Article published online: 2024-07-04

Original Article

Radiologic Decision-Making for Imaging in Pulmonary Embolism: Accuracy and Reliability of Large Language Models—Bing, Claude, ChatGPT, and Perplexity

Pradosh Kumar Sarangi[1] Suvrankar Datta[2] M. Sarthak Swarup[3] Swaha Panda[4] Debasish Swapnesh Kumar Nayak[5] Archana Malik[6] Ananda Datta[6] Himel Mondal[7]

[1] Department of Radiodiagnosis, All India Institute of Medical Sciences Deoghar, Deoghar, Jharkhand, India
[2] Department of Radiodiagnosis, All India Institute of Medical Sciences New Delhi, New Delhi, India
[3] Department of Radiodiagnosis, Vardhman Mahavir Medical College and Safdarjung Hospital New Delhi, New Delhi, India
[4] Department of Otorhinolaryngology and Head and Neck Surgery, All India Institute of Medical Sciences Deoghar, Deoghar, Jharkhand, India
[5] Department of Computer Science and Engineering, SOET, Centurion University of Technology and Management, Bhubaneswar, Odisha, India
[6] Department of Pulmonary Medicine, All India Institute of Medical Sciences Deoghar, Deoghar, Jharkhand, India
[7] Department of Physiology, All India Institute of Medical Sciences Deoghar, Deoghar, Jharkhand, India

Address for correspondence Pradosh Kumar Sarangi, MD, PDF, EDiR, Department of Radiodiagnosis, All India Institute of Medical Sciences, Deoghar 814152, Jharkhand, India (e-mail: drpksarangi@gmail.com).

Indian J Radiol Imaging

Abstract

Background Artificial intelligence chatbots have demonstrated potential to enhance clinical decision-making and streamline health care workflows, potentially alleviating administrative burdens. However, the contribution of AI chatbots to radiologic decision-making for clinical scenarios remains insufficiently explored. This study evaluates the accuracy and reliability of four prominent Large Language Models (LLMs)—Microsoft Bing, Claude, ChatGPT 3.5, and Perplexity—in offering clinical decision support for initial imaging for suspected pulmonary embolism (PE).

Methods Open-ended (OE) and select-all-that-apply (SATA) questions were crafted, covering four variants of case scenarios of PE in-line with the American College of Radiology Appropriateness Criteria®. These questions were presented to the LLMs by three radiologists from diverse geographical regions and setups. The responses were evaluated based on established scoring criteria, with a maximum achievable score of 2 points for OE responses and 1 point for each correct answer in SATA questions. To enable comparative analysis, scores were normalized (score divided by the maximum achievable score).

Result In OE questions, Perplexity achieved the highest accuracy (0.83), while Claude had the lowest (0.58), with Bing and ChatGPT each scoring 0.75. For SATA questions, Bing led with an accuracy of 0.96, Perplexity was the lowest at 0.56, and both Claude and ChatGPT scored 0.6. Overall, OE questions saw higher scores (0.73) compared to

Keywords
- large language model
- American College of Radiology Appropriateness Criteria
- pulmonary embolism
- Bing
- ChatGPT
- Claude
- Perplexity

DOI https://doi.org/10.1055/s-0044-1787974.
ISSN 0971-3026.

Thieme Medical and Scientific Publishers Pvt. Ltd., A-12, 2nd Floor, Sector 2, Noida-201301 UP, India

SATA (0.68). There is poor agreement among radiologists' scores for OE (Intraclass Correlation Coefficient [ICC] = −0.067, $p = 0.54$), while there is strong agreement for SATA (ICC = 0.875, $p < 0.001$).

Conclusion The study revealed variations in accuracy across LLMs for both OE and SATA questions. Perplexity showed superior performance in OE questions, while Bing excelled in SATA questions. OE queries yielded better overall results. The current inconsistencies in LLM accuracy highlight the importance of further refinement before these tools can be reliably integrated into clinical practice, with a need for additional LLM fine-tuning and judicious selection by radiologists to achieve consistent and reliable support for decision-making.

Introduction

In the rapidly evolving landscape of health care, artificial intelligence (AI) has emerged as a transformative force, particularly in the field of radiology.[1] The integration of large language models (LLMs) into radiologic decision-making processes has the potential to enhance accuracy and efficiency.

There has been extensive research exploring the capabilities of ChatGPT in the broader field of medicine, specifically in radiology. In radiology, ChatGPT and other LLMs have demonstrated promising and innovative applications. These applications encompass supporting medical writing and research,[2,3] structuring and organizing radiology reports,[4–7] protocoling radiology exams,[8] offering recommendations for screening,[9,10] addressing patient questions,[7,10,11] simulating text-based radiology board-style examinations,[12,13] providing differential diagnoses based on imaging patterns,[14,15] impressions,[16] and suggesting follow-up imaging by established guidelines,[17–20] among other functionalities.

LLMs like Microsoft Bing, Claude, ChatGPT, and Perplexity, are trained with large volumes of data and textual information. Their proficiency extends beyond mere report generation. LLMs can interpret textual reports and coherently present them. This capability not only aids radiologists in understanding and synthesizing vast amounts of data efficiently but also contributes to the overall accuracy of diagnostic assessments. Furthermore, LLMs can serve as valuable decision-support tools by suggesting additional investigations or follow-up procedures based on their comprehensive understanding of clinical information.[17–20] By facilitating precise and context-aware interpretations, LLMs emerge and have the potential to enhance diagnostic accuracy in the field of radiology.

The American College of Radiology (ACR) guidelines are a set of evidence-based recommendations and standards developed by the American College of Radiology. These guidelines serve as a comprehensive framework for radiologists to make informed decisions related to diagnostic imaging and medical procedures. It recommends the appropriate use of different imaging techniques based on clinical scenarios, patient characteristics, and evidence from medical literature and helps minimize unnecessary imaging procedures, reduce radiation exposure, and improve diagnostic accuracy. Nevertheless, variations persist in clinical practices, particularly in determining the necessity of imaging, the choice of modality, and the use of contrast material, leading to a significant number of inappropriate imaging procedures.[17]

Several clinical decision support (CDS) tools, such as iGuide by the European Society of Radiology (https://www.myesr.org/esriguide) and CareSelect Imaging by Change Healthcare (https://www.changehealthcare.com/clinical-decision-support/careselect/imaging), have been introduced to enhance adherence to published guidelines and have proven effective in reducing inappropriate examinations.[21–23] However, these tools often involve significant human interaction and may lose relevant clinical information due to limitations in handling free-text inputs.[23] LLMs present a promising solution by allowing input of free text and engaging in unrestricted interactions, potentially addressing the limitations of other CDS tools.

Since 1989, The Royal College of Radiologists has provided iRefer recommendations to assist in appropriate referrals to radiology departments. The eighth and current version, released in 2017, offers evidence-based guidance for referring physicians to suitable imaging tests or investigations (https://www.irefer.org.uk/why-irefer/about-irefer). iRefer is recognized as a crucial tool for advancing evidence-based imaging; however, it is not freely available, with subscription costs ranging from £120 to £4,200 per year.

Pulmonary embolism (PE) poses a significant challenge for radiologists due to its elusive and diverse clinical presentation. The symptoms of PE can be nonspecific and overlap with various other medical conditions, leading to diagnostic uncertainty. Moreover, the range of imaging modalities available for diagnosing pulmonary embolism, such as computed tomography pulmonary angiography, ventilation–perfusion scans, and chest radiographs, requires careful consideration of each patient's clinical context to determine the most appropriate and effective approach. Hence, ACR has defined its criteria to follow for better imaging decisions.

This study aimed to evaluate and compare the accuracy and reliability of four LLMs—Microsoft Bing, Claude, ChatGPT, and Perplexity, in the context of radiological decision-making for PE. By assessing the performance of these language models, the study aimed to provide insights into

their effectiveness in determining the suitability of initial imaging procedures based on preliminary clinical presentations, adhering to established standards such as those outlined by the American College of Radiology.

Methods

Type and Setting

This was a cross-sectional, observational study where we tested the accuracy of four LLMs about their accuracy and reliability in radiologic decision-making for PE according to ACR criteria. The study was conducted from September 2023 to November 2023.

Large Language Models

Four LLMs, namely Microsoft Bing (creative; https://www.bing.com/), Claude (https://claude.ai/), ChatGPT-3.5 (https://chat.openai.com), and Perplexity (https://www.perplexity.ai) were chosen based on their relevance, popularity, and contributions to medical science. All of the LLMs are freely accessible at the time of this study as a chatbot on the websites.

American College of Radiology Appropriateness Criteria

The ACR Appropriateness Criteria® (ACR AC) are evidence-based criteria that help referring doctors and other clinicians make the best imaging or therapy decisions for a given clinical condition. Using these principles helps practitioners improve the quality of treatment and contributes to the most effective use of radiology. For PE, there are four variants as shown in ►**Table 1**.

Questions

We designed two sets of questions. The first question was open-ended (OE) and the second question was to select all that apply (SATA). In the OE question, the case scenario is presented with suggestions for any imaging modality. In SATA, the case scenario is presented with closed-ended options and a question is asked about selecting all the suitable options.

Prompts

The prompts are designed for OE and SATA and saved for getting responses from each LLM separately. For OE, we used the prompt—"determine the single most appropriate initial imaging procedure according to ACR Appropriateness Criteria." For example, this is a full prompt of an OE—"Suspected pulmonary embolism. Low or intermediate pretest probability with a negative D-dimer. Determine the single most appropriate initial imaging procedure according to American College of Radiology (ACR) Appropriateness Criteria."

For SATA, the LLMs are provided with options to choose from. It can also state that none is suitable. For example, this is a full prompt of a SATA—"Suspected pulmonary embolism. Low or intermediate pretest probability with a negative D-dimer. Assess the appropriateness of the following initial imaging procedures procedure according to American College of Radiology (ACR) Appropriateness criteria in a concise manner: [Options]."

To avoid the influence of prior answers on model output, a new chat session was started for each prompt. In case of confusion in rating the responses by the user, it was sorted out in a virtual meeting as the LLMs have no standard responses to the same prompt and it needs human interpretation. For example, responses like "limited role," "rarely used," and "least appropriate" come across as LLM responses which were considered "usually not appropriate" as per ACR AC.

Observers

Three radiologists were given identical prompts from three different locations on the same day. Each radiologist was instructed to enter the same set of prompts without any textual modifications. The radiologists were designated as RAD1, RAD2, and RAD3. Recruiting three radiologists to ask the same questions was used to observe potential variations in LLM responses.

Scoring of Output

The scoring method is shown in ►**Table 2**. For OE, one response could get a maximum score of 2. For SATA, one correct response is scored 1. As there are 12 to 14 options depending on SATA variants, the maximum achievable score of each variant in the SATA question ranges from 12 to 14.

Statistical Methods

The results of the study are presented using descriptive statistics, including numbers and percentages. The overall score was normalized by dividing it by the maximum achievable score. For example, if an LLM gets a total of 3 $(0+1+2+0)$ in four variants from the prompt of RAD1, a total of 5 $(2+2+1+0)$ in four variants from the prompt of RAD2, and a total of 8 $(2+2+2+2)$ in four variants from the prompt of RAD3, then the normalized score $=(3+5+8)/24=0.67$. This was done to compare the scores among the

Table 1 Variants for suspected pulmonary embolism according to American College of Radiology Appropriateness Criteria®

Variant	Description
Variant 1	Suspected pulmonary embolism. Low or intermediate pretest probability with a negative D-dimer. Initial imaging.
Variant 2	Suspected pulmonary embolism. Low or intermediate pretest probability with a positive D-dimer. Initial imaging.
Variant 3	Suspected pulmonary embolism. High pretest probability. Initial imaging.
Variant 4	Suspected pulmonary embolism. Pregnant patient. Initial imaging.

Source: https://acsearch.acr.org/docs/69404/Narrative/.

Table 2 Scoring method of output of large language model

Response in open-ended prompt	Score	Response in select all that apply	Score
"Usually not appropriate" according to ACR criteria	0	Classifies an imaging procedure as "inappropriate" and ACR criteria state that it is "usually appropriate" or "may be appropriate" according to ACR criteria	0
"Is not listed" by ACR criteria	0	Classifies an imaging procedure as "appropriate" or "may be appropriate" and ACR criteria state that it is "usually not appropriate" according to ACR criteria	0
"May be appropriate" according to ACR criteria	1	Suggests imaging not listed by ACR	0
"Usually appropriate" according to ACR criteria	2	Classifies an imaging procedure as "Usually appropriate" and ACR criteria state that it is "usually appropriate" or "may be appropriate"	1
LLM suggested no further imaging which is correct as per ACR. (see Variant 1)	2	Classifies an imaging procedure as "inappropriate" and ACR criteria state that it is "usually not appropriate"	1

Abbreviation: ACR, American College of Radiology; LLM, large language model.
The maximum achievable score in open-ended response is 2. For select all that apply, the maximum achievable score may vary from 12 to 14 according to the options provided as for each correct prediction, the maximum achievable number is 1.

LLMs. We used GraphPad Prism 9.5.0 and Microsoft Excel 2010. A p-value <0.05 was considered statistically significant.

Ethical Issues

As this study was conducted with data available in the public domain and we only audited responses of AI that are being offered as free services. This study is exempted from ethical review according to National Ethical Guidelines for Biomedical and Health Research Involving Human Participants (2017).

Results

The accuracy level in the OE question was the highest in Perplexity (0.83) and lowest in Claude (0.58). The scores of LLMs in four variants received by three radiologists are shown in ►**Table 3**.

The accuracy level in the SATA question was the highest in Bing (0.96) and lowest in Perplexity (0.56). The scores of LLMs in four variants received by three radiologists are shown in ►**Table 4**.

Comparative accuracy among the LLMs is expressed in both OE and SATA in ►**Fig. 1**. Overall, the score in OE (0.73) was higher than the score in SATA (0.68).

The agreement among the scores obtained by three radiologists in OE was Intraclass Correlation Coefficient (ICC) $=-0.067$, $p=0.54$. The agreement among the scores obtained by three radiologists in SATA was ICC $=0.875$, $p<0.001$.

Only Claude and ChatGPT in SATA showed statistically significant ICC. The ICC of scores according to LLMs are shown in ►**Table 5**.

Discussion

The results of our study found that Bing had the highest overall accuracy and Claude had the lowest overall accuracy. However, perplexity had the highest accuracy for OE format prompts and Bing had the highest accuracy for SATA format prompts. It is to be noted that SATA prompts need higher order thinking (analyze, synthesize) where Bing performs the best. No previous studies have analyzed Bing for assessing ACR AC.

For, Variant 1 (Suspected pulmonary embolism. Low or intermediate pretest probability with a negative D-dimer. Initial imaging.), imaging tests are not necessary as per ACR AC. The suggestion of any imaging for this prompt will lead to low-value diagnostic imaging which has little or no impact on the management of the individual patient but may add

Table 3 Accuracy of all four large language models and annotation score on open-ended prompt

Variant	Bing			Claude			ChatGPT			Perplexity		
	RAD1	RAD2	RAD3	RAD1	RAD2	RAD3	RAD1	RAD2	RAD3	RAD1	RAD2	RAD3
V1 (2)	2	0	2	2	0	0	2	0	0	0	0	2
V2 (2)	2	0	2	2	0	2	2	2	0	2	2	2
V3 (2)	2	0	2	0	2	2	2	2	2	2	2	2
V4 (2)	2	2	2	0	2	2	2	2	2	2	2	2
Total (8)	8	2	8	4	4	6	8	6	4	6	6	8
Normalized score	0.75			0.58			0.75			0.83		

Abbreviations: RAD, radiologist; V, number; Variant, maximum achievable number.

Table 4 Accuracy of all four large language models and annotation score on "select all that apply" prompt

Variant	Bing			Claude			ChatGPT			Perplexity		
	RAD1	RAD2	RAD3	RAD1	RAD2	RAD3	RAD1	RAD2	RAD3	RAD1	RAD2	RAD3
V1 (12)	12	12	12	4	4	3	8	7	4	4	6	4
V2 (13)	13	12	13	9	8	10	7	7	7	9	5	12
V3 (13)	13	11	13	6	11	8	7	7	6	7	10	12
V4 (14)	13	12	14	9	13	9	12	10	12	6	3	9
Total (52)	51	47	52	28	36	30	34	31	29	26	24	37
Normalized score	0.96			0.6			0.6			0.56		

Abbreviations: RAD, radiologist; V, number; Variant, maximum achievable number.

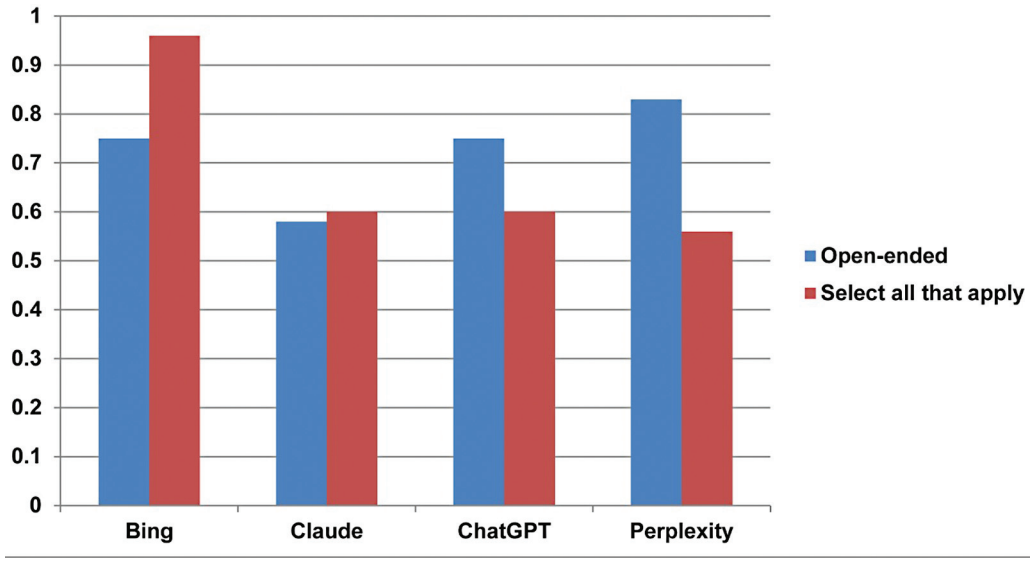

Fig. 1 Accuracy scores of answers of four large language models in open-ended and select all that apply type of questions.

Table 5 Intraclass Correlation Coefficients of individual large language models

Statistics	Bing		Claude		ChatGPT		Perplexity	
	OE	SATA	OE	SATA	OE	SATA	OE	SATA
ICC	–	0.45	−4	0.86	0.73	0.91	1	0.65
p-Value	–	0.24	0.89	0.01[a]	0.16	0.007[a]	–	0.13

Abbreviations: ICC, Intraclass Correlation Coefficient; OE, open-ended; SATA, select all that apply.
–, Could not be computed.
[a]Statistically significant *p*-value of ICC.

costs and an unnecessary risk to patients due to exposure to ionizing radiation and/or contrast media.[24] For this variant, Bing performed excellently for SATA prompts (accuracy 100%) and more than average (accuracy 66.66%) for OE prompts. Other chatbots mostly adopt a maximalist approach in clinical decision-making and lack the capability to recognize scenarios where imaging is unnecessary. This finding is similar to as shown by Rao et al.[19]

Prior studies have explored and compared the performance of ChatGPT and Bard chatbots in evaluating responses to ACR AC for liver lesions,[18] conducting breast cancer screening and pain assessments,[19] and 21 urologic clinical situations.[20] Studies by Patil et al[18] showed that ChatGPT-4 was able to correctly identify usually appropriate imaging per the AC more often than Bard in contrast to Doddi et al,[20] where both performed similarly. However, it is to be noted the subject domain was different (liver imaging vs. urologic imaging).

This is a first-of-its-kind study where we have evaluated four widely available free LLMs.

Our study revealed that all chatbots provided inconsistent responses across three different radiologists, with Bing exhibiting the highest inconsistency in OE responses. Statistically significant ICCs were observed only in Claude and

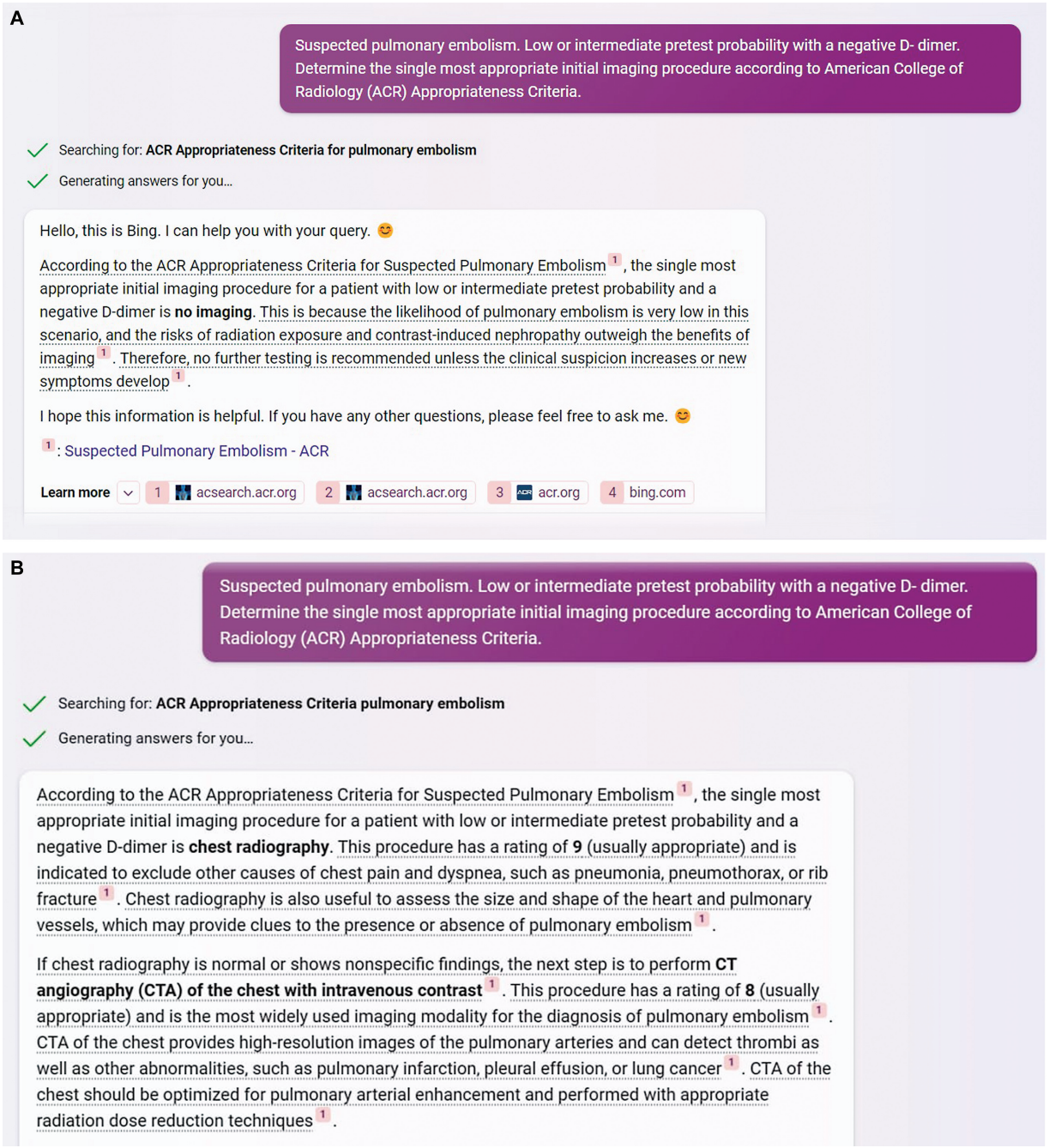

Fig. 2 (**A**, **B**) Response to variant 1 OE prompt for Bing for RAD1 (Accurate, score is 2) whereas inaccurate answer for the same prompt for RAD 2 (score is 0). OE, open-ended; RAD, radiologist.

ChatGPT for SATA questions (►**Table 5**), although their accuracy was lower compared to Bing. To the best of our knowledge, this is the first study to analyze intraclass correlation among observers to identify variations in responses. It is worth noting that Doddi et al mentioned this aspect in their paper, albeit without explicit analysis.[20] An instance of the variation in responses to the OE prompt for variant 1, as provided to Bing and obtained by RAD1 and RAD2, is depicted in ►**Fig. 2A, B**. Despite the inconsistency in responses from the chatbots, they hold promise in assisting health care providers in identifying the most suitable imaging modality.

Our study demonstrates the potential use of LLMs as an adjunct for radiologic decision-making at the point of care. The finding that Bing displays overall good accuracy in determining appropriate imaging steps for patients suspected of PE suggests a promising application for AI in the medical field.

The mention of the intricacy of radiologic decision-making and the emphasis on appropriate imaging utilization based on initial clinical presentations highlight the challenges in health care that AI can potentially address. However, it is important to consider factors such as the dataset used for training, the generalizability of the model, and ethical considerations surrounding the use of AI in health care.

Future research and validation studies will likely be crucial to further establish the reliability and effectiveness of LLMs in aiding medical decision-making. Additionally, collaboration between health care professionals and AI experts will be important to ensure the responsible and ethical integration of AI technologies in the field of radiology and beyond. In the future, the creation of an interactive chatbot that incorporates all appropriate ACR AC will prove beneficial in terms of both time and cost savings.

Limitations

The generated responses in the future may vary from those gathered during this study as LLMs are continually evolving. We have used only free untrained LLMs, however, studies have reported that fine-tuned GPT 3.5 and GPT 4 models have shown better responses.[25] Even with these free LLMs, our study shows that the accuracy level in the SATA question was the highest in Bing (0.96) which is a free LLM based on GPT-4. It is also noteworthy to mention that SATA questions need higher order thinking than OE questions. GPT-4 which is known to have better reasoning capabilities would likely yield even better results, although it is only available through OpenAI ChatGPT as a paid update. Highlighting the performance of free LLM versions, particularly Bing's impressive accuracy in SATA questions demonstrates the potential of these tools in clinical decision-making.

It is important to acknowledge that the responses of LLMs are influenced by the prompt's structure.[26] While our study maintained a consistent prompt by all radiologists, rewording prompts could lead to varied responses.

We believe that fine-tuning these free LLMs, utilizing Retrieval Augmented Generation and other techniques to ground the LLM's outputs to a particular knowledge base may give consistent and accurate results but warrants further research.

Conclusion

In conclusion, when evaluating various models of LLMs for both OE and SATA questions, significant variations in accuracy were observed. Notably, perplexity emerged as the top performer in OE questions, while Bing exhibited superior performance in SATA questions. It is noteworthy that, overall, LLMs demonstrated higher proficiency in OE questions compared to closed-ended questions and Bing exhibited the highest overall accuracy when considering responses for both OE and SATA questions. Despite these observations, it is crucial to acknowledge that the current LLMs lack consistent accuracy. Therefore, further development is imperative to enhance their reliability and effectiveness in clinical settings. These findings underscore the importance of additional training for LLMs and the necessity for careful selection by radiologists when considering their implementation.

Conflict of Interest
None declared.

References

1 Panayides AS, Amini A, Filipovic ND, et al. AI in medical imaging informatics: current challenges and future directions. IEEE J Biomed Health Inform 2020;24(07):1837–1857
2 Bera K, O'Connor G, Jiang S, Tirumani SH, Ramaiya N. Analysis of ChatGPT publications in radiology: literature so far. Curr Probl Diagn Radiol 2024;53(02):215–225
3 Tippareddy C, Jiang S, Bera K, et al. Radiology reading room for the future: harnessing the power of large language models like ChatGPT. Curr Probl Diagn Radiol 2023 (e-pub ahead of print). Doi: 10.1067/j.cpradiol.2023.08.018
4 Amin KS, Davis MA, Doshi R, Haims AH, Khosla P, Forman HP. Accuracy of ChatGPT, Google Bard, and Microsoft Bing for simplifying radiology reports. Radiology 2023;309(02):e232561
5 Jeblick K, Schachtner B, Dexl J, et al. ChatGPT makes medicine easy to swallow: an exploratory case study on simplified radiology reports. Eur Radiol 2024;34(05):2817–2825
6 Elkassem AA, Smith AD. Potential use cases for ChatGPT in radiology reporting. Am J Roentgenol 2023;221(03):373–376
7 Sarangi PK, Lumbani A, Swarup MS, et al. Assessing ChatGPT's proficiency in simplifying radiological reports for healthcare professionals and patients. Cureus 2023;15(12):e50881
8 Gertz RJ, Bunck AC, Lennartz S, et al. GPT-4 for automated determination of radiological study and protocol based on radiology request forms: a feasibility study. Radiology 2023;307(05): e230877
9 Haver HL, Ambinder EB, Bahl M, Oluyemi ET, Jeudy J, Yi PH. Appropriateness of breast cancer prevention and screening recommendations provided by ChatGPT. Radiology 2023;307(04): e230424
10 Haver HL, Lin CT, Sirajuddin A, Yi PH, Jeudy J. Use of ChatGPT, GPT-4, and Bard to improve readability of ChatGPT's answers to common questions about lung cancer and lung cancer screening. Am J Roentgenol 2023;221(05):701–704
11 Rahsepar AA, Tavakoli N, Kim GHJ, Hassani C, Abtin F, Bedayat A. How AI responds to common lung cancer questions: ChatGPT vs Google Bard. Radiology 2023;307(05):e230922
12 Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. Radiology 2023;307(05):e230582
13 Sarangi PK, Narayan RK, Mohakud S, Vats A, Sahani D, Mondal H. Assessing the capability of ChatGPT, Google Bard, and Microsoft Bing in solving radiology case vignettes. Indian J Radiol Imaging 2023;34(02):276–282
14 Sarangi PK, Irodi A, Panda S, Nayak DSK, Mondal H. Radiological differential diagnoses based on cardiovascular and thoracic imaging patterns: perspectives of four large language models. Indian J Radiol Imaging 2023;34(02):269–275
15 Kottlors J, Bratke G, Rauen P, et al. Feasibility of differential diagnosis based on imaging patterns using a large language model. Radiology 2023;308(01):e231167
16 Sun Z, Ong H, Kennedy P, et al. Evaluating GPT4 on impressions generation in radiology reports. Radiology 2023;307(05): e231259
17 Rau A, Rau S, Zoeller D, et al. A context-based chatbot surpasses trained radiologists and generic ChatGPT in following the ACR Appropriateness Guidelines. Radiology 2023;308(01):e230970
18 Patil NS, Huang RS, van der Pol CB, Larocque N. Using artificial intelligence chatbots as a radiologic decision-making tool for liver imaging: Do ChatGPT and Bard communicate information consistent with the ACR Appropriateness Criteria? J Am Coll Radiol 2023;20(10):1010–1013
19 Rao A, Kim J, Kamineni M, et al. Evaluating GPT as an adjunct for radiologic decision making: GPT-4 versus GPT-3.5 in a breast imaging pilot. J Am Coll Radiol 2023;20(10):990–997

20 Doddi S, Hibshman T, Salichs O, et al. Assessing appropriate responses to ACR urologic imaging scenarios using ChatGPT and Bard. Curr Probl Diagn Radiol 2024;53(02):226–229
21 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(01):45
22 European Society of Radiology (ESR) Methodology for ESR iGuide content. Insights Imaging 2019;10(01):32
23 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(06):531–537
24 Kjelle E, Andersen ER, Krokeide AM, et al. Characterizing and quantifying low-value diagnostic imaging internationally: a scoping review. BMC Med Imaging 2022;22(01):73
25 Gamble JL, Ferguson D, Yuen J, Sheikh A. Limitations of GPT-3.5 and GPT-4 in Applying Fleischner Society Guidelines to incidental lung nodules. Can Assoc Radiol J 2023;75:412–416
26 Sarangi PK, Mondal H. Response generated by large language models depends on the structure of the prompt. Indian J Radiol Imaging 2024 (e-pub ahead of print). Doi: 10.1055/s-0044-1782165

