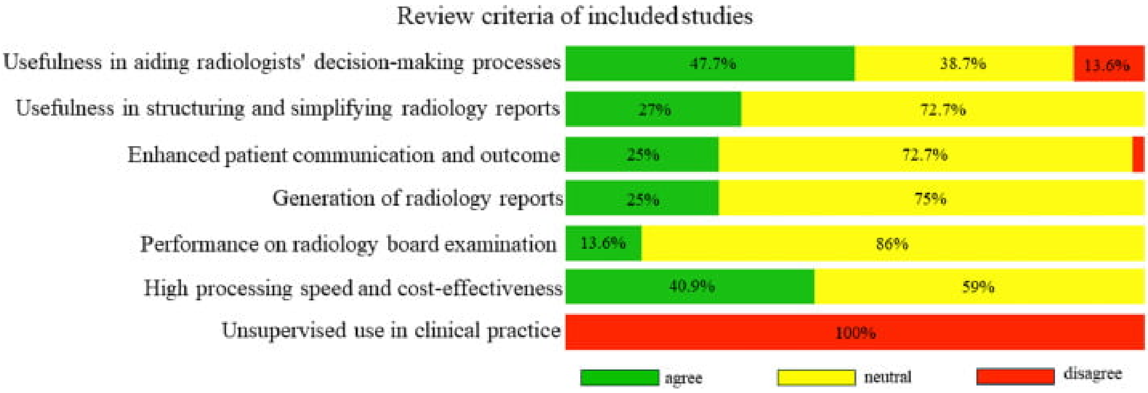
HealthManagement
Promoting Management and Leadership

# ChatGPT in Radiology: Potential, Limitations & Future in Clinical Radiology

Generative pre-trained transformer (ChatGPT), developed by OpenAI, is an advanced web-based chatbot utilising a large language model (LLM) AI. Trained extensively on vast text datasets, it excels in understanding and generating human-like responses, primarily in English. Initially released as ChatGPTv3.5 in November 2022 and updated to ChatGPTv4 in March 2023, it quickly gained attention for medical applications, particularly in radiology-related domains. Studies have explored its utility in clinical radiology writing, training, report generation, patient education, and disease screening. However, debates have arisen regarding the benefits and drawbacks of integrating AI like ChatGPT into routine clinical practice, including concerns about potential biases in its training data. A recent study published in Diagnostic and Interventional Imaging conducted a systematic review of ChatGPT's performance, identified limitations, and discussed future integration, optimisation, and ethical considerations in radiology applications.

## Methodology to review AI-based conversational LLMs in clinical radiology

The systematic review, conducted according to PRISMA guidelines, aimed to comprehensively gather studies up to January 1, 2024, concerning the utilisation of AI-based conversational large language models (LLMs), specifically focusing on ChatGPT, in clinical radiology applications. The search strategy involved querying major online databases such as PubMed, Web of Science, Embase, and Google Scholar from their inception without any temporal restrictions. The search employed a combination of medical subject headings (MeSH) and search terms, utilising Boolean operators (AND, OR) to refine the search. Keywords used included "ChatGPT," "Generative Pre-trained Transformer," "Large Language Model," "LLM," "Open AI Chat*," and "Radiology." Additionally, potential eligible studies were identified by manually examining the references of included studies.



Two independent reviewers, each with more than five years of clinical radiology research experience, screened the titles and abstracts identified in the search. They subsequently retrieved and evaluated the full texts of potentially relevant studies. The inclusion criteria encompassed various types of original studies focusing on ChatGPT in clinical radiology, including its use in radiology reports, recommendations, guidelines, and responses to radiology-related questions. Studies that did not meet these criteria, as well as correspondence, conference abstracts, and non-academic sources like magazines, newspapers, and internet websites, were excluded. Moreover, studies focusing on general medical education rather than specifically on radiology education were also excluded.

Data extraction was performed using standardised templates, created with commercially available software such as Microsoft Excel. Disagreements between reviewers at any screening stage were resolved through consensus or, if necessary, through discussions involving a third reviewer. The extracted data included various elements such as the first author's name, study region and design, characteristics of study targets (e.g., images, radiology reports, questions), comparison groups (ChatGPT vs. study targets/other LLM-based chatbots), ChatGPT version, sample size, study prompts, and qualitative and/or quantitative performance metrics (e.g., accuracy, sensitivity, specificity).

ChatGPT's performance was typically evaluated based on four criteria, with a primary focus on accuracy or agreement with radiologists' decisions and guidelines. Studies reporting high overall effectiveness were categorised as demonstrating "high performance," while those reporting inaccuracies were labelled as showing "low performance." Performance assessments were often quantified using percentage data from the included studies.

### Selection Process and Study Characteristics in Assessing ChatGPT's Role in Radiology

From an initial pool of 2263 studies, 1402 duplicates were removed, and 553 more were excluded due to focusing on ChatGPT applications outside clinical medicine. This left 308 studies for full-text screening. Of these, 264 were discarded as they primarily focused on other applications of ChatGPT, such as medical education, publication writing, review studies, correspondence, and ethics. Ultimately, 44 original publications meeting the inclusion criteria were selected. These studies examined ChatGPT's potential across five clinical radiology areas: diagnosis and clinical decision support, transformation and simplification of radiology reports, enhanced patient communication and outcomes, generation of radiology reports, and performance on radiology board examinations. Additionally, 11 studies compared ChatGPT versions 3.5 and 4. Among these, 11 studies were rated as high-quality, while the rest were evaluated as fair quality based on the Newcastle-Ottawa Scale criteria for risk of bias assessment.

### Performance Assessment and Comparative Analysis of ChatGPT in Radiology Practice

Out of 44 studies evaluating ChatGPT's performance in radiology practice, 37 (84.1%) reported high performance, while 7 (15.9%) indicated lower performance, with generally high accuracy noted. Among the 24 studies reporting the proportion of ChatGPT's performance, 19 (79.2%) recorded a median accuracy of 70.5%, and 5 (20.8%) reported a median agreement of 83.6% with radiologists’ decisions or guidelines. Concerns regarding inaccuracies were raised in 7 studies, particularly in diagnosis and clinical decision support in 6 studies (13.6%), and in patient communication in 1 study (2.3%). However, ChatGPT notably improved aiding radiologists' decision-making processes (70%), structuring and simplifying radiology reports (100%), patient communication and outcomes (83.3%), generation of radiology reports (100%), and performance on radiology board examinations (100%). None of the studies suggested ChatGPT's unsupervised use in clinical practice. Additionally, 32 studies (72.7%) reported the specific prompts employed. Comparisons between ChatGPT versions revealed that ChatGPTv4 generally outperformed v3.5 in 10 out of 11 studies (90.9%), particularly in addressing complex questions and enhancing advanced reasoning capabilities. One study suggested that software updates do not consistently improve performance, especially in terms of readability.

### ChatGPT's promises, limitations, and performance across key domains

Findings suggest that ChatGPT shows promise in 84.1% of studies, significantly contributing to five broad clinical areas: diagnostic and clinical decision support, transforming and simplifying radiology reports, patient communication and outcomes, and performance on radiology board examinations. However, this promise is tempered by numerous limitations, including biases, inaccuracies, hallucinations, misinformation, and ethical, legal, and technical aspects of integrating chatbots into clinical practice.

Twenty studies evaluated ChatGPT's decision-support capabilities in clinical radiology, demonstrating moderate to high accuracy in suggesting appropriate imaging modalities and protocols, though neuroradiologists significantly outperformed ChatGPT. Similarly, studies assessing its diagnostic performance found varying levels of accuracy across subspecialties, suggesting its role as a supplementary tool rather than a replacement for experienced professionals.

Eight studies assessed ChatGPT's ability to transform complex radiology reports into structured formats, highlighting its potential to simplify reports and enhance patient-provider communication. However, concerns were raised about oversimplification and data privacy issues. Several studies evaluated ChatGPT's performance in answering radiology board-style questions, with varying levels of accuracy observed. Updates to ChatGPTv4 showed improved performance, particularly in addressing higher-order thinking questions.

### Challenges and ethical considerations in integrating ChatGPT into clinical radiology practice

Despite its potential, limitations such as biased responses, restricted originality, incorrect citations, cybersecurity risks, and privacy concerns were noted. Strategies to mitigate these risks and ensure data privacy and security are crucial. Ethical dilemmas regarding the automation of human tasks, accountability, and legal liability for errors need to be addressed. Multi-disciplinary committees should develop evidence-based guidelines to manage the burden of AI-related errors. Technical challenges, including variations in outputs and inconsistencies in accuracy metrics, must be addressed before the implementation of LLM-based chatbots in clinical radiology practice. Establishing standards for the use of AI and LLM-based chatbots is necessary, alongside continued rigorous research and validation. Professional safeguards are crucial to ensure physicians and patients are well-informed about the technology's capabilities and limitations.

While ChatGPT shows promise in revolutionising radiology, further research, extensive multicenter studies, and validation efforts are required to confirm its proficiency and accuracy in diverse clinical settings.

**Source & Image Credit:** **Diagnostic and Interventional Imaging**

Published on : Thu, 2 May 2024