Journal of Magnetic Resonance Imaging

REVIEW

The Role of Large Language Models (LLMs) in Breast Imaging Today and in the Near Future

Simone Schiaffino[1,2] | Tianyu Zhang[3,4,5] | Ritse M. Mann[3,4] | Katja Pinker[6]

[1]Imaging Institute of Southern Switzerland (IIMSI), Ente Ospedaliero Cantonale (EOC), Lugano, Switzerland | [2]Faculty of Biomedical Sciences, Università della Svizzera Italiana, Lugano, Switzerland | [3]Department of Radiology, Netherlands Cancer Institute, Amsterdam, the Netherlands | [4]Department of Diagnostic Imaging, Radboud University Medical Center, Nijmegen, the Netherlands | [5]GROW School for Oncology and Development Biology, Maastricht University, Maastricht, the Netherlands | [6]Department of Radiology, Columbia University, Vagelos College of Physicians and Surgeons, New York, New York, USA

Correspondence: Katja Pinker (kp3124@cumc.columbia.edu)

Received: 21 January 2025 | **Revised:** 14 April 2025 | **Accepted:** 15 April 2025

Keywords: breast neoplasms | large language models | magnetic resonance imaging | radiology

ABSTRACT

This narrative review focuses on the integration of large language models (LLMs), such as GPT-4 and Gemini, into breast imaging. LLMs excel in understanding, processing, and generating human-like text, with potential applications ranging widely from decision-making to radiology reporting support. LLMs show promise in addressing current critical challenges, including rising demands for imaging services concurrent with an increasing shortage in the radiologist workforce. Their ability to integrate clinical guidelines and generate standardized, evidence-based reports has the potential to improve diagnostic consistency and reduce inter-reader variability. Emerging multimodal capabilities further extend their utility, enabling the integration of textual and visual data for tasks such as tumor classification and decision-making. Despite these advancements, significant challenges remain. LLMs often suffer from limitations such as hallucinations, biases in training datasets, and domain-specific knowledge gaps. These issues can affect their reliability, particularly in nuanced tasks like Breast Imaging Reporting and Data System categorization and multimodal image assessment. Moreover, ethical concerns about data privacy, biased outputs, and regulatory compliance must be addressed before effective deployment in the clinical setting. Current studies suggest that while LLMs can complement human expertise, their performance still lags behind that of radiologists in key areas, particularly in tasks requiring complex medical reasoning or direct image analysis. Looking ahead, LLMs are poised to play a crucial role in breast imaging by optimizing workflows, supporting multidisciplinary meetings, and improving patient education. However, their successful integration will depend on proper context training, robust validation, and ethical oversight, with human supervision as a crucial safeguard.

Evidence Level: 5.

Technical Efficacy: Stage 2.

Simone Schiaffino and Tianyu Zhang are co-first authors.

Ritse M. Mann and Katja Pinker are co-senior authors.

https://doi.org/10.1002/jmri.29807

1 | Introduction

Recent advancements in artificial intelligence (AI) are impacting numerous domains, including healthcare, with radiology as a particularly fertile ground for innovation [1]. Among the available AI technologies, large language models (LLMs) have demonstrated exceptional skills in understanding, processing, and generating human-like text, with their development advancing at an unprecedented pace [2]. Indeed, since their inception, LLMs have grown exponentially in both scale and function, fueled by advancements in deep learning architectures, enhanced computational capabilities, and the availability of vast datasets [3]. Such rapid growth has enabled them to evolve from general-purpose text generators into specialized tools capable of addressing complex challenges in domains such as breast imaging [4, 5]. In breast imaging, LLMs show promise for addressing critical challenges, both in the clinical setting (where there is growing demand for imaging services amidst a persistent shortage of radiologists) and in the research setting [4, 6].

LLMs, such as GPT-4 (OpenAI), Gemini (Google), Claude (Anthropic), LLaVA (Meta), and their contemporaries, exhibit capabilities that extend far beyond basic text generation. They excel at understanding context, processing large volumes of information, and producing highly structured outputs [7, 8]. In breast imaging, their potential applications span almost every stage of the clinical workflow, from decision-making support to the assessment and interpretation of breast imaging examinations [9, 10]. They can enhance the clarity and standardization of breast imaging reports, generate patient-specific recommendations, and synthesize complex findings from multimodal imaging studies [4, 11]. Moreover, by integrating clinical guidelines and evidence-based protocols directly into reporting workflows, they have the potential to reduce inter-reader variability, thereby improving diagnostic consistency [10, 12].

Despite their transformative potential, integrating LLMs into breast imaging presents significant challenges [10, 13]. Common limitations of LLMs—such as hallucinations, lack of in-context training, and domain-specific knowledge gaps—can affect their reliability [14]. Additionally, biases in training datasets, ethical concerns regarding AI deployment in clinical settings, and stringent requirements for data privacy and regulatory compliance must be addressed carefully [15–18].

This narrative review provides an overview of the presently investigated applications of LLMs in breast imaging, highlighting their transformative potential while reviewing the challenges that must be addressed for successful integration (Table 1). The text follows the patient's journey in the context of breast imaging, from decision-making before performing the exam, to the direct evaluation of the acquired images, to support in drafting the imaging report, and to the multidisciplinary discussion of the case at the end of the imaging assessment. A crucial note, particularly given the rapid evolution of these models, is the reliance of this article on published peer-reviewed articles, with preprints cited only in specific instances. The performance of LLMs as reported in the cited

TABLE 1 | Key findings from the literature regarding LLM applications in breast imaging.

Application	Key findings
Decision-making support	GPT-4 outperformed GPT-3.5 in breast cancer screening (98.4% vs. 88.9%) and breast pain (77.7% vs. 58.3%) cases, reducing over recommendations [19] GPT-4 and Glass AI showed promise in recommending imaging examinations for complex scenarios; Glass AI, which incorporates context-specific training, showed better results than GPT-4 [20]
Multimodal Analysis	GPT-4V and Gemini Pro Vision achieved modest accuracy (41%–49% and 29%–39% respectively) in interpreting Diagnosis Please cases published in the Radiology journal, underperforming compared with radiologists (61%) [21] GPT-4V performed better on text-based (81.5%) than image-based questions (47.8%) from ACR Diagnostic Radiology In-Training Examination questions, highlighting limitations in visual data interpretation [22] Combining LLMs with AI models like CNNs shows promise, achieving up to 92% accuracy in breast cancer classification [23]
Radiology Reporting	GPT-4 achieved 73.6% overall accuracy in assigning BI-RADS categories across factitious breast imaging reports; accuracy was highest for mammography reports (92%) and lowest for MRI reports (58%) [20] GPT-3.5, GPT-4 and Google Bard showed moderate agreement and frequently produced discordant BI-RADS assignments to human readers, potentially leading to clinically detrimental management changes [4] The TECRR dataset provided benchmark results for the use of LLMs in BI-RADS classification and highlighted the need for complex, multilingual, and modality datasets [24]

(Continues)

TABLE 1 | (Continued)

Application	Key findings
Multidisciplinary Meeting Decision Support	GPT-4 outperformed GPT-3.5 and Claude2 in tumor board decision-making tasks but showed limitations, such as inaccuracies in following nuanced clinical guidelines and a lack of adaptation to regional medical practices [25] Partial concordance between GPT models and tumor board recommendations has been found (e.g., 50%–70% concordance based on different studies) [10, 26] Specialized, continuously updated LLMs like Med-PaLM, LLaVA Med, and Glass AI are being investigated for enhancing tumor board efficiency and decision-making accuracy [27–29]

Abbreviations: AI, artificial intelligence; BI-RADS, Breast Imaging Reporting and Data System; CNN, convolutional neural network; LLM, large language model; MRI, magnetic resonance imaging.

peer-reviewed articles are reflective of the models' capabilities at the time of evaluation, underscoring the importance of contextualizing findings in a field where LLMs, such as GPT-4, are updated in near real-time. This dynamic nature necessitates a forward-thinking approach to interpret and apply published results in clinical practice.

2 | Decision-Making Support

Low-value imaging, defined as diagnostic imaging examinations providing little-to-no impact on patient management, accounts for an estimated 20%–50% of general imaging examinations, with rates varying between and within different countries [30]. In breast imaging, this is particularly evident in the management of benign breast tumors (e.g., follow-up and biopsy) and in the follow-up of patients with breast cancer [30]. To address the resource misuse associated with low-value imaging, guidelines such as the American College of Radiology (ACR) Appropriateness Criteria [31] have been promoted. These guidelines assist clinicians in selecting the most appropriate imaging or treatment option for specific clinical conditions, thereby improving patient care while minimizing unnecessary procedures. However, their reported impact remains limited [32].

Patients, referring physicians, radiologists, and schedulers could benefit from decision-making support tools. By triaging patients and recommending appropriate imaging examinations—whether screening, diagnostic, or advanced-level examinations—for specific clinical scenarios, decision-making tools can enhance resource allocation, avoid mis-scheduling, reduce unnecessary examinations, and improve the overall patient experience.

GPT-3.5 and GPT-4 as decision-making support tools have been investigated by Rao et al. [19] to recommend appropriate imaging examinations for two common clinical indications: breast cancer screening and breast pain, with the authors comparing the models' recommendations with that provided my the ACR Appropriateness Criteria. Two prompt formats were used, open-ended and "select all that apply" (SATA). With the open-ended format, the model was asked to identify "the single most appropriate imaging procedure" for a specific clinical scenario (e.g., "the single most appropriate imaging procedure" in average risk women); with the SATA format, a list of imaging modalities was provided to the model and it was asked about the level of appropriateness of each of them. As expected, GPT-4 outperformed GPT-3.5, with high accuracy for breast cancer screening SATA prompts (98.4% vs. 88.9%) and moderate accuracy for breast pain SATA prompts (77.7% vs. 58.3%). Notably, GPT-3.5 demonstrated a tendency to over-recommend imaging examinations, while GPT-4 showed improvements in reducing unnecessary imaging examinations. Further refinement and evaluation are necessary before broad clinical implementation, according to the authors.

Breast clinical scenarios were included among the 1075 ones from 11 ACR expert panels in a study by Zaki et al. [27] to explore the ability of GPT-4 and Glass AI (a GPT-4-based model fine-tuned on medical texts) to recommend the most appropriate imaging examination for a specific clinical scenario. The responses were compared with the ACR Appropriateness Criteria, and each output was scored from 1 to 3 based on the criteria. Moreover, Pearson's *R* correlation coefficient between GPT-4 and Glass AI scores was calculated across the topics for a given panel. The results suggest that both models show promise as decision-making support tools for recommending imaging examinations in complex clinical settings, especially Glass AI which incorporates context-specific training. A moderately positive agreement between the two models was observed.

These initial investigations highlight the potential of LLMs to indirectly assist radiologists by providing referring physicians with decision-making tools to identify the most appropriate imaging examinations for specific clinical scenarios. This could prioritize resources for patients truly in need of radiological evaluations while reducing the prevalence of low-value imaging. Moreover, integrating these capabilities directly into the Radiology Information System (RIS), for example, by directly proposing a diagnostic test based on the clinical scenario or the clinical request into a specific text box in the RIS, could streamline workflows and improve overall efficiency and patient experience in clinical practice.

3 | Expanding Beyond Text: The Multimodal Evolution of LLMs

Early LLMs were primarily text-based, limiting their application to tasks like natural language processing and generation. Recent advancements, however, have pushed LLMs into the realm of multimodality capability, enabling them to process and interpret multiple types of input, including text and images. This evolution enables direct image analysis, bridging textual and visual domains, potentially impacting diagnostic and clinical applications in radiology.

In radiology, the first published studies on LLMs' processing multimodal input consisting of both text and images have used inputs from publicly available case datasets. Suh et al. [21] assessed the diagnostic accuracy of GPT-4V (a general-purpose preview model of GPT-4 with vision capabilities, without specific training on medical images) and Gemini Pro Vision to interpret Diagnosis Please cases published in the Radiology journal. Of 190 cases, five were breast imaging cases. Patient history and original images were extracted from the PDF files, and uploaded to the chatbots with specific prompts, which included "Based on this information, present three different possible disease candidates …". Different temperature settings were used. Technically, the temperature is a parameter that can be modified in the development and use of LLMs. It controls the randomness of a LLM output (the response provided by the model), whereby lower temperature values make responses more deterministic, while higher ones introduce more diversity and creativity, often resulting in less predictable responses. Case solutions from eight board-certified radiologists were considered as reference standard. The authors found an improvement in the overall accuracy (including all 190 cases) with higher temperature settings (GPT-4V, 41%–49%; Gemini Pro Vision, 29%–39%). However, radiologists outperformed both LLMs, achieving 61% accuracy overall. Interestingly, the gap between human and LLMs performance narrowed for breast imaging cases, where Gemini Pro Vision performed comparably to radiologists (GPT-4V, 60%; Gemini Pro Vision, 60%–80%), without significant impact of the temperature variation.

Another study by Mukherjee et al. [33] compared GPT-4V's performance with that of radiologists and residents in interpreting 72 Cases of the Day from the Radiological Society of North America 2023 annual meeting. A screenshot of the case was provided to GPT-4V twice, with prompts classified in two categories: imaging-dependent, where image interpretation was required to determine a correct response, and imaging-independent, where image interpretation was not required to determine a correct response. The same screenshots were assessed by five radiologists and three residents. The potential diagnostic improvement with GPT-4V assistance to radiologists and residents was also assessed, providing them with the GPT-4V response, with the option to modify their initial response. The correct diagnosis, provided by the authors of the case of the day, was considered the reference standard for calculating accuracy. GPT-4V achieved a modest 43% accuracy overall, relying heavily on textual context for decision-making. Its accuracy for imaging-dependent questions (39%) was notably lower than for imaging-independent ones (70%). Radiologists and residents did not achieve significantly higher accuracies (59%–76%) than GPT-4V, and their performance did not improve significantly with GPT-4V assistance. These findings highlight the limitations of GPT-4V's multimodal capabilities and suggest the need for improved training and prompting strategies for medical imaging tasks.

Similarly, Hayden et al. [22] tested GPT-4V on ACR Diagnostic Radiology In-Training Examination questions, which included both text- and image-based questions. Screenshots of the questions were provided to GPT-4V, and the ACR-determined correct choice was the reference standard. No comparison with radiologists was performed in this study. The model achieved 65.3% overall accuracy, with better performance on text-based questions (81.5%) compared with image-based ones (47.8%). Structured prompts improved text-based responses but had no significant effect on image-based responses, highlighting the limitations of current LLMs in analyzing visual data.

Despite these challenges, multimodal LLMs continue to show promise in radiology compared with traditional models. At the 2024 Institute of Electrical and Electronics Engineers (IEEE) 12th International Conference on Healthcare Informatics, Guo and Wan [34] compared two LLMs, LLaVA and GPT-4, with the classic Visual Geometry Group model (VGG, a well-known deep learning model) for tumor classification in brain magnetic resonance imaging, breast ultrasound, and kidney computed tomography scans. Publicly available datasets from Kaggle were used, split into training, validation, and test sets. Through a dialogue-based method with the prompt "Is there a tumor in this picture? Please answer yes or no," the LLMs tumor classification capabilities were compared to the results obtained from the VGG. With prompt engineering, the LLMs achieved their highest accuracies (98%, 112%, and 69%, respectively, compared to the baseline VGG model), but the authors emphasized the need for caution due to issues like hallucinations, privacy concerns, and accountability challenges.

To date, the full potential of multimodal LLMs in breast imaging remains largely untapped. While LLM advancements in other fields, such as chest radiography, demonstrate limited effectiveness [21], studies focused on breast imaging are still in their infancy. Preliminary research, including conference presentations, suggests that combining LLMs with other AI models, such as convolutional neural networks (CNNs), could enhance diagnostic accuracy. For instance, integrating LLMs with CNNs for breast cancer classification achieved an accuracy of 92% in recent tests [23]. The potential integration of LLMs with commercially available AI software for mammography, digital breast tomosynthesis, and ultrasound remains unexplored at present.

4 | Improving Radiology Reporting with LLMs

LLMs were initially developed to process and generate natural language, making their application to radiology reporting a *sine qua non* [35]. In breast imaging reporting, their applications are particularly promising, given the complexity and standardization required in interpreting findings from different modalities. Potential main applications of LLMs in breast imaging reporting are structuring free-text reports, generating impressions from findings, and classifying findings based on the ACR's Breast Imaging and Reporting Data System (BI-RADS).

Structured reporting, aimed at standardizing and improving the quality of radiology reports, has been a longstanding goal in radiology, with the aim of reducing interobserver variability and enhancing report comprehensiveness. However, the widespread adoption of structured radiology reporting remains hindered due to challenges such as the rigidity of templates and user resistance [36]. LLMs have demonstrated the ability to convert free-text radiology findings into structured reports, potentially adapting to specific institutional needs and workflows. LLMs are capable of multilingual applications and could support radiologists across various imaging modalities, but the published literature is still limited, especially in breast imaging [4]. For instance, the performance of GPT-4 and Gemini in creating structured reports from free-text positron emission tomography/computed tomography (PET/CT) breast cancer reports has been assessed recently by Chen et al. [37], with nuclear medicine physicians' structured reports as the reference standard. Two free-text reports per patient were needed to be included and to provide lesion progression data. Three different outputs were asked from the LLMs. The first two were structured tables including primary breast cancer and metastatic lesions data respectively, while the third one was an analysis of the primary and metastatic lesions in terms of treatment modality, SUV, and size, and the overall progression using standardized criteria. The results showed that GPT-4 outperformed Gemini in data mining for lesion characteristics (e.g., size, metastatic sites) and in ascertaining lesion progression status, achieving higher semantic similarity scores (F1—a metric used to evaluate the balance between precision and recall in classification or information retrieval tasks, ranging from 0 to 1, where 1 indicates perfect precision and recall: 0.930 vs. 0.907). GPT-4 adapted dynamically and optimized results, learning from user-uploaded information, highlighting its potential in the automatization and standardization of radiological reporting in the lesion progression assessment.

Liu et al. [38] evaluated the ability of GPT-4 and Microsoft Bing (based on GPT-4) to convert free-text breast ultrasound reports into structured formats, assign BI-RADS categories, and provide management recommendations. Free-text breast ultrasound reports were provided to the models with this standardized prompt: "Please use the following text in ultrasound medicine to generate a structured medical ultrasound report with patient information, clinical history, ultrasound findings, impression, ACR-BI-RADS category, management recommendations based on ACR-BI-RADS guidelines." GPT-4 outperformed Bing in report quality, diagnostic accuracy, and recommendations, but remained inferior to senior radiologists.

The complexity and specialization of radiology make generating impressions a challenging task for LLMs. Zhang et al. developed in their study [39] a model fine-tuned on medical and radiology-specific data to generate impressions across various imaging modalities (including mammography) and anatomical sites (including the breast). Expert panels scored the impressions highly on linguistic appropriateness and clinical value. While effective in most domains, limitations were noted in specific diagnosis capabilities, underscoring the need for further refinement and context integration. Despite the fact that no mammography-specific results were provided by the authors, the conclusions can potentially be generalized to breast cancer patients, acknowledging the limitation of generalization.

The standardization of radiology reports has been a widely discussed topic for decades, with breast imaging serving as a pioneer in this field, with the first BI-RADS edition developed by the ACR in 1993 [40]. BI-RADS is a standardized system to interpret and report breast imaging findings, but its reproducibility is still considered low to moderate [4, 5].

Variability in inter-reader agreement for BI-RADS category assignments has driven the application of natural language processing (NLP) tools in this domain [41, 42]. Two different systems were used by two different groups, both utilizing advanced NLP techniques. The first group introduced BI-RADS BERT, a specialized deep learning system based on the BERT architecture, fine-tuned on clinical radiology texts. BI-RADS BERT leverages pre-trained neural language models on extensive text corpora, capturing a deep semantic understanding of the clinical context embedded in mammography reports. Thanks to its advanced contextual comprehension, BI-RADS BERT extracts relevant BI-RADS descriptors and assigns final BI-RADS categories [41]. The second group introduced an NLP approach employing semantic term embeddings to automatically extract BI-RADS descriptors from unstructured mammography reports. Specifically, the proposed system combines traditional NLP techniques with vector-based word representations (word embeddings) to precisely identify and classify key findings such as lesion morphology, breast density, and clinical recommendations [42]. Both groups demonstrated that NLP tools effectively extract BI-RADS descriptors, assign accurate BI-RADS categories, and predict pathologic outcomes, including biopsy-confirmed cancer diagnoses [41, 42]. These advancements underline the potential of NLP tools to enhance consistency and precision in breast imaging evaluations. Such tools offer a promising approach to address diagnostic variability, support standardization, and improve clinical workflows in breast imaging reporting and data interpretation.

More recently, generically trained LLMs have been investigated for this task. Two hundred and fifty factitious breast imaging reports, representing screening and diagnostic mammograms, ultrasounds, and MRIs, were generated by Haver et al. to assess the performance of GPT-4 in assigning BI-RADS categories [20]. Overall, GPT-4 achieved 73.6% accuracy, excelling in assigning BI-RADS categories across diagnostic mammogram reports (92%) but struggling with MRI reports (58%) and complex cases requiring nuanced interpretation. Reproducibility was also a challenge, with the model providing consistent BI-RADS categories in only 70% of repeated tests. These findings underscore the need for refined prompt engineering and domain-specific training to enhance clinical applicability.

Similarly, the agreement between LLMs and human radiologists in assigning BI-RADS categories from free-text findings was investigated by Cozzi et al. across 2400 clinical breast imaging reports written in English, Italian, and Dutch [4]. The authors hypothesized a scenario in which a patient or a referring physician sought a second opinion from a publicly available LLM, assessing the potential clinical impact of the recommendation provided by the model based on the assigned BI-RADS category. Human–human agreement was nearly perfect (quite unexpected, considering the known variability in inter-reader agreement in BI-RADS assignments), while GPT-4 achieved moderate agreement, outperforming GPT-3.5 and Google Bard (now Gemini). The general-purpose LLMs frequently produced discordant BI-RADS assignments, potentially leading to clinically detrimental management changes in 10.6% of cases for GPT-4, compared with 1.5% for human reviewers. Disagreements often involved upgrades or downgrades in BI-RADS categories, potentially affecting clinical management, for example, reducing the level of suspicion of a lesion requiring percutaneous biopsy. These findings highlight the limited reliability of general-purpose LLMs in performing complex medical reasoning, emphasizing the need for regulatory oversight and task-specific training before clinical implementation.

What emerges is that generically trained LLMs are missing precise domain-specific training in many fields, including the handling of radiology reports. An important step to overcome these limitations is to provide large, annotated databases to the models for fine-tuned training, as proposed by Hussain et al. in their paper published in 2024 [24]. Their study introduced the TECRR dataset, consisting of 5046 curated breast imaging reports from TecSalud hospitals in Mexico for BI-RADS classification. Reports were collected in Spanish, translated into English, and annotated by radiologists. Preprocessing ensured quality and consistency by removing duplicates and non-radiological data. Word-embedding techniques like term frequency/inverse document frequency (TF-IDF, a statistical measure weighting the importance of words based on their frequency in a document) and Word2Vec (a neural network-based model able to generate vector representations of words, capturing both semantic and syntactic relationships) were used for feature extraction. Model evaluation of machine learning models, deep learning models, and LLMs showed that BioGPT (a specialized transformer model for biomedical text generation [43]) outperformed others in sensitivity (0.60) on preprocessed data, while XGBoost (a machine learning model) achieved the highest accuracy (0.86). These results provide a benchmark for BI-RADS classification using structured reports, aiding future AI-driven advancements in breast imaging analysis and diagnostic standardization. However, the value of these results is limited by the dataset's including reports from a single institution only, with reports translated from Spanish to English using Google Translate, and the lack of MRI reports (MRI is the imaging modality in which LLMs tend to show the worst performance and thus where more training is needed). Future studies including reports in different languages and representing all breast imaging modalities should be promoted.

5 | LLMs in the Multidisciplinary Meeting: Aiding Management Decision-Making

Multidisciplinary meetings, or tumor boards, are essential in the management of patients with cancer, bringing together expertise from radiology, oncology, radiation therapy, pathology, and surgery, among others, to develop personalized therapeutic approaches. In breast cancer care, radiologists play a pivotal role by providing imaging assessments that guide diagnosis, staging, and treatment planning [44]. LLMs offer the potential to assist radiologists and other specialists in preparing for tumor board discussions, but their integration into this domain remains a work in progress.

Studies have begun to evaluate the role of LLMs in supporting tumor board decision-making. Deng et al. [25] tested GPT-3.5, GPT-4.0, and Claude2 in simulated breast cancer scenarios, including assessment and diagnosis, treatment decision-making, postoperative care, psychosocial support, and prognosis and rehabilitation. Quality, relevance and applicability of the provided responses were systematically assessed. A score ranging from 1 (insufficient) to 4 (excellent) was used to assess quality, identifying as insufficient a response that fails to adequately address the medical query, potentially lacking medical accuracy or omitting important medical details. Excellent responses fully respond to the medical query and provides additional relevant medical information or insights. The relevance of the response was also given a grade from 1 (insufficient) to 4 (excellent). Insufficient responses are potentially off-topic, medically inaccurate, or providing information that does not pertain to the medical question. Excellent responses instead address the medical query directly, with high relevance to all aspects of the medical question. The applicability of the responses was instead classified into three grades, from not applicable, partially to fully applicable. The response potentially does not fit the medical context or situation described in the query (not applicable), or can be directly applied to the medical context or situation described in the query, without any modifications or additions needed (fully applicable). GPT-4 outperformed GPT-3.5 and Claude2 in terms of quality, relevance, and applicability, especially in psychosocial support and treatment planning. Claude2 demonstrated strength in diagnostic assessment but fell short in other areas. Expert evaluations highlighted GPT-4's superior ability to provide comprehensive and clinically relevant responses across tasks, while GPT-3.5 lagged behind GPT-4 significantly. However, limitations still emerged with GPT-4, such as inaccuracies in following nuanced clinical guidelines and a lack of adaptation to regional medical practices.

Similar conclusions come from Griewing et al. [45] through their investigation of the concordance of five different publicly available LLMs with the recommendations of a multidisciplinary tumor board regarding treatment recommendations for complex breast cancer cases. They concluded: “At present, safe and evidenced use of LLM in clinical breast cancer care is not yet feasible.”

A similar approach was shared by Sorin et al. [46], evaluating GPT-3.5 as a support for breast tumor board decision-making. Ten consecutive female patients who were presented at a breast tumor board were analyzed. The goal was to compare GPT-3.5's management recommendations with those made by the tumor board. Of the 10 cases, 70% of GPT-3.5's recommendations were in line with the tumor board's decisions. The study points out that while GPT-3.5 shows promise in supporting clinical decision-making, there are limitations such as missing key clinical details, lack of referral to additional imaging, and potential bias in GPT-3.5's training data.

Using 20 patient profiles reflecting diverse pathological and molecular subtypes, another study based on GPT-3.5 found partial concordance (50% overall, 58.8% for invasive cancers) with tumor board final recommendations [26]. GPT-3.5 showed potential in recognizing key factors like hormonal status and metastatic scenarios but demonstrated inconsistencies, such as misusing genetic risk data.

As tumor boards rely increasingly on comprehensive and collaborative decision-making, LLMs could play a supporting role by synthesizing complex information, offering evidence-based recommendations, and streamlining preparatory tasks. For example, we can hypothesize a breast cancer case with biological and dimensional characteristics that may suggest a possible neoadjuvant therapy, but whose indication remains uncertain. The patient's clinical data are provided to the LLM, which provides the multidisciplinary group with an indication of which figures should be present at the discussion (radiologist to provide precise information on the extent of the disease, pathologist to provide more details on the histotype and receptor characteristics, oncologist and surgeon to discuss the two different options that is, surgery immediately or after neoadjuvant therapy). Furthermore, the LLM can already provide the group with an indication of the relevant guidelines, perhaps highlighting the relevant differences between the different sources. In relation to the different characteristics of the patient, for example, a young age or a diagnosis during pregnancy, the LLM can suggest the presence of a psychologist during the discussion in those groups where the presence of one is not constant. The potential applications are many and, also depending on the preferences of the various groups, are adapted case by case.

The future of LLMs in tumor boards likely lies in highly specialized, continuously updated models trained on expansive datasets. With further refinement, these tools have the potential to enhance the efficiency and accuracy of multidisciplinary decision-making, ultimately improving outcomes for patients.

6 | LLM Limitations

The limitations of LLMs are beyond the primary aim of this review, but we provide here discussion on the main ones that pose critical challenges to the safe and effective deployment of LLMs in breast imaging [17].

One of the most concerning limitations is the phenomenon of hallucinations [8, 14], where LLMs generate incorrect or fabricated responses. These inaccuracies often arise from insufficient pre-training, lack of domain-specific knowledge, or biases in the training data. Hallucinations can lead to the generation of plausible-sounding but erroneous conclusions, which could mislead clinicians [4].

Beyond hallucinations, LLMs can also struggle with understanding nuanced clinical scenarios or integrating complex, multimodal data [21, 22, 33]. This gap in capability is particularly problematic in tasks like BI-RADS categorization, where consistency is critical and where even small errors can significantly impact patient management [4].

Moreover, LLMs can amplify existing biases present in the datasets used for training. These biases may stem from demographic imbalances, underrepresentation of specific clinical conditions, or regional variations in practice guidelines, potentially resulting in outputs that are skewed or unrepresentative of diverse patient populations.

Another key limitation of LLMs is now their reliance on static, historical data. While LLMs can process vast amounts of data, they still do not incorporate real-time updates now. This lag in adaptability can be a significant drawback in rapidly evolving fields such as medicine, where new evidence emerges daily.

To address these challenges, significant advancements are required in both the training and implementation of LLMs. Fine-tuning on high-quality datasets is essential to improve their accuracy and relevance in clinical contexts [24]. The need for specific medical training has been listed by several authors, and now studies with this focus are emerging, including with Google Med-PaLM [26], LLaVA Med (Meta) [28], and the already cited Glass AI [27].

Equally important is the development of robust validation frameworks and governance mechanisms to ensure that LLM outputs are accurate, unbiased, and aligned with ethical and regulatory standards. Ethical concerns in deploying LLMs for breast imaging include risks of biased outputs due to imbalanced training data, compromised patient privacy from handling sensitive health information, and accountability challenges in clinical decision-making and must be addressed. Ultimately, while LLMs offer immense potential, these limitations underscore the necessity of human oversight, with clinicians serving as the final decision-makers to mitigate risks and ensure patient safety.

7 | The Role of LLMs in Breast Imaging in the Near Future

Despite being in their early stages and facing multiple challenges, LLMs have several potential applications in healthcare, including breast imaging, and they could have an impact even in a very short time considering the speed of growth we are observing.

With their increasing ability to understand complex medical contexts across diverse patient pathways and multiple imaging examinations, LLMs will play a critical role in optimizing logistics within breast imaging departments. It can be anticipated that this will enable efficient triaging of patients with similar complaints to appropriate imaging examinations while flagging cases requiring alternative measures. This would reduce scheduling errors, minimize workflow delays, and ultimately enhance the overall patient experience. The accuracy of such predictions will continue to improve over time, depending predominantly on the willingness of stakeholders to provide data for training and validation. Once integrated into the RIS, LLMs can perform many tasks autonomously with little-to-no human intervention. When fully connected to large-scale hospital systems and referring physicians' databases, LLMs may not only suggest imaging tests based on clinical prompts but also automate the scheduling of imaging examinations, further enhancing institutional efficiency. With the recent rise of visual language models (VLMs), these models could also provide recommendations for additional imaging examinations when initial ones are inconclusive or insufficient to answer the proposed questions.

Regarding radiology reporting, although the BI-RADS framework has successfully introduced a global language for breast imaging, variability persists across international, national, and even inter- and intra-institutional practices. LLMs have been shown to excel at prepopulating text fields, but radiologists are often hesitant to rely on such systems due to perceived descriptive errors in descriptive accuracy, which stem from both local variability and LLM limitations. However, it is envisioned that

LLMs have the potential to provide globally driven standardized reporting templates and terminology for breast imaging examinations. Over time, this could eventually lead to a whole new BI-RADS lexicon and likely more consistency from LLM to produce meaningful reports on a global scale.

Current studies discussing treatment recommendations show that LLMs often still fall short in providing recommendations for further diagnostics and treatments [19, 27]. This can be partially overcome by training on larger, more diverse datasets that better reflect individual patient contexts and nuanced recommendations [24]. However, in the present studies, the human ground truths are guideline-based and suffer from long update times. LLMs could consider the latest clinical guidelines, research findings, and drug interactions, ensuring that recommendations are both current and comprehensive.

Additionally, LLMs could help patients understand their treatment options in plain language, fostering better communication with healthcare providers, enabling more informed decision-making and treatment compliance. Patients could also use LLM-powered secure platforms to seek second opinions by uploading reports, imaging studies, or treatment recommendations [4]. These systems could analyze the information to provide detailed, evidence-based insights, highlighting potential concerns or alternative interpretations. By simplifying complex medical jargon and offering a broader context, LLMs could empower patients to engage in more informed discussions with their physicians, ultimately enhancing shared decision-making and ensuring more personalized care.

Ultimately, for LLMs to be effectively integrated into patient care, physicians must act as critical learned intermediaries. They must validate, refine, and contextualize LLM-generated recommendations with their clinical expertise and understanding of the patient's unique circumstances. This collaborative approach ensures that decisions remain patient-centered, evidence-based, and ethically sound, leveraging the speed and breadth of LLMs while safeguarding against inaccuracies or biases.

8 | Conclusions

In conclusion, the integration of LLMs into breast imaging is poised to bring transformative advancement, offering significant potential to enhance clinical practice. LLMs, exemplified by GPT-4, Gemini, and others, have demonstrated their capability in providing decision-making support, optimizing workflows, and supporting radiological reporting. From triaging patients to generating structured reports and integrating clinical guidelines, LLMs present a promising avenue to address the growing demands and persistent challenges within breast imaging. Their ability to standardize radiological interpretations and synthesize complex data could improve diagnostic consistency, reduce inter-reader variability, and ultimately enhance patient management.

The emergence of multimodal LLMs, capable of integrating textual and visual data, introduces a new frontier in breast imaging. While these models have shown early promise in tasks such as tumor classification and diagnostic support, they currently lag behind human experts in accuracy and reliability. Looking ahead, the role of LLMs in breast imaging will likely expand beyond the radiology department, providing support for decision-making, multidisciplinary meetings, and patient education. However, successful integration will require a collaborative approach, with radiologists and other healthcare providers serving as critical intermediaries to use LLM-generated outputs.

Acknowledgments

The authors thank Joanne Chin, MFA, ELS, for editing the manuscript.

Disclosure

LLM support was employed to build this review, including for researching content. However, no portion of the text was generated directly by these models; every part of the manuscript was written by the authors (tested by zeroGPT, https://www.zerogpt.com/).

References

1. T. Tan, A. Rodriguez-Ruiz, T. Zhang, et al., "Multi-Modal Artificial Intelligence for the Combination of Automated 3D Breast Ultrasound and Mammograms in a Population of Women With Predominantly Dense Breasts," *Insights Into Imaging* 14 (2023): 10.

2. Yahoo Finance, "ChatGPT on Track to Surpass 100 Million Users Faster than TikTok or Instagram: UBS," https://finance.yahoo.com/news/chatgpt-on-track-to-surpass-100-million-users-faster-than-tiktok-or-instagram-ubs-214423357.html, 2023.

3. P. Hager, F. Jungmann, R. Holland, et al., "Evaluation and Mitigation of the Limitations of Large Language Models in Clinical Decision-Making," *Nature Medicine* 30 (2024): 2613–2622.

4. A. Cozzi, K. Pinker, A. Hidber, et al., "BI-RADS Category Assignments by GPT-3.5, GPT-4, and Google Bard: A Multilanguage Study," *Radiology* 311, no. 1 (2024): e232133, https://doi.org/10.1148/radiol.232133.

5. G. Irmici, A. Cozzi, G. Della Pepa, et al., "How Do Large Language Models Answer Breast Cancer Quiz Questions? A Comparative Study of GPT-3.5, GPT-4 and Google Gemini," *Radiologia Medica* 129, no. 10 (2024): 1463–1467, https://doi.org/10.1007/s11547-024-01872-1.

6. L. C. Almeida, E. M. J. M. Farina, P. E. A. Kuriki, N. Abdala, and F. C. Kitamura, "Performance of ChatGPT on the Brazilian Radiology and Diagnostic Imaging and Mammography Board Examinations," *Radiology Artificial Intelligence* 6, no. 1 (2024): e230103, https://doi.org/10.1148/ryai.230103.

7. P. Kumar, "Large Language Models (LLMs): Survey, Technical Frameworks, and Future Challenges," *Artificial Intelligence Review* 57 (2024): 260.

8. A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large Language Models in Medicine," *Nature Medicine* 29 (2023): 1930–1940.

9. T. Zhang, T. Tan, X. Wang, et al., "RadioLOGIC, A Healthcare Model for Processing Electronic Health Records and Decision-Making in Breast Disease," *Cell Reports Medicine* 4 (2023): 101131.

10. V. Sorin, B. S. Glicksberg, Y. Artsi, et al., "Utilizing Large Language Models in Breast Cancer Management: Systematic Review," *Journal of Cancer Research and Clinical Oncology* 150 (2024): 140.

11. R. Doshi, K. S. Amin, P. Khosla, S. Bajaj, S. Chheang, and H. P. Forman, "Quantitative Evaluation of Large Language Models to Streamline Radiology Report Impressions: A Multimodal Retrospective Analysis," *Radiology* 310, no. 3 (2024): e231593, https://doi.org/10.1148/radiol.231593.

12. E. Sblendorio, V. Dentamaro, A. Lo Cascio, F. Germini, M. Piredda, and G. Cicolini, “Integrating Human Expertise & Automated Methods for a Dynamic and Multi-Parametric Evaluation of Large Language Models' Feasibility in Clinical Decision-Making,” *International Journal of Medical Informatics* 188 (2024): 105501.

13. S. A. Haider, S. M. Pressman, S. Borna, et al., “Evaluating Large Language Model (LLM) Performance on Established Breast Classification Systems,” *Diagnostics* 14 (2024): 1491.

14. J. A. Omiye, H. Gui, S. J. Rezaei, J. Zou, and R. Daneshjou, “Large Language Models in Medicine: The Potentials and Pitfalls,” *Annals of Internal Medicine* 177 (2024): 210–220.

15. S. Tripathi, K. Gabriel, S. Dheer, et al., “Understanding Biases and Disparities in Radiology AI Datasets: A Review,” *Journal of the American College of Radiology* 20 (2023): 836–841.

16. H. R. Kirk, B. Vidgen, P. Röttger, and S. A. Hale, “The Benefits, Risks and Bounds of Personalizing the Alignment of Large Language Models to Individuals,” *Nature Machine Intelligence* 6 (2024): 383–392.

17. B. Murdoch, “Privacy and Artificial Intelligence: Challenges for Protecting Health Information in a New Era,” *BMC Medical Ethics* 22 (2021): 122.

18. B. Meskó and E. J. Topol, “The Imperative for Regulatory Oversight of Large Language Models (Or Generative AI) in Healthcare,” *npj Digital Medicine* 6 (2023): 120.

19. A. Rao, J. Kim, M. Kamineni, et al., “Evaluating GPT as an Adjunct for Radiologic Decision Making: GPT-4 Versus GPT-3.5 in a Breast Imaging Pilot,” *Journal of the American College of Radiology* 20 (2023): 990–997.

20. H. L. Haver, P. H. Yi, J. Jeudy, and M. Bahl, “Use of ChatGPT to Assign BI-RADS Assessment Categories to Breast Imaging Reports,” *American Journal of Roentgenology* 223 (2024): e2431093.

21. P. S. Suh, W. H. Shim, C. H. Suh, et al., “Comparing Diagnostic Accuracy of Radiologists Versus GPT-4V and Gemini pro Vision Using Image Inputs From Diagnosis Please Cases,” *Radiology* 312 (2024): e240273.

22. N. Hayden, S. Gilbert, L. M. Poisson, B. Griffith, C. Klochko, and S. Wolfe, “Performance of GPT-4 with Vision on Text- and Image-Based ACR Diagnostic Radiology in-Training Examination Questions,” *Radiology* 312 (2024): e240153.

23. K. K. Harini, R. Nandhini, A. M. Rajeswari, and R. Deepalakshmi, “Breast Cancer Image Classification: Leveraging Deep Learning and Large Language Models for Semantic Integration,” in *IEEE International Conference on Contemporary Computing and Communications (InC4)* (Bangalore: IEEE, 2024), 1–6.

24. S. Hussain, U. Naseem, M. Ali, et al., “TECRR: A Benchmark Dataset of Radiological Reports for BI-RADS Classification With Machine Learning, Deep Learning, and Large Language Model Baselines,” *BMC Medical Informatics and Decision Making* 24 (2024): 310.

25. L. Deng, T. Wang, Z. Zhai, et al., “Evaluation of Large Language Models in Breast Cancer Clinical Scenarios: A Comparative Analysis Based on ChatGPT-3.5, ChatGPT-4.0, and Claude2,” *International Journal of Surgery* 110 (2024): 1941–1950.

26. S. Griewing, N. Gremke, U. Wagner, M. Lingenfelder, S. Kuhn, and J. Boekhoff, “Challenging ChatGPT 3.5 in Senology—An Assessment of Concordance With Breast Cancer Tumor Board Decision Making,” *Journal of Personalized Medicine* 13 (2023): 1502.

27. H. A. Zaki, A. Aoun, S. Munshi, H. Abdel-Megid, L. Nazario-Johnson, and S. H. Ahn, “The Application of Large Language Models for Radiologic Decision Making,” *Journal of the American College of Radiology* 21 (2024): 1072–1078.

28. K. Singhal, S. Azizi, T. Tu, et al., “Large Language Models Encode Clinical Knowledge,” *Nature* 620 (2023): 172–180.

29. C. Li, C. Wong, S. Zhang, et al., “LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day,” *Advances in Neural Information Processing Systems* 36 (2023): 28541–28564.

30. E. Kjelle, E. R. Andersen, A. M. Krokeide, et al., “Characterizing and Quantifying Low-Value Diagnostic Imaging Internationally: A Scoping Review,” *BMC Medical Imaging* 22 (2022): 73.

31. A. Y. Sheng, A. Castro, and R. E. Lewiss, “Awareness, Utilization, and Education of the ACR Appropriateness Criteria: A Review and Future Directions,” *Journal of the American College of Radiology* 13 (2016): 131–136.

32. J. H. Barth, S. Misra, K. M. Aakre, et al., “Why Are Clinical Practice Guidelines Not Followed?,” *Clinical Chemistry and Laboratory Medicine (CCLM)* 54 (2016): 1133–1139.

33. P. Mukherjee, B. Hou, A. Suri, et al., “Evaluation of GPT Large Language Model Performance on RSNA 2023 Case of the Day Questions,” *Radiology* 313 (2024): e240609.

34. Y. Guo and Z. Wan, “Performance Evaluation of Multimodal Large Language Models (LLaVA and GPT-4-Based ChatGPT) in Medical Image Classification Tasks,” in *IEEE 12th International Conference on Healthcare Informatics (ICHI)* (Orlando, FL: IEEE, 2024), 541–543.

35. F. Busch, L. Hoffmann, D. P. dos Santos, et al., “Large Language Models for Structured Reporting in Radiology: Past, Present, and Future,” *European Radiology* 35, no. 5 (2024): 1–14.

36. L. J. Grimm, A. L. Anderson, J. A. Baker, et al., “Interobserver Variability Between Breast Imagers Using the Fifth Edition of the BI-RADS MRI Lexicon,” *American Journal of Roentgenology* 204 (2015): 1120–1124.

37. K. Chen, W. Xu, and X. Li, “The Potential of Gemini and GPTs for Structured Report Generation Based on Free-Text 18F-FDG PET/CT Breast Cancer Reports,” *Academic Radiology* 32, no. 2 (2024): 624–633.

38. C. Liu, M. Wei, Y. Qin, 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 in Medicine & Biology* 50 (2024): 1697–1703.

39. L. Zhang, M. Liu, L. Wang, et al., “Constructing a Large Language Model to Generate Impressions From Findings in Radiology Reports,” *Radiology* 312 (2024): e240885.

40. E. A. Sickles and C. J. D'Orsi, “ACR BI-RADS Follow-Up and Outcome Monitoring,” in *ACR BI-RADS Atlas, Breast Imaging Reporting and Data System*, 5th ed. (American College of Radiology, 2013).

41. G. Kuling, B. Curpen, and A. L. Martel, “BI-RADS BERT and Using Section Segmentation to Understand Radiology Reports,” *Journal of Imaging* 8 (2022): 131.

42. I. Banerjee, S. Bozkurt, E. Alkim, H. Sagreiya, A. W. Kurian, and D. L. Rubin, “Automatic Inference of BI-RADS Final Assessment Categories From Narrative Mammography Report Findings,” *Journal of Biomedical Informatics* 92 (2019): 103137.

43. R. Luo, L. Sun, Y. Xia, et al., “BioGPT: Generative Pre-Trained Transformer for Biomedical Text Generation and Mining,” *Briefings in Bioinformatics* 23, no. 6 (2022): bbac409.

44. A. Di Pilla, M. R. Cozzolino, A. Mannocci, et al., “The Impact of Tumor Boards on Breast Cancer Care: Evidence From a Systematic Literature Review and Meta-Analysis,” *International Journal of Environmental Research and Public Health* 19 (2022): 14990.

45. S. Griewing, J. Knitza, J. Boekhoff, et al., “Evolution of Publicly Available Large Language Models for Complex Decision-Making in Breast Cancer Care,” *Archives of Gynecology and Obstetrics* 310 (2024): 537–550.

46. V. Sorin, E. Klang, M. Sklair-Levy, et al., “Large Language Model (ChatGPT) as a Support Tool for Breast Tumor Board,” *npj Breast Cancer* 9, no. 1 (2023): 44, https://doi.org/10.1038/s41523-023-00557-8.