Original Research

Canadian Association of Radiologists' Journal
2023, Vol. 0(0) 1–5
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/08465371231171125
journals.sagepub.com/home/caj
SAGE

Accuracy of Information and References Using ChatGPT-3 for Retrieval of Clinical Radiological Information

Matthias W. Wagner, MD[1,2], and Birgit B. Ertl-Wagner, MD, PhD, MHBA[1,2]

Abstract

Purpose: To assess the accuracy of answers provided by ChatGPT-3 when prompted with questions from the daily routine of radiologists and to evaluate the text response when ChatGPT-3 was prompted to provide references for a given answer. **Methods:** ChatGPT-3 (San Francisco, OpenAI) is an artificial intelligence chatbot based on a large language model (LLM) that has been designed to generate human-like text. A total of 88 questions were submitted to ChatGPT-3 using textual prompt. These 88 questions were equally dispersed across 8 subspecialty areas of radiology. The responses provided by ChatGPT-3 were assessed for correctness by cross-checking them with peer-reviewed, PubMed-listed references. In addition, the references provided by ChatGPT-3 were evaluated for authenticity. **Results:** A total of 59 of 88 responses (67%) to radiological questions were correct, while 29 responses (33%) had errors. Out of 343 references provided, only 124 references (36.2%) were available through internet search, while 219 references (63.8%) appeared to be generated by ChatGPT-3. When examining the 124 identified references, only 47 references (37.9%) were considered to provide enough background to correctly answer 24 questions (37.5%). **Conclusion:** In this pilot study, ChatGPT-3 provided correct responses to questions from the daily clinical routine of radiologists in only about two thirds, while the remainder of responses contained errors. The majority of provided references were not found and only a minority of the provided references contained the correct information to answer the question. Caution is advised when using ChatGPT-3 to retrieve radiological information.

Résumé

Objectif : Évaluer l'exactitude des réponses fournies par ChatGPT-3 en réaction à des questions de routine quotidienne des radiologistes et évaluer les réponses textes quand ChatGPT-3 est invité à fournir les références pour une réponse donnée. **Méthodes :** ChatGPT-3 (San Francisco, OpenAI) est un robot conversationnel (chatbot) utilisant une intelligence artificielle et s'appuyant sur un LLM ou grand modèle linguistique (Large Language Model) qui a été conçu pour générer un texte d'aspect naturel, c'est-à-dire produit par un humain. Un total de 88 questions a été soumis à ChatGPT-3 au moyen de messages textuels. Ces 88 questions étaient réparties de manière égale dans 8 domaines de sous-spécialités de la radiologie. La pertinence des réponses fournies par ChatGPT-3 a été évaluée en les recoupant avec des listes de référence PubMed revues par des pairs. De plus, l'authenticité des références fournies par ChatGPT-3 a été évaluée. **Résultats :** Un total de 59 réponses sur 88 (67 %) à des questions de radiologie étaient correctes, tandis que 29 réponses (33 %) contenaient des erreurs. Sur les 343 références fournies, seulement 124 (36,2 %) étaient disponibles après recherche sur le Web, tandis que 219 références (63,8 %) semblaient avoir être générées par ChatGPT-3. Après l'examen des 124 références identifiées, seulement 47 (37,9 %) ont été jugées comme procurant assez de contexte pour répondre correctement à 24 questions (37,5 %). **Conclusion :** Dans cette étude pilote, ChatGPT-3 n'a fourni des réponses correctes à des questions en rapport avec la routine clinique des radiologistes que dans seulement deux tiers des cas et le tiers restant des réponses contenait des erreurs. La majorité des références fournies n'a pas été trouvée et seulement une minorité de ces références contenait les informations correctes permettant de répondre aux questions. Il est donc conseillé de faire preuve de prudence lors de l'utilisation de ChatGPT-3 pour obtenir des informations relatives à la radiologie.

[1] Department of Diagnostic Imaging, Division of Neuroradiology, The Hospital for Sick Children, Toronto, Canada
[2] Department of Medical Imaging, University of Toronto, Toronto, Canada

Corresponding Author:
Matthias W. Wagner, Department of Diagnostic Imaging, Division of Neuroradiology, The Hospital for Sick Children, 555 University Ave, Toronto, ON M5G 1X8, Canada.
Email: m.w.wagner@me.com

Keywords
ChatGPT, AI, radiology, decision support, large language models

Introduction

Recently, large language models (LLMs) received much attention in scientific literature including radiology.[1-6] Probably, the most well-known model is OpenAI's GPT-3, which is based on generative pre-trained transformer (GPT) 3.5 with over 175 billion parameters.[2,7] ChatGPT-3 is an AI chatbot fine-tuned using reinforcement learning from human feedback to reward desired behavior and punish "toxic" text.[1,8,9] Within two months after its launch in November 2022, ChatGPT-3 has allowed over 100 million users to access and dialogue with its autoregressive LLM.[10] Several recent reports have shown promising performances of ChatGPT-3 including in an MBA degree exam[11] and in the United States Medical Licensing Exam.[12]

A recent pre-print study by Rao et al[4] evaluated ChatGPT-3 as an adjunct for radiologic decision-making. ChatGPT-3 achieved moderate scores when determining appropriate imaging modalities for various clinical presentations of breast cancer screening and breast pain.[4] Given the ability to generate text responses to complex input criteria,[4] we hypothesized that ChatGPT-3 is able to provide accurate answers to questions pertaining to clinical radiology.

The aim of our study was to assess the accuracy of answers provided by ChatGPT-3 when prompted with questions from the daily routine of radiologists. In addition, we aimed to evaluate the text response when ChatGPT-3 was prompted to provide references for a given answer.

Materials and Methods

ChatGPT-3

ChatGPT-3 (San Francisco, OpenAI) is an AI chatbot based on a LLM that has been designed to generate human-like text.[3] The language model is trained on approximately 570 GB of internet texts including books, articles, websites, and Wikipedia all limited to until 2021.[2,4] ChatGPT-3 does not search the internet when prompted to respond to user inputs. Instead, it employs a prediction algorithm that identifies the most likely "token" to succeed the previous one, based on patterns from its training data. Consequently, it does not duplicate existing information. All ChatGPT-3 model output was collected from the February 13, 2023, version of ChatGPT-3.

Model Input

A total of 88 questions were asked using textual prompt. These 88 questions were equally dispersed throughout 8 radiology sections including: Neuroradiology, pediatric radiology, gastrointestinal and genitourinary radiology, interventional radiology, cardiac radiology, musculoskeletal radiology, breast radiology, and chest wall and pulmonary radiology. Questions were randomly selected based on clinical experience. Questions were not systematically stratified into varying levels of difficulty. Six major categories of questions were created: 1) imaging findings of a condition (n = 15, 17%), 2) imaging findings of two differential diagnoses (n = 29, 33%), 3) modality-related questions (n = 11, 12.5%), 4) questions related to indications and contraindications (n = 7, 8%), 5) prognostic questions (n = 5, 5.7%), and 6) mixed content including differential diagnosis, anatomical questions, reference values, and complications (n = 21, 23.8%). All textual prompts are available in the supplementary table. New sessions were created for each question. Since the study used questions dispersed throughout 8 radiology sections, approval by the Ethics Committee and requirement for individual consent were not required.

Assessment of Responses

Each question was inputted once. Each response was assessed for the accuracy of the response using a 5-point Likert scale (1—incorrect, 2—some correct content, 3—approximately half correct content, 4—largely correct content, and 5—entirely correct content). Subsequently, a second textual prompt was provided consisting of the command: "Provide references for the answer above." Responses were then assessed for their number and accuracy (0—inaccurate and 1—accurate) and whether the cited article was indexed in the PubMed database (0—no and 1—yes). Last, in the event that a cited article was indexed, the article was accessed and a 5-point Likert scale was used to assess if the article or articles provided enough background to answer the initial question (1—no pertinent information, 2—some pertinent information, 3—approximately half pertinent information, 4—largely pertinent information, and 5—entirely pertinent information).

Statistical Analysis

Categorical data are presented as integers and percentages (%). Data were plotted using Microsoft Excel ® 2021.

Results

Responses to Questions

Out of 88 questions, 59 responses were correct (67%), 15 were largely correct (17%), 6 were approximately half correct (7%), 4 were mostly incorrect (4.5%), and 4 were entirely incorrect (4.5%). All questions and categorization of answers are available in the supplementary table.

Responses to Reference Prompt

For 88 reference prompts, a total of 343 references were provided. The median number of references per prompt was 4 (range: 1–6). Out of 343 references, 124 real references

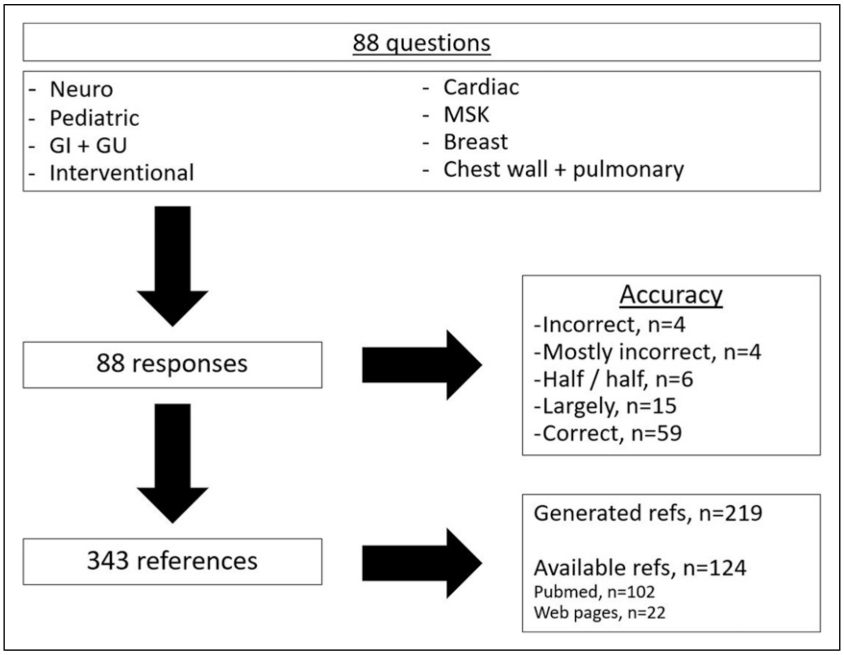

Figure 1. Schematic overview of the workflow. Textual prompts consisting of 88 questions from the daily routine of radiologists were created and responses were assessed for accuracy. After the prompt "Provide references for the answer above," 343 references were evaluated.

(36.2%) were available through internet search. The remainder of the references (n = 219, 63.8%) were unavailable through internet search and appeared to be generated by ChatGPT-3. References which appeared to be generated were not assessed systematically. Upon review of generated references, only a thorough internet search could prove that they were not authentic (example of reference number 1 to question number 1: Pascual-Castroviejo I, Roche C, Martínez-Bermejo A, et al. Polymicrogyria in tuberous sclerosis: a report of 14 patients and a review of the literature. Neuropediatrics. 2008; 39(1): 1–9. doi: 10.1055/s-2008-1038372). Of the 124 real references, 102 references (82.3%) were indexed in the PubMed database. The remainder of the identified references (n = 22, 17.7%) consisted of hyperlinks to various web pages (Figure 1).

Background to Answer Questions

The 124 identified references were distributed across 64 questions. A total of 47 references (37.9%) were considered to provide enough background to correctly answer 24 questions (37.5%). Two references (1.6%) were considered to provide most of the background needed to correctly answer one question (1.6%). Five references (4%) were considered to provide approximately half the background needed to correctly answer 5 questions (7.8%) and seven references (5.6%) were considered to provide little background needed to correctly answer 6 questions (9.4%). Interestingly, 16 references (12.9%) were provided by ChatGPT-3, which did not provide any background to answer 10 questions (15.6%). For the latter calculation, 4 questions were excluded since there were more references available potentially providing enough background. Furthermore, 19 questions were correctly answered although no identifiable reference was provided.

Discussion

In this pilot study investigating the accuracy of answers provided by ChatGPT-3 when prompted with 88 questions from the daily routine of radiologists, we found that only 59 responses (67%) were correct and that the remaining 29 responses (33%) had errors. We also assessed the response when ChatGPT-3 was prompted to provide references for a given answer. We found that out of 343 references provided, only 124 references (36.2%) were available through internet search, while 219 references (63.8%) appeared to be generated by ChatGPT-3. Furthermore, when examining those "true" references, we found that only 47 references (37.9%) were considered to provide enough background to correctly answer 24 questions (37.5%). Furthermore, 19 questions were correctly answered although no "true" reference was provided.

Given the ease of access and the ability to input individualized requests, it is tempting to consult ChatGPT-3 for information or problem-solving. In fact, few studies already investigated the abilities of ChatGPT-3 in the context of questions in daily clinical routine. Hirosawa et al[6] assessed the accuracy of differential diagnoses lists created by ChatGPT-3 compared to physicians for clinical histories with common presentation complaints. Authors found that the rate of correct diagnosis by ChatGPT-3 within the differential diagnoses lists was 83% compared to 93% correct physician's diagnosis. Rao et al[4] evaluated ChatGPT's capacity for clinical decision support in radiology. They compared ChatGPT's responses to the American College of Radiology Appropriateness Criteria for breast pain and breast cancer screening. Authors used an open-ended format to prompt ChatGPT to provide the single most appropriate imaging procedure and a select-all-that-apply format when ChatGPT was provided with a list of imaging modalities to assess. ChatGPT achieved moderate to good results for both breast pain and breast cancer screening. Rao et al listed some shortcomings of language models, which probably limited their results,[4] namely, the tendency or inability to relate facts to sources and the fabrication of information presented, aka "hallucinations." We also encountered these shortcomings in our study: Only 124 of 338 references (36.7%) were in fact "true" references. The remainder of the references were likely fabricated. In addition, 29 of 88 responses (33%) appeared to have also been fabricated.

In this study, 88 questions across 8 radiology subspecialties were submitted to ChatGPT-3. Questions covered common reading room scenarios including imaging findings of a condition, imaging findings of two differential diagnoses, modality-related questions, and questions related to prognosis, indications and contraindications. While the results of this study inform about the accuracy of ChatGPT-3's responses and the authenticity of references, there remain a number of unanswered questions. First, given that a third of questions were answered incorrectly and only 36% of its references are in fact authentic, what is the future role of ChatGPT-3 and other chatbots for radiologists? How can the user ensure that the medical information presented is accurate, complete, and robust to slight alterations of the input question? There is a need for further studies to address these questions also in the context of the newly released ChatGPT-4. In addition, users should keep in mind that any AI-powered chatbot is only as knowledgeable as the data they have been trained with.[6] ChatGPT-3 has been trained on data until 2020 with limited knowledge of events in 2021. Consequently, any radiologic study or textbook published in 2022 or later cannot be considered "training data." It is also important to note that the 88 questions submitted to ChatGPT-3 in this study do not represent the entire depth of radiology knowledge. In order to consider a certain (differential) diagnosis, radiologists first need to be trained to describe and recognize imaging findings and appreciate these findings in their clinical context.

ChatGPT-4 and future variants of LLMs might be able to address some of the questions and concerns mentioned above and if utilized correctly, AI-powered chatbots might have a positive impact in the reading room for example as problem-solving tools. However, users must be cognizant about the potential fabrication of incorrect information.

Our study has limitations that need to be taken into account when interpreting the results. Since the content of ChatGPT-3's training data is unknown to the public,[4] it is unclear which journal articles and web pages contributed to ChatGPT-3's answer to the question and prompt to provide references. In addition, the number of questions per radiological field was limited and the questions were not graded by difficulty. Future studies could also systematically analyze the effect of slight alterations of an input question on content of an answer and the authenticity of references.

Conclusion

In this pilot study, we demonstrated that ChatGPT-3 provided correct responses to questions from the daily clinical routine of radiologists in 67%, while 33% of answers had errors. When ChatGPT-3 was prompted to provide references for a given answer, only 36.7% of references were in fact "true," identifiable references, whereas 63.3% appear to have been generated by ChatGPT-3. Finally, we showed that only 37.9% of "true," identifiable references provided enough background for 37.5% of correctly answered questions. These results serve as a stark reminder of the limitations of ChatGPT-3 that must be considered when consulting ChatGPT-3 for clinically oriented questions. Caution is advised when using ChatGPT-3 to retrieve radiological information, especially in a clinical context.

Summary Statement

ChatGPT-3 provides correct responses to radiologic questions in only about two thirds. References are often not found. Only a minority of references contain information to answer the question. Caution is advised when using ChatGPT-3 to retrieve radiological information.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Matthias W. Wagner https://orcid.org/0000-0001-6501-839X

Supplemental Material

Supplemental material for this article is available online.

References

1. Kitamura FC. ChatGPT is shaping the future of medical writing but still requires human judgment. *Radiology*. 2023:230171.
2. Shen Y, Heacock L, Elias J, et al. ChatGPT and other large language models are double-edged swords. *Radiology*. 2023:230163.

3. Biswas S. ChatGPT and the future of medical writing. *Radiology*. 2023:223312.
4. Rao AS, Kim J, Kamineni M, Pang M, Lie W, Succi M. Evaluating ChatGPT as an adjunct for radiologic decision-making. *medRxiv*. 2023:2023. doi:10.1101/2023.02.02.23285399
5. Chavez MR, Butler TS, Rekawek P, Heo H, Kinzler WL. ChatGPT(Generative Pre-trained Transformer): Why we should embrace this technology. *Am J Obstet Gynecol*. 2023.
6. Hirosawa T, Harada Y, Yokose M, Sakamoto T, Kawamura R, Shimizu T. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: A pilot study. *Int J Environ Res Public Health*. 2023;20(4):3378.
7. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. arXiv preprint arXiv: 220302155. 2022.
8. Christiano PF, Leike J, Brown T, Martic M, Legg S, Amodei D. Deep reinforcement learning from human preferences. *Adv Neural Inf Process Syst*. 2017;30.
9. https://time.com/6247678/openai-chatgpt-kenya-workers//. Accessed March 20, 2023.
10. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app/. Accessed March 20, 2023.
11. Terwiesch C. Would Chat GPT get a Wharton MBA. A prediction based on its performance in the operations management course University of Pennsylvania. Mack Institute for Innovation Management at the Wharton School; 2023.
12. Kung T, Cheatham M, Medenilla A, et al. Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models. *PLOS Digit Health*. 2023; 2(2):e0000198.