Radiology

ORIGINAL RESEARCH • COMPUTER APPLICATIONS

A Context-based Chatbot Surpasses Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines

Alexander Rau, MD • Stephan Rau, MD • Daniela Zöller, PhD • Anna Fink, MD • Hien Tran, MD • Caroline Wilpert, MD • Johanna Nattenmüller, MD • Jakob Neubauer, MD • Fabian Bamberg, MD • Marco Reisert, PhD • Maximilian F. Russe, MD

From the Department of Diagnostic and Interventional Radiology (A.R., S.R., A.F., H.T., C.W., J. Nattenmüeller, J. Neubauer, F.B., M.F.R.), Department of Neuroradiology (A.R.), Medical Physics, Department of Diagnostic and Interventional Radiology (M.R.), and Department of Stereotactic and Functional Neurosurgery (M.R.), Medical Center–University of Freiburg, Faculty of Medicine, University of Freiburg, Breisacher Str 64, 79106 Freiburg, Germany; Institute of Medical Biometry and Statistics, Faculty of Medicine and Medical Center–University of Freiburg, Freiburg, Germany (D.Z.); and Freiburg Center for Data Analysis and Modelling, University of Freiburg, Freiburg, Germany (D.Z.). Received April 19, 2023; revision requested May 16; final revision received June 25; accepted July 11. **Address correspondence to** M.F.R. (email: *maximilian.russe@uniklinik-freiburg.de*).

A.R. supported by Berta-Ottenstein-Programme for Clinician Scientists, Faculty of Medicine, University of Freiburg.

Conflicts of interest are listed at the end of this article.

Radiology 2023; 308(1):e230970 • https://doi.org/10.1148/radiol.230970 • Content code: AI

Background: Radiologic imaging guidelines are crucial for accurate diagnosis and optimal patient care as they result in standardized decisions and thus reduce inappropriate imaging studies.

Purpose: To investigate the potential to support clinical decision-making using an interactive chatbot designed to provide personalized imaging recommendations from American College of Radiology (ACR) appropriateness criteria documents using semantic similarity processing.

Materials and Methods: The authors used 209 ACR appropriateness criteria documents as a specialized knowledge base and used LlamaIndex, a framework for connecting large language models with external data, and ChatGPT-3.5-turbo to create an appropriateness criteria context aware chatbot (accGPT). Fifty clinical case files were used to compare the performance of accGPT with that of general radiologists at varying experience levels and to generic ChatGPT-3.5 and 4.0.

Results: The performance of all chatbots reached at least that of humans. For the 50 case files, accGPT performed best in providing correct recommendations that were "usually appropriate" according to the ACR criteria and also provided the highest proportion of consistently correct answers in comparison with the generic chatbots and radiologists. Furthermore, the chatbots provided substantial time and cost savings, with an average decision time of 5 minutes and a cost of €0.19 ($0.21) for all cases, compared with 50 minutes and €29.99 ($33.24) for radiologists (both $P < .01$).

Conclusion: ChatGPT-based algorithms have the potential to substantially improve the decision-making for clinical imaging studies in accordance with ACR guidelines. Specifically, the performance of a context-based algorithm was superior to that of its generic counterpart, demonstrating the value of tailoring artificial intelligence solutions to specific health care applications.

Supplemental material is available for this article.

The increasing demand for accurate and efficient diagnostic imaging in modern health care requires standardized criteria for decision-making. The American College of Radiology (ACR) provides such recommendations since 1994 that streamline decision-making processes for referring clinicians (1). This allows for optimized patient care, improved diagnostic accuracy, and lower radiation exposure and reduces health care costs *(https://www.acr.org/Clinical-Resources/ACR-Appropriateness-Criteria)*. However, variability in clinical routine still persists regarding the requirement of imaging itself, modality, and need for contrast material, leading to a substantial amount of inappropriate imaging procedures (2–5).

Adherence to appropriateness guidelines is highly dependent on clinician and radiologist experience and hampered by a lack of awareness (3,6–8). In addition, the rapid evolution of imaging technology and the progress of clinical evidence necessitate continuous updates to the recommendations, further complicating their usage.

Several clinical decision support tools were introduced to improve the use of published guidelines (eg, iGuide [European Society of Radiology, *https://www.myesr.org/esriguide*] or CareSelect Imaging [Change Healthcare, *https://www.changehealthcare.com/clinical-decision-support/careselect/imaging*]) (7,9). These tools have proven valuable in improving the diagnostic management of patients with various clinical conditions, resulting in lower rates of inappropriate examinations (2,10). However, they require substantial human interaction while potentially losing relevant clinical information as a free-text input is not feasible (2).

Abbreviations

accGPT = appropriateness criteria context aware chatbot, ACR = American College of Radiology, AI = artificial intelligence, OR = odds ratio

Summary

A chatbot with specific knowledge of imaging appropriateness guidelines reliably provided recommendations for imaging requirement and modality and outperformed generic chatbots and radiologists at varying levels of experience.

Key Results

- Large language model–based chatbots reliably provided clinical decision support for imaging appropriateness and modality according to the guidelines of the American College of Radiology.
- The performance of an appropriateness criteria context aware chatbot (accGPT) with specific knowledge of the appropriateness criteria documents was superior to that of generic ChatGPT versions and radiologists; accGPT provided the highest proportion of "usually appropriate" answers and also the highest consistency.
- All chatbots induced substantial time and cost savings compared with the radiologists.
- Interactive chatbots have great potential to support clinical decision-making in radiologic routine while on the other hand, the authors corroborate the need for health care–tailored solutions.

Artificial intelligence (AI)–based algorithms using large language models can address these limitations by comprehending and interpreting human language. OpenAI introduced ChatGPT to a wide audience in November 2022. ChatGPT is a chatbot specifically trained for conversation and is based on a generative pretrained transformer using the latest version of GPT-3.5-turbo (11). ChatGPT, and especially GPT-4 (12) (released in March 2023, with at the moment still limited access), were shown to provide substantial medical knowledge being able to pass the U.S. Medical Licensing Examination (13,14).

ChatGPT enables rapid processing of complex information, and its potential applications in clinical radiology routines have been extensively explored and published in preprints. These comprise the preparation of radiologic reports (15), transferring radiologic reports into plain language by simplifying them (16,17), and providing clinical decision support on differential diagnoses, diagnostic procedures, final diagnosis, and treatment (18).

The community is already discussing the potential of ChatGPT to provide recommendations regarding imaging as clinical decision support (19), and initial data on breast cancer screening and breast pain are promising (20). However, only the respective description of the clinical conditions in the ACR guidelines were presented in the latter study (no clinical files were used) and the comparison to human performance is lacking. Furthermore, ChatGPT is limited to its training data (GPT- 3.5-turbo and GPT-4 were trained on data up to September 2021) and thus may not have access to the latest and most specialized knowledge. In addition, ChatGPT may be biased due to the large number of different sources of training data. Therefore, ChatGPT could provide incorrect or incomplete information.

Incorporating specialized knowledge to create an appropriateness criteria context aware GPT (accGPT) should provide more accurate and relevant responses to user queries. To explore this approach, we compared the performance of general radiologists of varying experience levels and publicly available generic chatbots GPT-3.5-turbo and GPT-4 against that of the accGPT chatbot built on GPT-3.5-turbo and enhanced with knowledge of the ACR Appropriateness Criteria using semantic similarity processing.

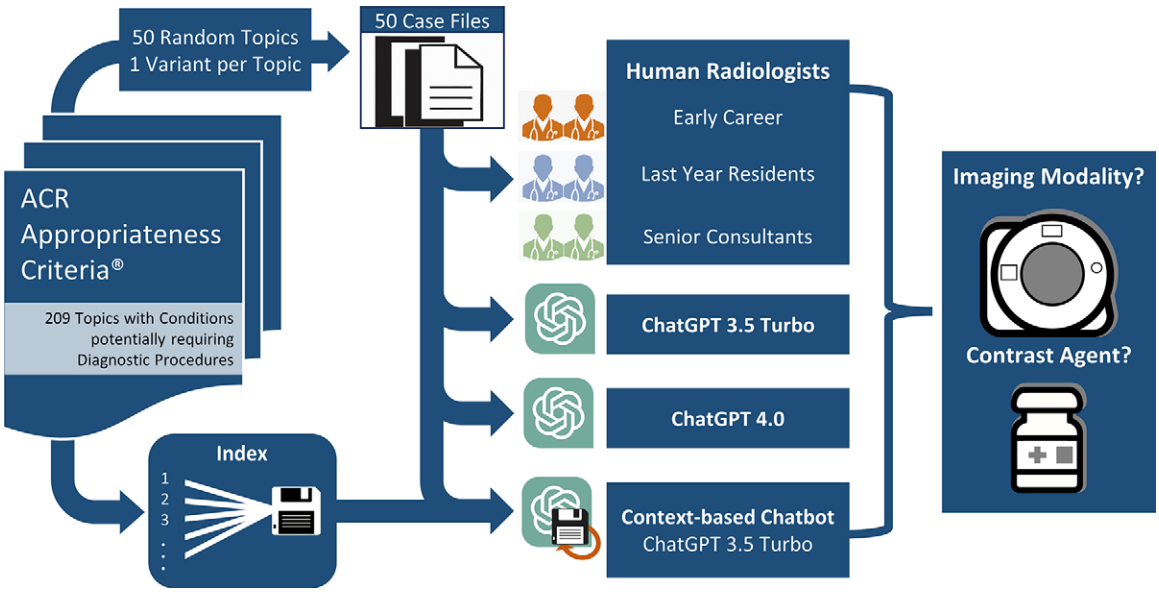

Figure 1: Schematic illustration of the creation and function of the American College of Radiology (ACR) appropriateness criteria context aware chatbot (accGPT).

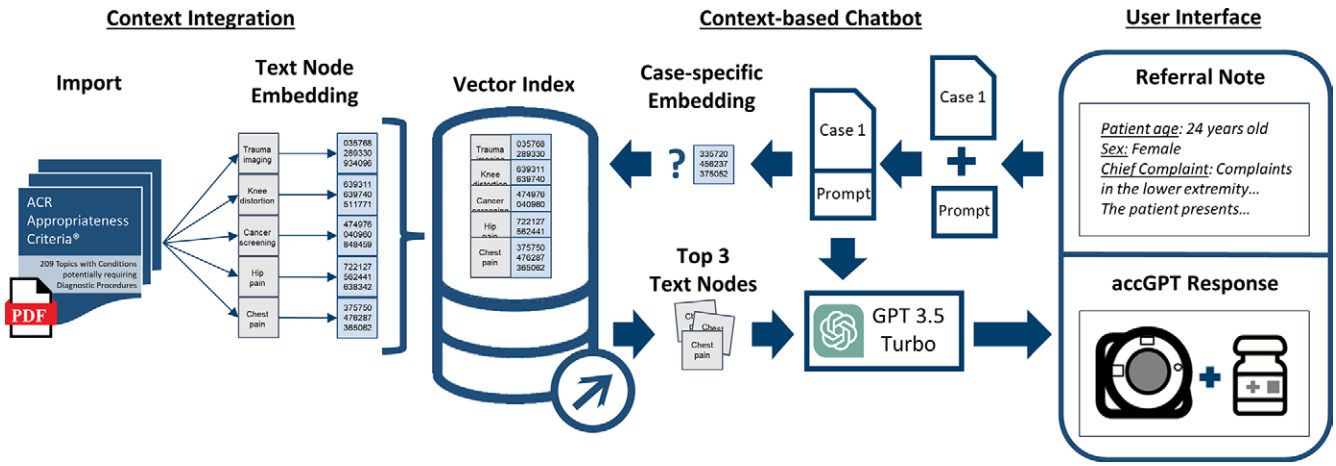

Figure 2: Schematic of the workflow of the case file creation, indexing of the American College of Radiology (ACR) appropriateness criteria context aware chatbot (accGPT), and performance analysis.

Materials and Methods

The workflow is depicted in Figure 1.

Technical Implementation of the Chatbots

Data preparation and indexing.—To develop and evaluate the proposed accGPT, we relied on the comprehensive collection of 209 topics with conditions potentially requiring diagnostic procedures from the ACR Appropriateness Criteria as a foundational knowledge base (a schematic is presented in Fig 2).

For semantic similarity processing, we used LlamaIndex (*https://github.com/jerryjliu/llama_index*; version 0.5.0) as an interface between external data from the ACR appropriateness guidelines and GPT-3.5-turbo. Text information was extracted from the ACR guidelines and embeddings were created using the GPTSimpleVectorIndex of LlamaIndex. For this, the document texts were divided into smaller chunks of up to 512 tokens (a measure of text content) and converted to text nodes. These nodes were then transformed into numerical vectors through a process known as embedding (text-embedding-ada-002-v2 model by OpenAI) and the result was stored in a dictionary-like structure, the so-called vector index.

Prompting strategy and answer synthesis.—To customize the output of all chatbots for the case-based scenarios, the question posed to the system in each case was as follows: "Is imaging typically appropriate for this case? If so, please specify the most suitable imaging modality and whether a contrast agent is required. Exclude 'May Be Appropriate' and 'Usually Not Appropriate' as potential responses."

For GPT-3.5-turbo and GPT-4, the direct output was captured as the response. For our context-based chatbot, the three best matching data nodes from the index were retrieved based on the embedding of the prompt. These nodes were used in a multistep answer creation and refinement method using GPT-3.5-turbo, and the final output was then captured.

Preparation of Case Files

We sought to test the chatbot's accuracy in comparison to human performance in a scenario resembling clinical routine (the workflow is depicted in Fig 2). For this, 50 clinical case files were created based on the ACR Appropriateness Criteria. From the 209 topics with clinical conditions potentially requiring diagnostic procedures, 50 were randomly chosen. Subsequently, in each of those, one variant was randomly selected. Based on those 50 clinical conditions, case files were created resembling clinical routine referral notes by an experienced radiologist not involved in the reading. Clinical files included information on age, sex, chief complaint, medical history, and results from clinical testing while excluding the suspected pathology in most cases. The case files encompassed a wide range of topics and medical conditions, some of which are rarely encountered in a radiologist's routine clinical practice (please see Appendix S1 for more details).

Assessment of Human and Chatbot Performance

The 50 case files were presented to six general radiologists at different experience levels who were familiar with the ACR guidelines: two early career radiologists (A.F. and S.R., in the 1st and 2nd year of training), two advanced residents in their last year of training (H.T., C.W.), and two board-certified radiologists (J. Nattenmüller and J. Neubauer, with 11 and 12 years of experience, respectively). For each case file, appropriateness of imaging and, if required, the most appropriate imaging modality and need for contrast agent administration was evaluated by all radiologists independently. During this assessment, no consultation of colleagues or guidelines was allowed.

We used a script-based approach on the 50 case files to perform six-fold repetition testing for all three chatbots and assess their performance. Again, appropriateness of imaging and, if required, the most appropriate imaging modality and need for contrast agent administration was evaluated.

Table 1: Correct Answers on the Case Files according to the ACR Appropriateness Criteria with regard to Radiologist Experience and Run Number

A: Experience Level of Radiologists

Appropriateness	Junior Radiologist 1	Junior Radiologist 2	Advanced Resident 1	Advanced Resident 2	Senior Radiologist 1	Senior Radiologist 2
"Usually appropriate"	33 (66) [53, 79]	32 (64) [51, 77]	30 (60) [46, 74]	33 (66) [53, 79]	37 (74) [62, 86]	33 (66) [53, 79]
"Usually appropriate" and "May be appropriate"	38 (76) [64, 88]	36 (72) [60, 84]	38 (76) [64, 88]	38 (76) [64, 88]	39 (78) [72, 89]	37 (74) [62, 86]

B: Run No. for Chatbots

Chatbot and Appropriateness	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6
GPT-3.5-turbo						
"Usually appropriate"	32 (64) [51, 77]	35 (70) [57, 83]	36 (72) [60, 84]	35 (70) [57, 83]	35 (70) [57, 83]	38 (76) [64, 88]
"Usually appropriate" and "May be appropriate"	36 (72) [60, 84]	41 (82) [71, 93]	40 (80) [69, 91]	39 (78) [72, 89]	39 (78) [72, 89]	40 (80) [69, 91]
GTP-4						
"Usually appropriate"	40 (80) [69, 91]	38 (76) [64, 88]	41 (8) [71, 93]	39 (78) [72, 89]	37 (74) [62, 86]	41 (82) [71, 93]
"Usually appropriate" and "May be appropriate"	45 (90) [82, 98]	43 (86) [76, 96]	44 (88) [79, 97]	43 (86) [76, 96]	42 (84) [74, 94]	43 (86) [76, 96]
accGPT						
"Usually appropriate"	41 (82) [71, 93]	41 (82) [71, 93]	41 (82) [71, 93]	41 (82) [71, 93]	43 (86) [76, 96]	42 (84) [74, 94]
"Usually appropriate" and "May be appropriate"	42 (84) [74, 94]	42 (84) [74, 94]	42 (84) [74, 94]	41 (82) [71, 93]	44 (88) [79, 97]	43 (86) [76, 96]

Note.—Data are numbers of correct answers (n = 50), with percentages in parentheses and 95% CIs in brackets. Junior radiologists 1 and 2 were in their 1st and 2nd year of training, respectively, the advanced residents were in their last year of training, and the two board-certified radiologists had 11 and 12 years of experience, respectively. accGBT = appropriateness criteria context aware chatbot, ACR = American College of Radiology.

Accuracy and Agreement of Radiologists and Chatbots

The respective human- and chatbot-derived recommendations for imaging in the 50 case files were evaluated regarding their appropriateness according to the ACR guidelines, that is, whether they met "usually appropriate" or "may be appropriate" criteria. To compare human raters, GPT-3.5-turbo, GPT-4, and accGPT, we fitted a generalized linear mixed model (binomial family and logit link) for the binary outcome "correct rating yes/no." Concerning the outcome, we fitted separate models for "may be appropriate" included in the category "yes" or "no," respectively. The rating method (human, GPT-3.5-turbo, GPT-4, accGPT) was included as a fixed factor while allowing for random intercepts for the case and the rater. We alternated the reference group of the method to obtain all pairwise comparisons. The estimation was performed using the glmer () function from the lm4 package (version 1.1–33) in GNU R software (version 4.2.1; R Foundation for Statistical Computing). The P values for the pairwise comparisons were calculated using asymptotic Wald tests.

To obtain an impression on the consistency of the rating via the GPT algorithms, we presented each case six times and calculated the proportion of cases with 100% (six of six) correct rating and with at least 66.66% (four of six) correct ratings.

Assessment of Cost-effectiveness of Radiologists and Chatbots

Furthermore, time to decision was assessed per reader to assess cost-effectiveness. Calculation of radiologists`s costs was based on the publicly available salary information for medical doctors working in university hospitals in Germany *(https://oeffentlicher-dienst.info/aerzte/uniklinik/)*.

The costs using the GPT models were monitored on the billing output of the OpenAI webpage *(https://platform.openai.com/account/usage)* and validated with the token usage during the first run and price list.

Code Availability

The source code for our chatbot implementation is publicly available on GitHub under the open-source MIT License *(https://github.com/maxrusse/accGPT)*. The use of the code for research and other projects must be in accordance with the terms of the license.

Results

Human and Chatbot Performance

Chatbots performed at least as well as humans, as shown in Table 1 and Figure 3. Case-based answer details are provided in Tables 1 and S1.

The performance of accGPT in providing correct recommendations for imaging in the 50 case files meeting the "usually appropriate" criteria was significantly better than that of radiologists and GPT-3.5-turbo (odds ratio [OR], 3.76 and 2.93, respectively; both $P < .001$). Compared with GPT-4, accGPT performed better only on a trend level (OR, 1.54; $P = .08$). More details are provided in Table 2.

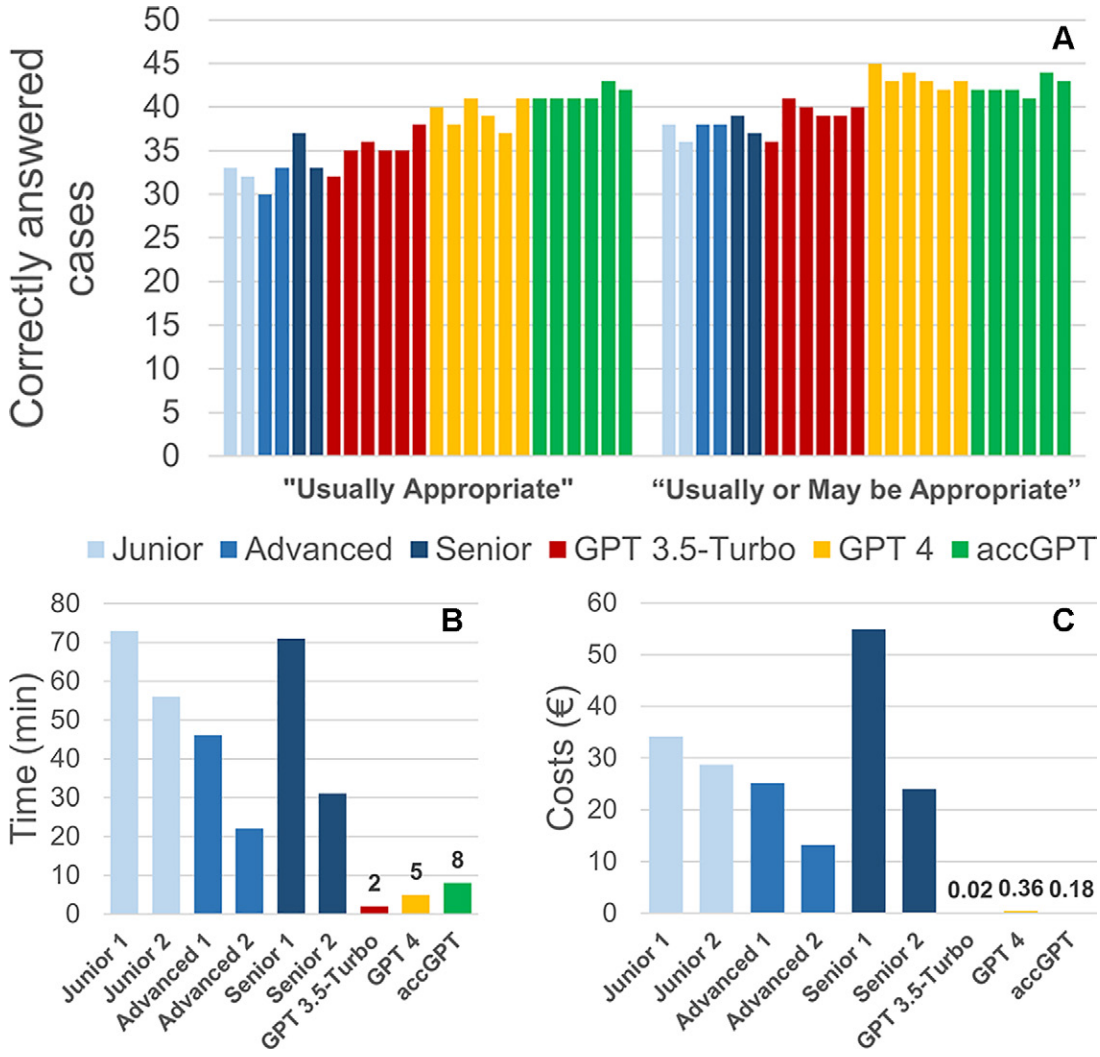

Figure 3: Bar graphs show comparison of radiologist and chatbot performance regarding **(A)** correctly provided recommendations in the 50 presented case files, **(B)** time required to process the 50 cases, and **(C)** costs incurred during the processing of the cases.

In the comparison of the recommendations meeting the "usually appropriate" and "may be appropriate" criteria, the performance of accGPT was again superior to that of radiologists and GPT-3.5-turbo (OR, 2.35 [P = .001] and 2.02 [P = .006], respectively), while the performance in comparison with GPT-4 was not different (OR, 0.78; P = .39). Please see Table 3 for details.

When we evaluated radiologists according to the different experience levels, we did not observe a robust difference in correct answers as the variance of the random intercept for rater was almost zero and excluding the intercept did not change the results for the fixed effect, implying that the experience level did not have an impact.

Consistency of Radiologists and Chatbots

As shown in Table 1, generic GPT 3.5-turbo and GPT-4 provided a higher count of correct "may be appropriate" recommendations compared with accGPT over the six repetitive runs of the respective chatbot. When considering "may be appropriate" as a false recommendation, accGPT was correct for all six runs in 74% (95% CI: 62, 86) of all cases and at least four times correct in 82% (95% CI: 71, 93) of all cases. GPT-3.5-turbo provided 42% (95% CI: 28, 56) and 68% (95% CI: 55, 81) correct recommendations, respectively, and GPT-4 64% (95% CI: 51, 77) and 78% (95% CI: 67, 90). Upon consideration of "may be appropriate" as correct, accGPT was correct for all six runs in 74% (95% CI: 62, 86) of all cases and at least four times correct in 84% (95% CI: 74, 94) of all cases. GPT-3.5-turbo provided 54% (95% CI: 40, 68) and 76% (95% CI: 64, 88) correct recommendations, respectively, and GPT-4 provided 76% (95% CI: 64, 88) and 86% (95% CI: 76, 96).

Analysis on Cost-Effectiveness of Radiologists and Chatbots

The radiologists spent varying amounts of time evaluating the case files, with a mean duration of 49 minutes 48 seconds (SD, 19 minutes; range, 22–73 minutes). This resulted in a mean cost of €29.99 ($33.24; SD, €12.78 [$14.16]; range, €13.2–€54.86 [$14.16–$60.80]). GPT 3.5-turbo required 2 minutes with €0.02 cost ($0.02), GPT-4 took 5 minutes and cost €0.36

Table 2: Comparison of the Performance of the Radiologists and Chatbots for "Usually Appropriate" Recommendations

	Reference			
Rater	Human	GPT-3.5-Turbo	GPT-4	accGPT
Human	…	0.78 (.23)	0.41 (<.001)	0.27 (<.001)
GPT-3.5-turbo	1.29 (.23)	…	0.53 (.004)	0.34 (<.001)
GPT-4	2.44 (.001)	1.90 (.004)	…	0.65 (.08)
accGPT	3.76 (<.001)	2.93 (<.001)	1.54 (.08)	…

Note.—Data are odds ratios, with *P* values in parentheses. accGBT = appropriateness criteria context aware chatbot.

Table 3: Comparison of the Performance of the Radiologists and Chatbots for "Usually Appropriate" and "May Be Appropriate" Recommendations

	Reference			
Rater	Human	GPT-3.5-Turbo	GPT-4	accGPT
Human	…	0.86 (.53)	0.33 (<.001)	0.43 (.001)
GPT-3.5-turbo	1.16 (.55)	…	0.39 (<.001)	0.49 (.007)
GPT4	3.00 (<.001)	2.59 (<.001)	…	1.28 (.39)
accGPT	2.35 (.001)	2.02 (.005)	0.78 (.39)	…

Note.—Data are odds ratios, with *P* values in parentheses. accGBT = appropriateness criteria context aware chatbot.

($0.40), and accGPT needed 8 minutes and the token costs amounted to €0.18 ($0.20). Overall, time and costs were significantly lower for the chatbots compared with the radiologists (both $P = .003$).

Discussion

Our results demonstrate the potential of the context-based appropriateness criteria context aware chatbot (accGPT) in making imaging recommendations based on the American College of Radiology (ACR) guidelines as it accepts standard clinical referral notes and provides concise recommendations on imaging in an end-to-end solution. When we compared the performance of accGPT with that of general radiologists with varying levels of experience and two publicly available generic chatbots (GPT-3.5-turbo and GPT-4.0), we noted a superior accuracy and higher consistency in meeting the "usually appropriate" ACR recommendations. In addition, all chatbots were substantially more time efficient and less expensive than human radiologists in the evaluation of the case files.

The topic of our study is of great relevance as the insufficient use of guidelines results in a large amount of inappropriate imaging (2,5). Inappropriate imaging is associated with increased costs for the health care systems, prolonged waiting lists, incorrect or delayed diagnoses, and potentially unnecessary exposure to ionizing radiation (6).

Notably, all evaluated chatbots achieved at least a human-level performance, with accGPT providing consistently the highest proportion of "usually appropriate" recommendations. It is worth emphasizing that accGPT provided "may be appropriate" recommendations less frequently than human raters and the other chatbots, which reflects the quality of the recommendations. The significant improvement in performance from GPT version 3.5 to 4.0 is corroborated by other studies focusing on the simplification of radiologic reports (17) and passing U.S. Medical Licensing Examination questions (14). Thus, it is reasonable to anticipate a substantial performance increase in a context-based chatbot using GPT-4. In addition, the early stages of development for a dedicated combination of commercially available vector stores and up-to-date OpenAI models are underway in a limited alpha phase *(https://openai.com/blog/chatgpt-plugins)*.

Nevertheless, ChatGPT itself was not explicitly designed for clinical decision-making. This might result in incorrect recommendations and induce the "out-of-vocabulary" problem, potentially yielding outputs that are not entirely accurate or consistent with reality. However, incorporating specialized context-based knowledge, such as the ACR guidelines, addresses this constraint. This adaptability also enables the chatbot's knowledge base to be updated or supplemented with additional guidelines or recommendations from other reputable sources. This also addresses concerns related to fake news and false data to ensure the reliability of the chatbot's output (19,21).

We observed challenges for both the radiologists and the chatbots, particularly in the cases in which no imaging was appropriate (eg, case files 6, 23, and 35). Rao et al (20) noted this limitation, too, as in their study ChatGPT recommended imaging in the only case where it was not necessary.

Because the ACR Appropriateness Criteria mainly comprise clinical conditions that necessitate some kind of imaging, further investigation including more case vignettes without the need for imaging (eg, dermatologic pathologies) is of interest.

Existing tools that are based solely on trusted sources, such as the European Society of Radiology iGuide, have the drawback that clinical information from case vignettes needs to be simplified for the input. This precludes assessing information from previous examinations and diagnostic workups (9). In contrast, the investigated chatbots allow for the direct input of the clinician's referral text. Furthermore, ChatGPT-based approaches allow for a deeper insight into the decision-finding process, as recently shown (14). This is also an advantage over less advanced natural language processing approaches that have already been developed as clinical decision support systems (22).

Another benefit of the proposed approach is the observed reduction in evaluation duration as well as the marked cost efficiency of the algorithms compared with radiologists. From a technical standpoint, the integration into clinical information systems or radiologic appointment scheduling systems is feasible. Considering the ever-increasing workload of

radiologists (23), integrated tools like accGPT might constitute a relief in clinical routine.

When introducing AI-based solutions like accGPT into health care, addressing ethical, legal, and data security considerations is crucial. In addition, the use of AI in clinical decision-making raises concerns about transparency, accountability, and potential biases (19,24). In the case of context-based chatbots, the use as a decision support tool, with an additional step of displaying the references, can enhance transparency and trust.

Further optimization of the proposed accGPT should aim to include detailed assessment of the costs, availability, and potential radiation dose. We see the potential for accGPT to be beneficial to both radiologists and referring physicians, albeit with different anticipated primary use cases. Radiologists would primarily use it as an information retrieval tool for rare cases, whereas ordering physicians might find it useful as a quick reference to guide decision-making in the ordering process. This is in alignment with current users of the manual search in the ACR guidelines. Because the performance is currently good but not excellent, a fully automated review of referral notes is not yet feasible. However, it is conceivable that these will be screened by accGPT and supportive recommendations will be made for referring physicians and radiologists.

The limited overall performance of the human raters is most likely due to the fact that the ACR guidelines were not available for consultation during the evaluation. In addition, the case files were assessed by general radiologists and many of the case files dealt with rare constellations. In these cases, better performance by subspecialized radiologists could be expected. However, including general radiologists allowed for covering the whole breadth of the ACR guidelines. This broad range may also explain the lack of superiority of experienced versus junior radiologists, as these rare cases would have required consultation of the ACR guidelines, which was not possible in this study. The analysis of the performance of human raters with access to the ACR guidelines during the evaluation is lacking. Here, a high performance can be assumed, but also a substantially higher expenditure of time. Further limitations include the fact that the data set is rather small and that we used synthetic case files. However, this allowed us to randomly select clinical conditions from the ACR guidelines and thus cover the full range of the ACR guidelines, including rare clinical conditions.

In summary, ChatGPT-based algorithms have the potential to substantially improve the decision-making for clinical imaging in accordance with American College of Radiology guidelines. Specifically, a context-based algorithm was superior to the generic chatbots, demonstrating the value of tailoring artificial intelligence solutions to specific health care applications and the potential to enhance generic chatbots to perform specific tasks.

Author contributions: Guarantors of integrity of entire study, **A.R., S.R., M.F.R.**; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **A.R., J. Nattenmüller, M.R., M.F.R.**; clinical studies, **A.R., S.R., J. Neubauer**; experimental studies, **A.R., A.F., H.T., C.W., J. Neubauer, F.B., M.F.R.**; statistical analysis, **A.R., D.Z., M.R., M.F.R.**; and manuscript editing, **A.R., S.R., D.Z., H.T., C.W., J. Nattenmüller, J. Neubauer, F.B., M.R., M.F.R.**

Disclosures of conflicts of interest: **A.R.** No relevant relationships. **S.R.** No relevant relationships. **D.Z.** Beirat IBS-DR; member ECB Subcommittee ISCB; deputy head Technical Committee Biometry GMDS. **A.F.** No relevant relationships. **H.T.** No relevant relationships. **C.W.** No relevant relationships. **J. Nattenmüller** No relevant relationships. **J. Neubauer** No relevant relationships. **F.B.** Research grants from Siemens Healthineers, Bayer Healthcare, and Bracco Healthcare; speakers bureau for Siemens Healthineers, Bayer Healthcare, and Bracco Healthcare; consultant for Bayer Healthcare; payment or honoraria for lectures, presentations, speakers bureaus, manuscript writing, or educational events from Siemens Healthineers and Bayer Healthcare; participation on a DataSafety Monitoring Board or Advisory Board, Siemens Healthineers; leadership or fiduciary role in Deutsche Röntgengesellschaft. **M.R.** No relevant relationships. **M.F.R.** No relevant relationships.

References

1. Cascade PN. Setting appropriateness guidelines for radiology. Radiology 1994;192(1):50A–54A.
2. Gabelloni M, Di Nasso M, Morganti R, et al. Application of the ESR iGuide clinical decision support system to the imaging pathway of patients with hepatocellular carcinoma and cholangiocarcinoma: preliminary findings. Radiol Med (Torino) 2020;125(6):531–537.
3. European Society of Radiology (ESR); American College of Radiology (ACR). European Society of Radiology (ESR) and American College of Radiology (ACR) report of the 2015 global summit on radiological quality and safety. Insights Imaging 2016;7(4):481–484.
4. Young GJ, Flaherty S, Zepeda ED, Mortele KJ, Griffith JL. Effects of Physician Experience, Specialty Training, and Self-referral on Inappropriate Diagnostic Imaging. J Gen Intern Med 2020;35(6):1661–1667. [Published correction appears in J Gen Intern Med 2020 May 6. 10.1007/s11606-020-05761-x.]
5. Francisco MZ, Altmayer S, Verma N, et al. Appropriateness of Computed Tomography and Ultrasound for Abdominal Complaints in the Emergency Department. Curr Probl Diagn Radiol 2021;50(6):799–802.
6. European Society of Radiology (ESR). Summary of the proceedings of the international forum 2016: "Imaging referral guidelines and clinical decision support - how can radiologists implement imaging referral guidelines in clinical routine?". Insights Imaging 2017;8(1):1–9.
7. Markus T, Saban M, Sosna J, et al. Does clinical decision support system promote expert consensus for appropriate imaging referrals? Chest-abdominal-pelvis CT as a case study. Insights Imaging 2023;14(1):45.
8. Saban M, Sosna J, Singer C, et al. Clinical decision support system recommendations: how often do radiologists and clinicians accept them? Eur Radiol 2022;32(6):4218–4224.
9. European Society of Radiology (ESR). Methodology for ESR iGuide content. Insights Imaging 2019;10(1):32.
10. Bookman K, West D, Ginde A, et al. Embedded Clinical Decision Support in Electronic Health Record Decreases Use of High-cost Imaging in the Emergency Department: EmbED study. Acad Emerg Med 2017;24(7):839–845.
11. GPT-3.5-turbo. OpenAI. https://platform.openai.com/docs/models/gpt-3.5. Accessed May 15, 2023.
12. GPT-4. OpenAI. https://openai.com/gpt-4. Accessed May 15, 2023.
13. 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 2023;2(2):e0000198.
14. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on Medical Challenge Problems. arXiv 2303.13375 [preprint] https://arxiv.org/abs/2303.13375. Posted March 20, 2023. Accessed May 15, 2023.
15. Buvat I, Weber W. Nuclear Medicine from a Novel Perspective: Buvat and Weber Talk with OpenAI's ChatGPT. J Nucl Med 2023;64(4):505–507.
16. Jeblick K, Schachtner B, Dexl J, et al. ChatGPT Makes Medicine Easy to Swallow: An Exploratory Case Study on Simplified Radiology Reports. arXiv 2212.14882 [preprint] https://arxiv.org/abs/2212.14882. Posted December 30, 2022. Accessed May 15, 2023.
17. Lyu Q, Tan J, Zapadka ME, et al. Translating Radiology Reports into Plain Language using ChatGPT and GPT-4 with Prompt Learning: Promising Results, Limitations, and Potential. arXiv 2303.09038 [preprint] https://arxiv.org/abs/2303.09038. Posted March 16, 2023. Accessed May 15, 2023.

18. Rao A, Pang M, Kim J, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow. medRxiv 2023.02.21.23285886. Posted February 26, 2023. Accessed May 15, 2023.
19. Shen Y, Heacock L, Elias J, et al. ChatGPT and Other Large Language Models Are Double-edged Swords. Radiology 2023;307(2):e230163.
20. Rao A, Kim J, Kamineni M, Pang M, Lie W, Succi MD. Evaluating ChatGPT as an Adjunct for Radiologic Decision-Making. medRxiv 2023.02.02.23285399. Posted February 7, 2023. Accessed May 15, 2023.
21. Else H. Abstracts written by ChatGPT fool scientists. Nature 2023;613(7944):423.
22. Chaudhari GR, Chillakuru YR, Chen TL, et al. Clinical language search algorithm from free-text: facilitating appropriate imaging. BMC Med Imaging 2022;22(1):18.
23. Kwee TC, Kwee RM. Workload of diagnostic radiologists in the foreseeable future based on recent scientific advances: growth expectations and role of artificial intelligence. Insights Imaging 2021;12(1):88.
24. Carter SM, Rogers W, Win KT, Frazer H, Richards B, Houssami N. The ethical, legal and social implications of using artificial intelligence systems in breast cancer care. Breast 2020;49:25–32.