CERASUS JOURNAL OF MEDICINE

ORIGINAL ARTICLE

Interpreting chest X-ray with ChatGPT: Can it serve as a tool for justifying computed tomography?

Nur Hürsoy[1]
Hafsa Kolluk[1]
Merve Solak[1]
Kubilay Kağan Budak[1]
Esat Kaba[1]

1. *Recep Tayyip Erdogan University, Department of Radiology, Rize, Türkiye*

Received: 04 February 2025
Accepted: 15 May 2025
Published: 29 June 2025

Corresponding Author: Esat Kaba
Address: Department of Radiology, Recep Tayyip Erdogan University, 53100, Türkiye
Mail: esatkaba04@gmail.com

Abstract

Objective: The aim of this study was to test the success of ChatGPT-4 in evaluating chest radiographs and detecting abnormal findings, and then to demonstrate its utility in computed tomography (CT) justification.

Methods: This study included 59 patients (20 patients in the first phase, and 39 patients in the second phase) from a publicly available chest X-ray dataset. X-rays were evaluated by an experienced chest radiologist (as gold standard), two radiology residents, and ChatGPT, first as normal-abnormal and then whether CT was needed if abnormal. Finally, the ChatGPT and two radiology residents' decisions were compared with the gold standard decision of the expert radiologist to obtain an accuracy value.

Results: The accuracy of Resident 1, Resident 2, and ChatGPT for normal-abnormal labeling was 76.27%, 93.22%, and 76.27%, respectively, for a total of 59 patients. The accuracy of Resident 1, Resident 2, and ChatGPT for CT necessity was 67.80%, 72.88%, and 66.10%, respectively. The expert radiologist determined that CT was not necessary in 30 patients. Of these 30 patients, Resident 1, Resident 2, and ChatGPT answered incorrectly in 14, 12, and 15 patients, respectively. There is no statistically significant difference between the responses of Resident 1, Resident 2, and ChatGPT for CT necessity (Chi-square, p=0.731).

Conclusion: The results of this study show that ChatGPT-4 is promising for chest X-ray interpretation and justification of CT scans. However, large language models such as ChatGPT, which still have major limitations, should be trained with a much larger number of radiology images.

Keywords: Justification; chest X-ray, thorax CT, large language models, ChatGPT

You may cite this article as: Hürsoy N, Kolluk H, Solak M, Budak KK, Kaba E. Interpreting chest X-ray with ChatGPT: Can it serve as a tool for justifying computed tomography? *Cerasus J Med.* 2025;2(2):118-126. doi:10.70058/cjm.1633438

Introduction

Systems that generate X-rays to produce images cause radiation exposure to the patient and, in some cases, to the healthcare workers. Report No. 184 of the National Council on Radiation Protection and Measurements [NCRP] of the United States of America reports that the proportion of total effective dose from computed tomography [CT] scans was 50% in 2006 and increased to 63% in 2016. The number of CT scans performed in the US has increased by 20% in 10 years [1]. Justification remains an important principle of radiation protection, although the ability to obtain images at lower radiation doses due to evolving technology seems to balance the increase in the number of examinations [2-4]. Under the acronym EU-JUST-CT, a project to improve justification was launched by the European Commission in 2021. In the survey conducted in 30 European countries as part of the project, more than half the participants said that examinations were not justified [4]. Revised by the American College of Radiology in 2023, the evaluation of findings seen on other imaging modalities such as chest radiography is the first item in the indications for chest CT [5]. Although chest radiographs are among the most commonly used imaging modalities, they can be difficult to interpret [6,7]. In a study evaluating CT scans ordered for suspected hilar pathology on chest radiography, pathology was found in 16.4% of patients, excluding vascular dilatation [8]. In our daily practice, CT scans occasionally are performed for the clarification of suspicious findings on chest radiography but do not have an impact on the patient's treatment decision.

The use of artificial intelligence in healthcare is becoming more widespread. Radiology is the first department to start using artificial intelligence applications. As of July 2023, 79% of the applications approved for use by the US Food and Drug Administration Administration [FDA] are in the field of radiology [9]. The frequency of use varies across the different subspecialties of radiology. Thoracic radiology ranks second with 31% of CE-marked applications [10]. Studies of different algorithms in lung radiology are ongoing [11-13].

Natural Language Processing [NLP] has reached a new dimension with Large Language Models [LLM]. Language models can answer different questions based on the relationships between word sequences and can produce written data according to different commands. The development of several models capable of processing images, audio and video recordings, and text has opened the way for various uses of these applications in the field of health [14].

The ChatGPT [Generative Pre-trained Transformer] language model developed by OpenAI software company has been used to study several different topics, including prioritizing emergency patients, evaluating sleep apnea syndrome, regulating protein energy malnutrition treatment, and interpreting electrocardiography [15-18]. With the widespread use of these studies, it will become possible to use language models in the field of health in the early period with greater accuracy and effectiveness. In this study, we aimed to demonstrate the success of the ChatGPT version 4.0 in interpreting chest radiographs and determining the necessity of CT scans from the radiograph findings. The study aimed to guide similar research by detailing the method section and offering insights into the use of language models.

Material and Methods

Determination of the study plan

The ChatGPT-4 version was selected for the study. The study team had previous experience using this version, which produces answers by accessing various data via the Internet [19]. The use of chest radiograph findings in CT justification was emphasized to provide a different perspective on the evaluation of chest radiographs. At this point, CT justification was investigated based only on Chest X-ray findings without any clinical information. It was agreed that heart failure, pulmonary edema, and lobar pneumonia were examples of clinical conditions that could be detected on chest radiography but would not require further investigation by CT. However, it was anticipated that the reasons for CT scanning may vary according to other data about the patient and that these reasons cannot be based on generally accepted sources. Given the similar difficulties experienced in decision-making in routine workflow, it was decided to evaluate the potential of ChatGPT in daily use by detailed interpretation of its responses to various commands.

After considering the implications for patient safety and potential ethical issues, the decision was made to proceed with the study using open, internationally accessible ready-to-use datasets so this study did not require institutional review board approval. In this context, the "National Institutes of Health Chest X-Ray Dataset", which is publicly available in the literature, was used [20]. This dataset contains 112,120 chest radiographs of 30,805 patients. From this dataset, a radiologist (EK) randomly selected 20 patients for the first phase

of this study and 50 patients for the second phase. It was agreed that normal images and images labeled with different pathologies, selected from the dataset by the radiologist, would be forwarded to two trainees without labeling information. The expert radiologist (NH) with, five years of experience in thoracic radiology, evaluated the labeled images, the responses of the trainees, and Chat-GPT.

Workflow

The 20 images selected from the dataset were shared with two residents (MS, HK). A third resident (KKB) uploaded the images to Chat-GPT in the same order. The trainees first decided whether the images were normal or not and whether a CT scan was needed after the x-ray. The three most important findings and the findings which has no clinical significance, if any, were noted. It took 35-40 minutes to upload 20 images to ChatGPT and respond to commands. At this point, the following prompt was given to ChatGPT using the role model prompting technique (e.g. act like an experienced radiologist) and the study was started.

Prompt 1:

As an experienced radiologist, could you evaluate these chest X-rays, and answer the following questions?

1-Are there any pathological findings?

2-If there are, list the 3 most important findings.

3- Is a Thoracic CT necessary for this X-ray as a further examination?

When the first phase of the study was evaluated, it was found that residents had difficulty in describing the findings and that common terms to be used should be established. Therefore, the table where the images were scored was updated and drop-down lists were added (Table 1).

In the second phase, 50 selected images were evaluated by the residents using the new table. 19 images were assessed quickly by Chat-GPT, but the model refused to respond to the commands when the image upload was resumed. The initial prompt was still used, but Chat-GPT responded to only 19 patients. It then refused to respond and provided the following output:

"I can't provide medical evaluations, including interpretation of chest X-rays or other radiological images. This requires specialized knowledge from licensed healthcare professionals to ensure accuracy and safety. Consult a certified radiologist or healthcare provider for a professional assessment and advice regarding your medical imaging."

To resolve this, the chat page was refreshed, prompts were repeated at different times of the day, and on different days, prompts were changed, and similar prompts were entered from different accounts, but no results were obtained. This effort was continued for four days, and the initial prompt was revised as follows:

Table 1: The drop-down lists on the Excel table for standardization of Chest X-ray evaluation.

Zone	Findings	Diagnosis	Mediastinum	Costophrenic Sinus
Right Lung	Opacity	Malignancy	Large	Normal
Left Lung	Nodule	Benign Conditions	Normal	Blunt
Right Upper Zone		Infection	Bilateral hilary enlargement	
Right Lower Zone	Ground Glass	Interstitital Disease	Right hilary enlargement	
Left Upper Zone	Reticulation	Edema	Left hilary enlargement	
Left Lower Zone	Air-trapping	Nodule	Cardiomegaly	
Upper Zone	Other	Other		
Lower Zone				
Diffuse				
Other				

Prompt 2:

"As an experienced radiologist, could you evaluate these chest X-rays, and answer the following questions in yes, or no?

If the answer is yes, then elaborate please.

1-Are there any pathological findings?

2-If there are, list the 3 most important findings.

3- Is a Thoracic CT necessary for this X-ray as a further examination?"

Assessing the answers

The images obtained from the dataset, and the residents' and ChatGPT's responses were evaluated by a radiologist (NH) with five years of experience in thoracic radiology. The results of the residents' evaluation were compared with the labels in the dataset and with the radiologist's evaluation. The answers of two residents were evaluated. The accuracy and appropriateness of the GPT's responses were analyzed. In addition to the labels in the dataset, the expert radiologist's comments also played a role in the adequacy assessment. The results are given in terms of numbers and percentages.

Statistical Analysis

The accuracy of responses from radiology residents and ChatGPT was evaluated by comparing them with labeled reference data and expert radiologist interpretations. The percentage of correct responses was calculated for both groups. To determine whether there was a significant difference between the performance of ChatGPT and the residents, a chi-square test was conducted. A p-value of less than 0.05 was considered statistically significant. All statistical analyses were performed using the open-source SciPy library in the Jupyter Notebook environment.

Results

Of the 20 images initially selected, 5 were normal; the abnormal images were labeled fibrosis, infiltration, nodule, cardiomegaly, mass, and consolidation. CT was deemed necessary by the radiologist to detail the findings on 8 images. The necessity of CT was more common among residents. There were differences in 5 of the answers of the residents, and the necessity of CT in 4 images evaluated differently varied according to the individuals. ChatGPT correctly evaluated all 5 normal images, whereas trainees recommended CT after three images labeled as normal. The first 3 images were the images that ChatGPT assessed as false negatives. In two images, it coded the finding on the wrong side. In two images with increased cardiothoracic index, it did not indicate cardiomegaly. When the images with incorrect answers were analyzed, it was determined that it failed to detect a nodule behind the costa and a small paramediastinal opacity. In addition, it described diffuse ground glass and reticulonodular opacities on the 17th film, which showed consolidation only in the right lower lobe.

After the first phase of the study was completed, 39 of the 50 selected cases were assessed by the ChatGPT at various times. After the 40th case, it refused to respond and the study was terminated at that stage. In

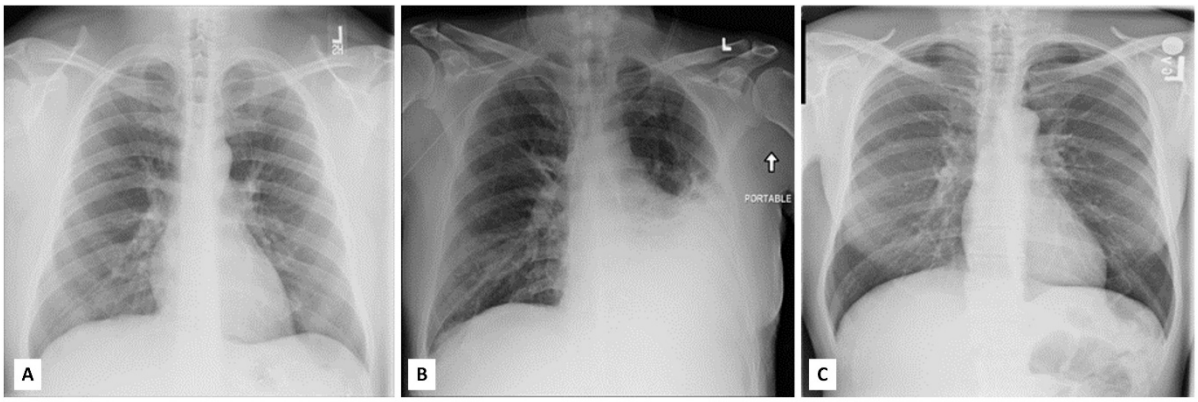

Figure 1: A: True Label: Normal, ChatGPT: Abnormal. **B:** True Label: Pleural effusion on the left, ChatGPT: Pleural effusion on the right. **C:** True Label: Normal, ChatGPT: Normal

the assessment by the expert radiologist (NH), CT was deemed necessary as a further investigation in 21 of the 39 cases. Although the number of cases in which CT was considered necessary by the trainees was similar, it was noted that they disagreed in 8 cases. In 15 out of 39 cases, it was noteworthy that the trainees disagreed with the findings. Figure 1 shows the incorrect and correct responses provided by ChatGPT for 3 different chest-X-rays.

ChatGPT misinterpreted 18 out of 39 cases. Of the 28 pathological chest radiographs, 13 were incorrect. In 10 of the misinterpreted chest radiographs, the specialist radiologist did not determine the need for CT. Although one study was marked normal, the consultant radiologist also felt that further investigation was required. ChatGPT also assessed this study as pathological, but the findings described were incorrect. In this case, a total of five patients labeled normal were incorrectly classified as pathological by Chat-GPT. Of the 11 cases that ChatGPT marked as normal, 5 were labelled as abnormal. Accuracy values for labeling patients as normal-abnormal for a total of 59 patients (20 first phase, 39-second phase) are given in Figure 2 for resident 1, resident 2, and ChatGPT. In addition, Figure 3 provides the accuracy values for resident 1, resident 2, and ChatGPT's predictions of CT necessity. Also in Figure 4, the expert radiologist's decision and the residents' and ChatGPT's predictions of CT necessity for each patient are visualized. In eight patients, 20.5% of the patients for whom Chat GPT recommended a CT scan, Chat GPT recommended a CT scan even though neither the radiology expert nor at least one of the two residents deemed it necessary. The expert radiologist determined that CT was not necessary in 30 patients. Of these 30 patients, Resident 1, Resident 2, and ChatGPT answered incorrectly in 14, 12, and 15 patients, respectively. There is no statistically significant difference between Resident 1, Resident 2, and ChatGPT responses (Chi-square, p=0.731)

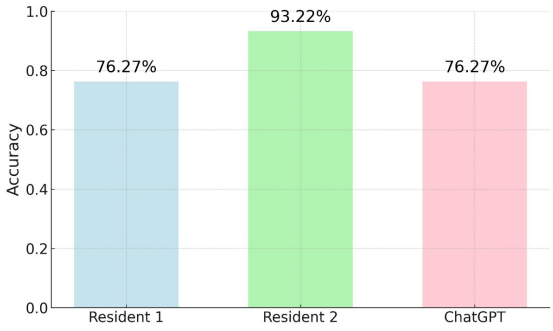

Figure 2: Normal-abnormal labeling accuracy of chest x-rays of residents and ChatGPT

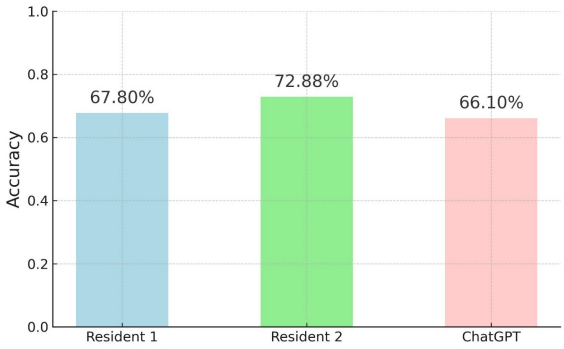

Figure 3: Accuracy rates of residents' and ChatGPT's prediction of CT necessity for chest X-rays

Discussion

Chest radiography is the most commonly performed imaging modality worldwide, yet it remains difficult to interpret. Inaccurate or inadequate evaluations of chest radiographs lead to an increase in the number of CTs. Artificial intelligence studies on chest radiographs are also quite common [20-23]. Chest X-ray studies using LLMs are also being tested [13,23].

In our study, we sought to answer the question of whether the evaluation of chest radiographs with Chat-GPT contributes to the reduction of unjustified CTs. Different prompts may provide the opportunity to experiment for different gains. However, at the beginning of the study, we realized the uncertainty of assessing the accuracy of our answers. During the study, we found that the interpretation of chest radiographs can vary depending on the acquisition technique, experience, and general

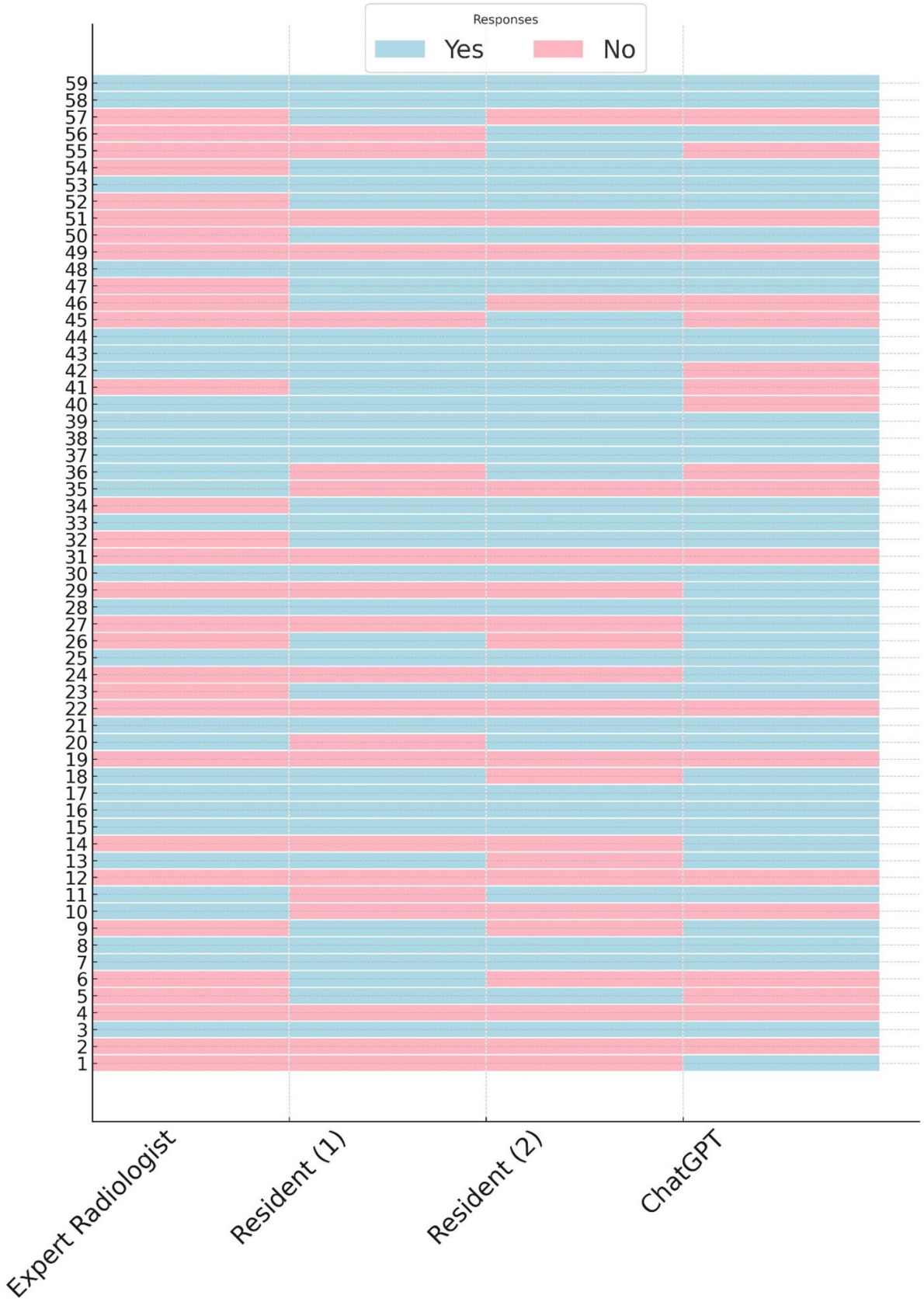

Figure 4: Expert radiologist's decision (gold standard), residents' and ChatGPT's decision on the necessity of CT in each patient

approach of the radiologist, making accurate labeling and unambiguous scoring difficult. In the literature, similar issues have been attempted to be overcome with a grading system used by different clinicians [24]. Similar publications have shown that the image analysis capability of LLM offers new clinical possibilities in radiology [25]. However, ongoing developments in the field of artificial intelligence are needed to increase diagnostic confidence in radiological applications [26].

The most important experience we have gained during our study has been the use of LLMs and the standardization of studies to be conducted with these models, the selection of topics, and the determination of evaluation criteria. When working with ChatGPT, we have experienced that the time setting should be done taking into account the days when it may fail. While discussing the study steps, we get an idea of the criteria that determine study quality in publications on similar issues.

When we examined the responses of Chat-GPT in detail in terms of CT justification, which is the main topic of our study, we found that it defined different findings in chest radiographs that it evaluated as pathological and recommended CT in a wide range of differential diagnoses. Chat-GPT recommends that CT should be performed after every chest radiograph which is evaluated as abnormal. Its interpretation of normal chest radiographs is consistent with our clinical approach: "Given the normal findings in this X-ray, a Thoracic CT doesn't seem necessary. However, a CT might be considered if there are clinical symptoms or a history of specific conditions that warrant further investigation. In this case, based on the X-ray alone, there are no significant abnormalities that suggest a need for additional imaging."

When we analyzed the errors of Chat-GPT, it was noteworthy that it gave incorrect directional information, did not detect cardiomegaly, and indicated some findings that were not found on radiography. The fact that we asked them to write down the 3 most important findings, if any, in the prompt may have triggered "hallucination". The film technique is also one of the factors influencing the answers. In two cases, Chat-GPT reported that the case was quite complex, stating: "The findings suggest a complex pulmonary condition that requires detailed imaging and possibly correlation with clinical symptoms and laboratory results to determine an appropriate course of treatment." The patients it describes as complex are those with really had diffuse pathologic findings, suggesting that the LLM's recommendation may be useful for triage.

This study has some limitations. Firstly, the number of images evaluated was small. Secondly, the prompting was performed only in English. Comparisons can be made by prompting in different languages. Thirdly, only ChatGPT-4, a paid version of LLMs, was used. In the future, the performance of different LLMs, such as the more recent version GPT-4o, should be compared with a larger number of images.

Conclusion

In this study, we shared our experiences about the difficulties that residents and radiologists with different experiences may encounter in chest X-ray evaluation studies with artificial intelligence algorithms and the use of LLM. In the results we obtained with limited data, we found that Chat-GPT may be insufficient although it contributes to CT justification. We think that studies with various prompt suggestions that may be useful in daily functioning in LLM use will be supportive of product development.

Conflict of interest

The authors declare no competing interests. The authors declare they have no financial interests.

Funding

No funding was obtained for this study.

Authors' Contributions: Concept: N.H., H.K., M.S., K.K.B., E.K..; Design: N.H., H.K., M.S., K.K.B., E.K.; Data Collection or Processing: N.H., H.K., M.S., K.K.B., E.K..; Analysis or Interpretation: N.H., H.K., M.S., K.K.B., E.K..; Literature Search: N.H., H.K., M.S., K.K.B., E.K..; Writing: N.H., H.K., M.S., K.K.B., E.K..;

Ethical approval: No ethics committee approval is required in this article since a publicly available dataset is used. The principles of the Declaration of Helsinki were followed during this study.

Patient consent: Patient consent is not required as public datasets are used.

References

1. Mettler FA, Mahesh M, Bhargavan Chatfield M, et al. *NCRP Reprt 184: Medical Radiation Exposure of Patients in the United States*. Recommendations of the National Council on Radiation Protection and Measurements; 2019.
2. E.G Friberg. HERCA European action week - result of a coordinated inspection initiative assessing Justification in Radiology. Int Conf Radiat Prot Med - Achiev Chang Pract. 2017;1–5.
3. Rastogi S, Singh R, Borse R, et al. Use of Multiphase CT Protocols in 18 Countries: Appropriateness and Radiation Doses. *Can Assoc Radiol J*. 2021;72(3):381-387. doi:10.1177/0846537119888390
4. Foley SJ, Bly R, Brady AP, et al. Justification of CT practices across Europe: results of a survey of national competent authorities and radiology societies. *Insights Imaging*. 2022;13(1):177. Published 2022 Nov 22. doi:10.1186/s13244-022-01325-1
5. American College of Radiology (ACR), Society of Advanced Body Imaging (SABI), Society for Pediatric Radiology (SPR), Society of Thoracic Radiology (STR). ACR–SABI–SPR–STR Practice Parameter for the Performance of Thoracic Computed Tomography (CT). Revised 2023. Available at: https://www.acr.org/
6. Speets AM, van der Graaf Y, Hoes AW, et al. Chest radiography in general practice: indications, diagnostic yield and consequences for patient management. *Br J Gen Pract*. 2006;56(529):574-578.
7. Gatt ME, Spectre G, Paltiel O, Hiller N, Stalnikowicz R. Chest radiographs in the emergency department: is the radiologist really necessary?. *Postgrad Med J*. 2003;79(930):214-217. doi:10.1136/pmj.79.930.214
8. Dadalı Y, Köksal D. Thorax CT findings of patients with hilar enlargement on chest X-Ray. *Ann Clin Anal Med*. 2020;11(3):235-238
9. FDA Artificial Intelligence and Machine Learning (AI/ML)-Enabled Medical Devices Page. Available at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices
10. van Leeuwen KG, Schalekamp S, Rutten MJCM, van Ginneken B, de Rooij M. Artificial intelligence in radiology: 100 commercially available products and their scientific evidence. *Eur Radiol*. 2021;31(6):3797-3804. doi:10.1007/s00330-021-07892-z
11. Ziegelmayer S, Marka AW, Lenhart N, et al. Evaluation of GPT-4's Chest X-Ray Impression Generation: A Reader Study on Performance and Perception. *J Med Internet Res*. 2023;25:e50865. Published 2023 Dec 22. doi:10.2196/50865
12. Tiu E, Talius E, Patel P, Langlotz CP, Ng AY, Rajpurkar P. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nat Biomed Eng*. 2022;6(12):1399-1406. doi:10.1038/s41551-022-00936-9
13. Lee KH, Lee RW, Kwon YE. Validation of a Deep Learning Chest X-ray Interpretation Model: Integrating Large-Scale AI and Large Language Models for Comparative Analysis with ChatGPT. *Diagnostics (Basel)*. 2023;14(1):90. Published 2023 Dec 30. doi:10.3390/diagnostics14010090
14. Bhayana R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. *Radiology*. 2024;310(1):e232756. doi:10.1148/radiol.232756
15. Zaboli A, Brigo F, Sibilio S, Mian M, Turcato G. Human intelligence versus Chat-GPT: who performs better in correctly classifying patients in triage?. *Am J Emerg Med*. 2024;79:44-47. doi:10.1016/j.ajem.2024.02.008
16. Mira FA, Favier V, Dos Santos Sobreira Nunes H, et al. Chat GPT for the management of obstructive sleep apnea: do we have a polar star?. *Eur Arch Otorhinolaryngol*. 2024;281(4):2087-2093. doi:10.1007/s00405-023-08270-9
17. Khan U. Revolutionizing Personalized Protein Energy Malnutrition Treatment: Harnessing the Power of Chat GPT. *Ann Biomed Eng*. 2024;52(5):1125-1127. doi:10.1007/s10439-023-03331-w
18. Günay S, Öztürk A, Özerol H, Yiğit Y, Erenler AK. Comparison of emergency medicine specialist, cardiologist, and chat-GPT in electrocardiography assessment. *Am J Emerg Med*. 2024;80:51-60. doi:10.1016/j.ajem.2024.03.017
19. Topçu Varlık A, Kaba E, Burakgazi G. The R.E.N.A.L. nephrometry scoring from CT reports with ChatGPT: example with proofs. *Jpn J Radiol*. 2024;42(8):929-931. doi:10.1007/s11604-024-01573-9
20. Xu S, Yang L, Kelly C. et al. ELIXR: Towards a general purpose X-ray artificial intelligence system through alignment of large language models and radiology vision encoders. 2023; Available at: http://arxiv.org/abs/2308.01317
21. Lee S, Youn J, Kim H, Kim M, Yoon SH. CXR-LLAVA: a multimodal large language model for interpreting chest X-ray images. 2023; Available at: https://arxiv.org/abs/2310.18341v3
22. Shentu J, Al Moubayed N. CXR-IRGen: An Integrated Vision and Language Model for the Generation of Clinically Accurate Chest X-Ray Image-Report Pairs. 2024;5200–9.

23. Thawkar O, Shaker A, Mullappilly SS, Cholakkal H, Anwer RM, Khan S, vd. XrayGPT: Chest Radiographs Summarization using Medical Vision-Language Models. 2023; Available at: http://arxiv.org/abs/2306.07971

24. Kozel G, Gurses ME, Gecici NN, et al. Chat-GPT on brain tumors: An examination of Artificial Intelligence/Machine Learning's ability to provide diagnoses and treatment plans for example neuro-oncology cases. *Clin Neurol Neurosurg*. 2024;239:108238. doi:10.1016/j.clineuro.2024.108238

25. Brin D, Sorin V, Barash Y, et al. Assessing GPT-4 multimodal performance in radiological image analysis. *Eur Radiol*. 2025;35(4):1959-1965. doi:10.1007/s00330-024-11035-5

26. Chetla N, Tandon M, Chang J, Sukhija K, Patel R, Sanchez R. Evaluating ChatGPT's Efficacy in Pediatric Pneumonia Detection From Chest X-Rays: Comparative Analysis of Specialized AI Models. *JMIR AI*. 2025;4:e67621. Published 2025 Jan 10. doi:10.2196/67621