—Original—

Performance of a Large Language Model on Japanese Emergency Medicine Board Certification Examinations

Yutaka Igarashi[1], Kyoichi Nakahara[1], Tatsuya Norii[2], Nodoka Miyake[1], Takashi Tagami[3] and Shoji Yokobori[1]

[1]Department of Emergency and Critical Care Medicine, Nippon Medical School, Tokyo, Japan
[2]Department of Emergency Medicine, University of New Mexico, NM, United States of America
[3]Department of Emergency and Critical Care Medicine, Nippon Medical School Musashi Kosugi Hospital, Kanagawa, Japan

Background: Emergency physicians need a broad range of knowledge and skills to address critical medical, traumatic, and environmental conditions. Artificial intelligence (AI), including large language models (LLMs), has potential applications in healthcare settings; however, the performance of LLMs in emergency medicine remains unclear.

Methods: To evaluate the reliability of information provided by ChatGPT, an LLM was given the questions set by the Japanese Association of Acute Medicine in its board certification examinations over a period of 5 years (2018-2022) and programmed to answer them twice. Statistical analysis was used to assess agreement of the two responses.

Results: The LLM successfully answered 465 of the 475 text-based questions, achieving an overall correct response rate of 62.3%. For questions without images, the rate of correct answers was 65.9%. For questions with images that were not explained to the LLM, the rate of correct answers was only 52.0%. The annual rates of correct answers to questions without images ranged from 56.3% to 78.8%. Accuracy was better for scenario-based questions (69.1%) than for stand-alone questions (62.1%). Agreement between the two responses was substantial (kappa = 0.70). Factual error accounted for 82% of the incorrectly answered questions.

Conclusion: An LLM performed satisfactorily on an emergency medicine board certification examination in Japanese and without images. However, factual errors in the responses highlight the need for physician oversight when using LLMs. (J Nippon Med Sch 2024; 91: 155–161)

Key words: artificial intelligence, emergency medicine, language, medicine, specialty boards

Introduction

Emergency medicine encompasses prehospital assistance, disaster readiness, proficiency in basic and advanced resuscitation techniques, and management of critical medical, traumatic, and environmental conditions that demand immediate attention in persons of all age groups[1]. Thus, emergency physicians need to have broad knowledge, advanced technical skills, the ability for rapid decision-making, good communication skills, and extensive experience in responding to various situations.

Artificial intelligence (AI) has the potential to reduce the burden of emergency physicians by fulfilling many expected roles[2–4]. Large language models (LLMs) are an advanced form of AI that learns from large quantities of textual information and engages in natural interactions with humans. Although LLMs are not models specially trained in a particular domain, they can answer questions related to medical expertise. For example, applications are expected in fields such as computer-aided diagnosis, data summarization, communication, and education[5–12].

To evaluate the performance of LLMs in applications

Correspondence to Yutaka Igarashi, MD, PhD, Department of Emergency and Critical Care Medicine, Nippon Medical School, 1–1–5 Sendagi, Bunkyo-ku, Tokyo 113–8603, Japan
E-mail: igarashiy@nms.ac.jp
https://doi.org/10.1272/jnms.JNMS.2024_91-205
Journal Website (https://www.nms.ac.jp/sh/jnms/)

requiring medical expertise, studies have been conducted using medical examinations for undergraduates and postgraduates worldwide. These evaluations included national medical licensing examinations in Japan[13–15] and the United States[10,16] and board certification examinations in neurology[17], nephrology[18], family medicine[19], general surgery[20], neurosurgery[21,22], orthopedic surgery[23], urology[24], plastic surgery[25], obstetrics and gynecology[26], anesthesiology[27], radiology[28], dermatology[29], ophthalmology[30], and otorhinolaryngology[31]. Earlier versions of ChatGPT had a correct answer rate of 30%-60%, i.e., mostly failing scores. However, performance has significantly improved since the release of ChatGPT-4. In studies comparing ChatGPT-3.5 and 4, ChatGPT-4 achieved higher scores in all examinations, with the score increasing by an average of approximately 20%, thus reaching the passing standard for many examinations.

However, except for cardiopulmonary resuscitation courses[32], there has been no evaluation of LLM performance in the field of emergency medicine. In addition, LLM performance on board certification examinations remains unclear. Therefore, we evaluated LLM performance on emergency medicine board certification examinations administered by the Japanese Association of Acute Medicine (JAAM).

Methods

JAAM Board Certification Examinations

The board certification examinations for emergency medicine conducted by the JAAM are divided into 3 parts: career as an emergency physician (10 points), record of medical practice (10 points), and written examination (80 points). A score of ≥70 points out of a total of 100 points is required to pass the board certification examination[33]. A score of 50 (62.5%) out of 70 points is required to pass the written examination. Each year, examinees are asked to answer 100 stand-alone and scenario-based questions by selecting 1 to 3 of the 5 options provided. Inappropriate questions are eliminated, and correct answers are published in the official JAAM journal. Since answers to older questions tend to change because of guideline updates and the accumulation of new evidence, the board certification examinations developed during a recent 5-year period (2018-2022) were used in this study.

LLM

ChatGPT-4 (May 24, 2023 Version, OpenAI, San Francisco, CA, USA) was used to answer the questions. This LLM utilizes self-attention mechanisms and extensive training data to produce natural language responses in conversational settings (**Fig. 1**). Similar studies excluded questions with images because ChatGPT could not process images. However, some scenario-based medical questions can be answered without using images by assuming the most probable disease. If ChatGPT was unable to answer a question because of the inclusion of images, that question was excluded. Because ChatGPT sometimes provides different answers to the same question, each question was entered into the ChatGPT interface by 2 independent raters (K.N. and Y.I.) to estimate accuracy and assess self-agreement in ChatGPT's response. A previous study on LLMs used similar methods, making the LLM answer the same questions twice and evaluating the agreement[27].

The primary outcome was the proportion of correct answers to the questions without images, under the same conditions as those encountered by examinees. Secondary outcomes included correct answers to all answerable questions, those with images, and those with stand-alone and scenario-based items. To assess robustness, agreement between the 2 sets of responses was evaluated using both kappa and simple agreement metrics. Since ChatGPT is language-independent and provides answers based on information, use of the Japanese language has no impact on the answers, and all questions were input using Japanese characters.

Categories of ChatGPT Errors

As in previous studies, incorrectly answered questions were classified into 4 primary categories: comprehension, factualness, specificity, and inference[34].

- *Comprehension errors* account for failures in understanding the question context and intent and often occur in the presence of grammar mistakes or ambiguity. For instance, misinterpretation of question intent was observed when the model failed to correctly address questions containing incorrect interrogative pronouns.
- *Factualness errors* arise when the model lacks the necessary supporting facts to produce an accurate answer, which can be attributed to the model's limited knowledge of specific entities, attributes, or events. These errors are straightforward and are a significant proportion of the model's errors.
- *Specificity errors* happen when the model fails to answer a problem at the correct level of specificity by providing answers that are too general or too detailed. For example, the model might categorize different subgenres of music as distinct genres, incorrectly claiming that they are not in the same genre.

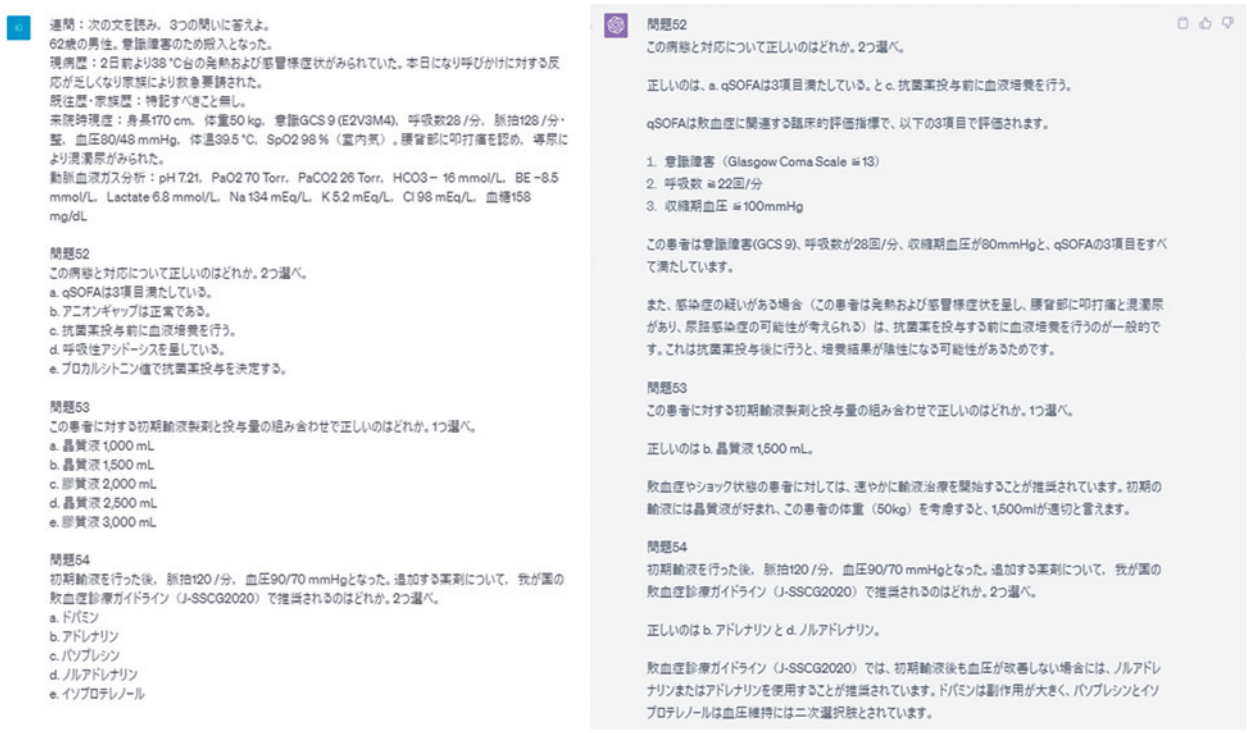

Fig. 1 List of sample questions
The questions are scenario-based questions on sepsis. ChatGPT answered each question with an explanation. On the basis of the latest guidelines, questions on infusion volume calculation and drug selection were correctly answered.

・*Inference errors* occur when the model, despite having the necessary knowledge, fails to reason effectively with the facts to derive the correct answer. This can happen when the model struggles to make predictions based on common sense or cannot correctly compare or analyze data.

We used these 4 categories to categorize the reasons for incorrect answers to questions without images.

Statistical Analyses

Categorical variables were analyzed with the chi-square test. A P value of <0.05 was considered to indicate statistical significance. The rate of agreement between 2 sets of responses was calculated using Cohen's kappa statistic, and statistical processing was conducted using R software version 4.3.0 (R Foundation for Statistical Computing, Vienna, Austria). The kappa accounts not only for mere agreement but also for agreement by chance[35].

Ethics statement

This study only used publicly available data, did not include any patient information, and did not affect patient safety. Therefore, the requirement of review by the Institutional Review Board was waived.

Results

Of the 475 questions with text choices, the LLM was able to answer 465 (**Fig. 2**). Ten questions could not be answered because the LLM was unable to process information from the images. Of the 342 questions without images administered to examinees and the LLM under the same conditions, 65.9% were answered correctly by the LLM, and a passing score of 62.5% was obtained. Of the 246 questions with images, 52.0% were answered correctly, even in the absence of any information about the image. This score was significantly lower than the score for questions without images ($p < 0.001$). Of all answerable questions, 62.3% were answered correctly.

For questions without images, the annual rates of correct answers ranged from 56.3% to 78.8% (**Table 1**). Moreover, the accuracy rate for scenario-based questions without images was 69.1%, which was higher than the rate for stand-alone questions (62.1%), although the difference was not significant ($p = 0.06$) (**Table 2**).

Each question was asked twice, and 72.5% elicited the same responses after 2 repetitions. The kappa value was 0.70, indicating moderate agreement[35].

Of the 233 incorrectly answered questions that did not include images, factual errors accounted for 191 (82%), inference errors for 36 (15%), and comprehension errors for 6 (3%).

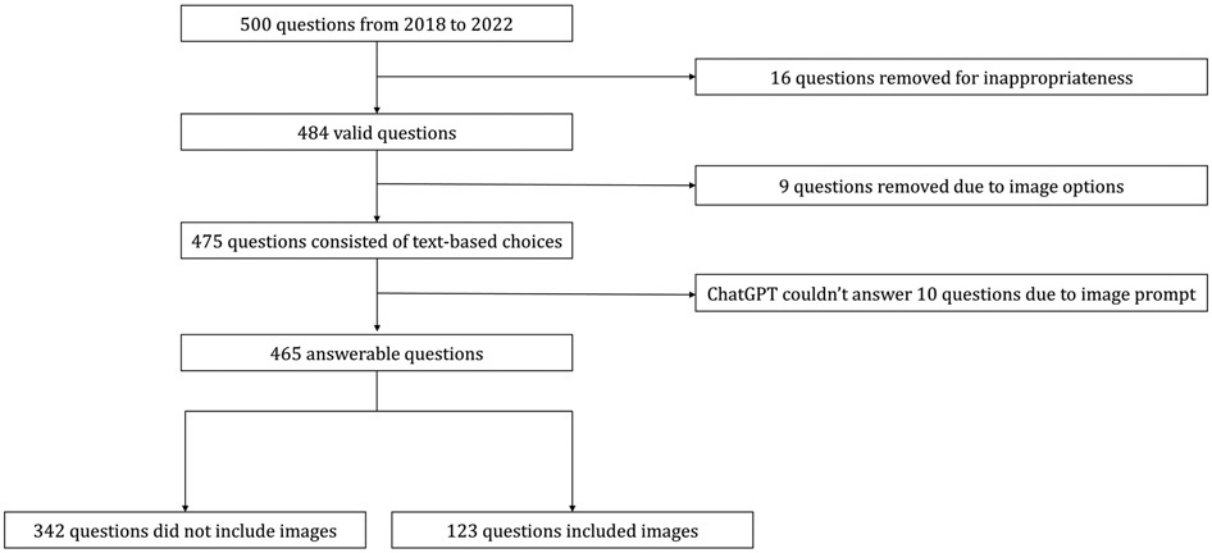

Fig. 2 Question flowchart

Table 1 Correct answers rates provided by ChatGPT-4

Year	Questions without images	Questions with images	Total
2022	104/132 (78.8%)	18/56 (32.1%)	122/188 (71.3%)
2021	67/118 (56.8%)	45/68 (66.2%)	112/186 (60.2%)
2020	89/134 (66.4%)	22/50 (44.0%)	111/184 (62.0%)
2019	101/140 (72.1%)	30/54 (55.6%)	131/194 (67.0%)
2018	90/160 (56.3%)	13/18 (72.2%)	103/178 (60.7%)
Total	451/684 (65.9%)	128/246 (52.0%)	579/930 (62.3%)

Table 2 Correct answers rates to stand-alone and scenario-based questions

Type	Stand-alone	Scenario-based
Questions without images	195/314 (62.1%)	256/ 370 (69.2%)
Questions with images	0/2 (0%)	128/244 (52.5%)
Total	195/316 (61.7%)	384/614 (62.5%)

Discussion

To our knowledge, this is the first study to use an emergency medicine board certification examination to evaluate an LLM. The findings demonstrated the considerable potential of LLMs to achieve a passing score on these examinations. Additionally, we assessed the inferencing capabilities of LLMs, including questions with images, and it nearly achieved a passing score for all answerable questions. To evaluate its robustness, we conducted the examinations twice and calculated the consistency rate, which showed moderate agreement.

The LLM had a high inference capability for emergency scenarios. In this study, after screening all questions, the scenario-based questions accounted for 25% of items. Of the 133 questions with images, 123 (92.5%) were answerable without image information, and 65 (52.8%) were answered correctly. Accuracy was greater for scenario-based problems than for stand-alone questions. In most previous studies, questions with images were excluded because the LLM could not use information from images and answered questions under the conditions encountered by the examinees. However, this result may reflect the fact that questions include enough textual information that a single diagnosis can be determined. Additionally, only 10% of incorrect answers resulted from inference errors. In a study of clinicopathological cases, which have been used since the 1950s to evaluate differential diagnosis generators, an LLM demonstrated a high clinical inference ability and provided correct diagnoses for the differential diagnosis of 64% of

challenging cases; in 39% of those, it provided the most likely diagnosis[9].

Although the performance of the present LLM was comparable to the level expected of board-certified doctors in emergency medicine, most incorrect answers were due to errors in factualness. Similar findings were observed in other board certification examinations[27]. In general, highly specialized models with questions that are answerable by yes or no answers are more accurate[36], whereas highly versatile models are less accurate[37]. An LLM is a multifunctional model that has not been trained in a specific domain; however, it is possible to fine-tune these models on specific tasks or domains to improve their performance in those areas.

A concern in emergency medicine is the potential for patients to use LLMs for self-diagnosis, as an LLM tends to "write plausible sounding but incorrect or nonsensical answers"[38]. It is possible for patients to be exposed to incorrect medical knowledge generated by LLMs. Currently, AI errors in medical knowledge require intervention from physicians to correct, as AI itself has a limited ability to self-correct errors. In contrast to another study, where self-agreement was close to 0.9[27], this study did not find a high level of agreement, indicating potential challenges in terms of robustness. It is important to exercise caution with ChatGPT's responses, as it may provide statements such as "I'm not a doctor and can't provide medical advice" or "Don't delay seeking medical attention."

This study has several limitations. First, the LLM may have already learned the questions published. However, the LLM's knowledge was based on information published up to September and it performed best on the latest examinations, i.e., those after that date—no performance degradation was observed. Second, although ChatGPT-4 was trained in various languages, including Japanese, its proficiency in non-English languages may not be equivalent to its proficiency in English. Third, the research showed that inference errors accounted for a small proportion of errors; however, this does not necessarily mean the AI is good at inference. It could be that the types of questions asked did not sufficiently challenge this skill.

Data availability statement: The data that support the findings of this study are available upon request from the corresponding author.

Author Contribution: YI, KN, and TN contributed to the design of the research. YI and KN collected data. YI, KN, TN, TT, and SY contributed to the interpretation of the results. YI and KN wrote the original draft of the manuscript. TN, TT, and SY critically reviewed the manuscript for important intellectual content. All authors read and approved the final manuscript to be published.

Conflict of Interest: The authors declare no conflict of interest for this article.

References

1. Tintinalli JE, Ma O, Yealy DM, et al. Tintinalli's emergency medicine: A comprehensive study guide. 9th ed. New York (NY): McGraw Hill; 2020.
2. Otaguro T, Tanaka H, Igarashi Y, et al. Machine learning for prediction of successful extubation of mechanical ventilated patients in an intensive care unit: A retrospective observational study. J Nippon Med Sch [Internet]. 2021 Nov 17;88(5):408–17. Available from: https://www.ncbi.nlm.nih.gov/pubmed/33692291
3. Igarashi Y, Nishimura K, Ogawa K, et al. Machine learning prediction for supplemental oxygen requirement in patients with COVID-19. J Nippon Med Sch [Internet]. 2022 May 12;89(2):161–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/34526457
4. Kitano S, Ogawa K, Igarashi Y, et al. Development of a machine learning model to predict cardiac arrest during transport of trauma patients. J Nippon Med Sch [Internet]. 2023 May 30;90(2):186–93. Available from: https://www.ncbi.nlm.nih.gov/pubmed/36823128
5. Will ChatGPT transform healthcare? Nat Med [Internet]. 2023 Mar;29(3):505–6. Available from: https://www.ncbi.nlm.nih.gov/pubmed/36918736
6. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. Lancet Digit Health [Internet]. 2023 Apr;5(4):e179–81. Available from: https://www.ncbi.nlm.nih.gov/pubmed/36894409
7. Ayers JW, Poliak A, Dredze M, et al. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. JAMA Intern Med [Internet]. 2023 Jun 1;183(6):589–96. Available from: https://www.ncbi.nlm.nih.gov/pubmed/37115527
8. Dahdah JE, Kassab J, Helou MCE, Gaballa A, Sayles S 3 rd, Phelan MP. ChatGPT: A valuable tool for emergency medical assistance. Ann Emerg Med [Internet]. 2023 Sep;82(3):411–3. Available from: https://www.ncbi.nlm.nih.gov/pubmed/37330721
9. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA [Internet]. 2023 Jul 3;330(1):78–80. Available from: https://www.ncbi.nlm.nih.gov/pubmed/37318797
10. 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 [Internet]. 2023 Feb 9;2(2):e0000198. Available from: https://www.ncbi.nlm.nih.gov/pubmed/36812645
11. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health [Internet]. 2023 Mar;5(3):e107–8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/36754724
12. Sabry Abdel-Messih M, Kamel Boulos MN. ChatGPT in clinical toxicology. JMIR Med Educ [Internet]. 2023 Mar 8;

9:e46876. Available from: https://www.ncbi.nlm.nih.gov/pubmed/36867743
13. Kaneda Y, Tanimoto T, Ozaki A, Sato T, Takahashi K. Can ChatGPT pass the 2023 Japanese National Medical Licensing Examination? Preprints [Preprint]. 2023 Mar;0191. Available from: https://doi.org/10.20944/preprints202303.0191.v1
14. Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. arXiv [Preprint]. 2023 Mar;18027. Available from: https://doi.org/10.48550/arXiv.2303.18027
15. 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 Jun 29;9:e48002.
16. Gilson A, Safranek CW, Huang T, 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 [Internet]. 2023 Feb 8;9:e45312. Available from: https://www.ncbi.nlm.nih.gov/pubmed/36753318
17. Giannos P. Evaluating the limits of AI in medical specialisation: ChatGPT's performance on the UK Neurology Specialty Certificate Examination. BMJ Neurol Open. 2023 Jun 15;5(1):e000451.
18. Noda R, Izaki Y, Kitano F, Komatsu J, Ichikawa D, Shibagaki Y. Performance of ChatGPT and Bard in Self-Assessment Questions for Nephrology Board Renewal. medRxiv [Preprint]. 2023 Jun 6;23291070. Available from: https://doi.org/10.1101/2023.06.06.23291070
19. Weng TL, Wang YM, Chang S, Chen TJ, Hwang SJ. ChatGPT failed Taiwan's Family Medicine Board Exam. J Chin Med Assoc. 2023 Aug 1;86(8):762−6.
20. Oh N, Choi GS, Lee WY. ChatGPT goes to the operating room: Evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. Ann Surg Treat Res [Internet]. 2023 May;104(5):269−73. Available from: https://www.ncbi.nlm.nih.gov/pubmed/37179699
21. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. Neurosurgery. 2023 Nov;93(5):1090−8.
22. Ali R, Tang OY, Connolly ID, et al. Performance of ChatGPT and GPT-4 on neurosurgery written board examinations. Neurosurgery. 2023 Dec 1;93(6):1353−65.
23. Lum ZC. Can artificial intelligence pass the American Board of Orthopaedic Surgery examination? Orthopaedic residents versus ChatGPT. Clin Orthop Relat Res. 2023 Aug 1;481(8):1623−30.
24. Huynh LM, Bonebrake BT, Schultis K, Quach A, Deibert CM. New artificial intelligence ChatGPT performs poorly on the 2022 Self-assessment Study Program for Urology. Urol Pract. 2023 Jul;10(4):409−15.
25. Gupta R, Herzog I, Park JB, et al. Performance of ChatGPT on the plastic surgery inservice training examination. Aesthet Surg J. 2023 Nov 16;43(12):NP1078−82.
26. Li SW, Kemp MW, Logan SJS, et al. ChatGPT outscored human candidates in a virtual objective structured clinical examination in obstetrics and gynecology. Am J Obstet Gynecol [Internet]. 2023 Aug;229(2):172.e1-12. Available from: https://www.ncbi.nlm.nih.gov/pubmed/37088277
27. Shay D, Kumar B, Bellamy D, et al. Assessment of ChatGPT success with specialty medical knowledge using anaesthesiology board examination practice questions. Br J Anaesth. 2023 Aug;131(2):e31−4.
28. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: Insights into current strengths and limitations. Radiology [Internet]. 2023 Jun;307(5):e230582. Available from: https://www.ncbi.nlm.nih.gov/pubmed/37191485
29. Passby L, Jenko N, Wernham A. Performance of ChatGPT on dermatology Specialty Certificate Examination multiple choice questions. Clin Exp Dermatol [Internet]. 2023 Jun 2;llad197. Available from: https://www.ncbi.nlm.nih.gov/pubmed/37264670
30. Teebagy S, Colwell L, Wood E, Yaghy A, Faustina M. Improved performance of ChatGPT-4 on the OKAP examination: A comparative study with ChatGPT-3.5. J Acad Ophthalmol (2017). 2023 Sep 11;15(2):e184−7.
31. Hoch CC, Wollenberg B, Luers JC, et al. ChatGPT's quiz skills in different otolaryngology subspecialties: An analysis of 2576 single-choice and multiple-choice board certification preparation questions. Eur Arch Otorhinolaryngol [Internet]. 2023 Sep;280(9):4271−8. Available from: https://www.ncbi.nlm.nih.gov/pubmed/37285018
32. Fijacko N, Gosak L, Stiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American heart association course? Resuscitation [Internet]. 2023 Apr;185:109732. Available from: https://www.ncbi.nlm.nih.gov/pubmed/36775020
33. Japanese Association for Acute Medicine. [Application for board-certified doctor of emergency and critical care medicine] [Internet]. Tokyo: Japanese Association for Acute Medicine. 2022 Dec 2 [cited 2023 Jun 10]. Available from: https://www.jaam.jp/info/2021/info-20211124.html. Japanese.
34. Zheng S, Huang J, Chang KC-C. Why does ChatGPT fall short in answering questions faithfully? arXiv [Preprint]. 2023 Apr;10513. Available from: https://doi.org/10.48550/arXiv.2304.10513
35. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) [Internet]. 2012;22(3):276−82. Available from: https://www.ncbi.nlm.nih.gov/pubmed/23092060
36. Jin Q, Dhingra B, Liu Z, Cohen W, Lu X. PubMedQA: A Dataset for Biomedical Research Question Answering. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019. Hong Kong, China: Association for Computational Linguistics; 2019. p. 2567−77.
37. Jin D, Pan E, Oufattole N, Weng W-H, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci [Internet]. 2021;11(14):6421. Available from: https://www.mdpi.com/2076-3417/11/14/6421
38. Brainard J. Journals take up arms against AI-written text. Science [Internet]. 2023 Feb 24;379(6634):740−1. Available from: https://www.ncbi.nlm.nih.gov/pubmed/36821673

(Received, August 17, 2023)
(Accepted, October 4, 2023)
(J-STAGE Advance Publication, March 2, 2024)