applied sciences

MDPI

Article

Beyond Answer Accuracy: An Audit-First Framework for Claim-Level Faithfulness Evaluation of Multimodal Retrieval-Augmented Generation in Chest X-Ray Question Answering

Becky Firomssa Gudeta and Saurabh Singh *

Department of Artificial Intelligence, Woosong University, Daejeon 34606, Republic of Korea; begu010@student.kristiania.no
* Correspondence: singh.saurabh@wsu.ac.kr

Abstract

Radiology vision–language models can generate fluent chest X-ray answers that contain unsupported assertions when individual claims are not verifiably grounded in retrieved evidence. Answer-level similarity and coarse label agreement cannot detect such failures because they do not preserve the evidence chain behind each output. This paper proposes an audit-first evaluation framework that logs, for each query, the retrieval trace, a per-claim citation map, schema enforcement actions, and system telemetry. Claim-level faithfulness is measured by a Claim Support Score and a Hallucination Rate that sum to one by construction over the claims actually delivered. On a de-identified, credentialed-access chest radiograph dataset with three radiology vision–language model backbones, reference-aligned Claim Support Score without retrieval ranges from 0.605 to 0.655, and evidence-linked Claim Support Score with retrieval ranges from 0.724 to 0.760. These use different reference standards and are reported as complementary measurements, not compared directly. With retrieval, Hallucination Rate falls from 0.332 to 0.388 in the lowest retrieval-support band to 0.148 to 0.210 in the highest, suggesting evidence coverage as a candidate lever. The automated support labels have not yet been validated by independent clinical expert annotation, so findings are evaluation results rather than clinical safety evidence.

Keywords: retrieval-augmented generation; multimodal evaluation; radiology; chest X-ray; faithfulness metrics; hallucination detection; evidence attribution; claim-level verification; auditable artificial intelligence

Check for updates

Academic Editors: George Drosatos and Gang Wei

Received: 5 June 2026
Revised: 12 October 2026
Accepted: 14 October 2026
Published: 22 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Radiology vision–language models (VLMs) have reached a level of fluency that makes chest X-ray question answering appear clinically useful [1–4]. However, fluency is not a safety guarantee. A model can produce confident statements about laterality, device positioning, or finding severity that are not supported by any available evidence, and such errors are invisible when evaluation relies on answer-level similarity or coarse aggregate label metrics, which assess final text without preserving the evidence chain that produced it. This paper argues that radiology question answering evaluation must be auditable: each clinical claim should be traceable to evidence the system actually used at generation time, and system behavior, including retrieval quality, routing decisions, and schema enforcement policies, must be measured alongside answer quality.

1.1. Motivation and Evaluation Problem

Large-scale datasets such as MIMIC-CXR [5] and MIMIC-CXR-JPG [6] provide de-identified, credentialed-access paired radiographic studies and reports that enable systematic study of radiology question answering under conditions resembling clinical documentation. In this setting, the potential cost of unsupported assertions is high: a single incorrect laterality claim or an invented device position in a fluent answer can mislead clinicians operating under time pressure, and the radiology community has cautioned about such risks [7]. An assessment of language-model outputs on radiology reports documented incorrect statements and potentially harmful passages [8].

Existing evaluation approaches fall into three categories, each with important gaps. Proxy extractor methods such as CheXpert [9] and RadGraph [10] measure content alignment at the label or entity level but cannot verify that a generated claim is supported by the specific evidence shown to the model during generation. General retrieval-augmented generation evaluation frameworks, such as RAGAS [11], compute faithfulness scores over final outputs but do not log which evidence blocks were cited by which claims and do not account for the structured output schemas, repair mechanisms, and drop policies that are standard in deployed clinical systems. Hallucination detection methods for language models, such as SelfCheckGPT [12] and Chain-of-Verification [13], operate on generated text without access to the retrieval trace and therefore cannot distinguish between a claim hallucinated despite strong supporting evidence and a claim that is unsupported because retrieval failed to surface relevant passages.

Two design requirements follow from these gaps. First, evaluation must be evidence-linked: each delivered claim must be evaluated against the specific evidence blocks shown to the generator at generation time. An important clarification applies here: the retrieved evidence blocks are drawn from a corpus of radiology reports from other patients, not from the current query patient's own record. A claim labeled Supported means it is consistent with findings in retrieved similar cases, not that it has been independently verified against the current patient's radiograph. This distinction between evidence-linked support and patient-specific clinical accuracy is an inherent property of corpus-retrieval evaluation and is stated formally at the opening of Section 3. Second, evaluation must be repair-aware: many deployed radiology question answering systems include a schema requiring per-claim citations, a repair mechanism for schema violations, and a drop policy that removes uncited claims from the delivered answer. Scoring hallucinations before these enforcement steps are applied measures a system that does not exist in deployment; hiding enforcement behind post-processing obscures system behavior. The proposed framework scores faithfulness on the final delivered output after at most one repair attempt and reports enforcement actions as first-class telemetry.

Scope of claims. Two further boundaries apply throughout this paper. First, the non-retrieval and retrieval conditions are scored against different reference standards (the patient-specific reference report and the retrieved corpus evidence, respectively), so they are reported as complementary measurements and are not subtracted from one another. Second, the operational support labels are assigned by an automated procedure that has not yet been validated through independent clinical expert annotation. Statements about clinical safety or deployment are therefore framed as hypotheses for future validation, not as findings of this study.

1.2. System Context

The evaluation framework proposed here is system-agnostic and can be applied to any multimodal radiology retrieval-augmented generation pipeline that exposes equivalent audit artifacts. The specific tiered multimodal pipeline used as the evaluation target in

Section 4 is described in full architectural detail in the master's thesis from which this work is derived [14]. The present paper focuses exclusively on evaluation methodology: the audit bundle specification, the repair-aware claim-level faithfulness metrics, the operational support-label procedure, and the diagnostic analyses they enable. System details relevant to reproducing the evaluation conditions are summarized in Section 4.2.

1.3. Contributions

This paper makes the following contributions:

(1) We describe an audit-first evaluation harness that logs a structured audit bundle per query capturing the retrieved evidence blocks, packed context, per-claim citation map, schema enforcement outcome, and system telemetry. The bundle enables exact replay of the evidence window for every test case and supports diagnostic attribution of hallucinations to retrieval failure, generation failure, or enforcement policy.
(2) Repair-aware claim-level faithfulness metrics, the Claim Support Score (CS) and Hallucination Rate (HR), are defined over the final delivered claim set after schema enforcement and at most one repair pass. By construction, CS and HR are mutually exclusive and exhaustive over the delivered claim set and therefore sum to one for every query, so that scoring reflects what the system actually delivers rather than intermediate outputs before enforcement.
(3) We develop an operational support-label procedure and retrieval diagnostic that assigns each delivered claim one of three evidence-linked labels (Supported, Unsupported, and Contradicted) with full evidence provenance stored per decision, together with Retrieval Support at k (RS@k), a query-evidence concept coverage metric computed independently of the support labels and used to stratify hallucination behavior by retrieval quality.
(4) A within-condition diagnostic finding, visible only through the audit framework, is described. In the retrieval-augmented condition on MIMIC-CXR, HR decreases as RS@5 increases, from 0.332 to 0.388 in cases with RS@5 below 0.40 to 0.148 to 0.210 in cases with RS@5 above 0.60, across all three backbones. This is an associational, within-corpus observation that suggests evidence coverage as a candidate lever for hallucination reduction. It is not a causal claim; no cross-condition subtraction between non-retrieval and retrieval conditions is made, and generalizability beyond MIMIC-CXR requires multi-site validation as discussed in Section 6.4. Claim-level counts additionally separate unsupported from contradicted claims within each condition, a decomposition that aggregate proxy metrics cannot provide.
(5) We establish a coverage-penalized faithfulness metric, $\text{CPF} = \text{CS}_{\text{post}} \times (1 - \text{drop_rate})$, that jointly measures faithfulness and completeness and prevents metric inflation through aggressive claim suppression, together with a characterization of three radiology VLM backbones using absolute within-condition measurements with bootstrap confidence intervals, retrieval stack ablations, corpus composition ablations, enforcement strictness comparisons, and qualitative failure modes surfaced by the audit trace.

Together, these contributions reframe multimodal radiology question answering evaluation as an evidence-linked audit problem and provide instrumentation for building, comparing, and monitoring retrieval-augmented generation pipelines using automated faithfulness metrics, pending independent clinical validation of the support-label procedure.

The remainder of this paper is organized as follows. Section 2 reviews related work on radiology VLMs, retrieval-augmented generation evaluation, clinical proxy metrics, and hallucination detection. Section 3 presents the audit-first evaluation framework. Section 4 describes the experimental design. Section 5 reports experimental results. Section 6 discusses failure modes, limitations, and future work. Section 7 concludes the paper.

2. Related Work

This section reviews work most relevant to evaluating multimodal radiology question answering systems, covering radiology vision–language models, retrieval-augmented generation evaluation frameworks, structured clinical proxy metrics, and hallucination detection methods for large language models. For each line of work, we identify what it measures well and what remains absent for auditable claim-level evaluation. Table 1 summarizes how the proposed framework differs from the most directly related approaches across six evaluation dimensions.

Table 1. Comparison of evaluation approaches across six dimensions. Partial: the dimension is addressed incompletely or only in special cases.

Framework	Claim Level	Evidence Linked	Repair Aware	Schema Enforcement	Telemetry	Radiology Specific
CheXpert/CheXbert [9,15]	No	No	No	No	No	Yes
RadGraph [10]	Partial	No	No	No	No	Yes
RAGAS [11]	No	Partial	No	No	No	No
SelfCheckGPT [12]	Yes	No	No	No	No	No
Chain-of-Verif. [13]	Yes	No	No	No	No	No
FActScore [16]	Yes	Partial	No	No	No	No
This work	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**	**Yes**

Bold indicates the framework proposed in this paper.

2.1. Radiology Vision–Language Models

General instruction-tuned vision–language models have made open-ended visual question answering feasible across many domains [2], and radiology-oriented variants adapt these capabilities to chest imaging tasks. LLaVA-Rad [1] adapts the LLaVA visual instruction tuning paradigm to radiology, demonstrating that lightweight models can generate clinically styled report text when fine-tuned on paired chest imaging data. CheXagent [3] introduces a multimodal backbone trained on radiology-specific data covering report summarization, abnormality detection, and natural language question answering. CXR-LLaVA [17] extends multimodal generation to chest radiography with emphasis on alignment to radiological terminology and findings structure. At a larger scale, Med-PaLM [18] showed that a large language model adapted to medical question answering can approach expert performance on examination-style benchmarks, and assessments of language-model outputs in radiology have reported both useful performance and instances of incorrect or potentially harmful statements [7,8]. Tanno et al. studied a clinician-in-the-loop setting in which VLM outputs are refined by radiologists, illustrating how collaborative human–AI workflows can be evaluated [4].

These systems demonstrate that radiology question answering can be performed in natural language and that outputs can appear clinically plausible and task-oriented. However, model capability does not imply evaluability. Published evaluations remain focused on answer-level similarity scores, coarse label agreement, or selected qualitative cases, which can obscure rare but clinically important errors and do not preserve evidence traceability when retrieval or post-processing is used during inference.

2.2. Retrieval-Augmented Generation Evaluation

Retrieval-augmented generation was introduced by Lewis et al. [19] to ground language model generation in retrieved evidence, improving factuality on knowledge-intensive tasks. Dense passage retrieval [20] and the BM25 probabilistic relevance model [21] form the standard sparse–dense hybrid retrieval stack, with cross-encoder reranking via monoT5 [22] further improving evidence precision before context packing. Scalable nearest-neighbor

search over dense representations is supported by FAISS [23], which enables large-corpus retrieval at the scale required for de-identified, credentialed-access radiology datasets.

On the evaluation side, RAGAS [11] introduced automated metrics for faithfulness and answer relevance over retrieved context, and the survey by Yu et al. [24] highlights the need to measure retrieval quality and generation groundedness jointly. However, neither addresses the constraints specific to deployed clinical retrieval-augmented generation systems: structured output schemas requiring per-claim citations, repair mechanisms for schema violations, and drop policies that remove uncited claims from the final delivered answer. Most existing retrieval-augmented generation evaluation methods assume a single free-form output and do not incorporate these enforcement steps into the definition of hallucination, meaning that evaluations measure a system that does not match what is actually deployed.

Two further limitations concern the reference standard. First, corpus-retrieved evidence originates from different patients than the query patient, so claim-level support against retrieved evidence measures within-corpus consistency rather than patient-specific clinical accuracy. Second, when a system without retrieval is scored against a patient-specific reference report and a retrieval-augmented system is scored against retrieved evidence, the two scores rest on different reference standards. Differences between such scores cannot be read as the effect of retrieval alone, and evaluation designs should either use a common reference standard or report the two conditions as separate measurements.

2.3. Structured Clinical Proxy Metrics

Structured proxy extractors are widely used to score radiology text at scale because manual clinical evaluation is expensive. CheXpert [9] provides uncertainty-aware labels across fourteen clinical observation categories and supports macro-F1 benchmarking. CheXbert [15] extends this labeling using BERT-based encoding, improving robustness to linguistic variation. RadGraph [10] extracts clinical entities and relations, enabling comparisons that go beyond finding presence or absence to include relational structure.

These proxies are valuable for capturing broad content alignment trends across model families. Their systematic limitation for evidence-linked evaluation is that they operate on the final output text and cannot verify that a claim is supported by the evidence shown to the model during generation. A system can improve its CheXpert macro-F1 while still producing unsupported or misattributed claims, because both the generated text and the reference may agree on a finding category while disagreeing on a clinically consequential modifier such as laterality, severity, or polarity. A second limitation is that category-level proxy metrics collapse all errors into a single aggregate score, making individual hallucination types, such as direct contradictions and unsupported assertions, indistinguishable. The audit-first framework proposed in this paper addresses both limitations by verifying each claim against its cited evidence and recording the specific support label assigned.

2.4. Hallucination Detection in Large Language Models

SelfCheckGPT [12] detects hallucinations in a zero-resource black-box setting by sampling multiple outputs and measuring inter-sample consistency: claims that are inconsistent across samples are flagged as likely hallucinated. Although effective for models in general-domain, it does not use the retrieval trace and cannot distinguish between a claim hallucinated despite strong supporting evidence and a claim that is unsupported because the retrieval failed to surface relevant passages, a distinction that is critical for diagnosing retrieval-augmented generation pipelines. Chain-of-Verification [13] reduces hallucinations by prompting models to generate and independently answer verification questions, but does not produce an evidence-linked audit trail and grounds verification in model-generated

questions rather than an indexed evidence corpus. FActScore [16] decomposes long-form text into atomic claims and verifies each against a reference knowledge source, which is closely related to the claim-level evaluation approach proposed here. The key differences are that the knowledge source in this work is a query-time retrieved evidence set rather than a static knowledge base; citation binding is part of the generation schema rather than a post hoc step; and enforcement behavior such as repair and dropping is treated as part of the system being evaluated rather than as an external verification tool, with full evidence provenance stored per label decision to support clinician review.

None of the above methods log the retrieval trace or the enforcement actions of a deployed pipeline, so they cannot separately attribute hallucinations to retrieval failure, generative behavior, or policy behavior. This attribution capability is the central diagnostic contribution of the proposed framework. Clinician validation of automated support labels is a recognized gap across all automated evaluation approaches in this domain and is identified as the priority for future work in Section 6.5.

2.5. Gap Summary

Three gaps motivate the proposed framework. First, existing radiology evaluation methods verify content at the label or entity level but do not tie each claim to the retrieved evidence available during generation, and do not account for the fact that corpus-retrieved evidence originates from other patients rather than the current query patient. Second, general retrieval-augmented generation evaluation frameworks do not account for structured output schemas, enforcement policies, or repair-and-drop behavior standard in deployed clinical systems. Third, large language model hallucination detection methods do not log the retrieval trace and therefore cannot diagnose whether hallucinations originate from retrieval failure or from generative behavior despite available evidence, nor can they distinguish directly contradicted claims from merely unsupported claims. Table 1 shows that the proposed framework is the only approach among those compared that satisfies all six evaluation dimensions simultaneously. These dimensions concern evaluation instrumentation; whether the resulting metrics track the judgment of clinical experts is not established by this comparison and is addressed as a limitation in Section 6.4.

3. Audit-First Evaluation Framework

The framework is designed to measure whether each clinical claim is supported by evidence that was available to the system at generation time, and how system policies including schema enforcement, repair, and claim dropping affect the delivered output. Evaluation is treated as a logged process over structured artifacts rather than a post hoc score computed from free text. Figure 1 provides an overview of the pipeline from query input through audit bundle logging to metric computation.

Table 2 defines the core notation used throughout the paper.

Scope clarification: Evidence-linked support versus patient-specific accuracy. Evidence blocks $e_i \in E$ are passages drawn from a corpus $\mathcal{D}$ of de-identified, credentialed-access radiology reports from other patients, not from the current query patient's own record. A claim labeled Supported means it is consistent with findings in retrieved similar cases; it does not mean the claim has been independently verified against the current patient's radiograph. This is an inherent property of corpus-retrieval evaluation and distinguishes the evidence-linked CS and HR metrics from patient-specific accuracy measures such as sensitivity and specificity. Patient-specific content alignment is addressed complementarily through reference-aligned proxy metrics (CheXpert macro-F1 and RadGraph F1) reported in Table 3, which compare generated text to the reference report for the same study. Both

measurement types are necessary and complementary; neither is a substitute for the other, and neither has been validated here against independent clinical expert annotation.

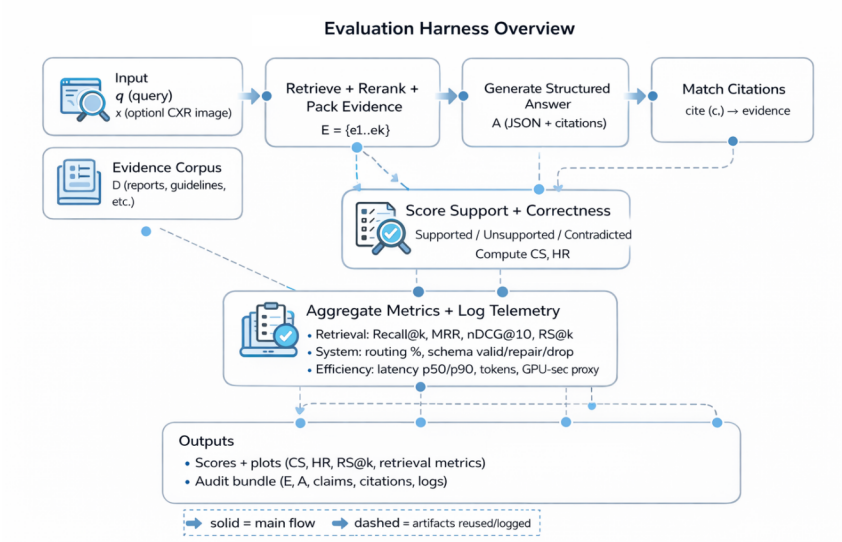


Figure 1. Audit-first evaluation pipeline. For each query q and optional image x, the harness routes the input to an inference tier, retrieves and reranks evidence blocks E, generates a structured answer $\hat{A}$, applies the validate-repair-drop enforcement policy π, and delivers the final answer A. Every intermediate artifact is stored in a persistent audit bundle. Downstream, the bundle feeds claim-level scoring (CS, HR, CPF), retrieval diagnostics (RS@k, Recall@k, nDCG@10), and enforcement and efficiency telemetry.

Table 2. Core notation.

Symbol	Definition
q, x	Natural language query; optional chest X-ray image
$\mathcal{D}$	Retrieval corpus of indexed radiology report passages
$E = \{e_1, \ldots, e_k\}$	Packed evidence set of k reranked blocks shown to the generator
$\hat{A}, A$	Raw generator output; final delivered answer after enforcement policy π
$\hat{C}$	Claims present after the repair pass and before the drop policy (pre-drop claim set)
$C = \{c_1, \ldots, c_m\}$	Atomic clinical claims in A after enforcement and dropping (delivered claim set)
$\text{cite}(c_i)$	Citation map: indices of evidence blocks cited for claim c_i
$C_{\text{sup}}, C_{\text{unsup}}, C_{\text{contra}}$	Subsets of Supported, Unsupported, and Contradicted claims in C
$t \in \{T_1, T_2, T_3\}$	Routing tier assigned to query q

Table 3. Stage 1 reference-aligned clinical proxy metrics (2021-case set). All values are computed against the patient-specific MIMIC-CXR reference report for each study. They are distinct from, and complementary to, the claim-level CS and HR in Table 4.

Model	Condition	Precision	Recall	CheXpert Macro-F1	RadGraph F1
LLaVA-Rad	Non-RAG	0.602	0.676	0.734	0.688
LLaVA-Rad	RAG	0.708	0.778	0.772	0.724
CheXagent	Non-RAG	0.632	0.704	0.744	0.693
CheXagent	RAG	0.724	0.790	0.770	0.721
CXR-LLaVA	Non-RAG	0.570	0.650	0.712	0.671
CXR-LLaVA	RAG	0.676	0.752	0.746	0.700

Table 4. Stage 1 claim-level results from audit-bundle logs (2021-case set; pooled over cases). CS = $|C_{sup}|/|C|$; HR = $(|C_{unsup}|+|C_{contra}|)/|C|$; CS + HR = 1.000 in every row by construction. Non-RAG rows: reference-aligned scoring against the patient-specific MIMIC-CXR report. RAG rows (bold): evidence-linked scoring against retrieved corpus evidence. The two conditions use different scoring bases and are complementary measurements, not a before-after comparison.

Model	Condition	$\|C\|$	$\|C_{sup}\|$	$\|C_{unsup}\|$	$\|C_{contra}\|$	CS	HR
LLaVA-Rad	Non-RAG (ref-aligned)	8890	5515	2190	1185	0.620	0.380
LLaVA-Rad	**RAG (evidence-linked)**	**8620**	**6290**	**1700**	**630**	**0.730**	**0.270**
CheXagent	Non-RAG (ref-aligned)	9450	6190	2220	1040	0.655	0.345
CheXagent	**RAG (evidence-linked)**	**9150**	**6950**	**1680**	**520**	**0.760**	**0.240**
CXR-LLaVA	Non-RAG (ref-aligned)	8250	4990	2050	1210	0.605	0.395
CXR-LLaVA	**RAG (evidence-linked)**	**7960**	**5760**	**1620**	**580**	**0.724**	**0.276**

3.1. Logged Artifacts and the Audit Bundle

For each query, the evaluation harness stores a structured audit bundle capturing all intermediate artifacts needed to reproduce the output and diagnose failures. Table 5 lists the six record groups stored per query, the key fields within each group, and their diagnostic purpose.

Table 5. Audit bundle record groups. Every field is stored for each test case, enabling exact replay of the evidence window and per-component attribution of failures.

Record Group	Key Fields Stored	Diagnostic Purpose
Input and context	Query, image ID, prompt template version, decoding config	Reproduces exact generation inputs; isolates prompt-wording effects
Routing decision	Assigned tier t, intent flags, RS@5 pre-score, conflict flags, escalation triggers	Attributes quality differences to routing policy; supports threshold calibration
Retrieved evidence	Full text of E, block IDs, source type, BM25/dense/reranker scores, index metadata	Separates retrieval failures from generation failures; enables post hoc coverage analysis
Generated output	Raw output, parsed $\hat{A}$, pre-drop claims $\hat{C}$, citation map cite(c_i), schema-normalized version	Preserves pre-enforcement content for comparison with the delivered answer
Enforcement trace	Schema validity flag, repair indicator, repaired $\hat{A}$, dropped claim indices, enforcement path taken, final A	Makes repair and drop behavior measurable and visible as first-class telemetry
System telemetry	Stage-level and end-to-end latencies, input/output token counts, GPU-seconds	Enables joint quality-efficiency reporting

Audit bundles are serialized as compressed JSON with a mean size of 9.4 KB per query, adding approximately 3.8 ms of overhead and requiring roughly 94 MB of storage per 10,000 queries, negligible relative to clinical PACS imaging volumes. The bundle enables two properties essential for scientific reproducibility: the evidence window and final output can be exactly reconstructed from stored artifacts, and failures can be attributed to retrieval, generation, or enforcement policy rather than to the system as a whole.

3.2. Claim Parsing and Citation Binding

The framework evaluates outputs at the level of atomic clinical claims. Each answer is decomposed into a claim set, where each claim c_i corresponds to a single clinically meaningful proposition such as "No pneumothorax is identified" or "Right-sided pleural effusion is present." The pre-drop claim set $\hat{C}$ consists of the claims parsed from the raw output $\hat{A}$ after the repair pass (when one is applied) and before the drop policy; the delivered

claim set $C = \{c_1, \ldots, c_m\}$ is the subset of $\hat{C}$ that survives the drop policy and constitutes the delivered answer A.

Atomic claim representation. The structured output schema requires the generator to produce claims as a list of atomic propositions. When the raw output contains multi-clause statements, a deterministic splitter separates conjunctions and attribute bundles into atomic claims wherever the split does not change clinical meaning. Each claim retains four lightweight clinical attributes: polarity (Present, Absent, or Uncertain), laterality (left, right, bilateral, or unspecified), severity modifier (for example small, moderate, or large, when stated), and an uncertainty marker (hedged or definitive). These attributes are necessary because many clinically consequential radiology errors involve laterality, severity, or polarity changes that are invisible to label-level proxy metrics.

Citation binding and integrity validation. For each claim c_i, the citation map $\text{cite}(c_i)$ records the integer indices of evidence blocks cited by the generator, creating a direct link between each claim and the exact passages available during generation. The harness validates that all indices in $\text{cite}(c_i)$ fall within $\{1, \ldots, k\}$. Out-of-range indices, missing citations, and syntactically invalid citation formats trigger the repair-and-drop procedure: a single structured repair pass re-invokes the generator with a correction instruction, and claims that remain invalid after repair are removed from A by the drop policy and do not enter the delivered claim set C used for post-enforcement metric computation. Figure 2 illustrates the full claim-level processing pipeline.

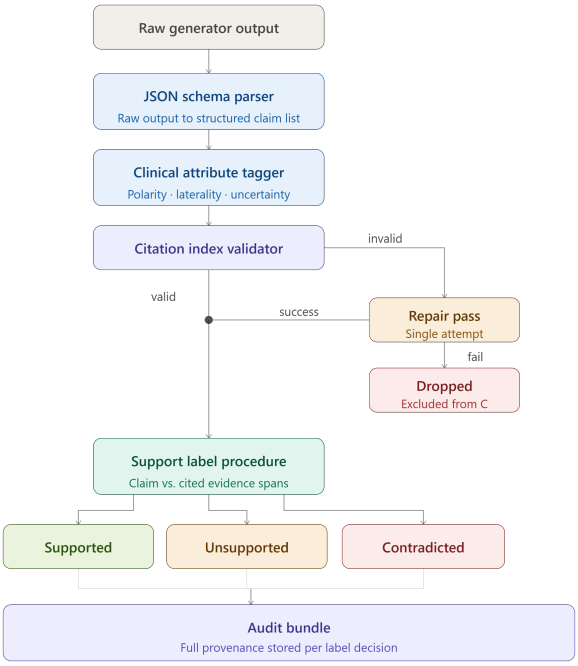


Figure 2. Claim-level processing pipeline. The raw generator output is parsed by the JSON schema validator. Each claim receives clinical attribute tags (polarity, laterality, severity, uncertainty). Citation indices are validated against the evidence set E. Invalid citations trigger one repair pass; claims remaining invalid after repair are dropped and excluded from C. The surviving claim set C is passed to the support-label procedure, which assigns each claim one of three labels (Supported, Unsupported, and Contradicted) by comparing the claim against its cited evidence spans, with full provenance stored in the audit bundle per decision.

3.3. Operational Support Labels

Each claim $c_i \in C$ is assigned one of three operational support labels by comparing the claim to the evidence blocks identified by $\text{cite}(c_i)$:

- Supported: The clinical proposition in c_i is entailed by at least one cited evidence block, consistent in polarity, laterality, severity, finding type, and uncertainty level.
- Unsupported: No cited evidence block contains information sufficient to entail c_i. This includes citations that are topically irrelevant, evidence that is relevant but incomplete, and cited evidence that says nothing about the finding or anatomical region named in the claim.
- Contradicted: At least one cited evidence block explicitly asserts a proposition inconsistent with c_i for the same finding and region, namely the opposite polarity, an incompatible laterality, or an incompatible severity modifier.

The distinction between the last two labels is deliberate. The absence of information in cited evidence yields Unsupported, not Contradicted. For example, a claim about the right lung that cites a block describing only the left pleural space is Unsupported, because the block neither entails nor refutes the claim. Contradicted requires an explicit inconsistent statement in a cited block, for example a claim of large effusions citing a block that states small effusions.

Hybrid label assignment procedure. Labels are assigned in two stages. In the first stage, clinical entities and findings in c_i and each cited block e_j are extracted using a radiology-domain named entity recognizer, and a finding-level overlap score is computed requiring the same finding type and anatomical region to appear in both. Claims reaching no positive overlap with any cited block, and claims with no valid citation, are labeled Unsupported. In the second stage, claims reaching a positive overlap score undergo attribute checking against the overlapping block: opposite polarity, incompatible laterality, or an incompatible severity modifier stated in the block results in Contradicted; consistent attributes yield Supported. Uncertainty handling is conservative: a definitively stated claim is labeled Unsupported when the evidence supports only a hedged assertion, and Contradicted when the evidence asserts the opposite polarity.

Repair and labeling independence. Schema repair attaches a citation to a claim but does not verify that the claim is semantically supported by its cited evidence. The support-label procedure is applied after enforcement, independently of whether the output required a repair pass; the entailment check runs on the claim text and cited evidence spans regardless of repair history. Repaired claims are flagged in the enforcement trace to enable targeted operator review.

Known limitations of the automated procedure. Three categories of clinical expression challenge the automated label assignment. Complex negation scope: Statements such as "no evidence of pneumothorax or pleural effusion" require correct parsing of negation across a conjunction; errors can flip polarity for one or both findings. Hedged uncertainty gradients: The procedure treats hedging as binary and does not model the gradient of certainty expressed by phrases such as "possible," "cannot be excluded," or "suspicious for." Implicit spatial reference: Findings described relative to anatomical landmarks without explicit laterality markers cannot be matched to explicit laterality terms in the claim and are labeled Unsupported with a flag for human review. Every support-label decision is stored in the audit bundle with the claim text, cited evidence indices, matched spans, and attribute values used to reach the decision, enabling a clinician to confirm or override any automated label without re-running the pipeline. The automated support labels are proxies for clinical expert judgment, have not been validated against independent radiologist annotation in this study, and are not equivalent to radiologist adjudication. CS and HR should therefore be interpreted as evidence-linked groundedness proxies rather than clinical accuracy or clinical safety measures. Large-scale clinician annotation of support labels is identified as the highest-priority future work in Section 6.5.

3.4. Metrics and Reporting

Table 6 provides the full metric glossary.

Table 6. Metric glossary for the audit-first evaluation framework.

Category	Metric	What It Measures	Source
Claim groundedness	CS	Proportion of delivered claims that are Supported	Audit bundle support labels
Claim groundedness	HR	Proportion of delivered claims that are Unsupported or Contradicted	Audit bundle support labels
Coverage-adjusted faithfulness	CPF	Faithfulness penalized for completeness loss from claim dropping	CS_{post} and drop rate
Retrieval quality	Recall@k, MRR, nDCG@10	Retrieval ranking effectiveness	Retrieval lists and relevance signal
Retrieval diagnostic	RS@k	Concept coverage of top-k blocks relative to query	Query concepts vs. evidence concepts
Enforcement telemetry	Routing %, schema valid %, repair rate, drop rate	Enforcement behavior and routing distribution	Audit bundle fields
Efficiency	Latency $p50/p90$, tokens, GPU-sec	Deployability and operational cost	Runtime logs
Clinical proxy	CheXpert macro-F1, RadGraph F1	Reference-aligned clinical content alignment	Proxy extractors vs. reference reports

3.4.1. Claim Support Score and Hallucination Rate

Let $C = \{c_1, \ldots, c_m\}$ be the delivered claim set after enforcement, with dropped claims excluded before any post-enforcement metric is computed. The three support labels are mutually exclusive and exhaustive over C, giving the partition:

$$C = C_{\text{sup}} \cup C_{\text{unsup}} \cup C_{\text{contra}}, \quad C_{\text{sup}} \cap C_{\text{unsup}} = C_{\text{sup}} \cap C_{\text{contra}} = C_{\text{unsup}} \cap C_{\text{contra}} = \emptyset. \quad (1)$$

The Claim Support Score and Hallucination Rate are defined as

$$\text{CS} = \frac{|C_{\text{sup}}|}{|C|}, \qquad \text{HR} = \frac{|C_{\text{unsup}}| + |C_{\text{contra}}|}{|C|}. \quad (2)$$

It follows directly from the partition in Equation (1) that

$$\text{CS} + \text{HR} = \frac{|C_{\text{sup}}| + |C_{\text{unsup}}| + |C_{\text{contra}}|}{|C|} = 1 \quad (3)$$

for every query with $|C| > 0$. Dataset-level values in this paper are reported as pooled (claim-level) ratios, in which the counts in Equation (2) are summed over all test cases before division; the pooled values also satisfy CS + HR = 1, as does the mean of per-query values. A test case with no delivered claim ($|C| = 0$) has no defined per-query CS or HR and contributes no claims to the pooled ratios, so pooled values are unaffected.

Relationship between CS and proxy F1. Suppose a reference-aligned precision and recall were defined as $P = |C_{\text{sup}}|/|C_{\text{gen}}|$ and $R = |C_{\text{sup}}|/|C_{\text{ref}}|$, where C_{ref} is a reference claim set. Then $\text{F1} = 2|C_{\text{sup}}|/(|C_{\text{gen}}| + |C_{\text{ref}}|)$, which equals $\text{CS} = |C_{\text{sup}}|/|C_{\text{gen}}|$ only when $|C_{\text{gen}}| = |C_{\text{ref}}|$. CS is therefore not an F1 score in general. In this paper, CS is computed from the support-label partition of the audit bundle, whereas the Precision and Recall reported in Table 3 are reference-aligned proxy measures computed by a different matching procedure, so F1 values derived from them need not equal CS. The two are reported in separate tables for this reason (Section 3.4.2).

3.4.2. Non-RAG and RAG Scoring Bases

Non-RAG scoring basis. Under the non-RAG condition, no evidence is retrieved and the structured output schema does not require per-claim citations. CS and HR for non-RAG

outputs are computed using a reference-aligned procedure that checks each claim against the MIMIC-CXR reference report for the same study, using the same hybrid label assignment procedure defined in Section 3.3 applied to reference-report passages rather than to retrieved evidence blocks. Because non-RAG outputs carry no citations, the whole reference report for the study serves as the evidence set for every claim and the valid-citation requirement is not applied: a claim is Unsupported when the report contains no entailing statement and Contradicted when the report explicitly asserts an inconsistent proposition. Non-RAG CS and HR therefore reflect content correspondence with a patient-specific reference report.

RAG scoring basis. Under the RAG condition, CS and HR are computed using the evidence-linked procedure defined in Section 3.3: each claim is evaluated against the specific evidence blocks retrieved for the same query. RAG CS and HR reflect whether generated claims are grounded in corpus-retrieved evidence.

Construct distinction. The two conditions use different reference standards and assess different system properties. They are complementary measurements of different constructs, not before-after measurements of the same construct. Accordingly, no difference, ratio, or percentage change between a non-RAG value and an RAG value is computed or interpreted anywhere in this paper, including for stratified values. Where both conditions appear side by side, each condition's own values are described separately, and the presentation is descriptive only. Evidence-linked and reference-aligned CS and HR from audit-bundle claim counts are reported in Table 4; reference-aligned proxy metrics (Precision, Recall, CheXpert macro-F1, RadGraph F1) are reported separately in Table 3. Statistical comparisons are performed only between configurations that share the same scoring basis (Section 3.5).

3.4.3. Coverage-Penalized Faithfulness

Post-enforcement CS can be inflated by strict enforcement policies that remove unsupported claims before scoring, because each removed claim leaves the denominator $|C|$. Three quantities are therefore reported for conditions with an enforcement policy. Pre-enforcement values CS_{pre} and HR_{pre} are computed with the same procedure over the pre-drop claim set $\hat{C}$ (all claims after the repair pass and before the drop policy), where claims without a valid citation are Unsupported because they cannot be linked to evidence. Post-enforcement values CS_{post} and HR_{post} are CS and HR over the delivered set C as defined above. The drop rate is drop_rate $= (|\hat{C}| - |C|)/|\hat{C}|$.

The Coverage-Penalized Faithfulness score is defined as

$$\text{CPF} = \text{CS}_{\text{post}} \times (1 - \text{drop_rate}). \quad (4)$$

Because $1 - \text{drop_rate} = |C|/|\hat{C}|$, CPF equals $|C_{sup}|/|\hat{C}|$, the fraction of all generated claims that both survive enforcement and are Supported. A system that drops every unsupported claim can reach $CS_{post} = 1.0$, but its CPF falls with every claim it removes, so CPF cannot be raised by suppression alone. CPF is reported together with the pre- and post-enforcement values in Table 7.

Table 7. Coverage-adjusted enforcement comparison (CheXagent, hybrid retrieval, 500-case Stage 2 set). CS_{pre} and HR_{pre}: after repair, before dropping. CS_{post} and HR_{post}: delivered claims only. CPF $= CS_{post} \times (1 - \text{drop_rate})$. Relaxed enforcement applies no citation validation, repair, or dropping.

Enforcement	CS_{pre}	HR_{pre}	CS_{post}	HR_{post}	Drop Rate	CPF
Strict (validate-repair-drop)	0.740	0.260	0.755	0.245	0.02	0.740
Relaxed (no citation check)	0.727	0.273	0.727	0.273	0.00	0.727

3.4.4. Retrieval Support at k and Its Independence from Support Labels

Let $\mathcal{F}(q)$ denote the set of clinical finding concepts extracted from query q and $\mathcal{F}(e_i)$ the concepts extracted from evidence block e_i. Retrieval Support at k is defined as:

$$\text{RS@}k = \frac{1}{|\mathcal{F}(q)|} \sum_{f \in \mathcal{F}(q)} \mathbb{I}[\exists\, i \in \{1, \ldots, k\} : f \in \mathcal{F}(e_i)]. \quad (5)$$

RS@k is the proportion of query clinical concepts covered by at least one of the top-k evidence blocks.

Independence from support labels. RS@k is computed from the query q and the retrieved evidence set E only, using the concept extractor applied independently to both. It uses neither the generated claims nor the support labels assigned to them, and it does not depend on the generator backbone. The support-label procedure uses the generated claims and their cited evidence spans; it does not use RS@k. There is therefore no circular dependency between RS@k and the support-label procedure: RS@k is a retrieval-side sufficiency estimate computed before generation; support labels are a generation-side grounding assessment computed after generation from disjoint inputs. Because RS@k is backbone independent, RS@5 band membership is identical across backbones on a given case set.

RS@k values reported in this paper are within-corpus estimates on MIMIC-CXR and reflect the concept distribution of the PubMedBERT-based encoder on that dataset. They should not be assumed to transfer to other corpora, institutions, or patient populations without recalibration, and RS@k should not be interpreted as an operational risk signal outside the controlled experimental conditions of this study, as discussed further in Section 6.4.

3.4.5. Clinical Proxy Metrics

CheXpert macro-F1 and RadGraph F1 are computed by applying the respective extractors to both the generated answer and the corresponding MIMIC-CXR reference report for the same study, then computing F_1 between the extracted label sets or entity-relation graphs. Precision and Recall are reported alongside them in Table 3 as reference-aligned proxy metrics, clearly separated from the evidence-linked CS and HR in Table 4. These proxy metrics are complementary signals that capture broad content alignment and support comparison with prior work that reports only proxy metrics. They do not verify evidence linkage and cannot detect the laterality, severity, and polarity errors that the claim-level procedure is designed to surface.

3.5. Statistical Protocol

Stage 1: within-condition analysis. Because non-RAG and RAG conditions use different scoring bases and assess different constructs (Section 3.4.2), no cross-condition paired difference tests are performed. Stage 1 results are reported as absolute pooled CS and HR values with 95% bootstrap confidence intervals computed independently for each condition. Each bootstrap replicate resamples test cases with replacement ($B = 10{,}000$), recomputes the pooled ratios from the resampled claim counts, and the two-sided interval is taken from the 2.5th and 97.5th percentiles of the replicate distribution.

Stage 2: paired analysis within a shared scoring basis. Stage 2 ablations compare configurations that share the same scoring basis: all are retrieval-augmented conditions scored with the evidence-linked procedure on identical cases. Each configuration is evaluated on the same 500 Stage 2 cases and per-case differences are computed before aggregation. Differences in CS and HR are summarized by the mean paired difference with a 95% confidence interval from a paired bootstrap ($B = 10{,}000$ resamples of cases) and a two-sided p-value

from a sign-flip permutation test ($B = 10{,}000$ permutations). Nonparametric procedures are used because per-case CS and HR are bounded and skewed.

Descriptive stratification. Values stratified by RS@5 band are reported separately within each condition and are descriptive; no inferential test and no cross-condition difference is computed for them.

Significance and effect sizes. A threshold of $\alpha = 0.05$ is applied to Stage 2 tests without correction for multiple comparisons, so p-values should be read as exploratory. Effect sizes are reported as absolute differences in CS or HR with their confidence intervals, which are directly interpretable on the 0 to 1 scale of the metrics.

4. Experimental Design

4.1. Dataset and Evaluation Splits

All experiments use the MIMIC-CXR dataset [5] and its MIMIC-CXR-JPG release [6], which provide 227,835 de-identified chest radiographic studies with paired free-text reports from Beth Israel Deaconess Medical Center. MIMIC-CXR and MIMIC-CXR-JPG are de-identified, credentialed-access resources distributed under the PhysioNet Credentialed Health Data License 1.5.0. Access requires completion of a PhysioNet data use agreement and Collaborative Institutional Training Initiative (CITI) human research training. Image inputs are restricted to frontal projections (posteroanterior and anteroposterior views) to maintain a consistent imaging context. The retrieval corpus $\mathcal{D}$ is constructed from the findings and impression sections of all training and validation split reports, yielding approximately 210,000 passage-level blocks indexed for both BM25 and dense retrieval. Test-split reports are excluded from $\mathcal{D}$ to prevent information leakage.

The Stage 1 primary evaluation set contains 2021 cases after frontal-projection filtering of the official MIMIC-CXR test split and is used for all backbone measurements. For Stage 2 analyses, a stratified subsample of 500 cases is drawn from the Stage 1 set with approximately 100 cases per finding stratum (pleural effusion, pneumonia, cardiomegaly, pneumothorax, and no finding), ensuring coverage of pathological and radiologically normal cases. A guideline-dependent subset of 180 cases drawn from the Stage 2 set contains prompts requiring normative information such as follow-up recommendations, contraindication checks, or dose thresholds; results on this subset are reported separately to isolate the effect of dynamic external evidence. All comparisons within Stage 2 are strictly paired: every configuration is evaluated on the same 500 cases and per-case differences are computed before aggregation. Values computed on the 2021-case set and on the 500-case set are not interchangeable, and each table caption states which set it uses.

The evaluation suite covers three task types: radiology QA (clinically phrased questions about findings, severity, and device positioning drawn from 30 prompt templates); findings summarization (structured summaries aligned with reporting conventions); and guideline-dependent QA (the 180-case subset requiring external normative knowledge). Prompt templates are held constant across all configurations.

4.2. System Under Evaluation

The evaluation target is a tiered multimodal retrieval-augmented generation pipeline described in full architectural detail in the master's thesis [14]. The system routes each query to one of three inference tiers: T_1 (rule-based template responses, approximately 2% of queries); T_2 (text-only retrieval-augmented generation with hybrid retrieval and monoT5 reranking, applied when RS@5 exceeds 0.55, approximately 79% of queries); and T_3 (full multimodal retrieval-augmented generation with image-aware inference, applied when RS@5 falls below 0.55, evidence conflicts are detected, or image-dependent intent is signaled—approximately 19% of queries). Within T_2 and T_3, the validate–repair–drop

enforcement policy described in Section 3.1 is applied to every output. The hybrid retrieval stack uses Elasticsearch 8.12.0 for BM25 sparse retrieval, a PubMedBERT-based dense encoder indexed with FAISS IVF1024 [23] for dense retrieval, and monoT5-3B [22] for cross-encoder reranking. Candidate retrieval returns the top 40 passages per modality, merged by reciprocal rank fusion before reranking. The top eight reranked passages are packed into a context window capped at 1200 tokens.

4.3. Model Backbones and Experimental Conditions

Three radiology vision–language model backbones are evaluated to check that observations are not specific to a single model. Table 8 summarizes the specifications. All models are loaded in FP16 precision with greedy decoding (temperature 0.0, maximum 512 output tokens) and no additional fine-tuning.

Table 8. Radiology vision–language model backbones evaluated. All models use greedy decoding ($T = 0.0$, max 512 output tokens), FP16 precision, and no additional fine-tuning beyond the published checkpoint.

Backbone	Parameters	Checkpoint or Source
LLaVA-Rad [1]	7B	Hugging Face: microsoft/llava-rad
CheXagent [3]	8B	Hugging Face: StanfordAIMI/CheXagent-8b
CXR-LLaVA [17]	7B	GitHub: ECOFRI/CXR_LLAVA

Stage 1 conditions are measured for all three backbones on the 2021-case set. (i) Non-RAG: The generator receives only q and x with no retrieved evidence and no citation schema. CS and HR are computed with the reference-aligned procedure against the MIMIC-CXR reference report for the same study. (ii) RAG (reports only): The generator receives the top-8 reranked evidence blocks from $\mathcal{D}$ with the citation schema and enforcement policy applied. CS and HR are computed with the evidence-linked procedure against the retrieved corpus evidence. These two conditions use different scoring bases and assess different system properties. They are reported as complementary measurements and are not subtracted from one another, as described in Section 3.4.2.

Stage 2 analyses fix CheXagent as the generator unless stated otherwise and examine four axes on the 500-case set: (a) corpus composition (reports only versus reports plus 14,200 PubMed abstracts and 830 ACR guideline sections); (b) retrieval stack configuration (BM25 only, dense only, and hybrid retrieval, evaluated for all three backbones); (c) schema enforcement strictness (strict, with one repair pass and drop on failure, versus relaxed, with no citation validation); and (d) routing policy (tiered versus always-T_3). All Stage 2 comparisons use the evidence-linked scoring procedure and share one scoring basis, which permits paired analysis within the retrieval-augmented condition. A stratified description of HR by RS@5 band, which also displays the non-RAG condition, is descriptive only (Section 5.2).

4.4. Implementation Details

All experiments run on a single node with one NVIDIA A100-80GB GPU and 256 GB system RAM under Ubuntu 22.04 LTS, using the Transformers library (v4.40.0) with all generators loaded in FP16 precision. Elasticsearch 8.12.0 provides BM25 retrieval with a 16 GB JVM heap; the FAISS IVF1024-PQ64 index is built from 210,000 passage embeddings of dimension 768 (index build time approximately 45 min on CPU). The monoT5-3B reranker is batched at 16 passages per forward pass on the same GPU as the generator. All prompt templates, audit-bundle schema definitions, and support-label rule sets are versioned

alongside experiment configuration files. Code and templates will be released publicly upon publication.

5. Results

Section 5.1 reports Stage 1 as two complementary measurements. Section 5.2 describes HR by retrieval-support band. Sections 5.3 and 5.4 report retrieval stack and corpus ablations. Section 5.5 reports coverage-adjusted enforcement analysis. Section 5.6 reports system telemetry, and Section 5.7 presents qualitative failure modes. No result in this section subtracts a non-RAG value from a RAG value.

5.1. Stage 1: Reference-Aligned and Evidence-Linked Measurements Across Backbones

Stage 1 reports two complementary measurements on the 2021-case set. Non-RAG outputs are scored with the reference-aligned procedure against the patient-specific MIMIC-CXR reference report; RAG outputs are scored with the evidence-linked procedure against retrieved corpus evidence. As established in Section 3.4.2, the two use different reference standards, so each condition is described by its own CS and HR with its own 95% bootstrap confidence interval, and no cross-condition difference is computed.

5.1.1. Claim-Level Results

Table 4 gives claim counts from the stored audit bundles for the 2021 Stage 1 cases. CS is $|C_{\text{sup}}|/|C|$, and HR is $(|C_{\text{unsup}}| + |C_{\text{contra}}|)/|C|$ from the support-label partition, so that $|C_{\text{sup}}| + |C_{\text{unsup}}| + |C_{\text{contra}}| = |C|$, and CS + HR = 1.000 in every row.

5.1.2. Within-Condition Confidence Intervals

Table 9 gives 95% bootstrap confidence intervals for each row of Table 4, computed independently within each condition. Because CS + HR = 1 in every bootstrap replicate, each HR interval is the complement of the corresponding CS interval.

Table 9. Within-condition 95% bootstrap confidence intervals for CS and HR (Stage 1, 2021-case set, B = 10,000 case-level resamples). Non-RAG rows use reference-aligned scoring; RAG rows use evidence-linked scoring. No cross-condition differences are reported because the scoring bases differ.

Model	Condition	CS	95% CI (CS)	HR	95% CI (HR)
LLaVA-Rad	Non-RAG	0.620	[0.610, 0.630]	0.380	[0.370, 0.390]
LLaVA-Rad	RAG	0.730	[0.720, 0.740]	0.270	[0.260, 0.280]
CheXagent	Non-RAG	0.655	[0.645, 0.665]	0.345	[0.335, 0.355]
CheXagent	RAG	0.760	[0.751, 0.769]	0.240	[0.231, 0.249]
CXR-LLaVA	Non-RAG	0.605	[0.594, 0.616]	0.395	[0.384, 0.406]
CXR-LLaVA	RAG	0.724	[0.714, 0.734]	0.276	[0.266, 0.286]

5.1.3. Reference-Aligned Clinical Proxy Metrics

Table 3 reports proxy metrics computed against the patient-specific MIMIC-CXR reference report, separately from the claim-level results in Table 4.

5.1.4. Observations from Claim-Level Data

Two observations follow from Table 4. Both are made within a condition or between two quantities computed for the same row, and neither compares a non-RAG value with an RAG value as an effect.

Observation 1: CS is not a proxy F1 score. The F1 score implied by the Precision and Recall in Table 3 differs from CS in every row. CS minus implied F1 is −0.017 (LLaVA-Rad,

non-RAG), −0.011 (LLaVA-Rad, RAG), −0.011 (CheXagent, non-RAG), +0.004 (CheXagent, RAG), −0.002 (CXR-LLaVA, non-RAG), and +0.012 (CXR-LLaVA, RAG). The sign varies across rows. The two quantities are produced by different matching procedures against different references (claim-level support labels versus reference-aligned proxy matching), and the sign of the gap is not interpreted further here; doing so would require case-level error analysis that is not performed in this study.

Observation 2: Composition of incorrect claims within each condition. Within each condition, the three-way partition separates Unsupported from Contradicted claims. In the reference-aligned (non-RAG) rows, Contradicted claims are 11.0 to 14.7 percent of delivered claims and Unsupported claims are 23.5 to 24.8 percent, so Contradicted claims make up 31.9 to 37.1 percent of incorrect claims. In the evidence-linked (RAG) rows, Contradicted claims are 5.7 to 7.3 percent of delivered claims and Unsupported claims are 18.4 to 20.4 percent, so Contradicted claims make up 23.6 to 27.0 percent of incorrect claims. Numerically, the Contradicted share is lower in the evidence-linked rows than in the reference-aligned rows for all three backbones. Because the two sets of labels are produced against different references, this ordering is reported descriptively and is not interpreted as an effect of retrieval. It illustrates the kind of decomposition that aggregate proxy metrics cannot provide. The labels are automated and have not been validated by independent clinical annotation, so no statement about the clinical consequence of either label type is made.

5.2. Retrieval Support Stratification

Table 10 reports HR by RS@5 band for each backbone on the 500-case Stage 2 set. RS@5 is computed from the query and the retrieved evidence set without involvement of the generator, so band membership is assigned once from the shared retrieval stack and is identical across backbones (n = 110, 178, and 212). Each condition is described separately. No difference between conditions is computed, and the table contains no cross-condition column.

Table 10. HR by RS@5 band, shown separately for each condition (500-case Stage 2 set; RAG column: default Stage 2 configuration with hybrid retrieval, monoT5 reranking, and strict enforcement). n is the number of cases per band and is identical across backbones because RS@5 does not depend on the generator. HR values are pooled within each band. The non-RAG column uses reference-aligned scoring and the RAG column uses evidence-linked scoring; values are not subtracted across conditions. The table is descriptive, and no inferential test is applied.

Model	RS@5 Band	*n*	HR (Non-RAG)	HR (RAG)
LLaVA-Rad	RS@5 < 0.40	110	0.364	0.361
LLaVA-Rad	0.40 ≤ RS@5 ≤ 0.60	178	0.367	0.279
LLaVA-Rad	RS@5 > 0.60	212	0.360	0.194
CheXagent	RS@5 < 0.40	110	0.334	0.332
CheXagent	0.40 ≤ RS@5 ≤ 0.60	178	0.336	0.253
CheXagent	RS@5 > 0.60	212	0.331	0.148
CXR-LLaVA	RS@5 < 0.40	110	0.393	0.388
CXR-LLaVA	0.40 ≤ RS@5 ≤ 0.60	178	0.396	0.308
CXR-LLaVA	RS@5 > 0.60	212	0.390	0.210

Within the non-RAG condition, HR is nearly constant across RS@5 bands (0.331 to 0.396), as expected for a condition that does not use retrieved evidence. Within the RAG condition, HR decreases monotonically with RS@5 for all three backbones: 0.332 to 0.388 in the lowest band, 0.253 to 0.308 in the middle band, and 0.148 to 0.210 in the highest band. This stratified pattern is not visible in aggregate HR, which averages over cases with different retrieval support, and it is observable only because the audit bundle stores the retrieval trace beside the generated output for every case. It is an associational, within-

corpus observation. The RS@5 thresholds of 0.40 and 0.60 reflect the concept distribution of the PubMedBERT-based encoder on MIMIC-CXR text and should not be assumed to transfer to other corpora without recalibration (Section 6.4).

5.3. Retrieval Stack Ablation

Table 11 reports the retrieval stack ablation for all three backbones on the 500-case Stage 2 set under reports-only RAG with strict enforcement. All rows share the evidence-linked scoring basis.

Table 11. Retrieval stack ablation (500-case Stage 2 set, reports-only RAG, strict enforcement). Prec@8 pre and post: evidence precision before and after monoT5 reranking. ΔCS and ΔHR are relative to the BM25-only row of the same backbone (both rows evidence-linked). Descriptive; no inferential test is reported.

Model	Retrieval	R@40	MRR	nDCG @10	RS@5	P@8 Pre	P@8 Post	ΔCS	ΔHR
LLaVA-Rad	BM25	0.80	0.48	0.64	0.60	0.56	0.63	ref	ref
LLaVA-Rad	Dense	0.76	0.44	0.60	0.57	0.53	0.61	−0.018	+0.018
LLaVA-Rad	Hybrid	0.88	0.55	0.71	0.66	0.58	0.69	+0.031	−0.031
CheXagent	BM25	0.81	0.49	0.65	0.61	0.57	0.64	ref	ref
CheXagent	Dense	0.77	0.45	0.60	0.57	0.54	0.61	−0.021	+0.021
CheXagent	Hybrid	0.87	0.54	0.70	0.65	0.58	0.68	+0.028	−0.028
CXR-LLaVA	BM25	0.79	0.47	0.63	0.59	0.55	0.62	ref	ref
CXR-LLaVA	Dense	0.75	0.43	0.59	0.56	0.52	0.60	−0.019	+0.019
CXR-LLaVA	Hybrid	0.86	0.53	0.69	0.64	0.57	0.67	+0.030	−0.030

Hybrid retrieval scores highest on every retrieval metric for all three backbones, and monoT5 reranking raises Precision@8 by 7 to 11 percentage points in every configuration. Hybrid retrieval is also associated with higher CS (by 0.028 to 0.031) and correspondingly lower HR than BM25-only retrieval within the evidence-linked condition.

The audit bundle stores the retrieval trace and the RS@5 value for every case, so the association between retrieval quality and CS can be examined per case and by band (Table 10). We make no causal claim: establishing a causal relationship between retrieval quality and CS would require randomized assignment of retrieval quality, which is not feasible in this observational setting. The pattern is consistent with a causal interpretation but does not constitute causal evidence.

5.4. Corpus Composition

Table 12 reports CS and HR for reports-only versus reports-plus-dynamic retrieval on the full Stage 2 set and on the 180-case guideline-dependent subset. Both rows use the evidence-linked scoring procedure with strict enforcement and therefore share one scoring basis.

Table 12. Corpus composition ablation (CheXagent, hybrid retrieval, strict enforcement). Top panel: full 500-case Stage 2 set. Bottom panel: 180-case guideline-dependent subset. Δ is reports-plus-dynamic minus reports-only. 95% bootstrap CI from 10,000 resamples; two-sided *p*-values from a permutation test.

Eval Set	Corpus	CS	HR	ΔCS	95% CI	*p*
Full ($n = 500$)	Reports only	0.755	0.245	ref	–	–
Full ($n = 500$)	Reports + dynamic	0.758	0.242	+0.003	[−0.004, +0.010]	0.31
Guideline ($n = 180$)	Reports only	0.731	0.269	ref	–	–
Guideline ($n = 180$)	Reports + dynamic	0.773	0.227	+0.042	[+0.018, +0.066]	0.001

Adding dynamic sources gives no statistically detectable aggregate difference in CS ($\Delta = +0.003$, $p = 0.31$) and a larger difference on the guideline-dependent subset ($\Delta = +0.042$, 95% CI [+0.018, +0.066], $p = 0.001$, uncorrected for multiple comparisons). This pattern is consistent with external normative sources helping selectively with queries that require normative evidence, and it is not an effect that has been independently verified.

5.5. Schema Enforcement Strictness and Coverage-Penalized Faithfulness

Table 7 reports three quantities for CheXagent with hybrid retrieval on the 500-case Stage 2 set, all with the evidence-linked procedure. Pre-enforcement values (CS_{pre}, HR_{pre}) are computed over all claims after the repair pass and before the drop policy, with claims lacking a valid citation counted as Unsupported. Post-enforcement values (CS_{post}, HR_{post}) are computed over the delivered claims. CPF is defined in Equation (4).

Three points follow from Table 7. First, CS_{post} is 0.028 higher under strict than under relaxed enforcement. Because decoding is deterministic, both runs start from the same raw outputs, so the 0.013 difference between the relaxed value (0.727) and CS_{pre} (0.740) is attributable to the repair pass, and the remaining 0.015 between CS_{pre} and CS_{post} to dropping. Second, CS_{post} exceeds CS_{pre} by construction: the dropped claims have no valid citation, are counted as Unsupported before enforcement, and leave the denominator afterward. The size of this gap (0.015) is therefore a measure of how much CS_{post} is raised by removal alone, and it is not evidence that the drop policy selects incorrect claims. Under strict enforcement, CPF equals CS_{pre} for the same reason (Section 3.4.3). Third, CPF is 0.740 under strict and 0.727 under relaxed enforcement, so the strict policy remains higher after the completeness penalty of a 2% drop rate.

5.6. System-Level Telemetry

Table 13 reports the operational profile of four routing configurations (CheXagent, hybrid retrieval, 500-case Stage 2 set). The overall $p90$ is the 90th percentile of the full mixed-tier latency distribution. With 2% of queries at T_1 and 79% at T_2, 81% of queries complete in roughly 200 ms or less, so the 90th percentile falls at about the 47th percentile of the T_3 distribution, near its median of 62 s. The sub-population $p90$ values for T_2 and T_3 are reported in separate columns.

Tiered routing resolves 79% of queries at T_2, where median latency is 70 ms against 62 s for the always-T_3 baseline, a reduction of more than 99% for those queries. The overall $p90$ under tiered routing remains about 62 s because 19% of queries still reach T_3. CS under tiered routing (0.755) and under always-T_3 (0.752) differ by 0.003, and no inferential test is reported for this comparison. The text-only configuration has the lowest CS (0.718), which is consistent with image-aware inference being associated with higher evidence-linked CS in this setting; the table does not isolate the contribution of image-aware inference for the queries that receive it.

Table 13. System-level telemetry under four routing configurations (CheXagent, hybrid retrieval, 500-case Stage 2 set). $p50$ values are medians. T2 $p90$ and T3 $p90$: 90th percentile within the Tier 2 or Tier 3 sub-population only. Overall $p90$: 90th percentile of the full mixed-tier distribution, which falls in the Tier 3 range when more than 10% of queries are routed to Tier 3.

Configuration	T1/T2/T3 (%)	Lat. $p50$	T2 $p90$	T3 $p90$	Overall $p90$	CS	HR
Always-T_3 (baseline)	0/0/100	62 s	–	74 s	74 s	0.752	0.248
Tiered (reports only)	2/79/19	70 ms (T_2) 62 s (T_3)	110 ms	74 s	≈62 s	0.755	0.245
Tiered (reports + dynamic)	2/76/22	72 ms (T_2) 63 s (T_3)	115 ms	76 s	≈63 s	0.756	0.244
Text-only T_2 (lower bound)	2/98/0	68 ms	108 ms	–	108 ms	0.718	0.282

5.7. Qualitative Failure-Mode Analysis

The audit bundle allows individual hallucinations to be traced to a pipeline component. The cases below illustrate how labels are assigned under the definitions in Section 3.3. They are stylized examples that reflect failure patterns of the kind the audit traces are designed to expose; they are not verbatim MIMIC-CXR text, which cannot be redistributed under the data use agreement. This section does not quantify how often each pattern occurs. A quantified attribution of failures to pipeline components needs a defined annotation protocol and is left to the clinician study in Section 6.5.

Failure mode 1: Retrieval-side failures. (a) Retrieval miss, labeled Unsupported. A query asks about right-sided pneumothorax, but all top-retrieved blocks describe only the left lung field. The model generates "No right-sided pneumothorax is identified" citing a block that describes the left pleural space. The label is Unsupported: the cited block contains no information about the right lung, so it neither entails nor refutes the claim. It is not Contradicted, because Contradicted requires an explicit inconsistent statement in a cited block. (b) Temporal evidence conflict, labeled Contradicted. Reports of a single other patient in the corpus may differ across time points. An earlier block states "moderate left pleural effusion" and a later block states "resolved effusion." A claim that "left pleural effusion is present" citing both blocks is Contradicted because one cited block explicitly asserts the opposite. Both cases arise from the retrieval configuration rather than from the generator, an attribution made possible by the stored retrieval trace.

Failure mode 2: Severity overgeneralization. Retrieved evidence states "small bilateral pleural effusions are present, left greater than right." The model generates "Large bilateral pleural effusions are present" citing this passage. The label is Contradicted, because the cited block explicitly states a severity modifier ("small") incompatible with "large." A category-level proxy such as CheXpert macro-F1 would score this claim as agreeing on the finding "pleural effusion present," so the error is invisible to it.

Failure mode 3: Schema repair recovering a valid claim. The model outputs a claim about cardiomegaly without a citation index. The repair pass attaches a citation to the retrieved block containing "cardiac silhouette is enlarged." The label procedure, applied independently of repair history, assigns Supported. The audit bundle records the output as "valid after one repair," which makes the compliance difficulty visible as a signal for prompt or schema refinement.

6. Discussion

6.1. What the Audit Framework Reveals

The framework surfaces two kinds of structure that aggregate proxy metrics do not show, and both depend on logging evidence traces and claim-level labels for every test case.

The first is the stratified pattern in Section 5.2. Within the retrieval-augmented condition, HR falls from 0.332 to 0.388 in the lowest RS@5 band to 0.148 to 0.210 in the highest band for all three backbones, whereas within the non-RAG condition HR is nearly constant across bands (0.331 to 0.396). Each condition is read on its own scoring basis and no difference between them is computed. Aggregate HR averages over the whole RS@5 distribution and cannot separate high-support from low-support cases, and CheXpert macro-F1, RadGraph F1, RAGAS, SelfCheckGPT, and FActScore do not link the evidence window shown to the generator to the claims delivered, so none of them can produce this stratification.

The second is the three-way decomposition of incorrect claims. Within each condition, the partition separates Unsupported from Contradicted claims (Observation 2 in Section 5.1). Two systems with the same HR can differ in this composition, and only claim-level labels with a three-way partition expose the difference. Because the labels are automated and not

clinically validated, the decomposition is offered as an analytic capability of the framework, not as a ranking of clinical risk.

Together, these capabilities (stratification by retrieval support, decomposition of incorrect claims, and attribution of individual claims to evidence gaps or conflicts in the audit trace) follow from the audit bundle design and are not available from evaluation methods that treat the final output as a black box.

6.2. Retrieval Coverage as a Candidate Lever in This Experimental Setting

Three observations from the retrieval-augmented condition point the same way. First, HR differs by 0.167 to 0.184 between the lowest and highest RS@5 bands within a single backbone (Table 10). Second, the three backbones differ by only 0.036 in evidence-linked CS (0.724 to 0.760, Table 4). Third, switching from BM25-only to hybrid retrieval is associated with CS differences of 0.028 to 0.031 (Table 11). Taken together, and within the MIMIC-CXR setting, variation in retrieval support is associated with larger differences in HR than variation in backbone. These quantities come from different case sets (500 and 2021 cases) and are compared only in rough magnitude, so the statement is a hypothesis for testing, not an estimate of relative effect.

Two qualifications apply. First, the backbones span a narrow capability range (7B to 8B parameter radiology models), and the relative importance of retrieval and generation may differ for more capable generators or for multi-step clinical reasoning. Second, the RS@5 thresholds of 0.40 and 0.60 and the absolute HR values reflect the MIMIC-CXR concept distribution, the PubMedBERT-based encoder, and the reporting style of a single institution. The finding is an associational pattern within a controlled setting, not a causal claim (Section 5.3), and its generalizability requires the multi-site validation described in Section 6.5.

6.3. Enforcement Behavior and Coverage-Penalized Faithfulness

Enforcement policies are usually reported as post-processing details. Table 7 shows why they should be reported as results. Post-enforcement CS rises when claims are removed because the removed claims leave the denominator, so CS alone cannot distinguish better grounding from heavier suppression. Here the dropped claims are uncited, count as Unsupported before enforcement, and raise CS from 0.740 to 0.755; this gap measures the effect of removal and is not evidence that dropped claims were selectively incorrect. CPF removes the incentive: it equals the fraction of all generated claims that survive and are Supported, so removing a claim cannot raise it unless the claim was Supported.

Relative to relaxed enforcement, strict enforcement has CS 0.028 higher and CPF 0.013 higher at a 2% drop rate. Under deterministic decoding, 0.013 of the CS difference corresponds to the repair pass and 0.015 to dropping. In this configuration, the gain therefore comes about equally from recovering citations through repair and from discarding uncited claims; this is observed in one backbone and one corpus and rests on automated labels. Reporting drop rate, pre-enforcement CS, and CPF together makes the faithfulness and completeness trade-off visible to those who operate or evaluate such systems.

6.4. Limitations

Non-RAG and RAG conditions use different reference standards. Non-RAG CS is reference-aligned, measuring correspondence with the patient-specific MIMIC-CXR report, while RAG CS is evidence-linked, measuring groundedness against corpus-retrieved evidence from other patients. The paper therefore does not quantify the benefit of retrieval: no difference, ratio, or percentage change between the two conditions is computed or

interpreted. Doing so would require one reference standard shared by both conditions (Section 6.5).

Automated labels without clinical validation. The support labels are assigned by an automated procedure and have not been compared with independent radiologist annotation. Three known failure modes (complex negation scope, hedged uncertainty gradients, and implicit spatial reference) will produce incorrect labels, and the boundary between Unsupported and Contradicted is itself sensitive to extraction errors. Label errors could be similar across configurations, which would limit their effect on relative comparisons, but they could also differ between conditions or retrieval settings; this has not been tested. CS and HR are therefore evidence-linked groundedness proxies. They are not measures of clinical accuracy, clinical safety, or deployment readiness, and nothing in this paper should be read as evidence of clinical safety. Every label decision is stored with its evidence provenance so that a clinician can confirm or override it without rerunning the pipeline.

Single dataset and exploratory statistics. All experiments use MIMIC-CXR, which reflects the reporting conventions, disease prevalence, and equipment of one US academic medical center over 2011 to 2016. RS@5 thresholds, absolute metric values, and routing distributions are specific to this corpus and encoder and should not be assumed to transfer without recalibration. Stage 2 tests are uncorrected for multiple comparisons and are exploratory. Table captions state the case set and configuration used because values from the 2021-case and 500-case sets are not interchangeable.

Static corpus, single modality, single-turn queries. The corpus is a static index of MIMIC-CXR reports and a fixed set of guideline documents, so incremental indexing, corpus drift, and distribution shift are not addressed. Inputs are frontal chest radiographs; other modalities would need their own claim decomposition rules and corpora, although the audit bundle design and the CS and HR definitions do not depend on modality. Queries are evaluated one turn at a time, and multi-turn dialog raises further questions about evidence coherence and claim consistency.

Coverage loss under strict enforcement. The 2% drop rate includes correctly stated claims that were removed because the repair pass could not recover a valid citation. CPF is designed to surface this cost, but a softer policy that recovers weakly cited claims has not been evaluated.

6.5. Future Work

Four directions follow from the findings and limitations.

(1) Clinician annotation and a common reference standard. A stratified study in which radiologists or radiology residents review automated labels, with inter-annotator agreement measured by Cohen's kappa, is the first priority. The sample should span all RS@5 bands and include the Unsupported and Contradicted boundary, negation, and implicit laterality cases. Clinician-verified claims could also serve as a single reference against which both non-RAG and RAG outputs are scored, which would permit a valid direct comparison of the two conditions. The audit bundle holds the information needed without rerunning the pipeline.
(2) Multi-site validation. Applying the harness to CheXpert [9], PadChest [25], and NIH ChestX-ray14 [26] with the same retrieval stack would show whether RS@5 thresholds and metric values transfer across institutions or require per-site recalibration.
(3) Temporal filtering. Failure mode 1(b) in Section 5.7 arises when passages from different time points of one patient are retrieved together. Date-aware filtering or provenance-aware ranking could be evaluated as a retrieval ablation within the existing harness.
(4) Softer enforcement and monitoring. Policies that recover weakly cited claims instead of dropping them could reduce coverage loss. Audit telemetry such as RS@5, drop

rate, and repair rate could feed a monitoring dashboard; any such use would require the clinical validation described in item (1) first.

7. Conclusions

This paper presented an audit-first framework for evaluating multimodal radiology question answering. It logs a structured audit bundle for each query, binds atomic claims to the evidence blocks they cite, and scores the claims actually delivered after schema enforcement with a Claim Support Score and a Hallucination Rate that sum to one by construction. A Coverage-Penalized Faithfulness score reports faithfulness together with completeness, so that strict enforcement cannot raise the score through removal alone.

On de-identified, credentialed-access MIMIC-CXR data with three radiology vision–language model backbones, the framework gave two complementary measurements. Reference-aligned CS without retrieval ranged from 0.605 to 0.655 (Table 4), and evidence-linked CS with retrieval ranged from 0.724 to 0.760. Because the two use different reference standards, they were reported side by side and not subtracted, and the paper does not estimate the benefit of retrieval.

Within the retrieval-augmented condition, HR decreased as retrieval support increased, from 0.332 to 0.388 in the lowest RS@5 band to 0.148 to 0.210 in the highest. This associational, within-corpus pattern marks evidence coverage as a candidate lever for further study. Claim-level labels also decomposed incorrect claims into Unsupported and Contradicted within each condition. In the enforcement analysis, CPF was 0.740 under strict and 0.727 under relaxed enforcement at a 2% drop rate, with the CS difference split about equally between the repair pass and dropping. Tiered routing gave nearly identical evidence-linked CS to always-on multimodal inference (0.755 and 0.752) with a median latency of 70 ms for the 79% of queries resolved at Tier 2.

These results rest on automated support labels that have not been validated by independent clinical annotation, on one dataset, and on exploratory statistics. They show that the framework can instrument and stratify retrieval-augmented radiology question answering; they do not establish clinical safety or deployment readiness. Clinician annotation of the labels, a common reference standard for both conditions, and multi-site validation are the immediate next steps.

Author Contributions: Conceptualization, B.F.G. and S.S.; methodology, B.F.G.; software, B.F.G.; validation, B.F.G. and S.S.; formal analysis, B.F.G.; investigation, B.F.G.; data curation, B.F.G.; writing—original draft preparation, B.F.G.; writing—review and editing, B.F.G. and S.S.; visualization, B.F.G.; supervision, S.S.; project administration, S.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Woosong University (Woosong University Research Fund, 2026). No grant number was assigned. The funder had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Institutional Review Board Statement: This study used only the MIMIC-CXR (https://doi.org/10.13026/C2JT1Q) and MIMIC-CXR-JPG (https://doi.org/10.13026/8360-t248) datasets, which are de-identified, credentialed-access resources governed by the PhysioNet Credentialed Health Data License 1.5.0. Access requires completion of a PhysioNet data use agreement and Collaborative Institutional Training Initiative (CITI) human research training. The Institutional Review Board of Woosong University issued Exemption Determination IRB-2026-4582 on 11 January 2026, determining that this study does not constitute human subjects research under the Bioethics and Safety Act of the Republic of Korea, as no individually identifiable patient data were accessed, no patient contact occurred, and all data accessed are de-identified under a signed PhysioNet data use agreement.

Informed Consent Statement: Not applicable. This study used only de-identified, credentialed-access datasets and involved no patient contact.

Data Availability Statement: The MIMIC-CXR dataset analyzed in this study is a de-identified, credentialed-access resource available subject to a PhysioNet data use agreement at https://physionet.org/content/mimic-cxr/2.0.0/ (accessed on 1 April 2026). The MIMIC-CXR-JPG release is available under the same terms at https://physionet.org/content/mimic-cxr-jpg/2.0.0/ (accessed on 1 April 2026). The evaluation harness code, prompt templates, audit-bundle JSON schema definitions, and support-label rule sets will be deposited in a public repository upon publication and are available from the corresponding author on reasonable request in the interim. Derived claim-level counts underlying Table 4 are available from the corresponding author on reasonable request, subject to the PhysioNet data use agreement.

Acknowledgments: The authors thank the PhysioNet team and the MIMIC-CXR contributors for maintaining credentialed access to the dataset that made this study possible. Computational resources were provided by the Graduate School of Woosong University. B.F.G. thanks the members of the Woosong University Artificial Intelligence Laboratory for feedback on earlier versions of this work. During the preparation of this manuscript, the authors used an AI assistant (Claude, Anthropic) under the authors' direction for writing and editing the manuscript text. The authors designed all experiments, hypotheses, and analyses; specified and reviewed all code; ran every experiment on their own servers; verified every reported number against the experiment logs; and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Zambrano Chaves, J.M.; Huang, S.-C.; Xu, Y.; Xu, H.; Usuyama, N.; Zhang, S.; Wang, F.; Xie, Y.; Khademi, M.; Yang, Z.; et al. A clinically accessible small multimodal radiology model and evaluation metric for chest X-ray findings. *Nat. Commun.* **2025**, *16*, 3108. [CrossRef] [PubMed]
2. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2023; Volume 36.
3. Chen, Z.; Varma, M.; Xu, J.; Paschali, M.; Van Veen, D.; Johnston, A.; Youssef, A.; Blankemeier, L.; Bluethgen, C.; Altmayer, S.; et al. A vision-language foundation model to enhance efficiency of chest X-ray interpretation. *arXiv* **2024**, arXiv:2401.12208.
4. Tanno, R.; Barrett, D.G.T.; Sellergren, A.; Ghaisas, S.; Dathathri, S.; See, A.; Welbl, J.; Lau, C.; Tu, T.; Azizi, S.; et al. Collaboration between clinicians and vision-language models in radiology report generation. *Nat. Med.* **2025**, *31*, 599–608. [CrossRef] [PubMed]
5. Johnson, A.E.W.; Pollard, T.J.; Berkowitz, S.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.; Mark, R.G.; Horng, S. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **2019**, *6*, 317. [CrossRef] [PubMed]
6. Johnson, A.E.W.; Pollard, T.J.; Greenbaum, N.R.; Lungren, M.P.; Deng, C.; Peng, Y.; Lu, Z.; Mark, R.G.; Berkowitz, S.J.; Horng, S. MIMIC-CXR-JPG, a large publicly available database of labeled chest radiographs. *arXiv* **2019**, arXiv:1901.07042.
7. Shen, Y.; Heacock, L.; Elias, J.; Hentel, K.D.; Reig, B.; Shih, G.; Moy, L. ChatGPT and other large language models are double-edged swords. *Radiology* **2023**, *307*, e230163. [CrossRef] [PubMed]
8. Jeblick, K.; Schachtner, B.; Dexl, J.; Mittermeier, A.; Stüber, A.T.; Topalis, J.; Weber, T.; Wesp, P.; Sabel, B.; Ricke, J.; et al. ChatGPT makes medicine easy to swallow: An exploratory case study on simplified radiology reports. *Eur. Radiol.* **2024**, *34*, 2817–2825. [CrossRef] [PubMed]
9. Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proc. AAAI Conf. Artif. Intell.* **2019**, *33*, 590–597. [CrossRef]
10. Jain, S.; Agrawal, A.; Saporta, A.; Truong, S.Q.; Duong, D.N.; Bui, T.; Chambon, P.; Zhang, Y.; Lungren, M.P.; Ng, A.Y.; et al. RadGraph: Extracting clinical entities and relations from radiology reports. *arXiv* **2021**, arXiv:2106.14463.
11. Es, S.; James, J.; Espinosa-Anke, L.; Schockaert, S. RAGAS: Automated evaluation of retrieval augmented generation. *arXiv* **2023**, arXiv:2309.15217.
12. Manakul, P.; Liusie, A.; Gales, M.J.F. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv* **2023**, arXiv:2303.08896.
13. Dhuliawala, S.; Komeili, M.; Xu, J.; Raileanu, R.; Li, X.; Celikyilmaz, A.; Weston, J. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 3563–3578. [CrossRef]

14. Gudeta, B.F. Auditable Claim-Level Faithfulness Evaluation for Multimodal Retrieval-Augmented Generation in Radiology Question Answering. Master's Thesis, Graduate School of Woosong University, Daejeon, Republic of Korea, 2025.
15. Smit, A.; Jain, S.; Rajpurkar, P.; Pareek, A.; Ng, A.Y.; Lungren, M.P. CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1500–1519. [CrossRef]
16. Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; Yih, W.; Koh, P.W.; Iyyer, M.; Zettlemoyer, L.; Hajishirzi, H. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2023; pp. 12076–12100. [CrossRef]
17. Lee, S.; Youn, J.; Kim, H.; Kim, M.; Yoon, S.H. CXR-LLaVA: A multimodal large language model for interpreting chest X-ray images. *Eur. Radiol.* **2025**, *35*, 4374–4386. [CrossRef] [PubMed]
18. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. Large language models encode clinical knowledge. *Nature* **2023**, *620*, 172–180. Correction in *Nature* **2023**, *620*, E19. [CrossRef] [PubMed]
19. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2020; Volume 33, pp. 9459–9474.
20. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 6769–6781. [CrossRef]
21. Robertson, S.; Zaragoza, H. The probabilistic relevance framework: BM25 and beyond. *Found. Trends Inf. Retr.* **2009**, *3*, 333–389. [CrossRef]
22. Nogueira, R.; Jiang, Z.; Pradeep, R.; Lin, J. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 708–718. [CrossRef]
23. Johnson, J.; Douze, M.; Jégou, H. Billion-scale similarity search with GPUs. *IEEE Trans. Big Data* **2021**, *7*, 535–547. [CrossRef]
24. Yu, H.; Gan, A.; Zhang, K.; Tong, S.; Liu, Q.; Liu, Z. Evaluation of retrieval-augmented generation: A survey. *arXiv* **2024**, arXiv:2405.07437.
25. Bustos, A.; Pertusa, A.; Salinas, J.M.; de la Iglesia-Vaya, M. PadChest: A large chest X-ray image dataset with multi-label annotated reports. *Med. Image Anal.* **2020**, *66*, 101797. [CrossRef] [PubMed]
26. Wang, X.; Peng, Y.; Lu, L.; Lu, Z.; Bagheri, M.; Summers, R.M. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2017; pp. 2097–2106. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.