**Int. J. Sci. R. Tech., 2026 3(9)**
A Multidisciplinary peer-reviewed Journal
www.ijsrtjournal.com [ISSN: 2394-7063]

*Original Article*

# AI-Generated Radiology Reports: Opportunities For Efficiency, Diagnostic Support, And Risks To Patient Safety

**Adil Ahmad Wani[1], Vyshak M.[2], Usha G.[3], Junaid ul Islam[4], Zakir Hussain Parray[5]***

[1]*Radiology and Imaging Technology, Swami Vivekanand Institute of Engineering & Technology, India*
[2]*Medical Imaging Technology, SRM Medical College Hospital and Research Centre, Chennai, India*
[3]*KMC Institute of Health Sciences, India*
[4]*Medical Radiology and Imaging Technology, School of Science, Maulana Azad National Urdu University, India*
[5]*A & OTT, Emversity Bangalore*

## ABSTRACT

Generative artificial intelligence (AI) and large language models (LLMs) are increasingly being embedded into radiology reporting workflows, promising to ease a mismatch between imaging volumes that grow 8-12% annually and a radiology workforce that expands by only 2-3% per year. This paper reviews the current evidence on AI-generated radiology reports across three dimensions: efficiency, diagnostic support, and patient safety. Across nine recent clinical studies, AI-assisted drafting reduced radiologist reporting time by 7.8% to 43.8%, with the largest gains observed for structured, high-volume examinations such as outpatient CT. As of March 2026, radiology accounted for 1,163 of 1,524 (76.3%) FDA-cleared artificial-intelligence medical devices, with representative diagnostic-support tools reporting sensitivity and specificity above 90% for intracranial haemorrhage, pulmonary embolism, large-vessel-occlusion stroke, and fracture detection. However, systematic evaluations also document clinically meaningful hallucination and omission rates — up to 5.5% for major omissions and 1.0% for malignant hallucinations in some report categories — alongside documented automation bias and unresolved medico-legal liability when radiologists over-rely on AI output. We synthesise this evidence into a risk-mitigation framework spanning mandatory human-in-the-loop review, hallucination-detection tooling, transparent error-rate disclosure, and regulatory post-market surveillance. We conclude that AI-generated radiology reports offer substantial, reproducible efficiency and diagnostic gains, but that these gains are only safely realisable within governance structures that actively guard against automation bias and hold hallucination rates to clinically negligible levels.

**Keywords**: artificial intelligence; radiology reporting; large language models; diagnostic accuracy; patient safety; automation bias; hallucination; clinical workflow.

## INTRODUCTION

Radiology departments worldwide face a widening gap between the supply of diagnostic imaging and the workforce available to interpret it. Imaging volumes increase by approximately 8-12% annually, while the radiology workforce grows by only 2-3% per year, a mismatch that has intensified reporting backlogs and burnout across health systems [1]. Digitalisation and teleradiology have already reduced report turnaround time by up to 50% relative to film-based workflows, and the diffusion of artificial intelligence (AI) — from early computer-aided detection (CAD) systems to modern large language models (LLMs) and vision-language models (VLMs) — represents the next stage of this transformation [2].

Two distinct but converging applications of AI have emerged in radiology reporting. The first is diagnostic-support AI, typically a convolutional neural network trained to detect a specific finding (for example, intracranial haemorrhage, pulmonary embolism, or a fracture) and flag it for radiologist review. The second, more recent development is

**Relevant conflicts of interest/financial disclosures**: 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.

generative report-writing AI, in which an LLM or VLM drafts free-text findings and impressions from images, structured keywords, or dictated fragments, which the radiologist then edits and signs [3][4]. Both applications are now deployed at scale: as of March 2026, the U.S. Food and Drug Administration (FDA) had cleared 1,163 radiology-specific AI algorithms, comprising more than three-quarters of all FDA-cleared AI medical devices [5].

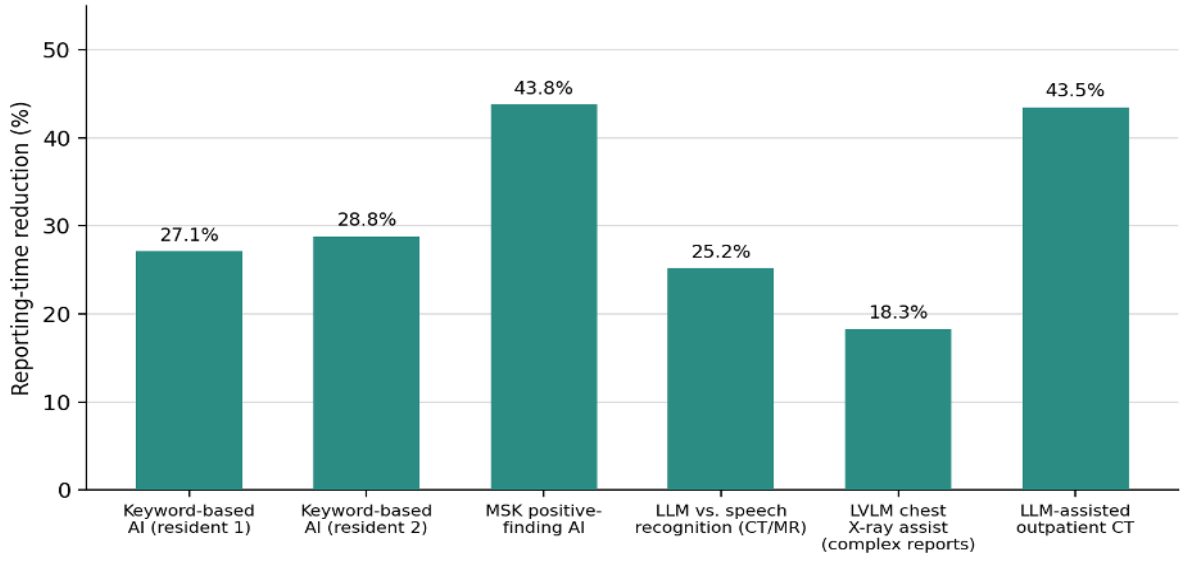
This paper reviews the peer-reviewed and pre-print literature on AI-generated radiology reports published primarily between 2023 and 2026, organised around three questions that determine whether this technology should be embraced, constrained, or redesigned: (i) What efficiency gains does the evidence actually support, and under what conditions do they hold? (ii) How well does AI perform as a diagnostic-support tool relative to unaided radiologist interpretation? and (iii) What specific, quantifiable risks does AI-generated reporting introduce for patient safety, and what governance measures mitigate them?

## II. OPPORTUNITIES FOR EFFICIENCY

The efficiency case for AI-generated reporting rests on a growing body of prospective and retrospective comparisons of reporting time, with and without AI assistance. A pilot study of keyword-based AI-assisted reporting found median reporting-time reductions of 27.1% and 28.8% for two radiology residents, with no significant difference in independently rated report quality ($p>0.50$) and mean primary-diagnosis accuracy of 72.0% for the AI-generated draft [3]. In musculoskeletal and neuroradiology cases, radiologists who entered key positive findings into an AI system that then drafted a full report reduced mean reporting time from 6.1 to 3.43 minutes, a 43.8% reduction ($p<0.0001$), across knee MRI, lumbar-spine MRI, head CT, and abdomen/pelvis CT examinations [6]. A separate European comparison of a general-purpose LLM against conventional speech recognition found shorter median report-generation times with the LLM (238 vs. 318 seconds, a 25.2% reduction) across CT and MR reports [7].

Gains are not uniform across modalities or case complexity. A proof-of-concept study of AI-assisted chest-radiograph reporting using a large vision-language model found a modest 7.8% reduction in mean writing time overall ($p=0.08$), but a significant 18.3% reduction for complex reports specifically ($p<0.05$), suggesting that AI assistance yields its greatest efficiency dividend where interpretive and dictation burden is already highest [8]. A retrospective cohort study at a single academic centre similarly found that LLM assistance significantly shortened inter-study intervals — a proxy for interpretation time — for outpatient CT with and without contrast (reductions of 10 and 11.5 minutes respectively, $p<0.01$), but produced no measurable benefit for MRI reporting, and required approximately ten hours of radiologist training over five days to reach proficiency [4].

**Figure 1. Reported reductions in radiology reporting time with AI-assisted report generation across recent studies**



**Summary of reporting-time findings**

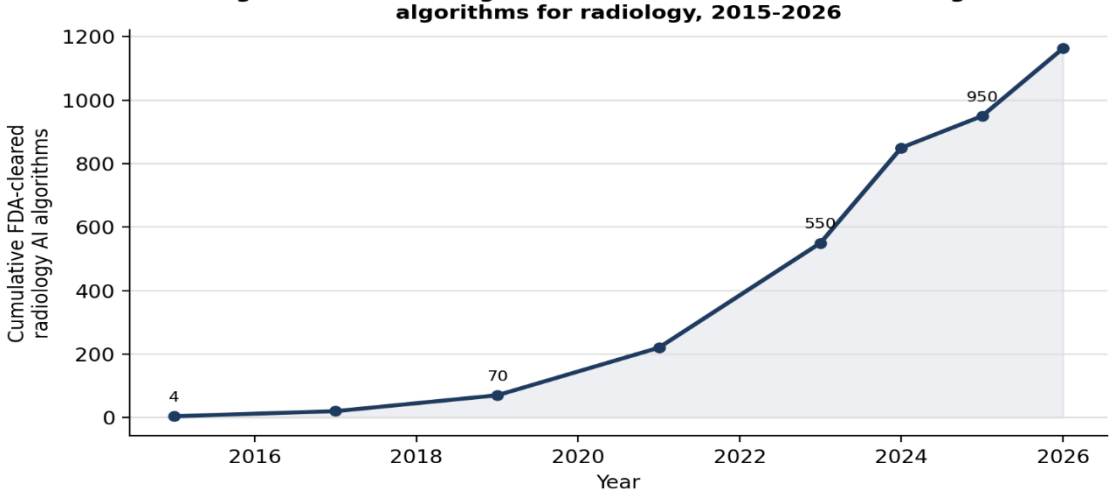
| Study / Setting | Modality | Reported Time Effect | Quality Impact |
|---|---|---|---|
| Keyword-based AI, resident 1 [3] | Mixed CT/MR | -27.1% reporting time | No significant change (p>0.50) |
| Keyword-based AI, resident 2 [3] | Mixed CT/MR | -28.8% reporting time | No significant change (p>0.50) |
| Positive-finding AI drafting [6] | MRI knee/spine, CT head/abdomen | 6.1 -> 3.43 min (-43.8%) | Improved reported confidence |
| LLM vs. speech recognition [7] | CT / MR | 318 -> 238 sec (-25.2%) | Comparable error rate |
| LVLM chest-X-ray assistant [8] | Chest radiograph | -7.8% overall; -18.3% complex | No quality loss reported |
| LLM-assisted outpatient workflow [4] | CT (contrast/non-contrast) | -10 to -11.5 min per study | No MRI benefit observed |

Taken together, these findings support three conclusions. First, AI-assisted drafting produces consistent, statistically significant reductions in reporting time for structured, high-volume examinations, typically in the range of 20-45%, without measurable loss of report quality when radiologists retain editing control. Second, benefit is modality- and complexity-dependent: gains concentrate in CT and in complex, multi-finding reports, while MRI and simple studies show smaller or non-significant effects. Third, realising these gains requires investment — in training (approximately ten hours in one implementation [4]), in workflow redesign, and in the radiologist's continued critical review of AI-generated text, a precondition that becomes central to the patient-safety discussion in Section IV.

## III. DIAGNOSTIC SUPPORT AND CLINICAL PERFORMANCE

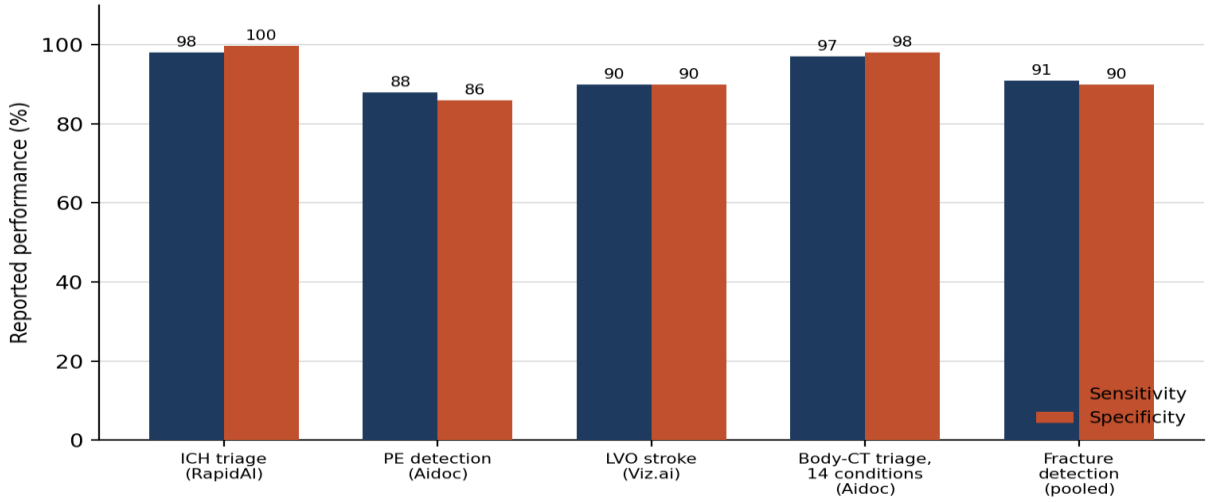
Beyond drafting text, AI increasingly functions as a diagnostic-support layer that flags, prioritises, or corroborates radiologist findings. Regulatory clearance of such tools has accelerated sharply: the FDA cleared roughly 21 AI algorithms per month in 2024, rising to about 30 per month by early 2026, bringing the cumulative total of FDA-cleared AI medical devices to 1,524, of which 1,163 (76.3%) are radiology-specific [5][9]. This trajectory reflects both technical maturation and growing clinical demand for tools that can triage time-critical findings around the clock [9].

**Figure 2. Cumulative growth of FDA-cleared artificial-intelligence algorithms for radiology, 2015-2026**




Reported diagnostic performance for individual tools is frequently strong. A commercially deployed intracranial-haemorrhage (ICH) triage algorithm reported sensitivity of 98.1% and specificity of 99.7% following its most recent FDA clearance, with the ability to detect haemorrhages as small as 0.4 mL [10]. Pooled meta-analytic estimates for pulmonary-embolism (PE) detection on CT pulmonary angiography report sensitivity of 0.88 and specificity of 0.86 [11], though a large single-centre validation of a commercially deployed PE algorithm found the tool detected 19 PE cases missed on the initial radiologist report, at the cost of a lower specificity that increased false positives and raised concerns about AI-driven overdiagnosis if used without radiologist oversight [12]. Large-vessel-occlusion (LVO) stroke detection software has been associated with a 44% reduction in diagnosis time, a 31-minute reduction in time-to-treatment, and a 40% reduction in 90-day disability in a 474-patient multicentre trial [13]. A 2026-cleared foundation-model-based body-CT triage system covering 14 conditions reported mean sensitivity of 97% and mean specificity of 98% in its pivotal study [13], and pooled reader studies of fracture-detection AI report sensitivity and specificity above 90%, with AI assistance improving unaided radiologist sensitivity by 6-8 percentage points without a corresponding loss of specificity [14].

**Figure 3. Reported sensitivity and specificity of representative FDA-cleared radiology AI diagnostic-support tools**




These figures should be read with two caveats. First, pivotal-study performance frequently exceeds real-world performance: an independent prospective registry study comparing three commercial fracture-detection tools on unselected clinical data found lower area-under-curve and sensitivity figures (for example, AUC 84.9% and sensitivity 79.5% for the best-performing tool) than vendor-reported pivotal-trial

statistics, underscoring the importance of local, prospective validation before deployment [15]. Second, algorithms validated predominantly on data from a single vendor's equipment or a narrow demographic range may under-perform when deployed on images from other scanners or in more diverse patient populations, a data-representativeness problem that disproportionately affects historically under-represented groups [9].

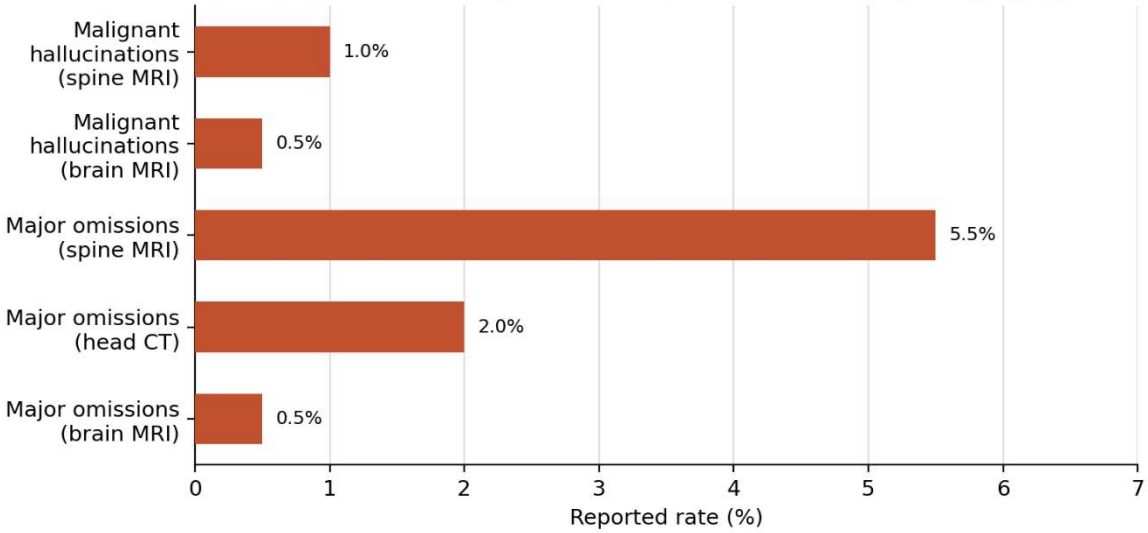
## IV. RISKS TO PATIENT SAFETY

### *A. Hallucinations and omissions*

Generative report-writing AI introduces a distinct failure mode largely absent from single-purpose detection algorithms: hallucination, defined as AI-generated content that is fictional, unfounded, or inconsistent with the underlying image or clinical input. In radiology-report generation, hallucinations manifest either as fabricated findings (for example, a vision-language model asserting a pleural effusion on a normal chest radiograph) or as critical omissions of true pathology, and both categories carry direct potential for patient harm [16][17]. A widely used clinical-impact taxonomy, the MediHall Score, stratifies hallucinations into catastrophic (fabrication or omission of a disease, scored zero for safety), critical (misjudging disease type), and attribute-level errors of lesser severity [18], a distinction that matters because not all AI errors carry equal clinical weight.

Quantitative evaluations give a concrete sense of scale. A dual-rater review of 500 AI-generated, patient-facing summaries of brain MRI, head CT, and spine MRI reports found major omissions in 0.5% of brain MRI, 2.0% of head CT, and 5.5% of spine MRI summaries, while malignant hallucinations — errors capable of altering clinical management — occurred in 0.5%, 0%, and 1.0% of the same report categories respectively [19]. In a prospective evaluation of a domain-specific multimodal generative AI model for chest-radiograph reporting, radiologists categorised unacceptable AI reports according to whether they contained false-positive or omitted findings, incorrect anatomical location, incorrect assessment of severity, or outright hallucination — including instances where the AI-generated report referenced prior imaging or CT findings that had never been supplied to the model [20]. Separately, evaluation of a general-purpose LLM asked to generate BI-RADS classifications and treatment recommendations from breast-ultrasound findings documented structured, fluent output that nonetheless contained frequent misdiagnoses, overdiagnoses, and internally inconsistent classifications [21].

**Figure 4. Documented rates of major omissions and malignant hallucinations in AI-generated report summaries, by imaging type**




These are not negligible rates in absolute terms: a 5.5% major-omission rate applied across a high-volume spine-MRI service reporting hundreds of studies per month implies a meaningful monthly count of clinically significant misses if AI output were accepted without independent radiologist verification. Detection tooling is emerging to address this gap — one hidden-state-based hallucination detector

reported an area-under-the-receiver-operating-characteristic-curve (AUROC) of 0.8751 across all findings and 0.8963 for clinically significant findings specifically, suggesting that automated hallucination screening is technically feasible but not yet a solved problem [16].

### *B. Automation bias and over-reliance*

A second, behavioural risk compounds the technical error rate: automation bias, the tendency of a human reviewer to over-trust an automated system's output over time. As radiologists observe an AI tool performing well across many cases, they may become less diligent in scrutinising studies the AI does not flag, or conversely give undue weight to an AI-flagged finding on an equivocal image that they might otherwise have attributed to artifact [22]. This dynamic is recognised in emerging regulation: Article 14 of the European Union's AI Act requires that high-risk AI systems be deployed in a way that keeps human overseers actively aware of the risk of automation bias, rather than assuming that human presence in the loop is by itself a sufficient safeguard [23].

Automation bias is difficult to counter because it is multifactorial — shaped by technical design, individual psychology, workload pressure, and institutional incentives simultaneously [23]. Experimental work with mock jurors illustrates how workflow design shapes both real and perceived accountability: in a scenario where a radiologist failed to detect a brain haemorrhage that an AI tool had correctly flagged, participants were substantially more likely to find the radiologist liable when the radiologist interpreted the scan only once, after seeing the AI output (74.7% found against the radiologist), compared with a double-read workflow in which the radiologist first interpreted the scan independently and only then reviewed the AI output (52.9%) [24][26]. This suggests that workflow sequencing — not only algorithmic accuracy — materially affects both patient safety and legal exposure.

### *C. Medico-legal liability and accountability*

Current malpractice doctrine in the United States generally continues to assign negligence liability to the treating physician even when an FDA-cleared AI tool contributed to an error, while liability exposure for the software vendor remains comparatively limited and legally heterogeneous across jurisdictions [25]. Documented cases include radiologists relying on FDA-cleared AI tools that nonetheless failed to detect early-stage lung nodules or fractures, leaving unresolved questions about whether responsibility should rest with the clinician, the institution, or the vendor [27]. Communicating a tool's known false-positive and false-negative rates to the radiologist — and, in some analyses, to the patient — has been shown experimentally to reduce perceived clinician culpability among mock jurors, reinforcing transparency as both an ethical and a medico-legal safeguard [24][26]. Because an AI system itself cannot be held legally accountable, radiologists who choose to integrate such tools into practice continue to bear the ultimate burden of clinical judgment, even as the tools shape that judgment in ways that are not always visible in the final signed report [28][29].

## V. DISCUSSION: TOWARD SAFE INTEGRATION

The evidence reviewed above does not support a binary verdict on AI-generated radiology reporting. Efficiency gains of 20-45% in reporting time, and diagnostic-support sensitivity and specificity frequently exceeding 90%, are real, reproducible, and clinically meaningful against a backdrop of a widening imaging-workforce gap [1][3][6][10][13]. At the same time, hallucination and major-omission rates of up to several percent, well-documented automation bias, and an unsettled liability landscape mean that unsupervised deployment of AI-generated reports would introduce a distinct and non-trivial category of patient-safety risk [19][22][25]. The two conclusions are compatible: AI-generated reporting is best understood not as a replacement for radiologist judgment but as a draft-and-verify workflow whose safety depends entirely on the rigour of the verification step.

| Risk Domain | Representative Finding | Mitigation Strategy |
|---|---|---|
| Fabricated / omitted findings | Up to 5.5% major omissions, 1.0% malignant hallucinations in spine MRI summaries [19] | Mandatory radiologist sign-off; automated hallucination screening (AUROC ~0.88-0.90) [16] |
| Automation bias | Radiologist diligence declines as trust in a well-performing AI tool grows over time [22][23] | Independent first read before viewing AI output; periodic blinded audits |
| Uneven real-world performance | Prospective registry AUC (84.9%) below vendor pivotal-trial figures [15] | Site-specific prospective validation before go-live; ongoing post-market surveillance |
| Liability ambiguity | Physician bears negligence liability despite AI involvement; vendor exposure limited [25][27] | Transparent disclosure of tool error rates; documented deviations from AI output [22] |
| Dataset / demographic bias | Tools may under-perform on scanners or populations outside training distribution [9] | Multi-vendor, multi-ethnic validation datasets; federated-learning consortia [9] |

Three governance priorities follow directly from this evidence base. First, human-in-the-loop review should be structurally protected rather than merely assumed: workflow design that preserves an independent first read, or that at minimum requires active radiologist engagement before report finalisation, measurably reduces both missed findings and perceived legal liability [24][26]. Second, hallucination-detection tooling and structured error taxonomies (such as the MediHall framework) should be integrated into deployment pipelines as a second line of defence, given that current AUROC performance of roughly 0.88-0.90 for clinically significant findings is promising but not yet sufficient to serve as a sole safeguard [16][18]. Third, regulatory and institutional frameworks need to catch up with the pace of deployment: most currently cleared tools are static, locked models, and adaptive systems that continue learning after deployment remain in a regulatory grey zone that existing FDA and EU frameworks were not originally designed to address [9][23].

## CONCLUSION

AI-generated radiology reports sit at a genuine inflection point. The efficiency case is empirically well-supported: across multiple prospective and retrospective studies, AI-assisted drafting reduces reporting time by roughly one-quarter to nearly one-half without measurable loss of report quality, and diagnostic-support algorithms now cover the large majority of FDA-cleared medical AI, with sensitivity and specificity frequently exceeding 90% for time-critical conditions such as intracranial haemorrhage, pulmonary embolism, and large-vessel-occlusion stroke. Against this, the patient-safety evidence is equally concrete: hallucination and major-omission rates, while low in absolute percentage terms, translate into clinically significant numbers of errors at scale, and automation bias plus unresolved medico-legal liability mean the human radiologist's verifying judgment remains the decisive safeguard rather than a formality. The path forward is not to choose between efficiency and safety but to engineer workflows, disclosure practices, and regulatory oversight that make the former conditional on the latter — preserving radiologist accountability and independent

clinical judgment as AI-generated drafting becomes a standard, rather than experimental, part of radiology practice.

## REFERENCES


1. Frontiers in Medicine (2025), "Utilization of large language models in decision-making for sustainability in radiology," Front. Med., 12:1632925.
2. Radiology Business (2026), "Radiology gets 68 new FDA-cleared algorithms," radiologybusiness.com, June 2026.
3. npj Digital Medicine (2025), "Keyword-based AI assistance in the generation of radiology reports: A pilot study," npj Digit. Med., 8, Nature Portfolio.
4. PubMed (2026), "Large language model-assisted radiology reporting in a single-radiologist implementation: a retrospective cohort study interpreted through a UTAUT lens," PMID:42009972.
5. IntuitionLabs (2025), "AI in Radiology: 2025 Trends, FDA Approvals & Adoption," intuitionlabs.ai.
6. Sun, X. et al., cited in medRxiv (2025), "Large Language Models in Radiology Reporting — A Systematic Review of Performance, Limitations, and Clinical Implications," medRxiv 2025.03.18.25324193.
7. European Radiology (2026), "Reporting efficiency in diagnostic imaging: Can plug-and-play general-purpose large language models outperform conventional speech recognition?," Eur. Radiol., Springer Nature.
8. Insights into Imaging (2026), "Enhancing radiology workflows through collaborative AI-assisted chest X-ray reporting using large vision-language models: a proof-of-concept study," Insights Imaging, Springer Nature.
9. ScienceDirect (2025), "Evaluating the Accuracy and Efficiency of AI-Generated Radiology Reports Based on Positive Findings — A Qualitative Assessment of AI in Radiology," Clin. Imaging, Elsevier.
10. Pinggy Blog (2026), "AI Medical Imaging in 2026: Best Radiology AI Tools, FDA Clearances, and Diagnostic Accuracy," pinggy.io.
11. medRxiv (2023), "A retrospective analysis of the diagnostic performance of an FDA approved software for the detection of intracranial hemorrhage," medRxiv 2023.11.02.23297974.
12. arXiv (2026), "Human-AI Collaboration in Radiology: The Case of Pulmonary Embolism," arXiv:2601.13379.
13. PMC (2023), "Retrospective batch analysis to evaluate the diagnostic accuracy of a clinically deployed AI algorithm for the detection of acute pulmonary embolism on CTPA," PMC10244304.
14. PMC (2025), "Artificial Intelligence in Fracture Diagnosis on Radiographs: Evidence, Pitfalls, and Pathways for Clinical Integration (2020-2025)," PMC12551796.
15. AZmed (2025), "How to evaluate FDA- or CE-Cleared AI for routine radiography," azmed.co, citing Luiken et al., "Evaluation of commercial AI algorithms for the detection of fractures, effusions, and dislocations on real-world clinical data: A prospective registry study."
16. arXiv (2025), "ReXTrust: A Model for Fine-Grained Hallucination Detection in AI-Generated Radiology Reports," arXiv:2412.15264.
17. arXiv (2026), "Hallucination in Medical Imaging AI: A Cross-Modality Analytical Framework for Taxonomy, Detection, and Mitigation under Regulatory Constraints," arXiv:2606.13211.
18. PMC (2025), "Mitigating hallucinations in healthcare AI: a systematic review of evidence-based strategies," PMC13470931.
19. Nature Scientific Reports (2026), "Patient-friendly simplification and translation of neuroradiology impressions using artificial intelligence," Sci. Rep., Nature Portfolio.
20. Radiology, RSNA (2024), "Diagnostic Accuracy and Clinical Value of a Domain-specific Multimodal Generative AI Model for Chest Radiograph Report Generation," Radiology, 10.1148/radiol.241476.
21. medRxiv (2025), cited in "Large Language Models in Radiology Reporting — A Systematic Review," medRxiv 2025.03.18.25324193 (Sun et al., GPT-3.5 BI-RADS evaluation).
22. ScienceDirect (2026), "Automation bias and overconfidence in artificial intelligence and associated legal implications," Curr. Probl. Diagn. Radiol., Elsevier.
23. arXiv (2025), "Automation Bias in the AI Act: On the Legal Implications of Attempting to De-Bias Human Oversight of AI," arXiv:2502.10036.

24. Nature Health (2026), "The radiologist-AI workflow and the risk of medical malpractice claims," Nat. Health, Springer Nature.
25. AJMC (2026), "FAQs About AI in Radiology: Legal Risks, Liability, and Malpractice," ajmc.com.
26. medRxiv (2025), "Radiologist-AI workflow can be modified to reduce the risk of medical malpractice claims," medRxiv 2025.06.14.25329278.
27. MDPI, Diagnostics (2025), "Governing Artificial Intelligence in Radiology: A Systematic Review of Ethical, Legal, and Regulatory Frameworks," Diagnostics, 15(18):2300.
28. PMC (2024), "AI in Radiology: Navigating Medical Responsibility," PMC11276428.

HOW TO CITE: Adil Ahmad Wani[1], Vyshak M.[2], Usha G.[3], Junaid ul Islam[4], Zakir Hussain Parray[5]*, AI-Generated Radiology Reports: Opportunities For Efficiency, Diagnostic Support, And Risks To Patient Safety, Int. J. Sci. R. Tech., 2026, 3 (9), 367-375. https://doi.org/ 10.5281/zenodo.22896938