Journal of Anesthesia (2024) 38:282–283
https://doi.org/10.1007/s00540-023-03275-4

LETTER TO THE EDITOR

ChatGPT's performance on JSA-certified anesthesiologist exam

Michiko Kinoshita[1] · **Mizuki Komasaka**[1] · **Katsuya Tanaka**[1]

Received: 24 May 2023 / Accepted: 12 October 2023 / Published online: 30 October 2023

Keywords ChatGPT · Artificial intelligence · Large language model · Anesthesiology

To the Editor:

One notable recent development in artificial intelligence (AI) is ChatGPT, a large language model trained on extensive textual datasets, including websites, books, and articles. Using sophisticated deep learning algorithms, ChatGPT can comprehend and generate human-like text with remarkable proficiency across diverse tasks. The Japanese Society of Anesthesiologists (JSA) certifies professionals who complete a prescribed educational program through written, oral, and practical examinations. This study assessed ChatGPT's performance on the JSA-Certified Anesthesiologist written examination.

Tests were conducted using ChatGPT (OpenAI, San Francisco) between May 5 and 17, 2023, including the two models available: GPT-3.5 and GPT-4 (both models were updated from May 3 to 12 versions). The 2021 and 2022 JSA-Certified Anesthesiologist examinations were procured from *Masui*, the official JSA journal [1, 2]. The examination uses a mark-sheet format and comprises questions in two categories: general and clinical (simulating authentic clinical scenarios). In our tests, we excluded questions containing figures or tables, given ChatGPT's inability to process such content, as well as items officially removed by JSA because of errors. Queries were input verbatim in Japanese, and ChatGPT's responses were similarly furnished in Japanese.

We assessed the accuracy of both GPT-3.5 and GPT-4 when taking the 2021 and 2022 examinations. Additionally, to verify performance consistency (i.e., choice agreement rather than correctness), both models were re-tested on the 2022 examination. We compared independent categorical data through Chi-square or Fisher's exact tests and matched pairs via McNemar's test, setting statistical significance at $p < 0.05$.

Of the 200 questions in the 2021 examination, only 163 items (132 general, 31 clinical) were considered after excluding 30 items containing figures/tables, with 7 removed by JSA. Similarly, of the 135 questions in the 2022 examination, 37 items with figures/tables were excluded, resulting in 98 usable items (71 general, 27 clinical).

GPT-3.5 scored 23.3% and 21.4% on the 2021 and 2022 examinations, respectively. In contrast, GPT-4 scored 51.5% and 49.0%. Consequently, GPT-4's accuracy significantly surpassed that of GPT-3.5 ($p < 0.001$ for both years). Neither GPT-3.5 nor GPT-4 exhibited significant differences in accuracy between general and clinical questions (general vs. clinical: 22.7% vs. 22.4%, $p = 1.000$; 51.2% vs. 48.3%, $p = 0.804$, respectively) (Supplemental_Table_1). Most incorrect responses were deemed plausible, and only 1.0% (GPT-3.5) and 2.3% (GPT-4) of the answers were categorized as "beyond my knowledge."

After retaking the 2022 examination, GPT-4 demonstrated a higher consistency level ($p = 0.010$), as it agreed with its initial responses in 65.3% of cases, compared to GPT-3.5's agreement rate of 45.9%. Discriminating by initially correct or incorrect responses, the agreement levels were, respectively, 81.2% and 50.0% for GPT-4, and 76.2% and 37.7% for GPT-3.5 ($p = 0.002$, $p = 0.003$, respectively) (Supplemental_Table_2).

Previous evaluations of GPT-3.5 have reported accuracies of approximately 60% on the United States Medical Licensing Exam (USMLE) and 65–75% on the American Heart Association (AHA) Basic Life Support (BLS) and Advanced Cardiovascular Life Support (ACLS) examinations, which correspond to passing and failing scores, respectively [3, 4]. The lower accuracy on the JSA examination than on the USMLE and AHA BLS/ACLS examinations could potentially be attributed to two factors: a more specialized and

✉ Michiko Kinoshita
michiko-kinoshita@tokushima-u.ac.jp

[1] Department of Anesthesiology, Tokushima University Hospital, 2-50-1 Kuramoto-cho, Tokushima-shi, Tokushima 770-8503, Japan

challenging nature of the JSA examination, and a reduced performance of ChatGPT in non-English languages.

Two papers, both preprints at the time of writing this paper, that report investigations on ChatGPT's performance on the Japanese Medical Licensing Exam (JMLE) offer valuable insights [5, 6]. Kasai et al. reported accuracies of approximately 55% and 75–80% for GPT-3.5 and GPT-4, respectively, on JMLE in Japanese [5]. Similarly, Tanaka et al. found GPT-4's accuracy on JMLE to be approximately 80%, with a slight performance enhancement in optimally translated English compared to Japanese [6]. These references suggest that a lower accuracy of GPT-3.5 and GPT-4 could be expected on the JSA examination than on JMLE.

Unlike conventional rule-based chatbots, generative AI models like ChatGPT can generate new but potentially inconsistent responses in real time. As our results show, ChatGPT does not always return the right answer even after previously doing so. Furthermore, ChatGPT rarely responded with "beyond my knowledge," i.e., it was also confident about incorrect answers. Additionally, ChatGPT may create plausible content that either corresponds to incorrect or nonsensical answers unsupported by training data (artificial hallucinations) or lacks real-world nuances and subtleties (overgeneralization).

The potential of generative AIs like ChatGPT to transform the medical practice, including anesthesiology, by providing quick access to extensive medical information for aiding decision-making on diagnosis, treatment, and medical management is noteworthy. However, as cautioned by OpenAI, ChatGPT is not currently tailored to medical data and should not be used for medical practice. Our findings underscore the imperative for continuous refinement, particularly within specialized medical fields. The integration of domain-specific medical data could increase the accuracy and reliability of generative AI, but its application to medicine still demands users' medical expertise. Furthermore, we must address the ethical, legal, and social implications of using AI and develop guidelines accordingly.

Some limitations apply to this study. First, because our findings capture only a specific juncture in the model's evolution, subsequent advancements in AI technology may yield differing outcomes. Second, because we provided Japanese input and ChatGPT's performance may vary by language, results could differ if processed in other languages. Third, ChatGPT's accuracy is inherently variable because it generates new responses at each iteration. Fourth, while JSA's official examination reviews—accessible only to JSA members—indicate yearly fluctuations of 45–67% in average scores for newly created questions, the absence of publicly available passing scores complicates the interpretation of knowledge disparity between ChatGPT and certified anesthesiologists.

In conclusion, this study shows that GPT-3.5 and GPT-4 achieved accuracies of approximately 20% and 50%, respectively, on the JSA-Certified Anesthesiologist examination. Further technical refinements and the formulation of ethical and legal guidelines are pivotal for applying generative AI, such as ChatGPT, in medical settings.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00540-023-03275-4.

Acknowledgements We would like to express our gratitude to the JSA and KOKUSEIDO CO., LTD. for permitting us to use data from the written JSA-Certified Anesthesiologist examination.

Author contributions Conceptualization: MK, MK, and KT; Methodology: MK, and KT; Investigation: MK, MK; Analysis: MK, MK; Writing-original draft preparation: MK; Writing-reviewing and editing: MK, and KT. All authors read and approved the final manuscript.

Data availability The datasets used and analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest All authors declare that they have no conflicts of interest.

References

1. 60th Written Examination for the Japanese Society of Anesthesiologists-Certified Anesthesiologist in 2021 (with answers) (in Japanese). Masui (Jpn J Anesthesiol). 2022;71:185–214.
2. 61st Written Examination for the Japanese Society of Anesthesiologists-Certified Anesthesiologist in 2022 (with answers) (in Japanese). Masui (Jpn J Anesthesiol). 2023;72:164–87.
3. Kung TH, Cheatham M, Medenilla A, Sillos C, De Leon L, Elepaño C, Madriaga M, Aggabao R, Diaz-Candido G, Maningo J, Tseng V. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. PLoS Digit Health. 2023;2: e0000198.
4. Fijačko N, Gosak L, Štiglic G, Picard CT, John Douma M. Can ChatGPT pass the life support exams without entering the American heart association course? Resuscitation. 2023;185:109732.
5. Kasai J, Kasai Y, Sakaguchi K, Yamada Y, Radev D. Evaluating GPT-4 and ChatGPT on Japanese Medical Licensing Examinations. arXiv:2303.18027. https://doi.org/10.48550/arXiv.2303.18027.
6. Tanaka Y, Nakata T, Aiga K, Etani T, Muramatsu R, Katagiri S, Kawai H, Higashino F, Enomoto M, Noda M, Kometani M, Takamura M, Yoneda T, Kakizaki H, Nomura A. Performance of generative pretrained transformer on the national medical licensing examination in Japan. medRviv. 2023. https://doi.org/10.1101/2023.04.17.23288603.