Article published online: 2025-08-21

From Chatbots to Agentic Workflows: Ensuring Responsible Deployment of Large Language Models in Radiology

Suvrankar Datta[1] Pradosh Kumar Sarangi[2]

[1] Koita Centre for Digital Health, Ashoka University, Sonepat, Haryana, India
[2] Department of Radiodiagnosis, All India Institute of Medical Sciences, Deoghar, Jharkhand, India

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

Indian J Radiol Imaging

Dear Editor,

The recent editorial by Alex and Kesavadas[1] eloquently highlights the transformative potential of large language models (LLMs) in radiological practice, from optimizing workflows to enhancing radiology education. Radiology departments globally are facing increasing demands, making the adoption of LLMs not just advantageous but inevitable. However, it is imperative that we also address critical considerations for the responsible deployment of these powerful tools into regular radiology workflows (see ►**Fig. 1** for an overview of major clinical LLM use cases in radiology and their evaluation metrics).

A fundamental consideration when deploying LLMs in clinical practice involves choosing appropriate deployment strategies. While popular proprietary models such as GPT-4 demonstrate impressive clinical reasoning capabilities, reliance on external cloud-based services inherently requires the transfer of protected health information, often across national borders, posing significant privacy and regulatory challenges. India's recently enacted Digital Personal Data Protection (DPDP) Act, 2023, currently mandates explicit consent for data storage and limits unregulated cross-border data transfers. These provisions encourage strategies that retain radiology data within national boundaries, enhancing legal safety and operational simplicity. Conversely, smaller, domain-specific, locally hosted models, fine-tuned with de-identified data, present viable alternatives, offering comparable task-specific accuracy with reduced risks to patient data privacy and improved compliance with stringent data protection regulations which countries including India are increasingly adopting.[2]

An equally important discussion is the necessity of seamless integration of LLM outputs with existing radiology information systems and electronic health records. The anticipated efficiency gains with LLM-driven tools may be undermined if radiologists need to navigate between disconnected interfaces. Thus, early and close collaboration between health care IT specialists, LLM developers, and radiologists is essential to ensure that LLM outputs are directly integrated into the existing clinical reporting interfaces.[3]

Further complexity arises with the advent of agentic workflows, in which LLMs autonomously perform multi-step clinical tasks such as guideline retrieval, decision support, or automated follow-up scheduling. For instance, in a recent multi-reader pilot at our center, we deployed an agentic workflow for generating radiology reports, using four sequential LLM agents capable of decomposing dictated keywords, retrieving relevant radiographic features via web search, synthesizing key findings, and validating the final structured reports. The agentic pipeline significantly reduced the time to report generation, supporting the feasibility and efficiency of agentic reporting systems in real-world practice. Despite their potential to markedly reduce cognitive burden, these automated systems remain computationally costlier and propagate errors across multiple interdependent tasks, challenging conventional evaluation metrics.[4]

Therefore, rigorous evaluation methods must move beyond traditional metrics like BLEU scores or token-matching accuracy. Comprehensive evaluation frameworks including specific ones for agentic workflows, such as those suggested by Jiang et al in their MedAgentBench paper[5] and combining with automated metrics, adversarial stress-testing also

DOI https://doi.org/10.1055/s-0045-1811264.
ISSN 0971-3026.

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

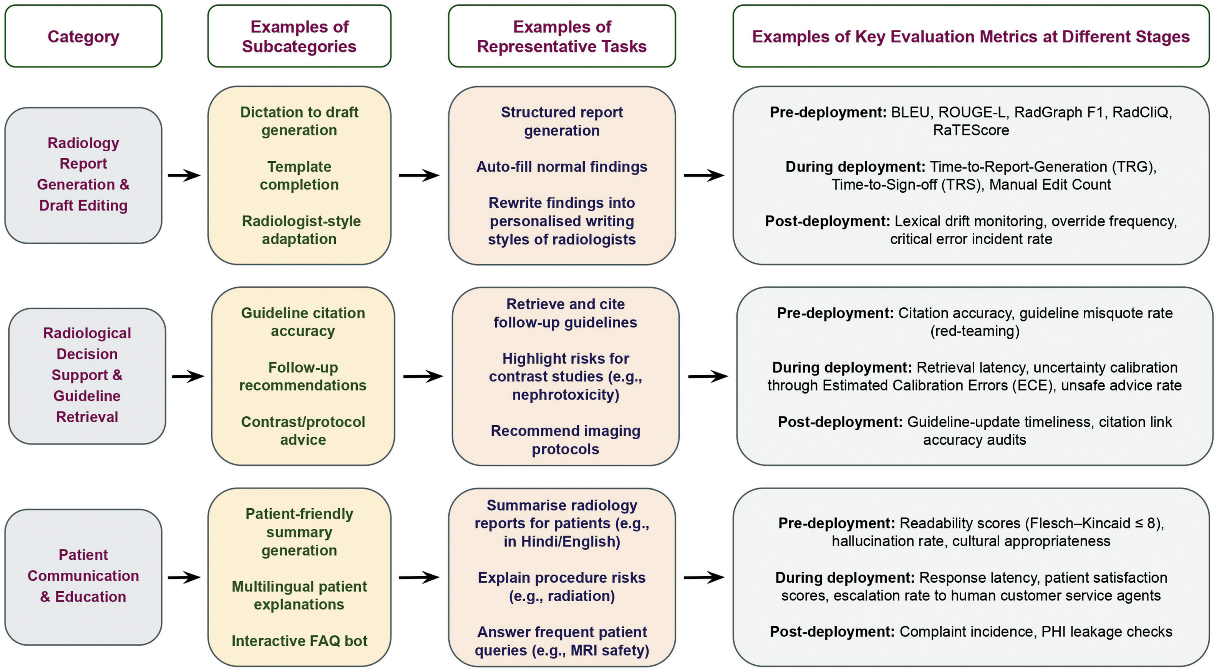

Fig. 1 Clinical applications of LLMs and agentic workflows in radiology: task taxonomy and evaluation metrics across deployment stages. BLEU, Bilingual Evaluation Understudy; LLMs, large language models; PHI, protected health information; RadGraph F1, F1 score metric based on overlap in clinical entities and relations; RadCliQ, radiology report clinical quality; RaTEScore, Radiological Report (Text) Evaluation Score; ROUGE-L, Recall-Oriented Understudy for Gisting Evaluation (Longest Common Subsequence).

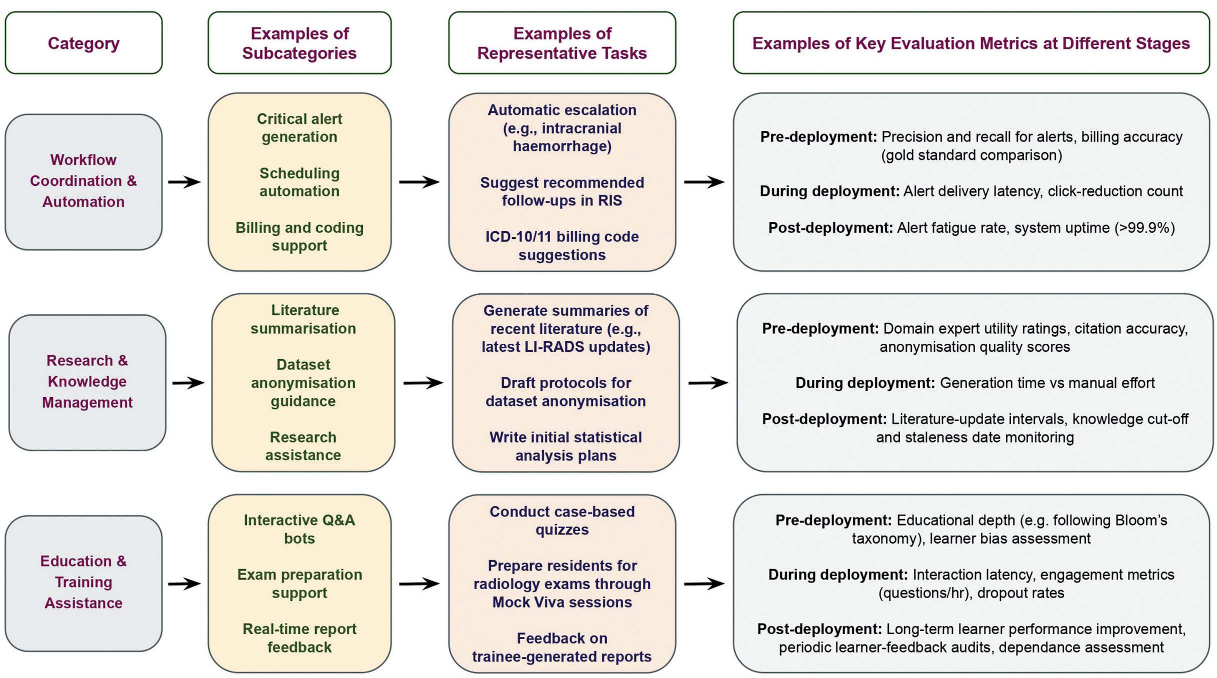

Fig. 2 Operational, research, and educational applications of agentic LLMs in radiology: task taxonomy and evaluation metrics across deployment stages. ICD-10/11, International Classification of Diseases, 10th and 11th Revisions; LI-RADS, liver imaging reporting and data system; LLMs, large language models; Q&A, question and answer; RIS, radiology information system.

known as "red-teaming" in which multidisciplinary reviewers deliberately attempt to "jailbreak" an LLM to output unsafe recommendations, identify worst-case failure modes and identify biases, along with expert radiologist validation, are urgently required to robustly assess both clinical utility and safety prior to broad deployment. ►**Fig. 2** details a stage-wise evaluation framework for operational, research, and educational deployments of LLMs and agentic workflows in radiology. Continuous monitoring of the real-world performance of LLM agents, with iterative feedback loops, will remain critical to maintaining high clinical standards and reliability.

Finally, while advocating for increased automation, it is paramount to maintain a symbiotic relationship between radiologists and LLM-driven software. Radiologists should always retain oversight of clinically impactful decisions, with automated outputs clearly being highlighted, and preferably indicating confidence intervals or uncertainty estimates, mandating radiologist review before clinical decisions are finalized.

In conclusion, the radiology community must proactively and urgently adopt structured, phased strategies that address privacy concerns, ensure interoperability, rigorously validate emerging agentic tools, and support ongoing performance monitoring. These steps are essential to responsibly harness the full potential of LLMs in augmenting radiologic practice, while safeguarding patient privacy and well-being.

Funding
None.

Conflict of Interest
None declared.

References

1 Alex A, Kesavadas C. Revolutionizing radiology: the role of large language models. Indian J Radiol Imaging 2024;35(01):1
2 Bluethgen C, Van Veen D, Zakka C, et al. Best practices for large language models in radiology. Radiology 2025;315(01):e240528
3 Tavakoli N, Kim D. AI-generated clinical histories for radiology reports: closing the information gap. Radiology 2025;314(02):e243910
4 Qiu J, Lam K, Li G, et al. LLM-based agentic systems in medicine and healthcare. Nat Mach Intell 2024;6(12):1418–1420
5 Jiang Y, Black KC, Geng G, et al. MedAgentBench: Dataset for Benchmarking LLMs as Agents in Medical Applications. Ithaca, NY: Cornell University; 2025

