*Article*

# Can ChatGPT Pass the 2023 Japanese National Medical Licensing Examination?

**Yudai Kaneda[1]*, Tetsuya Tanimoto[2], Akihiko Ozaki[3], Tomohiko Sato[4] and Kenzo Takahashi[5]**

[1] School of Medicine, Hokkaido University, Sapporo, Hokkaido, Japan; nature271828@gmail.com
[2] Department of Internal Medicine, Jyoban Hospital of Tokiwa Foundation, Iwaki, Fukushima, Japan; tetanimot@yahoo.co.jp
[3] Department of Breast and Thyroid Surgery, Jyoban Hospital of Tokiwa Foundation, Iwaki, Fukushima, Japan; ozakiakihiko@gmail.com
[4] Division of Transfusion Medicine and Cell Therapy, The Jikei University Hospital, Minato-ku, Tokyo, Japan; tomosatou@jikei.ac.jp
[5] Teikyo University Graduate School of Public Health, Itabashi-ku, Tokyo, Japan; kt.intl.hlth@gmail.com
* Correspondence: nature271828@gmail.com

**Abstract:** ChatGPT is gaining widespread acceptance for its ability to generate natural language sentences in response to various inputs and is expected to become a supplementary tool for diagnosing and determining treatment policies in clinical settings. ChatGPT was used to evaluate its ability to perform clinical inference and accuracy in answering questions on the 117th Japanese National Medical Licensing Examination held in February 2023. The exam questions were manually inputted into ChatGPT's window, and the accuracy of ChatGPT's responses was determined based on answers provided by a preparatory school. ChatGPT provided answers for 389 out of 400 questions, and its overall correct answer rate was 55.0%. The correct answer rate for 5-choice-1, 5-choice-2, and 5-choice-3 were 57.8%, 42.9%, and 41.2%, respectively. The highest correct answer rate was for the compulsory exam (67.0%), followed by the specific knowledge exam (54.1%) and the cross category exam (47.9%). The correct answer rate for non-image questions was 56.2% and for image questions, it was 51.5%. The study suggests that ChatGPT has potential to support healthcare professionals in clinical decision-making in Japanese clinical settings, but caution should be exercised in interpreting and using the answers generated by ChatGPT due to room for improvement in performance.

**Keywords: ChatGPT; Medical Licensing Examination; Clinical Settings; Japan**

## 1. Introduction

In recent years, the progress of artificial intelligence (AI) has propelled innovation in the healthcare industry [1, 2], with large language models (LLMs) known as self-regressive language models, in particular, receiving significant attention [3, 4]. ChatGPT, launched by OpenAI on November 30, 2022, is an accessible and refined LLM that has gained wide acceptance as a new level of service useful for retrieving information, answers, or solutions [5]. The distinguishing feature of ChatGPT is its training to generate natural language sentences that appear as if in conversation with a human, given input texts of various languages.

Despite not receiving specialized training in specific fields, ChatGPT has already achieved passing scores or results close to passing scores on exams that assume graduate-level specialization in fields such as law and business [6]. Moreover, it has been reported to achieve near-passing scores on the theoretical section of the United States

doi:10.20944/preprints202303.0191.v1

Medical Licensing Examination (USMLE) without additional training or learning [7]. Therefore, like AI for image diagnosis currently in use [8, 9], ChatGPT is expected to be utilized as a supplementary factor in diagnosing and determining treatment policies in clinical settings using language information in the future [10, 11]. Namely, the usefulness of ChatGPT in clinical settings can potentially be evaluated by having it solve various medical exams [12-15].

In this study, we examined the extent to which ChatGPT can be utilized in the Japanese clinical setting by investigating the 117th Japanese National Medical Licensing Examination (JNMLE) held in February 2023. This licensing examination using mostly Japanese language is required to be annually conducted based on the Japan's Medical Practitioners Act, which stipulates that medical doctors should possess the necessary knowledge and skills related to medicine and public health [16]. Indeed, the criteria for the National Medical Practitioner Examination questions conform to the basic knowledge and skills that medical doctors should possess at a minimum when taking the first step in the medical field, and are intended to ensure the minimum abilities that Japanese medical doctors should possess [17]. The passing rates for 2020, 2021, and 2022 were 92.1%, 91.4%, and 91.7%, respectively [18], and it is considered useful to examine how close ChatGPT's responses come to the passing standard. The purpose of this study is to evaluate the ability of ChatGPT, a non-domain specific LLM, to perform clinical inference and test the accuracy of its responses to questions on the 2023 JNMLE as of March, 2023.

## 2. Materials and Methods

### *ChatGPT*

ChatGPT (OpenAI, San Francisco, California) is a type of LLM that generates natural language responses to text inputs in a conversation context using self-attention mechanisms and a large amount of training data. Unlike the deep learning (DL) models that were prevalent before LLM, which were designed to learn and recognize patterns in data, LLM is a new type of AI algorithm trained to predict the likelihood of specific word sequences based on the context of preceding words. It is a language model included in servers that cannot browse or execute internet searches and is particularly effective in generating responses suitable for consistent contexts. ChatGPT is equipped with information up until September 2021 as of March, 2023.

### *Japanese National Medical Licensing Examination (JNMLE)*

JNMLE is composed of 6 blocks, divided into compulsory exam (blocks B and E), cross category exam (blocks C and F), and specific knowledge exam (blocks A and D), to be answered in a paper-based format over two consecutive days with 3 blocks per day once in a year. Each block consists of general questions that test basic medical knowledge and clinical questions that present specific case descriptions. The compulsory exam focus on primary care and cover basic questions that span multiple disciplines. The cross category and specific knowledge exams are limited to content that can be handled by any medical institution in Japan. Additionally, the exam includes many practical image questions that require interpretation such as CT or MRI images or electrocardiograms. Answers are provided in a multiple-choice format, where test-takers select one, two, or three correct answers from five options. While relatively rare, some problems may have more than five options or involve calculations.

There are three passing criteria for the Japanese national medical licensing examination: a total score of 80% or more in compulsory exam, a total score of approximately 70% in cross category and specific knowledge exam, and not making more than three mistakes in the forbiddance options. For compulsory exam, a total of 100 questions, consisting of 50 general questions and 50 clinical questions, are asked. The clinical questions are scored at 3 points per question, and 160 or more out of the total score of 200 is an absolute condition for passing. For cross category and specific knowledge exam, the total score of general and clinical questions is evaluated relative to other examinees, and the pass or fail is determined. In fiscal year 2020, 2021, and 2022, the border passing score rates were 72.6%, 69.7%, and 72.1%, respectively, and a score of approximately 70% is required [18].

Forbiddance options are choices that lead to the death of a patient or irreversible organ dysfunction or violate the laws that physicians should comply with. Choosing four of these options will result in failing the exam even if all other questions are correct. However, it is not disclosed whether the Ministry of Health, Labor, and Welfare (MHLW), which is in charge of the national medical examination, has set it as a forbiddance option, so it was not considered in this study. Also, there are cases where deletion questions are announced at the time of passing announcement due to reasons such as answers not being determined, but as of the time of writing this manuscript in March, 2023, it was unclear, and this study was conducted assuming that all questions will be scored.

### *The Measurement of Accuracy*

The examination questions for the 117th JNMLE, held on February 4 and 5, 2023, were permitted to be taken home and provided to us by medical student examinees for research purposes. The official answers to the exam questions, along with the questions themselves, are typically published on the website of the MHLW around April of each year.

The exam questions were manually inputted into ChatGPT's window on March 2, 2023, and we confirmed that they had not yet been made publicly available on the Internet at the time of input. In cases where the method of answering was unclear from the original question text, the phrase 'Provide one answer' was added as needed. For some questions that were part of a series, the second question was inputted following the text of the first question for ChatGPT to solve. In cases where questions included images, only the text of the question was inputted, and no image information was used. The accuracy of ChatGPT's responses was determined based on answers provided by a preparatory school specialized in the medical licensing examination, which are announced immediately after the exam [19].

### *Data Analysis*

In this study, we conducted three analyses. First, in order to evaluate ChatGPT's performance at JNMLE, we calculated the correct answer rate and score rate for each block and for the entire exam. Second, in order to assess how accurate ChatGPT's responses are based solely on linguistic information, we compared the correct answer rate when excluding image questions and when considering only image questions. Finally, in order to provide a more appropriate assessment of ChatGPT's ability to perform in Japanese clinical settings, we compared the correct answer rate for general problems and clinical problems when excluding image problems. We performed $\chi$-square test for each comparison. All analyses were performed using Stata/IC 15.0.

### *Ethical Approval*

Ethical considerations were not applied to this study as the data were in the public domain.

## 3. Results

Out of the 400 questions primarily presented in Japanese, 342 (85.5%) were in a format of a 5-choice single answer, with 2 (0.5%) of them presented in English, 35 (8.75%) were in the format of a 5-choice double answer, and 18 (4.5%) were in the format of a 5-choice triple answer. Among them, 110 (27.5%) questions included medical images. Only 2 (0.5%) questions had more than 6 answer choices, and 3 (0.75%) questions required calculations.

According to the results, ChatGPT generated some kind of answers for 389 (97.25%) out of the 400 questions that were presented. A total of 11 (2.75%) questions were excluded as ChatGPT did not generate any answer for them, and all of these questions were related to interpretation of medical images. Additionally, a total of 24 (6.0%) questions required the inclusion of phrases such as "answer one" to be solved, and if not included, there were no answers for 2 (0.5%) questions, and multiple answers were generated for 22 (5.5%) questions.

The overall simple correct answer rate was 55.0% (214/389). After excluding image questions, the correct answer rates for the 5-choice questions were 57.8% (192/332) for 5-choice-1, 42.9% (15/35) for 5-choice-2, and 41.2% (7/17) for 5-choice-3. There was no statistically significant difference in the correct answer rate between each types of the questions. Both of the questions with more than 6 answer choices were answered incorrectly. The response for the two English-language questions was correct, and all three calculation questions were answered incorrectly.

In terms of individual fields, the correct answer rate was highest for the compulsory exam (67.0%, 65/97), followed by the specific knowledge exam (54.1%, 80/148) and the cross category exam (47.9%, 69/144). There was a statistically significant difference in correct answer rates between the compulsory exam and the cross category exam (p=0.005).

Table 1 shows the correct answer rates, scores, and number of unanswered questions for each block. The total correct answer rate for the compulsory exam was 67.0%, with a score of 135/197 (68.5%). The total correct answer rate for the specific knowledge and the cross category exams was 51.0%, with a score of 149/292 (51.0%).

**Table 1.** Correct answer rates, scores, and number of unanswered questions for each block.

| | Correct answer rates | Score | Number of unanswered questions |
|---|---|---|---|
| A | 40/75 (53.3%) | 40/75 | 0 |
| B | 32/50 (64.0%) | 64/100 | 0 |
| C | 32/72 (44.4%) | 32/72 | 3 |
| D | 40/73 (54.8%) | 40/73 | 2 |
| E | 33/47 (70.2%) | 71/97 | 3 |
| F | 37/72 (51.4%) | 37/72 | 3 |
| Total | 214/389 (55.0%) | 284/489 (58.1%) | 11 |

Table 2 shows the correct answer rates for each block when excluding image questions and for image questions only independently. The overall correct answer rate for non-image questions was 56.2% (163/290), and the correct answer rate for image questions was 51.5% (51/99), with no statistically significant difference (p=0.488).

**Table 2.** Correct answer rates for image and non-image questions.

| | Non-image questions | Image questions |
|---|---|---|
| A | 19/35 (54.3%) | 21/40 (52.5%) |
| B | 29/45 (64.4%) | 3/5 (60.0%) |
| C | 26/58 (44.8%) | 6/14 (42.9%) |
| D | 27/47 (57.4%) | 13/26 (50.0%) |
| E | 29/41 (70.7%) | 4/6 (66.7%) |
| F | 33/64 (51.6%) | 4/8 (50.0%) |
| Total | 163/290 (56.2%) | 51/99 (51.5%) |

Table 3 shows the correct answer rates for general and clinical questions with image questions excluded from each block. The overall percentage of correct answers for general questions without image questions was 54.7% (75/137), and, that for clinical questions without image questions was 57.5% (88/153), with no statistically significant difference between them (p=0.722). Except for the D block, the percentage of correct answers for clinical questions was higher than that for general questions in all other blocks.

**Table 3.** Correct answer rates for general and clinical questions with image questions excluded.

| | General questions | Clinical questions |
|---|---|---|
| A | 8/15 (53.3%) | 11/20 (55.0%) |
| B | 15/24 (62.5%) | 14/21 (66.7%) |
| C | 12/31 (38.7%) | 14/27 (51.9%) |
| D | 10/13 (76.9%) | 17/34 (50.0%) |
| E | 13/21 (61.9%) | 16/20 (80.0%) |
| F | 17/33 (51.5%) | 16/31 (51.6%) |
| Total | 163/290 (56.2%) | 51/99 (51.5%) |

## 4. Discussion

This study evaluated the performance of ChatGPT as of March 2023, in answering the multiple-choice questions of the JNMLE conducted in February 2023. Although the random probability of correctly answering the questions is less than 20%, the overall accuracy rate was 55.0%, with a total score rate of 68.5% for the compulsory exam and 51.03% for the specific knowledge and the cross category exams. While ChatGPT did not meet the passing requirements of the JNMLE, it was suggested that it could be potentially useful in assisting clinical diagnosis and treatment decision-making in real-world Japanese-language settings if used with caution and an understanding of its characteristics.

Notably, there were differences in the accuracy rate by the exam. The accuracy rate was highest in the compulsory exam (67.0%), followed by the specific knowledge exam (54.1%) and the cross category exam (47.9%). There was a statistically significant difference in correct answer rates between the compulsory exam and the cross category exam (p=0.005), with higher accuracy rates in compulsory exam and lower rates in the cross category exam. In compulsory exam, basic multidisciplinary knowledge focused on primary care that is essential for becoming a resident is required, whereas specific knowledge about diseases and cross-disciplinary knowledge is required in cross category exam. This suggests that ChatGPT may excel at generating answers on more general topics or composite areas. Indeed, in a previous study evaluating the performance of ChatGPT in the field of ophthalmology, the highest results in general medicine were obtained, while there was room for improvement in sub-specialties such as neuro-ophthalmology, pathology, and intraocular tumors [20].

One notable point is that despite not using any image-related information, the correct answer rate for image questions alone was 51.5% (51/99). This is likely due to the fact that the questions were designed in a way that allows the answer to be inferred from the text even without visual information. However, combining the language information generated by ChatGPT with AI-based image diagnosis techniques may further improve clinical reasoning abilities in the near future.

The fact that there were 6.2% (24/389) of questions that required the phrase "Provide one answer" suggests the existence of a cultural nature of the exam in Japan. Namely, the

Japanese medical exam may require a significant amount of implicit consideration towards the examiners in order to obtain high scores, as the exam may not always explicitly indicate correct or incorrect answers. In fact, Japan is considered to be one of the countries with the highest-context communication in the world, where people communicate while reading the atmosphere, anticipating unspoken intentions due to their shared cultural background [21]. This cultural background may also affect the exam questions, requiring the examinees to read between the lines and infer implicit knowledge that is not explicitly stated in the exam questions. Thus, one possible reason why the performance of ChatGPT was not as high in the Japanese medical exam we examined here, as compared to a study evaluating the performance of ChatGPT in the USMLE [7], may be considered differences in cultural and linguistic communication.

Additionally, a study testing ChatGPT's performance on China's National Medical Licensing Examination also reported a failure to achieve a passing score [22]. This suggests that ChatGPT's performance may depend on differences in the amount of information conveyed through language. Indeed, it is estimated that the amount of information accessible in English on the internet is approximately 16.6 times greater than in Japanese and 37.7 times greater than in Chinese [23]. In this study, we input the questions from JNMLE directly into ChatGPT in Japanese and generated responses in Japanese. However, if the questions were translated into English and then answered, it is possible that ChatGPT would achieve a higher score even at its current performance level. In fact, both of the two questions in English that were included in this study were answered correctly by ChatGPT, which warrants further investigation.

ChatGPT is developed based on OpenAI's previous GPT-3.5 language model, with both supervised and reinforcement learning methods added. As user numbers increase, it is highly likely that answer accuracy will naturally improve [5]. In fact, although the data on which ChatGPT is based only goes up until September 2021 and the test data used in this study was from February 2023, ChatGPT achieved a noteworthy accuracy rate of 55.0% through its own learning despite being in the early stages of its release. However, in terms of implementation in the clinical setting, there is already available clinical decision support resource called UpToDate,[24] which aggregates human experiential knowledge and evidence-based knowledge, and it is important to accumulate further learning experience for ChatGPT to reach that level. Additionally, ChatGPT has been criticized for occasionally providing seemingly plausible but inaccurate or meaningless answers,[25, 26] and it is important for healthcare professionals who are users to rationally judge the accuracy of the information rather than relying blindly on it. In order to maximize patient benefits, it is important to further explore effective ways to utilize ChatGPT in various clinical situations.

This study has several limitations. First, the correct/incorrect answers for each question were calculated using preliminary answer reports released by a preparatory school, as the official answers from the MHLW were not yet available. Therefore, there is a possibility that the preparatory school's answers are incorrect, which could affect the accuracy of the results we calculated. Second, in this analysis, we did not investigate the basis for each answer. As most of the questions on the exam were multiple choice questions, it is possible to arrive at the correct answer by chance. Thus, it is important to verify the validity of ChatGPT's answers and investigate its performance in more detail in the future. Third, we did not evaluate the accuracy for each subject area in this study. By clarifying ChatGPT's strengths and weaknesses subject areas, we can identify the specific clinical scenarios where it can be utilized effectively. Fourth, the input of question text was done manually because as of March 2, 2023, when this study was conducted, the 2023 JNMLE questions were not yet available online, and the input had to be based on the paper questions distributed. Although multiple persons checked for confirmation, there is still a possibility of input errors. As ChatGPT generates answers in a natural language conversation, it has been pointed out that the correct answer can change depending on the context [27]. Hence, there is also a possibility that input errors may have affected the accuracy of the results. Despite these limitations, we believe that this study was conducted with sufficient research standards, as medical professionals and students were at the center of the research.

## 5. Conclusions

Our study demonstrated that ChatGPT can achieve a certain level of accuracy in the JNMLE, although it did not reach the passing level. This suggests that such large-scale language models have the potential to support healthcare professionals in the clinical decision-making process in Japanese clinical settings. However, there is still room for improvement in performance, and caution should be exercised in interpreting and using the answers generated by ChatGPT.

**Author Contributions:** Conceptualization, Y.K. and T.T.; methodology, Y.K.; validation, T.S. and K.T.; formal analysis, Y.K.; investigation, Y.K.; resources, Y.K.; data curation, Y.K.; writing—original draft preparation, Y.K.; writing—review and editing, T.T., A.O., T.S. and K.T.; visualization, K.Y.; supervision, T.T.; project administration, A.O. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data cannot be posted here as of March 9, 2023, due to copyright issues, but it will be publicly available on the website of the Japanese Ministry of Health, Labor and Welfare (link: https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryou/iryou/index.html) in April 2023.

**Acknowledgments:** I would like to express my wholehearted gratitude to Ms. Chiharu Kaneda for making this research setting possible.

**Conflicts of Interest:** TT received personal fees from MNES Inc and Bionics co., ltd. outside the submitted work. AO received personal fees from MNES Inc. and Kyowa Kirin co. ltd. outside the submitted work. TS received joint research funding from the Japanese Red Cross Society outside the submitted work.

## References

1. Dirican, C., *The impacts of robotics, artificial intelligence on business and economics.* Procedia-Social and Behavioral Sciences, 2015. **195**: p. 564-573.
2. Varghese, M., S. Raj, and V. Venkatesh, *Influence of AI in human lives.* arXiv preprint arXiv:2212.12305, 2022.
3. Brown, T., et al., *Language models are few-shot learners.* Advances in neural information processing systems, 2020. **33**: p. 1877-1901.
4. Savery, M., et al., *Question-driven summarization of answers to consumer health questions.* Sci Data, 2020. **7**(1): p. 322.
5. OpenAI. *ChatGPT*. 2022 [cited 2023 March 3]; Available from: https://openai.com/blog/chatgpt/
6. Kelly, S.M., *ChatGPT passes exams from law and business schools*. 2023.
7. Kung, T.H., et al., *Performance of ChatGPT on USMLE: Potential for AI-Assisted Medical Education Using Large Language Models.* medRxiv, 2022: p. 2022.12.19.22283643.
8. Sharma, P., et al., *Artificial Intelligence in Diagnostic Imaging: Status Quo, Challenges, and Future Opportunities.* Journal of Thoracic Imaging, 2020. **35**.
9. Lin, S.Y., M.R. Mahoney, and C.A. Sinsky, *Ten Ways Artificial Intelligence Will Transform Primary Care.* Journal of General Internal Medicine, 2019. **34**(8): p. 1626-1630.
10. Rao, A., et al., *Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow.* medRxiv, 2023.

11. Abdullah, I.S., A. Loganathan, and R.W. Lee, *ChatGPT & Doctors: The Medical Dream Team.* 2023.

12. Gilson, A., et al., *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. **9**: p. e45312.

13. Kung, T.H., et al., *Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models.* PLOS Digital Health, 2023. **2**(2): p. e0000198.

14. Huh, S., *Are ChatGPT's knowledge and interpretation ability comparable to those of medical students in Korea for taking a parasitology examination?: a descriptive study.* J Educ Eval Health Prof, 2023. **20**: p. 1.

15. Fijacko, N., et al., *Can ChatGPT pass the life support exams without entering the American heart association course?* Resuscitation, 2023. **185**: p. 109732.

16. Japanese Law Translation. *Medical Practitioners' Act*. 1948 [cited 2023 March 4]; Available from: https://www.japaneselawtranslation.go.jp/ja/laws/view/3992.

17. Ministry of Health, Labor and Welfare. *The National Examination for Medical Practitioners Question Criteria [in Japanese]*. 2018 [cited 2023 March 4]; Available from: https://www.mhlw.go.jp/stf/shingi2/0000128981.html.

18. MEC. *Medical Licensing Examination Data [in Japanese]*. 2023 [cited 2023 March 4]; Available from: https://www.gomec.co.jp/mec/kokushi/back_data/.

19. Medic Media. *117th National Medical Practitioners' Examination Preliminary Answers & Scoring Service*. 2023 [cited 2023 March 4]; Available from: https://kousoku.medilink-study.com/.

20. Antaki, F., et al., *Evaluating the Performance of ChatGPT in Ophthalmology: An Analysis of its Successes and Shortcomings.* medRxiv, 2023: p. 2023.01.22.23284882.

21. Meyer, E., *The culture map: Breaking through the invisible boundaries of global business*. 2014: Public Affairs.

22. Wang, X., et al., *ChatGPT Performs on the Chinese National Medical Licensing Examination.* 2023.

23. W3Techs.com. *Usage statistics of content languages for websites*. 2023 [cited 2023 March 3]; Available from: https://w3techs.com/technologies/overview/content_language.

24. Wolters Kluwer. *UpToDate*. 2023 [cited 2023 March 6]; Available from: https://www.uptodate.com/contents/search.

25. Anderson, N., et al., *AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in Sports & Exercise Medicine manuscript generation.* BMJ Open Sport Exerc Med, 2023. **9**(1): p. e001568.

26. Vincent, J., *AI-generated answers temporarily banned on coding Q&A site Stack Overflow*. 2022, Retrieved.

27. Mbakwe, A.B., et al., *ChatGPT passing USMLE shines a spotlight on the flaws of medical education.* PLOS Digit Health, 2023. **2**(2): p. e0000205.