MEDAGENT-X: AN AGENTIC FRAMEWORK FOR RETRIEVAL-GROUNDED AND AUDITABLE MEDICAL IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Automated chest radiograph classification is challenged by uncertain predictions, noisy report-derived supervision, and limited interpretability. We introduce MEDAGENT-X, a retrieval-grounded framework for auditable multi-label chest X-ray classification that explicitly separates visual prediction, evidence-based intervention, and verification. A fine-tuned RAD-DINO encoder predicts 12 radiographic findings, while label-specific thresholds identify uncertain predictions within predefined gray zones. Only these cases are routed to an image-to-report retrieval module that retrieves reports from visually similar training studies and applies a deterministic evidence-based fusion policy. A constrained LLM reviews proposed fusion changes, while a second read-only LLM independently assesses the evidence supporting finalized predictions without modifying diagnostic outputs. On a quality-controlled, patient-disjoint cohort of 5,299 CheXpert Plus studies, MEDAGENT-X achieves a macro-F1 of 0.7246 on the 906-study test set, compared with 0.6971 for the vision-only baseline, and improves gray-zone macro-F1 from 0.6310 to 0.7613. The framework additionally produces structured case-level records containing retrieved evidence, intervention history, and audit provenance. These results show that selective retrieval-grounded intervention can improve uncertain predictions while preserving explicit boundaries between prediction and LLM-based auditing.

1 INTRODUCTION

Chest radiographs are among the most widely used medical imaging examinations, yet reliable multi-label classification remains challenging. Large chest X-ray datasets such as CheXpert and MIMIC-CXR provide valuable supervision at scale, but their labels are commonly derived from radiology reports and therefore inherit uncertainty, missing information, and annotation noise. In this setting, a useful prediction system should provide more than a binary decision. It should also make clear how confident the model is, when additional evidence is needed, and what information contributed to a final prediction. These challenges become particularly important when several findings coexist in the same study, when their visual appearance differs across views, and when disease prevalence varies substantially across labels. A single decision threshold is therefore not necessarily appropriate for all findings Lipton et al. (2014). More importantly, predictions close to a label-specific decision boundary often correspond to cases in which the visual model itself is uncertain. Treating these threshold-adjacent predictions as unconditional decisions ignores useful uncertainty, whereas abstaining from all such cases can substantially reduce coverage. The central question we consider is therefore not only whether a model can classify chest radiographs accurately, but whether uncertain predictions can be selectively examined and revised in a way that remains transparent, traceable, and controlled.

To address this problem, we propose MEDAGENT-X, a retrieval-grounded framework for explainable and auditable multi-label chest radiograph classification. A fine-tuned RAD-DINO image encoder serves as the primary predictor and produces study-level probabilities for 12 predefined radiographic findings Pérez-García et al. (2025). Rather than allowing retrieval or language models to participate in every prediction, MEDAGENT-X introduces them only at explicit decision boundaries. Predictions sufficiently far from their validation-derived, label-specific thresholds retain the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

original vision-model decision. Predictions that fall within a predefined gray zone are instead routed to an image-to-report retrieval module, which retrieves reports associated with visually similar training studies. The retrieved evidence is then evaluated by a deterministic label-fusion policy, and only the resulting proposed changes are passed to a constrained LLM for review. A second, read-only LLM independently evaluates the evidence supporting the finalized prediction but has no authority to modify it.

A key design principle of MEDAGENT-X is the separation of prediction, intervention, and audit. Retrieved reports are treated as contextual evidence from similar historical cases rather than as labels for the target study. The deterministic fusion policy has the initial authority to modify gray-zone predictions, while the first LLM can only preserve, veto, or soften eligible changes under predefined constraints. The second LLM operates strictly as an evidence verifier, producing support assessments, evidence scores, rationales, and references to retrieved cases without changing the diagnostic output. Evaluation is further separated from inference through a deterministic Judge module that compares saved predictions against frozen ground-truth labels without accessing retrieved reports or participating in the prediction process. This separation allows predictive performance, intervention behavior, and evidence quality to be evaluated independently, while preserving a clear record of how each decision was produced.

The main contributions of this work are threefold. First, we introduce a selective retrieval-grounded inference framework in which only threshold-adjacent predictions are eligible for additional intervention, preserving the original vision-model decision outside predefined gray zones. Second, we separate deterministic label fusion, guarded LLM review, and read-only evidence verification so that prediction, intervention, and audit remain explicitly distinguishable. Third, we evaluate the framework on a quality-controlled, patient-disjoint CheXpert Plus cohort and show that retrieval-grounded deterministic intervention improves macro-F1 from 0.6971 to 0.7246 on the full test set and from 0.6310 to 0.7613 within the gray-zone subset, while producing structured evidence records for reviewed predictions.

2 Proposed Method

2.1 Framework Overview

MEDAGENT-X is organized as a two-stage framework for retrieval-grounded and auditable multi-label chest radiograph classification. As illustrated in Figure 1, Stage 1 constructs a quality-controlled, patient-disjoint study-level cohort from DICOM images, metadata, radiology reports, and report-derived labels. Stage 2 uses this cohort for visual prediction and selectively allows retrieved clinical context to influence threshold-adjacent decisions. The framework explicitly separates visual prediction, evidence-based intervention, evidence verification, and evaluation.

MEDAGENT-X is developed and evaluated using the CheXpert Plus dataset Chambon et al. (2024), which provides DICOM chest radiographs, patient- and study-level metadata, associated radiology reports, and CheXpert-style labels. A patient may contain multiple radiographic studies, and each study may contain one or more frontal or lateral views. We therefore formulate prediction at the study level. For a study s containing n_s available views,

$$X_s = \{x_{s,1}, x_{s,2}, \ldots, x_{s,n_s}\},$$

the framework predicts 12 radiographic findings: atelectasis, cardiomegaly, consolidation, edema, pleural effusion, pneumonia, pneumothorax, fracture, lung lesion, lung opacity, enlarged cardiomediastinum, and pleural other. A fine-tuned RAD-DINO encoder Pérez-García et al. (2025) serves as the primary visual predictor. Retrieval and language-model components do not replace the image model. Instead, label-specific thresholds identify predictions close to their decision boundaries, and only these threshold-adjacent predictions are eligible for label-changing intervention. Retrieved reports from visually similar training studies provide contextual evidence for a deterministic fusion policy. A constrained LLM subsequently reviews only proposed fusion changes, while a second read-only LLM evaluates the evidence supporting the finalized predictions. Evaluation is performed separately using frozen test labels that are never exposed to the inference pipeline.

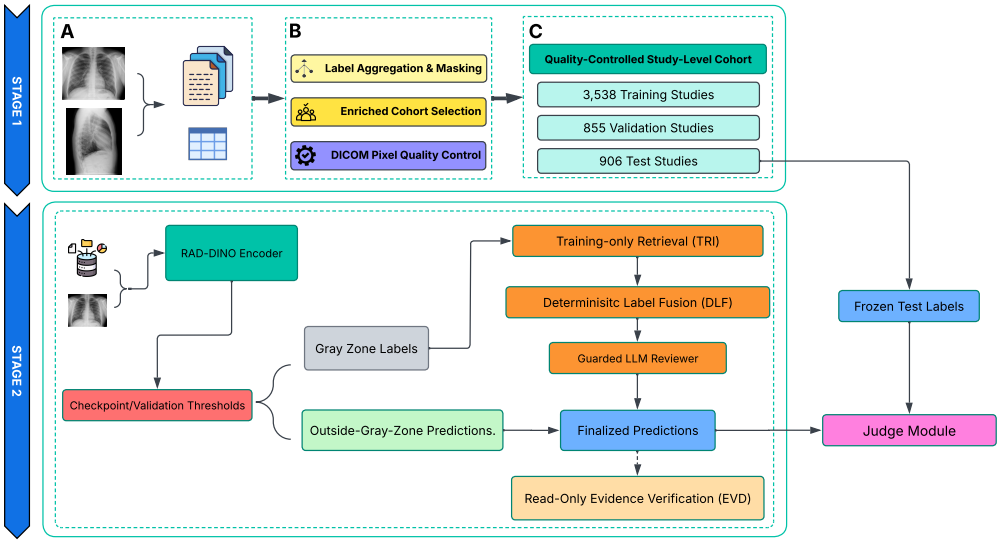


Figure 1: Overview of the MEDAGENT-X framework. Stage 1 constructs a quality-controlled, patient-disjoint study-level cohort through label aggregation and masking, label-aware cohort selection, and DICOM quality control. Stage 2 performs RAD-DINO visual prediction, training-only image-to-report retrieval (TRI), deterministic label fusion (DLF), guarded LLM review, and evidence verification (EVD). The Judge module is isolated from inference and compares finalized predictions with frozen test labels only during evaluation.

2.2 Study-Level Dataset Construction

Stage 1 converts the raw CheXpert Plus data into a quality-controlled study-level cohort for training and evaluation. Because patients may contain multiple studies and individual studies may contain multiple radiographic views, the data are organized hierarchically as patient $\rightarrow$ study $\rightarrow$ image view. Preprocessing decisions are retained at the corresponding levels to preserve traceability.

The complete set of 14 CheXpert-style report-derived labels is retained during preprocessing, while model prediction is restricted to the 12 radiographic findings defined above. *No Finding* and *Support Devices* are retained for consistency checks but are not included among the prediction targets.

Each report-derived finding is represented as one of four states: present, absent, uncertain, or unmentioned. For a study containing multiple views, the corresponding view-level states are aggregated using the precedence

$$\text{present} > \text{uncertain} > \text{absent} > \text{unmentioned}.$$

This rule prevents a positive or uncertain observation in one view from being overwritten by an absent or unmentioned state in another view. Cross-view disagreements are retained as conflict indicators rather than discarded. For supervised learning, present and absent states are converted to binary targets, whereas uncertain and unmentioned states are excluded from the training loss using a supervision mask. When *No Finding* is present and a target finding is otherwise unmentioned, that finding is treated as absent; contradictions involving *No Finding* are recorded separately.

To prevent patient-level leakage, patients are first assigned to training, validation, and test partitions using a deterministic 70/15/15 split. All studies belonging to the same patient remain within the assigned partition. Label-aware cohort selection and subsequent quality control are then performed independently within each partition; therefore, the final study counts need not preserve the exact 70/15/15 ratio. The training partition is used for vision-model optimization and retrieval-index construction, the validation partition for checkpoint, threshold, and policy selection, and the test partition exclusively for final evaluation.

Because the target findings have a long-tailed distribution, a deterministic label-aware enrichment procedure is applied independently within each split. Patients containing underrepresented positive findings are prioritized while limiting the number of selected studies per patient to preserve cohort diversity. Negative examples are subsequently added to improve representation of absent findings. The complete selection criteria and scarcity-weighting procedure are provided in Appendix A.1.

Each selected DICOM view is then subjected to technical quality control. Images are decoded using the appropriate DICOM intensity transformation and evaluated using intensity, contrast, entropy, sharpness, and noise-based indicators. Views with invalid pixel data or extreme technical abnormalities are removed, whereas warning-level views are retained. A multi-view study remains eligible when at least one view passes quality control and is removed only when all available views fail. Definitions of the quality metrics and filtering thresholds are provided in Appendix A.2.

The resulting representation contains quality-eligible image views, study-level targets and supervision masks, patient-disjoint partitions, and the provenance information required by the downstream inference pipeline.

2.3 Visual Prediction and Uncertainty Routing

For each eligible image view $x_{s,v}$, RAD-DINO produces a visual representation

$$z_{s,v} = f_\theta(x_{s,v})[\mathrm{CLS}] \in \mathbb{R}^d,$$

where f_θ denotes the vision encoder and d is the embedding dimension. To obtain a study-level representation from multiple views, we combine element-wise mean and maximum pooling,

$$\bar{z}_s = \frac{1}{n_s}\sum_{v=1}^{n_s} z_{s,v}, \qquad z_{s,j}^{\max} = \max_{1\le v\le n_s} z_{s,v,j},$$

and define

$$h_s = [\bar{z}_s; z_s^{\max}].$$

Mean pooling preserves information shared across views, while maximum pooling retains strong view-specific features. A linear classification head produces one sigmoid score for each target finding,

$$p_{s,y} = \sigma\left(W_y h_s + b_y\right),$$

where multiple labels are predicted independently because several findings may coexist within the same study.

Training uses the study-level supervision mask constructed in Stage 1. For target $t_{s,y}$ and mask $m_{s,y}$, only findings with valid binary supervision contribute to the masked binary cross-entropy objective,

$$\mathcal{L}_{\mathrm{masked}} = \frac{\sum_{s\in\mathcal{B}}\sum_y m_{s,y}\,\ell_{\mathrm{BCE}}(t_{s,y}, p_{s,y})}{\max\left(1, \sum_{s\in\mathcal{B}}\sum_y m_{s,y}\right)}.$$

Rather than applying a shared threshold of 0.5, MEDAGENT-X uses a label-specific operating threshold τ_y selected on the validation set to maximize F1. The complete threshold-selection procedure is provided in Appendix A.3. The corresponding vision-only prediction is

$$\hat{z}_{s,y}^{\mathrm{vis}} = \begin{cases} \text{present}, & p_{s,y} \ge \tau_y, \\ \text{absent}, & p_{s,y} < \tau_y. \end{cases}$$

To identify threshold-adjacent decisions, we define

$$g_{s,y} = \mathbb{I}\left[|p_{s,y} - \tau_y| \le \delta\right], \qquad \delta = 0.15.$$

Predictions outside the gray zone retain the original vision-only status and are not eligible for label-changing intervention:

$$g_{s,y} = 0 \quad \Longrightarrow \quad \hat{z}^{\text{det}}_{s,y} = \hat{z}^{\text{vis}}_{s,y}.$$

Only gray-zone predictions are routed to the retrieval-grounded intervention described in the following section.

2.4 Retrieval-Grounded Intervention and Verification

Only predictions within the gray zone are eligible for retrieval-grounded intervention. For a target study s, the study-level representation h_s produced by RAD-DINO is used to query a retrieval index containing only training-set studies. Each indexed record contains a study embedding and its associated radiology report; ground-truth label values are not included in the retrieval payload. Studies belonging to the same patient as the query are excluded to prevent longitudinal examinations from being used as independent supporting evidence.

Candidate studies are ranked using cosine similarity,

$$\text{sim}(s, r) = \frac{h_s^\top h_r}{\|h_s\|_2 \|h_r\|_2},$$

and the $K = 10$ highest-ranked eligible studies are retained. Retrieved reports are treated only as contextual evidence from visually similar training cases and are never interpreted as labels for the target study. Additional details on index construction and filtering are provided in Appendix A.4.

For each target finding y, a fixed rule-based extractor identifies affirmed and negated mentions in the retrieved reports. Let $c^+_{s,y}$ and $c^-_{s,y}$ denote the resulting positive and negative evidence counts, respectively. The extractor uses predefined finding-specific lexical terms, negation rules, and attributable report fragments so that the evidence entering the fusion policy remains deterministic and reproducible. The complete evidence-extraction procedure is provided in Appendix A.5.

Deterministic fusion operates independently for each finding and may modify only gray-zone predictions. For a visually absent finding, sufficient evidence for promotion is defined as

$$P_{s,y} = \mathbb{I}\left[c^+_{s,y} \ge 3 \ \wedge \ c^-_{s,y} = 0\right].$$

For a visually present finding, we define

$$D_{s,y} = \mathbb{I}\left[c^-_{s,y} \ge 1 \ \wedge \ c^-_{s,y} > c^+_{s,y}\right],$$

and

$$S^-_{s,y} = \mathbb{I}\left[c^+_{s,y} = 0 \ \wedge \ c^-_{s,y} \ge 3\right].$$

The deterministic post-retrieval status is then

$$\hat{z}^{\text{det}}_{s,y} = \begin{cases} \hat{z}^{\text{vis}}_{s,y}, & g_{s,y} = 0, \\ \text{present}, & g_{s,y} = 1,\ \hat{z}^{\text{vis}}_{s,y} = \text{absent},\ P_{s,y} = 1, \\ \text{absent}, & g_{s,y} = 1,\ \hat{z}^{\text{vis}}_{s,y} = \text{present},\ D_{s,y} = 1,\ S^-_{s,y} = 1, \\ \text{uncertain}, & g_{s,y} = 1,\ \hat{z}^{\text{vis}}_{s,y} = \text{present},\ D_{s,y} = 1,\ S^-_{s,y} = 0, \\ \hat{z}^{\text{vis}}_{s,y}, & \text{otherwise}. \end{cases}$$

This deterministic policy establishes the label-changing boundary before any language-model review.

The first LLM is invoked only for labels whose deterministic status differs from the original vision-only prediction. Defining

$$C_s = \left\{ y : \hat{z}_{s,y}^{\mathrm{det}} \neq \hat{z}_{s,y}^{\mathrm{vis}} \right\},$$

the reviewer receives the proposed deterministic change together with the vision score, decision threshold, retrieved evidence, and rule trace. Its role is restricted to reviewing an existing intervention rather than generating a new diagnosis. For each eligible label, the structured action is limited to *keep*, *veto*, or *uncertain*. Schema and policy validation are applied before any action is accepted; malformed, unsupported, or policy-inconsistent outputs fall back to the deterministic result. Detailed review constraints and fallback rules are provided in Appendix A.6.

After the finalized prediction $\hat{z}_{s,y}^{\mathrm{fus}}$ has been established, a second LLM performs read-only evidence verification. This verifier has no label-editing authority and cannot modify the finalized prediction. Instead, it assesses the degree of support and contradiction in the retrieved evidence and produces structured evidence scores, support assessments, rationales, and attributable case references. This separation prevents evidence interpretation from becoming an additional diagnostic action. The verification schema is described in Appendix A.7.

The inference pipeline is orchestrated using a typed LangGraph state graph LangChain (2024). LangGraph controls execution order, state transfer, and trace recording, while all prediction and label-changing decisions remain defined by the modules above. Additional orchestration and state-validation details are provided in Appendix A.8.

2.5 Evaluation Protocol

Evaluation is performed by a deterministic Judge module that remains strictly outside the inference pipeline. The Judge does not access retrieved reports, participate in LLM prompting, or modify predictions. It evaluates saved vision-only and finalized predictions against the frozen study-level test labels constructed in Stage 1; these labels are never exposed to the inference components.

Ground-truth findings labeled *uncertain* or *unmentioned* are excluded from binary performance evaluation. For scoreable positive findings, a finalized prediction of *uncertain* is counted as a missed positive rather than as a cost-free abstention. For scoreable negative findings, an *uncertain* prediction is treated as unresolved. This policy prevents the system from improving apparent performance simply by replacing difficult binary decisions with uncertainty. Performance is evaluated on two fixed scopes: the complete test set and the gray-zone subset

$$\mathcal{G}_{\mathrm{gray}} = \{(s, y) \in \mathcal{G} : g_{s,y} = 1\}.$$

The same evaluation scopes are applied to both vision-only and finalized predictions, allowing overall performance to be distinguished from performance on the threshold-adjacent cases for which intervention is permitted. We report macro and micro precision, recall, and F1, together with coverage. Label-level intervention records are retained to characterize promotions, demotions, uncertain outcomes, and whether each intervention improves, degrades, or leaves the scored prediction unchanged. Evidence-verification outputs are analyzed separately because the verifier is read-only and cannot alter predictive performance.

3 Experimental Results

Dataset and evaluation cohort. All experiments use the quality-controlled, patient-disjoint cohort produced by Stage 1. The finalized cohort contains 5,299 chest radiograph studies: 3,538 training studies, 855 validation studies, and 906 test studies. All studies belonging to a given patient are restricted to a single split.

The training partition is used to fine-tune RAD-DINO and construct the training-only image-to-report retrieval index. The validation partition is used for checkpoint selection, label-specific threshold selection, and development of the deterministic fusion policy. The test partition is held out for final evaluation and is unchanged across all reported experiments and ablations.

Table 1: Study-level cohort used throughout the experiments.

Split	Studies	Experimental role
Training	3,538	Vision-model training and retrieval index
Validation	855	Checkpoint, threshold, and policy selection
Test	906	Final evaluation
Total	5,299	–

Vision-model training. The final RAD-DINO checkpoint fine-tunes the last four transformer blocks while keeping the preceding encoder blocks frozen. Optimization uses AdamW with a learning rate of 10^{-5} for the pretrained backbone and 10^{-3} for the classification head, with weight decay 10^{-4}. A physical batch size of two studies is used with gradient accumulation over four batches, resulting in an effective batch size of eight studies. Training runs for at most 15 epochs with random seed 42 and early stopping after five validation epochs without improvement. The checkpoint with the highest validation masked macro-F1 is retained. Additional optimization details are provided in Appendix B.1.

Thresholds and gray-zone configuration. Each of the 12 target findings uses its own validation-selected operating threshold. The gray-zone margin is fixed to: $\delta = 0.15$

for all experiments. On the 906-study test set, this produces 1,812 gray-zone study-label instances among the 10,872 possible study-label pairs. The gray-zone mask is frozen across all configurations so that comparisons are performed on the same threshold-adjacent predictions.

Retrieval and language-model configuration. Retrieval-enabled configurations use the same training-only index with $K = 10$ nearest eligible studies. The main MEDAGENT-X configuration uses Qwen3-14B for both guarded review and evidence verification, executed locally through Ollama at temperature zero. To examine the sensitivity of guarded review to the choice of language model, an additional ablation uses Llama 3.1 8B as the guarded reviewer while leaving evidence verification disabled. The same validation policies are used in both configurations.

Evaluation protocol. All configurations are evaluated using the same deterministic Judge module. Performance is measured on both the complete test set and the frozen gray-zone subset. We report macro and micro precision, recall, and F1 together with coverage. The vision-only and finalized predictions are evaluated against the same frozen test labels. Evidence-verification outputs are analyzed separately from predictive performance because the verification module is read-only and cannot modify the final label. All reported ablations use the same cohort, checkpoint, label-specific thresholds, gray-zone mask, retrieval index, and Judge policy unless a component is explicitly removed by the ablation.

3.1 Main Results

Table 2 reports the performance of the complete MEDAGENT-X pipeline on the 906-study test set. Evaluation is performed on both the complete test set and the fixed gray-zone subset defined by the vision-only predictions. On the full test set, MEDAGENT-X achieves a macro-F1 of 0.7246, with macro precision of 0.9117 and macro recall of 0.6051. The corresponding micro-F1 is 0.7383. Compared with the vision-only RAD-DINO baseline, which achieves a macro-F1 of 0.6971, the complete pipeline improves macro-F1 by 0.0275. The improvement is more pronounced within the gray-zone subset, where MEDAGENT-X achieves a macro-F1 of 0.7613, compared with 0.6310 for the vision-only baseline. Macro recall increases from 0.5090 to 0.6950, indicating that retrieval-grounded intervention primarily improves the recovery of positive findings close to the visual decision boundary.

Performance varies across the 12 target findings. Lung lesion and lung opacity achieve the highest F1 scores, at 0.8348 and 0.8074, respectively, whereas pneumothorax and enlarged cardiomediastinum show lower F1 values of 0.5930 and 0.6078. These differences highlight variation in performance across individual radiographic findings.

Table 2: Performance of the complete MEDAGENT-X pipeline on the full test set and the fixed gray-zone subset.

Scope	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1	Coverage
Full test	0.9117	0.6051	0.7246	0.9136	0.6195	0.7383	0.1492
Gray zone	0.8773	0.6950	0.7613	0.8824	0.7386	0.8041	0.1837

Table 3: Ablation results on the 906-study test set.

Configuration	Macro P	Macro R	Macro F1	Micro F1	Coverage
Vision only	0.9108	0.5695	0.6971	0.7042	0.1494
Gray-zone abstention	0.9241	0.4619	0.6126	0.6240	0.1428
Deterministic retrieval fusion	0.9117	0.6051	0.7246	0.7383	0.1492
Guarded LLM review	0.9117	0.6064	0.7257	0.7389	0.1491

3.2 Ablation Study

We perform controlled ablations to isolate the contributions of gray-zone routing, retrieved report evidence, deterministic fusion, and guarded LLM review. All configurations use the same test cohort, RAD-DINO checkpoint, label-specific thresholds, gray-zone mask, and evaluation policy. Replacing every gray-zone prediction with an uncertain status does not improve overall performance. Although this strategy increases macro precision from 0.9108 to 0.9241, macro recall decreases substantially from 0.5695 to 0.4619, reducing macro-F1 to 0.6126. This shows that uncertainty routing alone is insufficient and that simply abstaining on threshold-adjacent cases removes many true positive predictions. Introducing training-only retrieval and deterministic label fusion produces the main performance improvement. Macro recall increases from 0.5695 to 0.6051 while macro-F1 improves from 0.6971 to 0.7246. Within the gray-zone subset, macro-F1 increases from 0.6310 to 0.7613. These results indicate that the improvement is primarily associated with retrieval-grounded deterministic intervention rather than uncertainty routing alone. Adding the guarded LLM reviewer results in only a small additional change in predictive performance. This behavior is consistent with its constrained role: the reviewer is designed to inspect proposed deterministic changes rather than function as an additional diagnostic classifier. The ablation therefore separates the predictive contribution of retrieval-grounded fusion from the auditing role of the language-model components.

3.3 Evidence Audit and Provenance

MEDAGENT-X records structured evidence for the finalized study-label decisions. Across the 906-study test cohort, the complete pipeline produces 2,989 label-level evidence records. Of these, 387 correspond to labels whose finalized status differs from the original vision-only status after guarded fusion. The remaining records document unchanged predictions for which evidence verification was performed. Each evidence record retains the visual score and status, retrieved supporting cases, intervention history, finalized label, evidence assessment, and verification provenance. Because the evidence-verification LLM is read-only, these records do not alter the predictive metrics reported in Table 2. Instead, they provide a separate mechanism for examining whether the evidence associated with a finalized prediction is supportive, contradictory, mixed, or insufficient.

3.4 Comparison with GLoRIA

To provide an additional evaluation on expert-labeled data, we compare MEDAGENT-X with the zero-shot GLoRIA vision–language model Huang et al. (2021) on the five-condition CheXpert competition cohort. This experiment is separate from the primary 12-label evaluation described above and contains 200 validation studies and 500 expert-labeled test studies covering atelectasis, cardiomegaly, consolidation, edema, and pleural effusion. Both systems use condition-specific operating thresholds selected on the validation partition and frozen before test evaluation. GLoRIA is evaluated zero-shot using its public ResNet-50 checkpoint and fixed text prompts. MEDAGENT-X uses the same supervised RAD-DINO checkpoint as the primary experiment, together with the

Table 4: Performance on the external five-condition expert-labeled CheXpert test cohort.

System	Macro P	Macro R	Macro F1	Micro P	Micro R	Micro F1
GLoRIA zero-shot	0.2416	0.3839	0.2904	0.2616	0.3845	0.3113
MEDAGENT-X vision only	0.2762	0.6332	0.3703	0.2832	0.6136	0.3875
MEDAGENT-X + fusion	0.2755	0.8022	0.3980	0.2874	0.8097	0.4242

training-only retrieval index and deterministic label-fusion module. Because the two approaches use different training paradigms, this experiment should be interpreted as a system-level comparison under a shared evaluation protocol rather than as a controlled architecture ablation.

As shown in Table 4, MEDAGENT-X with retrieval-grounded fusion achieves a macro-F1 of 0.3980 on this cohort, compared with 0.3703 for its vision-only prediction and 0.2904 for the zero-shot GLoRIA configuration. The increase from the MEDAGENT-X vision-only model to retrieval-grounded fusion is driven primarily by higher recall (0.6332 to 0.8022), while macro precision remains similar. Per-condition results show heterogeneous behavior. Retrieval-grounded fusion improves F1 relative to the vision-only MEDAGENT-X model for atelectasis, cardiomegaly, and pleural effusion, while changes are smaller for consolidation and edema. Complete per-condition results are provided in Appendix D.

4 Discussion and Limitations

The results support selective retrieval-grounded intervention rather than unrestricted language-model-based diagnosis. Deterministic retrieval fusion provides the main predictive improvement, increasing full-test macro-F1 from 0.6971 to 0.7246 and gray-zone macro-F1 from 0.6310 to 0.7613, whereas routing all gray-zone predictions to an uncertain state reduces macro-F1 to 0.6126. The guarded LLM reviewer produces only a small additional change in predictive performance, which is consistent with its intended role as a review mechanism rather than an independent classifier. The second LLM is strictly read-only and contributes evidence assessment and provenance without modifying finalized predictions. Several limitations remain. The primary experiments use a resource-constrained cohort of 5,299 CheXpert Plus studies rather than the complete dataset, which may limit coverage of rare findings. The main 12-label evaluation also relies on report-derived labels that may contain uncertainty and annotation noise, while the expert-labeled external comparison covers only five findings. In addition, the current evidence extractor uses fixed lexical and negation rules, and the gray-zone margin and deterministic fusion policy are validation-selected rather than learned jointly with the vision model. Finally, evaluation is limited to chest radiography and predominantly CheXpert-derived data, so broader generalization remains to be established. Future work should evaluate MEDAGENT-X on larger and more diverse clinical cohorts, extend expert-labeled external validation, and investigate more adaptive retrieval and evidence-fusion strategies while preserving explicit decision boundaries and auditability.

5 Conclusion

We presented MEDAGENT-X, a retrieval-grounded framework for auditable multi-label chest radiograph classification that explicitly separates visual prediction, evidence-based intervention, and evidence verification. A fine-tuned RAD-DINO encoder remains the primary predictor, while only threshold-adjacent predictions are eligible for retrieval-grounded deterministic fusion. Constrained LLM components subsequently review proposed changes and characterize the evidence supporting finalized predictions without acting as unrestricted diagnostic models. On the 906-study test set, retrieval-grounded fusion improves macro-F1 from 0.6971 for the vision-only baseline to 0.7246, with a larger improvement from 0.6310 to 0.7613 within the gray-zone subset. The ablation results indicate that the predictive gain is primarily associated with deterministic retrieval-grounded intervention, while the language-model components provide review, evidence assessment, and provenance. These results show that retrieval and language models can be incorporated into medical-image classification while preserving explicit boundaries between prediction, intervention, and audit.

AI USE STATEMENT

In this work, we used generative AI tools for language editing, grammatical corrections, and improving the clarity and readability of the manuscript. Generative AI tools were not used for developing the methodology, conducting experiments, analyzing data, generating results, or drawing scientific conclusions.

All AI-assisted edits were carefully reviewed and verified by the authors. We take full responsibility for the final content of this work, including any text produced with the aid of generative AI.

REFERENCES

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven Q. H. Truong, Chu The Chuong, and Curtis P. Langlotz. CheXpert Plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. *arXiv preprint arXiv:2405.19538*, 2024. doi: 10.48550/arXiv.2405.19538. URL https://arxiv.org/abs/2405.19538.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *International Conference on Machine Learning*, pp. 233–240, 2006. doi: 10.1145/1143844.1143874. URL https://doi.org/10.1145/1143844.1143874.

Shih-Cheng Huang, Liyue Shen, Matthew P. Lungren, and Serena Yeung. GLoRIA: A multimodal global-local representation learning framework for label-efficient medical image recognition. In *IEEE/CVF International Conference on Computer Vision*, pp. 3942–3951, 2021. doi: 10.1109/ICCV48922.2021.00391. URL https://doi.org/10.1109/ICCV48922.2021.00391.

LangChain. LangGraph overview. Software documentation, 2024. URL https://docs.langchain.com/oss/python/langgraph/overview. Accessed 2026-09-12.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. Optimal thresholding of classifiers to maximize F1 measure. In *Machine Learning and Knowledge Discovery in Databases*, pp. 225–239, 2014. doi: 10.1007/978-3-662-44851-9_15. URL https://doi.org/10.1007/978-3-662-44851-9_15.

Fernando Pérez-García, Harshita Sharma, Sam Bond-Taylor, Kenza Bouzid, Valentina Salvatelli, Maximilian Ilse, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Matthew P. Lungren, et al. Exploring scalable medical image encoders beyond text supervision. *Nature Machine Intelligence*, 7:119–130, 2025. doi: 10.1038/s42256-024-00965-w. URL https://doi.org/10.1038/s42256-024-00965-w.

Mingyuan Zhang, Harish Guruprasad Ramaswamy, and Shivani Agarwal. Convex calibrated surrogates for the multi-label F-measure. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 11246–11255, 2020. URL https://proceedings.mlr.press/v119/zhang20w.html.

A Additional Methodology and Framework Details

A.1 Label-Aware Cohort Selection

Because the CheXpert Plus label distribution is strongly imbalanced, uniform sampling of a resource-constrained subset may provide insufficient positive examples for less frequent findings. MEDAGENT-X therefore applies a deterministic label-aware enrichment procedure independently within the training, validation, and test partitions after patient-level splitting.

For cohort construction, a study s is considered positive for finding y when its aggregated study-level status is *present* and negative when its status is *absent*. Findings labeled *uncertain* or *unmentioned* do not contribute to either count. We define

$$I^+_{s,y} = \mathbb{I}[t_{s,y} = 1 \wedge m_{s,y} = 1], \qquad I^-_{s,y} = \mathbb{I}[t_{s,y} = 0 \wedge m_{s,y} = 1].$$

For patient i, the numbers of positive and negative studies for finding y are

$$P_{i,y} = \sum_{s \in S_i} I^+_{s,y}, \qquad N_{i,y} = \sum_{s \in S_i} I^-_{s,y}.$$

Let $A_{q,y}$ denote the number of available positive studies for finding y in partition q,

$$A_{q,y} = \sum_{i : q_i = q} P_{i,y}.$$

Patients are prioritized using a scarcity-weighted score,

$$R_i = \sum_{y \in \mathcal{Y}} \frac{\min(P_{i,y}, T_q)\, T_q}{\max(A_{q,y}, 1)},$$

where T_q denotes the target number of positive studies for the corresponding partition. This weighting assigns greater priority to patients containing findings that are scarce within that partition. Ties are resolved using the number of distinct positive findings, the total number of positive studies, and the patient identifier.

To preserve patient diversity, at most four studies are selected from a single patient. Studies are greedily selected according to the number of remaining label deficits they reduce. Before DICOM quality control, the target is 250 positive studies per finding in the training partition and 63 positive studies per finding in each of the validation and test partitions. These targets include an approximately 25% surplus to account for studies that may subsequently be removed during image quality control.

After positive enrichment, negative studies are added with a target ratio of approximately two negative examples per selected positive example. When remaining negative-label deficits cannot be satisfied by patients already containing positive target findings, background patients with no positive studies across the 12 target findings are considered. These patients are prioritized according to the number and magnitude of unresolved negative deficits covered by their studies.

The resulting cohort is subject to a soft cap of 5,000 patients. Any target that cannot be satisfied because of limited case availability or the patient and study limits is recorded explicitly rather than silently replaced.

A.2 DICOM Quality-Control Procedure

Technical quality control is performed independently for each downloaded DICOM view. Raw pixel intensities are first decoded using the DICOM rescale transformation

$$I_{\text{rescaled}} = a I_{\text{stored}} + b,$$

where a and b correspond to the DICOM `RescaleSlope` and `RescaleIntercept`, respectively. For `MONOCHROME1` images, pixel polarity is reversed before quality metrics are computed in accordance with DICOM image semantics.

The rescaled image is min–max normalized,

$$\widetilde{I} = \frac{I_{\text{rescaled}} - \min(I_{\text{rescaled}})}{\max(I_{\text{rescaled}}) - \min(I_{\text{rescaled}})}.$$

If the maximum and minimum intensities are identical, the resulting constant image is mapped to zero to avoid division by zero.

Six technical indicators are then computed: mean intensity, intensity standard deviation, percentile-based contrast, intensity entropy, sharpness, and a high-frequency residual-noise estimate. Specifically,

$$\mu_I = \text{mean}(\widetilde{I}), \qquad \sigma_I = \text{std}(\widetilde{I}),$$

$$C_I = Q_{0.95}(\widetilde{I}) - Q_{0.05}(\widetilde{I}),$$

$$H_I = -\sum_{k=1}^{256} p_k \log_2 p_k,$$

and the sharpness proxy is computed from the variance of a discrete Laplacian,

$$S_I = \text{Var}\left(\nabla^2 \widetilde{I}\right).$$

A high-frequency residual measure is additionally used to identify unusually noisy images.

Each quality feature k is evaluated relative to the downloaded cohort using a robust center and scale,

$$\widetilde{\mu}_k = \text{median}(v_k),$$

$$\widetilde{\sigma}_k = 1.4826\,\text{median}\,|v_k - \widetilde{\mu}_k|.$$

If the median absolute deviation yields a zero scale, the implementation falls back successively to the interquartile range divided by 1.349, the standard deviation, and finally a unit scale.

Mean intensity is evaluated in both tails. Standard deviation, contrast, entropy, and sharpness are evaluated for unusually low values, whereas the noise indicator is evaluated for unusually high values. A view receives a technical warning when the maximum directional deviation score is at least 3 and fails quality control when the score is at least 6.

Corrupted DICOM files, empty pixel arrays, and images for which the required quality metrics cannot be computed fail automatically. Warning-level images are retained because an unusual technical measurement alone is not considered sufficient evidence that the image is unusable. Failed views are removed, but a multi-view study remains eligible if at least one view passes quality control. A study is discarded only when all available views fail.

A.3 Label-Specific Threshold Selection and Gray-Zone Routing

MEDAGENT-X does not apply a common decision threshold of 0.5 across all findings.

Checkpoint thresholds. During training, each checkpoint stores a label-specific threshold τ_y^{ckpt} which is selected on supervised validation instances using the candidate threshold grid.

Given the candidate threshold grid

$$\mathcal{G} = \{0.05, 0.06, \ldots, 0.95\},$$

the threshold for finding y is selected as

$$\tau_y = \arg\max_{\tau \in \mathcal{G}} F_1\left(\{t_{s,y} : m_{s,y} = 1\}, \{\mathbb{I}[p_{s,y} \geq \tau] : m_{s,y} = 1\}\right).$$

The grid is traversed in ascending order, and the stored threshold is updated only after a strict improvement in validation F1. Consequently, when multiple thresholds achieve the same maximum F1, the smallest such threshold is retained.

Validation policy derived thresholds. Each label is subjected to a validation-only threshold policy which may help to refine the initial checkpoint threshold. Candidate thresholds include a dense grid over $[0.01, 0.99]$ and the observed validation scores. The candidate grid contains equally spaced values with a step size 0.005 and observed validation scores are rounded to six decimal places before candidate thresholds are evaluated. A candidate threshold is used only if it improves validation F1 while satisfying all guardrails: the threshold may decrease by at most 0.10, precision may fall by at most 0.02, the positive prediction rate may increase by at most 25% and F1 score must not decrease. The validation policy does not change the model score itself. All disease labels are eligible for this process.

Let τ_y^{pol} denote an accepted policy threshold. The active threshold at inference is

$$\tau_y^* = \begin{cases} \tau_y^{\text{pol}}, & \text{if the validation policy supplies an override for } y, \\ \tau_y^{\text{ckpt}}, & \text{otherwise.} \end{cases}$$

The resulting vision-only decision is

$$\hat{z}_{s,y}^{\text{vis}} = \begin{cases} \text{present}, & p_{s,y} \geq \tau_y, \\ \text{absent}, & p_{s,y} < \tau_y. \end{cases}$$

This procedure determines a label-specific operating point but does not calibrate $p_{s,y}$ as a clinical probability. The sigmoid output is therefore treated as a model score rather than as a calibrated estimate of disease likelihood.

Gray zone routing. Retrieval fusion is allowed to modify a vision decision only when its score is within a certain range to the active threshold.

$$g_{s,y} = \mathbb{I}\left[|p_{s,y} - \tau_y| \leq \delta\right], \qquad \delta = 0.15.$$

The gray zone is a fixed routing policy rather than a learned abstention or reject mechanism. When $g_{s,y} = 0$, deterministic fusion must preserve the original vision-only decision,

$$g_{s,y} = 0 \Longrightarrow \hat{z}_{s,y}^{\text{det}} = \hat{z}_{s,y}^{\text{vis}}.$$

Thus, label-changing authority is restricted to threshold-adjacent predictions, while predictions outside the gray zone remain unchanged by retrieval-grounded intervention.

A.4 Image-to-Report Retrieval

The retrieval index contains exclusively studies from the training partition. For every indexed training study r, the same fine-tuned RAD-DINO checkpoint and multi-view mean–max pooling procedure used during inference are applied to produce the study representation h_r.

Each retrieval record stores the study identifier, de-identified patient identifier, study embedding, and associated radiology report text. The report document is constructed from the available *Findings* and *Impression* sections,

$$D_r = \text{“Findings: ”}\|F_r\|\text{“Impression: ”}\|M_r,$$

where empty sections are omitted. Studies with no usable report text are not included in the retrieval index. For implementation consistency, the stored report payload is truncated to a maximum of 12,000 characters.

Ground-truth label values are deliberately excluded from the retrieval payload. Consequently, downstream intervention receives report text from visually similar historical training cases rather than explicit training labels.

Cosine similarity is used to compare a target study s with an indexed training study r,

$$d_{\cos}(s, r) = 1 - \frac{h_s^\top h_r}{\|h_s\|_2 \|h_r\|_2}, \qquad \text{sim}(s, r) = 1 - d_{\cos}(s, r).$$

The final neighborhood size is fixed to $K = 10$. Because eligibility filtering is applied after the initial vector search, the index first returns an expanded candidate set,

$$K' = \max(8K, K + 20) = 80.$$

Candidates are ranked by increasing cosine distance. The query study itself, when applicable, and all studies associated with the same de-identified patient are removed. The first K remaining eligible studies form the retrieval set $\mathcal{R}_s$.

Restricting the retrieval index to the training partition prevents validation or test reports from entering the inference process. Same-patient exclusion provides an additional safeguard against treating longitudinal examinations of the query patient as independent supporting evidence.

At runtime, the dimensionality of each query embedding is checked against the stored collection representation. A mismatch causes retrieval to terminate rather than querying an index constructed using an incompatible checkpoint or pooling configuration.

A.5 Label-Specific Report Evidence Extraction

Retrieved reports may contain affirmed, negated, uncertain, historical, or context-dependent descriptions of radiographic findings. MEDAGENT-X therefore uses a fixed rule-based extractor at the deterministic fusion boundary so that the numerical evidence influencing a label transition remains directly inspectable and reproducible.

For each finding y, a predefined dictionary $\mathcal{K}_y$ contains the finding name and selected radiological synonyms. For example, the dictionary for atelectasis includes expressions such as *atelectasis* and *volume loss*, while the edema dictionary includes expressions such as *pulmonary edema*, *interstitial edema*, and *vascular congestion*. A matched report fragment contributes at most one count to a given finding, even when multiple terms in $\mathcal{K}_y$ are matched within that fragment.

For a matched term beginning at character position j, the extractor examines the preceding 70 characters for predefined negation cues,

$$\mathcal{N} = \{\text{no}, \text{without}, \text{no evidence of}, \text{negative for}, \text{absent}, \text{clear of}, \text{free of}\}.$$

A match preceded by a negation cue is counted as negative; otherwise it is counted as positive. Across the retrieved reports in $\mathcal{R}_s$, the resulting positive and negative evidence counts are denoted

$$c_{s,y}^{+} \qquad \text{and} \qquad c_{s,y}^{-}.$$

Representative positive and negative report fragments are retained together with their source study identifiers to provide traceable evidence for the downstream fusion and LLM-review stages. Retained evidence fragments are truncated to 260 characters for storage and audit reporting, while the complete retrieved reports remain available to the language-model review components.

This extractor is intentionally simpler than a learned or dependency-based radiology report labeler. Its purpose is not to independently diagnose the target image, but to provide a deterministic and reproducible evidence signal for the fusion policy. The complete retrieved reports provide additional context to the guarded review and evidence-verification stages.

A.6 Deterministic Fusion and Guarded LLM Review

Deterministic fusion first produces a candidate post-retrieval status $\hat{z}_{s,y}^{\mathrm{det}}$ according to the label-specific evidence rules defined in Section 2.4. The guarded LLM reviewer is invoked only when deterministic fusion proposes at least one change to the vision-only prediction. Let

$$C_s = \left\{ y : \hat{z}_{s,y}^{\mathrm{det}} \neq \hat{z}_{s,y}^{\mathrm{vis}} \right\}$$

denote the set of labels proposed for modification.

For each $y \in C_s$, the review request includes the label name, vision-only status, deterministic status, vision score $p_{s,y}$, decision threshold τ_y, positive and negative retrieval counts, and the deterministic rule trace. The reviewer also receives the ranked retrieved cases, including their study identifiers, similarity scores, and report text.

The prompt explicitly states that the retrieved reports belong to visually similar historical studies rather than to the target patient and therefore constitute indirect contextual evidence. The reviewer is not permitted to introduce a new diagnosis or modify labels outside C_s.

The structured review action is restricted to

$$\mathcal{A} = \{\mathrm{keep}, \mathrm{veto}, \mathrm{uncertain}\}.$$

A *keep* action preserves the deterministic fusion result, a *veto* restores the original vision-only status, and an *uncertain* action replaces the proposed deterministic status with an uncertain status.

The structured response additionally includes confidence, an evidence assessment, a rationale, and identifiers of supporting or contradicting retrieved cases. Evidence assessment is restricted to

$$\{\mathrm{supporting}, \mathrm{contradictory}, \mathrm{mixed}, \mathrm{insufficient}\}.$$

Programmatic validation is applied before any review action can affect the finalized prediction. Responses containing missing, duplicated, or unrequested labels are rejected, and every cited study identifier must correspond to a retrieved case. A *keep* action must preserve the deterministic status, whereas a *veto* must restore the original vision-only status.

Non-keep actions are subject to additional evidence constraints. Low-confidence non-keep actions are rejected. A valid veto requires contradictory evidence and at least one contradicting retrieved case. An *uncertain* action requires mixed, insufficient, or contradictory evidence and must satisfy the policy constraints associated with the proposed deterministic transition.

If parsing, schema validation, or policy validation fails, the system issues one structured repair request. Valid label-level responses from the repaired output are retained, while invalid or missing responses fall back to the deterministic result. If the repair attempt also fails, the deterministic status is preserved for all affected labels.

The finalized post-review status is denoted $\hat{z}_{s,y}^{\mathrm{fus}}$. The corresponding audit record preserves the vision-only, deterministic, and finalized statuses together with the requested LLM action, confidence, evidence assessment, validation outcome, fallback reason when applicable, cited case identifiers, and rationale.

A.7 Evidence Verification

A second LLM invocation performs read-only evidence verification after the finalized prediction has been established. Its purpose differs from the guarded review: the first LLM evaluates whether a proposed deterministic intervention should be retained, whereas the verifier evaluates how well the finalized prediction is supported by the available evidence.

The verification set is

$$V_s = \left\{ y : \hat{z}_{s,y}^{\mathrm{fus}} \in \{\mathrm{present}, \mathrm{uncertain}\} \right\}.$$

For every $y \in V_s$, the verifier receives the finalized status, vision-only status, model score, decision threshold, gray-zone indicator, whether deterministic fusion modified the prediction, the corresponding fusion rationale, and the ranked retrieval evidence available to the fusion and review stages.

The structured response includes a label-level evidence score from 1 to 5, confidence, vision-support level, retrieval-support level, contradiction level, evidence assessment, rationale, and identifiers of supporting and contradicting retrieved cases. The verifier additionally produces an overall study-level evidence score together with summaries of the visual, retrieval, and fusion evidence.

Programmatic validation requires exact agreement with the requested label set and rejects duplicate labels or references to unknown retrieved cases. A contradictory assessment must specify the degree of contradiction and cite at least one contradicting retrieved case.

Critically, the verifier has no authority to modify $\hat{z}_{s,y}^{\mathrm{fus}}$. Its output is retained only as an evidence and audit artifact. This separation prevents the evidence narrative from becoming an additional diagnostic action.

If no eligible label exists, the verifier is not invoked and an explicit no-eligible-label record is produced. Verification failures are likewise retained as structured audit events rather than being silently omitted.

A.8 LangGraph Orchestration and State Validation

MEDAGENT-X executes the Stage 2 inference pipeline using a typed LangGraph state graph LangChain (2024). LangGraph is used for orchestration, state transfer, and execution tracing; it does not define the prediction or label-modification policies.

For study s, the graph is initialized with a study identifier and the corresponding eligible DICOM paths. The shared inference state is progressively populated with the visual prediction output, retrieved cases, deterministic fusion and guarded-review metadata, finalized labels, evidence-verification output, and the final evidence record. Deterministic fusion results and finalized post-review labels are retained separately so that the original intervention proposal and the post-guardrail outcome remain independently traceable.

The graph follows the fixed execution path

$$\mathrm{START} \to V \to R \to F \to E \to \mathrm{END},$$

where V, R, F, and E denote the visual-prediction, retrieval, fusion-review, and evidence-verification nodes, respectively.

If $\Delta_k(S)$ denotes the partial state produced by node k, state propagation is expressed as

$$S_s^{(k+1)} = S_s^{(k)} \cup \Delta_k\left(S_s^{(k)}\right).$$

The visual node requires the study identifier and eligible image paths and produces the study representation, per-label scores, thresholds, and vision-only statuses. The retrieval node consumes the study representation and returns the ranked training-only retrieval set. Before retrieval, the query embedding dimensionality is checked against the stored collection representation to prevent use of an incompatible index. The fusion node consumes the visual prediction and retrieval evidence,

computes the deterministic intervention, and conditionally invokes the guarded LLM reviewer. The evidence node receives the finalized prediction and retrieved cases and invokes the read-only evidence verifier.

Each node verifies that all required upstream information is present before execution. Missing, inconsistent, or incompatible state therefore produces an explicit error at the responsible node rather than propagating incomplete information through subsequent stages.

The resulting state supports reconstruction at two levels. At the study level, the stored trajectory captures visual prediction, retrieval, fusion review, and evidence verification. At the individual-label level, the stored model score, decision threshold, gray-zone indicator, retrieval counts, deterministic rule trace, LLM action, cited cases, validation outcome, and final status make each intervention traceable.

A.9 Judge Evaluator Details

For each finding y, the Judge computes

$$\mathrm{Precision}_y = \frac{\mathrm{TP}_y}{\mathrm{TP}_y + \mathrm{FP}_y}, \qquad \mathrm{Recall}_y = \frac{\mathrm{TP}_y}{\mathrm{TP}_y + \mathrm{FN}_y + \mathrm{MU}_y},$$

$$F_{1,y} = \frac{2\,\mathrm{Precision}_y\,\mathrm{Recall}_y}{\mathrm{Precision}_y + \mathrm{Recall}_y}, \qquad \mathrm{Coverage}_y = \frac{\mathrm{TP}_y + \mathrm{TN}_y + \mathrm{FP}_y + \mathrm{FN}_y + \mathrm{MU}_y}{N_y}.$$

Here, MU_y denotes scoreable positive instances for which the finalized prediction is *uncertain*, and N_y is the total number of evaluated instances for finding y. Undefined ratios remain undefined rather than being replaced by zero.

Macro metrics average the defined per-label values equally, whereas micro metrics pool counts across labels before computing precision, recall, and F1. Reporting both is important in the imbalanced multi-label setting: macro metrics retain the influence of less frequent findings, whereas micro metrics are weighted toward findings with larger numbers of scoreable instances Davis & Goadrich (2006); Zhang et al. (2020).

Per-label uncertainty counts and complete ground-truth-versus-prediction status records are also retained.

B Experimental Setup and Implementation Details

B.1 RAD-DINO Training Configuration

The selected RAD-DINO checkpoint fine-tunes the final four transformer blocks while freezing the preceding encoder blocks. A dropout layer with probability 0.1 and a linear classification head are applied to the study-level representation.

Optimization uses AdamW with separate learning rates for the pretrained backbone and newly initialized classification head. The backbone learning rate is set to 10^{-5}, the classification-head learning rate to 10^{-3}, and weight decay to 10^{-4}.

A physical batch contains two studies. Gradients are accumulated over four physical batches before each optimizer update, resulting in an effective batch size of eight studies. The accumulated gradient is clipped to a maximum Euclidean norm of 1.0.

The learning rate is linearly warmed up during the first 5% of optimizer steps and subsequently decayed to zero using a cosine schedule. Automatic mixed precision is used on CUDA devices.

Training proceeds for at most 15 epochs using random seed 42. Early stopping is triggered after five consecutive validation epochs without improvement. The checkpoint with the highest validation masked macro-F1 is retained for all subsequent experiments.

B.2 Final Label-Specific Operating Points

The final inference thresholds follow the target-label ordering

$$\begin{aligned}&(\text{Atelectasis}, \text{Cardiomegaly}, \text{Consolidation}, \text{Edema},\\ &\text{Enlarged Cardiomediastinum}, \text{Fracture}, \text{Lung Lesion}, \text{Lung Opacity},\\ &\text{Pleural Effusion}, \text{Pleural Other}, \text{Pneumonia}, \text{Pneumothorax}).\end{aligned}$$

The corresponding validation-selected threshold vector is

$$\boldsymbol{\tau} = (0.56, 0.50, 0.45, 0.453857, 0.35, 0.42, 0.39, 0.62, 0.60, 0.45, 0.28, 0.33).$$

Each target finding therefore uses an independently selected operating threshold rather than a shared threshold of 0.5. The final thresholds are selected using the validation partition and frozen before evaluation on the test set. No test-set labels are used for threshold selection.

B.3 Runtime and Retrieval Configuration

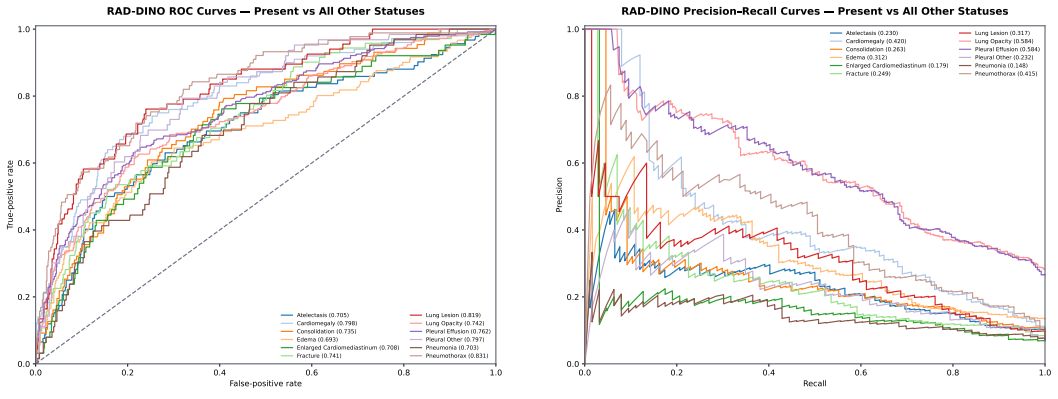
Vision inference is performed with a batch size of one study using automatic mixed precision and zero data-loading workers.

Retrieval-enabled configurations use the same training-only retrieval index described in Appendix A.4, with $K = 10$ nearest eligible studies retained for each query. The cohort, RAD-DINO checkpoint, label-specific operating thresholds, gray-zone margin, retrieval index, and test partition are held fixed across the evaluated configurations unless a component is explicitly removed by the corresponding ablation.

For language-model-enabled configurations, inference is executed locally through Ollama at temperature zero using structured outputs and the validation policies described in Appendix A.6. The primary MEDAGENT-X configuration uses Qwen3-14B, while the guarded-review ablation uses Llama 3.1 8B.

B.4 Vision-Only ROC and Precision–Recall Analysis

To complement the threshold-based evaluation in the main text, we analyze the continuous RAD-DINO scores before retrieval-grounded fusion or language-model review. Figure 2 reports receiver-operating-characteristic (ROC) and precision–recall (PR) curves for all 12 target findings.



(A) RAD-DINO AUROC curves.

(B) RAD-DINO precision-recall curves.

Figure 2: Continuous RAD-DINO discrimination across the 12 target findings. (A) Receiver-operating-characteristic curves. (B) Precision–recall curves. The plots evaluate vision-model scores before retrieval-grounded fusion and language-model review.

For this supplementary continuous-score analysis, *present* is treated as the positive class, while *absent*, *uncertain*, and *unmentioned* statuses are grouped into the negative class. This definition differs from the primary Judge evaluation, which excludes uncertain and unmentioned ground-truth labels from binary scoring, and the resulting AUROC and average-precision values should therefore be interpreted as a separate characterization of vision-model discrimination.

Across the 12 findings, the macro AUROC is 0.7528 and the macro average precision is 0.3276. Label-wise AUROC ranges from 0.6930 for edema to 0.8306 for pneumothorax, while average precision ranges from 0.1477 for pneumonia to 0.5842 for pleural effusion.

These curves characterize only the continuous RAD-DINO outputs. They do not represent the outputs of deterministic fusion or either LLM component, which operate on discrete label decisions downstream of the visual predictor.

C Additional Ablation Results

C.1 Ablation Configurations

We evaluate four configurations to examine the contributions of gray-zone routing, retrieved report evidence, deterministic fusion, and guarded language-model review. The test cohort, RAD-DINO checkpoint, label-specific thresholds, gray-zone margin, and Judge policy remain fixed across the configurations. Retrieval is enabled only in Experiments 3 and 4. Evidence verification is not included as a predictive ablation because it is strictly read-only and cannot alter the finalized study-label status. Table 5 summarizes the components enabled in each configuration.

Table 5: Components enabled in the four ablation configurations. Gray-zone abstention replaces each routed prediction with an *uncertain* status without using retrieved evidence.

Exp.	Retrieval	Label fusion	LLM review	Evidence verification
1	No	None	None	None
2	No	Gray $\rightarrow$ uncertain	None	None
3	Yes	Deterministic	None	None
4	Yes	Deterministic	Llama 3.1 8B	None

C.2 Detailed Ablation Results

C.2.1 Experiment 1: Vision-Only Baseline

The first configuration evaluates the fine-tuned RAD-DINO classifier without retrieval, deterministic fusion, guarded LLM review, or evidence verification. The selected checkpoint achieves a validation macro-F1 of 0.5957, macro AUROC of 0.8458, and macro average precision of 0.5935.

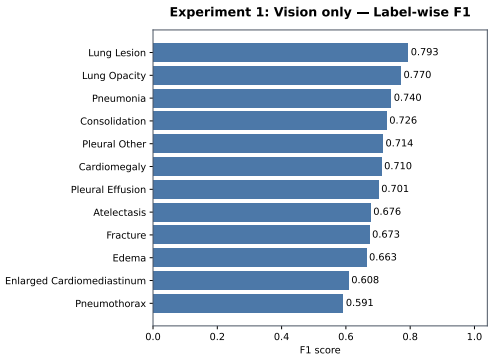
On the full test set, the vision-only model achieves macro precision of 0.9108, macro recall of 0.5695, and macro-F1 of 0.6971. The comparatively high precision and lower recall indicate that the baseline produces relatively precise positive predictions but misses a substantial fraction of positive findings.

Performance is lower within the fixed gray-zone subset. Macro-F1 decreases to 0.6310 and macro recall to 0.5090, while macro precision is 0.8748. This reduction is consistent with the gray zone containing threshold-adjacent predictions that are more difficult for the vision model than the full test set.

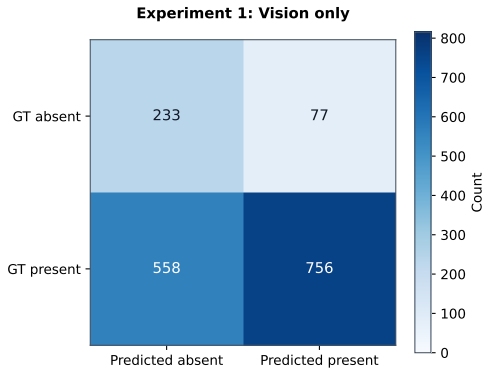
Figure 3 (A) reports the corresponding label-wise vision-only F1 scores, while Figure 3 (B) shows the vision-only confusion matrix.

C.2.2 Experiment 2: Gray-Zone Abstention Without Retrieval

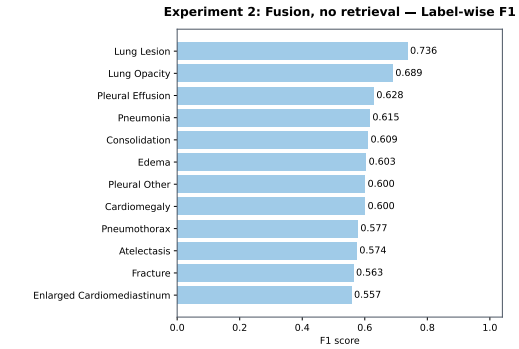
The second configuration evaluates whether uncertainty routing alone is sufficient to improve predictive performance. All 1,812 study-label predictions within the frozen gray-zone mask are re-



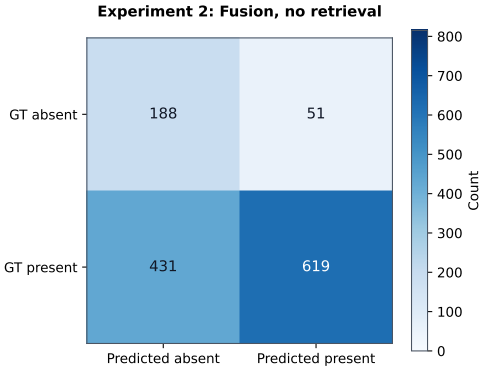
(A) Vision-only RAD-DINO baseline.



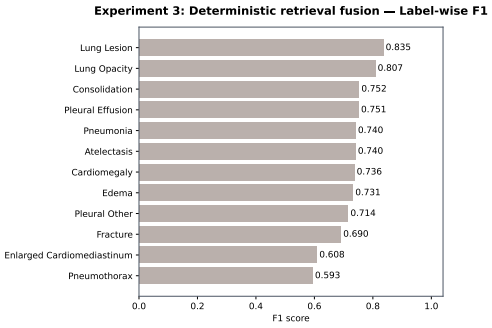
(B) Vision-only RAD-DINO baseline.



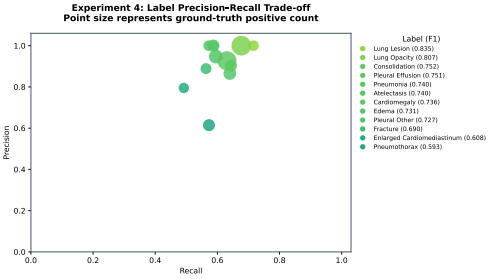
(C) Gray-zone abstention without retrieval.



(D) Gray-zone abstention without retrieval.



(E) Deterministic retrieval-grounded fusion.



(F) Guarded LLM fusion review.

Figure 3: Detailed results across the four ablation configurations. (A) Label-wise F1 scores for the vision-only RAD-DINO baseline. (B) Confusion matrix for the vision-only baseline. (C) Label-wise F1 scores after gray-zone abstention without retrieval. (D) Confusion matrix for the remaining binary predictions after gray-zone abstention. (E) Label-wise F1 scores after deterministic retrieval-grounded fusion. (F) Per-label precision–recall behavior after guarded LLM fusion review.

placed with an *uncertain* status, while the remaining 9,060 study-label decisions retain their original vision-only status. No retrieval, deterministic evidence fusion, or language-model review is used.

On the full test set, this strategy increases macro precision from 0.9108 to 0.9241 but reduces macro recall from 0.5695 to 0.4619. Consequently, macro-F1 decreases from 0.6971 to 0.6126. The increase in precision therefore comes at a substantial recall cost, showing that uncertainty routing alone is too conservative to improve overall predictive performance.

Within the gray-zone subset, all routed predictions are assigned an *uncertain* status. Under the Judge policy, an uncertain prediction for a scoreable positive finding is counted as a missed positive rather

than as a cost-free abstention. As a result, gray-zone recall falls to zero, while binary precision and F1 are undefined because no routed gray-zone prediction remains explicitly *present* or *absent*.

Figure 3 (C) shows the label-wise F1 behavior after gray-zone abstention. Figure 3 (D) summarizes the confusion matrix for the remaining binary predictions after the routed gray-zone cases have been assigned an *uncertain* status.

C.2.3 Experiment 3: Deterministic Retrieval-Grounded Fusion

The third configuration adds image-to-report retrieval and deterministic gray-zone fusion without language-model involvement. Deterministic fusion modifies 390 of the 1,812 routed study-label decisions, including 374 absent-to-present promotions and 16 demotions from a vision-positive status.

On the full test set, macro recall increases from 0.5695 for the vision-only baseline to 0.6051, while macro precision remains comparable at 0.9117. Macro-F1 consequently improves from 0.6971 to 0.7246, corresponding to an absolute increase of 0.0275.

The improvement is concentrated within the threshold-adjacent region for which intervention is permitted. On the fixed gray-zone subset, macro-F1 increases from 0.6310 to 0.7613, while macro recall rises from 0.5090 to 0.6950. Macro precision also increases slightly, from 0.8748 to 0.8773.

The largest beneficial promotion effects are observed for pleural effusion, lung opacity, edema, and atelectasis, which gain 16, 13, 10, and 7 true positives, respectively, with at most one additional false positive per finding. Pneumothorax is the only reported adverse case: deterministic fusion removes three false positives but also loses one true positive.

Figure 3 (E) reports the corresponding label-wise F1 scores after deterministic retrieval-grounded fusion.

C.2.4 Experiment 4: Guarded LLM Fusion Review

The fourth configuration adds the guarded LLM reviewer after deterministic retrieval-grounded fusion while leaving evidence verification disabled. Llama 3.1 8B is used as the reviewer in this ablation to examine the behavior of the guarded-review mechanism with a smaller language model.

The reviewer is invoked only for study-label decisions modified by deterministic fusion. In this run, review was requested for 264 studies; 261 LLM calls completed successfully and produced 372 label-level reviews. The resulting configuration achieves a full-test macro-F1 of 0.7257, compared with 0.7246 for deterministic retrieval fusion alone.

This small numerical difference should be interpreted cautiously because the upstream vision and retrieval path was rerun for this configuration, producing 392 deterministic changes compared with 390 in Experiment 3. The comparison therefore primarily demonstrates that guarded LLM review preserves the predictive behavior of the deterministic fusion pipeline rather than establishing an independent performance gain attributable to the reviewer.

This behavior is consistent with the deliberately constrained role of the LLM: it reviews proposed deterministic interventions rather than acting as an independent diagnostic classifier. Unsupported, malformed, or policy-inconsistent review actions are rejected according to the guardrails described in Appendix A.6.

Figure 3 (F) shows the corresponding per-label precision–recall behavior after guarded LLM review. Most findings retain high precision with more moderate recall, while pneumothorax and enlarged cardiomediastinum exhibit weaker F1 performance relative to the other target findings.

Overall, the ablation results show that gray-zone routing alone is not sufficient to improve predictive performance. Replacing threshold-adjacent predictions with an *uncertain* status increases precision but causes a substantial loss in recall. In contrast, training-only retrieval followed by deterministic fusion provides the main improvement in macro-F1 by recovering additional positive findings within the region where intervention is permitted.

Guarded LLM review produces only a small change in the reported predictive metrics in this run, consistent with its intentionally constrained role as a review mechanism rather than an additional diagnostic classifier. Together, the ablations indicate that the primary predictive contribution

Table 6: Per-condition F1 on the external five-condition expert-labeled test cohort.

Condition	GLoRIA	MEDAGENT-X Vision	MEDAGENT-X Fusion	Δ vs. GLoRIA
Atelectasis	0.1699	0.4068	0.4878	+0.3179
Cardiomegaly	0.3963	0.4789	0.5144	+0.1181
Consolidation	0.1188	0.1719	0.1700	+0.0512
Edema	0.4171	0.4128	0.4114	-0.0056
Pleural Effusion	0.3497	0.3810	0.4063	+0.0566

of MEDAGENT-X arises from retrieval-grounded deterministic intervention, while the language-model reviewer provides an additional controlled review layer.

D External Expert-Labeled Evaluation

D.1 Per-Condition GLoRIA Comparison

The external comparison uses the five-condition CheXpert competition cohort described in Section 3.5, consisting of 200 validation studies and 500 expert-labeled test studies. The evaluated findings are atelectasis, cardiomegaly, consolidation, edema, and pleural effusion.

GLoRIA is evaluated zero-shot using its public ResNet-50 checkpoint and fixed condition prompts. MEDAGENT-X uses the supervised RAD-DINO checkpoint from the primary experiment together with the training-only retrieval index and deterministic label-fusion module. Because the two approaches use different training paradigms, this analysis is interpreted as a system-level comparison under a shared evaluation protocol rather than as a controlled architecture ablation.

Table 6 reports the per-condition F1 results for the external expert-labeled test cohort. Retrieval-grounded fusion exceeds the evaluated GLoRIA configuration on four of the five findings.

Retrieval-grounded fusion exceeds the evaluated GLoRIA configuration on four of the five findings. The largest difference occurs for atelectasis, for which F1 increases from 0.1699 with GLoRIA to 0.4878 with MEDAGENT-X fusion. Cardiomegaly and pleural effusion also show higher F1 under MEDAGENT-X fusion, while the difference for consolidation is comparatively small. GLoRIA achieves slightly higher F1 for edema, with 0.4171 compared with 0.4114 for MEDAGENT-X fusion. These results indicate heterogeneous condition-level behavior despite the aggregate advantage reported in Section 3.5.