HHS Public Access

Author manuscript

J Am Coll Radiol. Author manuscript; available in PMC 2024 February 16.

Published in final edited form as:

J Am Coll Radiol. 2023 October ; 20(10): 990–997. doi:10.1016/j.jacr.2023.05.003.

Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot

Arya Rao, BA[a,b], **John Kim, BA**[a,b], **Meghana Kamineni, BS**[a,b], **Michael Pang, BS**[a,b], **Winston Lie, BA, MSc**[a,b], **Keith J. Dreyer, PhD, DO**[c], **Marc D. Succi, MD**[d]

[a]Harvard Medical School, Boston, Massachusetts.

[b]Medically Engineered Solutions in Healthcare, Innovation in Operations Research Center, Massachusetts General Hospital, Boston, Massachusetts.

[c]Harvard Medical School, Boston, Massachusetts; Medically Engineered Solutions in Healthcare, Innovation in Operations Research Center, Massachusetts General Hospital, Boston, Massachusetts; Department of Radiology, Massachusetts General Hospital, Boston, Massachusetts; and Chief Data Science Officer and Chief Imaging Information Officer for Mass General Brigham, Boston, Massachusetts.

[d]Harvard Medical School, Boston, Massachusetts; Medically Engineered Solutions in Healthcare, Innovation in Operations Research Center and Associate Chair of Innovation & Commercialization, Mass General Brigham Enterprise Radiology; Executive Director, MESH Incubator. Massachusetts General Hospital, Boston, Massachusetts; and Department of Radiology, Massachusetts General Hospital, Boston, Massachusetts.

Abstract

Objective: Despite rising popularity and performance, studies evaluating the use of large language models for clinical decision support are lacking. Here, we evaluate ChatGPT (Generative Pre-trained Transformer)-3.5 and GPT-4's (OpenAI, San Francisco, California) capacity for clinical decision support in radiology via the identification of appropriate imaging services for two important clinical presentations: breast cancer screening and breast pain.

Methods: We compared ChatGPT's responses to the ACR Appropriateness Criteria for breast pain and breast cancer screening. Our prompt formats included an open-ended (OE) and a select all that apply (SATA) format. Scoring criteria evaluated whether proposed imaging modalities were in accordance with ACR guidelines. Three replicate entries were conducted for each prompt, and the average of these was used to determine final scores.

Results: Both ChatGPT-3.5 and ChatGPT-4 achieved an average OE score of 1.830 (out of 2) for breast cancer screening prompts. ChatGPT-3.5 achieved a SATA average percentage correct of 88.9%, compared with ChatGPT-4's average percentage correct of 98.4% for breast cancer screening prompts. For breast pain, ChatGPT-3.5 achieved an average OE score of 1.125 (out of 2)

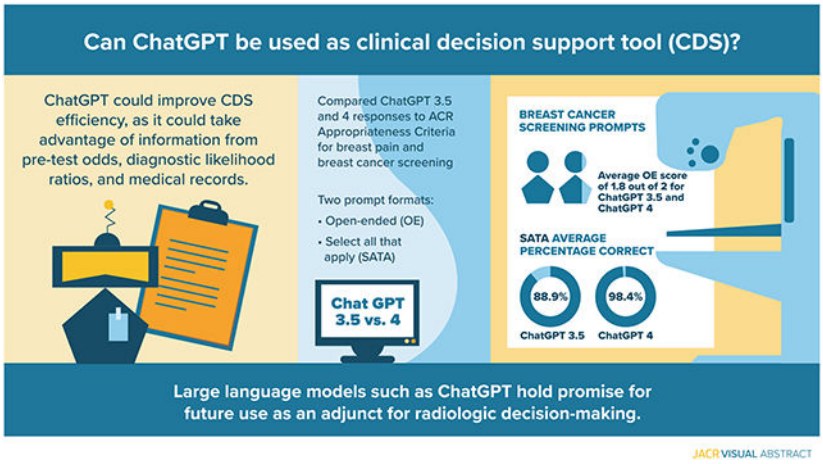
Corresponding author and reprints: Marc D. Succi, MD, Massachusetts General Hospital, Department of Radiology, 55 Fruit Street, Boston, MA 02114; msucci@mgh.harvard.edu. Follow this author via Twitter: Marc D. Succi @MarcSucciMD.

The authors state that they have no conflict of interest related to the material discussed in this article. The authors are non-partner/non-partnership track/employees.

and a SATA average percentage correct of 58.3%, as compared with an average OE score of 1.666 (out of 2) and a SATA average percentage correct of 77.7%.

Discussion: Our results demonstrate the eventual feasibility of using large language models like ChatGPT for radiologic decision making, with the potential to improve clinical workflow and responsible use of radiology services. More use cases and greater accuracy are necessary to evaluate and implement such tools.

Visual Abstract



Keywords

AI; breast imaging; ChatGPT; clinical decision making; clinical decision support

INTRODUCTION

Many artificial intelligence models have been utilized in health care to aid in clinical decision support (CDS), including for applications in radiology. Yet, use cases are mostly limited to imaging interpretation—in radiology, this includes identifying relevant image features and summarizing nonimage patient data, and reducing or removing barriers to patient care [1–6]. Other applications of AI in radiology focus on medical education, including providing real-time feedback to radiology trainees when providing diagnoses with radiology images [7]. In this study, we propose and evaluate a novel use of artificial intelligence in radiologic CDS: identifying appropriate imaging services based on an initial clinical presentation. We also compare the performance of two versions of the same artificial intelligence model to understand the performance differences across model updates.

ChatGPT (OpenAI, San Francisco, California) is a large language model (LLM) that is fine-tuned using supervised learning and reinforcement learning using human labels [8]. The two most recent versions of ChatGPT are built on Generative Pre-trained Transformer (GPT)-3.5 and GPT-4. Since its public release in November 2022, the chatbot from OpenAI has allowed over a million users to access and dialogue with its deep-learning, autoregressive LLM whose precursor (GPT-3) had at least 175 billion parameters under its training [9]. Despite several constraints, recent reports have shown promising ChatGPT

performance in the MBA degree examination [10], the United States Certified Public Accountant examination [11], the Uniform Bar examination [12], and the United States Medical Licensing Exam [13], demonstrating its ability to generate accurate textual responses to complex input criteria. ChatGPT's ability to simulate highly technical language has prompted researchers to credit the chatbot authorship in biomedical literature [14,15] with scientific journals scrambling to clarify publishing ethics in response [16–18]. The GPT model is evolving rapidly; no studies currently evaluate the latest model, GPT-4, in radiologic decision making.

Given ChatGPT's performance in these settings, we hypothesized that ChatGPT is capable of providing CDS in triaging patients for imaging services. To evaluate the feasibility of this application, we tested the accuracy of ChatGPT in determining appropriate imaging modalities for various clinical presentations of breast cancer screening and breast pain. Breast malignancies are a leading cause of cancer mortality in women, and their screening and diagnosis constitute significant imaging volume [19]. However, current practices of screening fall short of maximizing benefits for patients and reducing waste [20–22]. Similarly, mastalgia is a common complaint experienced by up to 70% of women in their lifetime [23]. Although only a significant minority presenting only with breast pain are associated with cancer, many will undergo diagnostic imaging, incurring ineffective health care resources [24]. To encourage responsible use of radiology services, the ACR has published various Appropriateness Criteria since 1993. Radiologists are often tasked with interpreting these guidelines, which classify patients into discrete demographic and risk groups. Integration of an AI-based tool into existing clinical workflows and systems could drastically improve efficiency, since such tools could take advantage of the wealth of information available from patient pretest odds, diagnostic likelihood ratios, and the medical records themselves.

Here, we explore a pilot for possible utility of ChatGPT in radiologic decision making, evaluating a variety of prompting mechanisms and clinical presentations to provide substantive performance metrics on possible use cases in clinical settings, such as by providers for CDS at the point of care.

We hypothesized that ChatGPT-3.5 would perform well in this use case. Since LLM models are being updated quickly, we also sought to evaluate how general model improvement translates to a specific clinical use case, and so we compared GPT-3.5 to GPT-4 performance. Substantial improvement in performance indicates that LLMs are quickly approaching readiness for use in the clinical setting and that continuous evaluation is necessary to keep pace with model progress.

METHODS

Artificial Intelligence

ChatGPT is a transformer-based language model that can generate humanlike text. ChatGPT features multiple layers of self-attention and feed-forward neural networks allowing it to capture the context and relationships between words in the input sequence. The language model is trained on diverse sources of text including websites, articles, and books but limited

to knowledge from current events until 2021 [25]. The ChatGPT model is self-contained in that it does not have the ability to search the Internet when generating responses. Instead, it predicts the most likely "token" (unit of text) to succeed the previous one based on patterns in its training data. Therefore, it does not explicitly search through existing information or copy existing information. When evaluating ChatGPT-3.5, all ChatGPT model output was collected from the January 9, 2023, version of ChatGPT. When evaluating ChatGPT-4, all output was collected from the April 10, 2023, version of ChatGPT.

ACR Criteria

The ACR Appropriateness Criteria for (1) breast pain and (2) breast cancer screening were selected as a ground truth comparison with ChatGPT. In brief, these criteria rate the clinical utility of diagnostic modalities given representative patient presentations in the context of breast pain and cancer. The appropriateness categories are "usually not appropriate," "may be appropriate," and "usually appropriate." Relative radiation level is also indicated for each modality on an ordinal scale; this information was not used in our study.

Model Input

We used two prompt formats for each ACR variant input into ChatGPT:

1. Open-ended (OE) format: Without providing the ACR list of imaging modalities, we asked ChatGPT to provide the "single most appropriate imaging procedure." This simulates how a user might actually interact with ChatGPT. Example: "For variant 'Breast cancer screening. Average-risk women: women with <15% lifetime risk of breast cancer,' determine the single most appropriate imaging procedure."
2. Select all that apply (SATA) format: We provided ChatGPT with the ACR list of imaging modalities for each variant and asked it to assess the appropriateness of each. This is in line with the actual usage of the ACR guidelines. Example: "For variant 'Breast cancer screening. Average-risk women: women with <15% lifetime risk of breast cancer,' assess appropriateness of the following procedures in a concise manner: mammography screening, digital breast tomosynthesis screening, US [ultrasound] breast, MRI breast without and with IV [intravenous] contrast, MRI breast without IV contrast, FDG-PET [PET positron emission tomography using fluorine-18-2-fluoro-2-deoxy-D-glucose imaging] breast dedicated, sestamibi MBI [molecular breast imaging]."

All prompts and outputs are available in the e-only Supplementary Data.

ChatGPT generates novel outputs with each input (even when the input is the same), and each output is informed by the context of the ongoing conversation. To avoid the influence of prior answers on model output, a new ChatGPT session was started for each prompt. To account for response-by-response variation, each prompt was tested three times, each time by a different user.

Workflow and Output Scoring

Each prompt was inputted three times, each time by a different user (three total replicates or outputs per prompt). Two scorers independently calculated an individual score for each output to confirm consensus on all output scores; there were no discrepancies between the scorers. The final score for each prompt was calculated as an average of the three replicate scores. A schematic of the workflow can be found in Figure 1, and scoring criteria can be found in Figure 2.

We reported the average of these raw scores for all variants. For SATA prompts, we also calculated the proportion of correct responses (the average of the raw scores divided by the maximum possible score for that variant). Since OE prompts should not yield answers that include the full spectrum of imaging options referenced by the ACR, the proportion of correct responses is not a valid metric for these outputs, thus we performed analysis on raw scores. All raw scores and additional statistical analysis can be found in the e-only Supplementary Data.

RESULTS

ChatGPT achieves moderate accuracy in radiologic decision making overall. Figures 3 and 4 show ChatGPT's performance on all variants tested. ChatGPT shows moderate accuracy on the whole—it achieved an average OE score of 1.83 (out of 2) and a SATA average percentage correct of 88.9% for breast cancer screening prompts and an average OE score of 1.125 (out of 2) and a SATA average percentage correct of 58.3% for breast pain prompts (Fig. 3, Fig. 4, e-only Supplementary Data).

ChatGPT also provided more reasoning in response to OE prompts, but more accuracy for SATA prompts. Qualitative analysis of the raw ChatGPT output shows that OE prompts tended to yield paragraph-form answers with rationale for the choice of imaging modality. Often, such rationale would reference key points made in the ACR guidelines [26], such as the degree of radiation exposure. Yet, SATA responses yielded a more complete picture of the imaging considerations—in the majority of cases, ChatGPT was able to identify not only appropriate imaging modalities but also inappropriate imaging modalities (SATA questions require the user to designate each option as appropriate or inappropriate) (Fig. 3, Fig. 4, e-only Supplementary Data).

ChatGPT achieved higher accuracy for breast cancer screening prompts than for breast pain prompts. Accuracy and precision were greater for breast cancer screening prompts than for breast pain prompts, as evidenced by higher scores and lower standard deviations across users in the former category (Fig. 3, Fig. 4, e-only Supplementary Data).

We also found that ChatGPT accuracy may vary with the severity of the initial presentation. For both the breast cancer screening and the breast pain variants, the variant number increases with the severity of the clinical presentation. For example, variant 1 for breast cancer screening references "average-risk women," and variant 3 references "high-risk women" [26]. Therefore, we were able to qualitatively examine trends in ChatGPT's triage ability. ChatGPT-3.5's accuracy for breast cancer screening variants decreased as

the severity of the clinical presentation increased (Fig. 4A), and its accuracy for breast pain variants increased as the severity increased (Fig. 3B, 4B). Notably, this trend does not hold for ChatGPT-4. Since model variability was not explicitly accounted for in our analysis, further investigation into the relationship between accuracy and severity is needed.

The newest version of ChatGPT, ChatGPT-4, is more capable of identifying opportunities for imaging stewardship than its predecessor. Variant 1 for breast pain was the only variant of those tested whose ACR recommendation was to decline imaging services entirely. Interestingly, for both OE and SATA prompts, ChatGPT-3.5 insisted on recommending imaging. Furthermore, ChatGPT provided multiple recommendations on occasion for OE prompts, even though all OE prompts asked specifically for “the single most appropriate” recommendation. These findings suggest that ChatGPT takes a maximalist approach in clinical decision making and is not well equipped to identify situations in which imaging is not indicated. Encouragingly, for one of the three breast pain variant 1 replicates tested in ChatGPT-4, the model recommended no imaging at all.

DISCUSSION

In this study, we provide first-of-its-kind evidence that LLMs hold promise for future use as an adjunct for radiologic decision making at the point of care. We show that ChatGPT displays impressive accuracy in identifying appropriateness of common imaging modalities for breast cancer screening and breast pain; this accuracy increases with the newest model, GPT-4. Given both the intricacy of radiologic decision making and the need for appropriate imaging utilization based only on initial clinical presentations, we believe this to be an impressive result. We also showcase the improved performance of ChatGPT-4 over ChatGPT-3.5, indicating that even generalized advances in model performance can lead to improvements in clinical use cases.

With an increasingly aging population and accessible imaging technologies, radiology imaging volumes are only expected to rise despite persisting concerns for low-value imaging [27–30]. Concomitant improvements in CDS infrastructure to enhance provider accuracy of appropriate imaging orders will emerge as a major priority and the involvement of ChatGPT-like AI is already being discussed [31].

Both ChatGPT-4 and ChatGPT-3.5 performed especially well when given a set of imaging options to evaluate (Fig. 3, Fig. 4, e-only Supplementary Data), consistently achieving >75% and >50% accuracy for breast pain SATA prompts and >95% and >80% for breast cancer screening SATA prompts across all prompts, respectively. This is consistent with possible use cases in the clinical setting—an emergency department, for example, will have a specific set of imaging modalities at its disposal, and a clinician is responsible for evaluating which of these is appropriate. ChatGPT’s performance on OE prompts was also encouraging, since it often provided extensive rationale for its recommendations, often in accordance with ACR recommendations. On OE prompts, ChatGPT-4 still outperformed ChatGPT-3.5, but the improvements were less substantial as compared with the improvements for SATA prompts. A hybrid approach, incorporating both a list of options for ChatGPT to evaluate and a request for ChatGPT to rationalize its choices, may provide optimal results in the clinical

setting. In addition, further prompt optimization and engineering could yield more accurate results.

Most notably, ChatGPT achieved impressive accuracy for breast cancer screening prompts (on average, 98.4% and 88.9% correct responses for SATA prompts for ChatGPT-4 and ChatGPT-3.5, respectively, and an average OE score of 1.830 for both ChatGPT versions). Given increased efforts to reduce overutilization of imaging services in this setting [32,33] and the high prevalence of breast cancer in the United States [34], this result is especially salient. For breast pain, ChatGPT-3.5 achieved an average OE score of 1.125 (out of 2) and a SATA average percentage correct of 58.3%, as compared with an average OE score of 1.666 (out of 2) and a SATA average percentage correct of 77.7%.

Some important limitations of our study involve the artificial intelligence model itself. ChatGPT is not free of the inherent limitations of language models: issues of alignment with user intent ("misalignment"), fabrication of information presented ("hallucinations"‘), and perhaps most arguably importantly in its potential clinical applications, inability to attribute factual information to a source. These limitations are reflected in ChatGPT-3.5's predilection toward providing more information than requested (multiple imaging modalities when just one was requested), recommending imaging in futile situations, providing incorrect rationale for incorrect imaging decisions, and, importantly, not distinguishing between imaging modalities with similar names but different applications (MRI with versus without IV contrast). However, many of these limitations were resolved in ChatGPT-4's output. For example, ChatGPT-4 correctly distinguished between MRI with versus without IV contrast and additionally correctly identified one clinical scenario in which imaging was not recommended according to ACR guidelines (variant 1, breast pain, SATA scoring). This indicates that some of the potential pitfalls of clinical implementation are readily solvable. These limitations must be considered when designing clinically oriented prompts for use with LLMs such as ChatGPT and when developing regulations for the use of artificial intelligence in clinical settings, including applicable approvals from agencies like the US FDA. Further investigation is necessary to evaluate the capacity of ChatGPT to determine whether or not imaging is required in a given clinical scenario. In addition, the use of replicate validation in our study, which we believe is unique relative to other ChatGPT studies, helps to identify limitations and reinforces accuracy; similar protocols should be adopted before clinical integration.

Because ChatGPT's training data are not public, it is not clear whether ChatGPT was trained on the ACR criteria before testing. However, given that this study is only concerned with the application of existing AI tools in radiologic decision making, it is inconsequential whether or not ChatGPT was trained on ACR guidelines—since ACR guidelines inform the standard of care, it is desirable that ChatGPT answers mirror existing guidelines and surprising that there is not complete concordance.

As artificial intelligence-based tools and LLMs specifically become more integrated with everyday use cases, we predict that specialized AI-based clinical decision-making tools will emerge. We believe that our study provides a critical data point in these endeavors,

identifying the surprising strengths of AI in determining appropriate diagnostic steps and highlighting weaknesses that need to be addressed in future iterations.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENT

The project described was supported in part by award Number T32GM144273 from the National Institute of General Medical Sciences. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of General Medical Sciences or the National Institutes of Health.

REFERENCES

1. Bizzo BC, Almeida RR, Michalski MH, Alkasab TK. Artificial intelligence and clinical decision support for radiologists and referring providers. J Am Coll Radiol 2019;16:1351–6. [PubMed: 31492414]
2. Witowski J, et al. MarkIt: a collaborative artificial intelligence annotation platform leveraging blockchain for medical imaging research. Blockchain Healthc Today 2021. 10.30953/bhty.v4.176.
3. Li MD, et al. Automated tracking of emergency department abdominal CT findings during the COVID-19 pandemic using natural language processing. Am J Emerg Med 2021;49:52–7. [PubMed: 34062318]
4. Kim D, et al. Accurate auto-labeling of chest X-ray images based on quantitative similarity to an explainable AI model. Nat Commun 2022;13:1867. [PubMed: 35388010]
5. Chung J, et al. Prediction of oxygen requirement in patients with COVID-19 using a pre-trained chest radiograph xAI model: efficient development of auditable risk prediction models via a fine-tuning approach. Sci Rep 2022;12:21164. [PubMed: 36476724]
6. Chonde DB, et al. RadTranslate: an artificial intelligence-powered intervention for urgent imaging to enhance care equity for patients with limited English proficiency during the COVID-19 pandemic. J Am Coll Radiol 2021;18:1000–8. [PubMed: 33609456]
7. Shah C, Davtyan K, Nasrallah I, Bryan RN, Mohan S. Artificial intelligence-powered clinical decision support and simulation platform for radiology trainee education. J Digit Imaging 2022. 10.1007/s10278-022-00713-9.
8. Ouyang L, et al. Training language models to follow instructions with human feedback. 2022. 10.48550/arXiv.2203.02155.
9. Brown TB, et al. Language models are few-shot learners. Available at: 10.48550/arXiv.2005.14165. Published 2020.
10. Terwiesch C Would Chat GPT3 get a Wharton MBA?
11. Bommarito J, Bommarito M, Katz DM, Katz J. GPT as knowledge worker: a zero-shot evaluation of (AI)CPA capabilities. Available at: 10.48550/arXiv.2301.04408. Published 2023.
12. Bommarito II M, Katz DM. GPT takes the bar exam. Available at: 10.48550/arXiv.2212.14402. Published 2022.
13. Kung TH, et al. Performance of ChatGPT on USMLE: potential for AI-assisted medical education using large language models. Available at: 10.1101/2022.12.19.22283643. Publihsed 2022.
14. Stokel-Walker C ChatGPT listed as author on research papers: many scientists disapprove. Nature 2023;613:620–1. [PubMed: 36653617]
15. Biswas S. ChatGPT and the future of medical writing. Radiology 2023. 10.1148/radiol.223312.
16. Flanagin A, Bibbins-Domingo K, Berkwits M, Christiansen SL. Nonhuman “authors” and implications for the integrity of scientific publication and medical knowledge. JAMA 2023. 10.1001/jama.2023.1344.
17. Thorp HH. ChatGPT is fun, but not an author. Science 2023;379. 313–313. [PubMed: 36701446]

18. Tools such as ChatGPT threaten transparent science; here are our ground rules for their use. Nature 2023;613:612–612. [PubMed: 36694020]
19. Lang M, et al. Imaging volume trends and recovery during the COVID-19 pandemic: a comparative analysis between a large urban academic hospital and its affiliated imaging centers. Acad Radiol 2020;27:1353–62. [PubMed: 32830030]
20. Sadigh G, et al. Downstream breast imaging following screening mammography in Medicare patients with advanced cancer: a population-based study. J Gen Intern Med 2018;33:284–90. [PubMed: 29139055]
21. Schonberg MA. Overutilization of breast cancer screening in the US: awareness of a growing problem. J Gen Intern Med 2018;33: 238–40. [PubMed: 29264700]
22. Habib AR, Grady D, Redberg RF. Recommendations from breast cancer centers for frequent screening mammography in younger women may do more harm than good. JAMA Intern Med 2021;181: 588–9. [PubMed: 33720278]
23. Goyal A. Breast pain. BMJ Clin Evid 2014;2014:0812.
24. Kushwaha AC, et al. Overutilization of health care resources for breast pain. AJR Am J Roentgenol 2018;211:217–23. [PubMed: 29792736]
25. What is ChatGPT?. OpenAI. Available at: https://help.openai.com/en/articles/6783457-chatgpt-general-faq. Accessed February 1, 2023.
26. American College of Radiology. ACR Appropriateness Criteria®. Available at: https://www.acr.org/Clinical-Resources/ACR-Appropriateness-Criteria.
27. Kjelle E, et al. Characterizing and quantifying low-value diagnostic imaging internationally: a scoping review. BMC Med Imaging 2022;22:73. [PubMed: 35448987]
28. Lee AH, et al. CT utilization in evaluation of skin and soft tissue extremity infections in the ED: retrospective cohort study. Am J Emerg Med 2023;64:96–100. [PubMed: 36502653]
29. Virji AZ, et al. Analysis of self-initiated visits for cervical trauma at urgent care centers and subsequent emergency department referral. Clin Imaging 2022;91:14–8. [PubMed: 35973271]
30. Succi MD, et al. Increased per-patient imaging utilization in an emergency department setting during COVID-19. Clin Imaging 2021;80:77–82. [PubMed: 34274685]
31. Shen Y, et al. ChatGPT and other large language models are double-edged swords. Radiology 2023. 10.1148/radiol.230163.
32. Sharma R, et al. Factors influencing overuse of breast cancer screening: a systematic review. J. Womens Health 2018;27: 1142–51.
33. Austin JD, et al. A mixed-methods study of multi-level factors influencing mammography overuse among an older ethnically diverse screening population: implications for de-implementation. Implement Sci Commun 2021;2:110. [PubMed: 34565481]
34. Giaquinto AN, et al. Breast cancer statistics, 2022. CA Cancer J Clin 2022;72:524–41. [PubMed: 36190501]

TAKE-HOME POINTS

- When provided with a list of possible imaging modalities, ChatGPT-4 achieves 98.4% accuracy for breast cancer screening recommendations and 77.7% accuracy for breast pain recommendations, and ChatGPT-3.5 achieves 88.9% accuracy for breast cancer screening recommendations and 58.3% accuracy for breast pain recommendations.
- ChatGPT-4 shows remarkable improvement in performance on radiology clinical decision-making tasks as compared with ChatGPT-3.5, despite the fact that the GPT models are generalized.
- ChatGPT-3.5 takes a maximalist approach to imaging recommendations, often recommending more imaging modalities than requested by user prompt and failing to identify situations in which imaging is futile. This trend is less pronounced for ChatGPT-4.
- For breast cancer screening, ChatGPT-3.5 performance increases with the severity of initial clinical presentation. For breast pain, ChatGPT-3.5 performance decreases with severity. This trend is less pronounced for ChatGPT-4.

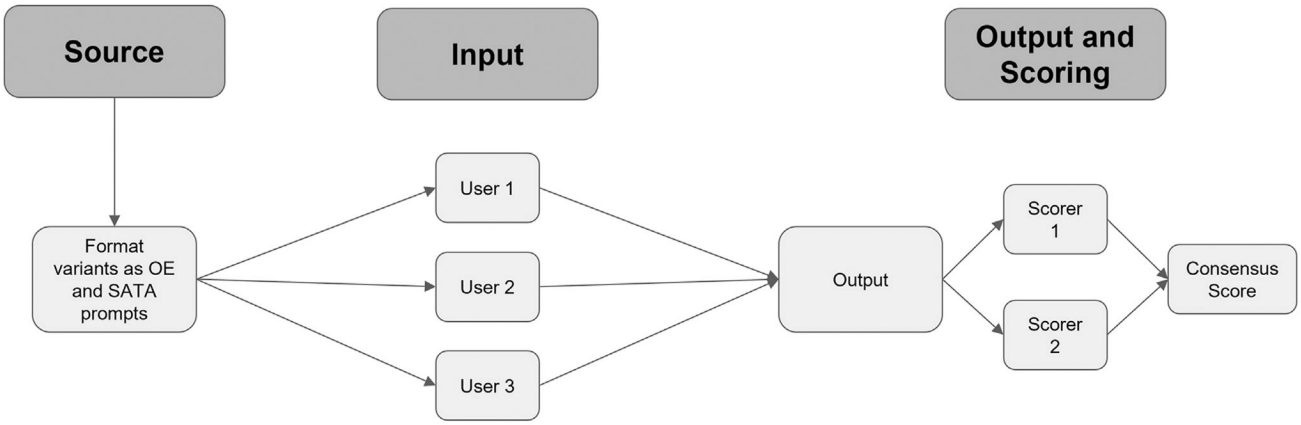


Fig. 1.

Schematic of experimental workflow. Prompts were developed from ACR variants for breast cancer screening and breast pain and converted to OE and SATA formats. Three independent users tested each prompt. Two independent scorers calculated scores for all outputs; these were compared to generate a consensus score. OE = open-ended; SATA = select all that apply.

OE Prompt Scoring Rubric	
ChatGPT-recommended Imaging Procedure is "Usually Not Appropriate" according to ACR criteria	0
ChatGPT-recommended Imaging Procedure "May Be Appropriate" according to ACR criteria	1
ChatGPT-recommended Imaging Procedure is "Usually Appropriate" according to ACR criteria	2
If ChatGPT outputted multiple imaging procedures, calculate the above score for each procedure given. The total score is the average of the individual procedure scores in these cases.	

SATA Prompt Scoring Rubric	
ChatGPT classifies an Imaging Procedure as appropriate and ACR criteria states that it is "Usually Appropriate" or "May Be Appropriate" according to ACR criteria	1
ChatGPT classifies an Imaging Procedure as inappropriate and ACR criteria states that it is "Usually Not Appropriate" according to ACR criteria	1
ChatGPT classifies an Imaging Procedure as inappropriate and ACR criteria states that it is "Usually Appropriate" or "May Be Appropriate" according to ACR criteria	0
ChatGPT classifies an Imaging Procedure as appropriate and ACR criteria states that it is "Usually Not Appropriate" according to ACR criteria	0
Imaging procedure listed by ACR is not listed by ChatGPT	0
For each imaging procedure listed by ChatGPT in a given output, add the designated score to the cumulative total score for that output. The maximum possible score for any output is equal to the number of imaging procedures evaluated in the respective ACR variant.	

Fig. 2.

Scoring criteria for OE and SATA prompts. Answers to OE prompts were scored on a 0 to 2 scale, in accordance with the ACR metrics for imaging appropriateness. If multiple imaging modalities were provided for a single prompt, an individual raw score was calculated for each modality, and these were averaged. Answers to SATA prompts were scored on a point or no point basis for each imaging modality provided. The maximum possible SATA score for a given variant was equal to the number of imaging procedures evaluated in the ACR criteria. OE = open-ended; SATA = select all that apply.

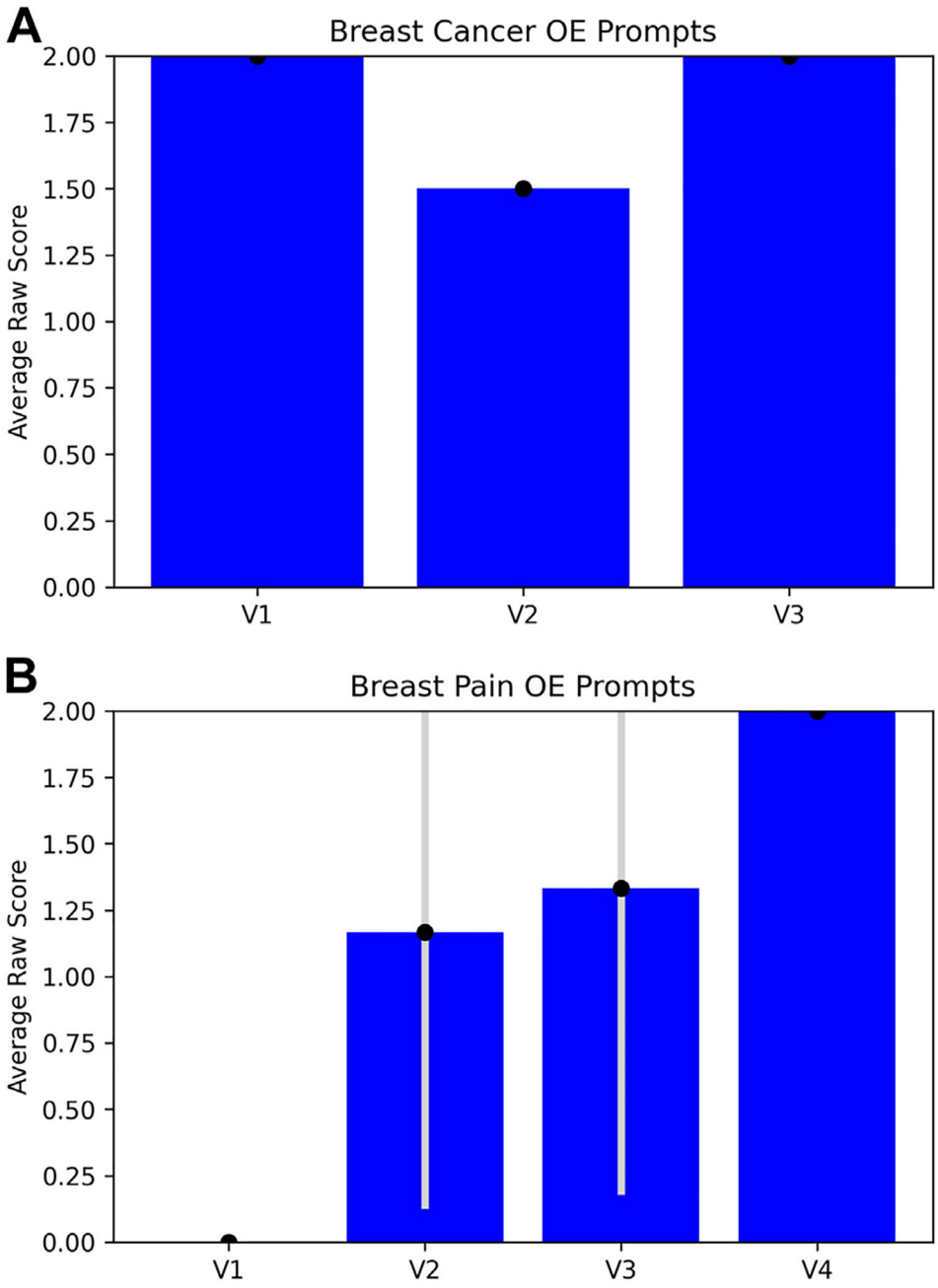


Fig. 3.

Performance of ChatGPT on OE prompts for breast cancer screening variants (A) and breast pain variants (B). OE performance was measured by the average raw score of the three replicate output scores for each variant (labeled according to the numbering in the ACR criteria variants). Error bars are ±1 standard deviation between the three replicate output scores. OE = open-ended; V1 = variant 1; V2 = variant 2; V3 = variant 3.

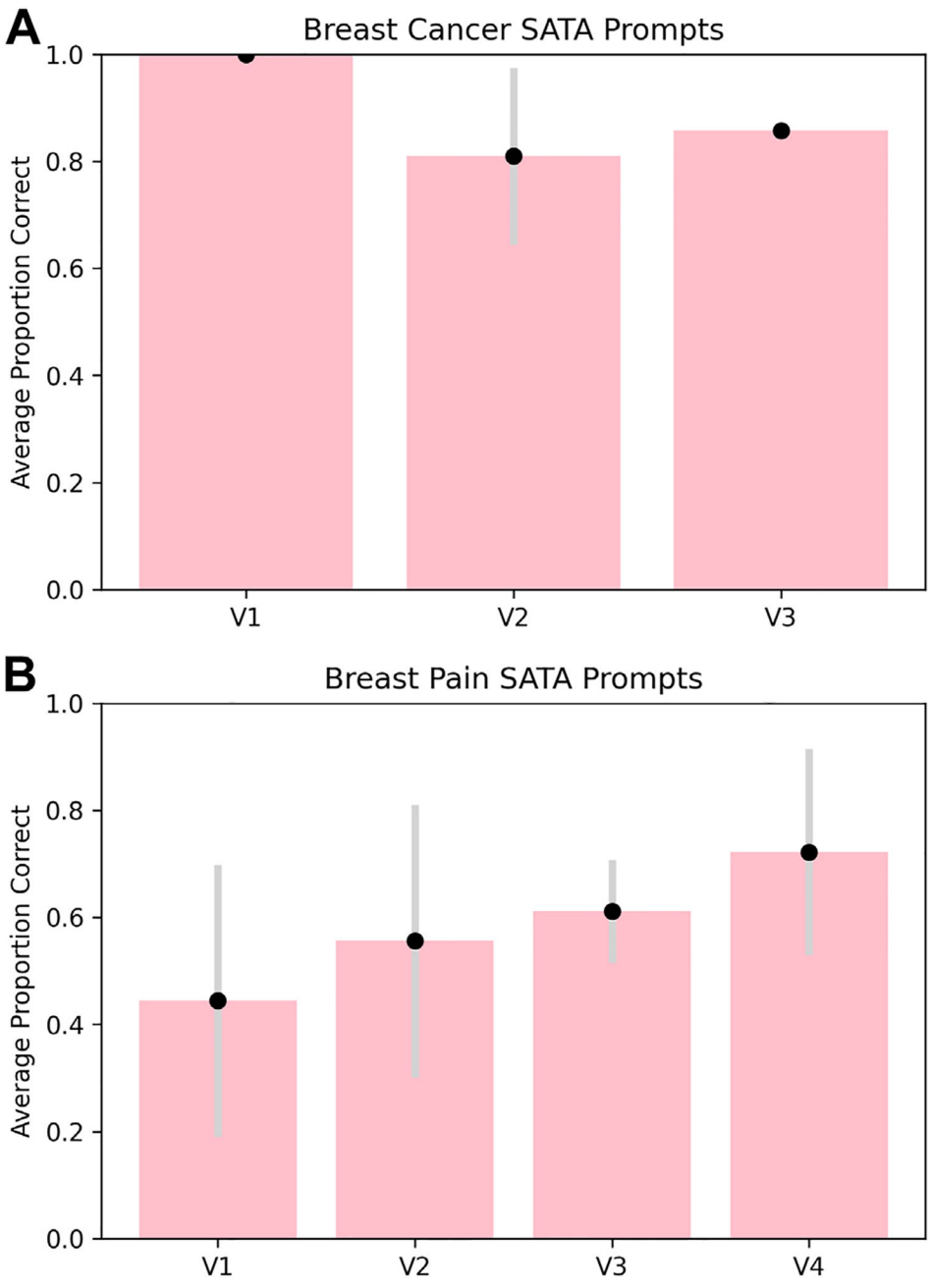


Fig. 4.

Performance of ChatGPT on SATA prompts for breast cancer screening variants (A) and breast pain variants (B). SATA performance was measured by the average proportion of correct answer selections for each variant from the three replicate output scores. Error bars for both prompt types are ±1 standard deviation between the three replicate output scores. SATA = select all that apply; V1 = variant 1; V2 = variant 2; V3 = variant 3; V4 = variant 4.