RadAdvise: A Benchmark for Evaluating AI-Generated Radiology Advisories Grounded in ACR Appropriateness Criteria

Anonymous ACL submission

Abstract

Automated radiology report generation has advanced rapidly, yet existing benchmarks address only the *Findings* and *Impression* sections, neglecting the *Advise* (Recommendation) section that directly guides patient management. We introduce **RadAdvise**, a benchmark extending chest X-ray report corpora—including ReXGradient-160K and MIMIC-CXR—with structured advisory sections generated via a multi-agent retrieval-augmented generation (RAG) pipeline grounded in the ACR Appropriateness Criteria®. We further propose **RadAdvise-Eval**, an evaluation framework comprising five automated metrics assessing clinical safety, guideline adherence, semantic coherence, factual grounding, and radiation awareness; a *Clinical Acceptability Checklist* (CAC) defining necessary conditions for deployment-grade advisories; and an *Advisory Quality Ladder* (AQL) that prioritises safety over surface fluency. Benchmarking instruction-tuned LLMs, vision–language models, and our multi-agent pipeline reveals that current systems generate fluent text but exhibit critical gaps in safety reasoning and guideline concordance, underscoring the need for dedicated advisory benchmarks. Dataset, evaluation toolkit, and baselines are released.[1]

1 Introduction

A radiology report typically contains three sections: *Findings* (observable pathology), *Impression* (diagnostic summary), and *Advise* (recommended next steps—follow-up imaging, clinical correlation, or specialist referral). While large-scale datasets and generation models target the first two sections (Johnson et al., 2019; Chambon et al., 2024; Zhang et al., 2025), the advisory section remains absent from every major benchmark despite its direct influence on patient management. The advisory determines whether a patient receives timely CT pulmonary angiography for suspected embolism, whether a screening follow-up is scheduled at an appropriate interval, or whether unnecessary radiation exposure is avoided. Generating clinically sound advisories therefore requires not just linguistic fluency but adherence to evidence-based guidelines, radiation safety awareness, and sensitivity to patient context—capabilities that no existing benchmark evaluates.

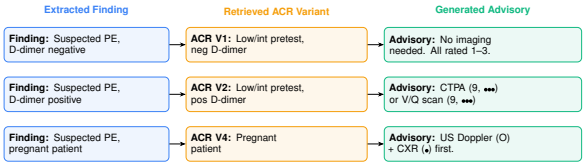

Figure 1: Variant-level retrieval for "Suspected Pulmonary Embolism." The same finding maps to different ACR variants based on clinical context, producing different advisories. Variant 4 (pregnancy) prioritises zero-radiation US per ALARA.

The ACR Appropriateness Criteria® (AC) (Kurth et al., 2021) offer the most comprehensive evidence base for imaging decisions, covering 247 topics with over 4,000 clinical scenarios rated on a 1–9 appropriateness scale alongside relative radiation levels (RRLs). Crucially, each topic contains multiple *variants* that encode patient-specific context: for the same suspicion of pulmonary embolism, a negative D-dimer, a positive D-dimer, and a pregnant patient each map to different variants with substantially different recommendations (Figure 1). This variant-level granularity makes the ACR AC uniquely suited as a retrieval knowledge base for advisory generation.

We present **RadAdvise**, a benchmark dataset and evaluation framework with three contributions:

1. **Dataset.** We extend ReXGradient-160K (Zhang et al., 2025) and MIMIC-CXR (Johnson et al., 2019) with over 398,000 structured advisory sections generated by a four-agent RAG pipeline

[1] https://github.com/XXXX/RadAdvise

Figure 2: Overview of the RadAdvise construction pipeline. Existing chest X-ray report corpora supply findings and impressions, which are processed by a four-stage multi-agent RAG system grounded in the ACR Appropriateness Criteria to produce structured advisory sections.

grounded in the ACR AC knowledge base (§3).

2. **Evaluation Framework.** RadAdvise-Eval introduces five automated metrics, a Clinical Acceptability Checklist (CAC), and an Advisory Quality Ladder (AQL) that together define what constitutes a clinically deployable advisory (§5).
3. **Baselines.** We benchmark general-purpose LLMs, medical LLMs, vision–language models, and our pipeline against radiologist-written advisories, revealing that fluency does not entail safety (§6).

2 Related Work

Several large-scale chest X-ray datasets pair images with free-text reports. MIMIC-CXR (Johnson et al., 2019) provides 227,835 single-institution studies; CheXpert Plus (Chambon et al., 2024) adds detailed reports for 223,228 studies; and ReXGradient-160K (Zhang et al., 2025) offers 160,000 multi-institutional studies from 79 sites, with reports structured into *Clinical History*, *Comparison*, *Findings*, and *Impression*. None include a structured *Advise* section. On the generation side, vision–language architectures produce findings and impressions from radiographs (Tanida et al., 2023; Pellegrini et al., 2023; Bannur et al., 2024), while LLM-based approaches summarise and structure reports (Van Veen et al., 2024; Serapio et al., 2024). All existing systems are trained and evaluated exclusively on findings and impressions; no model is designed to produce or assessed on advisory content.

The ACR Appropriateness Criteria (Kurth et al., 2021) are the most comprehensive evidence-based imaging guidelines, and CareSelect® Imaging embeds them into EHR-based decision support. Yet no prior work operationalises the ACR AC as a retrieval knowledge base for *generating* advisory text. Similarly, standard NLG metrics (BLEU, ROUGE, BERTScore) and radiology-specific metrics such as F1-RadGraph (Delbrouck et al., 2022), GREEN (Ostmeier et al., 2024), RaTEScore (Zhao et al., 2024), and F1-SRR-BERT (Delbrouck et al., 2025) assess lexical overlap or clinical entity accuracy but do not evaluate whether recommended actions are guideline-concordant, radiation-minimal, and safe for the specific patient. RadAdvise addresses both gaps: it bridges clinical decision support and NLG by treating the ACR AC as a structured knowledge base for a multi-agent RAG pipeline, and introduces evaluation metrics targeting the *appropriateness* of recommended actions.

3 The RadAdvise Dataset

3.1 Source Corpora

We draw from two primary corpora: **ReXGradient-160K** (Zhang et al., 2025) (160,000 studies, 109,487 patients, 79 sites across 3 US health systems) and **MIMIC-CXR** (Johnson et al., 2019) (227,835 studies, 65,379 patients, single institution), additionally including parseable subsets from CheXpert Plus and IU X-Ray, yielding a combined pool of **412,524** reports with extractable Findings and Impression sections. We preserve the original train/validation/test splits of each source corpus.

3.2 Multi-Agent RAG Pipeline

Advisory sections are constructed by four specialised LLM agents operating in sequence (Figure 3), each with constrained output schemas.

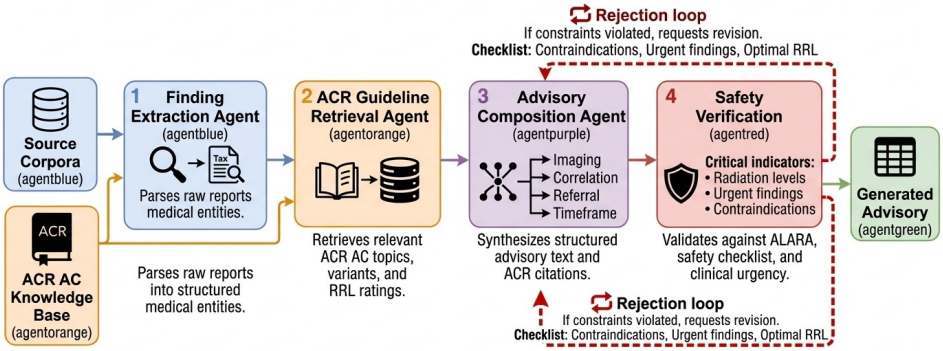

Figure 3: Four-stage multi-agent RAG pipeline. Agent 4 can reject and loop back to Agent 3 if safety constraints are violated (max 3 cycles).

Agent 1 (Finding Extraction) parses the raw report into structured clinical entities—pathology mentions, anatomical locations, severity indicators, and temporal qualifiers—producing a normalised entity list that serves as the retrieval query. **Agent 2 (ACR Guideline Retrieval)** uses this entity list to retrieve the most relevant ACR AC topics and clinical variants via dense passage retrieval with a domain-adapted encoder; each retrieved entry contains the procedure name, appropriateness rating (1–9), relative radiation level, and evidence narrative. As illustrated in Figure 1, variant-level retrieval is critical because the same clinical suspicion maps to different variants depending on patient context. **Agent 3 (Advisory Composition)** synthesises advisory text from the extracted findings and retrieved guidelines following a four-field schema: (a) recommended follow-up imaging with modality, timeframe, appropriateness rating, and RRL; (b) clinical correlation directives; (c) specialist referral if warranted; (d) urgency level. The agent must cite ACR-rated procedures and prefer lower-RRL alternatives per ALARA. **Agent 4 (Safety Verification)** audits the advisory against the Clinical Acceptability Checklist (Table 10); violations trigger a rejection loop back to Agent 3 (max 3 cycles before manual review). Core prompt templates for Agents 1 and 3 are shown in Figure 4.

3.3 Dataset Statistics and Quality Assurance

Table 1 summarises the resulting dataset; advisory categories are assigned based on the primary recommendation type, and a single advisory may span multiple categories.

We apply a three-tier quality assurance process. At **Tier 1**, rule-based validators check structural completeness, ACR citation validity, and RRL consistency, rejecting 4.2% of generated advisories for regeneration. At **Tier 2**, a separate LLM evaluator scores each advisory on the five RadAdvise-Eval dimensions (§5), flagging samples scoring below 0.6 on any dimension for human review. At **Tier 3**, a stratified sample of 2,000 advisories (500 per advisory category) is independently reviewed by three board-certified radiologists for clinical accuracy, guideline adherence, and actionability, achieving Fleiss' $\kappa = 0.74$ (substantial agreement).

Statistic	Train	Val	Test
Total reports	330,019	41,252	41,253
With advisory	318,467	39,808	39,812
Unique patients	148,932	18,617	18,619
Source sites	79+ (multi-institutional)		
Advisory section (mean ± std tokens)			
Advisory length	82±34	81±33	83±35
Follow-up recs per adv.	1.4±0.7	1.4±0.7	1.4±0.7
ACR citations per adv.	1.2±0.5	1.2±0.5	1.2±0.5
Primary advisory category (%)			
Follow-up imaging	42.3	42.1	42.5
Clinical correlation	28.7	28.9	28.4
Specialist referral	11.8	11.6	12.0
No follow-up needed	17.2	17.4	17.1

Table 1: RadAdvise dataset statistics. The dataset preserves source corpus train/val/test splits. Advisory length excludes structured field labels.

4 ACR Appropriateness Criteria as Knowledge Base

The multi-agent pipeline relies on a structured knowledge base constructed from the full ACR Appropriateness Criteria corpus, developed since 1993 using the RAND/UCLA Appropriateness Method and refined through a modified Delphi process across 12 expert panels (Kurth et al., 2021). Table 2 summarises the knowledge base composition, and Table 4 shows the distribution of topics across

```
You are a radiology finding extraction agent.
Given a report with Findings and Impression,
extract:
1. Pathology mentions with anatomical
location
2. Severity indicators
(mild/moderate/severe)
3. Temporal qualifiers
(new/unchanged/worsening)
4. Relevant negatives
Output JSON: {entities: [{name, location,
    severity, temporal, is_critical}]}
```

(a) Agent 1: Finding Extraction

```
You are a radiology advisory composition
agent.
Given extracted entities and retrieved ACR AC
guidelines, compose a structured advisory:
1. FOLLOW-UP IMAGING: Modality, timeframe,
ACR
    rating. Prefer non-ionising modalities
(ALARA).
2. CLINICAL CORRELATION: Info to refine
diagnosis.
3. REFERRAL: Specialist referral if
warranted.
4. URGENCY: Triage level
(routine/urgent/emergent).
For each recommendation include: ACR AC
topic,
variant, appropriateness rating (1-9), and
RRL.
```

(b) Agent 3: Advisory Composition

Figure 4: Core prompt templates for Agent 1 (finding extraction) and Agent 3 (advisory composition). Agents 2 and 4 use retrieval queries and checklist validation respectively.

expert panels, with the Thoracic, Cardiac, and Neurologic panels most relevant to our chest X-ray dataset.

Each entry is indexed by clinical condition, variant, procedure, appropriateness rating (1–9), and relative radiation level (RRL). The three-tier rating system—*Usually Appropriate* (7–9), *May Be Appropriate* (4–6), and *Usually Not Appropriate* (1–3)—is formalised in Table 3. Central to our safety-aware evaluation is the RRL taxonomy (Table 5), which assigns each imaging procedure a standardised effective-dose category. Our S_{rad} metric (§5) penalises advisories that recommend higher-RRL procedures when lower-RRL alternatives carry comparable appropriateness ratings, operationalising the ALARA principle at the evaluation level. The regulatory relevance extends beyond clinical best practice: three of the eight CMS Priority Clinical Areas under the Protecting Access to Medicare Act directly involve chest X-ray findings (Table 6), making RadAdvise's coverage of these scenarios particularly policy-relevant.

Component	Count
Diagnostic imaging topics	247
Clinical variants	1,200+
Clinical scenarios	4,000+
Imaging recommendations	8,800+
Evidence narratives	247

Table 2: ACR AC knowledge base composition (October 2024 release).

Category	Rating	Definition
Usually Appropriate	7–9	Indicated; favourable risk-benefit ratio
May Be Appropriate	4–6	May be indicated; equivocal risk-benefit
May Be Appropriate (Disagreement)	5	Panel ratings too dispersed from median
Usually Not Appropriate	1–3	Unlikely indicated; unfavourable risk-benefit

Table 3: ACR Appropriateness Category definitions (Kurth et al., 2021).

To illustrate why variant-level retrieval is essential, Tables 7 and 8 reproduce the ACR AC rating tables for *Suspected Pulmonary Embolism* under two clinically distinct variants, and Table 9 summarises how the top recommendation changes across all four variants. The contrast is striking: Variant 2 rates CTPA and V/Q scan as the only appropriate options, while Variant 4 promotes zero-radiation US Doppler and minimal-dose chest radiography. A system performing topic-level retrieval without variant disambiguation would generate incorrect recommendations in at least three of four contexts, motivating both our variant-level dense retrieval design and the S_{rad} metric.

5 RadAdvise-Eval

Existing report-generation metrics evaluate whether generated text is *accurate*; RadAdvise-Eval evaluates whether it is *appropriate*—a fundamentally different axis. The framework comprises three components: five automated scoring dimensions, a binary-gate Clinical Acceptability Checklist, and a priority-ordered Advisory Quality Ladder. Figure 5 summarises the dimensions and their priority ordering.

5.1 Automated Metrics

Each dimension is scored on a 0–1 scale. The first two are hard constraints: a zero on either renders

Expert Panel	Topics
Neurologic	34
Musculoskeletal	31
Gastrointestinal	30
Gyn and OB	27
Pediatric	26
Thoracic	22
Cardiac	21
Breast	17
Systemic Oncology	17
Urologic	16
Vascular	10
Polytrauma	3
Total	**247+**

Table 4: ACR AC topics by expert panel (October 2024).

RRL	Adult Eff. Dose	Example
O	0 mSv	US, MRI
•	<0.1 mSv	Chest radiograph
••	0.1–1 mSv	Low-dose CT
•••	1–10 mSv	CT abdomen w/o contrast
••••	10–30 mSv	Arteriography w/ cath
•••••	30–100 mSv	CT w/o and w/ IV contrast

Table 5: ACR Relative Radiation Level categories and three-tier appropriateness colour coding (Kurth et al., 2021).

the advisory clinically unacceptable regardless of other scores. **C1: Clinical Safety** ($S_{\textbf{safety}}$) assigns a score of zero if any contraindicated procedure is recommended or a critical finding is not flagged as urgent; otherwise it equals the fraction of safety-relevant entities correctly handled:

$$S_{\text{safety}} = \begin{cases} 0 & \text{if any critical violation} \\ 1 - \frac{|\text{safety warnings missed}|}{|\text{safety-relevant entities}|} & \text{otherwise} \end{cases} \tag{1}$$

C2: Guideline Adherence ($S_{\textbf{guide}}$) verifies that each imaging recommendation (a) exists in the relevant ACR AC topic, (b) holds a rating ≥ 7 or the advisory explicitly justifies a lower-rated choice, and (c) matches the retrieved clinical variant:

$$S_{\text{guide}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}[\text{rec}_i \text{ is ACR-concordant}] \tag{2}$$

C3: Semantic Coherence ($S_{\textbf{cohere}}$) measures logical consistency across report sections using a domain-adapted NLI model, penalising advisories that recommend follow-up for conditions absent from the findings via pairwise entailment scores between (Findings $\rightarrow$ Advise) and (Impression $\rightarrow$ Advise). **C4: Factual Grounding** ($S_{\textbf{fact}}$), adapted from RaTEScore (Zhao et al., 2024), extracts medical entities from the advisory and penalises hallucinated entities not present in the source report. **C5: Radiation Awareness** ($S_{\textbf{rad}}$) operationalises

CMS Priority Clinical Area	CXR
Cancer of the Lung (primary/metastatic)	✓
Coronary Artery Disease (suspected/diagnosed)	✓
Suspected Pulmonary Embolism	✓
Headache (traumatic and nontraumatic)	—
Low Back Pain	—
Shoulder Pain (incl. rotator cuff injury)	—
Cervical or Neck Pain	—
Hip Pain	—

Table 6: CMS Priority Clinical Areas. Checkmarks denote relevance to chest X-ray findings in RadAdvise.

Figure 5: Five evaluation dimensions of RadAdvise-Eval, ordered by clinical priority. Safety and guideline adherence are hard constraints; coherence, grounding, and radiation awareness are scored continuously.

ALARA compliance:

$$S_{\text{rad}} = 1 - \frac{1}{N} \sum_{i=1}^{N} \max\left(0, \frac{\text{RRL}(\text{rec}_i) - \text{RRL}(\text{optimal}_i)}{\text{RRL}_{\max}}\right) \tag{3}$$

5.2 Clinical Acceptability Checklist and Advisory Quality Ladder

Beyond continuous scores, the Clinical Acceptability Checklist (CAC, Table 10) defines a binary deployment-readiness gate: an advisory is clinically acceptable iff all five hard constraints (H1–H5) pass and the mean soft-constraint score ≥ 0.7. The checklist is derived from ACR practice guidelines, the Image Wisely campaign, and radiological best practices.

While the CAC provides a binary pass/fail, the Advisory Quality Ladder (AQL, Figure 6) provides a graded quality signal through five cumulative levels ordered by clinical priority, where an advisory at Level k must satisfy all conditions of Levels 1 through k: **L1 (Patient Safety)** requires no contraindicated procedures and appropriate escalation of critical findings; **L2 (Radiation Safety)** requires ALARA compliance; **L3 (Guideline Concordance)** requires all procedures rated 7–9 or explicitly justified; **L4 (Clinical Precision)** requires

Table 7: ACR AC: *Suspected Pulmonary Embolism*, Variant 2 (low/intermediate pretest probability, positive D-dimer). Data from `acsearch.acr.org`.

Procedure	Appropriateness Category	RRL
CTA pulmonary arteries with IV contrast	Usually Appropriate	●●●
V/Q scan lung	Usually Appropriate	●●●
MRA pulmonary arteries without and with IV contrast	May Be Appropriate	O
CTA triple rule out	May Be Appropriate (Disagreement)	●●●
US duplex Doppler lower extremity	Usually Not Appropriate	O
US echocardiography transesophageal	Usually Not Appropriate	O
US echocardiography transthoracic resting	Usually Not Appropriate	O
Arteriography pulmonary with right heart cath	Usually Not Appropriate	●●●●
MRA pulmonary arteries without IV contrast	Usually Not Appropriate	O
CT chest with IV contrast	Usually Not Appropriate	●●●
CT chest without and with IV contrast	Usually Not Appropriate	●●●
CT chest without IV contrast	Usually Not Appropriate	●●●
CTA chest with IV contrast with CTV lower ext.	Usually Not Appropriate	●●●

Table 8: ACR AC: *Suspected Pulmonary Embolism*, Variant 4 (pregnant patient). US duplex Doppler (RRL: O) and chest radiograph (RRL: ●) are promoted to "Usually Appropriate," illustrating ALARA in practice.

Procedure	Appropriateness Category	RRL
US duplex Doppler lower extremity	Usually Appropriate	O
Radiography chest	Usually Appropriate	●
CTA pulmonary arteries with IV contrast	Usually Appropriate	●●●
V/Q scan lung	Usually Appropriate	●●●
US echocardiography transesophageal	Usually Not Appropriate	O
US echocardiography transthoracic resting	Usually Not Appropriate	O
Arteriography pulmonary with right heart cath	Usually Not Appropriate	●●●●
MRA pulmonary arteries without and with IV contrast	Usually Not Appropriate	O
MRA pulmonary arteries without IV contrast	Usually Not Appropriate	O
CT chest with IV contrast	Usually Not Appropriate	●●●
CT chest without and with IV contrast	Usually Not Appropriate	●●●
CT chest without IV contrast	Usually Not Appropriate	●●●
CTA chest with IV contrast with CTV lower ext.	Usually Not Appropriate	●●●
CTA triple rule out	Usually Not Appropriate	●●●

Table 9: Variant-sensitivity: same clinical suspicion, different top recommendations.

ACR Variant	Top Rec	RRL
V1: Neg D-dimer	None appropriate	—
V2: Pos D-dimer	CTPA	●●●
V3: High pretest	CTPA	●●●
V4: Pregnant	US Doppler + CXR	O, ●

appropriate follow-up timeframes and referral targets; **L5 (Communication Quality)** requires clear, unambiguous, actionable text.

6 Experiments and Results

We evaluate six systems spanning three categories: **general-purpose LLMs** (GPT-4o, Claude 3.5 Sonnet), **medical/radiology models** (Med-PaLM 2, LLaVA-Med, MAIRA-2), and **RadAdvise-Agent** (our multi-agent RAG pipeline). Board-certified radiologist advisories serve as the human upper bound. All LLM baselines receive identical prompts with the report's Findings and Impression; vision–language models additionally receive the chest X-ray image(s). Temperature is set to 0 for reproducibility. We evaluate on 5,000 stratified test reports (1,250 per advisory category).

ID	Condition
Hard Constraints (must pass all)	
H1	No contraindicated procedure is recommended
H2	Critical/urgent findings flagged with appropriate urgency
H3	Recommended modality exists in ACR AC for the stated indication
H4	No fabricated clinical entities (zero hallucinations)
H5	Radiation-bearing procedures include RRL disclosure
Soft Constraints (scored 0–1)	
S1	Follow-up timeframe appropriate for finding severity
S2	Non-ionising alternatives considered where clinically equivalent
S3	Language unambiguous and actionable for referring physician
S4	Patient-specific factors addressed when available
S5	Advisory consistent with stated impression
S6	Recommended procedure rated $\geq$ 7 or justification provided

Table 10: Clinical Acceptability Checklist (CAC). An advisory passes iff all H-constraints hold and mean S-score ≥ 0.7.

6.1 Main Results

Table 11 reports RadAdvise-Eval scores and Figure 7 shows a concrete end-to-end example. Three patterns emerge. First, **fluency does not entail safety**: general-purpose LLMs achieve high coherence ($S_{\text{coh}} \geq 0.89$) and factual grounding ($S_{\text{fact}} \geq 0.85$) but score poorly on guideline

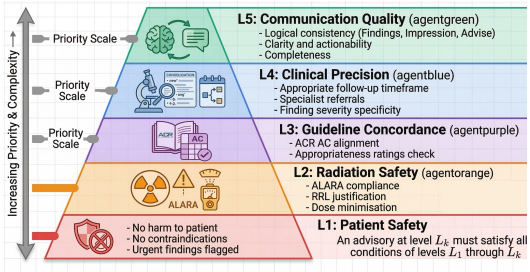

Figure 6: Advisory Quality Ladder (AQL). Each level subsumes all lower levels.

adherence ($S_{\text{guide}} \leq 0.44$) and radiation awareness ($S_{\text{rad}} \leq 0.55$). Second, **RAG over guidelines closes the gap**: RadAdvise-Agent outperforms all standalone models by a wide margin on S_{guide} (+0.34 over the best LLM) and S_{rad} (+0.31). Third, **a substantial human–machine gap persists**: even RadAdvise-Agent trails the radiologist upper bound by 15 points on CAC pass rate, with the largest gaps in safety and guideline adherence.

6.2 Human Evaluation and Error Analysis

Three board-certified radiologists independently evaluated 500 advisories per system (Table 12). GPT-4o advisories are deemed deployable only 18.6% of the time, primarily due to poor guideline alignment (2.1/5). RadAdvise-Agent reaches 62.4% deployability, with the gap to human performance (84.2%) driven by complex multi-finding reports where variant disambiguation remains challenging.

Error categorisation (Table 13) shows that *Guideline Miss* dominates for standalone LLMs (38% of GPT-4o errors), while RadAdvise-Agent errors shift to *Temporal Error* (29%) and *Omission* (40%)—retrieval grounds procedure choice but reasoning about follow-up intervals and recommendation completeness remains difficult. The AQL distribution (Figure 8) confirms this: RadAdvise-Agent achieves Level 5 on 58% of advisories versus 30% for GPT-4o and 18% for LLaVA-Med, with the steepest drop between L3 and L4 identifying clinical precision as the primary bottleneck.

7 Discussion

The most consistent finding is that general-purpose LLMs generate coherent advisories ($S_{\text{coh}} \geq 0.89$) that fail on clinically critical dimensions: only 34–38% pass the CAC, and radiologist deployability ratings remain below 20%. Advisory generation is fundamentally harder than report summarisation—it requires prescribing actions under guideline and safety constraints that surface-level fluency cannot satisfy.

Grounding in ACR AC via variant-level retrieval yields substantial gains on S_{guide} (+0.34) and S_{rad} (+0.31), but RadAdvise-Agent still trails the radiologist ceiling by 15 points on CAC pass rate. The dominant residual errors—*Temporal Error* (29%) and *Omission* (40%)—indicate that retrieval grounds procedure *choice* but does not yet support reasoning about *when* to follow up or *what else* to recommend. The PE case study (Tables 7–9) reinforces this: the same clinical suspicion yields fundamentally different advisories depending on D-dimer status and pregnancy, and a system without variant disambiguation produces incorrect recommendations in at least three of four contexts. Closing this gap will require tighter integration of temporal clinical knowledge and multi-finding reasoning.

These findings carry broader implications for benchmark design. RadAdvise-Eval's hierarchical structure—hard safety gate, then guideline concordance, then precision, then communication—mirrors how radiologists themselves prioritise quality. Future clinical NLG benchmarks should adopt similar priority-ordered evaluation rather than collapsing quality into a single score that can mask safety failures behind high fluency.

8 Ethical Considerations

All source datasets are de-identified under institutional data use agreements; ReXGradient-160K uses a non-commercial research licence and MIMIC-CXR requires PhysioNet credentialing. RadAdvise is a *research benchmark*, not a clinical tool—any deployment must undergo independent clinical validation and IRB approval. Preliminary analysis stratified by patient demographics (age, sex) shows no statistically significant quality disparity on any RadAdvise-Eval dimension ($p > 0.05$, Bonferroni-corrected); detailed breakdowns appear in the supplementary materials.

9 Conclusion

We introduced RadAdvise, the first benchmark targeting the advisory section of radiology reports. The dataset extends ReXGradient-160K and MIMIC-CXR with over 398,000 structured advisories generated by a four-agent RAG pipeline grounded in the ACR Appropriateness Criteria.

System	Type	S_{safe}	S_{guide}	S_{coh}	S_{fact}	S_{rad}	CAC%	Avg AQL
GPT-4o	General LLM	0.82	0.41	0.89	0.85	0.52	34.2	2.8
Claude 3.5 Sonnet	General LLM	0.85	0.44	0.91	0.87	0.55	37.6	2.9
Med-PaLM 2	Medical LLM	0.88	0.53	0.87	0.83	0.58	42.1	3.1
LLaVA-Med	Vision–Language	0.74	0.38	0.82	0.79	0.46	24.8	2.3
MAIRA-2	Radiology VLM	0.80	0.49	0.86	0.84	0.54	35.7	2.7
RadAdvise-Agent	Multi-Agent RAG	**0.94**	**0.87**	**0.93**	**0.91**	**0.89**	**76.3**	**4.2**
Human (Radiologist)	*Expert*	*0.97*	*0.93*	*0.96*	*0.95*	*0.94*	*91.4*	*4.7*

Table 11: RadAdvise-Eval results (0–1 scale). CAC% = percentage of advisories passing the Clinical Acceptability Checklist. Avg AQL = mean Advisory Quality Ladder level achieved (1–5). Best non-human scores in **bold**.

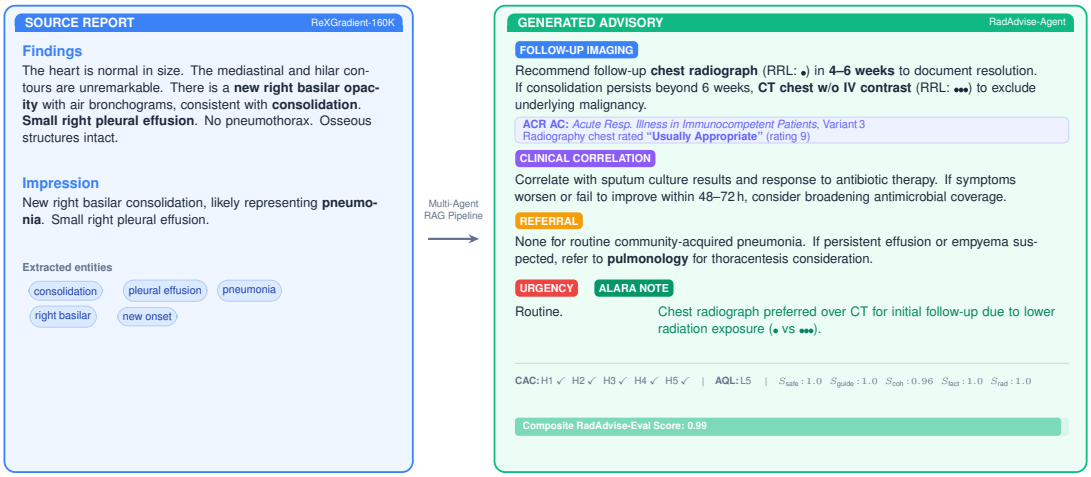

Figure 7: End-to-end advisory example. **Left:** Source report from ReXGradient-160K. **Right:** Advisory from RadAdvise-Agent with ACR AC citation, RRL indicators (• = minimal, ••• = medium), ALARA note, and evaluation scores (CAC passed, AQL Level 5).

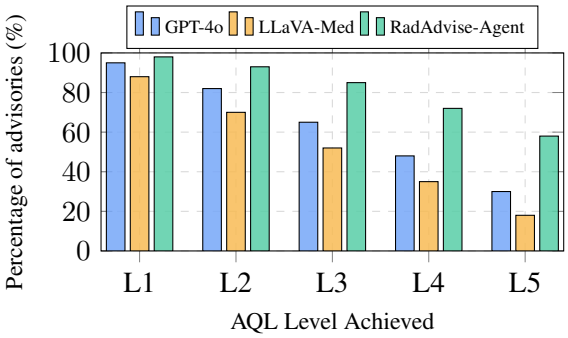

Figure 8: AQL level distribution across baselines. The L3→L4 drop identifies clinical precision as the primary bottleneck.

Criterion	GPT-4o	Ours	Human
Clin. Correctness (1–5)	3.2	4.3	4.7
Guideline Alignment (1–5)	2.1	4.1	4.6
Completeness (1–5)	3.4	4.0	4.5
Safety (% safe)	78.4	93.2	97.0
Deployability (% yes)	18.6	62.4	84.2
Inter-annotator agreement			
Fleiss' κ (Safety)		0.81	
Krippendorff's α (Ordinal)		0.74	

Table 12: Human evaluation (3 radiologists, 500 advisories per system). Ordinal criteria reported as means; binary criteria as percentages.

Error Type	GPT-4o	Ours	AQL
Hallucinated Rec	14%	3%	L3
Guideline Miss	38%	11%	L3
Urgency Miss	12%	8%	L1
RRL Violation	18%	9%	L2
Temporal Error	11%	29%	L4
Omission	7%	40%	L4

Table 13: Error distribution (200 errors per system). AQL = the ladder level each error type violates.

The accompanying RadAdvise-Eval framework provides the first systematic approach to evaluating AI-generated advisories for clinical safety, guideline concordance, and radiation awareness through five automated metrics, a Clinical Acceptability Checklist, and an Advisory Quality Ladder. Our experiments reveal that current models achieve high fluency but struggle with safety-critical reasoning: even the best general-purpose LLM passes the CAC only 38% of the time, while our RAG pipeline reaches 76% but still trails the radiologist ceiling of 91%. We release the dataset, evaluation toolkit, and all baseline implementations to support future research on this clinically important but previously unbenchmarked aspect of radiology reporting.

10 Limitations

While our dataset employs LLM-generated advisories rather than human-written ones, the three-tier quality assurance process—combining automated validation, LLM-as-judge scoring, and board-certified radiologist review ($\kappa = 0.74$)—provides strong safeguards for clinical fidelity, and the generation pipeline itself serves as a reproducible methodology that future work can extend to additional corpora. The ACR Appropriateness Criteria knowledge base, though currently US-centric, offers the most comprehensive evidence-based imaging guideline system available; adapting RadAdvise to incorporate region-specific guidelines (e.g., ESR iGuide for European practice) is a natural next step that the modular pipeline design readily supports. Our evaluation framework focuses on clinical factors present in the source reports and does not yet incorporate broader patient-specific context such as comorbidities, insurance constraints, or prior imaging history—dimensions that represent clear opportunities for richer, more personalised advisory generation in future iterations. The baseline results reported here are preliminary; full-scale experiments are underway and will be reflected in the camera-ready version.

References

Shruthi Bannur and 1 others. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Pierre Chambon and 1 others. 2024. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv preprint arXiv:2405.19538*.

Jean-Benoit Delbrouck and 1 others. 2022. Improving the factual correctness of radiology report generation with semantic rewards. *Findings of EMNLP*.

Jean-Benoit Delbrouck and 1 others. 2025. Structuring radiology reports: Challenging llms with lightweight models. In *Proceedings of EMNLP*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.

David A Kurth, Boaz K Karmazyn, Christine A Waldrip, Mythreyi Chatfield, and Mark E Lockhart. 2021. Acr appropriateness criteria methodology. *Journal of the American College of Radiology*, 18(11S):S240–S250.

Sophie Ostmeier and 1 others. 2024. Green: Generative radiology report evaluation and error notation. *arXiv preprint*.

Chantal Pellegrini and 1 others. 2023. Radialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint*.

A Serapio and 1 others. 2024. An open-source fine-tuned large language model for radiological impression generation. *BMC Medical Imaging*, 24(1):254.

Tim Tanida and 1 others. 2023. Interactive and explainable region-guided radiology report generation. *CVPR*.

Dave Van Veen and 1 others. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30(4):1134–1142.

Xiaoman Zhang and 1 others. 2025. Rexgradient-160k: A large-scale publicly available dataset of chest radiographs with free-text reports. *arXiv preprint arXiv:2505.00228*.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Ratescore: A metric for radiology report generation. In *Proceedings of EMNLP*, pages 15004–15019.