ChatGPT's Accuracy
in Answering the National Medical Licensing Examination in Japan

Takayuki Nakano
Department of Pulmonary Medicine,
Graduate School of Medical Science
Kyoto Prefectural University of Medicine
Kyoto, Japan
tnakano@koto.kpu-m.ac.jp

***Abstract*—When applying generative AIs to the healthcare field, it is necessary to evaluate their performance. Although there is a previous study on English, we know little about Japanese. We evaluated ChatGPT's accuracy on the Japanese Medical Licensing Examination without modification of its sentence. ChatGPT (-4) achieved an accuracy that was good enough to pass the exam, as long as questions did not contain images. ChatGPT(-4) also showed its ability to make reasonable clinical inferences. While ChatGPT may have potential in healthcare use, we need to know more about its capabilities with respect to healthcare fields.**

Keywords-generative AI; ChatGPT; Japanese; Medical Licesing Examination.

I. Introduction

Large Language Models (LLMs), which are constructed from deep learning techniques on huge Web data sets, have made remarkable progress in recent years. In November 2022, Open AI Inc. launched ChatGPT. They fine-tuned the LLM for dialog-generating AI. To apply generative AI, such as ChatGPT, to the healthcare field, it is absolutely important to assess whether or not they have sufficient and correct medical knowledge. In addition, languages other than English are widely used in the healthcare field globally. There is a need to evaluate the competency of generative AI in languages other than English regarding healthcare affairs. In particular, Japanese is one of the hardest languages to master. Therefore, if a generative AI can demonstrate sufficient medical knowledge even in Japanese, it is reasonable to expect that it can do the same in many languages other than English or Japanese. These results would be a great motivation to apply generative AI to healthcare in many countries.

ChatGPT has already demonstrated its great ability to cover various areas, including the medicine and healthcare fields, even though it did not use a language model specific to healthcare. A previous study showed that ChatGPT (-3.5) had been able to pass the United States Medical Licensing Examination (USMLE) in all three categories [1]. According to the study, ChatGPT obtained accuracy equal to that of those who actually passed the examination. Indeed, a previous study showed that ChatGPT can pass the Japanese Medical Licensing Examination [2]. However, this report allowed translation from Japanese to English or modification of question sentences if they were not suited for ChatGPT in 2023, such as images. Therefore, little is known about ChatGPT's capabilities in the Japanese healthcare field.

The researcher assumed that ChatGPT could answer medical questions correctly even if they were posed only in Japanese. The aim of this study was to evaluate the accuracy of ChatGPT in the Japanese Medical Licensing Examination. In addition, we compared the scores between ChatGPT and the average scores of students who actually took the examination.

II. Methods

In this section, we note about the Japanese Medical Licensing Examination and how to pose the question or its sentences and evaluate it.

A. Japanese Medical Licensing Examination

In Japan, the national examination for medical doctor candidates is held every year, and the actual posed questions and their correct answers are released on the website of the Ministry of Health, Labor and Welfare (MHLW. Almost all questions are written in Japanese. Examinees should select and answer from the presented choices; there is no descriptive question. The examination consists of three categories: general (e.g., Basic medicine such as anatomy, or public health), specific (e.g., Gastroenterology or cardiology), and content that must be mastered by a medical doctor. In addition, there are two main types of questions that could be solved with only knowledge (hereinafter referred to as "General question") and requiring clinical inference skills ('Clinical question'). The former gives the examinee one point per question if the answer is correct and the latter three points, and the latter is more similar to what medical doctors actually do. Candidates for the examination are required to exceed passing standards both in Contents have to be mastered and the others, and approximately 90% of all candidates pass the examination every year.

First, we collected all questions posed from 2018 to 2022 (N = 2000). Second, questions were excluded if they were classified as inappropriate by MHLW (N = 11) or contained images (e.g., photos of the patient, X-ray imaging, or figures) that ChatGPT cannot recognize in 2023 (N = 566). Overall, 1,423 questions were included in this analysis (Figure 1). For multiple-choice questions, a point was awarded only if all

Copyright (c) IARIA, 2024. ISBN: 978-1-68558-136-7

choices were correct. In principle, no modification of the question sentence was allowed; however, only when the

Number of ChatGPT choices differed from the answer, one prompt was allowed to be re-presented. ChatGPT scores were evaluated for both GPT-3.5 and GPT-4.

B. *Average number of medical students*

In this study, we defined "medical student average" as the average score of students who actually took the examination. Because this group consists of both passed and failed students, the scores reflect the performance of students who have completed the education process for doctoral studies in Japan. The medical student average was calculated from the percentages of correct answers written in the books that explain the Japanese Medical Licensing Examination every year. The percentages were derived from questionnaires that more than 90% of all examinees had answered, so the data were reliable enough.

C. *Statistical analysis*

Fisher's exact test was used to evaluate significance. We calculated using EZR, a globally recognized software for analyzing medical statistics.

III. RESULTS

In this section, we describe the percentage of correct answers and scores of both the ChatGPT and medical student averages by question type.

A. *Whole questions*

When analyzing all questions, the accuracy rate of GPT-3.5 was 58.0% (826/1423) and that of GPT-4 was 84.0% (1196/1423). The scoring rate of GPT-3.5 was 59.7% (1080/1809), GPT-4 was 85.0% (1537/1809), and the medical student average was 85.6% (1548.574/1809), respectively (Table I and Figure 2). GPT-3.5 showed a much lower score than the medical student average; on the contrary, GPT-4 showed equal to the medical student average (without any significance).

B. *By Questionare type*

When calculated by questionnaire type, GPT-4 showed an ability similar to the medical student average (Table II and Figure 3). GPT-4 scores improved in almost all areas compared with GPT-3.5. Furthermore, although there was no significant difference, GPT-4 scored better than the medical student average on the Specifics, a category that included questions related to diseases, tests, or treatments.

We also examined scores separating general questions from clinical questions (Figure 4). While GPT-4 performed slightly inferior to the medical student average in the Clinical question, which requires clinical inference skills, GPT-4 was superior to the medical student average on the General questions, which focus on medical knowledge. There were no significant differences in either case.

TABLE I. WHOLE QUESTIONS AS ANALYZED

	respondent		
	GPT-3.5	***GPT-4***	***Medical student average***
Number of correct answers	826	1196	N/A
Percentage of Correct answers (%)	58.0 (826/1423)	84.0 (1196/1423	N/A
Total score	1080	1537	1548.574
Scoring rate (%)	59.7 (1080/1809)	85.0 (1537/1809)	85.6 (1548.574/1809)
p value	<0.001	0.587	Ref.
Odds ratio	0.249	0.948	Ref.
(95% CI)	(0.211-0.293)	(0.786-1.145)	Ref.
Average time taken to	8.78 (1-57)	3.03 (1-93)	N/A

N/A. Not available, Ref. Reference.

TABLE II. BY QUESTIONNAIRE TYPE

Questionare type	**respondent**		
	GPT-3.5	***GPT-4***	***Medical student average***
General	57.3 (323/564)	80.5 (454/564)	82.2 (463.641/564)
Odds ratio (95% CI)	0.289 (0.217-0.385)	0.889 (0.651-1.214)	Ref.
Specifics	52.1 (222/426)	83.8 (357/426)	80.0 (340.766/426)
Odds ratio (95% CI)	0.272 (0.200-0.372)	1.289 (0.894-1.862)	Ref.
Content must be mastered	65.3 (535/819)	88.6 (726/819)	90.9 (744.167/819)
Odds ratio (95% CI)	0.190 (0.142-0.252)	0.787 (0.563-1.098)	Ref.
General question	58.6 (407/694)	85.0 (590/694)	83.7 (580.796/694)
Odds ratio (95% CI)	0.276 (0.212-0.357)	1.10 (0.817-1.490)	Ref.
Clinical question	60.4 (673/1115)	84.9 (947/1115)	86.8 (967.778/1115)
Odds ratio (95% CI)	0.231 (0.186-0.287)	0.856 (0.669-1.094)	Ref.

Ref. Reference.

IV. DISCUSSION

We assessed how much ChatGPT has knowledge about healthcare in Japanese sentences. Our research has shown that ChatGPT (-4) might have sufficient knowledge equal to that of medical doctor candidates. Moreover, ChatGPT (-4) could make clinical inferences only in Japanese and was almost as accurate as medical students who had graduated from medical school.

Conventionally, ChatGPT is less accurate in its products on non-English prompts. Indeed, generative AI is very useful, but this situation will prevent its application to the healthcare field outside English-speaking countries, such as Japan. We assessed the ChatGPT's medical knowledge and clinical inference skills in Japanese sentences. We considered the

Courtesy of IARIA Board and IARIA Press. Original source: ThinkMind Digital Library https://www.thinkmind.org

Medical Licensing Examination the most appropriate. First, the quality of the questions is guaranteed by a national institution. Second, it requires broad knowledge from basic medicine to internal medicine or surgery. Third, there was an ideal control group, the medical student average. There are also many medical specialist examinations in Japan, but most of them do not release actual questions or answers.

This study design was stricter than that of a previous report [2]. We had not allowed almost all modifications of the sentence, except for the number of choices. This fineness is partly used to evaluate the ability of ChatGPT in Japanese and to ensure that the questions are solved as closely as possible. Although there is skepticism about evaluating the significance of the results of generative AI, we believe that this rigorous study design allowed us to calculate significance.

We showed that ChatGPT (-4) can pass the Japanese Medical Licensing Examination if it excludes questions with images. Given the technical principles of generative AI, we could have presumed that it would perform better on questions requiring knowledge, but the fact that ChatGPT (-4) performed as well as the medical student average on questions requiring clinical inference skills was noteworthy. Inference skills are important in clinical practice and often take time for human medical students to master.

We have some limitations. First, we excluded more than a quarter of all questions, and most of the questions excluded contained images. Multimodal questions involving images may also be difficult for generative AI as human candidates. We consider that we can overcome this situation by collecting more than one thousand questions, and we hope that image recognition AIs will show us their desired performance. Second, this study specializes in medical doctor examinations. In addition to medical doctors, many other professions are involved in the healthcare field, and their contributions are significant. Therefore, you cannot simply judge that a generative AI can be applied to the healthcare field with only this study. However, it may be a milestone for the medical application of generative AI because knowledge of diseases and the ability to make clinical inferences are the basis for decision making in all healthcare fields.

Acknowledgment

I would like to thank Open AI Inc. and all its contributors. Furthermore, I am grateful to the reviewers for their useful suggestions.

References

[1] T. H. Kung and M. Cheatham, “Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models.” PLOS Digit Health, vol. 2, e0000198. Feb. 2023, doi: 10.1371/journal.pdig.0000198.

[2] Y Tanaka and A. Nomura, “Performance of Generative Pretrained Transformer on the National Medical Licensing Examination in Japan.” PLOS Digit Health, vol. 3, e0000433. Jan. 2024, doi: 10.1371/journal.pdig.0000433.

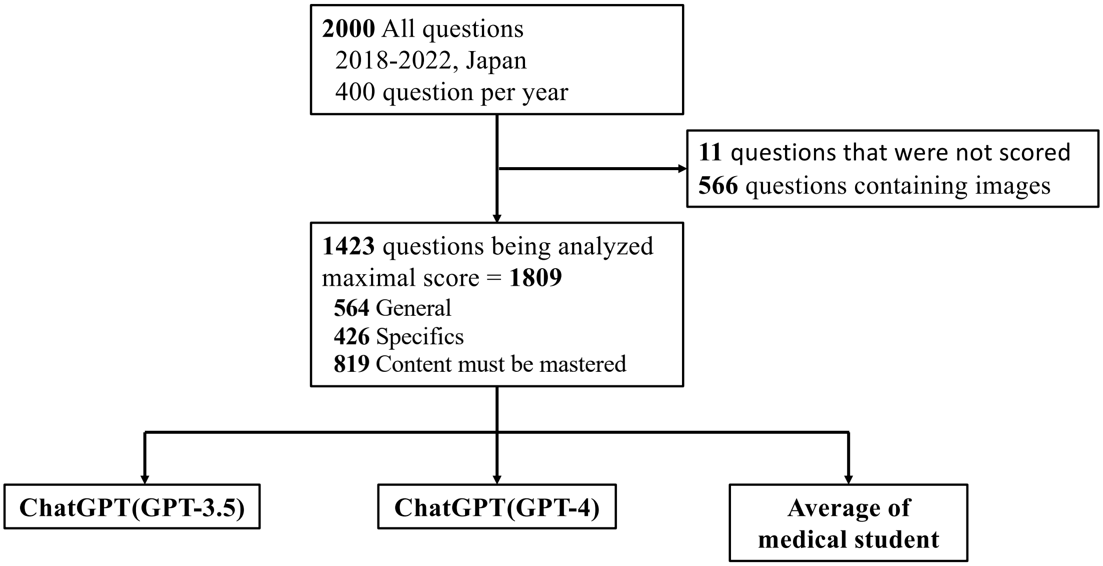
2000 All questions
2018-2022, Japan
400 question per year
11 questions that were not scored
566 questions containing images
1423 questions being analyzed
maximal score = **1809**
564 General
426 Specifics
819 Content must be mastered
ChatGPT(GPT-3.5)
ChatGPT(GPT-4)
Average of medical student

Figure 1. Questions regarding the inclusion and exclusion criteria of this study.

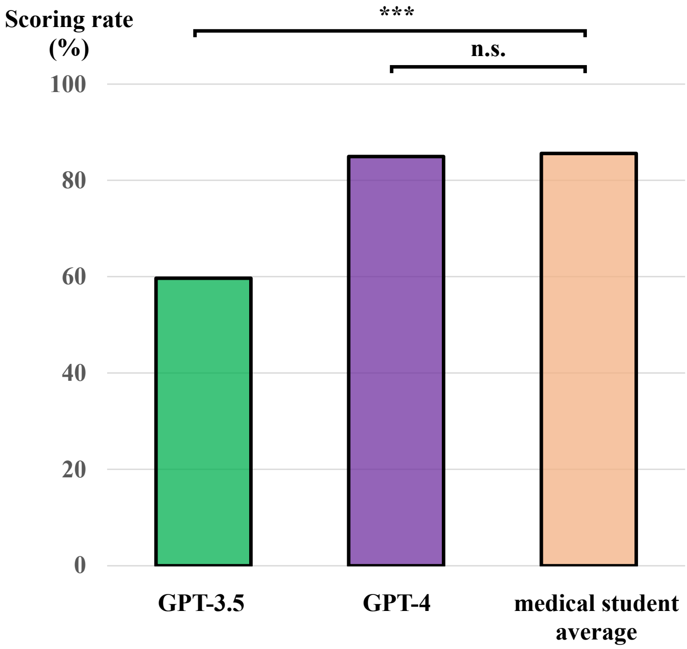

Figure 2. Total score in all questions. GPT-4 was superior to GPT-3.5 and equal to the medical student average.

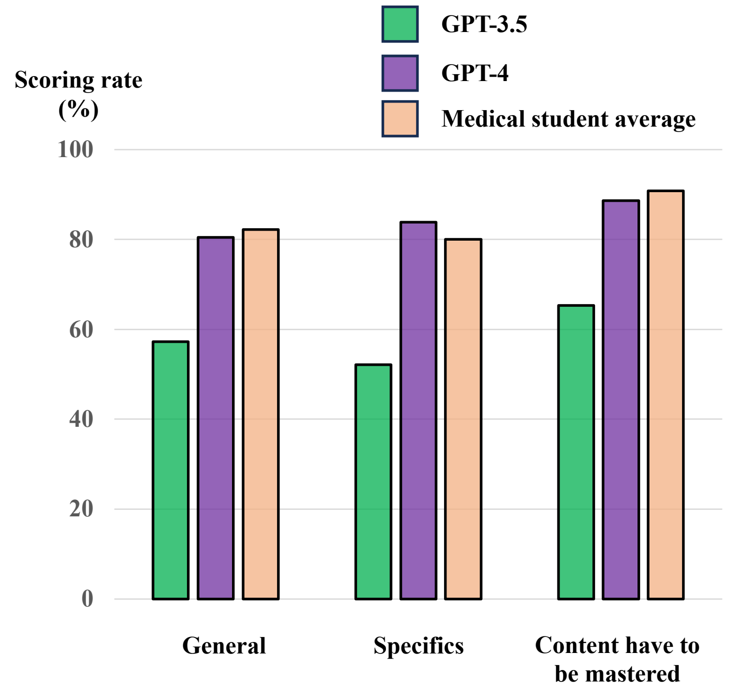

Figure 3. Score per category: General, Specific, and Content must be mastered. GPT-4 was generally similar to the medical student average. In particular, although there was no significance, ChatGPT (-4) was superior to the medical student average in Specific.

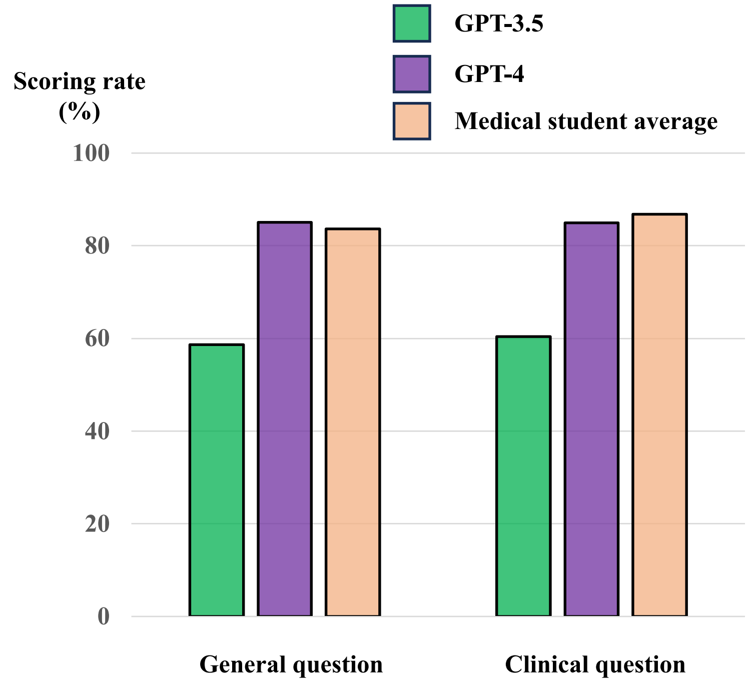

Figure 4. Score per category: General and Clinical questions. GPT-4 was generally similar to the medical student average. In particular, although there was no significant difference, ChatGPT (-4) was superior to the medical student average in the general question.