RadAgents: Multimodal Agentic Reasoning for Chest X-ray Interpretation with Radiologist-like Workflows

Kai Zhang* KAZ321@LEHIGH.EDU
Oracle Health AI & Lehigh University

Corey D Barrett COREY.BARRETT@ORACLE.COM
Jangwon Kim JANGWON.KIM@ORACLE.COM
Oracle Health AI

Lichao Sun LIS221@LEHIGH.EDU
Lehigh University

Tara Taghavi TARA.TAGHAVI@ORACLE.COM
Krishnaram Kenthapadi KRISHNARAM.KENTHAPADI@ORACLE.COM
Oracle Health AI

Abstract

Agentic systems offer a potential path to solve complex clinical tasks through collaboration among specialized agents, augmented by tool use and external knowledge bases. Nevertheless, for chest X-ray (CXR) interpretation, prevailing methods remain limited: (i) reasoning is frequently neither clinically interpretable nor aligned with guidelines, reflecting mere aggregation of tool outputs; (ii) multimodal evidence is insufficiently fused, yielding text-only rationales that are not visually grounded; and (iii) systems rarely detect or resolve cross-tool inconsistencies and provide no principled verification mechanisms. To bridge the above gaps, we present **RadAgents**, a multi-agent framework that couples clinical priors with task-aware multimodal reasoning and encodes a radiologist-style workflow into a modular, auditable pipeline. In addition, we integrate grounding and multimodal retrieval-augmentation to verify and resolve context conflicts, resulting in outputs that are more reliable, transparent, and consistent with clinical practice.

Keywords: Multi-agent system, multimodal reasoning, chest X-ray, image interpretation.

Data and Code Availability We use the following public datasets: MIMIC-CXR-JPG (Johnson et al., 2019), MS-CXR (Boecking et al., 2022), and MS-CXR-T (Bannur et al., 2023), which are accessible under their respective data use agreements. We plan to release the code after obtaining organizational approval.

Institutional Review Board (IRB) This work does not require IRB approval.

1. Introduction

Chest X-ray (CXR) imaging is a cornerstone of pulmonary screening, diagnosis, and follow-up, accounting for the largest share of diagnostic radiology examinations performed worldwide each year (Cid et al., 2024). Yet systematic assessment of thoracic structures remains labor-intensive, imposing a substantial time burden on radiologists (Fallahpour et al., 2025). The gradual introduction of AI into clinical practice shows promise for alleviating this workload (Zhang et al., 2024; Tanno et al., 2025). However, prevailing systems fall short on complex multimodal reasoning, such as integrating findings across disparate image regions, which is central to radiologists' practice; most adhere to end-to-end designs in which the visual encoder executes a *single, front-end pass* and subsequent reasoning proceeds *purely in text* (Wang et al., 2025). This encode-once, text-only paradigm decouples the reasoning trajectory from evolving visual evidence, leading to failures on tasks that require iterative re-inspection, precise measurements, and cross-comparisons (Liu et al., 2025).

A promising path for clinical reasoning is to *augment* large language models, including multimodal

* Work done as an intern at Oracle Health AI.

variants, with *external tools* (Lu et al., 2025). By delegating perceptual and classification subtasks such as organ or region segmentation and disease classification to validated modules, the language model can focus on planning and synthesis. Several agentic frameworks already explore this idea, ranging from training small models for limited tool use (Li et al., 2024; Nath et al., 2025) to pipeline systems that invoke general-purpose models for more flexible operations (Jiang et al., 2025; Schmidgall et al., 2024), although multiple agent coordination and/or debate introduces considerable computational overhead. In CXR interpretation, RadFabric (Chen et al., 2025) integrates diagnostic agents with a separate reasoning agent, and MedRAX (Fallahpour et al., 2025) expands task coverage by incorporating additional task specific models. Despite gains over single-model baselines, integration and reasoning steps are often opaque and not aligned with clinical workflow, which undermines trust and creates safety risks, and they still lack explicit verification and conflict resolution.

To address these limitations, we introduce **RadAgents**, a framework designed for complex multimodal reasoning in CXR, which has three primary technical contributions:

- We introduce *clinically aware interleaved reasoning*, a *multimodal* loop that combines clinical priors and study metadata with targeted image operations (segment, measure, localize, compare), so each hypothesis triggers tool use and returns inspectable evidence, mirroring how radiologists proceed from observation to measurement, comparison with prior studies, and conclusion (Fig. 6 and Appendix E).

- We propose a *training-free* multi-agent design, where five sub-agents operate in clean, task-scoped context compartments, coordinated by an *Orchestrator* and reconciled by a *Synthesizer*. This preserves visual grounding, supports parallel execution, and composes measurement, localization, characterization, comparison, and diagnosis (Figure 1).

- We add a lightweight context verifier and *visual retrieval-augmented generation* to detect, surface, and resolve inconsistencies before reporting, yielding more clinically aligned outputs.

- **RadAgents** achieves state-of-the-art performance, outperforming strong baselines by 10.2% on MS-CXR, 29.6% on MIMIC-CXR, and 21.5% on MS-CXR-T. Multimodal retrieval further contributes an average 8.0% boost over the no-retrieval variant, mitigating context conflicts and improving reliability.

2. Methods

RadAgents is a multi agent system with seven specialized agents (Figure 1). Five implement the clinical **ABCDE** review scheme (Hodler et al., 2019): ***A**irway, **B**reathing, **C**irculation, **D**iaphragm*, and ***E**verything else*. In addition, an *Orchestrator* agent analyzes each query and routes tasks to the appropriate specialists with the required patient context (for example, imaging view and prior studies), and a *Synthesizer* agent integrates their outputs, resolves conflicts, and produces the final output.

This design confines context to task specific compartments, reducing the information each agent must process and simplifying context compression by having each sub-agent produce an initial summary for downstream synthesis. It also allows parallel execution, lowering latency for long reasoning. For common and clinically significant CXR findings such as cardiomegaly and pleural effusion, we curate radiologist-like workflows, the predefined templates within RadAgents, to guide tool selection and clinically grounded reasoning (see the demonstration example in Appendix A). For out-of-template queries, the system invokes workflow-free reasoning, preserving flexibility. The design is extensible: new templates (e.g., reasoning or tool-chains) can be added, and some can generalize to tasks of similar scope or category.

2.1. Task-aware Sub-Agents

Each sub-agent, also called the ABCDE agent, has a defined purpose and domain of expertise. Each is governed by a custom system prompt derived from clinical guidelines and maintains its own context window (See details). The main scope and objectives of them are:

Airway agent: Systematically assess the central thorax for airway patency, alignment, and paratracheal lesions; for example, determine tracheal position (midline versus deviation).

Breathing agent: Survey the lungs and pleura for parenchymal and pleural pathology; for example, de-

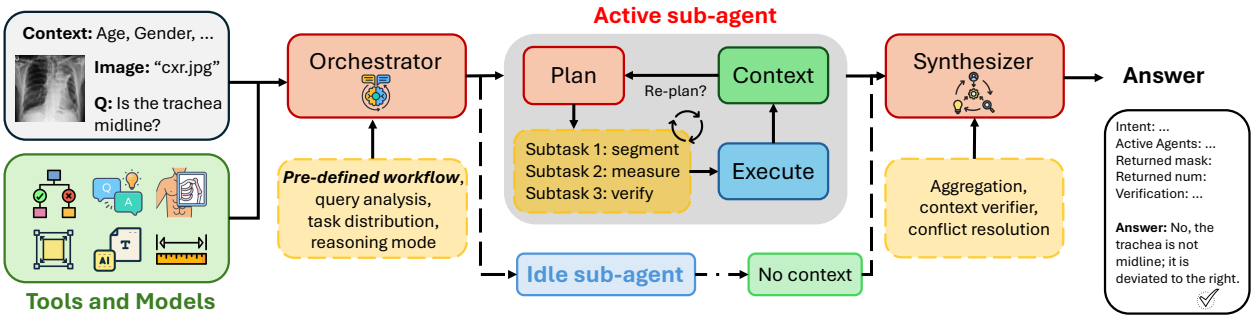


Figure 1: RadAgents framework. Each ABCDE subagent executes in parallel guided by clinical workflows, lowering latency, preserving isolation to avoid long-context drift, and improving trustworthiness.

tect opacities (alveolar, interstitial, nodular), infiltrates, and consolidation patterns.

Circulation agent: Evaluate the cardiac silhouette, mediastinum, and vessels; for example, compute the cardiothoracic ratio.

Diaphragm agent: Assess diaphragmatic integrity and look for subdiaphragmatic air; for example, compare right and left diaphragm height.

Everything Else agent: Identify other chest wall, soft tissue, and foreign body findings.

2.2. Radiologist-like Workflows

Solving different radiological tasks requires distinct reasoning modes and corresponding evidence. Figure 6 shows that multimodal reasoning, particularly interleaved quantitative measurements, strengthens reasoning for enlarged heart detection, whereas a simple grounding guided approach fails. Motivated by this, RadAgents curates clinically driven subtasks and workflow templates within the subagents' system prompts, organized into five modes: (M1) measurement, (M2) localization, (M3) characterization, (M4) relational and comparative reasoning, and (M5) diagnosis. Complex clinical reasoning composes these modes and triggers the appropriate tool calls within the agentic system; for example, judging progression of effusion volume uses M1 and M4. A detailed description of workflows can be found in Appendix E.

2.3. Global Controller Module

The global controller comprises the *Orchestrator* and the *Synthesizer*. The Orchestrator selects subagents and allocates tasks with appropriate patient context, and the Synthesizer integrates their outputs, verifies consistency, and resolves errors and conflicts. The major components are detailed below.

Query analysis. Given a query, the *Orchestrator* drafts a high level plan, selects the relevant subagents, and chooses the reasoning mode: ReAct when no workflow is specified (Yao et al., 2023), or Plan-and-Execute (P&E) (WANG et al., 2023) when a workflow template is available. This keeps the system language driven and adaptable across queries.

Tools. We employ a suite of models as tools for distinct CXR tasks: CheXagent (Chen et al., 2024b) for VQA, MAIRA-2 (Bannur et al., 2024) for grounding, the CheXpert Plus report generator (Chambon et al., 2024), and classification and organ segmentation models from TorchXRayVision (Cohen et al., 2022). In addition, we include unique programming tools that return zoomed-in quarter patches or serve for measurement and calculation purposes.

Context verifier. No tool is perfect, as their capabilities are constrained by model size and training data. When uncertainty arises, we trigger a verification step in which an advanced multimodal LLM serves as a judge (Chen et al., 2024a), filtering out incorrect outputs such as erroneous masks.

Retrieval-augmented conflic resolution. Tool outputs can conflict. On the *Synthesizer* side, we apply Visual Retrieval-Augmented Generation (V-RAG) (Chu et al., 2025): the agent retrieves clinically similar chest radiographs (based on image embeddings from Rad-DINO (Perez-Garcia et al., 2025)) and accompanying context like patient notes and uses these exemplars to adjudicate discrepancies among tools (Figure 2 and Appendix F). This mirrors routine radiologic practice, in which clinicians consult similar cases and content to calibrate interpretation.

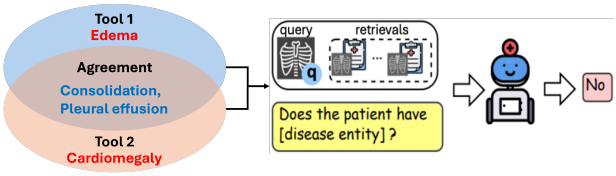


Figure 2: Resolving the conflicts via V-RAG (Chu et al., 2025).

3. Experiments

3.1. Experimental Setup

To demonstrate the generality of **RadAgents**, we evaluate it on three tasks with increasing reasoning complexity: VQA for existence and attributes (E&A), VQA for comparison and progression (C&P), and report generation. The data statistic and details are shown in Table 2.

Baselines. We instantiate all agents with GPT-4o. For comparison, we include (1) GPT-4o (Hurst et al., 2024), (2) GPT-4o with ReAct, where tools are available via function calling but without explicit workflow steering, (3) GPT-4o with monolithic Workflow where a MLLM executing ABCDE analysis end-to-end with the same tools, retrieval, and budges as RadAgents, and (4) two medical specialist models, CheXagent and MedGemma (Sellergren et al., 2025). Unless otherwise noted, the number of retrieved exemplars for V-RAG is set to $k = 3$ (see Appendix F for an ablation on k). We report results for two variants of RadAgents, with and without V-RAG.

Metrics. For VQA (E&A) and report generation, we use standard CXR text metrics (explanation in Appendix B): RadGraph F1 (Jain et al., 2021), CheXbert macro F1 across 14 labels (Smit et al., 2020), RaTE (Zhao et al., 2024), and GREEN (Ostmeier et al., 2024). Because the outputs are sentences, these metrics capture clinical correctness and consistency. For VQA (C&P), where the output is one of three choices, we report accuracy.

3.2. Existence and Attributes

The VQA questions cover seven common findings in CXR: atelectasis, cardiomegaly, consolidation, edema, lung opacity, pleural effusion, and pneumothorax, derived from the standard test split of MS–CXR. Each image receives the following prompt (details are stated in the Appendix D):

```
<image> Describe if [finding] is present; if
present, describe [attributes].
```

Table 1 shows that adding V-RAG to the agent improves all metrics over the ablation: +0.0298 CheXbert, +0.0389 RadGraph, +0.0611 RaTE, +0.0191 GREEN, raising *Avg.* from 0.4260 to 0.4632 (+0.0372, +8.7%). Among baselines, MedGemma is strongest (0.4205 *Avg.*) yet remains 10.2% below RadAgents. GPT-4o benefits from ReAct (+0.1023 *Avg.*) but still trails the full agent by 0.1010.

3.3. Comparison and Progression

We use MS–CXR–T to assess stability, improvement, or worsening of a specific positive finding (consolidation, edema, pleural effusion, or pneumothorax). We only retain cases where the metadata indicates a **consensus** among human reviewers. We pose a comparative question that explicitly references the prior study. The prompt template is:

```
Given current image <image>, and previous
image <image>, decide if [finding] is
improving, stable, or worsening.
```

Figure 3 shows that RadAgents+V-RAG achieves the best *overall accuracy* on MS-CXR-T, surpassing the ablated agent and all LLM baselines. The ordering mirrors E&A, underscoring the value of retrieval over similar studies for longitudinal reasoning.

3.4. Report Generation

We construct a MIMIC–CXR subset aligned with MS–CXR identities so that findings queried in VQA are represented in the corresponding reports. Prompts request generation of the *Findings* section, and all agents are activated by default. The prompt combines the template from Section 3.2 with a curated list of clinically significant findings, following prior work (Tu et al., 2024; Peng et al., 2025); details appear in Appendix D.

Table 1 shows that, RadAgents attains the best report quality (*Avg.* 0.4182). V-RAG contributes +0.0335 *Avg.* (+8.7%), with the largest gain on **GREEN** (+0.0706), plus lifts on CheXbert (+0.0315) and RaTE (+0.0406). GPT-4o+ReAct improves over GPT-4o (+0.0393) but remains 0.0956 below RadAgents, while MedGemma drops on this task (0.2686 *Avg.*).

Table 1: Evaluation on VQA (E&A) and report generation with baselines and RadAgents. The four metrics are commensurate (each normalized to the range $[0, 1]$).

Method	CheXbert-macro-F1(14)	RadGraph-F1	RaTE	GREEN	Avg.
VQA (E&A) on MS-CXR					
CheXagent	0.3321	0.1817	0.4526	0.3429	0.3273
MedGemma	0.3827	0.1624	0.5648	0.5723	0.4205
GPT-4o	0.3219	0.0928	0.4122	0.2127	0.2599
GPT-4o w/ ReAct	0.3613	0.1221	0.5034	0.4619	0.3622
GPT-4o w/ Workflow	0.4058	0.1617	0.5498	0.5351	0.4130
RadAgents wo/ V-RAG	0.4128	0.1925	0.5147	0.5841	0.4260
RadAgents	0.4426	0.2314	0.5758	0.6032	**0.4632**
Report Generation on MIMIC-CXR					
CheXagent	0.2916	0.1318	0.4129	0.1825	0.2547
MedGemma	0.2413	0.1189	0.4728	0.2416	0.2686
GPT-4o	0.2237	0.1324	0.4635	0.3138	0.2833
GPT-4o w/ ReAct	0.3521	0.1329	0.4731	0.3325	0.3226
GPT-4o w/ Workflow	0.4080	0.1556	0.5187	0.3841	0.3653
RadAgents wo/ V-RAG	0.4412	0.1826	0.5238	0.3821	0.3824
RadAgents	0.4727	0.1829	0.5644	0.4527	**0.4182**

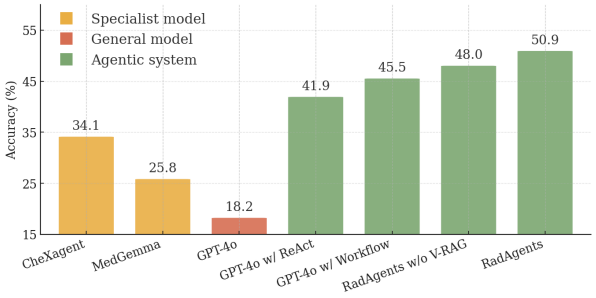


Figure 3: VQA results regrading progression.

3.5. Effectiveness of RadAgents Design

We evaluate the multi-agent design against a monolithic baseline in which a single LLM follows the ABCDE scheme (*GPT-4o w/ Workflow*), thereby separating improvements attributable to multi-agent coordination from those due to a structured workflow. Consistent with Table 1 and Figure 3, (i) introducing a workflow confers substantial gains over ReAct, and (ii) the multi-agent architecture further improves performance. These effects stem from contextual isolation within the workflow, which limits long-context drift in extended reasoning chains.

In addition, we validate the context verifier and conflict-resolution modules. On our 1,147-case dataset, the context verifier was triggered in 37.67% of instances. We observed a tool-conflict rate of 32.78%, of which 78.99% were resolved correctly.

4. Discussion

In this work, we propose the first radiologist-like agentic system, which demonstrates superior performance in CXR interpretation. Further improvements are required for broader application: (1) the context verifier can only detect errors but cannot modify outputs, especially for visual evidence such as bounding boxes, so stronger supporting models are needed; (2) current experiments are limited to frontal-view images, as most existing tools are incompatible with lateral views; (3) prompting strategies could be further optimized; (4) the framework could be extended to additional modalities, such as CT and MRI; and (5) our current implementation relies on large proprietary models and multiple tools, making it costly and difficult to reproduce; exploring smaller, open-source models is a valuable direction.

Acknowledgments

We would like to thank other members of Oracle Health AI for their support while developing our system, and Raefer Gabriel, Sri Gadde, Mark Johnson, Devashish Khatwani, Yuan-Fang Li, Anit Sahu, Praphul Singh, and Vishal Vishnoi for insightful feedback and discussions.

References

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Perez-Garcia, Maximilian Ilse, Daniel C Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, et al. Learning to exploit temporal structure for biomedical vision-language processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15016–15027, 2023.

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH 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.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024a.

Wenting Chen, Yi Dong, Zhaojun Ding, Yucheng Shi, Yifan Zhou, Fang Zeng, Yijun Luo, Tianyu Lin, Yihang Su, Yichen Wu, et al. Radfabric: Agentic ai system with reasoning capability for radiology. *arXiv preprint arXiv:2506.14142*, 2025.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, et al. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024b.

Yun-Wei Chu, Kai Zhang, Christopher Malon, and Martin Renqiang Min. Reducing hallucinations of medical multimodal large language models with visual retrieval-augmented generation. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*, 2025.

Yashin Dicente Cid, Matthew Macpherson, Louise Gervais-Andre, Yuanyi Zhu, Giuseppe Franco, Ruggiero Santeramo, Chee Lim, Ian Selby, Keerthini Muthuswamy, Ashik Amlani, et al. Development and validation of open-source deep neural networks for comprehensive chest x-ray reading: a retrospective, multicentre study. *The Lancet Digital Health*, 6(1):e44–e57, 2024.

Joseph Paul Cohen, Joseph D Viviano, Paul Bertin, Paul Morrison, Parsa Torabian, Matteo Guarrera, Matthew P Lungren, Akshay Chaudhari, Rupert Brooks, Mohammad Hashir, et al. Torchxrayvision: A library of chest x-ray datasets and models. In *International Conference on Medical Imaging with Deep Learning*, pages 231–249. PMLR, 2022.

Adibvafa Fallahpour, Jun Ma, Alif Munim, Hongwei Lyu, and BO WANG. Medrax: Medical reasoning agent for chest x-ray. In *Forty-second International Conference on Machine Learning*, 2025.

Juerg Hodler, Rahel A Kubik-Huch, and Gustav K von Schulthess. *Diseases of the chest, breast, heart and vessels 2019-2022: diagnostic and interventional imaging*. Springer Nature, 2019.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, et al. Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Yixing Jiang, Kameron C Black, Gloria Geng, Danny Park, James Zou, Andrew Y Ng, and Jonathan H

Chen. Medagentbench: a virtual ehr environment to benchmark medical llm agents. *NEJM AI*, 2(9): AIdbp2500144, 2025.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019.

Binxu Li, Tiankai Yan, Yuanting Pan, Jie Luo, Ruiyang Ji, Jiayuan Ding, Zhe Xu, Shilong Liu, Haoyu Dong, Zihao Lin, et al. Mmedagent: Learning to use medical tools with multi-modal agent. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8745–8760, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Chengzhi Liu, Zhongxing Xu, Qingyue Wei, Juncheng Wu, James Zou, Xin Eric Wang, Yuyin Zhou, and Sheng Liu. More thinking, less seeing? assessing amplified hallucination in multimodal reasoning models. *arXiv preprint arXiv:2505.21523*, 2025.

Pan Lu, Bowen Chen, Sheng Liu, Rahul Thapa, Joseph Boen, and James Zou. Octotools: An agentic framework with extensible tools for complex reasoning. *arXiv preprint arXiv:2502.11271*, 2025.

Yu A Malkov and Dmitry A Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018.

Vishwesh Nath, Wenqi Li, Dong Yang, Andriy Myronenko, Mingxin Zheng, Yao Lu, Zhijian Liu, Hongxu Yin, Yee Man Law, Yucheng Tang, et al. Vila-m3: Enhancing vision-language models with medical expert knowledge. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14788–14798, 2025.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Md, Michael Moseley, Curtis Langlotz, Akshay Chaudhari, et al. Green: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Cheng Peng, Kai Zhang, Mengxian Lyu, Hongfang Liu, Lichao Sun, and Yonghui Wu. Scaling up biomedical vision-language models: Fine-tuning, instruction tuning, and multi-modal learning. *arXiv preprint arXiv:2505.17436*, 2025.

Fernando Perez-Garcia, 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 (1):119–130, 2025.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *arXiv preprint arXiv:2405.07960*, 2024.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, et al. Medgemma technical report. *arXiv preprint arXiv:2507.05201*, 2025.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519, 2020.

Ryutaro Tanno, David GT Barrett, Andrew Sellergren, Sumedh Ghaisas, Sumanth Dathathri, Abigail See, Johannes Welbl, Charles Lau, Tao Tu, Shekoofeh Azizi, et al. Collaboration between clinicians and vision–language models in radiology report generation. *Nature Medicine*, 31(2):599–608, 2025.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira

Ktena, et al. Towards generalist biomedical ai. *NEJM AI*, 2024.

Lei WANG, Wanyu XU, Yihuai LAN, Zhiqiang HU, Yunshi LAN, and Roy Ka-Wei LEE. Lim, ee-peng. plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models.(2023). In *61st Annual Meeting of the Association for Computational Linguistics, ACL*, pages 9–14, 2023.

Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D Davison, Hui Ren, et al. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, 2024.

Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15004–15019, 2024.

Appendix A. Demonstrations

A.1. Full reasoning trajectory for tracheal deviation detection

Figure 5 shows the reasoning trajectory for the query "Is the trachea midline?" The sequence proceeds as follows: (a) the orchestrator agent analyzes the query, activates the airway agent, and initiates actions; (b) the returned tool context is visualized; (c) the synthesizer agent integrates the context but issues a warning due to low-confidence segmentation results, triggering the context verifier—specifically, GPT-4o is invoked to re-judge and answer the query; and (d) synthesis is completed, producing the final answer with an associated confidence score.

A.2. Failure case in CTR calculation

Most measurement reasoning relies on segmentation or grounding tools for support. For example, measuring heart width directly involves counting the number of pixels between boundary points. However, due to the high cost of annotation, ground-truth datasets for training robust segmentation or grounding models are limited. Moreover, most organ segmentation datasets are curated from normal images; when abnormalities obscure the organs, segmentation often fails, leading to downstream measurement errors. Figure 4(b) illustrates the segmentation masks for a normal CXR and a CXR with effusion. In the effusion case, both heart width and thoracic width are measured incorrectly due to inappropriate masks. These results further emphasize the need for a context verifier and resolver.

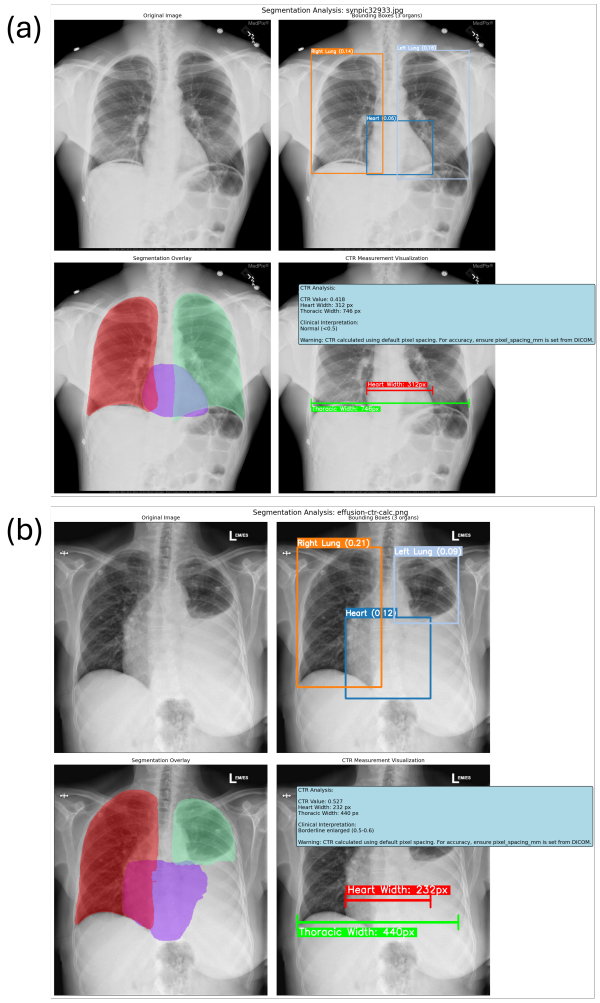
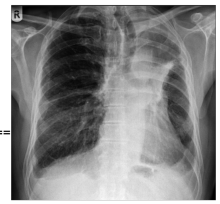


Figure 4: CTR calculation with segmentation masks on (a) normal and (b) effusion cases. Inaccurate masks in the abnormal case lead to incorrect heart and thoracic width measurements.

Appendix B. Clinical Evaluation Metrics

Evaluating the quality of generated radiology reports is non-trivial. Early works adopted general-domain natural language processing metrics such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002). While these metrics are widely used for text evaluation, they treat differences in wording the same as clinically significant errors, failing to reflect medical accuracy. To address this limitation, clinically informed evaluation metrics, such as CheXbert (Smit et al., 2020), RadGraph (Jain et al., 2021), GREEN (Ostmeier et al., 2024), and RaTEScore (Zhao et al., 2024), have been proposed to better assess clinical

(a)

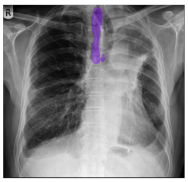


```
============================================================
ORCHESTRATOR: Analyzing query...
Query: Is the trachea midline?

Query Analysis (1.27s):
  Intent: localization
  Active Agents: ['airway']
  Analysis Focus: ['Tracheal position']
============================================================

ORCHESTRATOR: Executing agents...
Executing 1 agents sequentially...
Running airway...
  Calling segmentation with params: {'image_path':
'./demo/chest/tracheal_deviation.png', 'organs': ['Weasand']}
INFO: Starting segmentation process...
INFO: Validating and categorizing requested organs...
INFO: Loading and preprocessing image...
INFO: Running segmentation model...
INFO: Model prediction complete.
INFO: Generating segmentation visualization...
INFO: Visualization saved to temp/segmentation_53b03a54.png
INFO: Computing organ metrics...
INFO: Metrics computation complete.
INFO: Calculating Cardiothoracic Ratio (CTR)...
INFO: CTR status: CTR calculation requires successful segmentation of Heart,
Left Lung, and Right Lung.
INFO: Computing additional measurements...
  Segmentation tool returned tuple, extracting segmentation data
  Segmentation result keys: ['segmentation_image_path', 'metrics']
```

(b)



```
[DEBUG] airway - check_tracheal_position - segmentation
  Params: {
  "organs": [
    "Weasand"
  ]
}
  Result: {
  "success": true,
  "result": {
    "segmentation_image_path": "temp/segmentation_53b03a54.png",
    "metrics": {
      "Weasand": {
        "area_pixels": 12490,
        "area_cm2": 4.996,
        "centroid": [
          157.82834267413932,
          470.04891913530827
        ],
        "bbox": [14, 440, 301, 511
        ],
        "width": 71,
        "height": 287,
        "aspect_ratio": 4.042253521126761,
        "relative_position": {
          "top": 0.1809958058189671,
          "left": 0.5142767167782366,
          "center_dist": 0.31932350453271696
        },
        "mean_intensity": 118.73706965572458,
        "std_intensity": 24.211336890125622,
        "confidence_score": 0.032431717962026596
      },
    }
  },

INFO: Verifier analyzing original image independently (not using highlighted
segmentation)
✓ Completed airway in 24.76s

Agent execution completed in 24.76s
```

(c)

```
============================================================
ORCHESTRATOR: Running synthesis agent...
============================================================

SYNTHESIS AGENT: Performing clinical reasoning...

Collected 1 total findings from 1 agents

📋 DETAILED FINDINGS SUMMARY:
  1. [AIRWAY] tracheal_deviation (confidence: 9.0%)
     Evidence: Verifier independent assessment: The trachea appears deviated to the
right.. Disagrees with tool conclusion of 'midline' (position: 0.514)
     Measurements: tool_position=0.5142767167782366, tool_conclusion=midline,
vcot_direct_assessment=The trachea appears deviated to the right.

     ⚠️ Verifier DISAGREEMENT: Visual analysis contradicts tool result

     👁️ Direct Visual Assessment: The trachea appears deviated to the right.

     🔍 Context Verifier: Visual analysis contradicts tool result; Alternative:
ALTERNATIVE_ASSESSMENT: Right tracheal deviation is present, contrary to
tool's assessment of midline position.
```

(d)

```
📊 AGENT-SPECIFIC SUMMARY:
  🫁 AIRWAY: 1 findings, confidence=high
     📸 V-CoT triggered
     📋 Executed 1 tasks

🎯 SYNTHESIS STRATEGY:
   Pattern recognition: False
   Differential diagnosis: False

Synthesis completed in 8.15s

Analysis completed in 34.19 seconds

[DEBUG] Final results saved to:
logs/agent_debug/final_results_20250714_174214.json

============================================================
SUMMARY
============================================================
Answer: No, the trachea is not midline; it is deviated to the right.
Confidence: 80%

Activated Agents: ['airway']
```

Figure 5: RadAgents' reasoning trajectory for tracheal deviation detection.

correctness and utility. CheXbert is based on multi-label classification results for 5 or 13 diseases (along with one extra "normal" label). RadGraph considers literal entity agreement considering the positive or negative context of each entity. GREEN judges recall and precision errors by LLM prompting. RaTEScore is inspired by RadGraph but less sensitive to phrasing by an F1-like computation which allows semantic matching between entities based on a cosine similarity. The metrics are computed using their official and standardized implementations: RadGraph-F1[1], CheXbert-F1[2], RaTE Score[3], and GREEN[4].

Appendix C. Dataset

Operating an agentic system is costly because multimodal multi-step reasoning entails many LLM API calls; accordingly, we did not conduct very large scale experiments. We evaluate on MS–CXR and MS–CXR–T, whose annotations enable verification of intermediate outputs and support future reinforcement learning to incentivize tool use with open source models. For report generation, we condition on the current study and a single prior frontal image, rather than the full imaging history, to limit context length and processing cost. Dataset statistics are shown in Table 6.

Dataset	# Cases	# Images per case	Has prior?
MIMIC-CXR (subset)	181	2	Yes
MS-CXR (test set)	181	1	No
MS-CXR-T	785	2	Yes

Table 2: Details of datasets used in RadAgents.

Appendix D. Prompting

The input to **RadAgents** includes not only the image and query but also optional clinical context, such as patient demographics, indication, acquisition technique, comparison studies, and examination details (e.g., view/projection and region). We denote this aggregated context as Context in the templates for comprehensive CXR scanning and for report generation.

1. https://pypi.org/project/radgraph/0.1.2/
2. https://pypi.org/project/f1chexbert/
3. https://pypi.org/project/RaTEScore/0.5.0/
4. https://pypi.org/project/green-score/0.0.8/

```
Instruction: You are a helpful radiology
assistant. Describe what lines, tubes,
and devices are present and each of their
locations. Describe if pneumothorax is
present; if present, describe size on each
side. Describe if pleural eusion is present;
if present, describe amount on each side.
Describe if lung opacity (atelectasis,
brosis, consolidation, inltrate, lung mass,
pneumonia, pulmonary edema) is present;
if present, describe kinds and locations.
Describe the cardiac silhouee size. Describe
the width and contours of the mediastinum.
Describe if hilar enlargement is present;
if enlarged, describe side. Describe what
fractures or other skeletal abnormalities are
present.
  Given the [view] X-ray image(s) <images>,
  Q: Describe the findings in the image
following the Instructions, and Context:
{context}.
```

Appendix E. Reasoning modes and agentic workflows

Why interleaved reasoning? Classical vision and language systems compress the image once and then reason only in text. This single pass approach is ill suited to radiology, where clinicians iteratively survey the study, escalate suspicious findings to measurement, revisit earlier impressions as new evidence appears, and compare with priors. RadAgents operationalizes this practice by *interleaving* perception and reasoning: at intermediate points the system inspects additional visual evidence (for example, cropped regions or segmentation overlays) and updates its hypothesis before proceeding. A simple protocol is a two phase visual chain of thought (VCoT): first, a *pure visual assessment* that answers the question and cites observable evidence without tool outputs; second, an *evidence validation* step that reveals tool conclusions (for example, a measurement or a mask) and records agreement, disagreement, or uncertainty with a recalibrated confidence. This improves transparency and reduces anchoring on imperfect tools. Figure 6 illustrates how interleaved quantitative evidence aids enlarged heart detection, whereas a simple grounding guided approach fails.

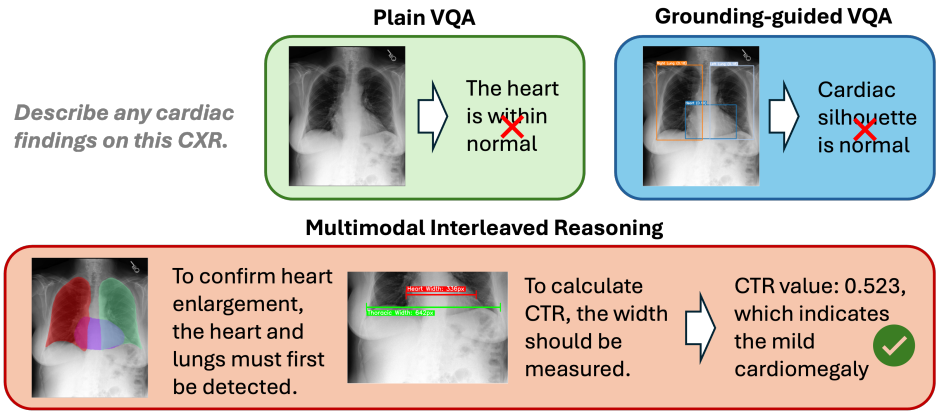


Figure 6: Different queries should trigger different reasoning modes. Simply cropping regions of interest and curating visual chain-of-thought reasoning is not a panacea.

(M1) Measurement

Goal. Provide objective and reproducible judgments for geometry constrained findings.
When. Explicit measurement requests or size abnormalities suggested by a sweep.
Tools. Segmentation yields organ masks and derived metrics; the cardiothoracic ratio (CTR) uses maximal cardiac width and thoracic width from bilateral lung extents, with projection recorded (PA versus AP) and raw pixel widths logged.
Evidence. Numeric values with projection, overlay visuals of the masks used, and brief caveats.

(M2) Localization

Goal. Localize small or subtle targets that benefit from high resolution crops.
When. Explicit localization queries or equivocal global signals.
Tools. Grounding proposes bounding boxes; segmentation constrains search when organ context matters; crop policies adapt to target scale.
Evidence. Boxes with confidence, region thumbnails, and landmark distances when relevant.

(M3) Characterization

Goal. Describe texture, morphology, distribution, and severity of parenchymal and pleural findings.
When. Opacity related queries or when a sweep suggests edema, atelectasis, pneumonia, or fibrosis.
Tools. Zone prioritization by a classifier, lung masks to focus attention, region crops from suspicious zones, VQA to standardize descriptors, and VCoT to justify labels.
Evidence. Pattern and distribution labels with severity, plus representative crops tied to the cited features.

(M4) Relational and comparative reasoning

Goal. Determine improvement, worsening, or stability relative to a prior study, as well as the relationships between each finding like complication.
When. Explicit comparison or association requests or any mention of a prior.
Tools. Temporal alignment of regions, adjustment for projection or view differences when possible, signed change statements in VCoT, and measurement deltas for quantitative targets.
Evidence. Paired crops and a concise comparison statement.

(M5) Differential Diagnosis

Goal. Synthesize evidence across modes to issue a patient level conclusion and differential with calibrated confidence.
When. After upstream modes have supplied sufficient evidence or when the query explicitly seeks a diagnosis.
Tools. The synthesizer aggregates structured findings, checks consistency, reconciles conflicts, and

Table 3: Retrieval quality at different k values, showing the trade-off between helpful and harmful rates.

k	Helpful-rate	Harmful-rate
1	0.48	0.09
3	0.62	0.14
5	0.65	0.20
7	0.66	0.25
9	0.67	0.28
11	0.67	0.31

maps evidence to diagnostic statements aligned with clinical guidance.

Evidence. A short justification that links key measurements, locations, and patterns to the final conclusion, plus uncertainty notes when appropriate.

Appendix F. Details of V-RAG

Multimodal retrieval. We retrieve images and their textual descriptions that align with the features of target medical images following (Chu et al., 2025). These references, rich in visual and textual medical details, guide response generation. To obtain embeddings, we use Rad-DINO, which provides robust representations across diverse CXR image types. For each image X_{img}, we extract its embedding $E_{img} = \mathbf{R}^d$, with $d = 768$, and store them in the embedding memory $\mathcal{M}$.

For efficient retrieval during inference, we build $\mathcal{M}$ using FAISS [5], a GPU-accelerated vector search system. We employ approximate kNN with the Hierarchical Navigable Small World (HNSW) algorithm (Malkov and Yashunin, 2018), enabling retrieval of the top-k most similar images in $\mathcal{M}$.

Sensitivity of k. We study how retrieval quality changes with the number of retrieved studies k with the sampled 100 cases from the MS-CXR test set used in this study. For each setting, we compute two metrics: *helpful-rate*, the percentage of retrieved studies that improve the answer, and *harmful-rate*, the percentage that hurt the answer. As shown in Table 3, As k increases, the harmful rate grows more quickly than the helpful, e.g., the helpful-rate increases from 0.48 at $k = 1$ to 0.65 at $k = 5$, while the harmful-rate also rises from 0.09 to 0.20. We hypothesize this is due to longer contexts imposing a heavier reasoning burden and increasing hallucination, consistent with prior LLM findings. This illustrates a trade-off: retrieving more studies provides greater chances of including helpful evidence but also increases the risk of introducing misleading content. To balance these effects, we choose $k = 3$ by default, achieving a helpful-rate of 0.62 with a moderate harmful-rate of 0.14.

Augmented Inference. In the inference stage, we encode the query image X_q to obtain its embedding. We then retrieve the top-k most similar images from $\mathcal{M}$, represented as $(I_1, \ldots, I_k)$ with their corresponding reports $(R_1, \ldots, R_k)$. These references are appended to the input of multimodal LLM to guide generation. The prompt is structured as:

```
This is the i-th similar image and its
report for your reference. [Reference]_i
... According to the query image and the
references, [Question] [Query Image].
```

where each reference is denoted as (I_i, R_i).

5. https://github.com/facebookresearch/faiss