

Prompt Engineering in Large Language Models for BI-RADS Classification of Imaging Reports: A Retrospective Evaluation

Wenjie Liu

Second Affiliated Hospital of Nanchang University

Hailong Wu

Second Affiliated Hospital of Nanchang University

Yuanyuan Lang

Second Affiliated Hospital of Nanchang University

Yan Luo

Second Affiliated Hospital of Nanchang University

Yan Li

Second Affiliated Hospital of Nanchang University

Xinyi Liu

Second Affiliated Hospital of Nanchang University

Yinping Leng

Second Affiliated Hospital of Nanchang University

Lianggeng Gong

`gong111999@126.com`

Second Affiliated Hospital of Nanchang University

Research Article

Keywords:

Posted Date: October 19th, 2025

DOI: <https://doi.org/10.21203/rs.3.rs-7526460/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Purpose

To evaluate how prompt engineering modulates large language models' (LLMs) accuracy in Breast Imaging Reporting and Data System (BI-RADS) classification of digital breast tomosynthesis (DBT) reports.

Materials and Methods

This retrospective study collected reports from 216 patients who underwent DBT for breast cancer screening or diagnosis. BI-RADS classifications were independently assigned to all reports by two experts. Three LLMs (GPT-4o, GPT-o3 mini, Qwen-2.5 max) were utilized to classify all reports using different prompts. Besides, six human readers independently assigned BI-RADS classifications. Agreement between experts and LLMs for BI-RADS categories was evaluated using Weighted Cohen's kappa (κ_w). Friedman and Nemenyi tests assessed κ_w differences among three prompt conditions. The frequencies of changed BI-RADS category assignments, which could impact clinical management, were also calculated.

Results

In prompt III, GPT-4o achieved near-perfect agreement with experts ($\kappa_w = 0.80$), surpassing GPT-o3 mini (0.76) and Qwen-2.5 max (0.79). Its κ_w was significantly higher in prompt III than in prompt II (0.69, $P, P < 0.05$) and prompt I (0.63, $P < 0.01$). While GPT-4o's κ_w remained lower than two mid-level radiologists (0.89 and 0.86), it exceeded two entry-level radiologists (0.76 and 0.79). Regarding clinical management changes, prompt III yielded a 14.8% discordance rate with experts, outperforming prompts I (29.6%) and II (28.2%), and aligning with entry-level radiologists (15.3%, 14.4%).

Conclusion

With optimized prompts, GPT-4o achieved near-perfect agreement and matched the clinical management performance of entry-level radiologists. These findings support the use of LLMs as an auxiliary tool for BI-RADS classification in breast cancer diagnosis by radiologists.

Summary Statement

Optimized prompting enabled best-performing GPT-4o to achieve near-perfect BI-RADS agreement with experts in DBT reports and low clinical management misclassification rate, indicating LLMs' potential as auxiliary tools for radiologists.

Key Results

In Prompt III, GPT-4o demonstrated near-perfect agreement with experts (kw, 0.80), outperforming both GPT-o3 mini (0.76) and Qwen (0.79). Notably, the kw of GPT-4o in Prompt III was significantly higher than that observed in Prompt II (0.69, $P<0.05$) and Prompt I (0.63, $P<0.01$).

The optimization of prompts altered the proportion of BI-RADS category assignments that would result in negative changes in clinical management, aligning with the performance of entry-level radiologists (GPT-4o: Prompt I, 29.6%; Prompt II, 28.2 %; Prompt III, 14.8%. Entry-level radiologists, 15.3% and 14.4%).

Introduction

In breast cancer screening and management, the Breast Imaging Reporting and Data System (BI-RADS) plays a critical role in guiding clinical decision-making, treatment selection, and prognostic assessment (1, 2). However, accurate BI-RADS classification requires extensive clinical expertise, often leading to interobserver variability among radiologists with varying levels of experience (3). Furthermore, the BI-RADS classification criteria vary across different imaging modalities, including mammography, breast ultrasound, and breast magnetic resonance imaging, further complicating the grading process (4). These challenges in standardization and consistency emphasize the need for a more efficient solution. In recent years, large language models (LLMs) have emerged, offering a promising new approach to address these limitations (5, 6).

LLMs are large-scale artificial intelligence systems trained through deep learning techniques, capable of understanding, processing, and generating natural language (7, 8). In recent years, the rapid development of LLMs, particularly within the healthcare sector, has attracted significant interest, with their potential applications becoming increasingly recognized (9–11). Models such as the GPT series, exemplify LLMs' emergence as powerful tools in critical healthcare processes, including medical diagnosis, treatment planning, and patient management, owing to their robust text generation and comprehension capabilities (12–16). Next-generation LLMs, such as GPT-4o, demonstrate enhanced precision and consistency, particularly in areas such as semantic understanding, reasoning, and specialized applications (17).

In recent years, researchers have explored the potential of LLMs for BI-RADS classification from free-text radiology reports (5, 6). Although pre-trained LLMs exhibit some accuracy gaps compared to expert radiologists, they have demonstrated significant potential in improving classification consistency and reducing inter-observer variability (18). Prompt engineering has emerged as a key technique for enhancing LLM performance (19, 20); however, its effectiveness in BI-RADS classification has not been comprehensively explored. This study aimed to evaluate LLM performance in BI-RADS classification of free-text digital breast tomosynthesis (DBT) reports, with an emphasis on improving classification accuracy and consistency through prompt optimization. Various prompt engineering strategies were explored to enhance the model's predictive capability for BI-RADS categorization and support AI-assisted radiologic diagnosis.

Materials and Methods

Study Design and Sample

This retrospective study received approval from the Institutional Review Board of the Second Affiliated Hospital of Nanchang University, which also waived the requirement for written informed consent. Radiology reports from adult women who underwent DBT examinations at the Second Affiliated Hospital of Nanchang University from January 2024 to December 2024 were collected. All reports were authored by radiologists who underwent specialized training and were written exclusively in Chinese. Exclusion criteria included (1) incomplete reports, (2) post-surgical examination, (3) patients aged under 18 years, (4) BI-RADS 0 and 6 classifications (as these require additional information not included in the imaging descriptions), and (5) BI-RADS 1 classifications (due to the small number of reports). To ensure an equal distribution across BI-RADS categories (2–5), cases were created under strict adherence to these exclusion criteria. Relevant clinical information, such as patient age, was also collected. At no point during the study were the images from the examinations evaluated. A flowchart detailing the patient selection process is shown in Fig. 1.

Report Processing and Evaluation

The clinical information, findings, impression, and BI-RADS categories sections from all exported imaging reports were extracted into a spreadsheet for subsequent analysis. Two certified senior breast radiologists (with 28 and 30 years of post-training experience, respectively) were then invited to independently review these reports according to the 6th edition of the American College of Radiology BI-RADS classification system. Each radiologist assigned a BI-RADS category, which served as the reference. During the evaluation process, the two experts were only provided with the "Findings" section of the reports.

BI-RADS Category Assignment by LLMs

This study evaluated the performance of three LLMs in BI-RADS classification: GPT-4o (OpenAI, USA), GPT-o3 mini (OpenAI, USA), and Qwen-2.5 max (Alibaba, China), which are newer iterations of widely recognized conversational artificial intelligence applications.

Each model was tested with three different prompts, with both the imaging findings and prompts written in Chinese (**Table S1**). Prompt I included elements of role-playing, grading limitations, and specified task requirements. Prompt II, an extension of Prompt I, suggested assigning a higher grade when the BI-RADS classification was unclear. Prompt III further expanded on Prompt II by recommending special attention to seemingly benign lesions and carefully considering whether the grade should be elevated, consistent with established clinical practice. The findings from the reports and prompts were input into GPT-4o, GPT-o3 mini, and Qwen-2.5 max. After each report was input and a response received, the chat session

was restarted. These models were tested using a zero-shot learning approach (i.e., no example data was provided prior to testing). To assess reproducibility, 10 reports were randomly selected from each BI-RADS category and re-entered into the LLMs for evaluation after 7 days. For this study, these models were accessed between February 3, 2025, and February 20, 2025, during which the versions of the three models remained unchanged.

BI-RADS Category Assignment by Human Readers

This study invited six human readers, categorized by their experience levels: two in-training radiologists (both had 6 months of post-training experience and BI-RADS classification training), two entry-level breast radiologists (with 6 and 7 years of post-training experience, respectively), and two mid-level breast radiologists (with 12 and 10 years of post-training experience, respectively), hereafter referred to as resident 1, resident 2, entry-level radiologist 1, entry-level radiologist 2, mid-level radiologist 1, and mid-level radiologist 2. Each reader independently evaluated and assigned a BI-RADS classification to each report.

Statistical Analysis

All statistical analyses were performed using Python (version 3.9.13), with the following packages used: Statsmodels (0.13.2), Seaborn (0.11.2), Matplotlib (3.5.2), and Scikit-learn (1.6.1).

The Weighted Cohen's kappa (κ_w) statistic was used to assess the consistency across multiple categories. The κ_w values and corresponding 95% confidence intervals (CIs) were calculated for the comparison between all LLMs and the reference standard, as well as between the six human readers and the reference standard. The strength of inter-rater agreement was interpreted as follows: $\kappa_w < 0.00$ indicates poor agreement, 0.00–0.20 indicates slight agreement, 0.21–0.40 indicates fair agreement, 0.41–0.60 indicates moderate agreement, 0.61–0.80 indicates substantial agreement, and 0.81–1.00 indicates almost perfect agreement (21). The significance of the kappa value was assessed using a two-tailed z-test based on asymptotic normal distribution theory. To evaluate significant differences between prompt conditions, the Friedman test (22) was employed to assess overall differences in κ_w values across three prompt conditions, followed by a Nemenyi post-hoc test (22) to determine inter-group significance. A P -value < 0.05 was considered statistically significant.

Besides, the assigned BI-RADS categories were grouped according to their associated clinical management pathways: BI-RADS 2 (normal or benign; no intervention required), BI-RADS 3 (probably benign; short-term follow-up), and BI-RADS 4 or 5 (suspicious or highly suggestive of malignancy; biopsy recommended). To determine whether reclassification (downgrade or upgrade) by the six human readers or LLMs would have a negative impact on clinical management, the reference standard was the established scenario for clinical management.

Results

Study Sample

A total of 216 imaging reports were included in the study, with 54 reports allocated to each BI-RADS category. Table 1 summarizes the characteristics of the 216 female patients and their corresponding radiology reports. The mean age of the patients was 50 ± 10 years. All 216 DBT imaging reports were entirely written in Chinese, and the extracted report sections contained an average of 103.24 words.

Table 1
Baseline characteristics of patients.

Characteristics	Value
Sex (n = 216)	100%
Female	
Age (years)	50.48 ± 10.07
No. of words in extracted portions of report	103.24 ± 31.44
Report language (n = 216)	
Chinese only	100%
Examination type (n = 216)	
DBT	100%
Reference-standard BI-RADS classification based on radiology report (sixth edition)	
BI-RADS 2	54 (25%)
BI-RADS 3	54 (25%)
BI-RADS 4	54 (25%)
BI-RADS 5	54 (25%)
Data are presented as mean $\pm$ standard deviation for normally distributed continuous variables, or number (%) for categorical variables.	
DBT = digital breast tomosynthesis. BI-RADS = Breast Imaging-Reporting and Data System.	

Evaluation of BI-RADS Classification of LLMs Across Different Prompts

Figure 2 illustrates the performance of the three LLMs in BI-RADS classification across different prompts. In Prompt III, GPT-4o demonstrated near-perfect agreement (κ , 0.80), outperforming both

GPT-o3 mini (0.76) and Qwen (0.79). Notably, the kw of GPT-4o in Prompt III was significantly higher than that observed in Prompt II (0.69, $P < 0.05$) and Prompt I (0.63, $P < 0.01$).

Figure 3 displays the confusion matrices of LLMs in BI-RADS classification across different prompts. With Prompt II (which incorporated an enhanced ambiguity case escalation mechanism building upon Prompt I), all LLMs significantly reduced the misclassification of BI-RADS 5 to 4. Specifically, GPT-4o, Qwen-2.5 max, and GPT-o3 mini exhibited a decrease in misclassified cases of 54% (28 to 13), 50% (12 to 6), and 68% (37 to 12), respectively. However, the improvement in the classification of BI-RADS 3 errors into 2 and BI-RADS 4 errors into 3 or 2 is not significant.

With Prompt III (incorporating a low-malignancy probability lesion attention mechanism based on Prompt II), all LLMs not only reduced misclassifications of BI-RADS 5 to 4, but also improved those of BI-RADS from 3 to 2. Specifically, GPT-4o, Qwen-2.5 max, and GPT-o3 mini experienced a reduction in such misclassifications of 61% (from 43 to 17), 65% (from 43 to 15), and (from 46 to 17), respectively. In addition, the misclassification of BI-RADS 4 to 2 was nearly eliminated (GPT-4o: from 11 to 0; Qwen-2.5 max: from 10 to 0, and GPT-o3: mini from 9 to 1).

For each prompt, there were no significant differences in the kw levels between the initial BI-RADS classifications and the BI-RADS classifications performed 7 days later for the three LLMs. The kw were nearly perfect with Prompt III (GPT-4o, 0.84; GPT-o3 mini, 0.85; Qwen-2.5 max, 0.90) (**Table S2**). Prompt II showed substantial to nearly perfect kw values (0.74–0.84), while Prompt I exhibited moderate to nearly perfect agreement (0.68–0.83).

Evaluation BI-RADS Classification of LLMs-Human and Human-Human

Figure 4 illustrates the performance of LLMs and six physicians with varying levels of experience in BI-RADS classification. The kw values of the two mid-level radiologists (0.90 and 0.86) were significantly higher than the performance of the LLMs under Prompt III. GPT-4o, when utilizing Prompt III (0.80), achieved a high kw, slightly surpassing entry-level radiologist 1 (0.76), and was comparable to entry-level radiologist 2 (0.79), while being notably superior to the two resident radiologists (0.54 and 0.61).

LLMs-Human and Human-Human Agreement according to Clinical Management

Table 2–3 and **Table S3** summarize the performance of various LLMs across different prompts regarding clinical management categories. The proportions of changed clinical management for the LLMs were comparable (with GPT-4o demonstrating the best overall performance). However, the proportion of changed clinical management in Prompt III was significantly lower than that of Prompts I and II. Table 4 summarizes the performance of six human readers in clinical management categories, with the top best performers being resident doctor 2, entry-level doctor 2, and mid-level doctor 1). The reference standards included cases requiring no treatment (BI-RADS 2, n = 54), follow-up in 6 months (BI-RADS 3, n = 54), and

biopsy or aspiration (BI-RADS 4 or 5, n = 108). Figure 5 illustrates the potential changes in clinical management between LLMs and human interpretations.

Table 2

Agreement between LLMs in Prompt I and the reference standard on clinical management reports.

Outcome	Standard-GPT-4o	Standard-GPT-o3	Standard-Qwen
Weighted Cohen’s kappa (95% CIs)	0.63 (0.57, 0.69)	0.60 (0.54, 0.66)	0.65 (0.59, 0.71)
Changes in management	64/216 (29.6%)	65/216 (30.1%)	69/216 (31.9%)
Upgraded	3/216 (1.4%)	2/216 (0.9%)	5/216 (2.3%)
From BI-RADS 2 to BI-RADS 3	2/3 (66.7%)	1/2 (50.0%)	2/5 (40.0%)
From BI-RADS 2/3 to BI-RADS 4/5	1/3 (33.3%)	1/2 (50.0%)	3/5 (60.0%)
Downgraded	61/216 (28.2%)	63/216 (29.2%)	64/216 (29.6%)
From BI-RADS 3 to BI-RADS 2	43/61 (70.5%)	46/63 (73.0%)	43/64 (67.2%)
From BI-RADS 4/5 to BI-RADS 2	11/61 (18.0%)	9/63 (14.3%)	10/64 (15.6%)
From BI-RADS 4/5 to BI-RADS 3	7/61 (11.5%)	8/63 (12.7%)	11/64 (17.2%)
Data are expressed as a numerator and a denominator with the percentage in parentheses.			
LLM = large language model. GPT-o3 = GPT-o3 mini. Qwen = Qwen-2.5 max. CIs = confidence intervals.			
BI-RADS = Breast Imaging Reporting and Data System.			

Table 3

Agreement between LLMs in Prompt III and the reference standard on clinical management reports.

Outcome	Standard-GPT-4o	Standard-GPT-o3	Standard-Qwen
Weighted Cohen's kappa (95% CIs)	0.80 (0.75, 0.85)	0.76 (0.71, 0.82)	0.79 (0.74, 0.84)
Changes in management	32/216 (14.8%)	41/216 (19.0%)	35/216 (16.2%)
Upgraded	2/216 (0.9%)	2/216 (0.9%)	7/216 (3.2%)
From BI-RADS 2 to BI-RADS 3	2/2 (100%)	2/2 (100%)	6/7 (85.7%)
From BI-RADS 2/3 to BI-RADS 4/5	0/2 (0%)	0/2 (0%)	1/7 (14.3%)
Downgraded	30/216 (13.9%)	39/216 (18.1%)	28/216 (13.0%)
From BI-RADS 3 to BI-RADS 2	17/30 (56.7%)	17/39 (43.6%)	15/28 (53.6%)
From BI-RADS 4/5 to BI-RADS 2	0/30 (0%)	1/39 (2.6%)	0/28 (0%)
From BI-RADS 4/5 to BI-RADS 3	13/30 (43.3%)	21/39 (53.8%)	13/28 (46.4%)
Data are expressed as a numerator and a denominator with the percentage in parentheses.			
LLM = large language model. GPT-o3 = GPT-o3 mini. Qwen = Qwen-2.5 max. CIs = confidence intervals.			
BI-RADS = Breast Imaging Reporting and Data System.			

Table 4

Agreement between human readers and the reference standard on clinical management reports.

Outcome	Standard-Resident 1	Standard-Resident 2	Standard-Entry-level 1	Standard-Entry-level 2	Standard-Mid-level 1	Standard-Mid-level 2
Weighted Cohen's kappa (95% CIs)	0.54 (0.47, 0.61)	0.61 (0.54, 0.67)	0.76 (0.71, 0.81)	0.79 (0.74, 0.84)	0.90 (0.86, 0.93)	0.86 (0.82, 0.90)
Changes in management	76/216 (35.2%)	55/216 (25.5%)	33/216 (15.3%)	31/216 (14.4%)	11/216 (5.1%)	23/216 (10.6%)
Upgraded	9/216 (4.2%)	9/216 (4.2%)	9/216 (4.2%)	8/216 (3.7%)	3/216 (1.4%)	6/216 (2.8%)
From BI-RADS 2 to BI-RADS 3	8/9 (88.9%)	5/9 (55.6%)	7/9 (77.8%)	6/8 (75.0%)	3/3 (100%)	5/6 (83.3%)
From BI-RADS 2/3 to BI-RADS 4/5	1/9 (11.1%)	4/9 (44.4%)	2/9 (22.2%)	2/8 (25.0%)	0/3 (0%)	1/6 (16.7%)
Downgraded	67/216 (31.0%)	46/216 (21.3%)	24/216 (11.1%)	23/216 (10.6%)	8/216 (3.7%)	17/216 (7.9%)
From BI-RADS 3 to BI-RADS 2	15/67 (22.4%)	13/46 (28.3%)	10/24 (41.7%)	11/23 (47.8%)	7/8 (87.5%)	11/17 (64.7%)
From BI-RADS 4/5 to BI-RADS 2	1/67 (1.5%)	0/46 (0%)	0/24 (0%)	0/23 (0%)	0/8 (0%)	0/17 (0%)
From BI-RADS 4/5 to BI-RADS 3	51/67 (76.1%)	33/46 (71.7%)	14/24 (58.3%)	12/23 (52.2%)	1/8 (12.5%)	6/17 (35.3%)
Data are expressed as a numerator and a denominator with the percentage in parentheses.						
Entry-level = Entry-level doctor; Mid-level = Mid-level doctor. CIs = confidence intervals. BI-RADS = Breast Imaging Reporting and Data System						

When the BI-RADS classifications from both the LLM and the reference standard lead to the same clinical decision (e.g., both BI-RADS 4/5 recommend biopsy), no change in management is required. For underclassification cases, where BI-RADS categories were inappropriately downgraded (e.g., from BI-RADS 4 or 5 to follow-up categories, or from BI-RADS 3, 4, 5 to no intervention), the underclassification rate for GPT-4o in Prompt III was 13.9% (n = 30/216), which was significantly lower than for Prompt I (28.2% [n = 61/216]) and Prompt II (27.8% [n = 60/216]). Entry-level radiologist 2 and mid-level radiologist 1 showed significantly lower underclassification rates than GPT-4o in all prompts (10.6% [n = 23/216] and 3.7% [n = 8/216], respectively), while resident 2 had a higher underclassification rate (21.3% [n = 46/216]) than GPT-4o in Prompt III. The underclassification by GPT-4o under Prompt I and Prompt II mainly involved downgrading from follow-up to no intervention (70.5% [n = 43/61] and 80.0% [n = 48/60],

respectively), whereas resident 1 primarily showed downgrading from biopsy-required categories to follow-up (76.1% [n = 51/67]). Conversely, for overclassification cases, where categories were inappropriately upgraded (e.g., from BI-RADS 2 to follow-up, or from BI-RADS 2, 3 to biopsy-required categories), the overclassification rate for GPT-4o in Prompt III was 0.9%, which was comparable to Prompt I (1.4%) and Prompt II (0.5%). Notably, all these overclassification rates were lower than those observed for the three human readers (4.2%, 3.7%, and 1.4%, respectively).

Discussion

This study represents the first systematic evaluation of LLM performance in BI-RADS classification within DBT free-text imaging reports through prompt engineering. Comparative analysis with expert radiologists revealed that strategic prompting significantly improved the agreement of BI-RADS classification evaluation in LLMs. These findings highlight the potential of prompt engineering to enhance LLM performance in supporting BI-RADS classification, indicating that strategic prompt design may be a crucial tool for improving consistency and accuracy in breast imaging interpretation.

Recent advancements in LLMs have demonstrated significant progress in the medical field. For instance, in medical imaging, LLMs have demonstrated capabilities in assisting clinicians with the interpretation of radiological, pathological, and electrocardiographic images, thereby enhancing both efficiency and accuracy in diagnostic workflows (23–26). Furthermore, LLMs exhibit substantial potential in drug discovery, facilitating the identification of novel therapeutic targets and accelerating the development of new pharmaceutical agents (27, 28). Beyond data extraction from free-text reports (29), LLMs can also perform tumor staging and grading through unstructured text analysis, such as leveraging ChatGPT to stage lung cancer based on free-text imaging reports (30). Despite these promising advancements, particularly in managing complex medical tasks, critical challenges, including instability, opacity, and inaccuracy in LLM outputs, remain substantial barriers to their clinical implementation (31). Besides, current research underscores the necessity of benchmarking LLM performance against human expertise across specialized domains as more advanced models are introduced.

Compared to prior studies, the GPT-4o model in this study demonstrated near-perfect agreement with senior radiologists (GPT-4o, kw = 0.80), significantly surpassing the moderate agreement reported in earlier research (GPT-4, Gwet AC1 = 0.52). Furthermore, the proportion of clinical management changes induced by GPT-4o (14.8%) was notably lower than that observed in previous investigations using GPT-4 (18.1%) (5). The performance of GPT-4o exceeded that of residents and yielded comparable results to those of entry-level radiologists with limited experience. The advancements in this study may be attributed to three primary innovations. First, the dual-principle prompt III design incorporated rule-based constraints aligned with American College of Radiology BI-RADS guidelines, reducing BI-RADS 5 to 4 misclassification rates by 53.6–67.6% across models through conservative upgrade principles (4). Besides, a low-heterogeneity attention mechanism mandated cross-validation of subtle imaging features, decreasing BI-RADS 3 to 2 misclassifications by 60.5–65.1% and eliminating 4 to 2 errors in GPT-4o and Qwen-2.5 max. This mechanism specifically addresses confirmation bias in the

interpretation of low-risk lesions, a known cognitive pitfall in diagnostic decision-making. Second, GPT-4o's architectural and training data optimizations enhanced its capacity to process complex linguistic tasks inherent to radiology reporting, outperforming earlier models. Third, restricting BI-RADS categorization outputs to categories 2 to 5 allowed the model to focus on ambiguous cases, improving classification accuracy.

Notably, while this study utilized the kw to assess agreement, prior research relied on Gwet's AC1 statistic. Although both metrics evaluate classification consistency, kw imposes penalties for ordinal misclassifications (e.g., adjacent category errors), enabling more comprehensive performance evaluation. This distinction likely contributes to the higher agreement observed in our study.

Notably, the proportion of clinical category management changes under GPT-4o with Prompt III was 14.8%, which was comparable to that of residents, yet significantly higher than that of mid-level radiologists. This suggests that, as an assistive tool, LLMs can facilitate rapid onboarding for novice practitioners and hold potential for reducing workload among entry-level radiologists, while also highlighting the necessity for human oversight in high-risk decision-making scenarios.

With the global proliferation and immediate availability of LLMs, particularly in the context of the current shortage of healthcare professionals, patients may upload imaging reports (only findings) to LLMs for interpretation. The process employed in this study, involving zero-shot training and no exposure of LLMs to images, closely simulates this behavior. Such scenarios could raise significant ethical and legal concerns, particularly if patients rely solely on LLM-generated results, which could potentially lead to misdiagnosis or treatment delays. Therefore, to prevent such outcomes, educating patients on the proper use of LLMs for interpreting imaging reports and critically evaluating the generated results is becoming increasingly important.

This study has several limitations that should be acknowledged. First, the sample was derived from a single institution and was conducted exclusively in Chinese, which limits the generalizability of the findings to some extent. Future studies should incorporate diverse languages and institutions to validate the applicability and performance of LLMs in varied contexts. Second, while prompt engineering improved LLM performance in classification consistency, this study did not compare LLMs with other artificial intelligence models, such as convolutional neural networks. Furthermore, this study primarily focused on the analysis of classification accuracy while neglecting factors critical in actual clinical settings, such as real-time performance and interpretability. Future research should place greater emphasis on optimizing LLM real-time feedback capabilities and addressing their transparency issues in clinical applications to enhance their operability and acceptance among clinicians.

In conclusion, the current study results indicate that although optimizing prompts can enhance the performance of LLMs in BI-RADS classification to some extent, their accuracy remains inferior to that of experienced clinicians. Given the critical role of BI-RADS classification in breast cancer diagnosis and management, LLMs optimized through prompt engineering are not yet capable of fully replacing the expertise of trained healthcare professionals. Therefore, future research should focus on developing a

multimodal framework driven by LLMs (e.g., integrating image feature vectors as input to the LLM and modeling them through joint visual-text encoding) to facilitate deeper collaboration between artificial intelligence and clinical expertise, ultimately enhancing healthcare quality and improving patient outcomes.

Declarations

Ethics approval

The studies involving human participants were reviewed and approved by the Institutional Review Board of the Second Affiliated Hospital of Nanchang University. Since data were evaluated retrospectively, pseudonymously and were solely obtained for treatment purposes, a requirement of informed consent was waived by the Institutional Review Board. All authors have confirmed that any experiments involving humans and/or the use of human tissue samples were conducted in accordance with relevant guidelines and regulations.

Consent for publication

Not applicable.

Availability of data and materials

The datasets used and analysed during the current study are available from the corresponding author on reasonable request.

Competing Interests

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Funding

No funding.

Author Contributions

Conception and design: W.L. and Y.L. Administrative support: L.G. Collection and assembly of data: W.L. Data analysis and interpretation: W.L. Partial human readers: Y.L., Y.L., H.W., Y.L., W.L., X.L. Manuscript writing: W.L., Y.L., and L.G. Final approval of manuscript: L.G. All authors contributed to the article and approved the submitted version.

Acknowledgements

Not applicable.

References

1. Timmers JM, van Doorne-Nagtegaal HJ, Zonderland HM, et al. The Breast Imaging Reporting and Data System (BI-RADS) in the Dutch breast cancer screening programme: its role as an assessment and stratification tool. *Eur Radiol* 2012;22(8):1717-1723. doi: 10.1007/s00330-012-2409-2
2. Spak DA, Plaxco JS, Santiago L, Dryden MJ, Dogan BE. BI-RADS((R)) fifth edition: A summary of changes. *Diagn Interv Imaging* 2017;98(3):179-190. doi: 10.1016/j.diii.2017.01.001
3. Ekpo EU, Ujong UP, Mello-Thoms C, McEntee MF. Assessment of Interradiologist Agreement Regarding Mammographic Breast Density Classification Using the Fifth Edition of the BI-RADS Atlas. *AJR Am J Roentgenol* 2016;206(5):1119-1123. doi: 10.2214/AJR.15.15049
4. American College of R. ACR BI-RADS atlas : breast imaging reporting and data system. Fifth edition ed. Reston, VA: American College of Radiology, 2013.
5. Cozzi A, Pinker K, Hidber A, et al. BI-RADS Category Assignments by GPT-3.5, GPT-4, and Google Bard: A Multilanguage Study. *Radiology* 2024;311(1):e232133. doi: 10.1148/radiol.232133
6. Haver HL, Yi PH, Jeudy J, Bahl M. Use of ChatGPT to Assign BI-RADS Assessment Categories to Breast Imaging Reports. *AJR Am J Roentgenol* 2024;223(3):e2431093. doi: 10.2214/AJR.24.31093
7. Yang J, Jin H, Tang R, et al. Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond. *ACM Transactions on Knowledge Discovery from Data* 2023;18:1-32. doi: 10.1145/3649506
8. Fan L, Li L, Zihui, Lee S, Yu H, Hemphill L. A Bibliometric Review of Large Language Models Research from 2017 to 2023. *ACM Transactions on Intelligent Systems and Technology* 2023. doi: 10.48550/arXiv.2304.02020
9. Will ChatGPT transform healthcare? *Nature Medicine* 2023;29(3):505-506. doi: 10.1038/s41591-023-02289-5
10. Mukherjee P, Hou B, Lanfredi RB, Summers RM. Feasibility of Using the Privacy-preserving Large Language Model Vicuna for Labeling Radiology Reports. *Radiology* 2023;309(1):e231147. doi: 10.1148/radiol.231147
11. Adams LC, Truhn D, Busch F, et al. 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
12. Liu J, Wang C, Liu S. Utility of ChatGPT in Clinical Practice. *J Med Internet Res* 2023;25:e48568. doi: 10.2196/48568
13. Yeo YH, Samaan JS, Ng WH, et al. Assessing the performance of ChatGPT in answering questions regarding cirrhosis and hepatocellular carcinoma. *Clin Mol Hepatol* 2023;29(3):721-732. doi: 10.3350/cmh.2023.0089
14. Rao A, Pang M, Kim J, et al. Assessing the Utility of ChatGPT Throughout the Entire Clinical Workflow: Development and Usability Study. *J Med Internet Res* 2023;25:e48659. doi: 10.2196/48659

15. Ozenbas C, Engin D, Altinok T, Akcay E, Aktas U, Tabanlı A. ChatGPT-4o's Performance in Brain Tumor Diagnosis and MRI Findings: A Comparative Analysis with Radiologists. *Acad Radiol* 2025. doi: 10.1016/j.acra.2025.01.033
16. Fatima A, Shafique MA, Alam K, Fadlalla Ahmed TK, Mustafa MS. ChatGPT in medicine: A cross-disciplinary systematic review of ChatGPT's (artificial intelligence) role in research, clinical practice, education, and patient interaction. *Medicine (Baltimore)* 2024;103(32):e39250. doi: 10.1097/MD.00000000000039250
17. Gallifant J, Fiske A, Levites Strekalova YA, et al. Peer review of GPT-4 technical report and systems card. *PLOS Digit Health* 2024;3(1):e0000417. doi: 10.1371/journal.pdig.0000417
18. Liu C, Wei M, Qin Y, et al. Harnessing Large Language Models for Structured Reporting in Breast Ultrasound: A Comparative Study of Open AI (GPT-4.0) and Microsoft Bing (GPT-4). *Ultrasound Med Biol* 2024;50(11):1697-1703. doi: 10.1016/j.ultrasmedbio.2024.07.007
19. Afshar M, Gao Y, Wills G, et al. Prompt engineering with a large language model to assist providers in responding to patient inquiries: a real-time implementation in the electronic health record. *JAMIA Open* 2024;7(3):ooae080. doi: 10.1093/jamiaopen/ooae080
20. Warren CJ, Edmonds VS, Payne NG, et al. Prompt matters: evaluation of large language model chatbot responses related to Peyronie's disease. *Sex Med* 2024;12(4):qfae055. doi: 10.1093/sexmed/qfae055
21. McHugh ML. Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 2012;22(3):276-282.
22. Demšar J. Statistical Comparisons of Classifiers over Multiple Data Sets. *J Mach Learn Res* 2006;7:1–30.
23. Zhao Z, Wang S, Gu J, et al. ChatCAD+: Toward a Universal and Reliable Interactive CAD Using LLMs. *IEEE Trans Med Imaging* 2024;43(11):3755-3766. doi: 10.1109/TMI.2024.3398350
24. Tian Y, Li Z, Jin Y, et al. Foundation model of ECG diagnosis: Diagnostics and explanations of any form and rhythm on ECG. *Cell Rep Med* 2024;5(12):101875. doi: 10.1016/j.xcrm.2024.101875
25. Wu SH, Tong WJ, Li MD, et al. Collaborative Enhancement of Consistency and Accuracy in US Diagnosis of Thyroid Nodules Using Large Language Models. *Radiology* 2024;310(3):e232255. doi: 10.1148/radiol.232255
26. Waqas A, Bui MM, Glassy EF, et al. Revolutionizing Digital Pathology With the Power of Generative Artificial Intelligence and Foundation Models. *Lab Invest* 2023;103(11):100255. doi: 10.1016/j.labinv.2023.100255
27. Tripathi S, Gabriel K, Tripathi PK, Kim E. Large language models reshaping molecular biology and drug development. *Chem Biol Drug Des* 2024;103(6):e14568. doi: 10.1111/cbdd.14568
28. Chakraborty C, Bhattacharya M, Lee SS. Artificial intelligence enabled ChatGPT and large language models in drug target discovery, drug discovery, and development. *Mol Ther Nucleic Acids* 2023;33:866-868. doi: 10.1016/j.omtn.2023.08.009
29. Le Guellec B, Lefevre A, Geay C, et al. Performance of an Open-Source Large Language Model in Extracting Information from Free-Text Radiology Reports. *Radiol Artif Intell* 2024;6(4):e230364. doi:

30. Lee JE, Park KS, Kim YH, Song HC, Park B, Jeong YJ. Lung Cancer Staging Using Chest CT and FDG PET/CT Free-Text Reports: Comparison Among Three ChatGPT Large Language Models and Six Human Readers of Varying Experience. AJR Am J Roentgenol 2024;223(6):e2431696. doi: 10.2214/AJR.24.31696
31. Ong JCL, Seng BJJ, Law JZF, et al. Artificial intelligence, ChatGPT, and other large language models for social determinants of health: Current state and future directions. Cell Rep Med 2024;5(1):101356. doi: 10.1016/j.xcrm.2023.101356

Figures

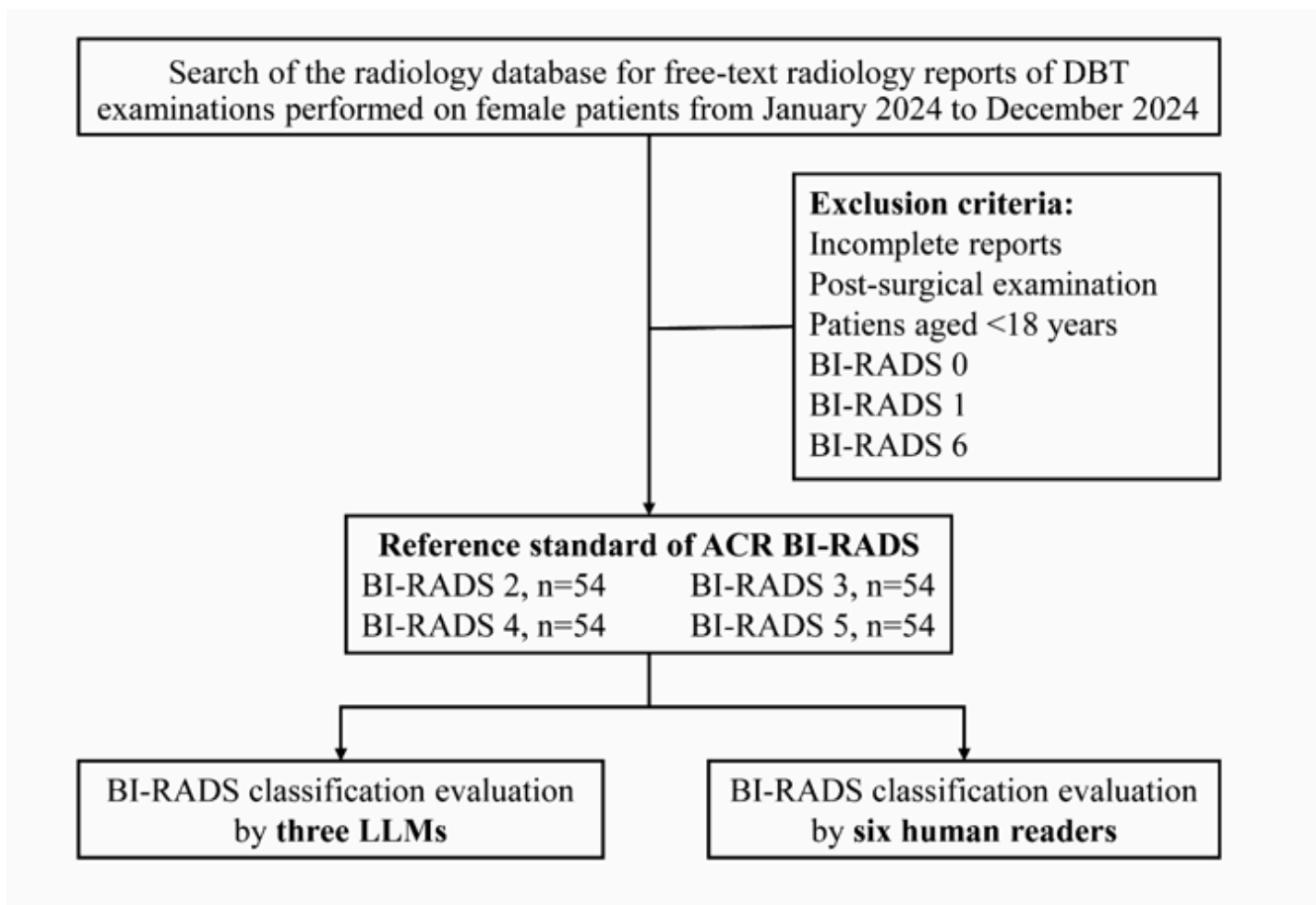


Figure 1

Flowchart depicting the patient selection process.

DBT= digital breast tomosynthesis, BI-RADS= Breast Imaging-Reporting and Data System, ACR= American College of Radiology, LLM= large language model.

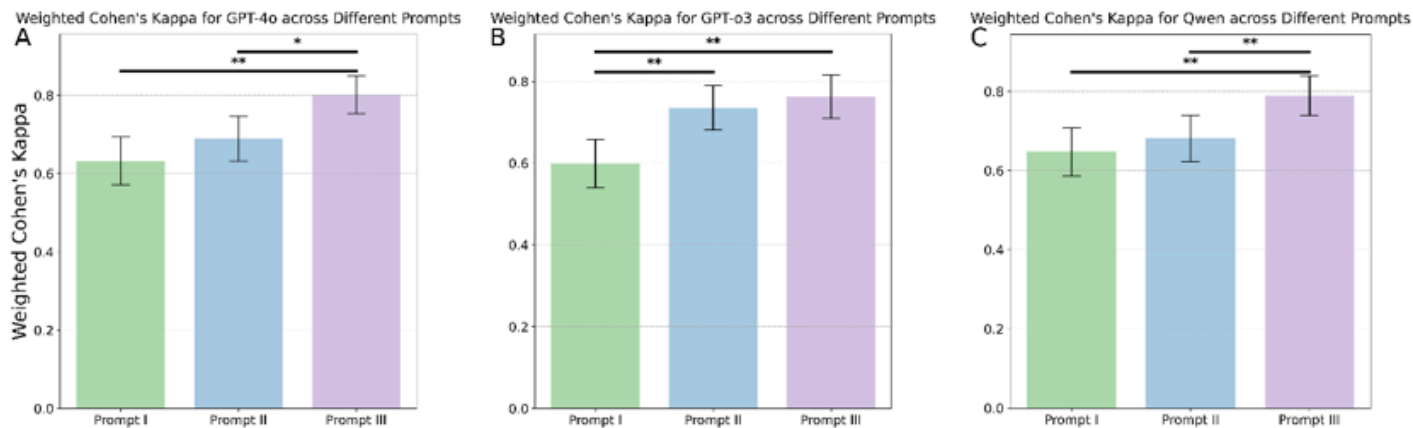


Figure 2

Weighted Cohen's kappa of large language models with different prompts. (A) GPT-4o. (B) GPT-o3 mini. (C) Qwen-2.5 max.

* $P < 0.05$ calculated as described by the Friedman test for the comparison of LLM-LLM κ_w values. ** $P < 0.01$ calculated as described by the Friedman test for the comparison of LLM-LLM κ_w values.

LLM= large language model.

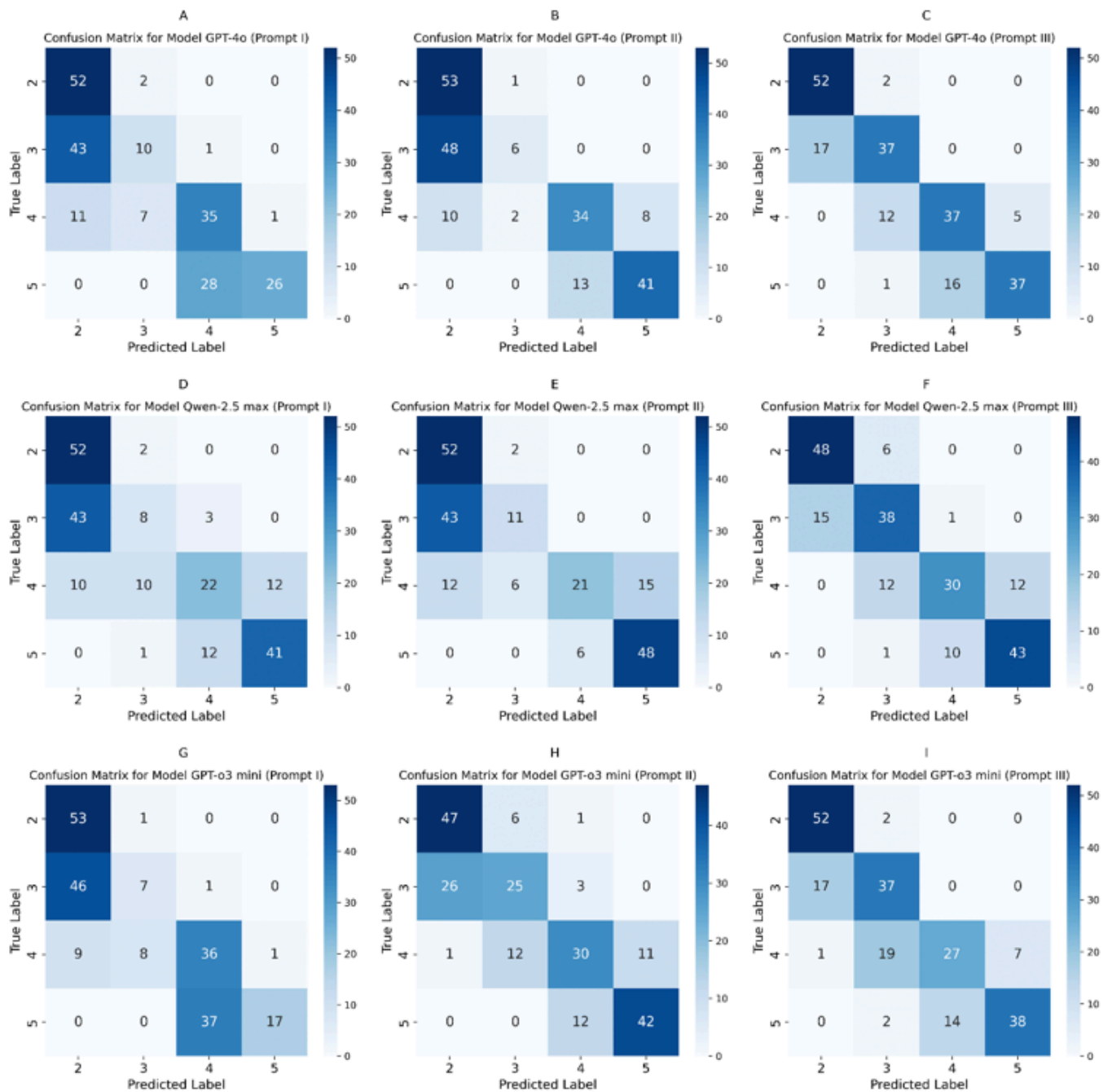


Figure 3

Confusion matrix of the large language models with different prompts. (A-C) Confusion matrix of the GPT-4o with different prompts. (D-F) Confusion matrix of the GPT-o3 with different prompts. (G-I) Confusion matrix of the Qwen-2.5 max with different prompts.

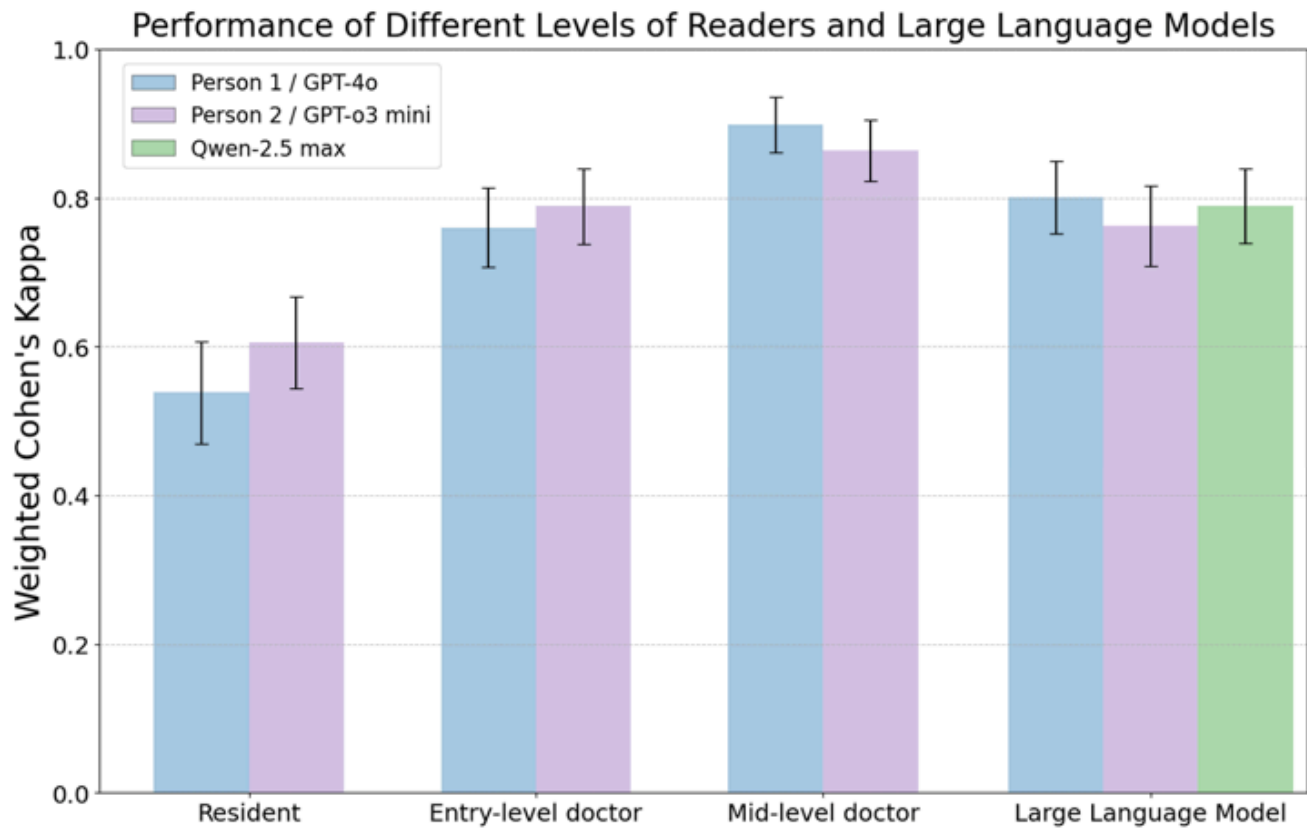


Figure 4

Weighted Cohen’s kappa of human readers with different levels of experience.

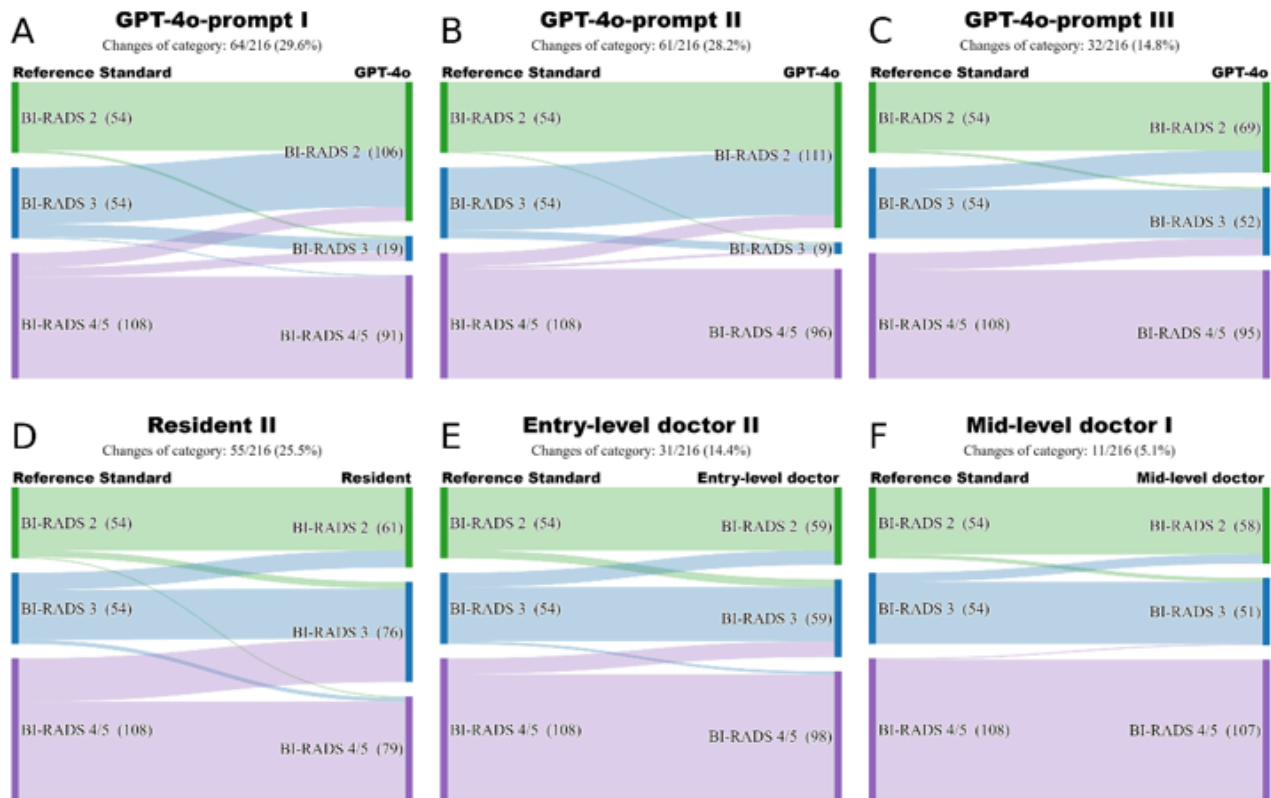


Figure 5

Sankey diagrams for GPT-4o and human readers with different levels of experience. (A-C) Sankey diagrams for GPT-4o with different prompts. (D) Sankey diagrams for resident doctors. (E) Sankey diagrams for entry-level doctors. (F) Sankey diagrams for mid-level doctors.

BI-RADS= Breast Imaging-Reporting and Data System.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementalmaterials.docx](#)