NEJM AI

DOI: 10.1056/AIp2300031

PERSPECTIVE

Use of GPT-4 to Diagnose Complex Clinical Cases

Alexander V. Eriksen, M.D.,[1,2] Sören Möller, M.Sc., Ph.D.,[3,4] and Jesper Ryg, M.D., Ph.D.[1,2]

Received: July 10, 2023; Revised: September 15, 2023; Accepted: September 29, 2023; Published: Month 00, 2023

Abstract

We assessed the performance of the newly released AI GPT-4 in diagnosing complex medical case challenges and compared the success rate to that of medical-journal readers. GPT-4 correctly diagnosed 57% of cases, outperforming 99.98% of simulated human readers generated from online answers. We highlight the potential for AI to be a powerful supportive tool for diagnosis; however, further improvements, validation, and addressing of ethical considerations are needed before clinical implementation. (No funding was obtained for this study.)

Introduction

The combination of a shortage of physicians and the increased complexity in the medical field partly due to the rapidly expanding diagnostic possibilities already constitutes a significant challenge for the timely and accurate delivery of diagnoses. Given demographic changes, with an aging population this workload challenge is expected to increase even further in the years to come, highlighting the need for new technological development. AI has existed for decades and previously showed promising results within single modal fields of medicine, such as medical imaging.[1] The continuous development of AI, including the large language model (LLM) known as the Generative Pretrained Transformer (GPT), has enabled research in exciting new areas, such as the generation of discharge summaries[2] and patient clinical letters. Recently, a paper exploring the potentials of GPT-4 showed that it was able to answer questions in the U.S. Medical Licensing Examination correctly.[3] However, how well it performs on real-life clinical cases is less well understood. For example, it remains unclear to what extent GPT-4 can aid in clinical cases that contain long, complicated, and varied patient descriptions and how it performs on these complex real-world cases compared with humans.

We assessed the performance of GPT-4 in real-life medical cases by comparing its performance with that of medical-journal readers. Our study utilized available complex clinical case challenges with comprehensive full-text information published online between January 2017 and January 2023.[4] Each case presents a medical history and a poll with six options for the most likely diagnosis. To solve the case challenges, we provided GPT-4

The author affiliations are listed at the end of the article.

Dr. Eriksen can be contacted at alexander.viktor.eriksen@rsyd.dk or at University of Southern Denmark Faculty of Health Sciences, Department of Clinical Research, Geriatric Research Unit, Kløvervænget 10, Odense, Syddanmark, Denmark 5000.

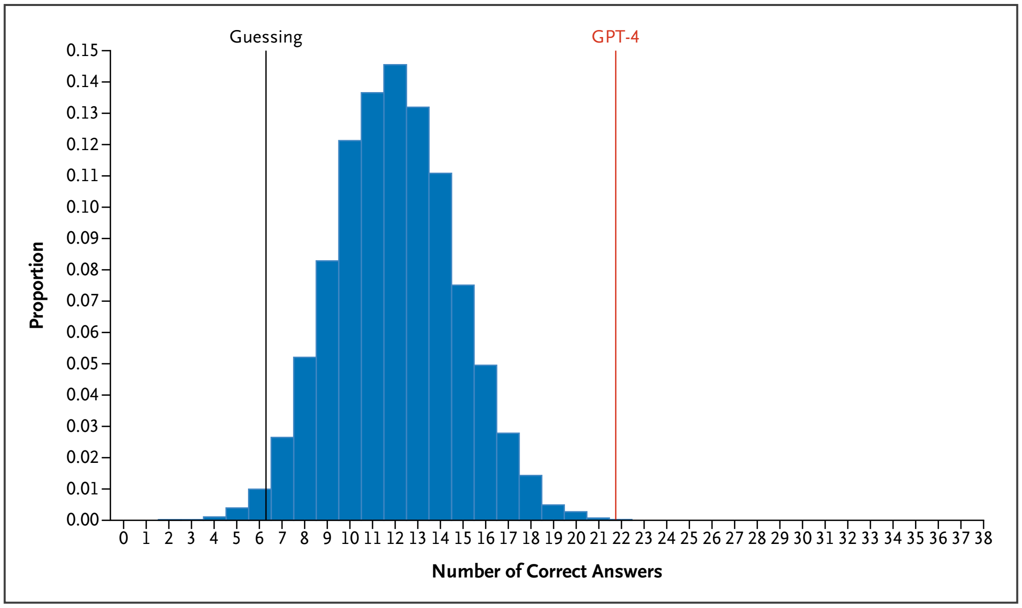

Figure 1. Number of Correct Answers of GPT-4 Compared with Guessing and a Simulated Population of Medical-Journal Readers.

Number of correct answers of GPT-4 (red line) to 38 multiple-choice real-world clinical case challenges compared with what would be expected by purely guessing with uniform probability for all answer possibilities (black line) and to the proportion of correct answers by a simulated population of 10,000 medical-journal readers (blue histogram).

with a prompt and a clinical case (see Supplementary Methods 1 in the Supplementary Appendix). The prompt instructed GPT-4 to solve the case by answering a multiple-choice question followed by the full unedited text from the clinical case report. Laboratory information contained in tables was converted to plain text and included in the case. The version of GPT-4 available to us could not accept images as input, so we added the unedited image description given in the clinical cases to the case text. The March 2023 edition of GPT-4 (maximum determinism: temp=0) was provided each case five times to assess reproducibility across repeated runs. This was also performed using the current (September 2023) edition of GPT-4 to test the behavior of GPT-4 over time. Because the applied cases were published online from 2017 to 2023 and GPT-4's training data include online material until September 2021, we furthermore performed a temporal analysis to assess the performance in cases before and after potentially available training data. For medical-journal readers, we collected the number and distribution of votes for each case. Using these observations, we simulated 10,000 sets of answers to all cases, resulting in a pseudopopulation of 10,000 generic human participants. The answers were simulated as independent Bernoulli-distributed variables (correct/incorrect answer) with marginal distributions as observed among medical-journal readers (see Supplementary Methods 2).

We identified 38 clinical case challenges and a total of 248,614 answers from online medical-journal readers.[4] The most common diagnoses among the case challenges were in the field of infectious disease, with 15 cases (39.5%), followed by 5 cases (13.1%) in endocrinology and 4 cases (10.5%) in rheumatology. Patients represented in the clinical cases ranged in age from newborn to 89 years old (median [interquartile range], 34 [18 to 57]), and 37% were female. The number of correct diagnoses among the 38 cases occurring by chance would be expected to be 6.3 (16.7%) due to the six poll options. The March 2023 edition of GPT-4 correctly diagnosed a mean of 21.8 cases (57%) with good reproducibility (55.3%, 57.9%, 57.9%, 57.9%, and 57.9%), whereas the medical-journal readers on average correctly diagnosed 13.7 cases (36%) (see Supplementary Table 1 and Supplementary Methods 1). GPT-4 correctly diagnosed 15.8 cases (52.7%) of those published up to September 2021 and 6 cases (75.0%) of those published after September 2021. Based on the simulation, we found that GPT-4 performed better than 99.98% of the pseudopopulation (Fig. 1). The September 2023 edition of GPT-4 correctly diagnosed 20.4 cases (54%).

Limitations

An important study limitation is the use of a poorly characterized population of human journal readers with unknown levels of medical skills. Moreover, we cannot assess whether the responses provided for the clinical cases reflect their maximum effort. Consequently, our results may represent a best-case scenario in favor of GPT-4. The assumption of independent answers on the 38 cases in our pseudopopulation is somewhat unrealistic, because some readers might consistently perform differently from others and the frequency at which participants respond correctly to the cases might depend on the level of medical skills as well as the distribution of these. However, even in the extreme case of maximally correlated correct answers among the medical-journal readers, GPT-4 would still perform better than 72% of human readers.

Conclusions

In this pilot assessment, we compared the diagnostic accuracy of GPT-4 in complex challenge cases to that of journal readers who answered the same questions on the Internet. GPT-4 performed surprisingly well in solving the complex case challenges and even better than the medical-journal readers. GPT-4 had a high reproducibility, and our temporal analysis suggests that the accuracy we observed is not due to these cases' appearing in the model's training data. However, performance did appear to change between different versions of GPT-4, with the newest version performing slightly worse. Although it demonstrated promising results in our study, GPT-4 missed almost every second diagnosis. Furthermore, answer options do not exist outside case challenges. However, a recently published letter reported research that tested the performance of GPT-4 on a closely related data set, demonstrating diagnostic abilities even without multiple-choice options.[5]

Currently, GPT-4 is not specifically designed for medical tasks. However, it is expected that progress on AI models will continue to accelerate, leading to faster diagnoses and better outcomes, which could improve outcomes and efficiency in many areas of health care.[1] Whereas efforts are in progress to develop such models, our results, together with recent findings by other researchers,[5] indicate that the current GPT-4 model may hold clinical promise today. However, proper clinical trials are needed to ensure that this technology is safe and effective for clinical use. Additionally, whereas GPT-4 in our study worked only on written records, future AI tools that are more specialized are expected to include other data sources, including medical imaging and structured numerical measurements, in their predictions. Importantly, future models should include training data from developing countries to ensure a broad, global benefit of this technology and reduce the potential for health care disparities. AI based on LLMs might be relevant not only for in-patient hospital settings but also for first-line screening that is performed either in general practice or by patients themselves. As we move toward this future, the ethical implications surrounding the lack of transparency by commercial models such as GPT-4 also need to be addressed,[1] as well as regulatory issues on data protection and privacy. Finally, clinical studies evaluating accuracy, safety, and validity should precede future implementation. Once these issues have been addressed and AI improves, society is expected to increasingly rely on AI as a tool to support the decision-making process with human oversight, rather than as a replacement for physicians.[1,3]

Disclosures

Author disclosures and other supplementary materials are available at ai.nejm.org.

No funding was obtained for the present study.

Author Affiliations

[1] Geriatric Research Unit, Department of Clinical Research, University of Southern Denmark, Odense, Denmark

[2] Department of Geriatric Medicine, Odense University Hospital, Odense, Denmark

[3] Open Patient data Explorative Network, OPEN, Odense University Hospital, Odense, Denmark

[4] Department of Clinical Research, University of Southern Denmark, Odense, Denmark

References

1. Haug CJ, Drazen JM. Artificial intelligence and machine learning in clinical medicine, 2023. N Engl J Med 2023;388:1201-1208. DOI: 10.1056/NEJMra2302038.

2. Patel SB, Lam K. ChatGPT: the future of discharge summaries? Lancet Digit Health 2023;5:e107-e108. DOI: 10.1016/S2589-7500(23)00021-3.

3. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med 2023;388:1233-1239. DOI: 10.1056/NEJMsr2214184.

4. The New England Journal of Medicine. Case challenges (https://www.nejm.org/case-challenges).

5. Kanjee Z, Crowe B, Rodman A. Accuracy of a generative artificial intelligence model in a complex diagnostic challenge. JAMA 2023; 330:78-80. DOI: https://doi.org/10.1001/jama.2023.8288.