Review

# Effectiveness, Safety, and Workflow Burden of Large Language Model–Based Medical Report Generation: Systematic Review

Jie-Lin Huang, MD; Ji-Qing Zhu, MD, PhD; Xiao-Guang Ni, MD, PhD

Department of Endoscopy, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China

**Corresponding Author:**

Xiao-Guang Ni, MD, PhD
Department of Endoscopy, National Cancer Center/National Clinical Research Center for Cancer/Cancer Hospital
Chinese Academy of Medical Sciences and Peking Union Medical College
No. 17 Panjiayuan South Lane, Chaoyang District
Beijing 100021
China
Phone: +86 10 8778 7606
Email: nixiaoguang@126.com

## Abstract

**Background:** Systems based on large language models (LLMs), multimodal LLMs, and vision-language foundation models are increasingly being evaluated for medical report generation in imaging and related clinical workflows. Existing reviews have summarized technical architectures, radiology applications, readability, and benchmark performance, but clinical readiness remains uncertain because safety, human oversight, and workflow outcomes are sparsely and inconsistently reported.

**Objective:** The aim of this study is to assess the effectiveness (expert acceptance and blinded preference), safety (clinically significant, omission, and commission errors), and workflow burden (reporting time, corrections, edit distance, and editing burden) of LLM-based medical report generation.

**Methods:** We searched PubMed/MEDLINE, Embase, Web of Science Core Collection, Scopus, and the Cochrane Library for studies published from January 1, 2016, through May 15, 2026. Eligible studies evaluated LLMs, multimodal LLMs, or vision-language foundation models for image-to-report generation, impression generation from findings, report drafting, or structured reporting in imaging workflows. Two reviewers performed screening, extraction, risk-of-bias assessment, and Grading of Recommendations Assessment, Development, and Evaluation–informed narrative certainty assessment. Outcomes were clinically significant error rate, omission error rate, commission error rate, reporting time, edit burden, expert acceptance, and blinded expert preference. Meta-analysis was not performed because no comparable outcome had at least 2 studies with compatible task structure and analyzable data.

**Results:** A total of 101 studies were included. Chest x-ray was the largest modality group (36 studies), followed by computed tomography, magnetic resonance imaging (MRI), ultrasound, endoscopy, pathology, ophthalmic, electrocardiographic, dental, and mixed-modality contexts. No study was judged at low risk of bias; 15 were moderate, 72 high, and 14 serious. Safety and workflow evidence remained heterogeneous and largely nonpoolable. In a chest x-ray study, AI report acceptance was similar to that of radiologist reports (6047/8580, 70.5% vs 6288/8580, 73.3%), but false-negative findings were slightly higher (1584/8580, 18.5% vs 1527/8580, 17.8%). In a clinician-collaboration chest x-ray study, AI reports were equivalent or preferred in 233 of 300 (77.7%) and 170 of 303 (56.1%) cases across 2 datasets; yet, clinically significant errors persisted. In a brain MRI study, AI assistance reduced reading time from 61 to 53 seconds, whereas impression drafting increased editing time and edit distance.

**Conclusions:** This review shifts the synthesis from plausible report generation to clinically interpretable effectiveness, safety, and workflow effects. Expert acceptance and preference suggested assistive value in selected supervised settings, but these signals were limited by inconsistent reporting of clinically significant errors, omissions, commissions, and failed generations. Workflow effects were mixed, with some studies reporting shorter reading time or drafting support, and others reporting greater editing time or edit distance. The evidence remains too heterogeneous, biased, and sparse on case-level end points to support a pooled meta-analysis or autonomous clinical-readiness claims. Adoption should remain locally validated,

clinician-supervised, and accompanied by standardized reporting of acceptance, preference, omissions, commissions, failed generations, reporting time, corrections, and editing burden.

**Trial Registration:** PROSPERO CRD420261302844; https://www.crd.york.ac.uk/PROSPERO/view/CRD420261302844

***J Med Internet Res 2026;28:e97007;*** doi: 10.2196/97007

**Keywords:** large language models; generative AI; medical report generation; radiology; pathology; endoscopy; workflow; systematic review

## Introduction

Medical report generation is a central bottleneck in imaging practice. In radiology, pathology, and endoscopy, clinicians must convert findings into coherent, clinically actionable reports under rising volume, workforce pressure, fatigue, and time constraints. Clinical documentation consumes substantial physician time, and reporting work is a major part of the information-transfer burden in image-based specialties [1]. Reports are also safety-critical communication artifacts: omissions, unsupported statements, ambiguous impressions, and delays in finalization can affect downstream clinical decisions. This reporting burden has therefore made automation and decision support an important target.

Earlier medical imaging report generation research predated the current large language model (LLM) wave and relied on neural architectures designed for specific text generation tasks rather than general purpose foundation models [2,3]. More recent systems based on LLMs have been reported for radiology impression drafting and lumbar spine magnetic resonance imaging (MRI) reporting [4,5], chest radiography and computed tomography (CT) report generation [6,7], endoscopy reporting [8,9], multimodal clinical or PACS (picture archiving and communication system)–linked applications [10,11], oral radiology and multimodal spinal MRI tasks [12,13], and ultrasound reporting [14]. These systems differ substantially in their inputs and intended roles: some generate complete reports from images, some draft impressions from existing findings, and others restructure or edit clinician-authored text. Treating these workflows as a single intervention can obscure important differences in risk, human oversight, and measurable benefit. Outcome interpretation, therefore, requires task-level classification rather than generic aggregation of all report-generation systems.

The central translational question remains unresolved: do these systems improve clinical reporting, or do they generate plausible but unsafe text? The literature is broad but fragmented across endoscopy [8,9], mixed-modality or workflow-linked settings [10,11], chest x-ray and cross-sectional imaging [15,16], pathology [17,18], and workflows ranging from autonomous generation to drafting, editing, and structuring support. This breadth does not yet translate into coherent clinical inference because studies differ in clinical input, generation task, comparator, human oversight, and outcome definition. A study that reports linguistic similarity to a reference report does not answer the same question as a study that measures missed findings, hallucinated statements, clinician editing burden, or final report acceptance.

The safety and workflow context for LLM-based report generation is broader than report text quality alone. Automation bias can reduce human verification of decision-support outputs when users overrely on automated suggestions, especially when verification is complex [19]. Clinical AI safety literature also emphasizes that bias, distribution shift, unclear failure modes, and insufficient monitoring can create risks even when model performance appears favorable in development settings [20]. For LLMs specifically, medical use raises concerns about fluent but unsupported outputs [21], and hallucination is a recognized limitation of natural language generation systems [22]. Recent radiology guidance on generative AI similarly emphasizes human oversight, local validation, monitoring, and workflow-specific safeguards before clinical deployment [23]. These risks make safety and workflow end points central to assessing clinical usefulness rather than secondary technical details.

Existing reviews show why a new synthesis is needed now. Reviews and commentaries have described the rapid emergence of AI report generation, prompt engineering and fine-tuning, structured reporting, foundation models in imaging, and endoscopy applications [24-28]. More recent systematic and scoping reviews have summarized automated radiology report-generation methods, structured-reporting applications, and deep learning trends [29-31], while others have focused on LLM use in radiology reports, report readability or simplification, and broader radiology adoption trajectories [32-34]. These reviews are valuable, but generally emphasize technical progress, radiology-centered use cases, readability, or benchmark-oriented evaluation. They do not resolve whether the peer-reviewed evidence base reports on clinically interpretable safety and workflow outcomes consistently enough to guide supervised implementation or cumulative clinical inference. In particular, prior syntheses have not consistently separated autonomous image-to-report generation from supervised drafting or report-structuring support, although those categories carry different clinical responsibilities and failure modes.

This gap is important because technical evaluation and clinical decision-making require different evidence. Many studies still emphasize benchmark language metrics such as BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), CIDEr (Consensus-Based Image Description Evaluation), and BERTScore (a BERT-based semantic similarity metric, where BERT stands for Bidirectional Encoder Representations from Transformers) [35-38], while clinically important outcomes are reported less consistently. For this review, effectiveness was defined as expert acceptance and

blinded expert preference; safety as clinically significant error rate, omission error rate, and commission error rate; and workflow burden as reporting time, number of corrections, edit distance, and editing burden. Here, omission errors are missed clinically relevant findings, and commission errors are unsupported or hallucinated findings introduced into the report. Benchmark similarity metrics help technical development but do not show whether a generated report missed a critical lesion, introduced a misleading statement, or reduced physicians' finalization effort. Image-aware evaluation frameworks have begun to address this gap, but they remain uncommon and unharmonized for clinical synthesis [39]. In addition, many reports provide reader ratings or preference judgments without case-level denominators, independent validation, or a clear account of failed generations. Consequently, individual positive studies do not yet establish whether the field is approaching clinical readiness or overstating progress through benchmark-dominated evaluation.

The objective of this systematic review was to assess the effectiveness (expert acceptance and blinded preference), safety (clinically significant errors, omission errors, and commission errors), and workflow burden (reporting time, corrections, edit distance, and editing burden) of LLM-based medical report generation across imaging modalities and related clinical reporting contexts. To support clinically meaningful interpretation, we separated task types and levels of human oversight so that autonomous image-to-report generation, supervised drafting, impression generation, and report-structuring support were not treated as equivalent interventions.

# Methods

## Protocol and Registration

This systematic review was conducted in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 statement (Checklist 1) [40]. Literature search reporting followed the PRISMA-S (Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension) [41], and narrative synthesis was reported with attention to the Synthesis Without Meta-Analysis (SWiM) guidance [42]. The review question, primary outcomes, and subgroup structure were specified a priori, and the review was registered with PROSPERO (International Prospective Register of Systematic Reviews) 2026 (CRD420261302844) before formal screening began. The prespecified review protocol and extraction backbone are available from the corresponding author upon reasonable request. The PRISMA-S checklist was completed item by item in the Checklist 1, and when a PRISMA-S item was not part of the search methodology or was not applicable, it was explicitly marked and justified.

## Eligibility Criteria

Eligible studies evaluated generative AI systems based on LLMs for generating an imaging report, report draft, or diagnostic impression from medical imaging, pathology whole-slide imaging, endoscopy, or multimodal, clinically relevant inputs. For eligibility, an intervention was considered eligible if an LLM, multimodal LLM, vision-language foundation model, or a language model component contributed directly to clinical report text generation, diagnostic impression generation, or clinical report drafting. Image classifiers, segmentation tools, retrieval systems, and image captioning systems were not eligible unless a language model or foundation model contributed directly to report generation.

Prespecified task types were report generation from images, generation of impressions from findings, and report drafting. Human oversight was categorized as autonomous generation, human-supervised drafting, or report editing or structuring support. Autonomous generation referred to systems that generated reports or report outputs from images or clinical inputs before human evaluation. Human-supervised drafting referred to workflows in which clinicians actively used, or iteratively collaborated on, AI-generated text during drafting. Report editing or structuring support referred to settings in which AI was applied after source report text, findings, or draft content was already available. We included randomized and nonrandomized comparative studies, reader studies, prospective workflow studies, retrospective validation studies, and single-arm technical validation studies if they used an explicit expert or clinical reference standard. Conference abstracts, reviews, editorials, letters without original data, dissertations, protocols, and preprints were excluded. Because preprints were excluded, this review was designed to evaluate peer-reviewed clinical evidence rather than the absolute frontier of technical model capability.

Effectiveness outcomes were expert acceptance rate and blinded expert preference. Safety outcomes were clinically significant error rate, omission error rate, and commission error rate. Workflow burden outcomes were reporting time, number of corrections, edit distance, and editing burden after generation. Omission errors were defined as missed clinically relevant findings, whereas commission errors were defined as unsupported or hallucinated findings introduced into the generated report. Exploratory outcomes included discrepancy rates between findings and impressions, and text similarity or language quality metrics, such as BLEU, ROUGE, CIDEr, and BERTScore.

## Information Sources

We searched PubMed/MEDLINE, Embase, Web of Science Core Collection, Scopus, and the Cochrane Library from January 1, 2016, through May 15, 2026. Europe PMC, OpenAlex, and reference lists of eligible studies and relevant reviews were used as supplemental sources through May 17, 2026, to identify additional records. Study registries were not used as bibliographic search sources because the eligibility criteria targeted completed peer-reviewed studies of medical report generation; this nonuse is reported explicitly in the PRISMA-S checklist. For studies that reported clinically important workflow or safety outcomes in nonconvertible formats, corresponding authors were contacted to request additional analyzable summary statistics or case-level denominator clarification.

## Search Strategy

The database strategy combined controlled vocabulary and free-text terms related to LLMs, multimodal LLMs, vision-language models, foundation models, and generative AI; report generation, report drafting, diagnostic impression generation, and structured reporting; and imaging domains including radiology, CT, MRI, radiography, pathology whole-slide imaging, ultrasound, endoscopy, and other medical image or signal interpretation settings. Exact database-specific search strings, interfaces, limits, dates, and record counts for PubMed/MEDLINE, Embase, Web of Science Core Collection, Scopus, and the Cochrane Library are provided in Multimedia Appendix 1; the completed PRISMA-S item-level reporting checklist is provided in the reporting guideline checklist file. Database searches were first completed on March 18, 2026, and were updated in May 2026 before final analysis. Across the original and updated database searches, 6936 records were identified before deduplication. Records retrieved in both searches were deduplicated by DOI, PMID, or normalized title and were counted once. After removal of 3008 duplicate records and 499 records already represented in the search corpus, 3429 database records remained for title and abstract screening.

## Study Selection

Title and abstract screening and full-text eligibility assessment were performed independently by 2 reviewers after deduplication, with disagreements resolved by discussion or adjudication by a third reviewer. No interrater reliability coefficient was generated; therefore, reviewer agreement is described procedurally rather than as a kappa statistic. Full-text exclusions were logged with one dominant reason per report.

## Data Extraction

Extraction frameworks at the study and outcome levels were used, and key extraction fields were checked in duplicate for studies that contributed clinically important safety or workflow outcomes. Extracted variables included study characteristics, clinical specialty, imaging modality, task type, level of human oversight, study design, validation setting, model characteristics, comparator type, reference standard, outcome definitions, and raw data sufficient for effect-size computation when available.

## Risk of Bias and Certainty Assessment

Risk of bias was assessed according to study design. Comparative and workflow studies were appraised using ROBINS-I (Risk of Bias in Non-Randomized Studies of Interventions) logic [43], with reporting appraisal informed by CLAIM (Checklist for Artificial Intelligence in Medical Imaging) as an adjunct for assessing AI reporting quality [44]. Retrospective technical validation studies that were not well captured by conventional intervention tools were assessed using a prespecified custom AI validation risk-of-bias framework. The core domains of that framework were representativeness of case selection, reference standard quality, blinding of expert assessment, validation independence, data leakage risk, model or prompt transparency, handling of failed or missing generations, and risk from outcome definition or selective reporting.

For the custom framework, moderate concern was assigned when a limitation was present but limited in scope or of limited consequence for clinical interpretation. High concern was assigned when a limitation was important and likely to influence interpretation, such as unclear validation independence, incomplete blinding, limited external validation, or incomplete failure reporting. Serious concern was assigned when a limitation could directly undermine clinical validity, such as likely data leakage, clearly nonrepresentative case selection, materially weak or unclear reference standards, the absence of handling of failed generations when failures could alter outcome interpretation, or strongly selective reporting of clinically important outcomes. The framework was used as a structured, qualitative, domain-based appraisal rather than a formally validated measurement instrument, and no interrater reliability coefficient was generated for this review. Its judgments should, therefore, be interpreted as transparent qualitative risk signals rather than as validated scale scores.

Because no outcomes met the criteria for meta-analysis, certainty was judged using a narrative certainty assessment informed by GRADE (Grading of Recommendations Assessment, Development, and Evaluation) domains rather than rating pooled effect estimates [45]. We summarized these judgments in a GRADE Summary of Findings table format, using outcome-domain study counts from the selected clinically interpretable end point studies, and reported relative effects as not pooled or not estimable when compatible effect structures were unavailable.

## Data Synthesis

Eligibility for quantitative synthesis was assessed at the outcome level according to clinical coherence, task structure, comparator or reference standard design, and availability of analyzable event counts or summary statistics. For outcomes amenable to direct quantitative reconstruction, planned effect measures were risk differences for dichotomous outcomes and mean differences for continuous outcomes, with 95% CIs when derivable from published data. Outcomes reported across reader-case pairs were not treated as independent case-level events when clustering could not be reconstructed.

Quantitative synthesis was prespecified only for outcomes reported by at least 2 studies with compatible definitions, task structures, comparator or reference standard designs, and analyzable data structures. Outcomes that did not meet these criteria were summarized using structured narrative synthesis. When individual studies provided sufficient data, targeted quantitative reconstruction was performed to support interpretation without pooling. Narrative synthesis was organized by modality, task type, outcome domain, and level of human oversight. Formal subgroup, sensitivity, heterogeneity, and reporting bias analyses were reserved for poolable synthesis sets.

## Deviations From Protocol

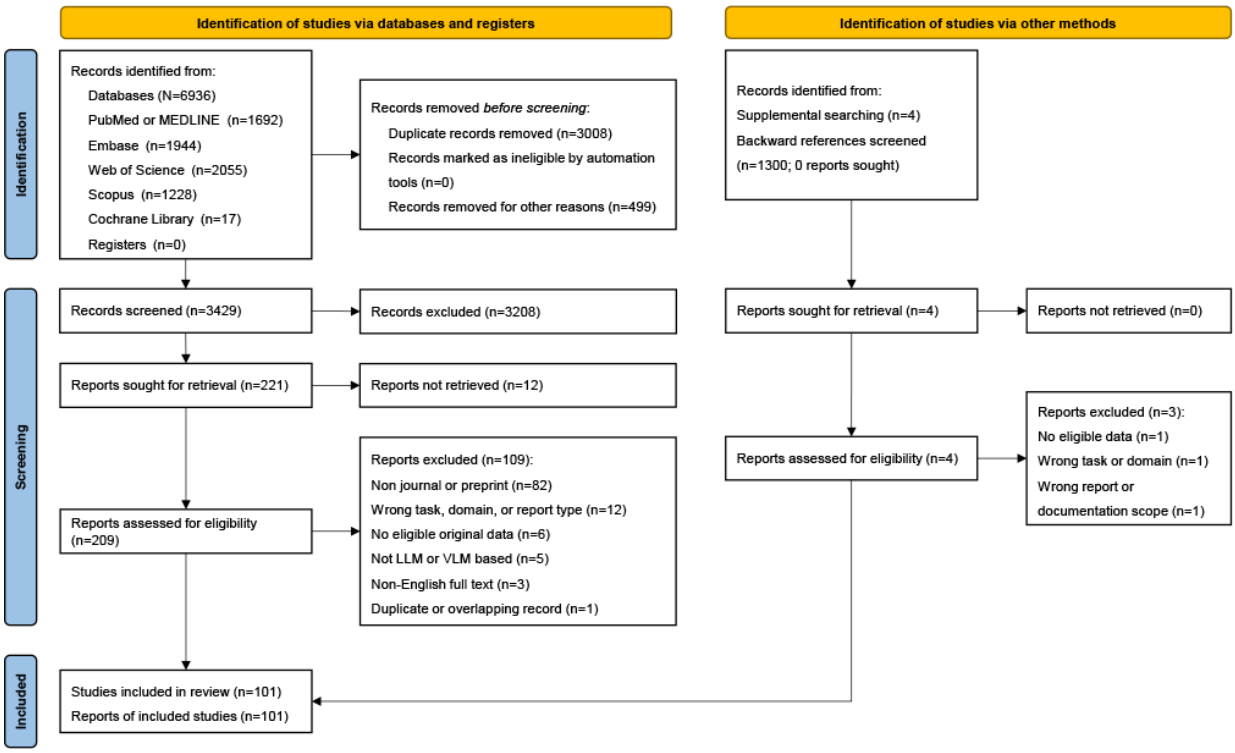
The protocol anticipated quantitative synthesis when clinically and methodologically compatible synthesis sets were available. Supplemental citation discovery and reference screening were added to improve retrieval completeness. These procedures did not change the review question, primary outcomes, eligibility framework, or planned subgroup structure. When outcomes did not meet pooling criteria, synthesis followed the prespecified nonpooled approach of structured narrative synthesis with targeted quantitative reconstruction of individual studies where defensible.

# Results

## Study Selection

The full study selection process is shown in Figure 1. Across the original and updated database searches, 6936 records were identified before deduplication. After removal of 3008 duplicate records and 499 records already represented in the search corpus, 3429 database records proceeded to title and abstract screening. Public supplemental screening and citation-discovery screening contributed 1 additional eligible study, whereas backward reference screening did not identify additional reports for full-text assessment. Overall, 101 studies met the inclusion criteria for this systematic review. Study-level records are provided in Supplementary Table S1 in Multimedia Appendix 2, and full-text exclusions are listed in Supplementary Table S2 in Multimedia Appendix 2.

**Figure 1.** PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) 2020 flow diagram for the systematic review. The figure summarizes database searching through May 15, 2026, other-method identification through supplemental citation discovery and backward reference screening, full-text assessment, and final inclusion of 101 studies. LLM: large language model; VLM: vision-language model.




## Study Characteristics

The included literature was broad in technical scope but remained concentrated in a limited set of clinical scenarios (Table 1). Of the 101 included studies, 68 (67.3%) were retrospective or technical validation studies, 28 (27.7%) were nonrandomized comparative, reader, or workflow studies, and 5 (5.0%) were prospective studies or real-world workflow or validation studies. Chest x-ray remained the largest modality group (36, 35.6%), followed by mixed-modality settings (20, 19.8%), CT, CT angiography, or positron emission tomography/CT (11, 10.9%), MRI (10, 9.9%), ultrasound (6, 5.9%), endoscopy (5, 5.0%), pathology reporting settings (5, 5.0%), other x-ray or radiography (5, 5.0%), ophthalmic imaging (2, 2.0%), and electrocardiogram (1, 1.0%). By task type, 61 (60.4%) studies evaluated report generation from images, 29 (28.7%) evaluated report drafting or structured report generation, and 11 (10.9%) evaluated impression or diagnostic conclusion generation from findings or reports. By human oversight, 66 (65.3%) studies were classified as autonomous

generation, 15 (14.9%) as human-supervised drafting, and 20 (19.8%) as report editing or structuring support. The relationship among modality, task type, and human oversight is shown in Figure 2, and the rapid expansion of the literature after 2023 is summarized in Figure 3A.

**Table 1.** Evidence profile of included studies.

| Domain and category | Studies, n (%) |
|---|---|
| Overall | |
| Total included studies | 101 |
| Study design | |
| Retrospective or technical validation | 68 (67.3) |
| Nonrandomized comparative, reader, or workflow study | 28 (27.7) |
| Prospective or real-world workflow or validation study | 5 (5.0) |
| Clinical source | |
| Chest x-ray | 36 (35.6) |
| Mixed modality | 20 (19.8) |
| CT[a], CTA[b], or PET/CT[c] | 11 (10.9) |
| MRI[d] | 10 (9.9) |
| Ultrasound | 6 (5.9) |
| Endoscopy | 5 (5.0) |
| Pathology report or whole-slide image | 5 (5.0) |
| Other x-ray or radiography | 5 (5.0) |
| Ophthalmic imaging | 2 (2.0) |
| Electrocardiogram | 1 (1.0) |
| Generation task | |
| Image to report generation | 61 (60.4) |
| Report drafting or structuring | 29 (28.7) |
| Findings or report to impression generation | 11 (10.9) |
| AI role in reporting workflow | |
| Autonomous generation | 66 (65.3) |
| Human-supervised drafting | 15 (14.9) |
| Report editing or structuring support | 20 (19.8) |
| Overall risk of bias | |
| Low | 0 (0) |
| Moderate | 15 (14.9) |
| High | 72 (71.3) |
| Serious | 14 (13.9) |

[a]CT: computed tomography.
[b]CTA: computed tomography angiography.
[c]PET: positron emission tomography.
[d]MRI: magnetic resonance imaging.

**Figure 2.** Sankey diagram linking clinical modality or report source, report generation task, and AI role in the reporting workflow for the 101-study evidence base. Flow width is proportional to the number of studies in each pathway. The figure shows continued concentration in x-ray or radiography, report generation from images, and autonomous generation, while also showing smaller evidence streams in computed tomography (CT), computed tomography angiography (CTA), or positron emission tomography (PET)/CT; magnetic resonance imaging (MRI); ultrasound; endoscopy; pathology; ophthalmic imaging or electrocardiogram (ECG); mixed modality; human-supervised drafting; and report editing or structuring workflows.

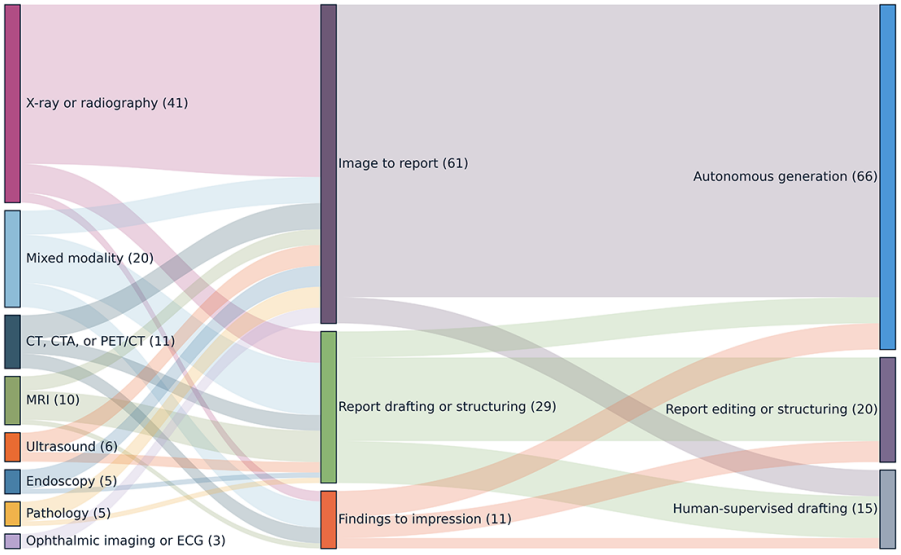
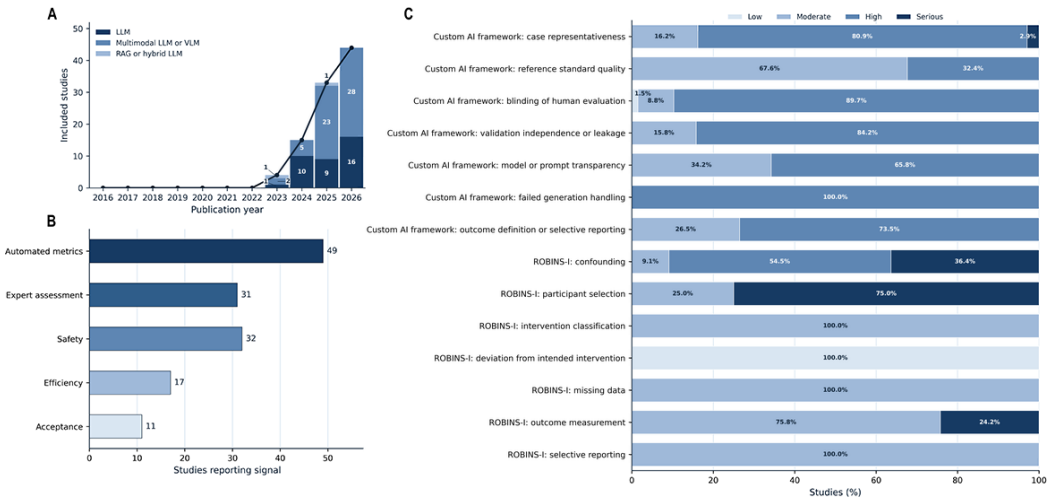



**Figure 3.** Temporal evolution, evaluation signals, and domain-level risk of bias. Panel A summarizes annual counts among studies for which a publication year was captured, stratified by broad model family or by multimodal or hybrid grouping. Panel B displays nonexclusive standardized evaluation signal categories across the 101-study evidence base; individual automated metric names were grouped under a single automated metrics category because studies used heterogeneous benchmark measures. Panel C summarizes domain-level risk-of-bias judgments across custom AI validation domains (n=68) and ROBINS-I-informed domains (n=33). LLM: large language model; RAG: retrieval-augmented generation; ROBINS-I: Risk of Bias in Non-Randomized Studies of Interventions; VLM: vision-language model.




Outcome reporting was imbalanced. Benchmark metrics such as BLEU, ROUGE, CIDEr, and BERTScore were common, especially in chest x-ray and technical validation studies, whereas clinically important outcomes, including significant errors, omissions, commissions, expert acceptance, blinded preference, reporting time, and edit burden, were inconsistently reported and used heterogeneous denominators. Thirty-six studies reported at least one clinically relevant human evaluation, safety, workflow, report quality, acceptability, or time outcome, but these outcomes differed by task, comparator, denominator, and reporting format. Benchmark-heavy chest x-ray studies were represented among the included studies [46-50]. Standardized evaluation signal reporting is summarized in Figure 3B. The subset of studies with clinically interpretable effectiveness, safety, and workflow-burden end points is summarized in Table 2.

**Table 2.** Selected studies with clinically interpretable evaluation end points[a].

| Study and year | Clinical source | Generation task | AI role in reporting workflow | Assessment method | Reported evaluation outcomes |
|---|---|---|---|---|---|
| Hong et al [51], 2025 | Chest x-ray | Image to report | Autonomous generation | Human reader assessment | Acceptance, safety |
| Tanno et al [52], 2025 | Chest x-ray | Image to report | Human-supervised drafting | Human reader assessment | Acceptance, safety |
| Wang et al [53], 2026 | MRI[b] | Findings to impression | Report editing or structuring support | Human reader assessment | Accuracy, efficiency |
| Serapio et al [54], 2024 | Mixed modality | Findings to impression | Human-supervised drafting | Workflow comparison | Efficiency |
| Li et al [55], 2025 | CT[c] | Image to report | Autonomous generation | Human expert assessment | Report quality, accuracy |
| Luo et al [56], 2026 | Endoscopy | Image to report | Autonomous generation | Human expert assessment | Report quality, accuracy |
| Ayaz et al [57], 2025 | Mixed modality | Image to report | Autonomous generation | Human expert assessment | Report quality |
| Li et al [58], 2026 | Mixed modality | Findings to impression | Human-supervised drafting | Clinical validation | Report quality, accuracy, efficiency |
| de Margerie-Mellon et al [59], 2026 | Mixed modality | Report drafting or structuring | Human-supervised drafting | Workflow comparison | Safety, efficiency |
| Jiang et al [60], 2026 | Endoscopy | Image to report | Autonomous generation | Clinical validation | Safety, accuracy, efficiency |
| Tekdemir et al [61], 2026 | CT | Findings to impression | Autonomous generation | Human expert assessment | Report quality, safety |
| Lorusso et al [62], 2026 | CTA[d] | Findings to impression | Autonomous generation | Reference standard comparison | Accuracy, efficiency |
| Dong et al [63], 2025 | MRI | Report drafting or structuring | Human-supervised drafting | Human expert assessment | Report quality, accuracy, efficiency |
| Pellegrini et al [64], 2026 | Chest x-ray | Image to report | Human-supervised drafting | Workflow comparison | Report quality, efficiency |
| Tan [65], 2026 | Mixed modality | Report drafting or structuring | Human-supervised drafting | Implementation assessment | Acceptance, efficiency |
| Wang et al [66], 2024 | Ultrasound | Report drafting or structuring | Report editing or structuring support | Human expert assessment | Clinician feedback, accuracy, efficiency |
| Hong et al [67], 2026 | Chest x-ray | Image to report | Autonomous generation | Robustness and sensitivity analysis | Acceptance, safety, accuracy |
| Zanardo et al [68], 2026 | MRI | Report drafting or structuring | Autonomous generation | Human expert assessment | Report quality |
| Xie et al [69], 2025 | MRI | Report drafting or structuring | Report editing or structuring support | Human expert assessment | Acceptance, accuracy, efficiency |
| Woźnicki et al [70], 2025 | Chest x-ray | Report drafting or structuring | Report editing or structuring support | Reference standard comparison | Report quality, accuracy |
| Shao et al [71], 2025 | Ophthalmic imaging | Image to report | Autonomous generation | Human expert assessment | Report quality |

[a]Categories are standardized for scanability and do not imply comparable outcome definitions across studies. Human-supervised drafting denotes AI involvement during report drafting, whereas report editing or structuring support denotes AI use after source report text, findings, or draft content were already available.
[b]MRI: magnetic resonance imaging.
[c]CT: computed tomography.
[d]CTA: computed tomography angiography.

## Risk of Bias and Certainty of Evidence

The risk-of-bias assessment framework is provided in Supplementary Table S3A in Multimedia Appendix 2. Overall risk of bias was moderate in 15 studies, high in 72, and serious in 14; no study was at low risk (Table 3). Although the evidence base included multicenter, external validation, and prospective validation studies, most studies were vulnerable to retrospective selection, incomplete external validation, nonrandomized workflow or reader comparisons, incomplete failure reporting, limited transparency in prompting or model handling, and selective emphasis on benchmark outcomes. Recurrent custom-framework

concerns involved validation independence or leakage risk, absent or unclear blinding, heterogeneous outcome definitions, and poor accounting for failed or unusable reports. Risk-of-bias judgments are summarized in Table 3; complete study-level judgments are provided in Supplementary Table S3B in Multimedia Appendix 2; and domain-level bias patterns are shown in Figure 3C.

**Table 3.** Risk-of-bias summary by assessment domain[a].

| Assessment set and domain | Risk-of-bias judgment distribution, n/N (%) |
|---|---|
| All studies | |
| Overall | Low: 0; moderate: 15/101 (14.9); high: 72/101 (71.3); serious: 14/101 (13.9) |
| Custom AI validation+CLAIM[b] | |
| Overall | Low: 0; moderate: 12/68 (17.6); high: 54/68 (79.4); serious: 2/68 (2.9) |
| Case selection | Low: 0; moderate: 11/68 (16.2); high: 55/68 (80.9); serious: 2/68 (2.9) |
| Reference standard | Low: 0; moderate: 46/68 (67.6); high: 22/68 (32.4); serious: 0 |
| Blinding | Low: 1/68 (1.5); moderate: 6/68 (8.8); high: 61/68 (89.7); serious: 0 |
| Validation or leakage | Low: 0; moderate: 9/68 (13.2); high: 57/68 (83.8); serious: 2/68 (2.9) |
| Model or prompt transparency | Low: 0; moderate: 41/68 (60.3); high: 27/68 (39.7); serious: 0 |
| Failure handling | Low: 0; moderate: 0; high: 66/68 (97.1); serious: 2/68 (2.9) |
| Outcome reporting | Low: 0; moderate: 18/68 (26.5); high: 50/68 (73.5); serious: 0 |
| ROBINS-I[c]+CLAIM | |
| Overall | Low: 0; moderate: 3/33 (9.1); high: 18/33 (54.5); serious: 12/33 (36.4) |
| Confounding | Low: 0; moderate: 3/33 (9.1); high: 18/33 (54.5); serious: 12/33 (36.4) |
| Participant selection | Low: 0; moderate: 23/33 (69.7); high: 0; serious: 10/33 (30.3) |
| Intervention classification | Low: 0; moderate: 33/33 (100); high: 0; serious: 0 |
| Intervention deviations | Low: 33/33 (100); moderate: 0; high: 0; serious: 0 |
| Missing data | Low: 0; moderate: 33/33 (100); high: 0; serious: 0 |
| Outcome measurement | Low: 0; moderate: 25/33 (75.8); high: 0; serious: 8/33 (24.2) |
| Reported result selection | Low: 0; moderate: 33/33 (100); high: 0; serious: 0 |

[a]This table summarizes risk-of-bias judgments across the 101-study evidence base. Complete study-level judgments are provided in Supplementary Table S3B in Multimedia Appendix 2.
[b]CLAIM: Checklist for Artificial Intelligence in Medical Imaging.
[c]ROBINS-I: Risk of Bias in Non-Randomized Studies of Interventions.

Certainty was judged very low across all 4 clinically important outcome domains in the GRADE Summary of Findings table (Table 4). This rating does not mean that all findings were clinically unfavorable; rather, it reflects a high or serious risk of bias, incompatible outcome definitions, indirectness across task types and clinical sources, sparse or nonreconstructable effect data, and concern that clinically unfavorable or failed report-generation outputs may be incompletely reported.

**Table 4.** GRADE[a] summary of findings for clinically important outcome domains[b].

| Outcome | Studies (n) | Study design | Risk of bias | Inconsistency | Indirectness | Imprecision | Other considerations | Impact | Certainty of the evidence (GRADE) | Importance |
|---|---|---|---|---|---|---|---|---|---|---|
| Expert acceptance | 6 | Nonrandomized studies | Serious | Serious | Serious | Serious | Strongly suspected publication or reporting bias | Not pooled. Acceptance measures varied; one clustered reader study reported acceptance of 70.5% (6047/8580) for AI-generated reports compared with 73.3% (6288/8580) for radiologist reports. | Very low | Important |

| Outcome | Studies (n) | Study design | Risk of bias | Inconsistency | Indirectness | Imprecision | Other considerations | Impact | Certainty of the evidence (GRADE) | Importance |
|---|---|---|---|---|---|---|---|---|---|---|
| Blinded expert preference | 2 | Nonrandomized studies | Serious | Serious | Serious | Serious | Strongly suspected publication or reporting bias | Not pooled. One study reported AI-generated reports as equivalent or preferred in 77.7% (233/300) and 56.1% (170/303) of 2 datasets; another study reported that 74.2% (49/66) were indistinguishable from human-written reports. | Very low | Important |
| Safety outcomes (clinically significant errors, omission errors, and commission errors) | 6 | Nonrandomized studies | Serious | Serious | Serious | Serious | Strongly suspected publication or reporting bias | Not pooled. Error definitions and denominators varied; one clustered reader study reported false-negative ratings of 18.5% (1584/8580) with AI compared with 17.8% (1527/8580) for radiologist reports, and false-positive ratings of 11.3% (971/8580) compared with 9.7% (831/8580). | Very low | Critical |
| Workflow burden | 11 | Nonrandomized studies | Serious | Serious | Serious | Serious | Strongly suspected publication or reporting bias | Not pooled. Workflow effects were mixed: AI shortened MRI[c] reading time (53 vs 61 s) but increased impression-editing time (18.29 vs 12.20 s) and edit distance (12.32 vs 5.74 words) in another study. | Very low | Important |

[a]GRADE: Grading of Recommendations Assessment, Development, and Evaluation.
[b]This table follows the GRADEpro (Evidence Prime Inc) or GRADE Summary of Findings structure for narrative evidence. The full review included 101 studies; 36 reported at least one clinically relevant human evaluation, safety, workflow, report quality, acceptability, or time signal; the GRADEpro table uses the 21 selected studies with clinically interpretable end points, as shown in Table 2.
[c]MRI: magnetic resonance imaging.

### Clinically Relevant Safety and Workflow Outcomes

Table 2 summarizes studies with clinically relevant human evaluation, safety, or workflow findings using standardized outcome domains, while key extractable quantitative findings are reported in this section, where they can be interpreted with study-specific context. In a large chest radiograph study of autonomous AI preliminary reports, acceptance was 70.5% (6047/8580) for AI-generated reports versus 73.3% (6288/8580) for radiologist reports, whereas false-positive findings were 11.3% (1527/8580) versus 9.7% (831/8580) and false-negative findings were 18.5% (1584/8580) versus 17.8% (1527/8580) in clustered reader evaluations [51]. These evaluations were repeated reader-report or reader-case ratings rather than independent case-level events. In a clinician collaboration study for chest radiograph report generation, AI-generated reports were equivalent to or preferred over clinician reports in 77.7% (233/300) of one dataset and 56.1% (170/303) of another when judged by at least half of the raters, but clinically significant errors occurred in both AI-generated and clinician-generated reports [52]. Outcome-level extraction details are provided in Supplementary Table S4 in Multimedia Appendix 2.

Other human evaluation studies of autonomous multimodal generation reported expert indistinguishability, acceptance, or quality ratings in 3D brain CT report generation, bronchoscopy report generation, and medical vision–language reporting across mixed devices [55-57]. Related autonomous report generation studies in chest x-ray, MRI, and fine-tuned LLaMA-based pipelines reported technical performance in additional settings [5,6,72].

Workflow outcomes differed across task types. In a brain MRI reader study, AI assistance improved diagnostic performance and reduced reading time from 61 to 53 seconds, with greater reported benefit among junior radiologists [53]. In contrast, in a study of radiological impression drafting with multiple readers, drafts generated by the model required more editing time than the radiologist baseline (18.29 vs 12.20 s) and a greater edit distance (12.32 vs 5.74 words changed) [54]. For this study of impression drafting, arm-specific means and CIs were sufficient to support the cautious quantitative reconstruction of single studies for sensitivity purposes, but no second clinically comparable workflow study was available for pooling (Supplementary Table S5 in Multimedia Appendix 2) [54]. Additional workflow studies in lumbar spine MRI, breast ultrasound, structured positron emission tomography/CT reporting, and practical drafting with ChatGPT reported on workflow integration in heterogeneous settings [73-76].

Additional studies in the evidence base evaluated blinded radiologist assessment, structured report reader experiments, emergency CT impression drafting with assessment of critical omissions, real-world or pilot reporting workflows, prospective endoscopy report validation, and reporting time comparisons with LLM assistance (Table 2). Reported findings varied by task: some studies reported improved reporting efficiency, structured report usability, or expert acceptability, whereas others reported persistent omissions, hallucinations, lower quality ratings from specialists compared with human reports, or workflow benefits limited to narrow, standardized settings. The characteristics of studies added to the final evidence set are summarized in Supplementary Table S1 in Multimedia Appendix 2 [58-71,77-130].

### Feasibility of Meta-Analysis and Narrative Synthesis

No prespecified effectiveness, safety, or workflow burden outcome had at least 2 studies with directly comparable definitions, coherent designs, and analyzable effect structures suitable for meta-analysis. The expanded evidence base broadened clinical contexts but also added heterogeneity in inputs, outputs, comparators, human oversight, and denominator definitions. The strongest quantitative workflow signal came from the single impression-drafting study in Table 2 and Supplementary Table S5 in Multimedia Appendix 2. Safety and expert evaluation outcomes were usually reported as clustered reader-level counts, threshold preferences, paired overlap summaries, qualitative ratings, or pilot metrics that could not be pooled. These reporting barriers were summarized in Supplementary Table S6 in Multimedia Appendix 2 and structured the narrative synthesis.

No additional unpublished outcome data were available by May 19, 2026; therefore, all quantitative reconstruction and narrative synthesis relied on publicly reported results.

## Discussion

### Principal Findings

This systematic review assessed the effectiveness (expert acceptance and blinded preference), safety (clinically significant errors, omission errors, and commission errors), and workflow burden (reporting time, corrections, edit distance, and editing burden) of LLM-based medical report generation. Relative to these objectives, the principal finding is that the field is active and clinically diverse, but the reported evidence remains too heterogeneous to support stable conclusions about overall effectiveness, safety, or workflow benefit. Effectiveness outcomes suggested possible assistive value in selected settings [51,52], but these outcomes were not consistently paired with case-level safety end points. Safety and workflow burden outcomes were reported with incompatible denominators, task structures, and comparison designs [53,54].

The findings, therefore, support a cautious, task-specific interpretation rather than a general claim that LLM-based report generation is effective or safe. Some supervised workflows appeared acceptable to experts or were useful for selected reporting tasks, but current evidence does not justify autonomous replacement of clinicians [20,23,131]. Benefit and risk differed by task and workflow because impression generation from findings, human-supervised drafting, report editing or structuring, and autonomous image-to-report generation involve different clinical inputs, opportunities for human correction, and failure modes. These task categories

should not be treated as interchangeable when interpreting acceptance, reporting errors, workflow burden, or implementation readiness.

## Interpretation and Comparison With Existing Literature

Prior reviews and commentaries have described the rapid emergence of AI report generation, structured reporting, prompt engineering, and foundation-model applications in imaging and endoscopy [24-28]. Recent systematic and scoping reviews have further summarized automated radiology report generation, structured reporting approaches, and deep learning trends [29-31], as well as LLM use in radiology reports, readability, simplification, and broader radiology adoption trajectories [32-34]. These publications establish that the technical literature is active, but leave a practical clinical question unresolved: whether published studies report clinically interpretable effectiveness, safety, and workflow outcomes in a way that can guide supervised implementation. This review differs from those prior reviews by prioritizing outcome definitions, the level of human oversight, and synthesis feasibility rather than technical capability alone. This emphasis allowed us to map not only what models can generate but also how reported outcomes relate to clinical usefulness.

The effectiveness findings need to be interpreted according to the outcomes that were actually reported. Expert acceptance, blinded preference, and report-quality judgments were the most common effectiveness signals, and several studies suggested that generated reports or AI-assisted reports could be acceptable, equivalent, or preferred in selected reader-evaluation settings [51,52]. However, acceptance and preference are not the same as clinical effectiveness unless they are accompanied by evidence that final reports are accurate, clinically complete, and not more burdensome to verify [20,23]. In this review, positive effectiveness signals were often limited by single-study designs, clustered reader ratings, heterogeneous rating scales, and incomplete linkage to omissions, commissions, or final clinician-edited report quality. Therefore, the evidence supports a narrower conclusion: LLM-generated or LLM-assisted report text may be useful in some supervised contexts, but published outcomes do not yet establish a generalizable effectiveness advantage across medical reporting workflows.

The safety and workflow outcomes further explain why broad effectiveness claims would be premature. In chest x-ray and clinician-collaboration settings, expert acceptance or preference could coexist with false-negative findings or clinically significant errors, showing that acceptable language does not necessarily imply safe reporting [51,52]. Workflow findings were also mixed: a brain MRI study reported shorter reading time with AI assistance, whereas an impression-drafting study reported longer editing time and greater edit distance than the radiologist baseline [53,54]. These findings suggest that workflow benefit depends on where the system enters the reporting process, whether clinicians start from images, findings, or draft text, and how much verification or correction the AI output requires.

The nonpoolability of the evidence base is itself an important synthesis finding. Meta-analysis was inappropriate not because publication volume was small, but because the evidence structure was incompatible across studies. Clustered reader-level counts, threshold preference summaries, paired workflow results without reconstructible dispersion, heterogeneous quality scales, pilot outcomes, and unclear denominators prevented valid pooling. This pattern aligns with SWiM principles: when effect estimates cannot be validly combined, transparent, structured narrative synthesis is preferable to forced quantitative aggregation [42]. The expanded search increased clinical breadth but did not create a coherent synthesis set for any primary or key secondary outcome, indicating a reporting-standardization problem that limits cumulative clinical inference.

The dominance of chest x-ray studies also needs careful interpretation. Public corpora such as MIMIC-CXR have accelerated model development and benchmarking [15,16], but repeated use of the same or related datasets can increase the risk of overlapping test sets, model and data leakage, and overly optimistic generalizability. This concern is especially relevant when training, validation, and external testing boundaries are not reported clearly. Endoscopy, pathology whole-slide imaging, CT, MRI, and ultrasound reporting involve different information densities, reference standard structures, and workflow constraints; consequently, evidence from templated chest radiography should not be generalized uncritically to more complex reporting domains.

The safety findings also require interpretation at the level of the clinical task. Human preference or acceptance of generated text should not be interpreted as equivalent to safety unless omissions, commissions, clinically significant discrepancies, and failed generations are measured explicitly [19,21,23]. A fluent report can still omit an important abnormality, introduce an unsupported finding, or increase the review burden for the clinician who must verify it [21-23]. This distinction is especially important for autonomous image-to-report systems, where errors may arise before a clinician's review, but it also matters for impression-drafting and report-structuring tools because they can shape the language and emphasis of the final report. Accordingly, implementation decisions should require direct measurement of error types in the specific workflow under consideration.

## Clinical and Research Implications

For clinical practice, the real-world implication is that LLM-based reporting systems should be evaluated as supervised assistive tools, not as unsupervised replacements [23,131]. Implementation decisions should be workflow specific because an AI tool that drafts an impression from existing findings has a different safety profile from an autonomous image-to-report model, and a tool that improves readability or structure does not necessarily reduce diagnostic error. Local validation should measure the task actually being deployed, the final clinician-edited report, and the workload created by reviewing or correcting generated text [20,23,131]. This means evaluating systems in the local modality, report template, patient mix, clinician review process, and

information-technology environment in which they would actually be used.

For developers and clinical investigators, these findings point toward a staged evidence pathway. Benchmark metrics may be appropriate for early technical screening, but progression toward clinical use should require prospective or externally validated studies with safety outcomes at the case level and reconstructable workflow statistics [20,23,131]. Evaluation should separate omissions from commissions, report failed or unusable generations, state case-level denominators, and measure reporting time, corrections, edit distance, acceptance, blinded preference, and final report quality in realistic workflows [23,44,131]. These reporting elements would make future evidence more useful for health systems because they connect apparent effectiveness to the safety and labor required to produce a clinically usable final report.

For evidence synthesis, the field needs more consistent reporting before meta-analysis can become informative. Future studies should define whether the unit of analysis is the patient, examination, report, reader-report pair, or generated text segment; specify whether evaluators were blinded; report how unusable generations were handled; and provide enough event counts or summary statistics to reconstruct effects [42,44,131]. They should also report effectiveness, safety, and workflow outcomes together rather than presenting acceptance, errors, and time savings as isolated signals. Without these elements, additional studies may increase publication volume without improving certainty or allowing health systems to judge whether LLM-assisted reporting improves outcomes that matter in practice.

## Limitations

This review has limitations. The evidence base was dominated by retrospective studies; prospective workflow research was sparse; and no main outcome supported a robust pooled estimate across multiple directly comparable studies. The custom AI validation framework was prespecified and transparently domain-based, but it was not a formally validated measurement instrument, and interrater agreement statistics were not generated; accordingly, those judgments should be interpreted as structured qualitative bias signals rather than as validated scale outputs. The narrative GRADE-informed certainty assessment was likewise constrained by the absence of pooled estimates; it used GRADE domains to structure certainty judgments but did not produce certainty ratings around pooled effect sizes [45].

Excluding preprints may underrepresent the newest model capabilities in a fast-moving field, but this choice was appropriate for a clinical review focused on peer-reviewed evidence. Some included and excluded studies may share public datasets or related benchmarks, reducing independence across studies [15,16]. We also could not always determine whether model training data overlapped with evaluation data, especially for studies using public radiology corpora or proprietary foundation models. Finally, structured narrative synthesis provides less statistical compression than meta-analysis but better reflects the current nonpoolable evidence structure [42].

## Conclusions and Broader Implications

This review provides a clinically oriented synthesis of LLM-based medical report generation across effectiveness, safety, workflow burden, and human oversight. It differs from radiology-specific, readability-oriented, and benchmark-focused reviews by asking whether peer-reviewed evidence can support clinical judgments about expert acceptance, blinded preference, clinically significant errors, omissions, commissions, reporting time, corrections, edit distance, and editing burden. Its main contribution is to show that the barrier is no longer simply whether models can produce plausible reports; it is whether studies report acceptance, preference, omissions, commissions, failed generations, time, corrections, edit distance, and human oversight well enough to guide practice. The broader real-world implication is that adoption should remain assistive, clinician-supervised, locally validated, and specific to the intended reporting workflow until stronger prospective safety and workflow evidence is available [20,23,131]. In practical terms, LLM-based reporting should be viewed as a candidate workflow aid whose value depends on the clinical task, validation setting, and human review process, not as a general replacement for expert report generation. The next stage of research should therefore move from benchmark-centered demonstrations toward standardized, case-level effectiveness, safety, and workflow burden evaluations that can support both local implementation decisions and future quantitative synthesis.

### Acknowledgments

The authors thank the corresponding authors of eligible studies who were contacted for clarification of potentially analyzable outcomes.

During the preparation of this manuscript, the authors used ChatGPT (GPT-5.5) to improve English language, grammar, and readability. The tool was not used for data analysis, interpretation of results, generation of scientific content, or reference retrieval. All AI-assisted output was reviewed, edited, and approved by the authors, who take full responsibility for the final content of the manuscript.

### Funding

This work was supported by the National High-Level Hospital Clinical Research Funding (grant: LC2024A04) and the Chinese Academy of Medical Sciences Innovation Fund for Medical Sciences (grant: 2022-I2M-C&T-B-059).

## Data Availability

All data extracted and analyzed in this systematic review are presented in the article and its multimedia appendices. The prespecified extraction framework, study-level extraction sheet, verified outcome-level extraction sheet, and conversion decision log are available in the Multimedia Appendices 1 and 2; additional clarifying materials are available from the corresponding author upon reasonable request.

## Authors' Contributions

Conceptualization: JLH, XGN
Methodology: JLH, JQZ, XGN
Literature screening: JLH, JQZ
Data curation: JLH
Formal analysis: JLH, JQZ
Investigation: JLH, JQZ
Supervision: XGN
Writing – original draft: JLH
Writing – review & editing: JQZ, XGN

## Conflicts of Interest

None declared.

## Multimedia Appendix 1

Exact search strategies and search update summary.
[DOCX File (Microsoft Word File), 42 KB-Multimedia Appendix 1]

## Multimedia Appendix 2

Supplementary evidence tables.
[DOCX File (Microsoft Word File), 76 KB-Multimedia Appendix 2]

## Checklist 1

Reporting guideline checklists: PRISMA 2020, PRISMA abstracts, PRISMA-S, and SWiM.
[DOCX File (Microsoft Word File), 37 KB-Checklist 1]

## References

1. Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. Ann Intern Med. Dec 6, 2016;165(11):753-760. [doi: 10.7326/M16-0961] [Medline: 27595430]
2. Jing B, Xie P, Xing E. On the automatic generation of medical imaging reports. Proc Assoc Comput Linguist Annu Meet. 2018:2577-2586. [doi: 10.18653/v1/P18-1240]
3. Chen Z, Song Y, Chang TH, Wan X. Generating radiology reports via memory-driven transformer. Proc Conf Empir Methods Nat Lang Process. 2020:1439-1449. [doi: 10.18653/v1/2020.emnlp-main.112]
4. Zhong T, Zhao W, Zhang Y, et al. ChatRadio-Valuer: a chat large language model for generalizable radiology impression generation on multi-institution and multi-system data. IEEE Trans Biomed Eng. Mar 2026;73(3):1050-1061. [doi: 10.1109/TBME.2025.3597325] [Medline: 40788800]
5. Yeasin M, Moinuddin KA, Havugimana F, Wang L, Park P. Auto-Rad: end-to-end report generation from lumber spine MRI using vision-language model. J Clin Med. Nov 23, 2024;13(23):7092. [doi: 10.3390/jcm13237092] [Medline: 39685549]
6. Lee RW, Lee KH, Yun JS, Kim MS, Choi HS. Comparative analysis of M4CXR, an LLM-based chest X-ray report generation model, and ChatGPT in radiological interpretation. J Clin Med. Nov 22, 2024;13(23):7057. [doi: 10.3390/jcm13237057] [Medline: 39685515]
7. Chen Z, Bie Y, Jin H, Chen H. Large language model with region-guided referring and grounding for CT report generation. IEEE Trans Med Imaging. Aug 2025;44(8):3139-3150. [doi: 10.1109/TMI.2025.3559923] [Medline: 40215158]
8. Massimi D, Stefano LD, Rizkala T, et al. Large language model-driven analysis and report generation of endoscopy videos-A pilot study. Dig Endosc. Mar 2026;38(3):e70134. [doi: 10.1111/den.70134] [Medline: 41804240]
9. Zhang X, Zheng X, Li Z, et al. A fine-tuning multimodal large language model for endoscopic report generation. Biomed Signal Process Control. Jun 2026;118:109737. [doi: 10.1016/j.bspc.2026.109737]
10. Joyson JL, Soundar KR, Nancy P. A multi-modal fusion-based deep learning with finetuned LLaMA 3 for lung disease diagnosis using PACS radiology reports. Biomed Signal Process Control. May 2026;117:109466. [doi: 10.1016/j.bspc.2026.109466]

11. Liu F, Zhou H, Wang K, et al. MetaGP: a generative foundation model integrating electronic health records and multimodal imaging for addressing unmet clinical needs. Cell Rep Med. Apr 15, 2025;6(4):102056. [doi: 10.1016/j.xcrm.2025.102056] [Medline: 40187356]
12. Dasanayaka C, Dandeniya K, Dissanayake MB, Gunasena C, Jayasinghe R. Multimodal AI and large language models for orthopantomography radiology report generation and Q&A. Appl Syst Innov. 2025;8(2):39. [doi: 10.3390/asi8020039]
13. Helmy H, Hosseini A, Ibrahim A, Baig-Mirza A, Sadek AR, Serag A. SPINE: segmentation-guided processing and integration of multimodal spinal MRI for natural-language enhanced report generation. Appl Artif Intell. 2026;40(1). [doi: 10.1080/08839514.2026.2626117]
14. Li Z, Li M, Wang W, Huang Q. Ultrasound report generation with fuzzy knowledge and multi-modal large language model. Expert Syst Appl. Nov 2025;292:128555. [doi: 10.1016/j.eswa.2025.128555]
15. Wang Z, Liu L, Wang L, Zhou L. R2GenGPT: radiology report generation with frozen LLMs. Meta-Radiology. Nov 2023;1(3):100033. [doi: 10.1016/j.metrad.2023.100033]
16. Liu Y, Li Y, Wang Z, et al. A systematic evaluation of GPT-4V's multimodal capability for chest X-ray image analysis. Meta-Radiology. Dec 2024;2(4):100099. [doi: 10.1016/j.metrad.2024.100099]
17. Kim KA, Hong S, Yoo S, Kang Y, Shim HS. Enhancing structured pathology report generation with foundation model and modular design. IEEE Access. 2025;13:121290-121299. [doi: 10.1109/ACCESS.2025.3588121]
18. Yan H, Shao D. Multimodal medical image analysis: integrating LLM and RAG deep learning strategies. J Adv Inf Technol. 2025;16(4):568-581. [doi: 10.12720/jait.16.4.568-581]
19. Lyell D, Coiera E. Automation bias and verification complexity: a systematic review. J Am Med Inform Assoc. Mar 1, 2017;24(2):423-431. [doi: 10.1093/jamia/ocw105] [Medline: 27516495]
20. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf. Mar 2019;28(3):231-237. [doi: 10.1136/bmjqs-2018-008370] [Medline: 30636200]
21. Lee P, Bubeck S, Petro J. Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. N Engl J Med. Mar 30, 2023;388(13):1233-1239. [doi: 10.1056/NEJMsr2214184] [Medline: 36988602]
22. Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. ACM Comput Surv. Dec 31, 2023;55(12):1-38. [doi: 10.1145/3571730]
23. Yi PH, Haver HL, Jeudy JJ, et al. Best practices for the safe use of large language models and other generative AI in radiology. Radiology. Sep 2025;316(3):e241516. [doi: 10.1148/radiol.241516] [Medline: 40985835]
24. Seah JCY, Tang JSN, Tran A. Drafting the future: the dawn of AI report generation in radiology. Radiology. Jul 2025;316(1):e243378. [doi: 10.1148/radiol.243378] [Medline: 40728403]
25. Kim TT, Makutonin M, Sirous R, Javan R. Optimizing large language models in radiology and mitigating pitfalls: prompt engineering and fine-tuning. Radiographics. Apr 2025;45(4):e240073. [doi: 10.1148/rg.240073] [Medline: 40048389]
26. Oda M. Generative AI and foundation models in medical image. Radiol Phys Technol. Dec 2025;18(4):937-948. [doi: 10.1007/s12194-025-00968-1] [Medline: 41051729]
27. Labaki C, Uche-Anya EN, Berzin TM. Artificial intelligence in gastrointestinal endoscopy. Gastroenterol Clin North Am. Dec 2024;53(4):773-786. [doi: 10.1016/j.gtc.2024.08.005] [Medline: 39489586]
28. Biswas S, Khan S, Awal SS. Can ChatGPT write radiology reports?Chin J Acad Radiol. Mar 2024;7(1):102-106. [doi: 10.1007/s42058-023-00132-x]
29. Sloan P, Clatworthy P, Simpson E, Mirmehdi M. Automated radiology report generation: a review of recent advances. IEEE Rev Biomed Eng. 2025;18:368-387. [doi: 10.1109/RBME.2024.3408456] [Medline: 38829752]
30. Busch F, Hoffmann L, Dos Santos DP, et al. Large language models for structured reporting in radiology: past, present, and future. Eur Radiol. May 2025;35(5):2589-2602. [doi: 10.1007/s00330-024-11107-6] [Medline: 39438330]
31. Izhar A, Idris N, Japar N. Medical radiology report generation: a systematic review of current deep learning methods, trends, and future directions. Artif Intell Med. Oct 2025;168:103220. [doi: 10.1016/j.artmed.2025.103220] [Medline: 40700862]
32. Lee RC, Hadidchi R, Coard MC, et al. Use of large language models on radiology reports: a scoping review. J Am Coll Radiol. Mar 2026;23(3):437-454. [doi: 10.1016/j.jacr.2025.10.005] [Medline: 41196263]
33. Patwardhan V, Balchander D, Fussell D, et al. Leveraging large language models to enhance radiology report readability: a systematic review. J Am Coll Radiol. Mar 2026;23(3):354-361. [doi: 10.1016/j.jacr.2025.09.004] [Medline: 40945554]
34. Al Zaabi A, Alshibli R, AlAmri A, AlRuheili I, Lutfi SL. Trends and trajectories in the rise of large language models in radiology: scoping review. JMIR Med Inform. Dec 9, 2025;13:e78041. [doi: 10.2196/78041] [Medline: 41364806]

35. Papineni K, Roukos S, Ward T, Zhu WJ. BLEU: a method for automatic evaluation of machine translation. In: Isabelle P, Charniak E, Lin D, editors. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2002:311-318. [doi: 10.3115/1073083.1073135]
36. Lin CY. ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop. Association for Computational Linguistics; 2004:74-81. URL: https://aclanthology.org/W04-1013/ [Accessed 2026-08-01]
37. Vedantam R, Zitnick CL, Parikh D. CIDEr: consensus-based image description evaluation. Proc IEEE Comput Vis Pattern Recognit Conf (CVPR). 2015:4566-4575. [doi: 10.1109/CVPR.2015.7299087]
38. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. BERTScore: evaluating text generation with BERT. Presented at: International Conference on Learning Representations; Apr 26-30, 2020; Addis Ababa, Ethiopia. 2020.URL: https://openreview.net/pdf?id=SkeHuCVFDr [Accessed 2026-08-01]
39. Gholipour Picha S, Al Chanti D, Caplier A. Trust but verify: image-aware evaluation of radiology report generators. Mach Learn Appl. Mar 2026;23:100851. [doi: 10.1016/j.mlwa.2026.100851]
40. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. Mar 29, 2021;372:n71. [doi: 10.1136/bmj.n71] [Medline: 33782057]
41. Rethlefsen ML, Kirtley S, Waffenschmidt S, et al. PRISMA-S: an extension to the PRISMA statement for reporting literature searches in systematic reviews. Syst Rev. Jan 26, 2021;10(1):39. [doi: 10.1186/s13643-020-01542-z] [Medline: 33499930]
42. Campbell M, McKenzie JE, Sowden A, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ. Jan 16, 2020;368:l6890. [doi: 10.1136/bmj.l6890] [Medline: 31948937]
43. Sterne JA, Hernán MA, Reeves BC, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. Oct 12, 2016;355:i4919. [doi: 10.1136/bmj.i4919] [Medline: 27733354]
44. Tejani AS, Klontzas ME, Gatti AA, et al. Checklist for artificial intelligence in medical imaging (CLAIM): 2024 update. Radiol Artif Intell. Jul 2024;6(4):e240300. [doi: 10.1148/ryai.240300] [Medline: 38809149]
45. Guyatt GH, Oxman AD, Vist GE, et al. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. Apr 26, 2008;336(7650):924-926. [doi: 10.1136/bmj.39489.470347.AD] [Medline: 18436948]
46. Wang X, Wang F, Wang H, et al. Activating associative disease-aware vision token memory for LLM-based X-ray report generation. IEEE Trans Med Imaging. Feb 2026;45(2):583-595. [doi: 10.1109/TMI.2025.3603416] [Medline: 40864571]
47. Noh WJ, Pi SW, Lee BD. Hybrid framework for lesion-aware, clinically coherent chest X-ray report generation using contrastive learning and large language models. Sci Rep. Jan 5, 2026;16(1):4645. [doi: 10.1038/s41598-025-34799-2] [Medline: 41492086]
48. Han P, Li X, Jing S, Wei J. Dynamic feature fusion guiding and multimodal large language model refining for medical image report generation. Expert Syst Appl. Mar 2026;299:130082. [doi: 10.1016/j.eswa.2025.130082]
49. Jia X, Xiong Y, Pei S, Zhang Y, Yan C, Fang Z. Semantic feedback-based RAG for radiology report generation. Big Data Min Anal. 2026;9(2):393-406. [doi: 10.26599/BDMA.2025.9020037]
50. Wang X, Li Y, Wang F, Wang S, Li C, Jiang B. R2GenCSR: mining contextual and residual information for LLM-based radiology report generation. IEEE J Biomed Health Inform. ;2026:1-14. [doi: 10.1109/JBHI.2026.3669539]
51. Hong EK, Ham J, Roh B, et al. Diagnostic accuracy and clinical value of a domain-specific multimodal generative AI model for chest radiograph report generation. Radiology. Mar 2025;314(3):e241476. [doi: 10.1148/radiol.241476] [Medline: 40131111]
52. Tanno R, Barrett DGT, Sellergren A, et al. Collaboration between clinicians and vision-language models in radiology report generation. Nat Med. Feb 2025;31(2):599-608. [doi: 10.1038/s41591-024-03302-1] [Medline: 39511432]
53. Wang ML, Zhang RP, Wu WJ, et al. Evaluation of large language models for diagnostic impression generation from brain MRI report findings: a multicenter benchmark and reader study. NPJ Digit Med. Jan 22, 2026;9(1):187. [doi: 10.1038/s41746-026-02380-4] [Medline: 41571872]
54. Serapio A, Chaudhari G, Savage C, et al. An open-source fine-tuned large language model for radiological impression generation: a multi-reader performance study. BMC Med Imaging. Sep 27, 2024;24(1):254. [doi: 10.1186/s12880-024-01435-w] [Medline: 39333958]
55. Li CY, Chang KJ, Yang CF, et al. Towards a holistic framework for multimodal LLM in 3D brain CT radiology report generation. Nat Commun. Mar 6, 2025;16(1):2258. [doi: 10.1038/s41467-025-57426-0] [Medline: 40050277]
56. Luo X, Huang X, Liang X, et al. Towards automated reporting: a bronchoscopy report dataset for enhancing multimodality large language models. Sci Data. Feb 3, 2026;13(1):339. [doi: 10.1038/s41597-026-06692-8] [Medline: 41634075]

57. Ayaz M, Khan M, Saqib M, Khelifi A, Sajjad M, Elsaddik A. MedVLM: medical vision–language model for consumer devices. IEEE Consum Electron Mag. 2025;14(5):75-83. [doi: 10.1109/MCE.2024.3522521]
58. Li M, Wang Y, Miao Z, et al. Fine-tuned large language model for automated radiology impression generation: a multicenter evaluation. Radiol Artif Intell. May 2026;8(3):e250714. [doi: 10.1148/ryai.250714] [Medline: 41983921]
59. de Margerie-Mellon C, Duron L, Fournier L, Ouakil L, Soulat G, Teixeira PG. Reporting efficiency in diagnostic imaging: can plug-and-play general-purpose large language models outperform conventional speech recognition? Eur Radiol. Aug 2026;36(8):6282-6291. [doi: 10.1007/s00330-026-12524-5] [Medline: 42009869]
60. Jiang R, Chen B, Dong Z, et al. Domain specific multimodal large language model for automated endoscopy reporting with multicenter prospective validation. NPJ Digit Med. 2026;9(1):394. [doi: 10.1038/s41746-026-02569-7] [Medline: 41904204]
61. Tekdemir H, Çıvgın E, Akpınar Ş, et al. Evaluating large language models for Turkish emergency CT impression drafting: quality, critical omissions, and readability. J Imaging Inform Med. May 6, 2026. [doi: 10.1007/s10278-026-01989-x] [Medline: 42091800]
62. Lorusso G, Ruscino G, Spitaleri A, et al. Comparative evaluation of large language models for generating CAD-RADS 2.0-compliant diagnostic conclusions in cardiac CT reports. Insights Imaging. Apr 22, 2026;17(1):112. [doi: 10.1186/s13244-026-02285-6] [Medline: 42018072]
63. Dong F, Nie S, Chen M, Xu F, Li Q. Keyword-based AI assistance in the generation of radiology reports: a pilot study. NPJ Digit Med. Aug 1, 2025;8(1):490. [doi: 10.1038/s41746-025-01889-4] [Medline: 40750683]
64. Pellegrini C, Özsoy E, Gassert FT, et al. Enhancing radiology workflows through collaborative AI-assisted chest X-ray reporting using large vision-language models: a proof-of-concept study. Insights Imaging. Apr 28, 2026;17(1):123. [doi: 10.1186/s13244-026-02292-7] [Medline: 42047956]
65. Tan N. Large language model–assisted radiology reporting in a single-radiologist implementation: a retrospective cohort study interpreted through a UTAUT lens. Abdom Radiol. 2026. [doi: 10.1007/s00261-026-05524-y]
66. Wang Z, Guo R, Sun P, Qian L, Hu X. Enhancing diagnostic accuracy and efficiency with GPT-4-generated structured reports: a comprehensive study. J Med Biol Eng. Feb 2024;44(1):144-153. [doi: 10.1007/s40846-024-00849-9]
67. Hong EK, Lee S, Song O, Leung A, Hammer M, Suh CH. Temperature setting of a multimodal generative artificial intelligence model: association with accuracy and quality of artificial intelligence-generated chest radiograph reports. AJR Am J Roentgenol. Feb 2026;226(2):e2533704. [doi: 10.2214/AJR.25.33704] [Medline: 41159788]
68. Zanardo M, Albano D, Molinari V, et al. Can AI write reports like a radiologist? A blinded evaluation of large language model-generated lumbar spine MRI reports. Eur Radiol Exp. Feb 23, 2026;10(1):16. [doi: 10.1186/s41747-026-00682-6] [Medline: 41729370]
69. Xie Y, Hu Z, Tao H, et al. Large language models for efficient whole-organ MRI score-based reports and categorization in knee osteoarthritis. Insights Imaging. May 14, 2025;16(1):100. [doi: 10.1186/s13244-025-01976-w] [Medline: 40366500]
70. Woźnicki P, Laqua C, Fiku I, et al. Automatic structuring of radiology reports with on-premise open-source large language models. Eur Radiol. Apr 2025;35(4):2018-2029. [doi: 10.1007/s00330-024-11074-y] [Medline: 39390261]
71. Shao A, Liu X, Shen W, et al. Generative artificial intelligence for fundus fluorescein angiography interpretation and human expert evaluation. NPJ Digit Med. Jul 2, 2025;8(1):396. [doi: 10.1038/s41746-025-01759-z] [Medline: 40603524]
72. Voinea ȘV, Mămuleanu M, Teică RV, Florescu LM, Selișteanu D, Gheonea IA. GPT-driven radiology report generation with fine-tuned Llama 3. Bioengineering (Basel). Oct 18, 2024;11(10):1043. [doi: 10.3390/bioengineering11101043] [Medline: 39451418]
73. Park J, Han K, Oh JS, et al. Prospective evaluation of artificial intelligence (AI) in lumbar spine magnetic resonance imaging (MRI) workflow: from deep learning (DL)–enhanced accelerated acquisition to simultaneous vision language model (VLM)–based automated report generation. Eur J Radiol. Mar 2026;196:112695. [doi: 10.1016/j.ejrad.2026.112695] [Medline: 41579672]
74. Soleimani M, Seyyedi N, Ayyoubzadeh SM, Kalhori SRN, Keshavarz H. Practical evaluation of ChatGPT performance for radiology report generation. Acad Radiol. Dec 2024;31(12):4823-4832. [doi: 10.1016/j.acra.2024.07.020] [Medline: 39142976]
75. Azhar K, Lee BD, Byon SS, Lee S, Cho KR, Song SE. Semiautomated breast ultrasound report generation using multimodal large language models and deep learning. Front Med (Lausanne). 2026;13:1679203. [doi: 10.3389/fmed.2026.1679203] [Medline: 41647521]
76. Chen K, Xu W, Li X. The potential of Gemini and GPTs for structured report generation based on free-text $^{18}$F-FDG PET/CT breast cancer reports. Acad Radiol. Feb 2025;32(2):624-633. [doi: 10.1016/j.acra.2024.08.052] [Medline: 39245597]

77. Raminedi S, Shridevi S, Won D. Multi-modal transformer architecture for medical image analysis and automated report generation. Sci Rep. Aug 20, 2024;14(1):19281. [doi: 10.1038/s41598-024-69981-5] [Medline: 39164302]
78. Bosbach WA, Heide MS, Gözlügöl N, et al. Conceptual proposal for LLM-generated FDG PET/CT follow-up reports in melanoma: a pilot study on model stability and blinded expert evaluation. Front Nucl Med. 2026;6:1723650. [doi: 10.3389/fnume.2026.1723650] [Medline: 41909529]
79. Tran M, Schmidle P, Guo RR, et al. Generating dermatopathology reports from gigapixel whole slide images with HistoGPT. Nat Commun. May 27, 2025;16(1):4886. [doi: 10.1038/s41467-025-60014-x] [Medline: 40419470]
80. Liu D, Li W, Xu N, et al. SpineVLM: a markdown-guided structured fine-tuning framework for spine X-ray report generation. IEEE J Biomed Health Inform. May 4, 2026. [doi: 10.1109/JBHI.2026.3689568] [Medline: 42081407]
81. Ye X, Shen Y, Chen Q, et al. Report generation system for slit-lamp image interpretation using vision-language models. Ophthalmol Ther. Apr 2026;15(4):1509-1522. [doi: 10.1007/s40123-026-01352-x] [Medline: 41831132]
82. Barzegar M, Rostami H, Sanati A, Afshoon R, Kosari A, Keshavarz A. Enhancing chest X-ray report generation with pathology-guided prompts and vision-language shortcut bias mitigation. Biomed Signal Process Control. Jul 2026;120:110173. [doi: 10.1016/j.bspc.2026.110173]
83. Allaberdiev S, Khan A, Mamarasulov S, Chen X. Chestxgen: dynamic memory–augmented vision-language transformer with context-aware gating for radiology report generation. J Artif Intell Soft Comput Res. Jan 1, 2026;16(1):55-72. [doi: 10.2478/jaiscr-2026-0003]
84. Lin KH, Chang CP, Kuo CT, et al. Automatic speech recognition and large language models for multilingual pathology report generation: proof-of-concept study. JMIR Form Res. May 13, 2026;10:e90814. [doi: 10.2196/90814] [Medline: 42127277]
85. Khatoon S, Mahmood A. Automated radiological report generation from breast ultrasound images using vision and language transformers. J Imaging. Feb 6, 2026;12(2):68. [doi: 10.3390/jimaging12020068] [Medline: 41745433]
86. Park YW, Kang M, Ryu H, et al. A robust vision language model for molecular status prediction and radiology report generation in adult-type diffuse gliomas. NPJ Digit Med. Apr 2, 2026;9(1):406. [doi: 10.1038/s41746-026-02581-x] [Medline: 41927936]
87. Hossain MZ, Ahmed M, Samu MSS, Islam MR. Privacy-preserving chest X-ray report generation via multimodal federated learning with ViT and GPT2. Biomed Mater Devices. 2025. [doi: 10.1007/s44174-025-00538-4]
88. Hosseini A, Ibrahim A, Serag A. M3: multimodal artificial intelligence for medical report generation and visual question answering from 3D abdominal CT scans. BJR Artif Intell. 2025;2(1):ubaf011. [doi: 10.1093/bjrai/ubaf011] [Medline: 42063998]
89. Zhang L, Liu M, Wang L, et al. Constructing a large language model to generate impressions from findings in radiology reports. Radiology. Sep 2024;312(3):e240885. [doi: 10.1148/radiol.240885] [Medline: 39287525]
90. Sun Z, Ong H, Kennedy P, et al. Evaluating GPT-4 on impressions generation in radiology reports. Radiology. Jun 2023;307(5):e231259. [doi: 10.1148/radiol.231259] [Medline: 37367439]
91. Huang J, Neill L, Wittbrodt M, et al. Generative artificial intelligence for chest radiograph interpretation in the emergency department. JAMA Netw Open. Oct 2, 2023;6(10):e2336100. [doi: 10.1001/jamanetworkopen.2023.36100] [Medline: 37796505]
92. Mallio CA, Bernetti C, Sertorio AC, Zobel BB. ChatGPT in radiology structured reporting: analysis of ChatGPT-3.5 Turbo and GPT-4 in reducing word count and recalling findings. Quant Imaging Med Surg. Feb 1, 2024;14(2):2096-2102. [doi: 10.21037/qims-23-1300] [Medline: 38415145]
93. Stephan D, Bertsch A, Burwinkel M, et al. AI in dental radiology improving the efficiency of reporting with ChatGPT: comparative study. J Med Internet Res. Dec 23, 2024;26:e60684. [doi: 10.2196/60684] [Medline: 39714078]
94. Kim C, Cho S, Yoon JH. Utility of multimodal large language models in analyzing chest X-rays with incomplete contextual information. Healthc Inform Res. Oct 2025;31(4):416-425. [doi: 10.4258/hir.2025.31.4.416] [Medline: 41265427]
95. Di Palma L, Darvizeh F, Alì M, Fazzini D. Structured transformation of unstructured prostate MRI reports using large language models. Tomography. Jun 17, 2025;11(6):69. [doi: 10.3390/tomography11060069] [Medline: 40560015]
96. Gupta A, Malhotra H, Garg AK, Rangarajan K. Enhancing radiological reporting in head and neck cancer: converting free-text CT scan reports to structured reports using large language models. Indian J Radiol Imaging. 2025;35(1):43-49. [doi: 10.1055/s-0044-1788589] [Medline: 39697521]
97. Lee S, Youn J, Kim H, Kim M, Yoon SH. CXR-LLaVA: a multimodal large language model for interpreting chest X-ray images. Eur Radiol. 2025;35(7):4374-4386. [doi: 10.1007/s00330-024-11339-6]
98. Pesapane F, Nicosia L, Rotili A, et al. A preliminary investigation into the potential, pitfalls, and limitations of large language models for mammography interpretation. Discov Oncol. Feb 24, 2025;16(1):233. [doi: 10.1007/s12672-025-02005-4] [Medline: 39992569]

99. Hartsock I, Araujo C, Folio L, Rasool G. Improving radiology report conciseness and structure via local large language models. J Imaging Inform Med. 2026;39(1):1005-1016. [doi: 10.1007/s10278-025-01510-w] [Medline: 40259201]
100. Silbergleit M, Tóth A, Chamberlin JH, et al. ChatGPT vs Gemini: comparative accuracy and efficiency in CAD-RADS score assignment from radiology reports. J Imaging Inform Med. Aug 2025;38(4):2303-2311. [doi: 10.1007/s10278-024-01328-y] [Medline: 39528887]
101. Adams LC, Truhn D, Busch F, et al. Leveraging GPT-4 for post hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study. Radiology. May 2023;307(4):e230725. [doi: 10.1148/radiol.230725] [Medline: 37014240]
102. Jiang H, Xia S, Yang Y, et al. Transforming free-text radiology reports into structured reports using ChatGPT: a study on thyroid ultrasonography. Eur J Radiol. Jun 2024;175:111458. [doi: 10.1016/j.ejrad.2024.111458] [Medline: 38613868]
103. Hasani AM, Singh S, Zahergivar A, et al. Evaluating the performance of generative pre-trained transformer-4 (GPT-4) in standardizing radiology reports. Eur Radiol. Jun 2024;34(6):3566-3574. [doi: 10.1007/s00330-023-10384-x] [Medline: 37938381]
104. Tsaniya H, Fatichah C, Suciati N, Obi T, Lee JS. Medical report generation with knowledge distillation and multi-stage hierarchical attention in vision transformer encoder and GPT-2 decoder. IEEE Access. 2025;13:132973-132989. [doi: 10.1109/ACCESS.2025.3588344]
105. Yan L, Zhou X, Wang Y, Chang X, Li Q, Han G. Automated ultrasound diagnosis via CLIP-GPT synergy: a multimodal framework for image classification and report generation. IEEE Access. 2025;13:107950-107960. [doi: 10.1109/ACCESS.2025.3578462]
106. López-Úbeda P, Martín-Noguerol T, Escartín J, Cabrera-Zubizarreta A, Luna A. Automated MRI pituitary structured reporting from free-text using a fine-tuned Llama model: a feasibility study. Jpn J Radiol. May 2025;43(5):770-778. [doi: 10.1007/s11604-024-01721-1] [Medline: 39730936]
107. Tian W, Huang X, Cheng T, et al. A medical multimodal large language model for pediatric pneumonia. IEEE J Biomed Health Inform. Sep 2025;29(9):6869-6882. [doi: 10.1109/JBHI.2025.3569361] [Medline: 40354198]
108. Li Y, Lin Z, Zhang Q, et al. Towards generalizable pathology reports via a multimodal LLM with the multicenter in-context learning. Med Image Anal. Jun 2026;111:104060. [doi: 10.1016/j.media.2026.104060]
109. García Hidalgo C, Consentino Hernández JA, Cayuela Espí JV, et al. Informe radiológico estructurado asistido por modelos de lenguaje en residentes de Radiología: piloto de implementación en Urgencias [Article in Spanish]. Rev Esp Edu Med. 2026;7(1). [doi: 10.6018/edumed.695571]
110. Tasneem N, van der Pol CB, Zahoor A, et al. From radiology findings to artificial intelligence–powered impressions: a retrospective study on the comparative performance of recent large language models. Intell Med. Apr 2026;6(2):154-165. [doi: 10.1016/j.imed.2025.11.003]
111. Sheng W, Wang Y, Xiao L, et al. BI-RADS-compliant structured mammography reporting using locally deployed large language models under privacy constraints. Eur Radiol. May 2026;36(5):3676-3686. [doi: 10.1007/s00330-025-12147-2] [Medline: 41258455]
112. Liu R, Bai Y, Yue X, Zhang P. Teaching multimodal LLMs to comprehend 12-lead electrocardiographic images. NPJ Digit Med. Mar 16, 2026;9(1):349. [doi: 10.1038/s41746-026-02551-3] [Medline: 41840182]
113. Lee C, Park S, Shin CI, et al. Read like a radiologist: efficient vision-language model for 3D medical imaging interpretation. Med Image Anal. Jun 2026;111:104077. [doi: 10.1016/j.media.2026.104077] [Medline: 41990528]
114. Alkhaldi A, Alnajim R, Alabdullatef L, Alyahya R, Chen J, Zhu D, et al. MiniGPT-MED: a unified vision-language model for radiology image understanding. Trans Mach Learn Res. 2026. URL: https://openreview.net/pdf?id=NenHFEg1Di [Accessed 2026-08-01]
115. Abdaoui H, Barbaria S, Dergaa I, et al. MedFusionT5: cross-modal attention boosts semantic quality and reduces hallucinations in dental AI. Int Dent J. Jun 2026;76(3):109404. [doi: 10.1016/j.identj.2025.109404] [Medline: 41771189]
116. Mao Y, Xu W, Qin Y, Gao Y. CT-Agent: a multimodal-LLM agent for 3D CT radiology question answering. Sci China Inf Sci. May 2026;69(5):150107. [doi: 10.1007/s11432-025-4818-7]
117. Rouhizadeh H, Sandralegar A, Yazdani A, et al. The detectability paradox: bilingual medical report generation with open-weight models and the limits of human oversight. J Am Med Inform Assoc. Jul 1, 2026;33(7):1303-1313. [doi: 10.1093/jamia/ocag070] [Medline: 42097830]
118. Zhang Y, Pan Y, Zhong T, et al. Potential of multimodal large language models for data mining of medical images and free-text reports. Meta-Radiology. Dec 2024;2(4):100103. [doi: 10.1016/j.metrad.2024.100103]
119. Bosbach WA, Senge JF, Nemeth B, et al. Ability of ChatGPT to generate competent radiology reports for distal radius fracture by use of RSNA template items and integrated AO classifier. Curr Probl Diagn Radiol. 2024;53(1):102-110. [doi: 10.1067/j.cpradiol.2023.04.001] [Medline: 37263804]
120. Fang J, Xing S, Li K, Guo Z, Li G, Yu C. Automated generation of chest X-ray imaging diagnostic reports by multimodal and multi granularity features fusion. Biomed Signal Process Control. Jul 2025;105:107562. [doi: 10.1016/j.bspc.2025.107562]

121. Edirisinghe D, Nimalsiri W, Hennayake M, Meedeniya D, Lim G. Chest X-ray report generation using abnormality guided vision language model. IEEE Access. 2025;13:157651-157673. [doi: 10.1109/ACCESS.2025.3606961]
122. Li H, Wang H, Sun X, He H, Feng J. Context-enhanced framework for medical image report generation using multimodal contexts. Knowl Based Syst. Feb 2025;310:112913. [doi: 10.1016/j.knosys.2024.112913]
123. Lin Z, Hong Z, Zhou Z, et al. DC-RRG: diagnosis-centered cascaded radiology report generation. Expert Syst Appl. Mar 2026;299:129884. [doi: 10.1016/j.eswa.2025.129884]
124. Izhar A, Idris N, Japar N. Engaging preference optimization alignment in large language model for continual radiology report generation: a hybrid approach. Cogn Comput. 2025;17(1). [doi: 10.1007/s12559-025-10404-6]
125. Leonardi G, Portinale L, Santomauro A. Enhancing radiology report generation through pre-trained language models. Prog Artif Intell. 2024. [doi: 10.1007/s13748-024-00358-5]
126. Aslam SM. Generative artificial intelligence for the automated generation of medical reports: a modified poor and rich optimization (MPRO)–based approach. Traitement du Signal. Feb 28, 2025;42(1):409-422. [doi: 10.18280/ts.420135]
127. Deng Q, Huang Z, Wang Y, et al. Grounded knowledge–enhanced medical vision–language pre-training for chest X-ray. Biomed Signal Process Control. Feb 2026;112:108416. [doi: 10.1016/j.bspc.2025.108416]
128. Li Y, Liu Y, Wang Z, et al. S-RRG-Bench: structured radiology report generation with fine-grained evaluation framework. Meta-Radiology. Dec 2025;3(4):100171. [doi: 10.1016/j.metrad.2025.100171]
129. Ucan M, Kaya B, Kaya M. Turkish chest X-ray report generation model using the Swin enhanced yield transformer (Model-SEY) framework. Diagnostics (Basel). Jul 17, 2025;15(14):1805. [doi: 10.3390/diagnostics15141805] [Medline: 40722555]
130. Rakibul Islam M, Zahid Hossain M, Ahmed M, Sharmin Sultana Samu M. Vision-language models for automated chest X-ray interpretation: leveraging ViT and GPT-2. Eng Rep. Jun 2025;7(6). [doi: 10.1002/eng2.70220]
131. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. May 2022;28(5):924-933. [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]


## Abbreviations

**BERT:** Bidirectional Encoder Representations from Transformers
**BLEU:** Bilingual Evaluation Understudy
**CIDEr:** Consensus-Based Image Description Evaluation
**CLAIM:** Checklist for Artificial Intelligence in Medical Imaging
**CT:** computed tomography
**GRADE:** Grading of Recommendations Assessment, Development, and Evaluation
**LLM:** large language model
**MRI:** magnetic resonance imaging
**PACS:** picture archiving and communication system
**PRISMA:** Preferred Reporting Items for Systematic Reviews and Meta-Analyses
**PRISMA-S:** Preferred Reporting Items for Systematic Reviews and Meta-Analyses Literature Search Extension
**PROSPERO:** International Prospective Register of Systematic Reviews
**ROBINS-I:** Risk of Bias in Non-Randomized Studies of Interventions
**ROUGE:** Recall-Oriented Understudy for Gisting Evaluation
**SWiM:** Synthesis Without Meta-Analysis


*Edited by Stefano Brini; peer-reviewed by Rajpal Singh Sisodiya, Yi Lu; submitted 02.Apr.2026; final revised version received 17.Jun.2026; accepted 02.Jul.2026; published 09.Sep.2026*

*Please cite as:*
*Huang JL, Zhu JQ, Ni XG*
*Effectiveness, Safety, and Workflow Burden of Large Language Model–Based Medical Report Generation: Systematic Review*
*J Med Internet Res 2026;28:e97007*
*URL: https://www.jmir.org/2026/1/e97007*
*doi: 10.2196/97007*