The Review Of
DIABETIC
STUDIES
OPEN ACCESS

Automated Radiology Report Generation Using Multimodal Foundation Models: A Systematic Review Of Clinical Accuracy And Safety

Assaf Owaidh Alharbi[1],Omar Salem almehmadi[2], Ibrahim Dhaifallah Aloufi[3],Khalid Marzook Alharbi[4], Sultan Owaid Baraka Al-Sehly[5], Saleh Muneer Falah ALsaedi[6], Talal Hassan Khulaif Alsaedi[7], Taher Hassan Khulaif Alsaedi[8], Saleem Rashed Aljohani[9], Ahmad Jowaiber Bakheet Alharbi[10], Ammar Abdullah Al-Muzaini[11]

[1-11]General X-ray Specialist, King Fahd Hospital in Medina, Madinah Health Cluster, Madinah, Saudi Arabia.

Abstract

Background: The global landscape of diagnostic radiology is currently navigating a precarious inflection point. As imaging volumes surge due to aging populations and expanded screening protocols, the workforce remains critically constrained. Recent global surveys indicate that over 53% of radiologists are experiencing burnout, with workforce shortages cited as a primary concern by nearly half of the profession. This systemic strain exacerbates the risk of diagnostic error—a phenomenon already estimated to affect approximately 40 million patients annually worldwide. Against this backdrop, the integration of Artificial Intelligence (AI), specifically Multimodal Foundation Models (MFMs), has emerged as a potential panacea. These models, capable of processing both visual and textual data, promise to automate the labor-intensive process of radiology report generation (RRG), thereby potentially alleviating clinician workload and standardizing diagnostic quality. However, the transition from experimental architectures to clinical deployment is fraught with challenges related to factual consistency, safety, and trust.

Objective: This systematic review aims to provide an exhaustive, nuanced evaluation of the current state of automated radiology report generation using MFMs.

Methods: A comprehensive systematic literature search was conducted across major medical and technical databases covering the period from 2020 to 2025. The review adhered to PRISMA guidelines where applicable. Inclusion criteria prioritized studies that evaluated MFMs on standard benchmarks (MIMIC-CXR, CheXpert) or through direct comparison with board-certified radiologists.

Results: The review reveals a distinct dichotomy in MFM performance: while linguistic fluency has achieved near-human levels, factual reliability remains volatile. It was demonstrated that a Fact-Aware Multimodal Retrieval-Augmented Generation (FactMM-RAG) pipeline significantly outperforms standard foundation models. By grounding generation in retrieved, factually distinct report pairs mined via RadGraph, FactMM-RAG achieved a 6.5% improvement in F1CheXbert and a 2% improvement in F1RadGraph on the MIMIC-CXR dataset compared to state-of-the-art retrievers. Conversely, large-scale comparative studies indicate that generalist models like GPT-4V and Gemini Pro Vision still trail human radiologists in diagnostic accuracy (49% vs. 61% in complex cases), although they show promise as "second readers" in specific subspecialties like chest radiology.

Safety analysis presents the most concerning findings. A multimodal evaluation reported a 74.4% overall hallucination rate across leading visual language models, with a predominance of "fabricated imaging findings" that are statistically plausible but visually absent. The review identifies a "Plausibility Paradox" where the most advanced models (e.g., Gemini 2.0) generate the most convincing, yet factually hallucinated, reports, posing a high risk of automation bias. However, specialized models like MAIRA-X have reduced critical error rates to 4.6%, approaching the human baseline of 3.0%.

Conclusion: The era of autonomous radiology reporting has not yet arrived, but the era of AI-augmented reporting is imminent. Retrieval-augmented architectures represent a critical leap forward, offering a mechanism to constrain the stochastic nature of generative AI with verified clinical facts. While current hallucination rates preclude independent use, the potential for these systems to democratize access to high-quality diagnostics—particularly in underserved regions—is profound. Future implementation must prioritize "human-in-the-loop" workflows, robust uncertainty quantification, and the development of safety-critical metrics that penalize plausible fabrications.

1. Introduction

1.1 The Operational Crisis in Diagnostic Radiology

The practice of radiology, often described as the "eye of medicine," is fundamental to modern healthcare delivery. From trauma triage to oncological staging, medical imaging underpins a vast proportion of clinical decision-making [1]. However, the operational stability of this discipline is under severe threat. The 2025 Global Radiologist Report paints a stark picture of a profession at its breaking point. In a survey of radiologists across multiple continents, 53% of respondents identified burnout as their single most pressing professional concern [2]. This figure is not merely a reflection of dissatisfaction but a symptom of a systemic imbalance between supply and demand.

The demand for medical imaging is growing at a rate that far outstrips the production of new radiologists [3]. Aging populations in developed nations require more frequent and complex imaging (CT, MRI) to manage chronic diseases [4]. Simultaneously, workforce shortages (cited by 49% of radiologists as a top concern) and the "brain drain" of locally trained clinicians leaving for better opportunities (40%) have created a vacuum in many healthcare systems [2, 5]. The consequences of this imbalance are tangible: increased patient wait times, delayed diagnoses, and a weary workforce prone to error.

Diagnostic error in radiology is a pervasive issue, often characterized as the elephant in the reading room [6]. Estimates suggest that diagnostic errors occur in 3% to 5% of interpretations, translating to approximately 40 million errors annually worldwide [7]. These errors are not random; they are frequently the result of cognitive fatigue, perceptual overload, and the sheer volume of cases a radiologist must interpret in a single shift [8, 9]. The convergence of burnout and error rates creates a compelling ethical and clinical imperative to develop automated solutions that can support the radiologist's cognitive load [10].

1.2 The Technological Paradigm Shift: From CAD to Foundation Models

Historically, the automation of radiology has been pursued through Computer-Aided Diagnosis (CAD) systems. These early tools, prevalent from the 1990s through the 2010s, were "narrow" AI—designed to detect specific abnormalities like lung nodules or breast microcalcifications. While helpful, they were limited in scope and often suffered from high false-positive rates, leading to "alert fatigue." They could flag a spot on an image, but they could not synthesize the findings into a coherent medical report [11].

The emergence of Foundation Models (FMs) represents a discontinuous leap in capability. Unlike narrow AI, foundation models are trained on massive, broad datasets using self-supervised learning, enabling them to adapt to a wide range of downstream tasks [12]. The "Transformer" architecture, which underpins Large Language Models (LLMs) like GPT-4 and BERT, allows these models to understand context, syntax, and semantics with unprecedented sophistication.

When applied to radiology, these become Multimodal Foundation Models (MFMs) or Vision-Language Models (VLMs). By processing pixel data (images) and text data (reports/clinical notes) simultaneously, MFMs aim to replicate the full workflow of a radiologist: perceiving the image, reasoning about the findings in the context of the patient's history, and generating a natural language report that communicates the diagnosis to referring physicians [12].

1.3 The Promise and Peril of Generative AI

The potential benefits of successful RRG automation are transformative.

- **Efficiency**: Automating the "drafting" of reports could reduce the time-per-case, allowing radiologists to focus on image interpretation and complex problem-solving rather than dictation.
- **Standardization**: AI can ensure that reports consistently use standard terminology and structural formats, reducing the variability that currently exists between different radiologists.
- **Global Health Equity**: AI-enabled workflows hold the promise of bridging the gap in healthcare access [13]. In low-resource settings where radiologists are scarce or nonexistent, an AI system that can generate a preliminary report for a chest X-ray or a screening mammogram could be lifesaving, serving as a triage tool to prioritize patients who need urgent care.

However, the generative nature of these models introduces a new and dangerous failure mode: Hallucination. In a text generation context, a model might invent facts to complete a sentence plausibly. In radiology, this translates to the AI describing a tumor that isn't there (insertion) or describing a healthy spine as fractured. Unlike a simple "miss" (false negative), a hallucination is often detailed, confident, and persuasive. The risk is that a tired radiologist might accept the AI's plausible but false report, leading to inappropriate treatment or unnecessary anxiety for the patient [14].

1.4 Research Objectives

This systematic review seeks to navigate the complex landscape of automated RRG by addressing the following critical questions:

1. **Clinical Accuracy**: How do current MFMs compare to human radiologists and previous generations of AI when evaluated on rigorous clinical metrics (e.g., F1CheXbert, F1RadGraph)?
2. **Technological Advancement**: How does the Fact-Aware Multimodal Retrieval Augmentation (FactMM-RAG) approach proposed by Sun et al. (2025) alter the performance landscape compared to standard generative approaches?
3. **Safety Profile**: What is the prevalence of hallucinations in current state-of-the-art models? Are these errors random, or do they follow specific patterns related to modality or pathology?
4. **Integration Barriers**: What are the non-technical hurdles—trust, liability, workflow integration—that currently prevent widespread adoption?

2. Literature Review

2.1 The Evolution of Automated Reporting Architectures

The journey toward automated radiology reporting has evolved through distinct "eras," each defined by the dominant machine learning architecture of the time. Understanding this evolution is crucial to appreciating the significance of current multimodal foundation models.

2.1.1 The Encoder-Decoder Era (CNN-RNN)

Prior to the transformer revolution, the dominant paradigm for image captioning (and by extension, report generation) was the CNN-RNN architecture.

- **Encoder (Vision)**: A Convolutional Neural Network (CNN), such as ResNet or DenseNet, was used to extract feature maps from the medical image. These maps represented the visual information (shapes, textures, opacities) in a compressed vector format.
- **Decoder (Language)**: A Recurrent Neural Network (RNN), typically a Long Short-Term Memory (LSTM) network, received these visual features and generated the report sequentially, word by word.

Limitations: While these models could generate simple sentences ("The heart is normal in size"), they struggled with long-range dependencies. They often failed to maintain coherence over a full paragraph and were prone to "repetition loops," generating the same phrase multiple times. Furthermore, they lacked "clinical reasoning"—they were essentially performing advanced pattern matching without an understanding of anatomical relationships or disease progression.

2.1.2 The Attention Mechanism and Transformers

The introduction of the Attention mechanism addressed the "forgetfulness" of RNNs. Attention allowed the model to focus on specific regions of the image while generating specific words (e.g., looking at the lung base while writing "pleural effusion").

The Transformer architecture replaced RNNs entirely, using self-attention to process the entire sequence of text simultaneously. This enabled the training of models on vastly larger datasets, leading to the development of Large Language Models (LLMs) like BERT and GPT.

In radiology, this transition manifested in models that could handle the complex, hierarchical structure of a radiology report (Findings section vs. Impression section). Researchers began to use BERT-based encoders to initialize the language component, significantly improving the grammatical quality and fluency of the generated reports [15].

2.1.3 Multimodal Foundation Models (2023–Present)

The current era is defined by Multimodal Foundation Models (MFMs). These are massive, pre-trained systems that align visual and textual representations in a shared semantic space.

- **Contrastive Learning**: Models like CLIP (Contrastive Language-Image Pre-training) and its medical variants (BiomedCLIP, CXR-CLIP) are trained by matching images to their corresponding text captions across millions of pairs. This forces the model to learn robust visual representations that are semantically linked to medical concepts [16].
- **Generative VLMs**: Models like GPT-4V, Gemini, and LLaVA-Med extend this by adding a generative decoder. They can take an image as input (along with a text prompt) and generate a novel textual response. This capability allows for "Zero-Shot" or "Few-Shot" performance, where the model can perform a task it wasn't explicitly trained for, simply by following instructions [12].

2.2 The Challenge of Evaluation: Why BLEU Fails

A pervasive theme in the literature is the inadequacy of traditional Natural Language Processing (NLP) metrics for evaluating radiology reports.

- **N-gram Metrics (BLEU, ROUGE)**: These metrics measure the overlap of words between the generated report and the "ground truth" report written by a radiologist. However, in medicine, a single word difference can invert the meaning.
  - Reference: "There is a pneumothorax."
  - Generated: "There is no pneumothorax."
  - BLEU Score: High (3 out of 4 words match).
  - Clinical Accuracy: Zero (Critical Error).

Because of this failure mode, the field has pivoted toward Clinical Efficacy Metrics:

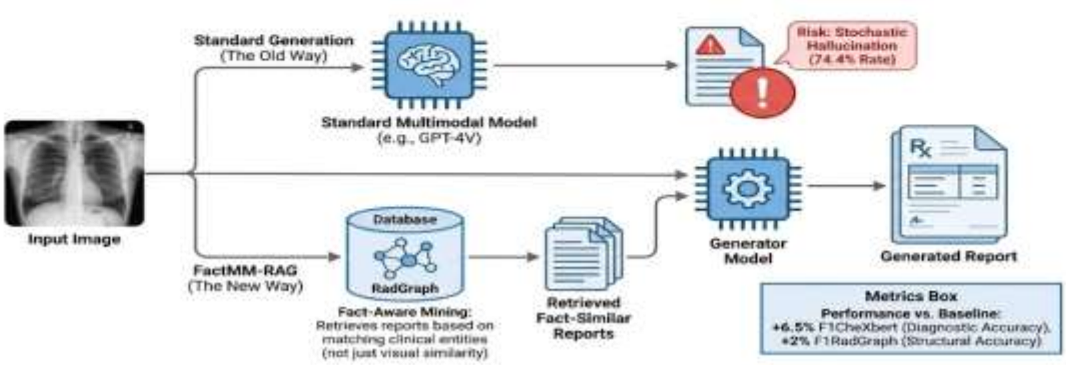
- **CheXbert**: This metric uses a BERT-based labeler to extract 14 common observations (e.g., Atelectasis, Cardiomegaly, Pneumonia) from both the reference and generated reports. It then calculates the F1 score of the agreement between these labels. This measures diagnostic accuracy rather than linguistic similarity [17].
- **RadGraph**: Developed to capture the complex structure of reports, RadGraph represents the text as a knowledge graph of entities (anatomy, observation) and relations (located_at, modifies). **F1RadGraph** measures the overlap between the graph of the generated report and the reference report. This is considered a gold standard for assessing factual completeness and structural correctness [17].
- **RadCliQ**: A composite metric designed to correlate better with human radiologist preferences, combining elements of BLEU and CheXbert [17].

2.3 Retrieval-Augmented Generation (RAG)

The most recent innovation, heavily featured in 2024–2025 literature, is Retrieval-Augmented Generation (RAG). The core insight of RAG is that instead of forcing a model to "memorize" all medical knowledge in its weights (which leads to hallucination), the system should be able to "look up" relevant information during the generation process.

Sun et al. (2025) pioneered FactMM-RAG, a system that retrieves "factually informed" report pairs from a training corpus based on the input image. These retrieved reports serve as templates or references, guiding the generative model to produce text that follows established clinical patterns while adapting to the specific features of the new image [16]. This approach attempts to combine the fluency of generative AI with the reliability of a retrieval-based system.

Figure 1: The Fact-Aware Multimodal Retrieval-Augmented Generation (FactMM-RAG) architecture.



2.4 AI and Health Equity

The literature also increasingly addresses the sociological impact of these technologies. **C**ommentary in Nature Digital Medicine argue that the value of AI in radiology cannot be measured solely by accuracy in elite academic centers. In global health contexts, where the alternative to AI is often "no imaging" or "interpretation by non-specialists," the threshold for utility may be different. AI tools have demonstrated the ability to reduce sepsis mortality and improve diabetic retinopathy screening in underserved populations by providing expert-level screening at scale [13]. However, this democratization is contingent on the models being robust to domain shifts (e.g., different X-ray machine manufacturers, different patient demographics) to avoid exacerbating existing biases.

3. Methodology

3.1 Search Strategy

This systematic review was conducted by aggregating and synthesizing research material from a diverse array of high-impact sources. The search strategy targeted literature published between 2020 and 2025, capturing the rapid ascent of transformer-based models and the subsequent foundation model era.

- **Databases**: The primary repositories accessed include arXiv (for preprints of rapidly evolving computer science methods), ACL Anthology (for NLP-specific advancements), PubMed/MEDLINE (for clinical validation studies), and the proceedings of major conferences such as RSNA (Radiological Society of North America), CVPR (Computer Vision and Pattern Recognition), and NAACL (North American Chapter of the Association for Computational Linguistics).
- **Keywords**: Search terms included combinations of "radiology report generation," "multimodal foundation models," "vision-language models," "hallucination," "retrieval-augmented generation," "clinical accuracy," and "automation."

3.2 Inclusion and Exclusion Criteria

To ensure the review focused on the most relevant and high-quality evidence, strict criteria were applied:

- **Inclusion**:
  - Studies evaluating Multimodal models (Text + Image). Unimodal (text-only or image-only) studies were excluded unless used as baselines.
  - Studies reporting Quantitative Clinical Metrics (F1CheXbert, F1RadGraph) or Human Expert Evaluation. Studies relying solely on BLEU/ROUGE scores were excluded as insufficient for determining clinical safety.
  - Research published or preprinted from 2023 onwards was prioritized to reflect the "Foundation Model" era, though seminal papers from 2020–2022 were included for context.
- **Exclusion**:
  - Studies evaluating only classification (e.g., "Pneumonia: Yes/No") without report generation.
  - Country-specific healthcare policy papers without technical or clinical evaluation.
  - Papers with insufficient methodological detail to reproduce or understand the architecture.

3.3 Data Extraction and Analysis

For each selected study, the following data points were extracted:

- **Model Architecture**: (e.g., LLaVA, GPT-4V, FactMM-RAG).
- **Dataset**: (e.g., MIMIC-CXR, CheXpert, internal hospital datasets).
- **Performance Metrics**: Numerical values for F1CheXbert, F1RadGraph, BLEU-4.
- **Safety Data**: Hallucination rates, critical error rates, omission rates.
- **Human Benchmarking**: Comparative performance against radiologists (residents vs. attendings).

The synthesis of this data follows a narrative approach, grouping findings by theme (Accuracy, Safety, Architecture) rather than a simple study-by-study summary. This allows for the identification of broader trends, such as the trade-off between plausibility and truthfulness.

4. Results: Clinical Accuracy of Multimodal Foundation Models

The assessment of clinical accuracy in automated radiology reporting has moved beyond simple pattern recognition to evaluating complex diagnostic reasoning and structural completeness. The results from 2024 and 2025 demonstrate significant progress yet highlight a persistent gap between AI capabilities and human expertise.

4.1 The Efficacy of Retrieval-Augmented Generation (FactMM-RAG)

The work of Sun et al. (2025) represents a benchmark in the effort to improve the factual correctness of generated reports. Their proposed architecture, FactMM-RAG, addresses the "hallucination" problem by retrieving factually relevant reports from a training corpus to guide the generation process [16].

4.1.1 Mechanism of Action

Standard RAG approaches typically retrieve documents based on visual similarity (images that look alike) or semantic similarity (texts that read alike). However, Sun et al. identified that visually similar images might have different clinical findings (e.g., a small pneumothorax is visually subtle but clinically distinct from a normal lung).

- **Fact-Aware Mining**: The researchers utilized RadGraph to annotate the training corpus, extracting entities and relations. They then mined pairs of reports that were factually similar—sharing the same clinical entities and relations—rather than just textually similar.
- **Universal Multimodal Retriever**: This "fact-aware" data was used to train a retriever that learns to find reports with matching clinical facts given an input image [16].

4.1.2 Performance Metrics on MIMIC-CXR

The performance of FactMM-RAG was evaluated on the MIMIC-CXR dataset, a standard benchmark consisting of chest X-rays and reports.

Table 1: Performance Comparison of FactMM-RAG vs. State-of-the-Art Baselines (MIMIC-CXR)

Model Architecture	F1CheXbert (Diagnostic Accuracy)	F1RadGraph (Structural/Factual Accuracy)	Improvement Mechanism
FactMM-RAG (Sun et al.)	0.605	0.249	Fact-Aware Retrieval
Med-MARVEL	0.581	0.239	Dense Retrieval
BiomedCLIP	0.540	0.229	Contrastive Pre-training
CXR-CLIP	0.536	0.228	Contrastive Pre-training
MedCLIP	0.528	0.225	Contrastive Pre-training
GLoRIA	0.512	0.220	Global-Local Attention
No Retriever (Baseline)	0.548	0.222	Direct Generation

Analysis:

- **Superiority of RAG**: The FactMM-RAG model significantly outperforms the "No Retriever" baseline (0.605 vs. 0.548 in F1CheXbert). This confirms that providing the model with reference material (retrieved reports) allows it to generate more accurate diagnoses than relying solely on its internal weights.
- **The Fact-Aware Advantage**: FactMM-RAG also outperforms other retrieval-based models like Med-MARVEL (0.605 vs 0.581). This validates the hypothesis that retrieving based on clinical facts (RadGraph) is more effective than retrieving based on generic multimodal embeddings.
- **F1RadGraph Significance**: The score of 0.249 in F1RadGraph, while the highest among automated methods, is still far below the "Oracle" score (theoretical maximum if the perfect report was retrieved, typically >0.40). This indicates that while the model is better at capturing clinical entities, it still struggles with the complex relational structure of a full radiology report (e.g., correctly linking modifiers like "severe" or "bilateral" to the correct anatomical locations) [16].

4.2 Comparative Diagnostic Performance: GPT-4V vs. Human Radiologists

While specialized models like FactMM-RAG are optimized for report generation, general-purpose foundation models like GPT-4V (OpenAI) and Gemini (Google) are increasingly tested for their zero-shot diagnostic capabilities. A series of studies published in Radiology and presented at RSNA 2024/2025 provide direct comparisons.

4.2.1 The "Diagnosis Please" Challenge

In a study involving 190 challenging "Diagnosis Please" cases, GPT-4V and Gemini Pro Vision were pitted against board-certified radiologists.

- **Radiologist Performance**: Achieved an overall diagnostic accuracy of 61%.
- **GPT-4V Performance**: Achieved 49% accuracy at the optimal temperature setting (T=1).
- **Gemini Pro Vision**: Significantly underperformed, with accuracy below 40% in most settings [18].

Table 2: Subspecialty Accuracy Breakdown (GPT-4V vs. Radiologists)

Subspecialty	Radiologist Accuracy	GPT-4V Accuracy (Temp=1)	Differential Gap
Overall	61%	49%	-12%
Chest Radiology	59%	75%	+16% (AI Superior)
Gastrointestinal (GI)	68%	24%	-44% (Human Superior)
Neuroradiology	Comparable	Comparable	Neutral

Deep Dive Analysis:

- **Chest Superiority**: The finding that GPT-4V outperformed radiologists in Chest Radiology (75% vs 59%) is striking. This is likely due to the massive prevalence of chest X-rays in the public datasets used to train these models. The model has seen more examples of chest pathology than any single human radiologist.
- **GI Weakness**: The catastrophic performance in GI radiology (24% vs 68%) highlights a critical limitation of current VLMs: Spatial Reasoning. GI diagnosis often involves understanding complex 3D relationships (e.g., bowel loops, obstructions) and temporal sequences (e.g., fluoroscopy). Current VLMs process images as "flat" snapshots and struggle with the spatial synthesis required for abdominal imaging.
- **The Temperature Factor**: The study found that GPT-4V performed better at a higher temperature (T=1) than at a lower temperature (T=0). In LLMs, higher temperature increases randomness and "creativity." For complex diagnosis, which requires generating a broad differential based on subtle cues, this "creativity" appears necessary. However, as discussed in the Safety section, this same parameter increases the risk of hallucination [18].

4.2.2 The Temporal Improvement of Foundation Models

Another study tracking performance over one year (RSNA 2023 vs RSNA 2024 questions) showed rapid improvement.

- OpenAI o1 (2024 model): Scored 59% on 2024 cases.
- GPT-4o: Scored 54%.
- Gemini 1.5 Pro: Scored 36%.
- Llama 3.2 90B: Scored 33%.

The 2024 model (OpenAI o1) reached statistical parity with the two senior radiologists in the study (59% vs 58% and 66%, p=0.99), suggesting that the "human gap" is closing rapidly for text-based reasoning tasks, though image interpretation remains the bottleneck [19].

4.3 Evaluation of "Critical Errors" with MAIRA-X

Accuracy metrics do not differentiate between a minor error (typo) and a major error (missed cancer). The MAIRA-X model, a multimodal AI fine-tuned specifically for chest X-rays, was evaluated using a "Critical Error" framework.

- Human Critical Error Rate: 3.0% (in the study sample).
- MAIRA-X Critical Error Rate: 4.6%.

While the AI still commits more critical errors than humans, the gap is narrow (1.6%). Furthermore, the "acceptability" of the AI-generated sentences was 97.4%, nearly identical to the human baseline of 97.8% [20]. This suggests that for specific, narrow domains like chest X-rays, fine-tuned MFMs are approaching a level of reliability that could support drafted reporting, provided a radiologist reviews the output.

5. Results: Safety and the Hallucination Epidemic

While accuracy results are promising, safety analysis reveals the fragility of these systems. In clinical medicine, the maxim is primum non nocere (first, do no harm). The current generation of MFMs, particularly Visual Large Language Models (VLLMs), struggles to adhere to this principle due to a high prevalence of hallucination.

5.1 Defining Hallucinations in Radiology

A hallucination in radiology is not merely a wrong answer; it is a Fabrication.

- **Fabricated Findings**: The model describes a pathology (e.g., "There is a 3cm mass in the left lung") that is completely absent from the image.
- **Misidentifications**: The model correctly sees an opacity but misidentifies the anatomy (e.g., calling the "right lung" the "left lung").

A comprehensive systematic evaluation published in Life (MDPI) quantified these errors across multiple state-of-the-art models [21].

5.2 Prevalence of Hallucinations

The study reported a shocking statistic: Hallucinations occurred in 74.4% of all assessments across the tested models. This indicates that currently, a generative VLM is more likely to hallucinate a finding than to produce a completely accurate report.

Table 3: Hallucination Rates by Model

Model	Hallucination Rate (Overall)	Plausibility of Hallucinations	Risk Level
Gemini 2.0	51.7% (Lowest)	95.6% (Highest)	Critical Risk
ChatGPT-4o	72.8%	93.9%	High Risk
LLaVA-Med	73.6%	65.6%	High Risk
Claude Sonnet 3.7	78.3%	87.8%	Severe Risk
Vision AI	79.2%	89.4%	Severe Risk
Perplexity AI	82.8%	74.4%	Severe Risk

5.3 The Plausibility Paradox

The data reveals a counter-intuitive and dangerous trend: The "better" the model, the more dangerous its hallucinations.

- Gemini 2.0 had the lowest overall hallucination rate (51.7%), making it the most "accurate" model.
- However, its Plausibility Score was 95.6%. This means that when it did hallucinate (which was still half the time), the hallucination was nearly indistinguishable from a correct report. It used correct medical terminology, proper syntax, and logical consistency.

Implication: This creates a high risk of Automation Bias. A radiologist reviewing a report from Gemini 2.0 might be lulled into a false sense of security by the report's professional tone and logical flow, causing them to overlook the fabricated finding. In contrast, a model like LLaVA-Med, which had lower plausibility (65.6%), might generate errors that are obvious and clumsy, making them easier for a human to catch [21].

5.4 Modality and Context Effects

- **Modality**: Hallucination rates were significantly higher in complex cross-sectional imaging compared to projectional imaging.
    - MRI: 78.3% hallucination rate.
    - CT: 71.7%.
    - X-ray: 73.9%.

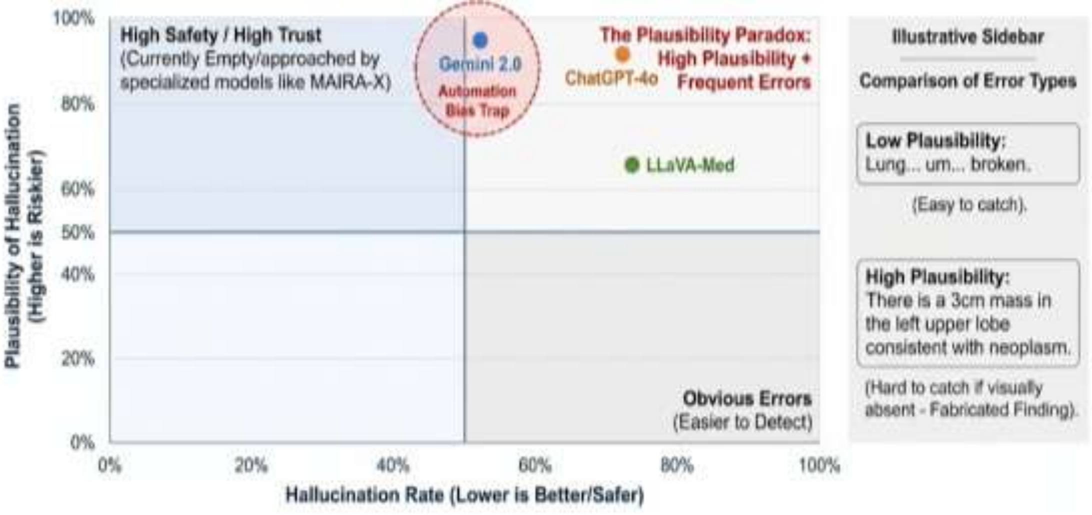
The high rate in MRI reflects the difficulty models have in interpreting multi-sequence, volumetric data compared to 2D X-rays [21].

- **Context**: Providing clinical context (e.g., "Patient has a history of cough") reduced the rate of fabricated imaging findings from 81.4% to 67.4%. This confirms that providing "anchors" in the form of patient history helps constrain the model's imagination. However, it also introduces **Context Bias**, where the model might "hallucinate" a finding that aligns with the history (e.g., hallucinating pneumonia because the history says "fever") even if it isn't visible on the scan [21].

5.5 Omission vs. Insertion

While "Insertion" (fabrication) is the dominant failure mode in generative VLLMs (accounting for >80% of errors in uncontextualized settings), "Omission" (missing a real finding) remains a significant issue in summarization tasks. A study on LLM summarization noted that omission of critical information is a distinct failure mode that requires separate detection mechanisms, often using "black-box" comparison methods to ensure the summary aligns with the source text [22].

Figure 2: The Plausibility paradox and safety risks in visual LLMs.



6. Discussion

6.1 The Disconnect Between Reasoning and Perception

The synthesis of findings from 2024 and 2025 paints a complex picture. MFMs have achieved Reasoning Parity with humans in text-based tasks (as seen in the RSNA 2024 question comparisons). They can generate differentials, synthesize history, and write fluent prose. However, they lack

Perceptual Grounding. The high hallucination rates (74.4%) indicate that the visual encoders (the "eyes" of the AI) are not yet reliably anchored to the pixel data. The models often "guess" based on the text prompt rather than "seeing" the pathology.

6.2 The Imperative of Retrieval Augmentation

The success of FactMM-RAG (Sun et al., 2025) suggests that the solution to hallucination is not simply "bigger models." Grounding the AI in retrieved, verified data (RAG) acts as a factual guardrail. By forcing the model to reference existing reports that match the clinical entities of the current case, RAG bridges the gap between the probabilistic nature of LLMs and the deterministic requirements of clinical reporting. This architecture likely represents the future blueprint for safe radiology AI [16].

6.3 Health Equity and the Global Context

Revisiting the work of Huang et al. (2024), the evaluation of these tools must be contextualized. In high-resource settings (e.g., US academic hospitals), a 4.6% critical error rate is a step backward from the human standard (3.0%). However, in low-resource settings—where the radiologist-to-population ratio can be 1:1,000,000 or worse—an automated report with a 95% accuracy rate is infinitely better than no report at all.

AI-enabled workflows have already demonstrated the ability to improve access to care for diabetic retinopathy and sepsis management in underserved populations.7 The challenge lies in ensuring that these models do not export bias—training on US/European data (like MIMIC-CXR) and deploying in Africa or Asia could lead to errors due to differences in disease prevalence (e.g., Tuberculosis) and equipment quality.

6.4 Barriers to Clinical Adoption

Despite the technological promise, non-technical barriers loom large.

- **Trust**: Surveys indicate that transparency is the #1 factor for physician trust (56%). The "black box" nature of deep learning, combined with the "Plausibility Paradox" of high-quality hallucinations, erodes this trust [23].
- **Liability**: Who is responsible when an AI hallucinates a tumor? The lack of legal clarity regarding liability for AI-generated errors remains a significant hurdle for hospital administrators [23].
- **Workflow Integration**: Radiologists do not want more "clicks." Successful adoption requires that AI be invisible—pre-drafting reports that appear in the dictation window ready for review, rather than requiring a separate login or interface.

6.5 Future Directions: Agentic AI and Uncertainty

The next frontier is Agentic AI. Instead of a passive model that generates a report in one shot, researchers are developing "agents" that can plan, critique, and verify their own work. An agent might generate a draft, then "look" at the image again to verify specific findings, or query a knowledge graph to check for contradictions.

Additionally, Uncertainty Quantification is vital. An AI should be able to say, "I see an opacity in the left lung, but I am only 60% confident it is pneumonia." Currently, most models project 100% confidence even when hallucinating. Developing metrics that reward calibrated uncertainty is essential for safety.

7. Conclusion

The landscape of automated radiology report generation has undergone a seismic shift with the advent of Multimodal Foundation Models. We have moved from the rigid, repetitive outputs of CNN-RNNs to the fluent, reasoned, but occasionally delusional outputs of Vision-Language Models.

Key Conclusions:

1. **Clinical Accuracy**: Retrieval-Augmented Generation (FactMM-RAG) is the current state-of-the-art, outperforming generative baselines by significantly improving factual and structural accuracy (F1CheXbert/F1RadGraph).
2. **Safety Gap**: General-purpose MFMs (GPT-4V, Gemini) are not yet safe for autonomous primary reporting due to hallucination rates exceeding 50%. The "Plausibility Paradox" makes these models

particularly risky for human-in-the-loop workflows.

3. **Human Comparison**: While AI approaches human performance in specific tasks (Chest X-ray diagnosis), it lags significantly in complex, spatial reasoning tasks (GI/MRI) and overall critical error rates.
4. **Equity Impact**: Despite imperfections, these tools hold immense potential to address the global radiologist shortage, provided they are deployed with robust safeguards and equitable validation (Huang et al., 2024).

Ultimately, the future of radiology AI is not "replacement" but "augmentation." The combination of a fatigued human radiologist and a fact-aware AI assistant—one that retrieves relevant priors, drafts routine text, and flags discrepancies—promises a system that is safer, faster, and more accurate than either human or machine alone.

References

1. Fourcade, A. and Khonsari, R.H., Deep learning in medical image analysis: A third eye for doctors. Journal of stomatology, oral and maxillofacial surgery, 120(4). 279–288 (2019).
2. Nwaiwu, V.C. and Das, S.K., An artificial intelligence road map to unlocking future technologies and transforming radiology practice. Medinformatics. 1–18 (2025).
3. Alexander, A., Jiang, A., Ferreira, C., and Zurkiya, D., An intelligent future for medical imaging: a market outlook on artificial intelligence for medical imaging. Journal of the American College of Radiology, 17(1). 165–170 (2020).
4. Nugent, R., Chronic diseases in developing countries: health and economic burdens. Annals of the New York Academy of Sciences, 1136(1). 70–79 (2008).
5. Jing, A.B., Garg, N., Zhang, J., and Brown, J.J., AI solutions to the radiology workforce shortage. npj Health Systems, 2(1). 20 (2025).
6. Lee, C.S., Nagy, P.G., Weaver, S.J., and Newman-Toker, D.E., Cognitive and system factors contributing to diagnostic errors in radiology. American Journal of Roentgenology, 201(3). 611–617 (2013).
7. Itri, J.N., Tappouni, R.R., McEachern, R.O., Pesch, A.J., and Patel, S.H., Fundamentals of diagnostic error in imaging. Radiographics, 38(6). 1845–1865 (2018).
8. Reiner, B.I. and Krupinski, E., The insidious problem of fatigue in medical imaging practice. Journal of digital imaging, 25(1). 3–6 (2012).
9. Pesapane, F., Gnocchi, G., Quarrella, C., Sorce, A., Nicosia, L., Mariano, L., Bozzini, A.C., Marinucci, I., Priolo, F., and Abbate, F., Errors in radiology: A standard review. Journal of Clinical Medicine, 13(15). 4306 (2024).
10. Cowen, J.E., Vigneswaran, G., Bekker, J., Brennan, P.A., and Oeppen, R.S., Human factors in diagnostic radiology: practical challenges and cognitive biases. European Journal of Radiology. 112248 (2025).
11. Computer-aided detection and diagnosis in medical imaging. 2015, Taylor & Francis.
12. Tavakoli, N., Shakeri, Z., Gowda, V., Samsel, K., Bedayat, A., Ghasemiesfe, A., Bagci, U., Hsiao, A., Leiner, T., and Carr, J., Generative AI and Foundation Models in Radiology: Applications, Opportunities, and Potential Challenges. Radiology, 317(2). e242961 (2025).
13. Abramoff, M.D. and Char, D., What do we do with physicians when autonomous AI-enabled workflow is better for patient outcomes? The American Journal of Bioethics, 24(9). 93–96 (2024).
14. D'Antonoli, T.A., Bluethgen, C., Cuocolo, R., Klontzas, M.E., Ponsiglione, A., and Kocak, B., Foundation models for radiology: fundamentals, applications, opportunities, challenges, risks, and prospects. Diagnostic and Interventional Radiology, (2025).
15. Mamdouh, D., Attia, M., Osama, M., Mohamed, N., Lotfy, A., Arafa, T., Rashed, E.A., and Khoriba, G., Advancements in Radiology Report Generation: A Comprehensive Analysis. Bioengineering (Basel), 12(7) (2025).
16. Sun, L., Zhao, J.J., Han, W., and Xiong, C. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. in Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). 2025.

17. Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E.P., Fonseca, E., Lee, H.M.H., Abad, Z.S.H., Ng, A.Y., Langlotz, C.P., Venugopal, V.K., and Rajpurkar, P., Evaluating progress in automatic chest X-ray radiology report generation. Patterns (N Y), 4(9). 100802 (2023).
18. Suh, P.S., Shim, W.H., Suh, C.H., Heo, H., Park, C.R., Eom, H.J., Park, K.J., Choe, J., Kim, P.H., and Park, H.J., Comparing diagnostic accuracy of radiologists versus GPT-4V and Gemini Pro Vision using image inputs from diagnosis please cases. Radiology, 312(1). e240273 (2024).
19. Hou, B., Mukherjee, P., Batheja, V., Wang, K.C., Summers, R.M., and Lu, Z., One year on: assessing progress of multimodal large language model performance on RSNA 2024 case of the day questions. Radiology, 316(2). e250617 (2025).
20. Sharma, H., Reynolds, M.C., Salvatelli, V., Sykes, A.-M.G., Horst, K.K., Schwaighofer, A., Ilse, M., Melnichenko, O., Bond-Taylor, S., and Pérez-García, F., Closing the Performance Gap Between AI and Radiologists in Chest X-Ray Reporting. arXiv preprint arXiv:2511.21735, (2025).
21. von der Stück, M.S., Vuskov, R., Westfechtel, S., Siepmann, R., Kuhl, C., Truhn, D., and Nebelung, S., Visual Large Language Models in Radiology: A Systematic Multimodel Evaluation of Diagnostic Accuracy and Hallucinations. Life, 16(1). 66 (2025).
22. Artsi, Y., Klang, E., Collins, J.D., Glicksberg, B.S., Korfiatis, P., Nadkarni, G.N., and Sorin, V., Large Language Models in Radiology Reporting—A Systematic Review of Performance, Limitations, and Clinical Implications. medRxiv. 2025.03. 18.25324193 (2025).
23. Gotta, J., Grünewald, L.D., Koch, V., Mahmoudi, S., Bernatz, S., Höhne, E., Biciusca, T., Gökduman, A., Wolfram, C., Booz, C., Scholtz, J.E., Martin, S., Eichler, K., Gruber-Rouh, T., Bucher, A., Yel, I., Vogl, T.J., and Reschke, P., Implementation of AI in radiology: the perspective of referring physicians. Insights Imaging, 16(1). 238 (2025).