Evaluating Large Language Models for Automated Reporting and Data Systems Categorization: Cross-Sectional Study

Qingxia Wu, Qingxia Wu, Huali Li, Yan Wang, Yan Bai, Yaping Wu, Xuan Yu, Xiaodong Li, Pei Dong, Jon Xue, Dinggang Shen, Meiyun Wang

Submitted to: JMIR Medical Informatics
on: December 25, 2023

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on it's website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressively prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Evaluating Large Language Models for Automated Reporting and Data Systems Categorization: Cross-Sectional Study

Qingxia Wu[1*]; Qingxia Wu[2, 3*]; Huali Li[4]; Yan Wang[5]; Yan Bai[5]; Yaping Wu[5]; Xuan Yu[5]; Xiaodong Li[1]; Pei Dong[2, 6]; Jon Xue[7]; Dinggang Shen[7, 8]; Meiyun Wang[5]

[1]Henan Provincial People’s Hospital & People’s Hospital of Zhengzhou University Zhengzhou CN
[2]Beijing United Imaging Research Institute of Intelligent Imaging Beijing CN
[3]Department of Radiology Luoyang Central Hospital Luoyang CN
[4]Department of Medical Imaging Henan Provincial People’s Hospital & People’s Hospital of Zhengzhou University Zhengzhou CN
[5]United Imaging Intelligence (Beijing) Co.,Ltd. Beijing CN
[6]Shanghai United Imaging Intelligence Co.,Ltd. Shanghai CN
[7]School of Biomedical Engineering Shanghai Tech University Shanghai CN
[*]these authors contributed equally

Corresponding Author:
Meiyun Wang
Department of Medical Imaging
Henan Provincial People’s Hospital & People’s Hospital of Zhengzhou University
No 7, Weiwu Road, Jinshui District
Zhengzhou
CN

Abstract

Background: Large language models (LLMs) show promise for improving radiology workflows, but their performance on structured radiological tasks such as Radiology Reporting and Data Systems (RADS) categorization remains unexplored.

Objective: To evaluate three LLM chatbots - Claude-2, GPT-3.5, and GPT-4 - on assigning Reporting and Data Systems (RADS) categories to radiology reports and assess the impact of different prompting strategies.

Methods: This cross-sectional study compared three chatbots using 30 radiology reports (10 per RADS criteria), utilizing a three-level prompting strategy: zero-shot, few-shot, and guideline PDF-informed prompts. The cases were grounded in LI-RADS® CT/MRI v2018, Lung-RADS® v2022, and O-RADS™ MRI, meticulously prepared by board-certified radiologists. Each report underwent six assessments. Two blinded reviewers assessed the chatbots' response at patient-level RADS categorization and overall ratings. The agreement across repetitions was assessed using Fleiss's kappa.

Results: Claude-2 achieved the highest accuracy in overall ratings with few-shot prompts and guideline PDFs (Prompt-2), attaining 57% (17/30) average accuracy over six runs and 50% (15/30) accuracy with k-pass voting. Without prompt engineering, all chatbots performed poorly. The introduction of a structured exemplar prompt (Prompt-1) increased the accuracy of overall ratings for all chatbots. Providing Prompt-2 further improved Claude-2’s performance, an enhancement not replicated by GPT-4. The inter-run agreement was substantial for Claude-2 (k=0.66 for overall rating, k=0.69 for RADS categorization), fair for GPT-4 (k=0.39 for both), and fair for GPT-3.5 (k=0.21 for overall rating and k=0.39 for RADS categorization). All chatbots showed significantly higher accuracy with LI-RADS v2018 compared to Lung-RADS v2022 and O-RADS (P<.05), with Prompt-2, Claude-2 achieved the highest overall rating accuracy of 75% (45/60) in LI-RADS v2018.

Conclusions: When equipped with structured prompts and guideline PDFs, Claude-2 demonstrated potential in assigning RADS categories to radiology cases according to established criteria such as LI-RADS v2018. However, the current generation of chatbots lags in accurately categorizing cases based on more recent RADS criteria.

(JMIR Preprints 25/12/2023:55799)
DOI: https://doi.org/10.2196/preprints.55799

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
✓ **Only make the preprint title and abstract visible.**
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?
✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**
Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

Original Manuscript

Evaluating Large Language Models for Automated Reporting and Data Systems Categorization: Cross-Sectional Study

Abstract

Background: Large language models (LLMs) show promise for improving radiology workflows, but their performance on structured radiological tasks such as Radiology Reporting and Data Systems (RADS) categorization remains unexplored.

Objectives: To evaluate three LLM chatbots - Claude-2, GPT-3.5, and GPT-4 - on assigning Reporting and Data Systems (RADS) categories to radiology reports and assess the impact of different prompting strategies.

Methods: This cross-sectional study compared three chatbots using 30 radiology reports (10 per RADS criteria), utilizing a three-level prompting strategy: zero-shot, few-shot, and guideline PDF-informed prompts. The cases were grounded in LI-RADS® CT/MRI v2018, Lung-RADS® v2022, and O-RADS™ MRI, meticulously prepared by board-certified radiologists. Each report underwent six assessments. Two blinded reviewers assessed the chatbots' response at patient-level RADS categorization and overall ratings. The agreement across repetitions was assessed using Fleiss's kappa.

Results: Claude-2 achieved the highest accuracy in overall ratings with few-shot prompts and guideline PDFs (Prompt-2), attaining 57% (17/30) average accuracy over six runs and 50% (15/30) accuracy with k-pass voting. Without prompt engineering, all chatbots performed poorly. The introduction of a structured exemplar prompt (Prompt-1) increased the accuracy of overall ratings for all chatbots. Providing Prompt-2 further improved Claude-2's performance, an enhancement not replicated by GPT-4. The inter-run agreement was substantial for Claude-2 (k=0.66 for overall rating, k=0.69 for RADS categorization), fair for GPT-4 (k=0.39 for both), and fair for GPT-3.5 (k=0.21 for overall rating and k=0.39 for RADS categorization). All chatbots showed significantly higher accuracy with LI-RADS v2018 compared to Lung-RADS v2022 and O-RADS (*P*<.05), with Prompt-2, Claude-2 achieved the highest overall rating accuracy of 75% (45/60) in LI-RADS v2018.

Conclusions: When equipped with structured prompts and guideline PDFs, Claude-2 demonstrated potential in assigning RADS categories to radiology cases according to established criteria such as LI-RADS v2018. However, the current generation of chatbots lags in accurately categorizing cases based on more recent RADS criteria.

Keywords

Radiology Reporting and Data Systems;LI-RADS; Lung-RADS; O-RADS; artificial intelligence; machine learning; large language model; ChatGPT; chatbot; chatbots; categorization; recommendation; recommendations; accuracy

Introduction

Since ChatGPT's public release in November 2022, large language models (LLMs) have attracted great interest in medical imaging applications [1]. Research indicated that ChatGPT showed promising applications in various aspects of the medical imaging process. Even without radiology-specific pretraining, LLMs can pass board examinations [2], provide radiology decision support [3], assist in differential diagnosis [3–6], and generate impressions from findings or structured reports [7-9].These applications not only accelerate the imaging diagnosis process, alleviate the workload of doctors but also improve the accuracy of diagnosis [10]. However, limitations exist, with one study showing ChatGPT-3 producing erroneous answers for a third of daily clinical questions and about 63% of provided references were not found [11]. ChatGPT's dangerous tendency to produce inaccurate responses is less frequent in GPT-4 but still limits usability in medical education and practice at present [12]. Tailoring LLMs to radiology may enhance reliability, as an appropriateness criteria context aware chatbot outperformed generic chatbots and radiologists [12].

The American College of Radiology (ACR) Reporting and Data Systems (RADS) standardizes communication of imaging findings. As of August 2023, there have been nine disease-specific systems endorsed by ACR, referring to products from the lexicons to report templates [13]. RADS reduces terminology variability, facilitates communication between radiologists and referring physicians, allows consistent evaluations, and conveys clinical significance to improve care. However, complexity and unfamiliarity limit adoption. Consequently, endeavors should be pursued to broaden the implementation of RADS. Therefore, we conducted this study to evaluate LLM's capabilities on a focused RADS assignment task for radiology reports.

A prompt serves as a directive or instruction given to LLMs to generate a particular response. The technique of "prompt tuning" has emerged as a valuable approach to refine the performance of LLMs, particularly for specific domains or tasks [14] . By providing structured queries or exemplary responses, the output of chatbots can be tailored for accurate and relevant answers. Such prompt tuning strategies leverage LLMs' knowledge while guiding appropriate delivery for particular challenges [14]. Given the complexity and specificity of the RADS categorization, our investigation emphasizes different prompt impacts to assess chatbot capabilities and potential performance enhancement through refined prompting tuning.

In this study, our primary objective was to meticulously evaluate the performance of three LLMs (GPT-3.5, GPT-4, and Claude-2) for RADS categorization using different prompt tuning strategies. We aimed to test their accuracy and consistency in RADS categorization and shed light on the potential benefits and limitations of relying on chatbot-derived information for the categorization of specific RADS.

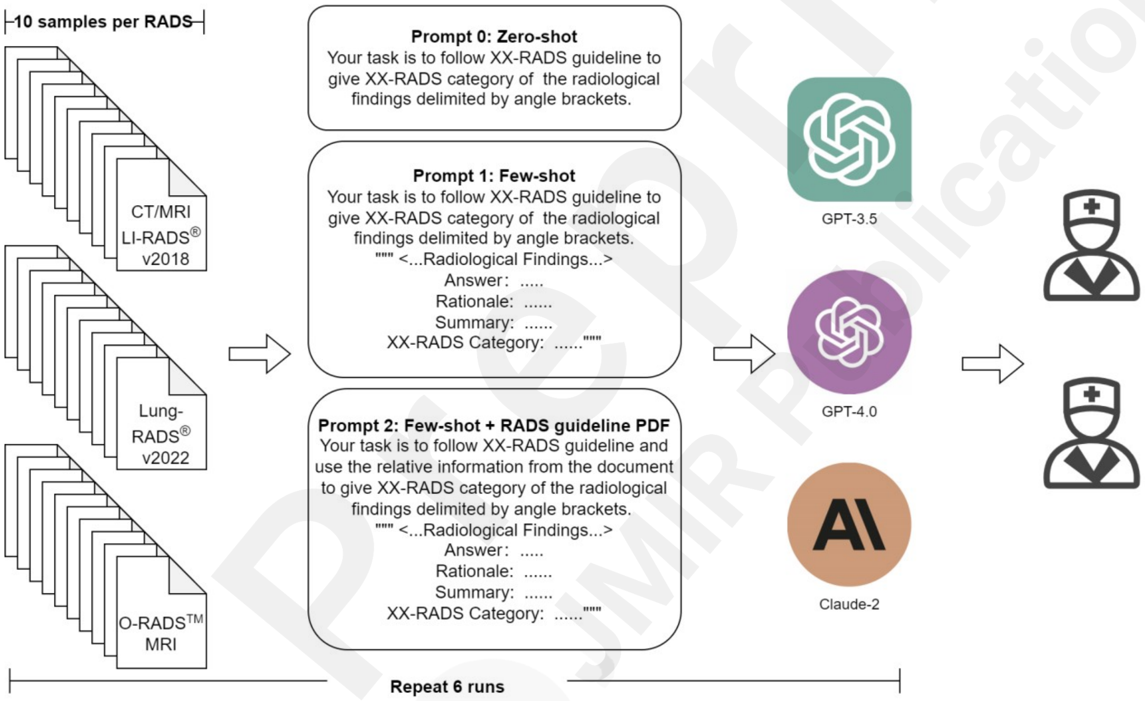
Methods

This study was deemed exempt by the Institutional Review Board, owing to the absence of human subject involvement.

Study Design

The workflow of the study is shown in Figure 1. We conducted a cross-sectional analysis in September 2023 to evaluate the competency of three chatbots - GPT-3.5, GPT-4 (OpenAI, August 30(th), 2023 version) [15], and Claude-2 (Anthropic) [16] - in the task of assigning three RADS categorizations to radiology reports. Given the chatbot's knowledge cessation was as of September 2021, we opted for Liver Imaging Reporting & Data System (LI-RADS®) CT/MRI v2018 [17], Lung

CT Screening Reporting & Data System (Lung-RADS®) v2022 [18], and Ovarian-Adnexal Reporting & Data System (O-RADS™) MRI (developed in 2022) [19] as the yardsticks to compare the responses engendered by GPT-3.5, GPT-4, and Claude-2. A total of thirty radiology reports for either CT or MRI examinations were composed for this analysis, with ten cases representing each of the three RADS reporting systems. The radiology reports used for testing were generated by radiologists with more than 10-year experiences to correct the wording styles from real-life cases based on respective RADS systems. For each RADS (i.e., LI-, Lung-, and O-RADS), we tried to reflect the complexity and diversity so that the reports cover typical cases in clinical practice. Therefore, reports with 2-3 simple cases and 7-8 challenging cases were generated for one RADS. These include scenarios such as prior exam comparison, the presence of multiple nodules, extensive categorization under different RADS systems, and updates from the most recent LI-RADS and Lung-RADS guidelines. The characteristics of radiology reports for each RADS and the distribution of the number of the reports across the three RADS were shown in Multimedia Appendix 1. The objective was to evaluate the performance of chatbots on a highly structured radiology workflow task involving cancer risk categorization based on structured report inputs. The study design focused on a defined use case to illuminate the strengths and limitations of existing natural language processing technology in this radiology sub-domain.



Prompt 0: Zero-shot
Your task is to follow XX-RADS guideline to give XX-RADS category of the radiological findings delimited by angle brackets.

Prompt 1: Few-shot
Your task is to follow XX-RADS guideline to give XX-RADS category of the radiological findings delimited by angle brackets.
""" <...Radiological Findings...>
Answer: .....
Rationale: ......
Summary: ......
XX-RADS Category: ......"""

Prompt 2: Few-shot + RADS guideline PDF
Your task is to follow XX-RADS guideline and use the relative information from the document to give XX-RADS category of the radiological findings delimited by angle brackets.
""" <...Radiological Findings...>
Answer: .....
Rationale: ......
Summary: ......
XX-RADS Category: ......"""

Figure 1. Flowchart of study design.

Prompts

We collected and analyzed responses from GPT-3.5, GPT-4, and Claude-2 for each case. To mitigate bias, the radiological findings were presented individually via separate interactions, with corresponding responses saved for analysis. Three prompt templates were designed to elicit each RADS categorization along with explanatory rationale:

Prompt-0 was a zero-shot prompt, merely introducing the RADS assignment task, such as “Your task is to follow Lung-RADS® v2022 guideline to give Lung-RADS category of the radiological findings delimited by angle brackets.”

Prompt-1 was a few-shot prompt, furnishing an exemplar of RADS categorization including the reasoning, summarized impression, and final category. For example: ” Your task is to follow Lung-RADS® v2022 guideline to give Lung-RADS category of the radiological findings delimited by angle brackets. """ < …Radiological Findings… > Answer：Rationale: {…} Overall: {…} Summary: {…} Lung-RADS Category: X """ ”.

Prompt-2 distinctly instructed chatbots to consult the PDF of corresponding RADS guidelines, compensating for these chatbots' lack of radiology-specific pretraining. For Claude-2, the PDF could be directly ingested, while GPT-4 required the use of an "Ask for PDF" plugin to extract pertinent information [20,21].

Each case was evaluated six times with each chatbot across the three prompt levels. The representative radiological reports and prompts are shown in Multimedia Appendix 2. The links to all the prompts and guideline PDFs are shown in Multimedia Appendix 3.

Evaluation of chatbots

Two study authors (Q.W. and H.L.) independently evaluated the following for each chatbot response in a blinded manner, with any discrepancies resolved by a third senior radiologist (Y. W.). The following were assessed for each response:

(1) Patient-level RADS categorization: Judged as correct, incorrect, or unsure. "Correct" denotes that the chatbot accurately identified the patient-level RADS category, irrespective of the rationale provided. "Unsure" denotes that the chatbot's response failed to provide a decisive RADS category. For example, a response articulating that "a definitive Lung-RADS category cannot be assigned" would be categorized as "unsure".

(2) Overall rating: Assessed as either correct or incorrect. A response is judged incorrect if any errors are identified, including:

E1 - Factual extraction error, denotes the chatbots' inability to paraphrase the radiological findings accurately, consequently misinterpreting the information.

E2 - Hallucination, encompassing the fabrication of nonexistent RADS categories (E2a) and RADS criteria (E2b).

E3 - Reasoning error, which includes the incapacity to logically interpret the imaging description (E3a) and the RADS category accurately (E3b). The subtype errors for reasoning imaging description include the inability to reason lesion signal (E3ai), lesion size (E3aii), and/or enhancement (E3aiii) accurately.

E4 - Explanatory error, encompassing inaccurate elucidation of RADS category meaning (E4a) and erroneous explanation of the recommended management and follow-up corresponding to the RADS category (E4b).

If a chatbot's feedback manifested any of the aforementioned infractions, it was labeled as incorrect, with the specific type of error documented. To assess the consistency of the evaluations, a k-pass voting method was also applied. Specifically, a case was deemed accurately categorized if it met the criteria in a minimum of 4 out of the 6 runs.

Statistical analyses

The accuracy of the patient-level RADS categorization and overall rating for each chatbot was compared using the chi-squared test. The agreement across the six repeated runs was assessed using

Fleiss's kappa. Agreement strength was interpreted as follows: <0 signified poor, 0-0.20 indicated slight, 0.21-0.40 represented fair, 0.41-0.60 was interpreted as moderate, 0.61-0.80 denoted substantial, and 0.81-1 was characterized as almost perfect. Statistical significance was defined as 2-sided *P* < .05. All analyses were performed using R statistical software version 4.1.2 (R Foundation for Statistical Computing).

Results

Performance of Chatbots

The performance of Chatbots is shown in Figure 2, Table 1 and Table 2, with the links to case-level details provided in Multimedia Appendix 4. For the overall rating (Table 1, Average row), Claude-2 with Prompt-2 demonstrated significantly higher average accuracy across the 30 cases compared to Claude-2 with Prompt-0 (odds ratio [OR] =8.16, p<.001). GPT-4 with Prompt-2 also showed improved average accuracy compared to GPT-4 with Prompt-0, but the difference was not statistically significant (OR=3.19, *P*=.13). When using the k-pass voting method (Table 1, K-pass voting row), Claude-2 with Prompt-2 had significantly higher accuracy than Claude-2 with Prompt-0 (OR=8.65, *P*=.002). Similarly, GPT-4 with Prompt-2 was significantly more accurate than GPT-4 with Prompt-0 (OR=11.98, *P*=.01). For the exact assignment of the patient-level RADS categorization (Table 2, Average row), Claude-2 with Prompt-2 showed significantly more average accuracy than Claude-2 with Prompt-0 (*P*=.04).

Table 1. Correct overall ratings of different chatbots and prompts.

Chatbots and Prompts	Prompt-0	Prompt-1	Prompt-2
GPT-3.5			
Run1	3(10)[3,28]	9(30)[15,50]	
Run2	3(10)[3,28]	9(30)[15,50]	
Run3	4(13)[4,32]	7(23)[11,43]	
Run4	4(13)[4,32]	5(17)[6,35]	
Run5	3(10)[3,28]	6(20)[8,39]	
Run6	3(10)[3,28]	4(13)[4,32]	
Average[a]	3(10)[3,28]	7(23)[11,43]	
K-pass voting[b]	1(3)[0,19]	2(7)[1,24]	
GPT-4			
Run1	4(13)[4,32]	11(37)[21,56]	12(40)[23,59]
Run2	4(13)[4,32]	7(23)[11,43]	8(27)[13,46]
Run3	4(13)[4,32]	9(30)[15,50]	9(30)[15,50]
Run4	2(7)[1,24]	9(30)[15,50]	13(43)[26,62]
Run5	5(17)[6,35]	11(37)[21,56]	8(27)[13,46]
Run6	6(20)[8,39]	9(30)[15,50]	8(27)[13,46]
Average[a]	4(13)[4,32]	9(30)[15,50]	10(33)[18,53]
K-pass voting[b]	1(3)[0,19]	6(20)[8,39]	**9(30)[15,50]**[c]
Claude-2			
Run1	4(13)[4,32]	10(33)[18,53]	19(63)[44,79]
Run2	5(17)[6,35]	8(27)[13,46]	16(53)[35,71]
Run3	5(17)[6,35]	7(23)[11,43]	15(50)[33,67]
Run4	5(17)[6,35]	6(20)[8,39]	17(57)[38,74]
Run5	3(10)[3,28]	7(23)[11,43]	18(60)[41,77]
Run6	3(10)[3,28]	7(23)[11,43]	14(47)[29,65]
Average[a]	4(13)[4,32]	8(27)[13,46]	**17(57)[38,74]**[c]
K-pass voting[b]	3(10)[3,28]	7(23)[11,43]	**15(50)[33,67]**[c]

Note. — Data are numbers of correct overall ratings (n=30), with percentages in parentheses and 95% CIs in brackets. [a] accuracy by the average method. [b] accuracy by k-pass voting (≥4/6 runs correct). [c] significant between Prompt-0 and Prompt-2.CI = Confidence interval

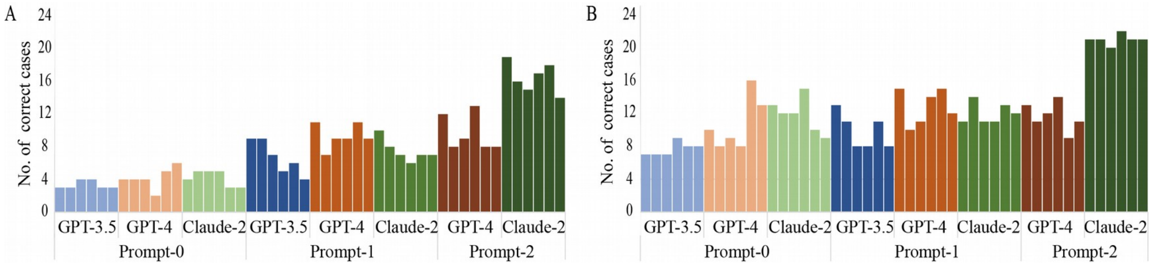


Figure 2. Bar graphs show the comparison of chabot performance across six runs regarding (A) overall rating, (B) patient-level RADS categorization. RADS = Reporting and Data Systems.

Table 2. The number of correct, incorrect, and unsure responses for patient-level RADS categorization across different chatbots and prompts.

Chatbots and Prompts	Run1	Run2	Run3	Run4	Run5	Run6	Average[a]	K-pass voting[b]
GPT-3.5								
Prompt-0	7/23/0	7/23/0	7/23/0	9/21/0	8/21/1	8/20/2	8/22/0	7/23/0
Prompt-1	13/15/2	11/19/0	8/21/1	8/21/1	11/19/0	8/22/0	10/20/0	7/23/0
GPT-4								
Prompt-0	10/20/0	8/19/3	9/20/1	8/22/0	16/14/0	13/15/2	11/18/1	8/22/0
Prompt-1	15/14/1	10/18/2	11/18/1	14/15/1	15/14/1	12/18/0	13/16/1	11/19/0
Prompt-2	13/16/1	11/18/1	12/18/0	14/16/0	9/21/0	11/16/3	12/18/0	11/19/0
Claude-2								
Prompt-0	13/17/0	12/18/0	12/18/0	15/15/0	10/20/0	9/21/0	12/18/0	13/17/0
Prompt-1	11/19/0	14/16/0	11/19/0	11/19/0	13/17/0	12/18/0	12/18/1	11/19/0
Prompt-2	21/9/0	21/9/0	20/10/0	22/8/0	21/9/0	2021/8/1	**21/9/0**[c]	21/9/0

Note. — Data are numbers of correct/incorrect/unsure patient-level RADS categories. [a] accuracy by the average method. [b] accuracy by k-pass voting (≥4/6 runs correct). [c] significant between Prompt-0 and Prompt-2. RADS = Reporting and Data Systems.

Consistency of Chatbots

As shown in Table 3, among the thirty cases evaluated in six runs, Claude-2 with Prompt-2 showed substantial agreement (k=0.65 for overall rating; k=0.74 for RADS categorization). GPT-4, when interfaced with Prompt-2, demonstrated moderate agreement (k=0.46 for overall rating; k=0.41 for RADS categorization). When evaluated with Prompt-1, GPT-4 presented moderate agreement (k=0.38 for overall rating; k=0.42 for RADS categorization). In contrast, Claude-2 showed substantial agreement (k=0.63 for overall rating; k=0.61 for RADS categorization), while GPT-3.5 exhibited a range from slight to fair agreement. With Prompt-0, Claude-2 showed moderate agreement (k=0.49) for overall rating, and substantial agreement for RADS categorization (k=0.65). GPT4 manifested slight agreement (k=0.19) for the overall rating and fair agreement for RADS categorization. Meanwhile, GPT-3.5 showed fair agreement (k=0.28) for the overall rating and moderate agreement (k=0.57) for RADS categorization.

Table 3. The consistency of different chatbots and prompts between six runs.

	Prompt-0	Prompt-1	Prompt-2	All
Patient-level RADS categorization				
GPT-3.5	0.57(0.48,0.65)	0.24(0.15,0.32)		0.39(0.33,0.46)
GPT-4	0.33(0.25,0.42)	0.42(0.34,0.5)	0.41(0.33,0.5)	0.39(0.34,0.44)
Claude-2	0.65(0.56,0.74)	0.61(0.52,0.7)	0.74(0.65,0.83)	0.69(0.64,0.74)
Overall rating				
GPT-3.5	0.28(0.19,0.37)	0.14(0.05,0.23)		0.21(0.14,0.27)
GPT-4	0.19(0.1,0.28)	0.38(0.29,0.47)	0.46(0.37,0.55)	0.39(0.34,0.45)
Claude-2	0.49(0.4,0.58)	0.63(0.53,0.72)	0.65(0.56,0.75)	0.66(0.61,0.72)

Note. — Data are the results from Fleiss's kappa. RADS = Reporting and Data Systems.

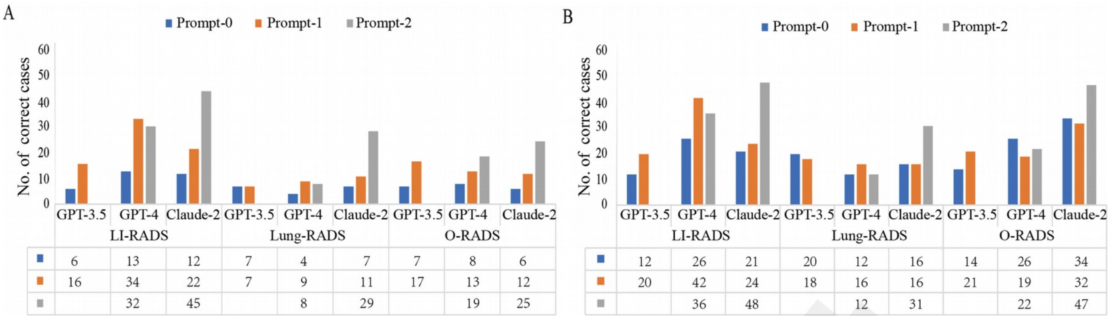
Subgroup analysis

Since the knowledge base for ChatGPT was frozen as of September 2021, accounting for the knowledge limitations of LLMs developed before the latest RADS guideline updates, we compared the responses of different RADS criteria. The total accurate responses across six runs were computed for all prompts. Both GPT-4 and Claude-2 demonstrated superior performance in the context of LI-RADS CT/MRI v2018 as opposed to Lung-RADS v2022 and O-RADS MRI (all p<0.05, Table 4). Figure 3 delineates the performance of various chatbots across different prompts and RADS categories. For the overall rating (Figure 3A), Claude-2 exhibited a progressive trend of enhancement of overall rating accuracy from Prompt-0 to Prompt-1 to Prompt-2, with 20.0% (12/60), 36.7% (22/60), and 75.0% (45/60) for LIRADS, 11.7% (7/60), 18.3% (11/60), and 48.3% (29/60) for Lung-RADS, and 10.0% (6/60), 20.0% (12/60), and 41.7% (25/60) for O-RADS, respectively. Notably, with Prompt-2, Claude-2 achieved the highest overall rating accuracy of 75% in older systems like LI-RADS v2018. Conversely, GPT-4 improved with Prompt-1/2 over Prompt-0, but Prompt-2 did not exceed Prompt-1. For the RADS categorization (Figure 3B), Prompt-1 and Prompt-2 outperformed Prompt-0 for LI-RADS, irrespective of chatbots. However, for Lung-RADS and O-RADS, Prompt-0 sometimes superseded Prompt-1.

Table 4. The performance of chatbots within different RADS criteria.

Chatbots and RADS	Develop time	RADS categorization [Correct/Incorrect/Unsure]	*P* value	Overall rating [Correct/Incorrect]	*P* value
GPT-3.5					
LI-RADS CT/MRI	2018	32/86/2	Reference	22/98	Reference
Lung-RADS	2022	38/78/4	0.83	14/106	0.15
O-RADS MRI	2022	35/84/1	0.46	24/96	0.87
GPT-4					
LI-RADS CT/MRI	2018	104/74/2	Reference	78/102	Reference
Lung-RADS	2022	40/128/12	**<.001**	21/159	**<.001**
O-RADS MRI	2022	67/110/3	**<.001**	40/140	**<.001**
Claude-2					
LI-RADS	2018	93/86/1	Reference	79/101	Reference
Lung-RADS	2022	63/117/0	**0.001**	47/133	**<.001**
O-RADS MRI	2022	113/67/0	**0.04**	43/137	**<.001**

Note. — Data are aggregate numbers across six runs. RADS = Reporting and Data Systems.



A

	LI-RADS			Lung-RADS			O-RADS		
	GPT-3.5	GPT-4	Claude-2	GPT-3.5	GPT-4	Claude-2	GPT-3.5	GPT-4	Claude-2
🟦	6	13	12	7	4	7	7	8	6
🟧	16	34	22	7	9	11	17	13	12
⬜		32	45		8	29		19	25

B

	LI-RADS			Lung-RADS			O-RADS		
	GPT-3.5	GPT-4	Claude-2	GPT-3.5	GPT-4	Claude-2	GPT-3.5	GPT-4	Claude-2
🟦	12	26	21	20	12	16	14	26	34
🟧	20	42	24	18	16	16	21	19	32
⬜		36	48		12	31		22	47

Figure 3. The performance of chatbots and prompts within different RADS criteria. (A) overall rating, (B) patient-level RADS categorization. RADS = Reporting and Data Systems.

Analysis of error types

A total of 1440 cases were analyzed for error types, with details provided in Multimedia Appendix 4. The bar plot illustrating the distribution of errors across the three chatbots is shown in Figure 4. A typical example of factual extraction error (E1) occurred in response to the 7th Lung-RADS question. The statement "The 3mm solid nodule in the lateral basal segmental bronchus is subsegmental" is inaccurate, as the lateral basal segmental bronchus represents one of the 18 defined lung segments, not a subsegment [22].

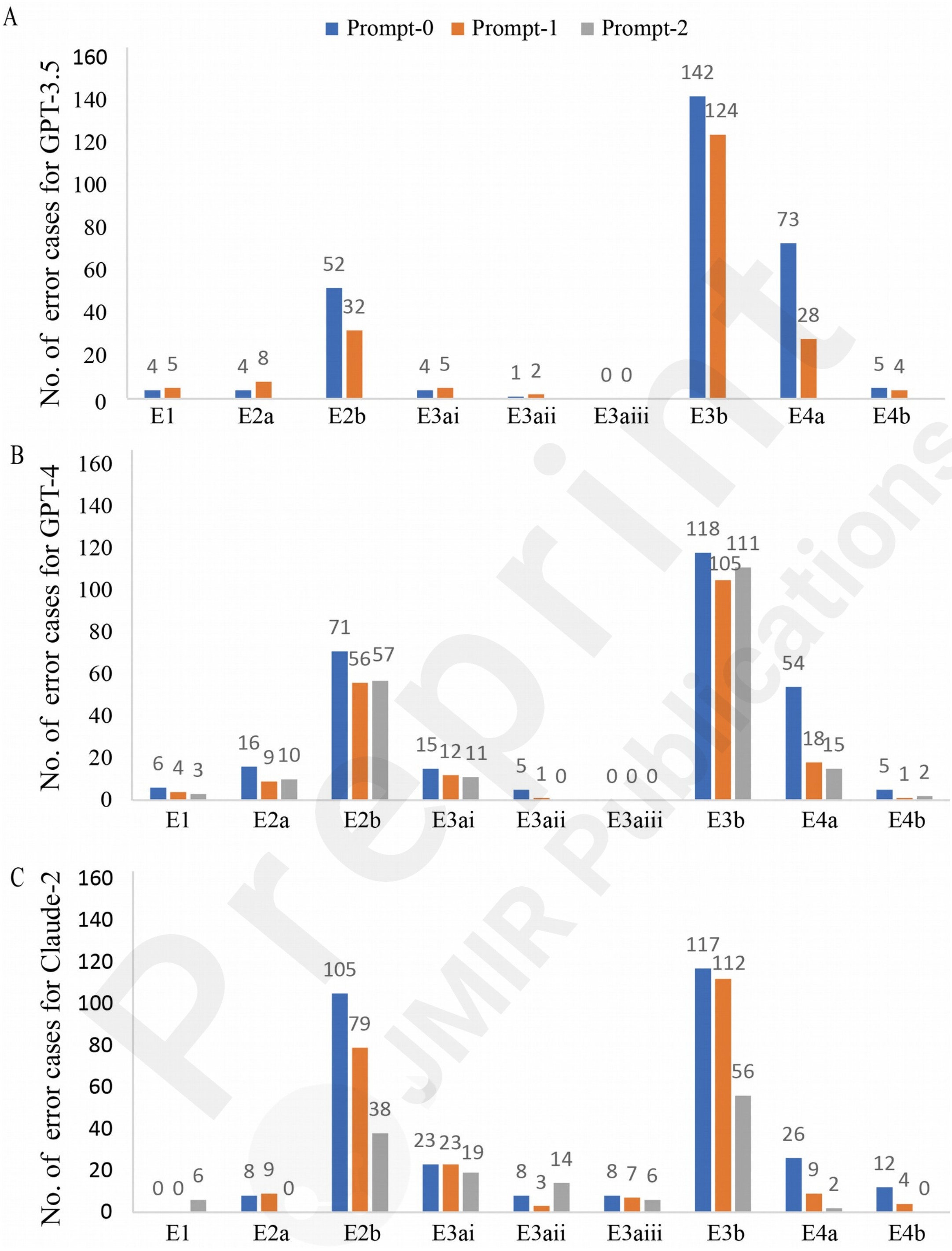


Figure 4. The number of error types for different chatbots. E1 - Factual extraction error, denotes the chatbots' inability to paraphrase the radiological findings accurately, consequently misinterpreting the information. E2 - Hallucination, encompassing the fabrication of nonexistent RADS categories (E2a) and RADS criteria (E2b). E3 - Reasoning error, which includes the incapacity to logically interpret the imaging description (E3a) and the RADS category accurately (E3b). The subtype errors for reasoning imaging description include the inability to reason lesion signal (E3ai),

lesion size (E3aii), and/or enhancement (E3aiii) accurately. E4 - Explanatory error, encompassing inaccurate elucidation of RADS category meaning (E4a) and erroneous explanation of the recommended management and follow-up corresponding to the RADS category (E4b). RADS = Reporting and Data Systems.

Hallucination of inappropriate RADS categories (E2a) occurred more frequently with Prompt-0 across all three chatbots. However, this error rate decreased to zero for Claude-2 when using Prompt-2, a trend not seen with GPT-3.5 or GPT-4. A recurrent E2a error in LI-RADS was the obsolete category LR-5V from the 2014 version, now superseded by LR-TIV in subsequent editions [23,24]. Furthermore, hallucination of invalid RADS criteria (E2b) was more prevalent than E2a. For instance, the LI-RADS second question response stating "T2 marked hyperintensity is a feature commonly associated with hepatocellular carcinoma (HCC)" is inaccurate, as T2 marked hyperintensity is characteristic of hemangioma, not HCC. Despite initial higher E2b rates, Claude-2 demonstrated a substantial reduction with Prompt-2 (105 to 38 instances), exceeding the decrement seen with GPT-4 (71 to 57 instances).

Regarding reasoning error, incorrect RADS category reasoning (E3b) was the most frequent error but decreased for all chatbots with Prompt-1 and Prompt-2 versus Prompt-0. Claude-2 reduced errors by almost half with Prompt-2, while the GPT-4 decrease was less pronounced. Lesion signal interpretation errors (E3ai) included misinterpreting hypointensity on diffusion-weighted imaging (DWI) as "restricted diffusion," rather than facilitated diffusion. Lesion size reasoning errors (E3aii) occurred in 34 out of 1440 cases, predominantly by Claude-2 (25/34, 73.5%), especially in systems like Lung-RADS and LI-RADS where size is critical for categorization. Examples were attributing a 12mm pulmonary nodule to the >=6mm but <8mm range, or assigning a hepatic lesion measuring 2.3 by 1.5cm to the 10-19mm category. Reasoning enhancement errors (E3aiii) were exclusive to Claude-2 in O-RADS, where enhancement significantly impacts categorization. Misclassifying images at 40 seconds post-contrast as early or delayed enhancement exemplifies this error.

Explanatory errors (E4) including incorrect RADS category definitions (E4a) and inappropriate management recommendations (E4b) also substantially declined with Prompt-1 and Prompt-2. For instance, in the first Lung-RADS question response, the statement "The 4X designation indicates infectious/inflammatory etiology is suspected. " is incorrect. Lung-RADS 4X means Category 3 or 4 nodules with additional features or imaging findings that increase suspicion of lung cancer [18].

Discussion

In this study, we evaluated the performance of three chatbots - GPT-3.5, GPT-4, and Claude-2 - in categorizing radiological findings according to RADS criteria. Using three levels of prompts providing increasing structure, examples, and domain knowledge, the chatbots' accuracies and consistencies were quantified across 30 cases. The best performance was achieved by Claude-2 when provided with few-shot prompting and the RADS criteria PDFs. Interestingly, the chatbots tended to categorize better for the relatively older LI-RADS v2018 criteria in contrast to the more recent Lung-RADS v2022 and O-RADS guidelines published after the chatbots' training cutoff.

The incorporation of RADS, which standardizes reporting in radiology, has been a significant advancement, although the multiplicity and complexity of these systems impose a steep learning curve for radiologists [13]. Even for subspecialized radiologists at tertiary hospitals, mastering the numerous RADS guidelines poses challenges, requiring familiarity with the lexicons, regular application in daily practice, and ongoing learning to remain current with new versions. While previous studies have shown that LLMs could assist radiologists in various tasks [2–5,7,11], their performance at RADS categorization from imaging findings is untested. We therefore evaluated LLMs for focused RADS categorization of testing cases.

Without prompt engineering (Prompt-0), all chatbots performed poorly. However, accuracy improved for all chatbots when provided an exemplar prompt demonstrating the desired response structure (Prompt-1). This underscores the utility of prompt tuning for aligning LLMs to specific domains like radiology. Further enriching Prompt-1 with the RADS guideline PDFs as a relevant knowledge source (Prompt-2) considerably enhanced Claude-2's accuracy, a feat not mirrored by GPT-4. This discrepancy could stem from ChatGPT's reliance on an external plugin to access documents, while Claude-2's architecture accommodates the direct assimilation of expansive texts, benefiting from its larger context window and superior long document processing capabilities.

Notably, we discerned performance disparities across RADS criteria. When queried on older established guidelines like LI-RADS v2018 [17], the chatbots demonstrated greater accuracy compared to more recent schemes such as Lung-RADS v2022 and O-RADS [18,19,25]. Specifically, GPT-4 and Claude-2 had significantly higher total correct ratings for LI-RADS versus Lung-RADS and O-RADS (all $p<0.05$). This could be attributed to their extensive exposure to the voluminous data related to the matured LI-RADS during their pretraining phase. With Prompt-2, Claude-2 achieved 75% (45/60) accuracy for overall rating LI-RADS categorization. The poorer performance on newer RADS criteria highlights the need for strategies to continually align LLMs with the most up-to-date knowledge.

A deep dive into the error-type analysis revealed informative trends. Incorrect RADS category reasoning (E3b) constituted the most frequent error across chatbots, decreasing with prompt tuning. Targeted prompting also reduced critical errors like hallucinations of RADS criteria (E2b) and categories (E2a), likely by constraining output to valid responses. During pretraining, GPT-liked LLMs predict the next word in the unlabeled dataset, risking learning fallacious relationships between RADS features. For instance, Lung-RADS v2022 lacks categories 5 and 6 [18], though some other RADS like Breast Imaging Reporting and Data System include them [26]. Using Prompt-0, chatbots erroneously hallucinated Lung-RADS category 5 and 6. Explanatory errors (E4) including inaccurate definition of the assigned RADS category (E4a) and inappropriate management recommendations (E4b) also substantially declined with prompt tuning. For instance, when queried on the novel O-RADS criteria with Prompt-0, chatbots hallucinated follow-up recommendations from other RADS criteria, and responded " O-RADS category 3 refers to an indeterminate adnexal mass and warrants short-interval follow-up". Targeted prompting appears to mitigate these critical errors like hallucination and incorrect reasoning. Careful prompt engineering is essential to properly shape LLM knowledge for radiology tasks.

There are also several limitations in this study. First, only the LI-RADS CT/MRI and O-RADS MRI were included, excluding LI-RADS ultrasound (US) and O-RADS US guidelines which are often practiced in an independent Ultrasound department [27,28]. Second, the chatbot's performance was heavily dependent on prompt quality. We only test three types of prompts, further prompt strategies studies are warranted to investigate the impact of more exhaustive engineering on chatbots' accuracy. Third, GPT-4-turbo was released on November 6, 2023, representing the latest GPT-4 model with improvements in instruction following, reproducible outputs, and more [29]. Furthermore, its training data extends to April 2023 compared to September 2021 for the base GPT-4 model tested here. We are uncertain about this newest GPT-4-turbo model's performance on the RADS categorization task. Evaluating GPT-4-turbo represents an important direction for future work. Fourth, our study focused on 3 out of 9 RADS [13], with a limited 10 cases for each RADS category. Although our choice ensured a blend of old and new guidelines and tried to cover all the RADS scores as much as possible, extending evaluations to all the RADS guidelines and incorporating more radiology reports from real clinical scenarios could offer deeper insights into potential limitations. Nonetheless, this initial study highlights critical considerations of prompt design and knowledge calibration required for safely applying LLMs in radiology. Fifth, evaluating the performance of the LLM in comparison to radiologists of varying expertise levels proves valuable for discerning its strengths and weaknesses in real-world applications. This comparative analysis will be undertaken in our forthcoming studies.

In conclusion, when equipped with structured prompts and guideline PDFs, Claude-2 demonstrates potential in assigning RADS categories to radiology cases according to established criteria such as LI-RADS v2018. However, the current generation of chatbots lags in accurately categorizing cases based on more recent RADS criteria. Our study highlights the potential of LLMs in streamlining radiological categorizations while also pinpointing the enhancements necessary for their dependable application in clinical practice for RADS categorization tasks.

Acknowledgments

This study has received funding by National Natural Science Foundation of China (82371934, 82001783), Joint Fund of Henan Province Science and Technology R&D Program (225200810062).

The authors thank Chuanjian Lv, MD, Zejun Wen, MM, and Jianghua Lou, MM for their help in drafting the radiology reports with regard to Lung-RADS, LI-RADS, and O-RADS respectively.

Authors' Contributions

QW, QW, HL, YW, YB, YW, XY, and MW contributed to study design. QW and QW contributed to statistical analysis. All authors contributed to the acquisition, analysis, or interpretation of data; the drafting of the manuscript; and critical revision of the manuscript.

Conflicts of Interest

QW and PD are senior engineers of Beijing United Imaging Research Institute of Intelligent Imaging and United Imaging Intelligence (Beijing) Co., Ltd. JX and DS are senior specialists of Shanghai United Imaging Intelligence Co., Ltd. The companies have no role in designing and performing the surveillance and analyzing and interpreting the data. All other authors report no conflict of interest relevant to this article.

Multimedia Appendix 1

The characteristics of radiology reports for each RADS and the distribution of the number of the reports across the three RADS.

Multimedia Appendix 2

Representative radiology reports and prompts.

Multimedia Appendix 3

Links to Prompts and guideline PDFs.

Multimedia Appendix 4

Links to Prompt Engineering Results.

References

1. Li R, Kumar A, Chen JH. How Chatbots and Large Language Model Artificial Intelligence Systems Will Reshape Modern Medicine: Fountain of Creativity or Pandora's Box? JAMA Intern Med. 2023;183(6):596. doi: 10.1001/jamainternmed.2023.1835

2. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a Radiology Board-style Examination: Insights into Current Strengths and Limitations. Radiology. 2023;307(5):e230582. doi: 10.1148/radiol.230582

3. Rao A, Kim J, Kamineni M, Pang M, Lie W, Dreyer KJ, Succi MD. Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot. J Am Coll Radio. 2023;S1546144023003940. doi: 10.1016/j.jacr.2023.05.003

4. Ueda D, Mitsuyama Y, Takita H, Horiuchi D, Walston SL, Tatekawa H, Miki Y. ChatGPT's Diagnostic Performance from Patient History and Imaging Findings on the Diagnosis Please Quizzes. Radiology. 2023;308(1):e231040. doi: 10.1148/radiol.231040

5. Kottlors J, Bratke G, Rauen P, Kabbasch C, Persigehl T, Schlamann M, Lennartz S. Feasibility of Differential Diagnosis Based on Imaging Patterns Using a Large Language Model. Radiology. 2023;308(1):e231167. doi: 10.1148/radiol.231167

6. Gertz RJ, Bunck AC, Lennartz S, Dratsch T, Iuga A-I, Maintz D, Kottlors J. GPT-4 for Automated Determination of Radiologic Study and Protocol Based on Radiology Request Forms: A Feasibility Study. Radiology. 2023;307(5):e230877. doi: 10.1148/radiol.230877

7. Adams LC, Truhn D, Busch F, Kader A, Niehues SM, Makowski MR, Bressem KK. Leveraging GPT-4 for Post Hoc Transformation of Free-text Radiology Reports into Structured Reporting: A Multilingual Feasibility Study. Radiology. 2023;307(4):e230725. doi: 10.1148/radiol.230725

8. Wagner MW, Ertl-Wagner BB. Accuracy of Information and References Using ChatGPT-3 for Retrieval of Clinical Radiological Information. Can Assoc Radiol J 2023;084653712311711. doi: 10.1177/08465371231171125

9. Ziegelmayer S, Marka AW, Lenhart N, Nehls N, Reischl S, Harder F, Sauter A, Makowski M, Graf M, Gawlitza J. Evaluation of GPT-4's Chest X-Ray Impression Generation: A Reader Study on Performance and Perception. J Med Internet Res. 2023;25:e50865. doi: 10.2196/50865

10. Bhayana R, Bleakney RR, Krishna S. GPT-4 in Radiology: Improvements in Advanced Reasoning. Radiology. 2023;307(5):e230987. doi: 10.1148/radiol.230987

11. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. J Med Internet Res. 2023;25:e48568. doi: 10.2196/48568

12. Rau A, Rau S, Zöller D, Fink A, Tran H, Wilpert C, Nattenmüller J, Neubauer J, Bamberg F, Reisert M, Russe MF. A Context-based Chatbot Surpasses Radiologists and Generic ChatGPT in Following the ACR Appropriateness Guidelines. Radiology. 2023;308(1):e230970. doi: 10.1148/radiol.230970

13. Reporting and Data Systems. Available from: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems [accessed August 26, 2023].

14. Meskó B. Prompt Engineering as an Important Emerging Skill for Medical Professionals: Tutorial. J Med Internet Res. 2023;25:e50638. doi: 10.2196/50638

15. OpenAI. Available from: https://openai.com [accessed November 8, 2023].

16. Claude 2. Anthropic.Available from: https://www.anthropic.com/index/claude-2 [accessed November 8, 2023].

17. Chernyak V, Fowler KJ, Kamaya A, Kielar AZ, Elsayes KM, Bashir MR, Kono Y, Do RK, Mitchell DG, Singal AG, Tang A, Sirlin CB. Liver Imaging Reporting and Data System (LI-RADS) Version 2018: Imaging of Hepatocellular Carcinoma in At-Risk Patients. Radiology. 2018;289(3):816–830. doi: 10.1148/radiol.2018181494

18. Martin MD, Kanne JP, Broderick LS, Kazerooni EA, Meyer CA. RadioGraphics Update: Lung-RADS 2022. RadioGraphics. 2023;43(11):e230037. doi: 10.1148/rg.230037

19. Sadowski EA, Thomassin-Naggara I, Rockall A, Maturen KE, Forstner R, Jha P, Nougaret S, Siegelman ES, Reinhold C. O-RADS MRI Risk Stratification System: Guide for Assessing Adnexal Lesions from the ACR O-RADS Committee. Radiology. 2022;303(1):35–47. doi: 10.1148/radiol.204371

20. AskYourPDF. Available from: https://askyourpdf.com [accessed November 8, 2023].

21. ChatGPT plugins. Available from: https://openai.com/blog/chatgpt-plugins [accessed November 8, 2023].

22. Jones J, Rasuli B, Vadera S, et al. Bronchopulmonary segmental anatomy. Radiopaedia.org. https://doi.org/10.53347/rID-13644 [accessed on 08 Nov 2023]

23. Mitchell DG, Bruix J, Sherman M, Sirlin CB. LI-RADS (Liver Imaging Reporting and Data System): Summary, discussion, and consensus of the LI-RADS Management Working Group and future directions. Hepatology. 2015;61(3):1056–1065. doi: 10.1002/hep.27304

24. Elsayes KM, Hooker JC, Agrons MM, Kielar AZ, Tang A, Fowler KJ, Chernyak V, Bashir MR, Kono Y, Do RK, Mitchell DG, Kamaya A, Hecht EM, Sirlin CB. 2017

Version of LI-RADS for CT and MR Imaging: An Update. RadioGraphics. 2017;37(7):1994–2017. doi: 10.1148/rg.2017170098

25. Suarez-Weiss KE, Sadowski EA, Zhang M, Burk KS, Tran VT, Shinagare AB. Practical Tips for Reporting Adnexal Lesions Using O-RADS MRI. RadioGraphics. 2023;43(7):e220142. doi: 10.1148/rg.220142

26. Breast Imaging Reporting & Data System. American College of Radiology..Available from: https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/Bi-Rads [accessed November 3, 2023].

27. Strachowski LM, Jha P, Phillips CH, Blanchette Porter MM, Froyman W, Glanc P, Guo Y, Patel MD, Reinhold C, Suh-Burgmann EJ, Timmerman D, Andreotti RF. O-RADS US v2022: An Update from the American College of Radiology's Ovarian-Adnexal Reporting and Data System US Committee. Radiology. 2023;308(3):e230685. doi: 10.1148/radiol.230685

28. Quaia E. State of the Art: LI-RADS for Contrast-enhanced US. Radiology. 2019;293(1):4–14. doi: 10.1148/radiol.2019190005

29. New models and developer products announced at DevDay Available from: https://openai.com/blog/new-models-and-developer-products-announced-at-devday [accessed November 9, 2023].

Abbreviations:

LLM = Large language model

ACR = American College of Radiology

RADS = Radiology Reporting and Data Systems

LI-RADS = Liver Imaging Reporting and Data System

Lung-RADS = Lung CT Screening Reporting and Data System

O-RADS = Ovarian-Adnexal Reporting and Data System

OR = Odds ratio

HCC = Hepatocellular carcinoma

DWI = Diffusion-weighted imaging

US = Ultrasound

CI = Confidence interval

Figure legends

Figure 1. Flowchart of study design.

Figure 2. The comparison of chabot performance across six runs. (A) overall rating, (B) patient-level RADS categorization. RADS = Reporting and Data Systems.

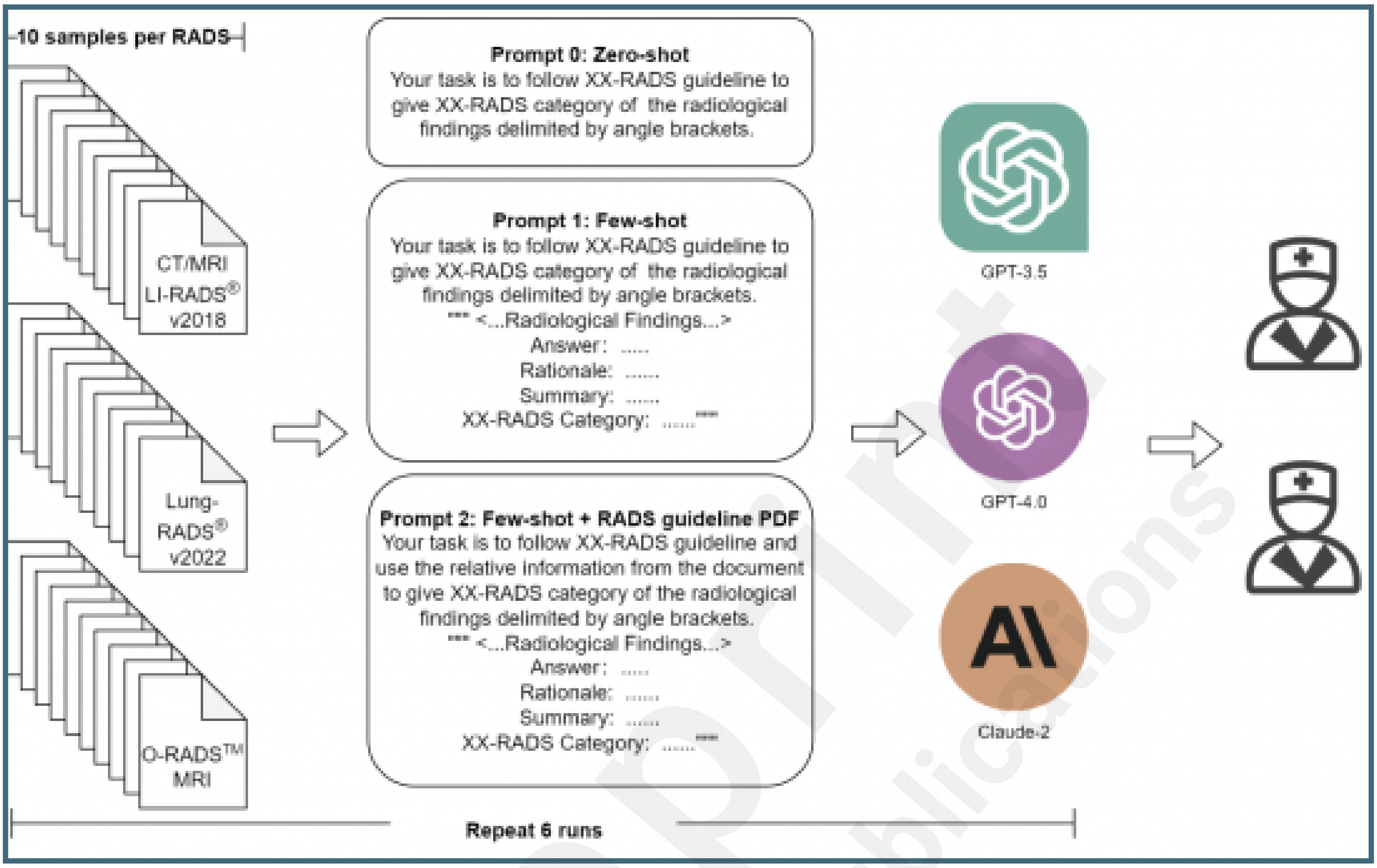
Figure 3. The performance of chatbots and prompts within different RADS criteria. (A) overall rating, (B) patient-level RADS categorization. RADS = Reporting and Data Systems.

Figure 4. The number of error types for different chatbots. E1 - Factual extraction error, denotes the chatbots' inability to paraphrase the radiological findings accurately, consequently misinterpreting the information. E2 - Hallucination, encompassing the fabrication of nonexistent RADS categories (E2a) and RADS criteria (E2b). E3 - Reasoning error, which includes the incapacity to logically interpret the imaging description (E3a) and the RADS category accurately (E3b). The subtype errors for reasoning imaging description include the inability to reason lesion signal (E3ai), lesion size (E3aii), and/or enhancement (E3aiii) accurately. E4 - Explanatory error, encompassing inaccurate elucidation of RADS category meaning (E4a) and erroneous explanation of the recommended management and follow-up corresponding to the RADS category (E4b). RADS = Reporting and Data Systems.

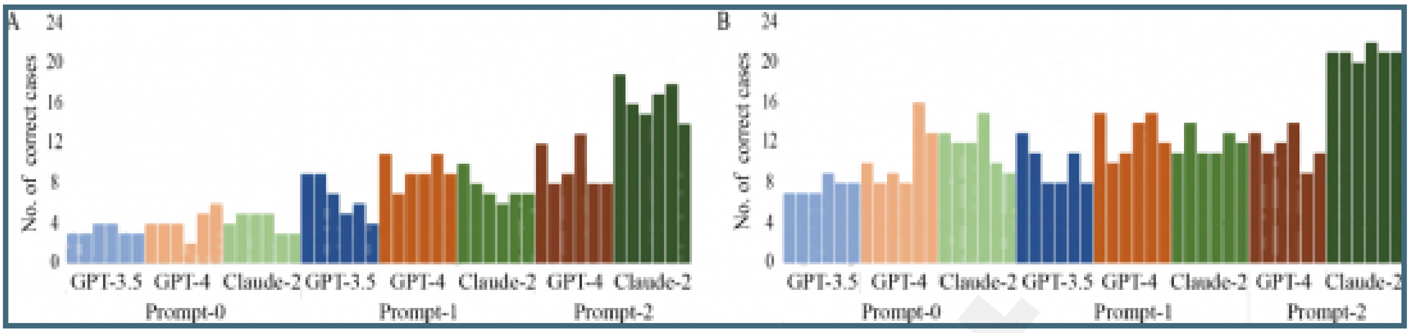
Supplementary Files

Figures

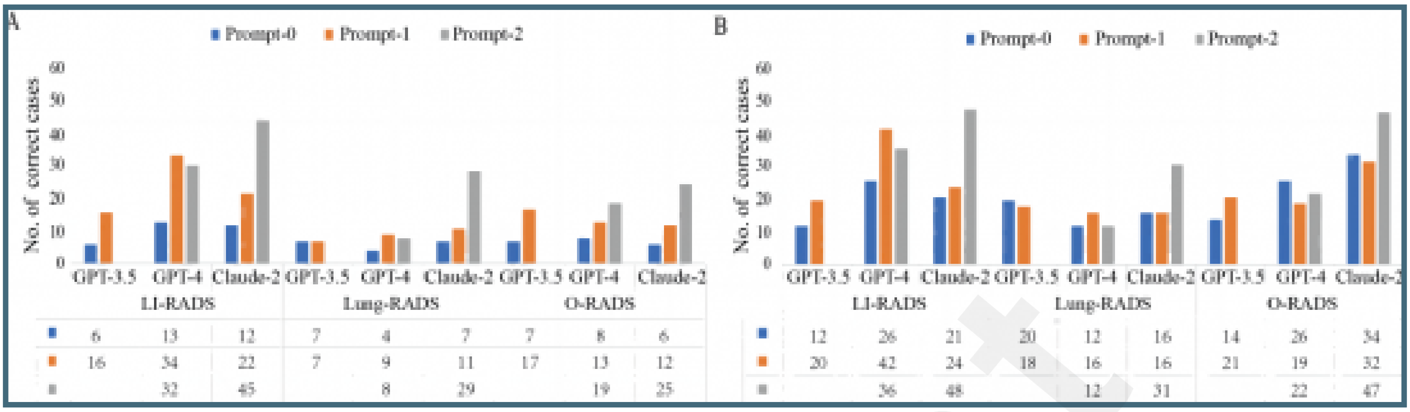
Flowchart of study design.


10 samples per RADS
CT/MRI LI-RADS® v2018
Lung-RADS® v2022
O-RADS™ MRI
Prompt 0: Zero-shot
Your task is to follow XX-RADS guideline to give XX-RADS category of the radiological findings delimited by angle brackets.
Prompt 1: Few-shot
Your task is to follow XX-RADS guideline to give XX-RADS category of the radiological findings delimited by angle brackets.
""" <...Radiological Findings...>
Answer: ......
Rationale: ......
Summary: ......
XX-RADS Category: ......"""
Prompt 2: Few-shot + RADS guideline PDF
Your task is to follow XX-RADS guideline and use the relative information from the document to give XX-RADS category of the radiological findings delimited by angle brackets.
""" <...Radiological Findings...>
Answer: ......
Rationale: ......
Summary: ......
XX-RADS Category: ......"""
GPT-3.5
GPT-4.0
Claude-2
Repeat 6 runs

The comparison of chabot performance across six runs.



The performance of chatbots and prompts within different RADS criteria.



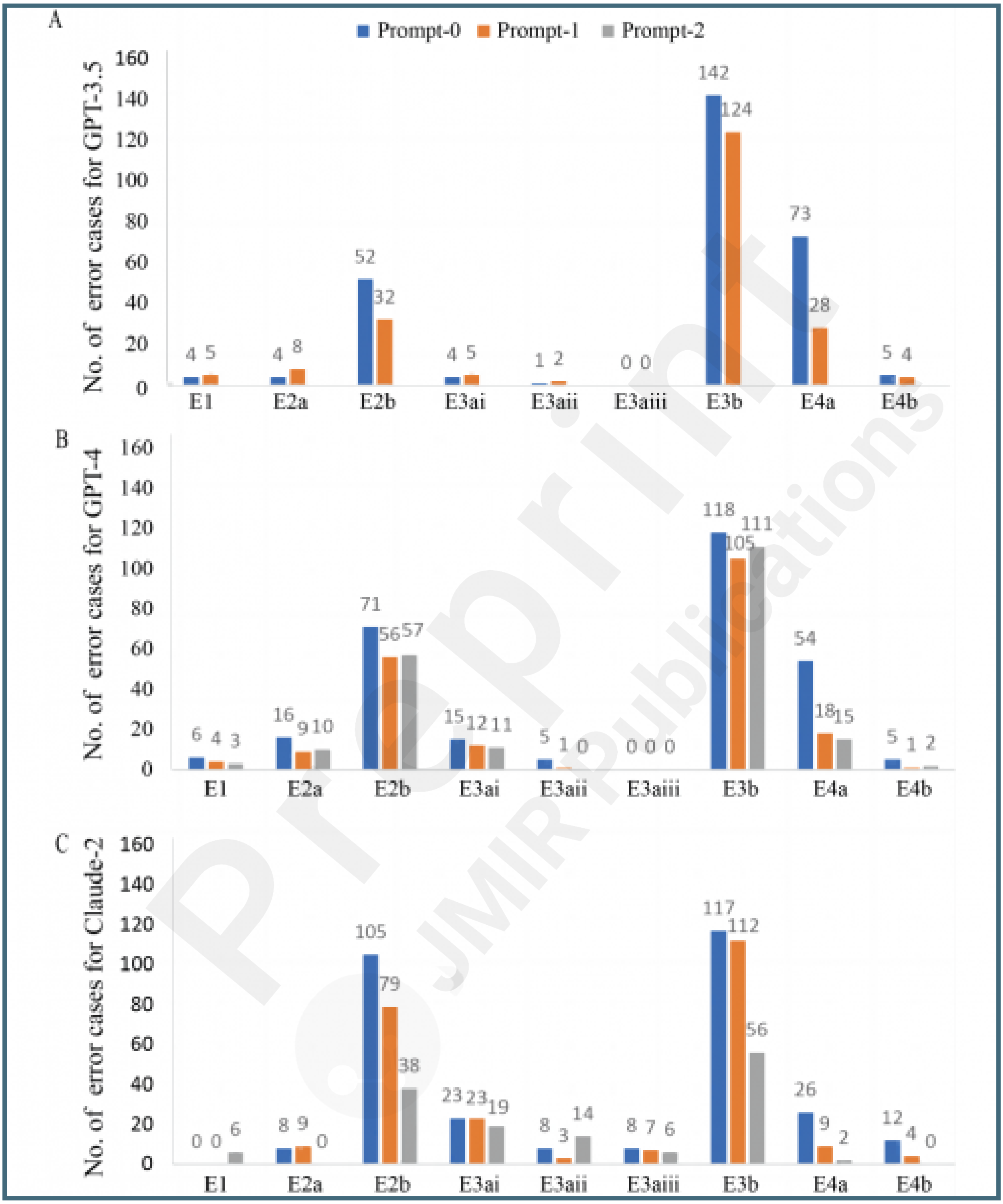
A

	LI-RADS			Lung-RADS			O-RADS		
	GPT-3.5	GPT-4	Claude-2	GPT-3.5	GPT-4	Claude-2	GPT-3.5	GPT-4	Claude-2
🟦	6	13	12	7	4	7	7	8	6
🟧	16	34	22	7	9	11	17	13	12
⬜		32	45		8	29		19	25

B

	LI-RADS			Lung-RADS			O-RADS		
	GPT-3.5	GPT-4	Claude-2	GPT-3.5	GPT-4	Claude-2	GPT-3.5	GPT-4	Claude-2
🟦	12	26	21	20	12	16	14	26	34
🟧	20	42	24	18	16	16	21	19	32
⬜		36	48		12	31		22	47

The number of error types for different chatbots.



Multimedia Appendixes

The characteristics of radiology reports for each RADS and the distribution of the number of the reports across the three RADS.
URL: http://asset.jmir.pub/assets/1372115bdaa1555b009df2c0f9752f3a.docx

Representative radiology reports and prompts.
URL: http://asset.jmir.pub/assets/a69616a01715cdc9d6b3630645662a26.docx

Links to Prompts and guideline PDFs.
URL: http://asset.jmir.pub/assets/cebb28af944abed48d299037b00f908e.docx

Links to Prompt Engineering Results.
URL: http://asset.jmir.pub/assets/d6899ec1ba4305591347f1e3834efbd7.docx