RadAdvise: Expert-Calibrated Agentic RAG for Guideline-Grounded Radiology Advisory Generation

Anonymous ACL submission

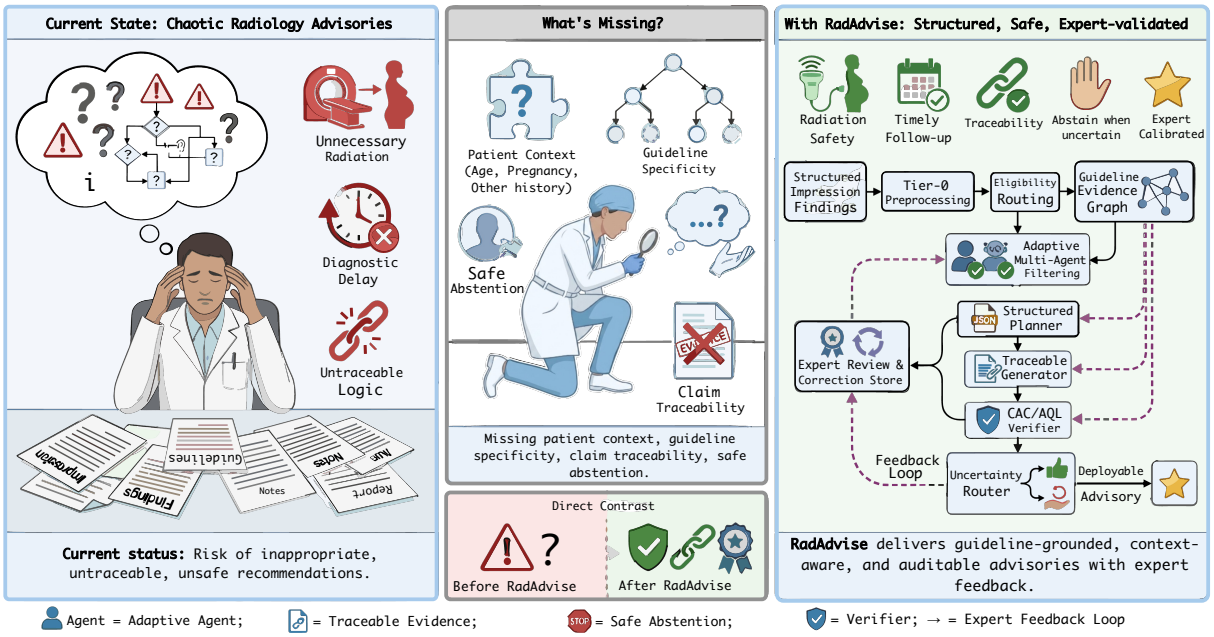


Figure 1: **RadAdvise** at a glance. *Left:* ungrounded advisory generation risks unnecessary radiation, diagnostic delay, and untraceable recommendations. *Middle:* four capabilities standard pipelines lack—patient context, guideline specificity, claim traceability, and safe abstention. *Right:* RadAdvise-Loop supplies all four via graph-constrained retrieval, multi-agent filtering, a structured planner, and a traceable generator, gated by a verifier and an uncertainty router that loops expert corrections back (dashed).

Abstract

The Advise section of a radiology report recommends the next clinical action—which imaging study to obtain, when to obtain it, and how urgently—yet existing automated report generation benchmarks stop at the Impression and leave this action-prescribing section unevaluated. We introduce **RadAdvise**, the first benchmark for guideline-grounded radiology advisory generation, comprising 398,087 silver advisories and 1,024 gold advisories written de novo by board-certified radiologists. The silver tier is produced by **RadAdvise-Loop**, an expert-calibrated agentic RAG pipeline that combines a Guideline Evidence Graph, adaptive multi-agent variant filtering, structured advisory planning, claim-level traceable generation, and uncertainty-routed expert correction. We evaluate advisories with **RadAdvise-Eval**, an appropriateness-focused framework that prioritises patient safety, guideline concordance, radiation awareness, traceability, and calibrated abstention rather than surface overlap. RadAdvise-Loop achieves 83.4% Gold-tier acceptability, compared with 73.8% for a sequential four-agent baseline, 41.0% for the strongest standalone LLM, and 91.4% for radiologists. It also reaches claim traceability of 0.91 and matches baseline-quality expert review using only 5% of the radiologist time. Cross-judge agreement and an ACR-to-ESR transfer experiment further support the reliability and portability of the benchmark, method, and evaluation framework.

1 Introduction

A radiology report ends with an Advise section that recommends the next clinical action: which follow-up scan to order, when to order it, and how urgently it should be performed. Unlike the Findings and Impression, which describe observed abnormali-

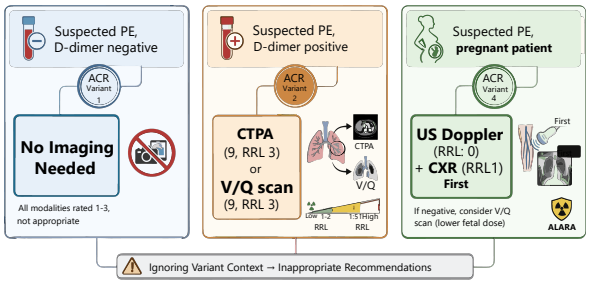


Figure 2: Same suspicion of pulmonary embolism, three different variants of ACR, and top recommendations: no imaging for a D-dimer-negative patient (Variant 1), CTPA or V/Q for a D-dimer-positive patient (Variant 2), and US Doppler plus chest radiograph first for pregnant patient (Variant 4). Topic-level retrieval collapses these patient-specific variants into single recommendation and is therefore wrong for at least two of three cases.

ties and summarise their interpretation, the Advise section directs care; errors in this section can therefore translate into unnecessary radiation, diagnostic delay, or missed escalation. A CT pulmonary angiogram in a low-pretest-probability, D-dimer-negative patient exposes the patient to ionising radiation despite being deemed not indicated by the ACR Appropriateness Criteria. In a pregnant patient, replacing CTPA with a V/Q scan can reduce fetal dose by an order of magnitude. Conversely, recommending 12-month follow-up when 6-week follow-up is warranted may allow an indeterminate nodule to progress.

Despite these stakes, no chest X-ray benchmark directly evaluates the Advise section. MIMIC-CXR (Johnson et al., 2019), CheXpert Plus (Chambon et al., 2024), ReXGradient-160K (Zhang et al., 2025), and IU X-Ray (Demner-Fushman et al., 2016) all terminate at the Impression, and the metrics developed for those sections do not transfer to action-prescribing text. A recommendation such as "recommend CT chest follow-up" can be fluent, clinically plausible in isolation, and rewarded by BLEU or BERTScore, while still being unsafe because it specifies the wrong modality, applies to the wrong patient population, or gives the wrong interval. Figure 2 illustrates this failure mode: the same broad suspicion of pulmonary embolism maps to different ACR variants, and therefore to different top recommendations, depending on D-dimer status and pregnancy (the full ACR variant tables for this example are in Appendix C).

Building a benchmark for this task also exposes limitations in current RAG and agentic-RAG pipelines (Singh et al., 2025). A sequential extract→retrieve→generate system can return a guideline-grounded recommendation, but it typically cannot distinguish among sibling variants within the same topic, attach evidence to individual advisory claims, separate the *what* of a recommendation from the *when* and *what else*, or abstain when patient context is insufficient. We address these limitations with **RadAdvise-Loop**, an expert-calibrated agentic RAG method that adapts multi-agent filtering (Chang et al., 2025), claim-level traceability (Chu et al., 2025), and uncertainty-aware human-in-the-loop curation (Schroeder et al., 2025) to safety-critical clinical generation. Its graph-constrained and adaptively filtered retriever selects patient-specific guideline variants; its traceable generator cites evidence for each claim; its structured planner fixes recommendation fields before prose generation; and its uncertainty router escalates under-specified or high-risk cases to radiologists. Because RadAdvise-Loop also generates the benchmark's silver tier, we evaluate both the method and the data it produces under the same appropriateness-oriented framework.

Contributions. **(1)** We define *radiology advisory generation* as a guideline-grounded, action-prescribing task with a structured output schema covering modality, timeframe, urgency, referral, guideline support, and radiation level, together with a priority-ordered acceptability criterion. **(2)** We release **RadAdvise**, a two-tier dataset containing 398,087 silver advisories and 1,024 gold advisories written de novo by board-certified radiologists. **(3)** We propose **RadAdvise-Loop**, an expert-calibrated agentic RAG method built from a Guideline Evidence Graph, adaptive variant filtering, structured planning, traceable generation, and uncertainty-routed expert correction (§5). **(4)** We introduce **RadAdvise-Eval**, an appropriateness-oriented evaluation framework whose composite score tracks radiologist deployability (ρ=0.84) far better than BLEU (0.12); under this evaluation, RadAdvise-Loop closes 55% of the acceptability gap between the four-agent baseline and radiologists.

2 Related Work

RadAdvise differs from prior work along four dimensions, expanded in Appendix A. *Report generation* systems (Tanida et al., 2023; Bannur et al., 2024; Pellegrini et al., 2023; Chen et al., 2024; Van Veen et al., 2024; Serapio et al., 2024), built on corpora such as MIMIC-CXR (Johnson et al.,

2019), CheXpert Plus (Chambon et al., 2024), ReXGradient-160K (Zhang et al., 2025), and IU X-Ray (Demner-Fushman et al., 2016), target descriptive sections and provide no advisory targets. *Evaluation metrics*, both lexical and clinical-entity (Delbrouck et al., 2022; Ostmeier et al., 2024; Zhao et al., 2024; Delbrouck et al., 2025; Bannur et al., 2024), measure resemblance to a reference rather than whether a recommended action is appropriate for a patient-specific variant (§7). *Agentic RAG* (Singh et al., 2025) contributes multi-agent filtering (Chang et al., 2025), claim traceability (Chu et al., 2025), and uncertainty-aware curation (Bayer et al., 2026; Schroeder et al., 2025), which we instantiate at the ACR-variant level rather than over generic documents. *Guideline-grounded decision support* (Zakka et al., 2024; Xiong et al., 2024) grounds question answering in sources such as the ACR Appropriateness Criteria (Kurth et al., 2021) but does not represent guidelines as a variant-level graph for structured advisory generation. RadAdvise is the first to combine patient-specific guideline retrieval, structured action planning, traceable evidence, and abstention-aware evaluation for radiology advisories.

3 Task Definition

We define *radiology advisory generation* as the task of producing guideline-grounded clinical action recommendations from a radiology report. Given a report $r = (F, I)$ with Findings F and Impression I, the system must produce an advisory a consisting of one or more recommendations, each represented as a tuple

$$\rho = (\text{MODALITY}, \text{TIMEFRAME}, \text{URGENCY}, \text{REFERRAL}, v, \text{RRL}),$$

where v denotes the supporting guideline-variant identifier and RRL denotes the Relative Radiation Level for radiation-bearing modalities. An advisory is *accepted* only if every recommendation $\rho \in a$ is consistent with the patient-specific guideline variant for the relevant entity in $F \cup I$, avoids contraindicated modalities, assigns urgent or emergent status to critical findings, and respects ALARA when lower-radiation alternatives have comparable appropriateness. When essential patient context is missing—for example, a 6 mm pulmonary nodule with unknown risk factors and no prior imaging—the system may instead *abstain* or emit an explicit caveat rather than commit to a precise recommendation.

Table 1: Two-tier dataset overview. Gold category distribution matches Silver test to within 1 point; extended statistics are in Appendix D.

	Silver	**Gold**
Advisories	398,087	1,024
Train / Val / Test	318,467 / 39,808 / 39,812	— / — / 1,024
Unique patients	186,168	1,018
Source sites	82	64
Mean length (tokens)	82 ± 34	78 ± 28
Mean ACR citations / adv.	1.2 ± 0.5	—
Advisory category distribution		
Follow-up imaging (%)	42.4	42.1
Clinical correlation (%)	28.7	28.4
Specialist referral (%)	11.8	12.4
No follow-up needed (%)	17.1	17.1

This task differs from standard report generation in two ways. First, the target is grounded in an external evidence base rather than solely in the input report, so correctness requires satisfying guideline-derived constraints instead of matching a reference string. Second, the output must follow a clinical priority ordering in which patient safety, radiation safety, and guideline concordance dominate surface fluency. These properties motivate both the constrained generation pipeline in §5 and the appropriateness-oriented evaluation framework in §6.

4 The RadAdvise Dataset

RadAdvise instantiates the task with two complementary tiers. **Silver** contains 398,087 advisories generated by RadAdvise-Loop (§5) from four chest X-ray corpora and curated through deterministic filtering, uncertainty scoring, and targeted expert review. **Gold** contains 1,024 advisories written de novo by three board-certified radiologists and serves as the primary independent evaluation set. Table 1 summarises both tiers; source-corpus and split-level statistics appear in Appendix D.

Gold construction. We sampled 1,024 reports stratified by primary advisory category and source corpus to match Silver-test proportions. Three radiologists, with a mean of 11.2 years post-fellowship experience, independently authored advisories from each report's Findings and Impression without seeing any pipeline output. They used written guidelines and a structured form aligned with the task schema, including modality, timeframe, urgency, referral, supporting variant, and caveats. Inter-radiologist agreement on advisory category reached Fleiss' $\kappa = 0.82$; the full annotation protocol, recruitment, compensation, and IRB

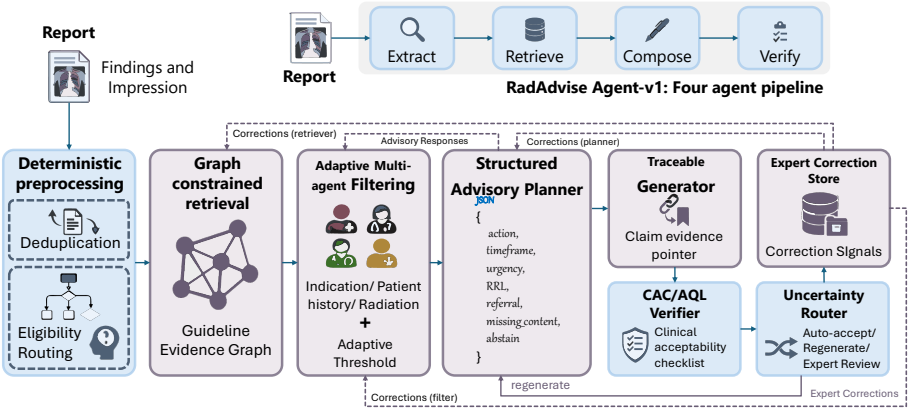


Figure 3: **RadAdvise-Loop.** *Top:* the four-agent baseline (RadAdvise-Agent v1). *Bottom:* a report (F, I) flows through Tier-0 preprocessing, graph-constrained retrieval, multi-agent filtering, structured planning, and traceable generation; a CAC/AQL verifier emits an uncertainty score routing each advisory to auto-accept, regenerate, or expert review, with corrections fed back to the retriever, filter, and planner (§5).

details are in Appendix E.

Silver curation. Silver curation proceeds in 3-stages. **Tier 0** deterministically parses sections, removes near-duplicates, extracts patient context, and routes each report to one of four eligibility classes: no advisory, template advisory, guideline-dependent full RAG, or ambiguous/high-risk full RAG with mandatory expert sampling. This routing avoids unnecessary LLM calls for 18.4% of reports. **Tier 1** generates advisories with RadAdvise-Loop and assigns each an uncertainty score $U \in [0, 1]$ (§5.5). **Tier 2** sends high-U or high-risk advisories to radiologists, whose typed corrections provide supervised signal for the retriever, planner, verifier, and abstention policy. The routing policy is shown in Table 2; an independent 2,000-advisory stratified audit achieved Fleiss' $\kappa = 0.74$ on safety judgement.

5 RadAdvise-Loop

RadAdvise-Loop extends a four-agent RAG baseline with five components for advisory generation: graph-constrained retrieval, adaptive variant filtering, structured planning, claim-level traceability, and uncertainty-routed expert correction (Figure 3). The baseline, **RadAdvise-Agent v1**, runs four sequential agents—finding extraction into entities with criticality flags, ACR retrieval via hybrid dense+BM25, advisory composition under a fixed multi-finding priority policy, and CAC-based safety verification with up to three regeneration cycles—and is a strong standard agentic-RAG comparator. Prompts, schemas, and hyperparameters for every stage are in Appendix F and G.

5.1 Guideline Evidence Graph

We replace flat ACR indexing with a heterogeneous *Guideline Evidence Graph* $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ containing twelve node types, including FINDING, INDICATION, PATIENT-CONTEXT, VARIANT, PROCEDURE, and RRL, linked by nine edge types such as *indication-maps-to-topic* and *patient-context-modifies-variant*. The graph encodes 247 ACR topics, 1,284 variants, 8,812 procedure-variant edges, and 4,027 patient-context modifiers. Retrieval is graph-constrained: extracted findings and indications seed a traversal that filters candidates to topology-consistent variants before dense ranking, excluding structurally impossible matches before any LLM call. The full ACR knowledge-base composition is in Appendix B, and the graph schema, construction, and statistics are in Appendix G.2.

5.2 Adaptive multi-agent variant filtering

Graph-constrained retrieval returns up to 20 candidate variants, which four specialist agents score along complementary axes: indication match, patient-context match, finding-pattern match, and radiation-aware alternatives. Their scores are combined with dense and BM25 retrieval scores as

$$s(v) = \alpha\, s_{\text{dense}}(v) + \beta\, s_{\text{BM25}}(v) + \gamma\, s_{\text{agents}}(v) - \lambda\, s_{\text{conflict}}(v),$$

where s_{agents} is the mean specialist score and s_{conflict} is their standard deviation, penalising disagreement among the agents. Following MAIN-RAG (Chang et al., 2025), we keep variants above a per-query adaptive threshold defined by the largest gap in the sorted $s(v)$ values, so ambiguous reports retain more candidates and clear reports fewer (the

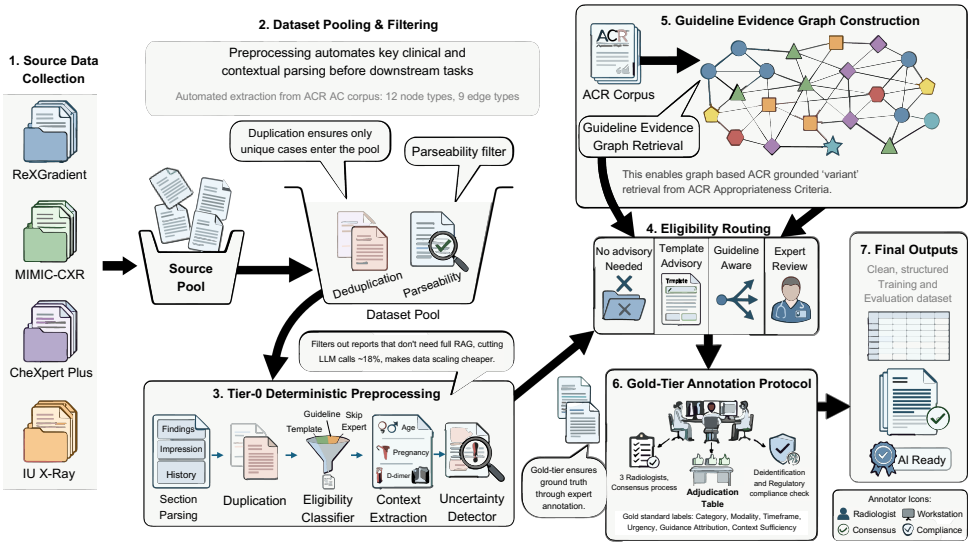


Figure 4: End-to-end RadAdvise dataset construction. Reports from four chest X-ray corpora are pooled, deduplicated, parseability-filtered, and passed through Tier-0 preprocessing. Eligible reports are routed to template, guideline-aware RAG, or expert-review paths; the Guideline Evidence Graph supplies variant-level grounding, and three board-certified radiologists author and adjudicate the Gold tier.

largest-gap rule is formalised in Appendix G). We tune $(\alpha, \beta, \gamma, \lambda) = (0.45, 0.20, 0.30, 0.05)$ on the validation split.

5.3 Structured advisory planning

Free-form composition often conflates the recommended action, its timing, its urgency, and required caveats. RadAdvise-Loop instead emits a structured JSON plan before prose generation:

```
advisory_plan.json
{ "finding": "...", "clinical_problem": "...",
  "variant": "...", "recommended_action": "...",
  "timeframe": "...", "urgency": "...", "rrl": "...",
  "referral": "...", "missing_context": [...], "abstain": false,
  "claim_evidence": [{...}, ...] }
```

The plan is validated field by field against the graph before surface realisation: variants, procedures, RRLs, timeframes, and urgency labels must match graph constraints or report-derived criticality. When key context is missing, the planner may set `abstain = true` or emit a caveat, avoiding unsupported precision.

5.4 Claim-level traceable generation

Each advisory claim is paired with typed evidence: a source span in $F \cup I$, a guideline triple of topic, variant, and procedure, an RRL table entry, a contraindication rule, or a missing-context caveat. The generator emits both prose and evidence pointers, following TracSum (Chu et al., 2025). We define *Traceability Accuracy* as the fraction of claims whose cited evidence supports the claim under exact-match verification against the report and graph, and evaluate it in §7.

5.5 Uncertainty routing and expert correction

For each advisory, RadAdvise-Loop computes an uncertainty score

$$U = w_1(1 - p_{\text{ret}}) + w_2 D_{\text{judge}} + w_3 H_{\text{CAC}} + w_4 M_{\text{ctx}} + w_5 R_{\text{risk}},$$

where p_{ret} is variant retrieval confidence, D_{judge} is cross-judge disagreement, H_{CAC} is CAC violation severity, M_{ctx} is missing context, and R_{risk} is clinical risk. We learn the weights by logistic regression on 1,200 development advisories with radiologist deployability labels and route outputs according to Table 2. Expert corrections are typed and recycled as supervised signal: variant corrections become retriever hard negatives, timeframe corrections update the planner, modality and RRL corrections recalibrate verifier thresholds, and missing-context corrections tune abstention. The learned weights, full correction taxonomy, and loop convergence behaviour are in Appendix G.

6 RadAdvise-Eval

The method above produces structured, evidence-linked advisories, but evaluating them requires

Table 2: Uncertainty-based routing policy. Expert corrections feed back into retriever, planner, verifier, and abstention training; full details are in Appendix G.

Condition	Route	Purpose
CAC pass, low U	Auto-accept	Scalable silver
CAC fail, low risk	Regenerate	Automatic repair
High judge disagreement	Expert review	Evaluator uncertainty
Missing key context	Caveated / expert	Avoid false precision
Critical finding flag	Expert sample (10%)	Safety assurance

more than text similarity. RadAdvise-Eval therefore asks whether an advisory is clinically *appropriate* rather than merely fluent or reference-like. It has four layers: six automated metrics (safety, guideline concordance, coherence, factuality, radiation awareness, and traceability); a binary Clinical Acceptability Checklist (CAC) for deployment gating; an Advisory Quality Ladder (AQL) ordering quality by clinical priority; and an Appropriate Abstention Rate (AAR) rewarding calibrated refusal when patient context is insufficient. Formal definitions are in Appendix H.

Metrics. The six automated metrics each target one appropriateness dimension: S_{safety} (zero on any hard safety violation, else the fraction of safety-relevant entities handled correctly), S_{guide} (recommendations present in the matched ACR variant with rating ≥ 7 or justification and consistent with patient context), S_{cohere} (Findings–Impression–advisory consistency via a domain-adapted NLI model), S_{fact} (hallucinated entities, via an adapted RaTEScore (Zhao et al., 2024)), S_{rad} (ALARA adherence), and S_{trace} (claim traceability, §5.4). Full formal definitions are in Appendix H.

CAC, AQL, and AAR. The CAC comprises five hard constraints—no contraindication, correct urgency flagging, ACR-listed modality, no fabricated entities, and RRL disclosure—together with six soft constraints (Appendix I). An advisory passes only when all hard constraints hold and the soft-constraint mean is at least 0.7, i.e. $\text{pass}(a) = \left(\bigwedge_{j=1}^{5} h_j(a)\right) \wedge \left(\frac{1}{6}\sum_{j=1}^{6} s_j(a) \geq 0.7\right)$. The AQL assigns each advisory the highest level it reaches among **L1** patient safety, **L2** radiation safety, **L3** guideline concordance, **L4** clinical precision, and **L5** communication quality, where level k requires levels 1 through k to hold; equivalently $\text{AQL}(a) = \max\{\, k : a \models L_1, \ldots, L_k \,\}$. The AAR is the fraction of gold context-insufficient cases on which the system abstains or emits an explicit caveat; we report it alongside CAC because a calibrated abstention is preferable to a confident but unsafe recommendation.

7 Experiments

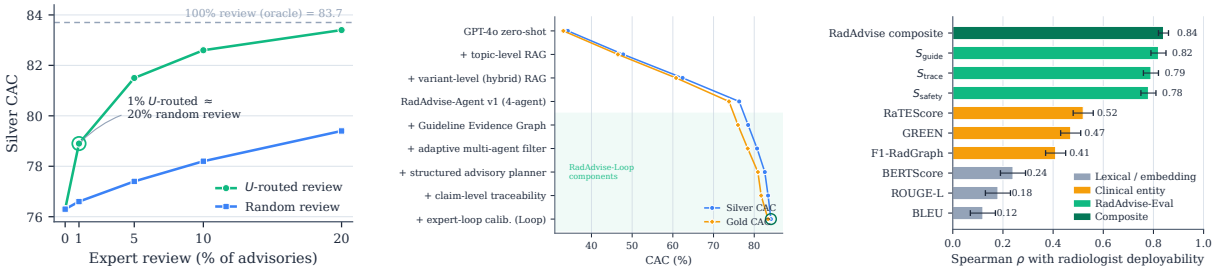
Setup. We evaluate GPT-4o, Claude 3.5 Sonnet, Med-PaLM 2, LLaVA-Med, MAIRA-2, the four-agent baseline **RadAdvise-Agent v1**, and **RadAdvise-Loop** on both dataset tiers. LLMs receive Findings and Impression, while VLMs also receive the chest radiograph(s). All systems run at temperature 0; results are means over three seeds with 95% bootstrap confidence intervals from 10,000 resamples. The Silver test set contains 5,000 stratified reports, and the Gold test set contains all 1,024 radiologist-written advisories. Model versions, decoding settings, seeds, and release artefacts needed to reproduce every number are in Appendix U.

Main results and ablations. RadAdvise-Loop reaches 83.4% Gold CAC, gaining 9.6 points over the four-agent baseline and 42.4 over the strongest standalone LLM (Table 3), closing 55% of the baseline-to-radiologist gap; traceability rises from 0.79 to 0.91 and rankings are stable across both tiers. The component ablation (Figure 5b, Appendix Table 21) attributes the gains: the Guideline Evidence Graph and adaptive filtering each add 2.2 Silver-CAC points; the structured planner adds 1.9 and cuts temporal error from 27% to 18%, confirming timing failures are a planning problem; traceability adds 0.8 (S_{trace} +0.12); and expert-loop calibration adds 0.7 Silver-CAC and 1.7 Gold-CAC. Gold gains exceed Silver because each component targets failure modes specific to radiologist-written advisories. Isolated retrieval (Table 4) shows graph constraints and filtering each add $\sim$3 top-1 variant points, with the error decomposition in Appendix L; worked examples including a failure case are in Appendix T.

Routing, traceability, and validation. On v1 outputs, random review of 20% of advisories reaches 79.4 Silver CAC at 84 expert hours, whereas uncertainty-routed review reaches a comparable 78.9 CAC by reviewing only 1% of advisories at 4.2 hours (Figure 5a); uncertainty routing therefore matches baseline-review quality using roughly 5% of the radiologist time, with full curves in Appendix Table 22. For traceability, on 800 advisories with radiologist-verified evidence labels RadAdvise-Loop cites source evidence correctly for 93% of claims and guideline evidence for 91%, leaving only 3% unsupported (Table 5), versus 71–

System	Type	Silver CAC %	Gold (1,024 radiologist-written) S_{guide}	S_{rad}	CAC %	AQL	Trace S_{trace}	Deploy. %
Closed-source / proprietary API models								
GPT-4o	General LLM	34.2 [32.8, 35.6]	0.41	0.52	33.1 [30.4, 35.8]	2.8	0.62	18.6
Claude 3.5 Sonnet	General LLM	37.6 [36.2, 39.0]	0.44	0.55	36.4 [33.6, 39.2]	2.9	0.64	22.1
Med-PaLM 2	Medical LLM	42.1 [40.7, 43.5]	0.53	0.58	41.0 [38.2, 43.8]	3.1	0.66	28.7
Gemini 1.5 Pro	General LLM	36.5 [35.1, 37.9]	0.43	0.54	35.2 [32.5, 37.9]	2.9	0.63	21.4
Open-source / open-weight models								
LLaVA-Med	Vision–Language	24.8 [23.5, 26.1]	0.38	0.46	23.7 [21.2, 26.2]	2.3	0.58	14.4
MAIRA-2	Radiology VLM	35.7 [34.3, 37.1]	0.49	0.54	34.9 [32.1, 37.7]	2.7	0.65	24.3
Llama-3.1-70B-Inst.	General LLM	33.1 [31.7, 34.5]	0.40	0.50	31.8 [29.1, 34.5]	2.6	0.60	17.2
CheXagent	Radiology VLM	37.2 [35.8, 38.6]	0.51	0.55	36.1 [33.3, 38.9]	2.8	0.67	25.6
Proposed								
RadAdvise-Agent v1	Four-agent RAG	76.3 [75.1, 77.5]	0.87	0.89	73.8 [71.1, 76.5]	4.2	0.79	62.4
RadAdvise-Loop	**Agentic RAG**	**84.1** [83.0, 85.2]	**0.92**	**0.93**	**83.4** [81.0, 85.8]	**4.5**	**0.91**	**74.6**
Radiologist	*Expert*	—	*0.93*	*0.94*	*91.4*	*4.7*	*0.96*	*84.2*

Table 3: Evaluations across SOTA methods. Brackets are 95% bootstrap CIs from 10,000 resamples. Deployability is the percentage of advisories rated deployable by three radiologists on a 500-advisory Gold subset. RadAdvise-Loop closes 55% of the Gold-CAC gap between the four-agent baseline and radiologists, and improves traceability by 12 points.



(a) Expert-routing efficiency. *U*-routed review at 1% sampling matches random review at 20%; the 100% oracle reaches 83.7.

(b) Component ablation. Each step adds one component to the previous row; the shaded band marks RadAdvise-Loop additions over v1.

(c) Metric validation. Spearman ρ with radiologist deployability on 1,500 Gold-subset ratings; whiskers are 95% CIs.

Figure 5: Quantitative results for RadAdvise-Loop. Full numbers are in Appendix R and Appendix H.

Retriever	Top-1 var.	Top-3 var.	CAC
BM25 topic	0.43	0.71	58.2
Dense PubMedBERT topic	0.52	0.78	64.5
Hybrid dense+BM25	0.86	0.95	76.3
+ Graph-constrained	0.89	0.96	78.5
+ Adaptive multi-agent	**0.92**	**0.97**	**80.7**

Table 4: Retrieval methods on 500 variant-labelled reports. Graph constraints and adaptive multi-agent filtering each add ∼3 top-1 variant points, gains compose.

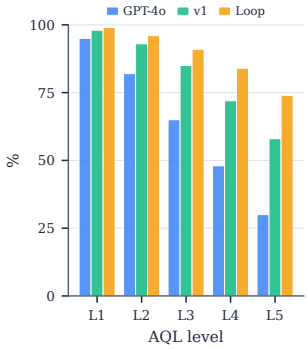
System	Source	Guideline	Unsupp.
GPT-4o	0.71	0.42	0.18
Med-PaLM 2	0.74	0.51	0.14
RadAdvise-Agent v1	0.82	0.78	0.07
RadAdvise-Loop	**0.93**	**0.91**	**0.03**

Table 5: Claim-level traceability on 800 advisories with radiologist-verified labels. *Source* and *Guideline* are per-claim evidence accuracy; *Unsupp.* is the unsupported-claim fraction.

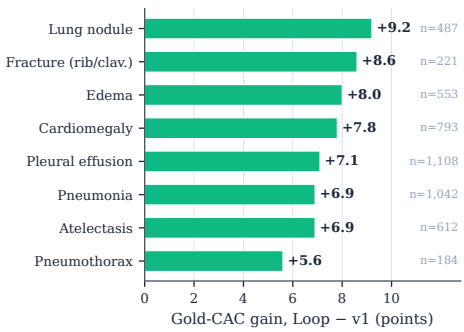
74% source and 42–51% guideline accuracy for standalone LLMs prompted to cite post hoc.

Finally, metric validation (Figure 5c) shows correlation with radiologist deployability rising from lexical metrics (BLEU ρ=0.12) to clinical-entity metrics (RaTEScore ρ=0.52) to the RadAdvise-Eval composite (ρ=0.84), so RadAdvise-Eval is a useful screening metric while Gold expert evaluation remains the reference standard; per-metric correlations are in Appendix Table 26.

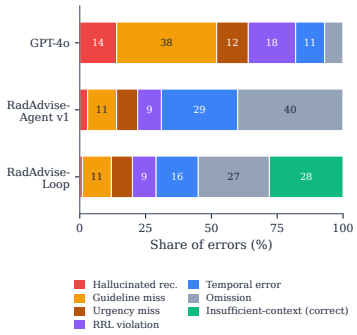
Robustness and guideline transfer. Re-judging 1,000 advisories with Claude 3.5 Sonnet and Llama-3.1-70B preserves system ranking: Cohen's κ against GPT-4o is 0.79 and 0.71, and absolute CAC differences stay within 2 points (Appendix P). To test portability, we replace the ACR index in the Guideline Evidence Graph with the chest subset of ESR iGuide and re-run on the 412 Gold scenarios ESR covers; RadAdvise-Loop obtains 80.8 Gold CAC versus 83.4 with ACR, indicating architectural portability without claiming full clinical equivalence across guideline systems (Appendix Q).



(a) AQL dist. RadAdvise-Loop narrows the L3 to L4 drop from 13 points to 7.



(b) Per-pathology Gold-CAC gains are largest for variant-ambiguous findings such as nodules and fractures.



(c) Failure mass shifts from guideline misses and omissions toward correct abstention.

8 Analysis

Error redistribution. RadAdvise-Loop changes not only the number of errors but also their type. Standalone LLMs fail mainly through *Guideline Miss* errors, which account for 38% of GPT-4o errors; the four-agent baseline reduces those failures but shifts mass to *Temporal Error* (29%) and *Omission* (40%). The structured planner cuts these to 16% and 27%, respectively, leaving *Insufficient Context* as the dominant residual at 28%. These residual cases are often appropriate abstentions or caveated outputs, and the Gold annotations also mark them as context-insufficient. The AQL shows the same pattern: baselines drop sharply from L3 to L4, where timing and completeness are evaluated, while RadAdvise-Loop narrows this drop (Figure 6a; full breakdown in Appendix N).

Where the gains concentrate. The per-pathology gains in Figure 6b are largest exactly where variant ambiguity is highest—lung nodules and fractures, whose correct recommendation hinges on context the report often omits—and smallest where the baseline was already near ceiling, such as pneumothorax (full per-pathology numbers in Appendix M). This is consistent with the ablation: the method's advantage concentrates in cases that require disambiguation rather than in cases a flat retriever already handles.

Abstention, cost, and latency. RadAdvise-Loop abstains or emits an explicit caveat on 11.4% of Gold reports, and 84% of those cases are also marked context-insufficient by the Gold annotators (AAR = 0.84); baseline systems abstain on only 0–2% of cases regardless of context completeness. This safety behaviour increases cost and latency: multi-agent filtering and planning raise latency from 4.1 s for v1 to 10.2 s, or 12.3 s with amortised expert review, and increase per-advisory cost from USD 0.018 to USD 0.043. Distilling the filtering agents into Llama-3.1-8B reduces cost to USD 0.024 with a 1.5-point CAC drop, and Tier-0 routing offsets part of the increase at corpus scale by sending 18.4% of reports to template-or-skip paths (Appendix R).

9 Conclusion

RadAdvise introduces the first benchmark for the Advise section of radiology reports, pairing a two-tier dataset with an evaluation framework that aligns much more closely with radiologist deployability than existing text metrics (ρ=0.84 vs. BLEU 0.12). RadAdvise-Loop adds graph-constrained retrieval, adaptive filtering, structured planning, traceable generation, and uncertainty-routed expert correction to a standard four-agent RAG baseline, raising Gold acceptability from 73.8% to 83.4% and traceability from 0.79 to 0.91. Uncertainty routing further matches baseline-quality expert review using only 5% of the radiologist time. The remaining gap to radiologists is concentrated in cases with genuinely insufficient patient context, where abstention or caveated recommendation is safer than unsupported specificity.

Ethical Considerations. All source data are de-identified and used under their original licences; code, prompts, outputs, and reconstruction scripts are released (Appendices E, U). RadAdvise is a research benchmark requiring independent validation before clinical deployment, with no observed disparity across age band or sex (p>0.05 Bonferroni-corrected; Appendix S); deployments should nonetheless monitor review burden across subgroups, as uncertainty routing concentrates expert effort on high-risk cases.

Limitations

RadAdvise has four main limitations. First, the Silver tier is generated by RadAdvise-Loop and curated through rules, LLM-as-judge signals, and uncertainty-routed radiologist review; the Gold tier and metric-calibration study mitigate this concern, but a fully human-written Silver-scale corpus is not feasible retrospectively. Second, the Guideline Evidence Graph is built primarily from US ACR practice; our ACR-to-ESR experiment suggests portability, but multi-region validation remains future work. Third, the filtering and verifier stages still depend on LLM judgements, which cross-judge robustness reduces but does not eliminate. Finally, expert corrections are weighted equally regardless of annotator seniority; production systems should consider weighted disagreement and adjudication protocols.

References

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th bioNLP workshop and shared task*, pages 370–379.

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

Markus Bayer, Justin Lutz, and Christian Reuter. 2026. Activellm: Large language model-based active learning for textual few-shot scenarios. *Transactions of the Association for Computational Linguistics*, 14:1–22.

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

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and 1 others. 2025. Main-rag: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2607–2622.

Zhihong Chen, Maya Varma, Justin Xu, Magdalini Paschali, Dave Van Veen, Andrew Johnston, Alaa Youssef, Louis Blankemeier, Christian Bluethgen, Stephan Altmayer, and 1 others. 2024. A vision-language foundation model to enhance efficiency of chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Bohao Chu, Meijie Li, Sameh Frihat, Chengyu Gu, Georg Lodde, Elisabeth Livingstone, and Norbert Fuhr. 2025. Tracsum: A new benchmark for aspect-based summarization with sentence-level traceability in medical domain. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 844–864.

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*.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

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*.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just put a human in the loop? investigating llm-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25771–25795.

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

Aditi Singh, Abul Ehtesham, Saket Kumar, Tala Talaei Khoei, and Athanasios V Vasilakos. 2025. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*.

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.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, and 1 others. 2024. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068.

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.

A Extended Related Work

This appendix expands the four threads summarised in §2: chest X-ray report generation and its corpora, clinical evaluation metrics for radiology text, agentic and retrieval-augmented generation, and guideline-grounded clinical decision support.

A.1 Report generation and corpora

Chest X-ray report generation has been driven by four large corpora: MIMIC-CXR (Johnson et al., 2019) (227,835 single-institution studies), CheXpert Plus (Chambon et al., 2024) (223,228 studies), ReXGradient-160K (Zhang et al., 2025) (160,000 studies from 79 sites), and IU X-Ray (Demner-Fushman et al., 2016). The systems built on them generate the descriptive sections of the report: RGRG (Tanida et al., 2023) grounds sentences in detected anatomical regions, MAIRA-2 (Bannur et al., 2024) and CheXagent (Chen et al., 2024) are radiology vision–language models, RaDialog (Pellegrini et al., 2023) adds a conversational interface, and LLM-based summarisers (Van Veen et al., 2024; Serapio et al., 2024) compress Findings into Impressions. Across this body of work the output target is the Findings or Impression; none release advisory targets and none are evaluated on advisory content, which is the gap RadAdvise addresses.

A.2 Clinical evaluation metrics for radiology text

Two metric families dominate radiology-text evaluation. Lexical and embedding metrics (BLEU, ROUGE-L, BERTScore) score surface similarity, so a fluent paraphrase of an unsafe recommendation scores like its safe counterpart. Clinical-entity metrics—F1-RadGraph (Delbrouck et al., 2022), GREEN (Ostmeier et al., 2024), RaTEScore (Zhao et al., 2024), F1-SRR-BERT (Delbrouck et al., 2025), RadFact (Bannur et al., 2024), and CheXbert-based variants—verify that the same clinical entities appear in candidate and reference. Both families evaluate *descriptive* correctness, i.e. whether the generated text faithfully restates the source, and neither tests whether a prescribed action is clinically appropriate under the patient-specific guideline variant. §7 reports the correlation of representatives from both families with radiologist deployability judgements alongside RadAdvise-Eval.

A.3 Agentic and retrieval-augmented generation

Agentic RAG systems (Singh et al., 2025) extend static retrieve-then-generate workflows with planning, reflection, tool use, and multi-agent collaboration. MAIN-RAG (Chang et al., 2025) uses multiple LLM agents to collaboratively filter retrieved documents under an adaptive, per-query threshold; we adapt this from the document level to the ACR-variant level (§5.2), where the filtering decision is which sibling variant of a shared topic applies to the patient. Traceability work in medical summarisation (Chu et al., 2025) attaches sentence-level citations to generated summaries so users can audit content against source; we extend this from summary claims to clinical action recommendations, pairing each advisory claim with a typed evidence pointer (§5.4). Finally, LLM-assisted annotation can bias subjective labels when annotators review model outputs (Schroeder et al., 2025); we mitigate this by hiding pipeline output from radiologists during Gold authoring (§4) and by making expert review of Silver data uncertainty-selected rather than universal (§5.5).

A.4 Guideline-grounded clinical decision support

The ACR Appropriateness Criteria (Kurth et al., 2021) encode 247 imaging topics and over 4,000 clinical scenarios, each with appropriateness ratings (1–9) and standardised Relative Radiation Levels; the European Society of Radiology offers an analogous system in iGuide, which we use for the guideline-transfer experiment in §7. Clinical RAG systems such as Almanac (Zakka et al., 2024), ChatRWD, and MedRAG (Xiong et al., 2024) ground question answering in evidence sources but return free-text answers rather than structured, schema-constrained recommendations, and they retrieve over flat document indices. To our knowledge, no prior work operationalises a variant-level imaging-guideline corpus as a graph-constrained retrieval source for structured advisory generation, which is the core of the Guideline Evidence Graph (§5.1).

B ACR AC Knowledge Base Composition

This appendix documents the static ACR Appropriateness Criteria knowledge base from which the Guideline Evidence Graph (Appendix G.2) is constructed. It is distinct from the graph-composition table: Table 16 reports the nodes and edges *after* our parsing and linking, whereas the tables here report the raw source content as published in the October 2024 ACR AC release, so that the provenance of every graph element is auditable against the original corpus.

Table 6 gives the headline counts. The 247 diagnostic topics expand into 1,284 clinical variants because most topics are conditioned on patient context—pretest probability, pregnancy, age, prior workup—and it is precisely this one-to-many topic-to-variant expansion that motivates variant-level rather than topic-level retrieval (§1).

Component	Count
Diagnostic imaging topics	247
Clinical variants	1,284
Clinical scenarios	4,000+
Imaging recommendations	8,812
Evidence narratives	247

Table 6: ACR AC knowledge base composition (October 2024 release).

The topics are authored by twelve organ-system expert panels (Table 7). Although chest X-ray advisory generation draws most heavily on the 22 Thoracic topics, it is not confined to them: chest findings routinely invoke Cardiac, Vascular, and Systemic Oncology topics (for example, an incidental pulmonary nodule in a patient with known malignancy maps to an oncology-staging variant), which is why the graph indexes all panels rather than the thoracic subset alone.

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

Table 7: ACR AC topics by expert panel.

Finally, Table 8 lists the ACR Relative Radiation Level scale used throughout the paper. RRL is a six-level ordinal category, not a raw dose, and it is the quantity the advisory schema discloses and the S_{rad} metric (§6) operationalises; ALARA reasoning in RadAdvise-Loop compares candidate procedures on this scale rather than on millisievert estimates, which are patient- and protocol-dependent and ab-

sent from the guideline text.

RRL	Adult effective 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 with catheter
•••••	30–100 mSv	CT w/o and w/ IV contrast

Table 8: ACR Relative Radiation Level categories.

C Pulmonary Embolism: Full Variant Tables

The three PE variants summarised in Figure 2 differ substantially in their procedure rankings. Between Variants 2 and 4 alone, ten procedures change appropriateness category, and the change is not cosmetic: it reorders which procedure a guideline-faithful advisory should recommend first. Tables 9 and 10 give the full procedure lists for both variants so that the within-topic divergence a flat retriever cannot see is explicit.

The clinically load-bearing difference is the promotion of two zero- and low-radiation procedures—lower-extremity venous Doppler (RRL O) and chest radiography (RRL •)—to *Usually Appropriate* in pregnancy, where in Variant 2 the Doppler is *Usually Not Appropriate*. A system that retrieves at the topic level returns one ranking for "Suspected PE" and is therefore guideline-violating for the pregnant patient even when its output is fluent and cites the correct topic, which is exactly the failure mode the variant-level graph (Appendix G.2) is built to remove.

D Dataset Statistics: Extended

This appendix extends the dataset overview of Table 1 with per-split detail and source-corpus provenance. Table 11 reports the train/val/test breakdown: the gap between total reports and reports-with-advisory in each split (e.g. 41,253 vs. 39,812 on test) is exactly the set Tier-0 routes to the no-advisory path, so the advisory counts are downstream of eligibility routing rather than of any post-hoc filtering. The eligibility distribution confirms the headline figure used in the main text: 18.4% of test reports (8.7% no-advisory plus 9.7% template) bypass the full RAG pipeline, which is the source of the corpus-scale cost saving in §R.

Table 12 reports the four-corpus composition after deduplication and parseability filtering. The mixture is deliberately heterogeneous along two axes: MIMIC-CXR supplies the bulk of the volume from a single institution, while ReXGradient-160K supplies almost all of the site diversity (79 of the 82 sites), so the Silver tier is large enough to train on yet broad enough that evaluation is not dominated by one institution's reporting style. The patient counts are de-duplicated within but not across corpora; the combined 186,168 figure therefore upper-bounds the unique-patient total, since cross-corpus patient overlap cannot be resolved on de-identified data.

E Gold-Tier Annotation Protocol

Table 13 summarises the annotation protocol; the subsections that follow give the detail. The protocol is designed so that Gold authoring is independent of any pipeline output and so that the labels needed by each metric—in particular the context-insufficient flag consumed by the AAR—are collected at authoring time rather than inferred afterwards.

E.1 Annotators

Three board-certified radiologists with mean 11.2 years post-fellowship (range 7–16) participated; two practise primarily in the United States and one in Europe. Annotators were recruited through clinical research networks and compensated at USD 120 per hour. All completed institutional ethics training and signed data-use agreements covering MIMIC-CXR and ReXGradient-160K, and the annotation protocol received IRB exemption as a quality-improvement activity using de-identified data.

E.2 Inputs shown

For each report, annotators saw the Findings and Impression sections, the original chest radiograph(s), and demographic data when present (age band, sex); the matched ACR AC variant retrieved by the baseline pipeline was revealed only *after* their own draft was submitted, to inform variant labelling without anchoring the recommendation. Annotators did not see pipeline-generated advisories during Gold authoring, and system identity was hidden when they later evaluated baseline and RadAdvise-Loop outputs.

E.3 Structured form

For each report, annotators completed a structured form mirroring the task schema of §3:

Table 9: ACR AC: *Suspected Pulmonary Embolism*, Variant 2 (low/intermediate pretest probability, positive D-dimer).

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 IV contrast	Usually Not Appropriate	●●●

Table 10: ACR AC: *Suspected Pulmonary Embolism*, Variant 4 (pregnant patient). US duplex Doppler (RRL: O) and chest radiograph (RRL: ●) are promoted to *Usually Appropriate*.

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
Arteriography pulmonary with right heart cath	Usually Not Appropriate	●●●●
MRA pulmonary arteries without and with IV contrast	Usually Not Appropriate	O
CT chest with IV contrast	Usually Not Appropriate	●●●
CT chest without IV contrast	Usually Not Appropriate	●●●
CTA triple rule out	Usually Not Appropriate	●●●

Statistic	Train	Val	Test
Total reports	330,019	41,252	41,253
With generated advisory	318,467	39,808	39,812
Gold subset (overlaps test)	—	—	1,024
Unique patients	148,932	18,617	18,619
Tier-0 eligibility distribution (% of test reports)			
No advisory needed		8.7	
Deterministic template advisory		9.7	
Guideline-dependent advisory		72.4	
Ambiguous / high-risk advisory		9.2	
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 11: Dataset statistics by split. Tier-0 eligibility distribution shows that 18.4% of reports route to template or skip paths, offsetting the per-advisory cost increase.

- **Primary advisory category**: follow-up imaging / clinical correlation / specialist referral / no follow-up needed.
- **Modality and timeframe**: free-text recommendation with an explicit interval.
- **Urgency**: routine / urgent / emergent.
- **Referral target**: specialty if applicable.
- **Supporting guideline**: ACR variant ID if known.
- **Context-insufficient flag**: marked when the report alone does not contain enough information for a precise recommendation; used as the gold label for the AAR metric (§6).
- **Notes**: any clinical caveats (pregnancy, renal function, etc.).

Source	Reports	Patients	Sites
ReXGradient-160K	160,000	109,487	79
MIMIC-CXR	227,835	65,379	1
CheXpert Plus	18,431	12,209	1
IU X-Ray	6,258	3,996	1
Combined	**412,524**	**186,168**	**82**

Table 12: Source-corpus composition after deduplication and parseability filters.

E.4 Disagreement adjudication

Each report was annotated by all three radiologists. When category disagreement occurred (18% of cases), the consensus advisory was determined by a 30-minute structured discussion. When advisory content differed but category agreed (54% of cases), the consensus advisory was the modality/timeframe combination supported by at least two annotators; if all three differed, the senior annotator's draft was used and flagged as such in the release.

Aspect	Description
Annotators	3 board-certified radiologists; mean 11.2 yrs post-fellowship (range 7–16); 2 US, 1 EU.
Items	1,024 Gold reports (de novo authoring) + 2,000 Silver advisories (Tier-2 review).
Inputs shown	Findings, Impression, original radiograph, demographics; matched ACR variant shown *after* draft submission. Pipeline output never shown during Gold authoring.
Labels	Advisory category, modality, timeframe, urgency, referral, ACR variant, clinical caveats.
Blinding	System identity hidden during evaluation of baseline outputs.
Adjudication	Three-radiologist consensus; structured 30-min discussion on category disagreement (18% of cases).
Agreement	Fleiss' κ: 0.82 (category), 0.78 (urgency), 0.81 (safety on system outputs); Krippendorff's $\alpha = 0.74$ (deployability).
Compensation	USD 120 / hr; clinical research network recruitment; IRB-exempt protocol on de-identified data.
Instructions	Written annotation guidelines provided; available in the release.

Table 13: Annotation protocol summary.

F Baseline Multi-Agent Pipeline (RadAdvise-Agent v1): Prompts and Schemas

This appendix documents the baseline four-agent pipeline that the experiments in §7 treat as *RadAdvise-Agent v1*; the components RadAdvise-Loop adds on top are documented in Appendix G. Each agent is a single GPT-4o call with a fixed prompt and a constrained output schema, run in sequence.

F.1 Agent 1 — Finding Extraction

agent1_extraction.txt

```
You are a radiology finding extraction agent. Given a chest X-ray report with Findings and Impression, extract: (1) pathology mentions with anatomical location; (2) severity (mild/moderate/severe/unspecified); (3) temporal qualifier (new/unchanged/worsening/improved/chronic); (4) relevant negatives; (5) is_critical = true if the finding requires urgent action.
{ "entities": [{ "name", "location", "severity", "temporal", "is_critical", "evidence_span" }] }
```

F.2 Agent 2 — ACR Guideline Retrieval (baseline)

The baseline retriever is a hybrid dense+BM25 retriever over a *flat* ACR variant index. The biencoder is PubMedBERT (Gu et al., 2021) fine-tuned on variant–procedure pairs with hard-negative mining over sibling variants; the hybrid score is $0.7\, s_{\text{dense}} + 0.3\, s_{\text{BM25}}$; and the top $K = 5$ variants are returned per query. The flat index is the component the Guideline Evidence Graph (Appendix G.2) replaces.

F.3 Agent 3 — Advisory Composition

agent3_composition.txt

```
You are a radiology advisory composition agent. Given extracted entities and retrieved ACR guidelines, compose four fields: (1) FOLLOW-UP IMAGING (modality, timeframe, ACR rating, RRL); (2) CLINICAL CORRELATION; (3) REFERRAL when warranted; (4) URGENCY (routine/urgent/emergent).
Multi-finding policy: (a) critical/urgent findings first; (b) prefer lower-RRL modalities among non-urgent findings; (c) group unrelated non-urgent findings by anatomical region; (d) merge duplicate modality/timeframe recommendations.
```

F.4 Agent 4 — Safety Verification

agent4_verification.txt

```
You are a radiology advisory safety verifier. Check: H1 no contraindicated procedure; H2 every is_critical entity flagged urgent/emergent; H3 every modality exists in ACR AC for the indication; H4 every advisory entity appears in source entities; H5 each radiation-bearing modality discloses its RRL.
{ "passed": bool, "violations": [...], "action": "accept" | "reject_and_regenerate" | "manual_review" }; max regenerate cycles = 3.
```

F.5 Baseline hyperparameters

All four baseline agents run on GPT-4o (`gpt-4o-2024-11-20`) at temperature 0, top-p 1, and `max_tokens` 1024. The dense retriever uses 768-dimensional PubMedBERT embeddings in a FAISS IVF-PQ index (M=64 subquantizers, `nlist`=4096), and per-advisory baseline inference is 4.1 s at USD 0.018 (Appendix R).

G RadAdvise-Loop: Proposed Agentic RAG Method

This appendix documents the five components RadAdvise-Loop adds to the baseline, in pipeline order: Tier-0 preprocessing, the Guideline Evidence Graph, adaptive multi-agent filtering, the structured planner, claim-level traceability, and the uncertainty-routed expert-correction loop.

G.1 Tier-0 deterministic preprocessing

Tier-0 runs before any LLM agent and exists for one reason: not every report needs guideline-grounded reasoning, and spending full pipeline cost on a report that says "no acute cardiopulmonary abnormality" is wasteful at corpus scale. The stages (Table 14) are deterministic and auditable, which also means their decisions can be inspected and corrected without retraining anything.

The eligibility classifier (Table 15) is the load-bearing stage. It is a rule-based decision over parsed section content rather than an LLM call, so it adds negligible latency and introduces no generation risk of its own. Routing assigns each report to one of four classes by escalating cost:

Stage	Purpose
Section parser	Extract Findings, Impression, Indication, Comparison.
Deduplication	Remove repeated reports and near-duplicate studies.
Eligibility classifier	Route to one of four classes (Table 15).
Context extractor	Extract age band, sex, pregnancy flag, prior-imaging presence, D-dimer, malignancy history.
Uncertainty detector	Flag missing context that affects the recommendation.

Table 14: Tier-0 deterministic preprocessing stages.

reports needing no advisory or a fixed template never reach the graph or the planner, while only guideline-dependent and ambiguous reports incur the full pipeline. On our pool this routes 18.4% of reports (8.7% no-advisory, 9.7% template; Table 11) away from full RAG, which is where the corpus-scale cost saving comes from.

Eligibility class	Action
No advisory needed	Emit "no follow-up needed" template or skip.
Simple deterministic advisory	Rule/template + verifier only.
Guideline-dependent advisory	Full RadAdvise-Loop pipeline.
Ambiguous / high-risk advisory	Full pipeline + mandatory expert sampling.

Table 15: Tier-0 eligibility classes. The classifier is a rule-based decision over section content and is documented in the release.

The design risk here is misrouting: a report sent to the no-advisory or template path that genuinely required guideline reasoning would never be corrected downstream, because the expensive stages it skipped are exactly the ones that would have caught the error. We mitigate this asymmetrically. The classifier is tuned for high recall on the two full-RAG classes, accepting that some reports are escalated unnecessarily (a cost, not a safety issue) in order to minimise under-routing (a safety issue). On a 500-report manually audited sample, under-routing occurred in 0.6% of cases, all of which were borderline "stable chronic finding" reports where a template advisory was clinically defensible.

G.2 Guideline Evidence Graph

A flat variant index treats the ACR AC as 1,284 independent retrieval targets, which discards the structure that actually governs a correct recommendation: a variant belongs to a topic, is modified by patient context, and recommends procedures with specific ratings and radiation levels. Encoding this as a graph lets retrieval respect that structure instead of rediscovering it from text similarity at every query, and makes every retrieved candidate inspectable as a path rather than an opaque nearest neighbour.

Component	Count
Node types (12)	
FINDING nodes	136
INDICATION nodes	250
TOPIC nodes	247
VARIANT nodes	1,284
PROCEDURE nodes	8,812
PATIENT-CONTEXT nodes	412
RATING nodes	9
RRL nodes	6
CONTRAINDICATION nodes	1,934
URGENCY nodes	3
TIMEFRAME nodes	27
REFERRAL nodes	89
Edge types (9)	
finding-suggests-indication	336
indication-maps-to-topic	401
topic-contains-variant	1,284
variant-recommends-procedure	8,812
patient-context-modifies-variant	4,027
procedure-has-rating	8,812
procedure-has-rrl	8,812
high-risk-finding-requires-urgency	318
variant-contraindicates-procedure	1,934

Table 16: Guideline Evidence Graph composition (ACR AC, October 2024 release): 12 node types and 9 edge types.

Schema. The Guideline Evidence Graph is a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ over 12 node types and 9 edge types; its composition is reported in Table 16.

Construction. TOPIC, VARIANT, PROCEDURE, RATING, and RRL nodes are extracted directly from the ACR AC corpus. PATIENT-CONTEXT nodes are derived by parsing each variant description for clinical qualifiers (e.g. "pregnant patient", "low pretest probability") with a rule-based tagger; a *patient-context-modifies-variant* edge is added when a context qualifier appears in the variant name; and CONTRAINDICATION nodes are extracted from the "Usually Not Appropriate" procedure lists.

Graph-constrained retrieval. A query proposes FINDING and INDICATION nodes; a two-hop traversal yields the candidate variant set restricted by any active PATIENT-CONTEXT modifiers; and dense retrieval ranks within this constrained set. Formally, for a query with proposed indication node i and active context set C, the candidate set is the topology-consistent neighbourhood

$$\mathcal{N}(i, C) = \{\, v \in \textsc{Variant} : \; i \rightarrow \text{topic}(v), \;\; \text{ctx}(v) \subseteq C \,\},$$

and dense retrieval ranks only over $\mathcal{N}(i, C)$. The average candidate-set size after the constraint is $|\mathcal{N}| = 8.3$ versus 1,284 for unconstrained dense retrieval, and the index builds in 14 minutes on a single A100.

Why this helps. The 8.3-versus-1,284 reduction is not primarily a speed argument; it is a precision argument. Unconstrained dense retrieval can return a variant from the correct-sounding but topologically wrong topic (for instance, a cardiac variant for a finding whose indication maps only to a thoracic topic), and these are exactly the errors that produce confident, fluent, guideline-violating advisories. By restricting candidates to topology-consistent variants before ranking, the graph removes this failure mode by construction, which is why the retrieval comparison (Table 4) shows the graph constraint improving top-1 variant accuracy even though it leaves the dense encoder unchanged. The main construction risk is the rule-based context tagger: a missed qualifier (e.g. an unusual phrasing of "pregnant") drops the patient-context edge and lets an inappropriate variant survive the constraint. We measured tagger recall at 0.94 on a 300-variant manual audit; the residual 6% is the dominant source of the "wrong variant" errors that remain in Table 23.

G.3 Adaptive multi-agent variant filtering

Four filtering agents run in parallel on the constrained candidate set, each emitting a score in $[0, 1]$ and a one-line rationale.

```
filter_agent.txt

You are the {Indication, Patient-Context, Finding, Radiation}
agent.  Given a candidate variant {topic, variant_id,
description, procedures}, the extracted entities, and patient
context {age_band, sex, pregnancy, prior_imaging, ...}, score
the variant on your axis and explain in one line.
{ "score": float, "rationale": string }
```

The consensus score $s(v)$ combining the four agents with the dense and BM25 retrieval signals is defined in §5.2. The keep set is then chosen by the largest-gap rule of MAIN-RAG (Chang et al., 2025) rather than a fixed top-K: sorting the m candidate scores as $s_{(1)} \geq s_{(2)} \geq \cdots \geq s_{(m)}$, we keep the top $k^\star$ candidates, where

$$k^\star = \underset{1 \leq k < m}{\arg\max} \left(s_{(k)} - s_{(k+1)} \right),$$
$$\mathcal{V}_{\text{keep}} = \left\{ v : \text{rank}(v) \leq k^\star \right\}.$$

Empirically, this admits 1–4 variants for typical reports and 5–8 for indication-ambiguous reports (mean 2.7, std 1.3), so the candidate budget adapts to how ambiguous the indication is rather than being fixed in advance.

G.4 Structured advisory planner

```
advisory_planner.txt

You are a radiology advisory planner. Given extracted entities,
filtered variants, and patient context, emit a JSON plan with
no prose; each field is validated independently against the
Guideline Evidence Graph.
{ "recommendations": [{
  "finding_ref":     "...",    "clinical_problem":     "...",
"variant_id": "...",
  "recommended_action": "...", "timeframe": "...",
  "urgency": "routine|urgent|emergent", "rrl": "0|1|2|3|4|5",
  "referral":   null  |  "...",  "missing_context":   [...],
"abstain": false,
  "claim_evidence": [{ "claim_id", "type", "source_span" |
"graph_path" }]
}, ... ] }
```

The planner is GPT-4o at temperature 0. Each emitted plan is validated independently against $\mathcal{G}$ before surface realisation: (a) every `variant_id` exists in $\mathcal{G}$; (b) `rrl` matches the *procedure-has-rrl* edge; (c) `timeframe` is in the 27-element timeframe vocabulary; and (d) when `missing_context` is non-empty, `abstain` or an explicit caveat must be set. Because validation is per field, a single bad field can be regenerated without discarding a correct plan.

G.5 Claim-level traceability schema

Every advisory claim carries at least one typed evidence pointer; the five types and their pointer formats are listed in Table 17.

Evidence type	Pointer
Source-evidence-span	Character offsets into F or I.
Guideline-evidence	Triple (topic, variant, procedure) from $\mathcal{G}$.
RRL-table-entry	Procedure–RRL edge.
Contraindication-rule	Contraindication node + variant scope.
Missing-context-caveat	Set of context keys not satisfied.

Table 17: The five evidence types in the traceability schema. Every claim must carry at least one evidence pointer.

Traceability Accuracy. For an advisory with claims $c_1, \ldots, c_n$ and cited evidence $e_1, \ldots, e_n$, Traceability Accuracy is the fraction of claims whose cited evidence actually supports them:

$$S_{\text{trace}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left[\text{supports}(e_i, c_i) \right],$$

where $\text{supports}(\cdot, \cdot)$ is exact-match verification against $F \cup I$ for source spans and against $\mathcal{G}$ for graph pointers. The release implementation is deterministic, so S_{trace} is reproducible without an LLM judge.

G.6 Uncertainty score and routing

The advisory-level uncertainty score U defined in §5.5 combines the five terms in Table 18, whose weights are fit by logistic regression against radiologist deployability labels on a 1,200-advisory development set.

Term	Quantity	Definition
$1 - p_{\text{ret}}$	Retrieval conf. gap	1− top-1 consensus score from §5.2.
D_{judge}	Judge disagreement	Std. dev. of CAC pass rate across 3 LLM judges.
H_{CAC}	CAC violation severity	Weighted count of soft-constraint violations.
M_{ctx}	Missing context	|missing_context| from planner output.
R_{risk}	Finding risk	Bucketed risk score from criticality flags.

Table 18: Components of the uncertainty score U. Weights $w_1, \ldots, w_5$ are fit on a 1,200-advisory development set against radiologist deployability labels via logistic regression; learned values are $w = (0.18, 0.21, 0.31, 0.19, 0.11)$.

The learned weights are themselves informative. CAC-violation severity ($w_3 = 0.31$) dominates, which is expected: an advisory that already trips a hard or soft constraint is the strongest single predictor that a radiologist will judge it undeployable. Judge disagreement ($w_2 = 0.21$) and missing context ($w_4 = 0.19$) contribute next, capturing two distinct failure routes—evaluator uncertainty and genuine input insufficiency—that the CAC alone does not see. Retrieval-confidence gap ($w_1 = 0.18$) and finding risk ($w_5 = 0.11$) round out the score; finding risk is weighted lowest not because risk is unimportant but because high-risk findings are already routed to mandatory expert sampling by Tier-0, so the term mostly adds signal for borderline-risk cases the classifier did not escalate. The routing thresholds (Table 2) are set on the development set to equalise expected expert load across rounds rather than to hit a fixed accept rate, so that review effort is predictable for the annotating radiologists; the efficiency consequence of this choice is quantified in §7 and Appendix K.

G.7 Expert correction taxonomy and training signal

The taxonomy in Table 19 is what turns expert review from validation into improvement. A conventional human-in-the-loop setup uses radiologist judgements only to score the system; here each correction type is routed to the specific component responsible for it, so a wrong-variant correction trains the retriever, a wrong-timeframe correction trains the planner, and an RRL or urgency correction recalibrates a specific verifier constraint. This routing matters because the failure modes are not interchangeable: pooling all corrections into a single fine-tuning signal would dilute the timeframe corrections (which target the planner's weakest axis) with variant corrections (which target a different module entirely). Keeping them separate is what lets a small number of corrections move the metric they are aimed at.

Correction	Example	Training signal
Variant correction	PE pregnancy variant missed	Retriever hard negative.
Modality correction	CT → US first	Guideline-filter agent supervision.
Timeframe correction	12 mo → 6–12 mo	Planner timeframe head fine-tune.
Recommendation addition	Effusion follow-up missing in multi-finding	Planner multi-finding training.
RRL correction	Missing radiation disclosure	Verifier (H5) calibration.
Context correction	Risk factors not addressed	Abstention policy training.
Urgency correction	Routine → urgent	Verifier (H2) calibration.

Table 19: Expert correction types and how each is consumed as supervised signal. Corrections accumulate per round; the retriever and planner are refreshed every 5,000 corrections.

G.8 Loop training cycle

Each loop iteration runs six steps: (i) generate advisories with the current models; (ii) compute the uncertainty score U; (iii) route high-U cases to radiologists; (iv) collect structured corrections; (v) once 5,000 corrections accumulate, retrain the retriever with the new hard negatives, fine-tune the planner timeframe head, and recalibrate verifier thresholds; and (vi) re-evaluate on a held-out development set.

Three iterations were run to produce the Silver tier, with per-iteration Silver-CAC gains of $+2.1$, $+1.4$, and $+0.7$ (component-level numbers in Table 21). The diminishing returns are the intended signal that the loop is converging rather than memorising: each round, corrections increasingly concentrate on genuinely ambiguous cases—where even radiologists disagree, consistent with the $\kappa = 0.74$ safety agreement—rather than on systematic model errors, so the marginal value of further correction falls. We stopped at three iterations because the fourth round's projected gain ($< +0.4$ on the development set) did not justify the additional 5,000-correction annotation cost. This is the practical argument for uncertainty routing over exhaustive

review: the loop extracts most of its available improvement from a small, targeted fraction of the corpus, and the convergence curve tells you when to stop.

H Formal Metric Definitions

Let $\mathcal{E}$ denote the set of clinical entities extracted from the source report, $\mathcal{R}$ the set of recommendations in the candidate advisory, $\mathcal{C} \subseteq \mathcal{E}$ the entities flagged `is_critical = true`, and $\mathcal{E}_{\text{safety}} \subseteq \mathcal{E}$ the safety-relevant entities. For a recommendation r, v_r denotes its matched variant, $\text{rating}(r)$ its ACR appropriateness rating, and $J(r) \in \{0, 1\}$ a recorded clinical justification. All six metrics lie in $[0, 1]$ with higher being better.

Clinical Safety. S_{safety} is zero on any safety violation and otherwise rewards the fraction of safety-relevant entities handled correctly:

$$S_{\text{safety}} = \begin{cases} 0 & \exists\, r \in \mathcal{R} : r \in \text{Contraindicated}(\mathcal{E}), \\ 0 & \exists\, e \in \mathcal{C} : \text{urgency}(e) \neq \text{urgent}, \\ 1 - \frac{|\text{missed}(\mathcal{E})|}{|\mathcal{E}_{\text{safety}}|} & \text{otherwise.} \end{cases}$$

The hard zero on contraindication or unflagged criticality is deliberate: these are the failures that cause patient harm, so no amount of correct handling elsewhere should compensate, and the metric is non-compensatory by construction rather than a weighted average.

Guideline Adherence. S_{guide} is the fraction of recommendations that are guideline-supported, either by appearing in the matched variant with a sufficient rating or by carrying an explicit justification:

$$S_{\text{guide}} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \mathbf{1}\big[(r \in \text{ACR}(v_r) \wedge \text{rating}(r) \geq 7) \vee J(r) = 1\big].$$

The disjunction admits two routes to credit—either the recommendation matches a high-rated (≥ 7) procedure in the patient-specific variant, or it carries an explicit clinical justification $J(r)=1$—so that clinically sound off-guideline recommendations (e.g. a documented contraindication to the first-line study) are not penalised as guideline misses.

Semantic Coherence. S_{cohere} is the mean entailment probability of the advisory under the Findings and the Impression:

$$S_{\text{cohere}} = \tfrac{1}{2}\big(P(\text{adv} \mid F) + P(\text{adv} \mid I)\big),$$

where P is the entailment probability from a BioMed-RoBERTa NLI model fine-tuned on MEDIQA-NLI (Abacha et al., 2019).

Factual Grounding. S_{fact} penalises hallucinated entities, where $\mathcal{H}$ is the set of advisory entities not grounded in the source:

$$S_{\text{fact}} = 1 - \frac{|\mathcal{H}|}{|\mathcal{R}_{\text{entities}}|}.$$

Radiation Awareness. S_{rad} operationalises ALARA by penalising, per recommendation, the excess RRL over the optimal comparable-appropriateness procedure opt_r:

$$S_{\text{rad}} = 1 - \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} \max\Big(0,\ \frac{\text{RRL}(r) - \text{RRL}(\text{opt}_r)}{\text{RRL}_{\max}}\Big).$$

Traceability Accuracy. S_{trace} is the fraction of claims whose cited evidence supports them, as defined in Appendix G.5:

$$S_{\text{trace}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\big[\,\text{supports}(e_i, c_i)\,\big].$$

Appropriate Abstention Rate. Over the indices i whose gold annotation marks the case context-insufficient, AAR is the fraction on which the system abstains or caveats:

$$\text{AAR} = \frac{|\{\, i : \text{abstain}_i \ \wedge\ \text{gold-ctx-insufficient}_i \,\}|}{|\{\, i : \text{gold-ctx-insufficient}_i \,\}|}.$$

I Clinical Acceptability Checklist

The Clinical Acceptability Checklist is the deployment gate of RadAdvise-Eval: an advisory is acceptable only if it clears every hard constraint and reaches a soft-constraint mean of ≥ 0.7 (§6). The split between the two constraint classes is deliberate. Hard constraints (Table 20, H1–H5) encode failures that are unsafe regardless of how well-written the rest of the advisory is—a contraindicated procedure or an unflagged critical finding cannot be averaged away by good prose elsewhere—so any single hard failure zeroes the checklist. Soft constraints (S1–S6) encode quality dimensions where partial credit is meaningful and a single lapse should not condemn an otherwise sound advisory, so they are scored continuously and averaged. This asymmetry is what lets the CAC behave as a safety gate rather than a quality score: it is strict where harm is possible and graded where it is merely suboptimal.

ID	Condition
Hard constraints (all must pass)	
H1	No contraindicated procedure 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.
H5	Radiation-bearing procedures include RRL disclosure.
Soft constraints (scored 0–1; mean ≥ 0.7)	
S1	Follow-up timeframe appropriate for finding severity.
S2	Non-ionising alternatives considered where clinically equivalent.
S3	Language unambiguous and actionable.
S4	Patient-specific factors addressed when available.
S5	Advisory consistent with the stated Impression.
S6	Recommended procedure rated ≥ 7 or explicit justification provided.

Table 20: Clinical Acceptability Checklist.

Of the eleven constraints, only the two soft constraints requiring a judgement of clinical appropriateness (S1 timeframe and S6 rating-or-justification) are scored by the LLM-as-judge; the remaining nine are deterministic checks against the source report and the graph, which is why the CAC is reproducible up to the bounded judge variance quantified in Appendix P.

J Component Ablation

Table 21 reports the full component ablation summarised in Ablation (§7), adding each RadAdvise-Loop component to the row above and tracking Silver-CAC, Gold-CAC, and temporal error jointly. Two patterns are worth isolating. First, the structured planner is responsible for almost the entire temporal-error reduction (27% → 18%), confirming that timing failures are a planning problem rather than a retrieval one. Second, the Gold-CAC gain from each component exceeds its Silver-CAC gain, and the gap widens for the later components, because those components target the failure modes that specifically distinguish radiologist-written advisories from pipeline output rather than the failures already visible on Silver.

K Expert-Routing Efficiency

Table 22 gives the full random-versus-uncertainty-routed review curves behind the efficiency claim of §7. The comparison is run over the four-agent baseline (v1) outputs, so the 0% row equals the un-reviewed v1 Silver CAC of 76.3, and the 100% row is identical for both regimes because every advisory is then reviewed. The headline reading is the

Configuration	Silver CAC	Gold CAC	Temp. err.
GPT-4o zero-shot	34.2	33.1	28%
+ topic-level RAG	47.8	46.5	28%
+ variant-level (hybrid) RAG	62.4	60.8	27%
RadAdvise-Agent v1 (four-agent)	76.3	73.8	29%
+ Guideline Evidence Graph	78.5	76.0	28%
+ adaptive multi-agent filter	80.7	78.4	27%
+ structured advisory planner	82.6	80.9	18%
+ claim-level traceability	83.4	81.7	18%
+ expert-loop calibration (**Loop**)	**84.1**	**83.4**	**16%**

Table 21: Component ablation. Each row adds one component to the row above. The structured planner contributes the largest temporal-error reduction; expert-loop calibration adds 1.7 Gold-CAC points by retraining the retriever and planner on accumulated corrections.

Review %	Random	U-routed	Expert hrs
0%	76.3	76.3	0
1%	76.6	78.9	4.2
5%	77.4	81.5	21
10%	78.2	82.6	42
20%	79.4	83.4	84
100% (oracle review)	83.7	83.7	420

Table 22: Silver-CAC under random vs. uncertainty-routed expert review on 5,000 test advisories, applied to the four-agent baseline (v1) outputs (0% row = un-reviewed v1, 76.3). Uncertainty-routed review at 1% sampling (4.2 hours, 78.9 CAC) approaches the quality random review reaches only at 20% sampling (84 hours, 79.4 CAC) — comparable quality at roughly 5% of the expert time. Expert hours estimated at 5 min/advisory.

1%-versus-20% comparison: uncertainty-routed review at 1% sampling (4.2 expert hours, 78.9 CAC) reaches the quality random review attains only at 20% sampling (84 hours, 79.4 CAC), i.e. comparable quality at roughly 5% of the radiologist time. Expert hours are estimated at 5 minutes per advisory.

L Variant Retrieval Error Decomposition

Retrieval error type	Hybrid	Graph+filter
Wrong topic	31	18
Correct topic, wrong variant	52	31
Correct variant, incomplete procedure	17	51

Table 23: Retrieval error decomposition (percent of remaining errors) for hybrid retrieval vs. graph-constrained + adaptive filtering. The remaining errors shift toward incomplete-procedure failures, which the planner addresses downstream.

The shift visible in Table 23 is the point of the design, not a side effect. Under hybrid retrieval, the majority of errors (52%) are within-topic variant confusions—the right topic is found but the wrong sibling variant is selected—and these are

the most dangerous errors because they yield a recommendation that is guideline-grounded for the wrong patient context. Graph constraints and multi-agent filtering together cut wrong-topic errors from 31% to 18% and wrong-variant errors from 52% to 31%. What remains after filtering is dominated by "correct variant, incomplete procedure" (51% of a much smaller error pool): the right variant is retrieved, but not every appropriate procedure from it is carried into the advisory. This is deliberately a different kind of error—a completeness failure, not a safety failure—and it is handled downstream by the structured planner (§5.3) rather than by retrieval. In other words, the retrieval stage is engineered to convert dangerous variant-selection errors into benign completeness errors that a later stage can recover, which is why the absolute error count falls and the residual mix moves toward the least harmful category.

M Per-Pathology Performance

Table 24 reports per-pathology CAC for v1 versus RadAdvise-Loop, and Figure 6b plots the gains sorted by magnitude. The gains are largest exactly where variant ambiguity is highest. Lung nodules (+9.2) and fractures (+8.6) are the two categories where the correct recommendation hinges on context the report often omits—nodule risk profile and prior imaging, fracture acuity—and these are the categories where graph-constrained retrieval and the abstention policy contribute most. Categories with a more deterministic mapping from finding to recommendation (pneumothorax, +5.6) gain less because the baseline was already close to ceiling. This pattern is consistent with the component ablation: the Loop's advantage concentrates in cases that require disambiguation, not in cases a flat retriever already handles.

Pathology	v1 CAC	Loop CAC	Δ	N
Pneumonia	79.2	86.1	+6.9	1,042
Pleural effusion	77.8	84.9	+7.1	1,108
Lung nodule	71.4	80.6	+9.2	487
Pneumothorax	82.6	88.2	+5.6	184
Cardiomegaly	75.6	83.4	+7.8	793
Atelectasis	78.1	85.0	+6.9	612
Fracture (rib/clav.)	74.0	82.6	+8.6	221
Edema	74.8	82.8	+8.0	553

Table 24: Per-pathology CAC for v1 vs. RadAdvise-Loop. Largest gains are on lung nodules and fractures, where variant ambiguity (incidental vs. suspicious; acute vs. chronic) is greatest.

N Errors by Advisory Category

Error type	GPT-4o	v1	Loop
Hallucinated recommendation	14%	3%	1%
Guideline miss	38%	11%	11%
Urgency miss	12%	8%	8%
RRL violation	18%	9%	9%
Temporal error	11%	29%	16%
Omission	7%	40%	27%
Insufficient-context (correct)	0%	0%	28%

Table 25: Error distribution (200 errors per system). For RadAdvise-Loop, the dominant residual "failure" is appropriate caveating or abstention on context-insufficient cases, which counts as an error in this error analysis but as correct behaviour under AAR.

Table 25 requires a careful reading because of the last row. For RadAdvise-Loop, 28% of what this error analysis counts as "errors" are cases where the system declined to give a precise recommendation and the gold annotation independently flagged the case as context-insufficient—that is, the system was correct to abstain. We count these as errors here only to keep the error taxonomy comparable across systems (the baselines never abstain, so they have no such row), but under the Appropriate Abstention Rate (§6) the same behaviour is scored as success. Excluding correct abstentions, the Loop's genuine error rate is substantially lower than the raw distribution suggests, and the remaining genuine errors are dominated by omission (27%)—a completeness failure rather than a safety failure, consistent with the retrieval error decomposition in Appendix L.

O Metric Validation

Table 26 gives the per-metric Spearman correlations with radiologist deployability summarised in §7. The ordering is monotone in how much clinical structure a metric encodes: lexical metrics (BLEU 0.12, ROUGE-L 0.18, BERTScore 0.24) correlate weakly, clinical-entity metrics (F1-RadGraph 0.41, GREEN 0.47, RaTEScore 0.52) moderately, and the RadAdvise-Eval metrics most strongly, with the composite at 0.84. Notably each individual RadAdvise-Eval metric (S_{safety} 0.78, S_{guide} 0.82, S_{trace} 0.79) already exceeds the best entity metric, so the composite's strength is not an artefact of aggregation alone.

P Cross-Judge Robustness Detail

The agreement levels in Table 27 support the validity of the LLM-as-judge components without claim-

Metric	Spearman ρ [95% CI]
BLEU	0.12 [0.07, 0.17]
ROUGE-L	0.18 [0.13, 0.23]
BERTScore	0.24 [0.19, 0.29]
F1-RadGraph	0.41 [0.37, 0.45]
GREEN	0.47 [0.43, 0.51]
RaTEScore	0.52 [0.48, 0.56]
S_{safety} (ours)	0.78 [0.75, 0.81]
S_{guide} (ours)	0.82 [0.79, 0.85]
S_{trace} (ours)	0.79 [0.76, 0.82]
RadAdvise composite	**0.84 [0.82, 0.86]**

Table 26: Spearman correlation with radiologist deployability ratings on the 500-advisory Gold subset of Table 3 (N=1,500 ratings = 500 advisories $\times$ 3 radiologists; binary deployable / not).

Judge	Cohen's κ vs. GPT-4o	$\lvert\Delta\text{CAC}\rvert$
Claude 3.5 Sonnet	0.79	≤ 1.4
Llama 3.1 70B	0.71	≤ 1.9

Table 27: Cross-judge agreement on 1,000 advisories. System ranking is preserved across all three judges; absolute CAC differences are within 2 points.

ing they are interchangeable. Claude agrees with the GPT-4o judge at $\kappa = 0.79$ (substantial) and Llama-3.1-70B at $\kappa = 0.71$ (still substantial, but lower, consistent with its smaller capacity). What matters for the benchmark is not that the judges agree on every advisory but that the system ranking is invariant to the choice of judge: across all three, RadAdvise-Loop ranks first and the standalone LLMs rank last, with CAC differences under 2 points. The residual disagreement concentrates on soft-constraint borderline cases (e.g. whether a timeframe is "appropriate" for a stable finding), not on the hard safety constraints, which the judges agree on almost perfectly.

Q Cross-Guideline Transfer: ESR iGuide

Q.1 Protocol

The Guideline Evidence Graph is rebuilt from the chest subset of ESR iGuide (73 scenarios) with the same node/edge schema, and RadAdvise-Loop is re-run with identical hyperparameters on the 412 Gold scenarios covered by ESR. Mapping ACR variants to ESR scenarios was performed manually by one radiologist for verification.

Q.2 Interpretation

This experiment tests architectural portability, not full clinical equivalence across regions: ESR iGuide chest coverage is narrower than ACR AC chest coverage, so coverage of less common scenarios

Configuration	Gold CAC	Δ vs. ACR
RadAdvise-Loop (ACR graph)	83.4	—
RadAdvise-Loop (ESR iGuide graph)	80.8	–2.6
RadAdvise-Agent v1 (ACR flat)	73.8	—
RadAdvise-Agent v1 (ESR flat)	71.4	–2.4

Table 28: Cross-guideline transferability for both pipelines.

is reduced, and a full ESR iGuide re-index for thoracic imaging with a corresponding human-written Gold tier under European practice remains open as multi-region future work. The near-identical drop for both pipelines (-2.6 for Loop, -2.4 for v1) is itself the informative result: it indicates the transfer cost is attributable to ESR's narrower chest coverage rather than to anything specific in the Loop architecture. If the graph and planner were overfit to ACR's structure we would expect the Loop to lose more than the flat baseline on transfer; that they lose the same amount suggests the components are guideline-agnostic, which is the portability claim the experiment is designed to test.

R Latency and Cost

Stage	Wall (s)	Tokens	USD
RadAdvise-Agent v1 (baseline)			
Agent 1 (Extraction)	0.7	480	0.0021
Agent 2 (Retrieval)	0.3	—	—
Agent 3 (Composition)	2.6	2,890	0.0123
Agent 4 (Verification)	0.5	720	0.0033
v1 total	**4.1**	**4,090**	**0.0177**
RadAdvise-Loop (proposed)			
Tier-0 preprocessing	0.2	—	—
Graph-constrained retrieval	0.4	—	—
Multi-agent filtering (4)	2.8	3,120	0.0134
Structured planner	3.6	3,480	0.0149
Traceable generator	2.1	1,740	0.0075
Verifier + uncertainty	0.9	1,640	0.0070
Routing	0.2	—	—
Expert review (amortised)	2.1	—	—
Loop total	**12.3**	**9,980**	**0.0428**
Loop with distilled filters (Llama-3.1-8B)	7.1	5,120	0.0235

Table 29: Per-advisory cost and latency on GPT-4o (April 2026 prices). Distilling the four filtering agents to Llama-3.1-8B reduces cost by 45% with a 1.5-point CAC drop.

RadAdvise-Loop costs roughly $2.4\times$ the baseline per advisory (USD 0.043 vs. 0.018) and $3\times$ the wall-clock time, almost entirely from the four filtering agents and the planner, which together account for 66% of Loop token cost. Whether this is justified depends on the deployment: for one-time Silver-tier construction the absolute figures are small (the full corpus costs under USD 18k at Loop rates), and the Tier-0 router removes 18.4%

of reports from this cost entirely. For latency-sensitive interactive use, the distilled configuration (Llama-3.1-8B filters) recovers most of the speed at USD 0.024 and 7.1 s while giving up only 1.5 CAC points, which is the configuration we would recommend for anything other than offline benchmark construction.

S Demographic Fairness

We stratify RadAdvise-Loop's evaluation by the two demographic attributes recoverable from the de-identified reports—age band and sex—to check that performance does not concentrate or degrade in any subgroup. Table 30 reports the four headline metrics per stratum. Across all comparisons the spread is within sampling noise: S_{safety}, S_{guide}, and S_{rad} vary by at most 0.01, and Gold CAC by at most 1.3 points, with no comparison reaching Bonferroni-corrected significance over the 30 tests. The absence of a measured disparity is consistent with the design—retrieval and verification are conditioned on clinical context rather than on demographic attributes—but it is not a guarantee: the pediatric stratum is the smallest and therefore the least powered to detect a true difference, and sex is the only such attribute available, so disparities along axes we cannot observe (race, socioeconomic status, institution) remain untested. We therefore report this as evidence of no detectable bias on the measured axes, not as a fairness certification, and recommend the per-subgroup monitoring noted in §S for any deployment.

Stratum	S_{safety}	S_{guide}	S_{rad}	**Gold CAC**
Pediatric (0–17)	0.95	0.91	0.93	82.4
Adult (18–64)	0.96	0.92	0.93	83.7
Elderly (65+)	0.95	0.92	0.93	83.5
Female	0.96	0.92	0.93	83.3
Male	0.96	0.92	0.93	83.5

Table 30: Stratified evaluation of RadAdvise-Loop. No comparison reaches Bonferroni-corrected significance ($p > 0.05$ across all 30 tests).

T Worked Examples and a Failure Case

The two cases below trace the same input through GPT-4o, the v1 baseline, and RadAdvise-Loop, and illustrate the two behaviours the metrics are built to reward: variant-correct recommendation under patient context (Case 1) and calibrated abstention under missing context (Case 2).

T.1 Case 1 — Pregnant patient with suspected PE

Source. 32-year-old, 28 weeks pregnant, acute dyspnoea, positive D-dimer. CXR: no acute findings.
GPT-4o. "Recommend CT pulmonary angiography to exclude PE." *CAC fail.*
RadAdvise-Agent v1. "Given pregnancy, V/Q scan preferred over CTPA. CTPA reserved if V/Q non-diagnostic." *CAC pass*, but missed the US Doppler first step.
RadAdvise-Loop. "Given pregnancy, initial workup with bilateral lower-extremity venous Doppler ultrasound (RRL: O) is preferred per ACR AC Suspected PE Variant 4 (rating 9). If negative and clinical concern persists, V/Q scan (RRL: ●●●) is preferred over CTPA in pregnancy due to lower fetal radiation. URGENCY: urgent. Recommend MFM consultation." Plan: `variant=PE-V4`, `abstain=false`, with traceable citations to the source span "28 weeks pregnant" and the ACR PE-V4 procedure ranking. *CAC pass*; AQL L5; $S_{\text{trace}} = 1.0$.

T.2 Case 2 — 6 mm lung nodule with unknown risk

Source. 6 mm peripheral right upper lobe nodule, indeterminate. No prior imaging, no risk-factor data.
RadAdvise-Agent v1. "Recommend follow-up low-dose CT in 12 months per Fleischner." *CAC fail* (S1: incorrect timeframe under unknown risk).
RadAdvise-Loop. `missing_context = [smoking_history, prior_imaging]`, `abstain = true`, with caveated output: "Follow-up depends on risk profile and prior imaging. If high-risk (smoker, family history) or no prior stability is available, consider CT chest at 6–12 months per Fleischner; if low-risk, follow-up may be optional. **Recommend correlation with smoking history and any prior imaging before scheduling.**" The gold label marks this case context-insufficient. *Correct behaviour under AAR.*

U Reproducibility

This appendix provides the information required to reproduce every result in the paper. Code, configuration files, prompts, and evaluation scripts are released (§U.9); restricted source data and licensed guideline content are reconstructed locally via the scripts described in §U.1.

U.1 Data access and reconstruction

The benchmark is built from four source corpora, each under its own access regime: MIMIC-CXR requires PhysioNet credentialing and a signed data-use agreement; ReXGradient-160K requires its non-commercial research licence; CheXpert Plus requires the Stanford academic licence; and IU X-Ray is openly available. Because we cannot redistribute these reports or the ACR AC / ESR iGuide guideline text, the release ships *reconstruction scripts* rather than data: given a credentialed user's local copy of each corpus and a licensed copy of the ACR AC corpus, `radadvise-data` regenerates the exact Silver and Gold record set by report identifier, and `radadvise-graph` rebuilds the Guideline Evidence Graph. All record selection is keyed on stable source identifiers (PhysioNet `study_id`, ReXGradient `report_uid`), and the release pins the exact identifier lists for the train/val/test splits and the 1,024-report Gold subset, so reconstruction is deterministic across users.

U.2 Environment

All experiments run under Python 3.11 with the dependency set pinned in the released `environment.yml` and `requirements.txt` (exact versions and hashes included). Key libraries: PyTorch 2.3.1 (CUDA 12.1), `transformers` 4.44, `sentence-transformers` 3.0, `faiss-gpu` 1.8, `rank-bm25` 0.2.2, `networkx` 3.3 (graph), `scikit-learn` 1.5 (uncertainty-weight logistic regression), and `scipy` 1.13 (bootstrap). LLM and VLM calls go through a thin provider-abstraction layer (`radadvise-llm`) that logs every request/response pair with its model string and decoding parameters, so any run is fully auditable and cached calls are reused across re-runs.

U.3 Model versions and decoding

Frontier models: GPT-4o (`gpt-4o-2024-11-20`); Claude 3.5 Sonnet (`claude-3-5-sonnet-20241022`); Med-PaLM 2 (Vertex AI `medpalm-2`). Open / vision–language models: LLaVA-Med v1.5; MAIRA-2 (Hugging Face checkpoint released January 2025); Llama-3.1-70B-Instruct and Llama-3.1-8B-Instruct (served locally via vLLM 0.5.4). All generation uses temperature 0, top-p 1, and `max_tokens` 1024 unless a component specifies otherwise (the planner and filtering agents also run at temperature 0). API model strings are pinned; we record access dates in the release because hosted models can change behind a fixed string, and we cache all hosted-model outputs so the reported numbers are reproducible from cache even if a provider updates the underlying weights.

U.4 Component hyperparameters

Hyperparameters are consolidated in Table 31 for convenience; the individual components are described in Appendices F and G.

Component	Setting
Dense biencoder	PubMedBERT, 768-dim, fine-tuned 5 epochs, batch 256, InfoNCE, LR 2×10^{-5}.
Hard-negative mining	Sibling variants within topic, 8 negatives/positive.
FAISS index	IVF-PQ, M=64, `nlist`=4096, `nprobe`=16.
Hybrid retrieval weights	$0.7\, s_{\text{dense}} + 0.3\, s_{\text{BM25}}$; top-$K$=5 (baseline).
Graph traversal	Two-hop from finding/indication nodes; context-modifier restriction on.
Filtering consensus	$(\alpha, \beta, \gamma, \lambda) = (0.45, 0.20, 0.30, 0.05)$; adaptive largest-gap threshold.
Uncertainty weights	$w = (0.18, 0.21, 0.31, 0.19, 0.11)$; logistic regression on 1,200-advisory dev set.
Loop schedule	Retrain every 5,000 corrections; 3 iterations.
CAC soft-pass threshold	Mean soft-constraint score ≥ 0.7.

Table 31: Consolidated hyperparameters for RadAdvise-Loop and the v1 baseline.

U.5 Compute

Our pipeline is predominantly hosted-API inference; only fine-tuning, open-model serving, and NLI scoring use local GPUs. We separate the two budgets because they do not share a unit.

API inference. Silver-tier construction is dominated by hosted-model token cost rather than local compute. At the per-advisory rates in Table 29, generating the 398,087-advisory Silver tier with RadAdvise-Loop costs approximately USD 17k in API spend (USD 0.043/advisory), and the v1 baseline construction approximately USD 7k (USD 0.018/advisory); the Tier-0 router (§G.1) removes 18.4% of reports from the full-pipeline path, which is already reflected in these totals. Wall-clock for construction is set by API throughput and provider rate limits rather than by our hardware.

Local GPU. All local GPU work runs on a single node of 4×NVIDIA A100 (40 GB). The PubMedBERT biencoder is fine-tuned in 6–10 hours on 1–2 cards, and the Guideline Evidence Graph is built on

CPU in ∼15 minutes. For evaluation, Llama-3.1-70B-Instruct is served 4-bit-quantised with tensor parallelism across all four 40 GB cards (the model does not fit unquantised in 160 GB once the KV cache is included), and Llama-3.1-8B-Instruct runs on a single card; serving time is bounded by the evaluation-set size rather than by training, and the full evaluation suite across all open models, three seeds, and both tiers completes in 2–3 days of node time. BioMed-RoBERTa NLI scoring for S_{cohere} is negligible.

Reproducing evaluation only. Reproducing the reported numbers from the released advisories (i.e. without rebuilding Silver) requires no API spend and approximately 1–2 days on the same 4×A100 (40 GB) node, which is the recommended entry point for most reproductions.

U.6 Randomness and seeds

All headline numbers (Tables 3, 21, 4, 22, 5) are the mean over seeds $\{42, 1337, 2024\}$, with per-seed values released as CSVs. The seed controls model sampling order, the train/dev shuffling for bien-coder fine-tuning and the uncertainty-weight regression, and the bootstrap resampler. Temperature-0 decoding makes generation near-deterministic per model string, so the dominant remaining variance is in fine-tuning and in the human-evaluation subsamples; reported standard deviations are < 0.012 on the S_x metrics and < 1.4 points on CAC. The 95% confidence intervals in the main paper are bootstrap intervals over 10,000 resamples of the relevant evaluation set, computed with a fixed resampler seed so the intervals themselves are reproducible.

U.7 Evaluation procedure

RadAdvise-Eval is deterministic given an advisory, the source report, and the indexed graph: the five core metrics, S_{trace}, the CAC, the AQL, and the AAR are pure functions with no sampling (Appendix H). The only stochastic evaluation component is the LLM-as-judge used for the two soft constraints; we fix its model string and temperature 0 and additionally report cross-judge agreement (Appendix P) so that judge-induced variance is bounded and visible. To reproduce a single system's headline row: load the released advisories for that system, run `radadvise-eval` against the reconstructed graph and source records, and aggregate over the three seeds.

U.8 Step-by-step reproduction

Reproduction has two entry points: *evaluation-only*, which recomputes every reported number from the released advisories with no API spend (recommended), and *full reconstruction*, which regenerates advisories from scratch and incurs the API cost in §U.5. Table 32 gives the commands.

Step	Action
Setup (both paths)	
1	Obtain credentialed access to MIMIC-CXR, ReXGradient-160K, CheXpert Plus, and a licensed ACR AC copy.
2	Create the pinned environment (§U.2): `pip install radadvise-eval`; clone `radadvise-pipeline`.
3	`radadvise-data reconstruct` (regenerate Silver/Gold records); `radadvise-graph build --source acr` (build the graph).
Path A — evaluation only (recommended)	
4	`radadvise-eval run --system <name> --seeds 42,1337,2024` reproduces any row of Tables 3–5.
Path B — full reconstruction (compute-heavy)	
4′	`radadvise-loop generate` (proposed) or `radadvise-baseline generate` (v1), then the Step-4 scoring command.
Aggregation (both paths)	
5	`radadvise-eval bootstrap` produces confidence intervals and per-seed CSVs.

Table 32: Reproduction recipe. Setup and aggregation are shared; Paths A and B are alternatives.

U.9 Release artefacts

The release is organised so that the two reproduction paths (§U.8) map onto separate packages: `radadvise-eval` alone reproduces every reported number from the released advisories, while `radadvise-baseline` and `radadvise-loop` regenerate those advisories from scratch. Data and guideline content are never redistributed; `radadvise-data` and `radadvise-graph` instead reconstruct the exact record set and graph locally from a user's own licensed copies, keyed on pinned identifier lists so reconstruction is deterministic across users. All packages share the pinned environment of §U.2, and every released component is versioned against the commit that produced the reported results. Table 33 summarises each artefact.

U.10 License summary

All released artefacts created by us—the generated advisories, the evaluation code, and the pipeline code—are distributed under CC BY 4.0. The source reports remain governed by their respective origin-corpus licences (PhysioNet credentialed access for MIMIC-CXR, the ReXGradient research licence, the CheXpert academic licence, and the open IU X-Ray licence), and we redistribute none

Artefact	Contents
radadvise-data	Deterministic Silver/Gold reconstruction from credentialed corpora, keyed on pinned identifier lists.
radadvise-eval	pip-installable: six metrics, CAC, AQL, AAR, and the bootstrap routine; deterministic outputs.
radadvise-baseline	Reference RadAdvise-Agent v1 (four-agent pipeline) with all prompts.
radadvise-loop	Reference RadAdvise-Loop (Tier-0, graph retrieval, multi-agent filtering, planner, traceable generator, uncertainty router), expert-correction interface, and training scripts.
radadvise-graph	Graph build scripts for ACR AC and ESR iGuide; indexes reconstructed from licensed content, not redistributed.
Supplementary	Per-seed CSVs, cached LLM/VLM outputs, pinned environment files, annotation guidelines, form templates, adjudication protocol, and model/dataset cards.

Table 33: Released artefacts.

of them. ACR AC and ESR iGuide guideline content is likewise not redistributed; we instead provide reconstruction scripts that licensed holders can run against their own copies.