5, GPT-4, and GPT-4V for the 33rd-35th Japanese Board of Orthopaedic Surgery Examination (JBOSE)

The accuracy was defined as the percentage of correct answers. All questions include both text-based and image-based questions.

For GPT-4, the average accuracies were 59% for all questions, 66% for text-based questions, and 38% for image-based questions. In comparison, GPT-3.5 achieved 30% for all questions, 32% for text-based questions, and 25% for image-based questions. A significant difference was observed between GPT-4 and GPT-3.5 in text-based questions ($p = 0.002$; Table 3).

	GPT-4	GPT-3.5	GPT-4V	p-value
	Number of correct answers/total questions, accuracy (%)			
All questions	172/294, (59)	88/294, (30)	N.A.	<0.001 ^{*§}
Text-based questions	142/214, (66)	68/214, (32)	N.A.	0.002 ^{*§}
Image-based questions without image	30/80, (38)	20/80, (25)	N.A.	0.1 [§]
Image-based questions with image	N.A.	N.A.	30/80, (38)	0.9 [†]

TABLE 3: Average of three-year total accuracy of GPT-4, GPT-3.5, and GPT-4V for all questions, text-based questions, and image-based questions

The accuracy was defined as the percentage of correct answers. All questions include both text-based and image-based questions.

* $p < 0.05$. † p-value was calculated between GPT-4 and GPT-4V. § p-value was calculated between GPT-4 and GPT-3.5. N.A.: not available.

Performance of GPT-4 for modified question

We made ChatGPT answer the modified questions that had no information about the number of correct options. Modified questions were made for all text-based questions in the 33rd to 35th examination ($n = 214$). Out of the questions correctly answered in the conventional questions ($n = 142$), 79 were also correctly answered in the modified questions, resulting in a reproducibility of 56%. Additionally, there were a few questions ($n = 6$) answered correctly in the modified questions, but answered incorrectly in the conventional questions (Figure 2).

All text-based questions in the 33rd to 35th examination (n=214)

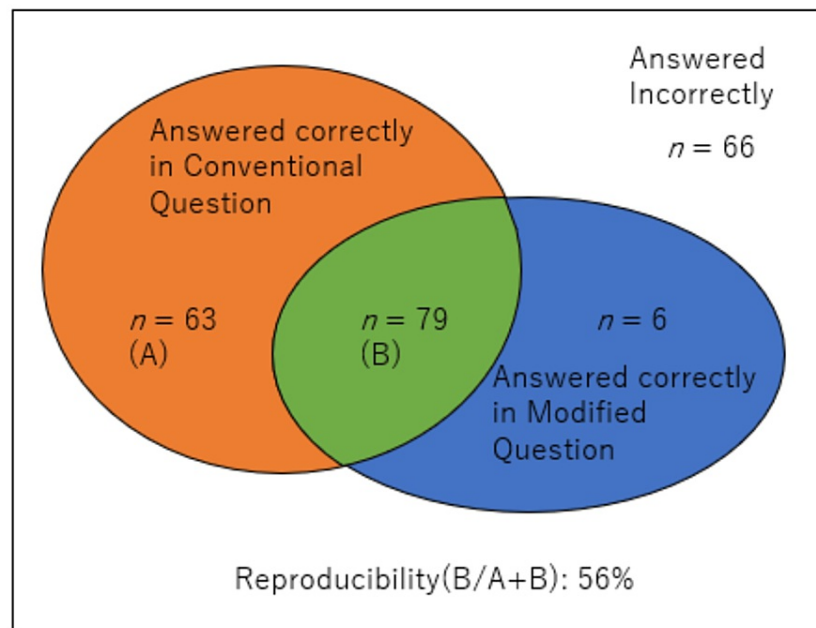


FIGURE 2: GPT-4's performance on the modified question and its reproducibility

To further assess the performance of GPT-4, a "Modified Question" was made for all text-based questions in the 33rd-35th examination (n = 214), concealing the number of correct options. Reproducibility was defined as the proportion of questions answered correctly in the modified question format (A) out of those correctly answered in the conventional question format (A+B).

Discussion

We evaluated ChatGPT's performance in the orthopedic field using the JBOSE. While GPT-3.5 did not reach the passing line for the exam, GPT-4 achieved the passing line despite including image-based questions. Although GPT-4 could not recognize the image, the GPT-4's accuracy for the image-based questions reached 38%. This fact suggests that GPT-4 may have advanced reasoning skills. It has been reported that GPT-4's advancements were not merely due to increased training data; it can now interpret complex texts and grasp nuanced differences [10]. We think that GPT-4 was able to deduce correct answers even from incomplete questions using its clinical reasoning skills. The percentage of correct answers on GPT-4 for all questions was categorized by subject type (Figure 3).

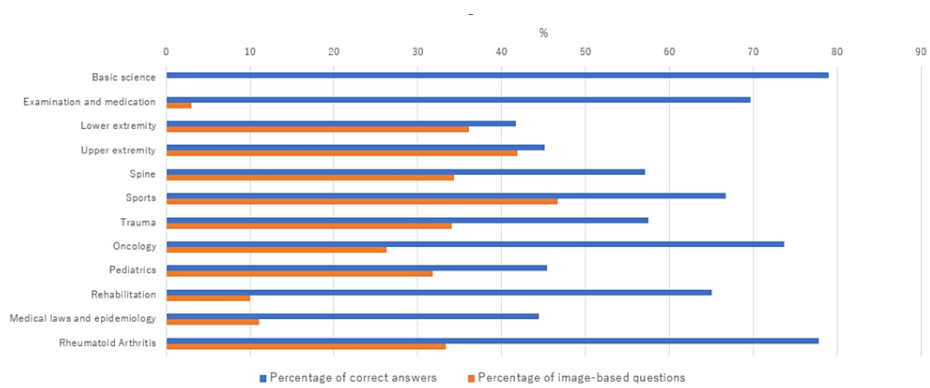


FIGURE 3: Percentage of correct answers and image-based questions categorized by subject type

The percentage of correct answers on GPT-4 for all questions was categorized by subject type. There was a tendency for the accuracy rate to increase with a lower proportion of image-based questions.

While the accuracy rate was higher for the basic science, which asked for simple knowledge, it was lower for questions involving image-based questions.

Since October 2023, GPT-4V has become available and features image recognition capabilities. However, the accuracy of image-based questions showed no improvement. There was a tendency for questions incorrectly answered by GPT-4 to also be answered incorrectly by GPT-4V. Specific examples of responses by GPT-4 and GPT-4V to image-based questions are shown in Figure 4.

(A)

Question 33:
A 4-year-old girl. She became unable to walk due to pain in her right lower leg without any obvious cause. Plain X-rays of both lower legs are provided. What is the most likely diagnosis?
a. Marble bone disease
b. Osteogenesis imperfecta
c. Child abuse syndrome
d. Congenital pseudarthrosis of the tibia
e. Congenital tibial dysplasia

(The correct answer is option d.)

(B)

GPT-4:
Without imaging or more clinical details, it is hard to make a conclusive diagnosis. However, based on the given description, if there are no obvious signs of bone deformity or fractures, **Child Abuse Syndrome (option c)** could be a consideration, particularly if other signs of abuse or neglect are present. This would warrant immediate further investigation.

(C)

GPT-4V:
From the image, there appears to be no abnormalities in both tibias, but there are findings suggestive of a fracture in the right fibula. In the case of a 4-year-old girl sustaining a fracture without a clear cause, the possibility of severe physical abuse should be considered. Therefore, the most likely diagnosis among the options is:
c. Child abuse syndrome.

FIGURE 4: The examples of responses of GPT-4 and GPT-4V to image-based questions.

Even if images were provided to GPT-4V, the responses were similar to GPT-4, in which no image was provided.

(A) An example of a question that includes both textual statements and images. X-ray images were omitted for copyright reasons. The textual statement was translated from Japanese.

(B) The answer obtained from GPT-4.

(C) The answer obtained from GPT-4V.

Image credits: Nozomu Nakajima.

While the presence of a fracture was recognized with the inclusion of an image, GPT-4V failed to identify the specific location of the fracture or the characteristic changes necessary for diagnosis. Consequently, the responses were similar to those given when no image was provided. We think that GPT-4V has been trained sufficiently for medical imaging due to the difficulties in acquiring extensive image data, given the concerns over personal privacy.

The utility of ChatGPT in the medical field is gathering attention, and several existing studies support this notion [11,12,16-21]. We summarized in Table 4 research compiling examinations that evaluated knowledge using ChatGPT.

	Author (year)	Exam	Language	Model	Result
General Medicine	Kung et al. (2023) [11]	United States Medical Licensing Examination	English	GPT-3.5	Pass
	Takagi et al. (2023) [16]	Japanese Medical Licensing Examination	Japanese	GPT-4	Pass
Orthopedics				GPT-4	Pass
	Kung et al. (2023) [12]	Orthopaedic In-Training Examination	English	GPT-3.5	Fail
	Saad et al. (2023) [17]	Orthopaedic Fellow of the Royal College of Surgeons	English	GPT-4	Fail
	Massey et al. (2023) [18]	ResStudy Orthopaedic Examination Question Bank	English	GPT-4	Fail
Neurosurgery				GPT-4	Pass
	Ali et al. (2023) [19]	Neurosurgery Written Board Examinations	English	GPT-3.5	Pass
Gastroenterology	Suchman et al. (2023) [20]	American College of Gastroenterology Self-Assessment Test	English	GPT-4	Fail
Radiology	Bhayana et al. (2023) [21]	Canadian Royal College and American Board of Radiology examinations	English	GPT-3.5	Pass

TABLE 4: The summary of existing studies compiling examinations that evaluate knowledge using ChatGPT

While there were slight variations in the prompts, we found that GPT-3.5 could pass the USMLE [11]. However, one of the notable limitations of GPT-3.5 was its lower accuracy in non-English languages [22], which was particularly evident in the Japanese medical licensing exam where GPT-3.5 could not achieve a pass level. In contrast, GPT-4 showed significant improvement, successfully passing the same exam [16]. This enhancement in GPT-4's performance in non-English languages marks a critical step in its applicability in global medical contexts. In studies of specialties, most of the exams were administered in English. It was reported that ChatGPT achieved the passing line in orthopedics, neurosurgery, and radiology examinations [12,19,21]. On the other hand, some reports reported that ChatGPT has not yet reached the level of a specialist [17,18,20]. Our study indicated that GPT-4 may have knowledge at the level of the Japanese license of an orthopedic specialist who has trained for five years and nine months after obtaining their medical license [10,23]. Although it is not possible to make a simple comparison with examinations in English, as the difficulty level and the passing line are different, it is apparent that ChatGPT has reached a certain level in the Japanese orthopedic field.

Expert systems, developed in the 1970s, stood as AI designed to mimic human knowledge in specific domains. These systems aimed to emulate experts by absorbing vast amounts of precise information. Mycin was the typical model. It could guess the identity of bacteria and suggest appropriate antibiotics from the programmed 200 rules by entering the characteristics of the patient (Figure 5) [8].

(A)

An example of a conversation with Mycin

Mycin: From what site was the specimen taken?
 User: Blood
 Mycin: Is it Gram-positive?
 User: No
 Mycin: Is it a rod or coccus?
 User: Rod
 Mycin: Did it grow aerobically?
 User: Yes
 User: What do you think the identity of this bacterium is at this point?
 Mycin: The identity of this bacterium is *Pseudomonas*(.333) *Klebsiella* (.333) *Enterobacter* (.333)

(B)

Rule047

IF
 1)The site of the culture is blood, and
 2)The identity of the organism is not known with certainty, and
 3)The stain of the organism is Gram-negative, and
 4)The morphology of the organism is rod, and
 5)The patient has been seriously burned
Then
 There is weakly suggestive evidence (.4) that the identity of the organism is pseudomonas.

FIGURE 5: An example of a conversation with Mycin

To enter the characteristics of the patient, Mycin can guess the identity of bacteria and suggest appropriate antibiotics from the 200 rules.

(A) An example of a conversation with Mycin.

(B) One example out of 200 rules.

Image credits: Nozomu Nakajima.

However, despite the innovative system, they were not widely used because they required experts to manually input knowledge, making updates difficult. Additionally, being rule-based, these systems could only process inputs in a predetermined manner, restricting their adaptability [4,5]. In contrast, ChatGPT boasts a wide-ranging knowledge base, allowing for flexible, rapid responses in conversational formats, highlighting its superiority [23,24]. While remaining adaptable, ChatGPT has rapidly improved in accuracy to a level that would pass the JBOSE [25]. ChatGPT has the potential to evolve into a state-of-the-art expert system by further increasing knowledge in specialized areas.

In addition to issues like response reliability, a challenge known as “Artificial Hallucination” is highlighted. This refers to AI generating incorrect or unsubstantiated information as if it were factual, a significant concern with large language models [26,27]. While the rate of hallucinations in GPT-4 is reportedly decreasing, it is still unsatisfactory [28]. These hallucinations, essentially AI “misconceptions” based on false information, can be hard to discern without expert knowledge. This leads to potential misreading and underscores the need for further refinement.

Although GPT-4 corresponded to the passing level, the performance on modified questions (where the number of correct options was concealed) was lower compared with conventional questions. In some cases, once a question was answered correctly, it would respond incorrectly when asked at a different time. There was some randomness in ChatGPT’s responses because it is based on transformer architecture, which works by predicting the next word in a sequence based on probabilities. It analyzes the context and previous words to estimate which word might logically follow. This process, drawing from extensive training data, results in some unpredictability or randomness in the responses [10,24].

This study had some limitations. First, ChatGPT is continuously evolving through user feedback, which means that outcomes might vary depending on the timing of the test. The evaluations of GPT-3.5 and GPT-4 were conducted in August 2023, while GPT-4V was tested in October. Similarly, the randomness of ChatGPT’s responses could also affect the results, but each model was tested only once in this test. Second, the quality of images of image-based questions was not very high. For this study, images were downloaded as monochrome PDFs from the society’s website. However, in the actual examination, more high-quality color images are available on computer screens. This may lead to GPT-4V’s failure to produce a better performance than GPT-4 in the image-based questions.

Conclusions

This study demonstrated that ChatGPT has the performance to pass the Japanese orthopedic specialist examination. However, even if ChatGPT can pass the exam, it cannot replace the clinician as is. “Artificial hallucinations” can cause harm to patients, and the use of such technology in medical practice should be carefully phased in.

However, with its extensive knowledge across a wide range of fields beyond medicine, the user-friendly chat format, and multimodal functionalities like image recognition and web search, ChatGPT has the potential to be a potent support tool. Therefore, we continuously need to evaluate and improve the evolving capabilities of AI to ensure its beneficial and safe use in the medical field.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Takahito Fujimori, Nozomu Nakajima, Masayuki Furuya, Yuya Kanie, Hirotatsu Imai, Kosuke Kita, Keisuke Uemura, Seiji Okada

Critical review of the manuscript for important intellectual content: Takahito Fujimori, Nozomu Nakajima, Masayuki Furuya, Yuya Kanie, Hirotatsu Imai, Kosuke Kita, Keisuke Uemura, Seiji Okada

Supervision: Takahito Fujimori

Acquisition, analysis, or interpretation of data: Nozomu Nakajima

Drafting of the manuscript: Nozomu Nakajima

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue.

Animal subjects: All authors have confirmed that this study did not involve animal subjects or tissue.

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

- Russell S, Norvig P: Artificial Intelligence: A Modern Approach. Third Edition. Prentice Hall, Hoboken, NJ; 2010.
- Ghahramani Z: Probabilistic machine learning and artificial intelligence. *Nature*. 2015, 521:452-9. [10.1038/nature14541](https://doi.org/10.1038/nature14541)
- Ramesh AN, Kambhampati C, Monson JR, Drew PJ: Artificial intelligence in medicine. *Ann R Coll Surg Engl*. 2004, 86:354-8. [10.1308/147870804290](https://doi.org/10.1308/147870804290)
- Duda RO, Shortliffe EH: Expert systems research. *Science*. 1985, 220:261-8. [10.1126/science.6540198](https://doi.org/10.1126/science.6540198)
- Adami C: A brief history of artificial intelligence research. *Artif Life*. 2021, 27:131-7. [10.1162/artl_a_00349](https://doi.org/10.1162/artl_a_00349)
- Davis R, Buchanan B, Shortliffe E: Production rules as a representation for a knowledge-based consultation program. *Artif Intell*. 1977, 8:15-45. [10.1016/0004-3702\(77\)90003-0](https://doi.org/10.1016/0004-3702(77)90003-0)
- Shortliffe EH: Mycin: a knowledge-based computer program applied to infectious diseases. *Proc Annu Symp Comput Appl Med Care*. 1977, 66-9.
- Shortliffe EH: Computer-Based Medical Consultations: Mycin. Elsevier, Amsterdam, Netherlands; 1976. [10.1016/B978-0-444-00179-5.X5001-X](https://doi.org/10.1016/B978-0-444-00179-5.X5001-X)
- Hasman A: My journey through the field of medical informatics. *Stud Health Technol Inform*. 2022, 300:38-52. [10.3233/SHTI220940](https://doi.org/10.3233/SHTI220940)
- Open AI. GPT-4 Technical Report. (2023). Accessed: January 10, 2024: <https://cdn.openai.com/papers/gpt-4.pdf>.
- 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. [10.1371/journal.pdig.0000198](https://doi.org/10.1371/journal.pdig.0000198)
- Kung JE, Marshall C, Gauthier C, Gonzalez TA, Jackson JB 3rd: Evaluating ChatGPT performance on the orthopaedic in-training examination. *JB JS Open Access*. 2023, 8:e23.00056. [10.2106/JBJS.OA.23.00056](https://doi.org/10.2106/JBJS.OA.23.00056)
- Yamada H, Hosogane N, Adachi N: About 35th orthopaedic specialist examination. *J Jpn Orthop Assoc*. 2023, 97:317-53.
- Yamada H, Taniguchi N, Hosogane N: About 34th orthopaedic specialist examination. *J Jpn Orthop Assoc*. 2022, 96:309-49.
- Matsuyama Y, Kawano H, Yamada H: About 33rd orthopaedic specialist examination. *J Jpn Orthop Assoc*. 2021, 95:305-47.
- Takagi S, Watari T, Erabi A, Sakaguchi K: Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: comparison study. *JMIR Med Educ*. 2023, 9:e48002. [10.2196/48002](https://doi.org/10.2196/48002)
- Saad A, Iyengar KP, Kurisunkal V, Botchu R: Assessing ChatGPT's ability to pass the FRCS orthopaedic part A exam: a critical analysis. *Surgeon*. 2023, 21:263-6. [10.1016/j.surge.2023.07.001](https://doi.org/10.1016/j.surge.2023.07.001)
- Massey PA, Montgomery C, Zhang AS: Comparison of ChatGPT-3.5, ChatGPT-4, and orthopaedic resident performance on orthopaedic assessment examinations. *J Am Acad Orthop Surg*. 2023, 31:1173-9. [10.5435/JAAOS-D-23-00396](https://doi.org/10.5435/JAAOS-D-23-00396)
- Ali R, Tang OY, Connolly ID, et al.: Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. *Neurosurgery*. 2023, 93:1353-65. [10.1227/neu.0000000000002632](https://doi.org/10.1227/neu.0000000000002632)
- Suchman K, Garg S, Trindade AJ: Chat Generative Pretrained Transformer fails the multiple-choice American College of Gastroenterology self-assessment test. *Am J Gastroenterol*. 2023, 118:2280-2.

[10.14509/ajg.0000000000002320](https://doi.org/10.14509/ajg.0000000000002320)

21. Bhayana R, Krishna S, Bleakney RR: Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. 2023, 307:e230582. [10.1148/radiol.230582](https://doi.org/10.1148/radiol.230582)
22. Seghier ML: ChatGPT: not all languages are equal. *Nature*. 2023, 615:216. [10.1038/d41586-023-00680-3](https://doi.org/10.1038/d41586-023-00680-3)
23. Brown T, Mann B, Ryder N, et al.: Language models are few-shot learners. *Adv Neural Inf Process Syst*. 2020, 33:1877-901.
24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L: Attention is all you need. *Adv Neural Inf Process Syst*. 2017, 30:1-11.
25. GPT-4 API general availability and deprecation of older models in the Completions API. (2023). Accessed: January 10, 2024: <https://openai.com/blog/gpt-4-api-general-availability>.
26. Goddard J: Hallucinations in ChatGPT: a cautionary tale for biomedical researcher. *Am J Med*. 2023, 136:1059-60. [10.1016/j.amjmed.2023.06.012](https://doi.org/10.1016/j.amjmed.2023.06.012)
27. Sallam M: ChatGPT utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. *Healthcare (Basel)*. 2023, 11:887. [10.3390/healthcare11060887](https://doi.org/10.3390/healthcare11060887)
28. Walters WH, Wilder EI: Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep*. 2023, 13:14045. [10.1058/s41598-023-41032-5](https://doi.org/10.1058/s41598-023-41032-5)