TRACE: A Multimodal Multi-Agent System for Transparent Chest X-Ray Assistance

by

Zhaoyu Wang

Thesis submitted to the University of Ottawa in partial
fulfillment of the requirements for the degree of
Master of Applied Science in Biomedical Engineering

Ottawa-Carleton Institute of Biomedical Engineering
School of Electrical Engineering and Computer Science
Faculty of Engineering
University of Ottawa

© Zhaoyu Wang, Ottawa, Canada, 2026

Declaration of Authorship

I certify that this thesis is my own original work, except where the contributions of others are explicitly acknowledged. I understand the University's regulations regarding plagiarism and the disciplinary consequences associated with it. Any ideas, words, data, or other material drawn from the work of others have been properly cited and acknowledged at the point of use.

Abstract

Chest X-ray (CXR) AI has shown strong progress in abnormality detection, report generation, and multimodal medical reasoning, yet many existing systems remain limited as user-facing assistive tools. In particular, report-centered or text-only systems often provide limited support for evidence inspection, follow-up interaction, and cross-study continuity. This thesis presents TRACE (Transparent Radiology Assistance for Chest X-ray Explanation), a multimodal multi-agent system for transparent CXR assistance. TRACE integrates detector-supported image analysis, multimodal report drafting, report refinement, transparency-oriented interaction, consultation support, and archive-assisted comparison within a unified workflow.

TRACE was evaluated from four complementary perspectives: objective report quality, transparency-supported consultation preparation, archive-assisted cross-study comparison, and subjective usability and workload. Objective report quality was assessed on a 2,000-image frontal CXR benchmark constructed from MIMIC-CXR and the Indiana University Chest X-rays dataset, using external comparisons against LLaVA-Med and the MedRAX report-generation tool as well as an internal ablation of the TRACE reporting pipeline. TRACE achieved the best overall profile across most reported semantic and label-oriented metrics, indicating that detector-supported finding integration and report refinement can improve report quality within a broader assistive workflow. Two controlled user studies and a questionnaire-based human-factors assessment were also conducted with 48 participants. The transparency-oriented interface significantly improved users' ability to identify verification-worthy report conclusions, justify why they warranted clinician verification, and formulate more specific verification-oriented consultation questions. Archive support significantly improved report-level difference identification and reduced cross-study comparison time. In addition, the full TRACE interface was associated with higher perceived usability and lower subjective workload than the text-only interface, with the largest workload reductions observed in mental demand and frustration.

Overall, this thesis shows that, within the controlled benchmark and user-study settings examined here, CXR AI can provide measurable assistive value when it is designed and evaluated not only for report generation, but also for evidence inspection, verification-oriented interaction, and continuity across encounters within a unified assistive workflow.

Clinical usefulness remains to be established through further validation with patients and clinicians.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Professor Abdulmotaleb El Saddik, for his continuous support, encouragement, and guidance throughout my master's studies and the completion of this thesis. His insight, trust, and academic leadership have been invaluable to this work.

I would also like to sincerely thank Dr. Roberto Martínez and Dr. Haopeng Wang for their technical support, project guidance, and generous assistance throughout the development of this research. Their advice and discussions greatly helped shape the direction of the project.

Contents

Declaration of Authorship **ii**

Abstract **iii**

Acknowledgements **v**

List of Figures **ix**

List of Tables **xi**

1 Introduction **1**
- 1.1 Background, Problem Statement, and Motivation 1
- 1.2 Research Questions, Thesis Statement, and Design Overview 4
- 1.3 Thesis Contributions 8
- 1.4 Scope, Positioning, and Practical Significance 10
- 1.5 Thesis Organization 11

2 Related Work **12**
- 2.1 Core CXR Perception and Structured Visual Understanding 12
- 2.2 Radiology Report Generation and Multimodal CXR Models 15
- 2.3 Grounding, Transparency, and Trustworthy CXR Assistance 18
- 2.4 Agentic and Workflow-Oriented CXR Assistance 20
- 2.5 Positioning of TRACE Against Prior Systems 23
- 2.6 Summary of Research Gaps 25

3 Proposed System Architecture and Workflow **27**
- 3.1 Chapter Overview 27
- 3.2 System Architecture 28

3.2.1 Agentic Control Principle . . . 31
3.2.2 Design Rationale . . . 34
3.3 Orchestrator Agent . . . 36
3.3.1 LangGraph-Based Orchestration . . . 36
3.3.2 Routing Logic . . . 37
3.3.3 Session State and Continuity . . . 38
3.4 Diagnostic Agent . . . 38
3.4.1 Diagnostic Supervisor and Internal State . . . 40
3.4.2 CXR Validation and Report Drafting . . . 41
3.4.3 Ark$^+$-Based Abnormality Detection . . . 43
3.4.4 Operating Threshold Determination for Deployment . . . 45
3.4.5 Grad-CAM-Based Visual Inspection . . . 47
3.4.6 Report Synthesis, Review, and Refinement . . . 50
3.5 Consultation Agent . . . 52
3.6 Archive Agent . . . 55
3.7 Frontend User Interface and Multi-Level Transparency . . . 57
3.7.1 Multi-Level Transparency Design . . . 58
3.7.2 Interaction Structure . . . 60
3.7.3 Analysis Workspace . . . 62
3.8 Chapter Summary . . . 66

4 Evaluation **68**
4.1 Evaluation Overview . . . 68
4.2 Study Design, Materials, and Outcome Measures . . . 69
4.2.1 Participants and Recruitment . . . 71
4.3 Baseline Justification . . . 73
4.4 Statistical Methodology . . . 75
4.5 Experiment I: Objective Report Quality Evaluation . . . 77
4.5.1 External Comparison . . . 77
4.5.2 Diagnostic-Pipeline Ablation . . . 78
4.6 Experiment II: Transparency-Supported Consultation Preparation . . . 79
4.7 Experiment III: Archive-Assisted Cross-Study Comparison . . . 81
4.8 Questionnaire-Based Workload and Usability Assessment . . . 83

5 Results and Discussion **84**
5.1 Chapter Overview . . . 84

5.2 Results of Experiment I: Objective Report Quality Evaluation 84
5.3 Results of Experiment II: Transparency-Supported Verification Awareness and Consultation Preparation 87
5.4 Results of Experiment III: Archive-Assisted Cross-Study Comparison . . . 90
5.5 Results of the Questionnaire-Based Workload and Usability Assessment . . 93
5.6 Discussion 97
5.7 Chapter Summary 102

6 Conclusion and Future Work **103**
6.1 Conclusion 103
6.2 Limitations 105
6.3 Future Work 107

A Representative User-Study Materials **111**
A.1 Experiment II Materials 112
A.1.1 Representative Case A 112
A.1.2 Representative Case B 114
A.2 Experiment III Materials 115
A.2.1 Pair X 116
A.2.2 Pair Y 117

References **120**

List of Figures

1.1 A representative clinical CXR example with its accompanying radiology report. . . . 2
1.2 Summary of the main research gap addressed in this thesis. . . . 4
1.3 High-level workflow of TRACE for transparent CXR assistance. . . . 8

3.1 Internal multi-agent architecture of the TRACE reasoning core. . . . 29
3.2 Illustrative single-turn execution through the TRACE diagnostic pathway. 33
3.3 Diagnostic agent workflow. . . . 40
3.4 An example of a CXR with a Grad-CAM overlay. . . . 49
3.5 The workflow of the consultation agent. . . . 53
3.6 Archive agent workflow. . . . 56
3.7 An example of trajectory content. . . . 59
3.8 The user interface of TRACE. . . . 61
3.9 The conversation workspace of the user interface. . . . 62
3.10 The visual analysis view of the analysis workspace. . . . 63
3.11 The medical report view of the analysis workspace. . . . 64
3.12 The label explanation view of the analysis workspace. . . . 65
3.13 The AI logic view of the analysis workspace. . . . 66

4.1 The age distribution of participants. . . . 72
4.2 The background distribution of participants. . . . 73
4.3 Workflow of the user study. . . . 74

5.1 Summary of Experiment III outcomes under the *without-archive* and *with-archive* conditions. . . . 91
5.2 Overall questionnaire results by condition. . . . 94
5.3 Raw TLX dimension scores by condition. . . . 96

A.1 Representative original image and heatmap overlay for Case A. . . . 113

A.2 Representative original image and heatmap overlay for Case B. 114

List of Tables

3.1 Label-wise AUC-ROC of the fine-tuned Ark$^+$ detector. 45
3.2 Per-label operating thresholds used by the deployed Ark$^+$ detector. 47

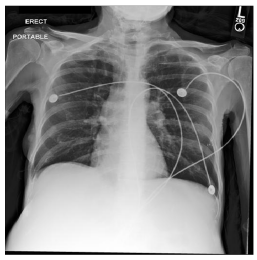
5.1 Objective report quality results for Experiment I. 85
5.2 Results of Experiment II. 88
5.3 Results of Experiment III. 90
5.4 Overall questionnaire results for usability and workload ($N = 48$, within-subject design). 94
5.5 Raw TLX dimension scores by condition ($N = 48$). 97

Chapter 1. Introduction

1.1 Background, Problem Statement, and Motivation

Chest X-ray (CXR) remains one of the most widely used imaging examinations in contemporary healthcare because it is fast, accessible, and clinically useful at large scale. According to the United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR), more than 4.2 billion radiological examinations are performed worldwide each year, and chest radiography accounts for a substantial share of this volume [1]. National statistics show the same pattern. In England, the National Health Service (NHS) reported 49.9 million imaging tests in 2024/25, of which 23.5 million were plain radiography procedures. Within the subset of imaging tests commonly used to diagnose or exclude cancer, CXR remained the most common examination [2]. These figures show that chest radiography is a routine part of clinical care, so even modest improvements in interpretation or workflow can affect a very large number of patients.

Its routine use should not be mistaken for interpretive simplicity. A clinically meaningful CXR reading requires systematic inspection of multiple anatomical structures, recognition and prioritization of relevant findings, differentiation of pathology from normal variation, and synthesis of those observations into a concise and clinically useful radiology



Findings

The lungs are clear of focal consolidation, pleural effusion or pneumothorax. The heart size is normal. The mediastinal contours are normal. Multiple surgical clips project over the left breast, and old left rib fractures are noted.

Impression

No acute cardiopulmonary process.

Figure 1.1: A representative clinical CXR example with its accompanying radiology report.

report [3, 4]. Figure 1.1 shows a representative clinical CXR together with its accompanying report, which is typically organized into sections such as *Findings* and *Impression*. The report condenses visual evidence into a form that supports subsequent clinical decision making.

Because CXR interpretation combines visual recognition, structured reasoning, medical summarization, and communication of findings, chest radiography has become a major application area in medical artificial intelligence (AI). Existing studies have reported strong results on abnormality classification, localization, segmentation, visual question answering, and automated report generation [3, 5–7]. When such tools are integrated into clinical workflows, they can improve reporting efficiency and reader agreement [8–11]. However, strong performance on an isolated benchmark does not by itself establish practical usefulness. A model may perform well on one narrowly defined task and still offer limited value in realistic use if its outputs are difficult to inspect, difficult to question, or difficult to integrate into a broader workflow.

Foundation models, including large language models (LLMs) and large multimodal

models (LMMs), have widened the scope of what CXR AI can do. Systems such as LLaVA-Med and CheXagent suggest that one model can support more unified forms of image-grounded medical assistance [12–16]. Yet these models still face important limitations in medical settings. They may hallucinate unsupported findings, produce inconsistent outputs, miss relevant image regions, or provide conclusions whose evidential basis is difficult for users to assess [12–14, 17]. For that reason, their direct use as end-to-end clinical systems remains difficult to justify.

A central limitation of the current literature is fragmentation. Many CXR AI systems are designed for a single downstream objective, such as disease prediction, report generation, or question answering, and they usually operate as separate tools. In practice, a CXR interpretation rarely ends with one output. Users often need a connected sequence of capabilities that includes diagnostic interpretation, visual evidence inspection, follow-up clarification, and, when prior information is available, cross-study comparison.

This fragmentation leads to two related research problems. The first is a system design problem: how can multiple AI capabilities be organized into one coherent CXR assistance workflow? The second is a transparency problem: how can such a workflow expose enough visual, textual, and procedural information for users to understand how results were produced and how they should be used? These problems matter most when the system is intended as a user-facing assistant that supports understanding, follow-up questioning, and communication rather than as a backend automation tool.

As shown in Figure 1.2, the main gap is not the lack of progress on individual CXR

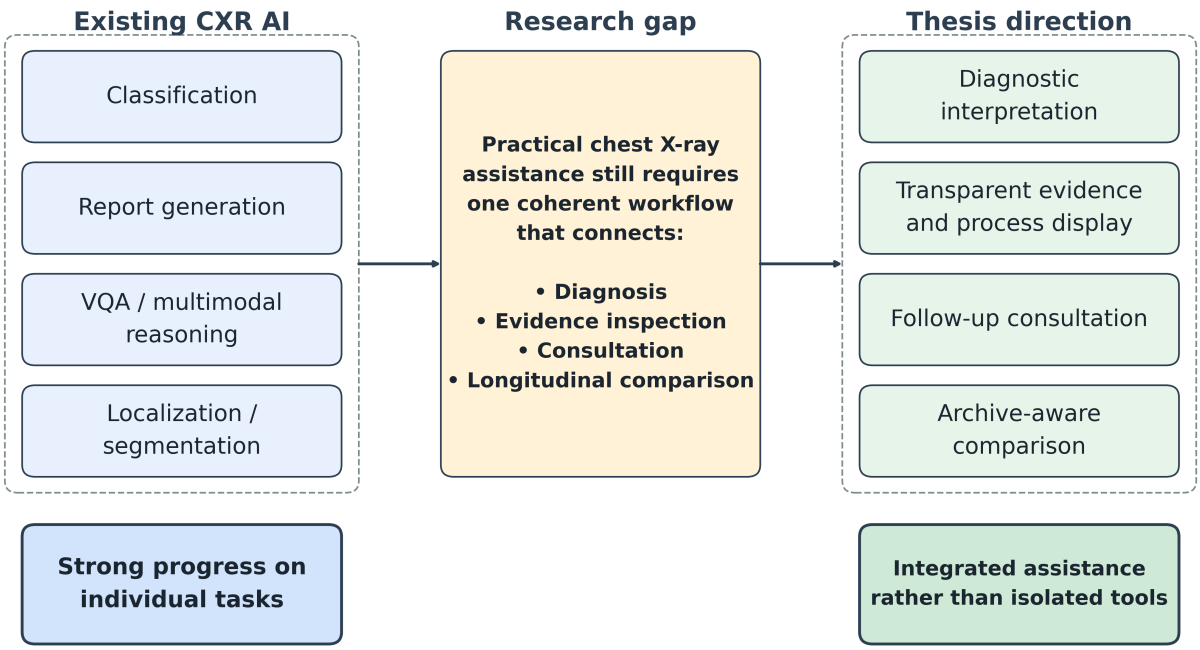


Figure 1.2: Summary of the main research gap addressed in this thesis.

AI tasks. The main gap is the lack of an integrated and transparent user-facing workflow that connects those tasks in practical use. This thesis addresses that gap by studying how diagnostic analysis, evidence inspection, follow-up consultation, and archive-assisted comparison can be combined within one assistive CXR framework.

1.2 Research Questions, Thesis Statement, and Design Overview

The objective of this thesis is to investigate whether multiple specialized AI capabilities can be integrated into a unified CXR assistance framework that improves report generation,

user-facing transparency, and workflow continuity. More specifically, this thesis examines whether a detector-supported report-generation pipeline can improve report quality, whether multi-level transparency can better support verification-oriented user interaction, and whether archive-assisted comparison can improve report-level comparison efficiency and difference identification.

The study is guided by the following research questions and hypotheses:

RQ1. *Does a detector-supported report-generation pipeline improve objective CXR report quality relative to direct report-generation baselines?*

H1. A detector-supported and review-enhanced report-generation pipeline will achieve stronger objective report quality than direct report-generation baselines, particularly on label-oriented abnormality-preservation metrics.

RQ2. *Does multi-level transparency improve users' ability to identify verification-worthy report conclusions and formulate verification-oriented consultation questions?*

H2. A transparency-oriented interface will improve verification-oriented follow-up preparation relative to a text-only interface.

RQ3. *Does archive-assisted cross-study comparison improve report-level comparison efficiency and difference identification?*

H3. Archive-assisted comparison will reduce comparison time and improve report-level difference identification relative to a condition without archive support.

The central claim tested in this thesis is that organizing CXR AI as a transparent assistive workflow can improve three measurable outcomes: objective report quality, verification-oriented user interaction, and report-level continuity across encounters. This claim is evaluated through three corresponding components: detector-supported report generation, a transparency-oriented interface, and archive-assisted cross-study comparison.

To test this claim, this thesis presents TRACE (*Transparent Radiology Assistance for Chest X-ray Explanation*), a multimodal multi-agent system for transparent CXR assistance. TRACE is designed as a user-facing assistive system that integrates detector-supported image analysis, multimodal report drafting, report refinement, transparency-oriented interaction, follow-up consultation, and archive-assisted comparison within one unified workflow. Rather than treating AI-based routing as the primary system objective, TRACE uses orchestration to coordinate complementary AI capabilities and assemble diagnostic, visual, textual, and longitudinal support into an inspectable response.

At interaction turn t, the active case context is represented as

$$s_t = (I_t, q_t, a_t, h_t, F_t, R_t),$$

where I_t denotes the current CXR image, q_t the current user query, a_t an optional patient alias for archive retrieval, h_t the conversation history, F_t the current structured findings, and R_t the current radiology-style report. The findings and report may initially be empty and are updated as the case progresses. Based on this context, the orchestrator applies

a routing policy $w_t = \pi(s_t)$ to select the diagnostic, consultation, or archive pathway according to the support required by the current interaction.

The selected specialist pathway contributes task-specific reasoning support. This support is summarized as

$$\Omega_t = \{F_t, H_t, R_t^0, K_t, \Delta_t, \tau_t\},$$

where H_t denotes Grad-CAM visual inspection cues, R_t^0 the initial image-conditioned report draft, K_t retrieved medical knowledge when required for consultation, Δ_t the current-versus-prior report comparison, and τ_t the recorded execution trajectory. Not every element is required at every interaction turn; the contents of the support bundle depend on the selected specialist pathway.

These elements are assembled into an inspectable response package,

$$y_t = \{R_t, F_t, H_t, \Delta_t, m_t, \tau_t\},$$

where m_t denotes the conversational message returned to the user. The response package is presented through the interactive virtual clinic and can be inspected through the report, structured findings, visual cues, comparison information, conversational output, and execution trajectory. The resulting outputs also become part of the evolving case context for subsequent follow-up interaction, forming a closed-loop assistance process.

Figure 1.3 summarizes this workflow from user input and active case context through AI-supported specialist reasoning to the final inspectable response package.

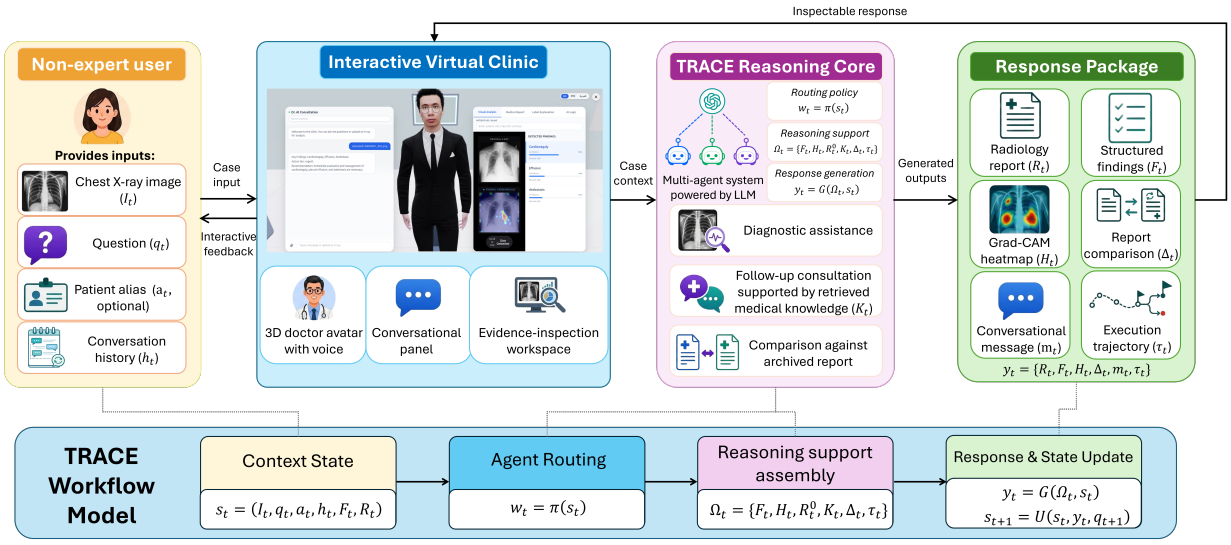


Figure 1.3: High-level workflow of TRACE for transparent CXR assistance. A user provides a chest X-ray image I_t, query q_t, optional patient alias a_t, and conversation history h_t through the interactive virtual clinic. The active context s_t informs specialist routing w_t, after which the TRACE reasoning core coordinates diagnostic, consultation, or archive support and assembles the available reasoning-support elements Ω_t. The resulting response package y_t may contain the finalized report R_t, structured findings F_t, Grad-CAM visual cues H_t, report comparison Δ_t, conversational message m_t, and execution trajectory τ_t. The returned outputs support inspection and subsequent follow-up interaction within the same case context.

1.3 Thesis Contributions

This thesis makes contributions at both the system and research levels.

At the system level, it presents TRACE, a multimodal multi-agent system that unifies diagnostic analysis, follow-up consultation, and archive-assisted cross-study comparison within one user-facing CXR assistance workflow. It also develops a detector-supported report-generation pipeline that combines structured abnormality detection, multimodal

report drafting, report review, and report refinement to produce radiology-style outputs. In addition, it integrates multi-level transparency through visual heatmap overlays, structured findings, report presentation, and process-level execution traces. Finally, it implements an archive-aware mechanism that supports retrieval and report-level comparison of prior cases when patient-level context is available.

At the research level, the thesis provides evidence that a detector-supported and review-enhanced pipeline can improve objective report quality relative to direct report-generation baselines. It further shows that multi-level transparency helps users identify verification-worthy report conclusions, justify why those conclusions warrant clinician verification, and formulate more specific verification-oriented consultation questions. The archive-assisted comparison study shows that report-level continuity support can improve difference identification and reduce comparison time. These findings support the view that practical CXR AI should be studied not only as a modeling problem, but also as a workflow, transparency, and interaction problem.

Relative to prior workflow-oriented CXR systems, the contribution of this thesis lies not merely in integrating multiple tools into one interface. The main research insight is that detector-supported reporting, transparency-oriented inspection, and archive-assisted comparison can be evaluated as complementary parts of one assistive CXR workflow. The resulting contribution is therefore both system-oriented and evaluation-oriented.

1.4 Scope, Positioning, and Practical Significance

This thesis positions TRACE as an assistive system for CXR analysis in user-facing settings. The goal is to study how AI can support report generation, evidence inspection, follow-up interaction, and cross-study comparison within a bounded CXR workflow.

The scope of the framework is intentionally bounded. TRACE is designed for detector-supported report generation, transparency-supported follow-up interaction, and archive-assisted report comparison within a CXR-specific workflow. It is not intended as a general-purpose medical reasoning agent for arbitrary clinical queries, and it does not attempt full longitudinal diagnosis across serial imaging studies.

The archive pathway should be understood within this bounded scope. Its role is to retrieve prior report context and support report-level comparison of current and prior findings when prior information is available. It therefore functions as a continuity aid within the assistive workflow rather than as a dedicated temporal diagnosis model.

The practical significance of this framing is that TRACE is evaluated as a user-facing assistive workflow, with attention to report quality, inspection support, follow-up preparation, and report-level continuity rather than standalone prediction alone.

1.5 Thesis Organization

The remainder of this thesis is organized as follows. Chapter 2 reviews related work in CXR perception, automated radiology report generation, multimodal medical foundation models, grounding and transparency, and agentic workflow-oriented medical AI, and positions TRACE with respect to prior systems. Chapter 3 presents the architecture, components, design rationale, and workflow of the proposed system. Chapter 4 describes the evaluation design, including objective report quality evaluation, transparency-supported interaction, archive-assisted cross-study comparison, baseline selection, and statistical methodology. Chapter 5 presents the experimental results, ablation analysis, and discussion. Chapter 6 concludes the thesis, discusses its limitations, and outlines directions for future work.

Chapter 2. Related Work

CXR AI has progressed rapidly in visual perception, report generation, multimodal modeling, grounding, explainability, and, more recently, agentic medical systems. However, these lines of work have not yet converged into a single assistance paradigm. Many studies still target one downstream task, such as abnormality classification, report generation, or visual question answering. In practical use, CXR interpretation usually involves a connected sequence of activities that includes image interpretation, report drafting, evidence inspection, follow-up clarification, and comparison with prior studies when available. For that reason, the most relevant prior work for this thesis includes not only individual CXR models, but also studies that examine how multiple capabilities can be organized into usable and inspectable assistance systems.

2.1 Core CXR Perception and Structured Visual Understanding

Core perception tasks provide the visual foundation of CXR AI. Before a system can generate a report, answer a follow-up question, or support comparison across studies, it must first extract clinically relevant information from the image. This is why a large part of the literature has focused on multi-label abnormality classification, weakly or strongly super-

vised localization, abnormality detection, and anatomical or disease-related segmentation [18–22].

Early progress was driven mainly by image-level classification benchmarks. ChestX-ray8 showed that hospital-scale chest radiograph collections could be repurposed into a benchmark for weakly supervised disease classification and localization, helping establish the multi-label formulation that has become standard in CXR interpretation [18]. CheXpert extended this line of work by introducing large-scale labels for common radiographic observations together with explicit treatment of uncertainty in report-derived annotations [19]. Around the same period, CheXNet drew wider attention to deep convolutional models for chest radiograph classification through strong results on pneumonia detection and broader abnormality prediction [23]. These studies established classification as the dominant early task in CXR AI.

Image-level prediction, however, answers only part of the radiological question. A model may indicate that an abnormality is present without showing where it is located, how extensive it is, or whether the prediction is grounded in a plausible anatomical region. This limitation led naturally to work on localization and detection. Early studies often relied on weak supervision, using class activation maps or attention mechanisms to infer candidate disease regions from image-level labels [18, 23]. These methods provided rough visual evidence, but their spatial precision remained limited. Later datasets addressed this problem more directly. VinDr-CXR introduced radiologist-generated annotations with both global labels and bounding boxes for local findings, making more rigorous evaluation of abnormality localization possible [20]. ChestX-Det10 likewise provided box-level annotations for

thoracic abnormalities and framed the task explicitly as disease detection rather than only classification [21]. This progression brought CXR AI closer to the visual search process performed by radiologists.

Segmentation extends this progression toward finer structural understanding. In chest radiography, segmentation may refer to delineation of anatomical regions such as the lungs or heart, or to disease-related regions. Compared with classification, it provides more explicit structural information and can support downstream tasks such as region-constrained explanation, organ-aware measurement, and anatomically grounded reasoning. Progress in this area was long constrained by the scarcity of large public mask annotations, but newer datasets have begun to change that. CheXmask, for example, introduced a large-scale multi-center dataset of anatomical segmentation masks derived from major public CXR sources, substantially expanding the scale at which anatomy-aware CXR models can be trained and evaluated [22].

As these tasks matured, the field increasingly treated classification, localization, detection, and segmentation as related forms of visual understanding rather than as fully separate problems. Foundation X is a representative example. It combines classification, localization, and segmentation through a unified pretraining strategy and shows that cross-task learning can improve generalization and fine-grained visual performance [24]. Ark$^+$ extends this broader foundation-model perspective by aggregating heterogeneous expert labels across multiple public datasets and demonstrating wide diagnostic scope, adaptability, and robustness [25]. Taken together, this literature shows that core perception has become substantially stronger and more structurally informed, although its primary contri-

bution remains the extraction of visual evidence rather than the full delivery of user-facing assistance.

2.2 Radiology Report Generation and Multimodal CXR Models

Automated radiology report generation marks a shift in CXR AI from visual prediction to clinically usable textual output. Unlike abnormality classification, this task requires a system not only to detect image content, but also to express that content in radiological language. The generated text must indicate which abnormalities are present, how they are described, where they are located, which findings deserve emphasis, and how they should be summarized. Recent surveys therefore treat report generation as a multimodal problem involving image understanding, language modeling, feature fusion, and clinically informed evaluation rather than as a simple form of image captioning [26].

Early report-generation work was largely shaped by encoder-decoder architectures adapted from natural image captioning. Among the influential studies in this line, R2Gen showed that report generation benefits from mechanisms that retain prior decoding context and longer-range textual structure, which improves the coherence of radiology-style narratives [27]. Around the same period, knowledge-enhanced approaches began to address a central limitation of purely data-driven generation, namely that fluent text was not always clinically dependable. Zhang et al. introduced a knowledge-graph-based framework that injected structured disease information into the generation process, showing that external medical structure could improve report quality beyond conventional captioning pipelines

[28].

As work on report generation matured, lexical fluency became an insufficient objective. A report may read smoothly while still omitting important findings, overstating abnormalities, or expressing clinically misleading relations. This led to greater emphasis on diagnosis-aware and clinically guided generation. PromptMRG is representative of this development because it uses diagnosis-driven prompts to steer generation toward clinically relevant content instead of relying on visual features alone [29]. More broadly, this line of work reflects a shift away from generic image-to-text generation toward report generation shaped by disease semantics, abnormality cues, and structured intermediate representations.

The rise of LLMs changed this landscape again. Instead of training all components from scratch, newer approaches increasingly align CXR visual features with pretrained language models so that broader linguistic knowledge can be reused in the radiology domain. R2GenGPT is a clear example of this shift, showing that frozen LLMs can be adapted to radiology report generation through lightweight vision-language alignment modules [30]. In a similar vein, *Bootstrapping Large Language Models for Radiology Report Generation* introduced in-domain instance induction and coarse-to-fine decoding to make more effective use of LLM priors for radiology-specific text generation [31]. At this stage, progress depends not only on stronger visual encoders, but also on how pretrained language models are adapted without losing clinical precision.

These developments connect naturally to the rise of multimodal foundation models for

chest radiography. Earlier systems were usually built for individual tasks, with separate models for classification, report generation, or question answering. More recent work instead explores whether image and language representations can be learned within a shared framework and reused across multiple chest radiograph tasks [32, 33]. LLaVA-Med helped establish biomedical multimodal instruction tuning as a viable direction, but later work increasingly moved toward CXR-specific vision-language models (VLMs) trained more directly on radiographic images, reporting conventions, and domain-specific instruction data [15, 32, 33]. CheXagent is one of the clearest examples of this transition, showing that a single VLM can support broader CXR interpretation rather than one narrow downstream task [16]. CXR-LLaVA moved in a similar direction from an open-source perspective, while LLaVA-Rad drew attention to the practical importance of accessibility and local deployment for radiology-oriented multimodal systems [34, 35].

Even with these advances, multimodal and report-generating CXR models still face recurring concerns about hallucinated findings, limited transparency, uneven factual reliability, and weak support for multi-step interaction [32, 33, 35]. This literature has substantially expanded what a single model can do, but it has also made evaluation criteria such as grounding, factual consistency, and practical usability more important than textual fluency alone.

2.3 Grounding, Transparency, and Trustworthy CXR Assistance

As CXR AI systems take on broader roles, questions of trustworthiness become central rather than secondary. In medical settings, it is not enough for a model to produce the correct answer on average. Users also need to judge whether an output is visually plausible, whether it is supported by the image, and whether possible error is communicated appropriately. Recent reviews of explainable AI in medical imaging therefore link interpretability directly to clinical acceptability, especially in high-stakes settings such as radiology [36, 37].

The most familiar approach to explainability in CXR AI has been post-hoc visual explanation. Saliency-based methods, especially class activation mapping techniques such as Grad-CAM, are widely used to highlight image regions that appear to contribute to a model's prediction [36, 38]. Their appeal is clear. They are relatively simple to apply, visually intuitive, and well suited to classification-style pipelines. For these reasons, they have become a common form of explanation in chest radiograph studies. At the same time, their limitations are now well recognized. Reviews have repeatedly noted that saliency maps often provide only coarse localization, may vary across architectures or perturbations, and do not explain how the model arrived at its decision in any fuller sense [36, 39]. Heatmaps can therefore make a system easier to inspect, but they do not by themselves establish trustworthy reasoning.

These limitations have pushed the field toward more explicit forms of grounding and more rigorous evaluation of explainability claims. REFLACX marked an important step

in that direction by combining CXRs with synchronized radiologist eye tracking, dictation, abnormality localization, and report information [40]. Because the dataset captures where radiologists actually looked while interpreting the image, it provides a more concrete basis for comparing model attention with human visual search. More recently, LATTE-CXR extended this line of work with locally aligned image-text pairs and eye-tracking-derived sentence-level regions for phrase grounding and explainability assessment [41, 42]. With resources of this kind, explainability is no longer framed only as informal visualization. It is increasingly framed as the question of whether textual claims and model behavior can be tied to radiologically meaningful regions.

The same issue appears in report generation. A report may be fluent and broadly plausible while still being unsupported by the image, or it may describe a genuine abnormality while assigning it to the wrong location. Grounded report generation responds to this problem by linking textual findings to image regions. MAIRA-2 is a notable example because it extends CXR report generation toward spatially grounded outputs [43]. PadChest-GR further strengthened this area by introducing a manually curated bilingual CXR dataset for grounded report generation with sentence-level finding descriptions and bounding-box localization [44]. These developments matter because they change what a report-producing system is expected to provide. The goal is no longer only a plausible narrative, but also some visible basis for its claims.

Trustworthiness also extends beyond grounding to factual reliability. Large multimodal models may generate convincing but incorrect findings, omit important abnormalities, or hallucinate unsupported statements. This has led to more explicit work on fact-checking

AI-generated CXR reports. Mahmood et al. proposed phrase-grounded fact-checking for automatically generated CXR reports, targeting errors in both findings and their locations [45]. Guidance on the use of LLMs in radiology has likewise argued for structured prompting, constrained outputs, and expert review instead of unconstrained autonomous use [37]. CXR-specific evaluations such as CXR-Agent have similarly shown that VLMs can still produce confident but unsupported statements, creating new interpretive risks rather than simply reducing workload [46]. This literature shows that trustworthy CXR assistance involves not only visual plausibility, but also explicit grounding, factual consistency, and practical conditions under which users remain able to inspect and correct system outputs.

2.4 Agentic and Workflow-Oriented CXR Assistance

The recent rise of agentic AI has introduced a different way of thinking about medical AI. Rather than treating a model as a passive input-output system, agentic frameworks treat it as an active controller that can plan, call external tools, retrieve information, maintain working memory, and refine its response over multiple steps. Recent surveys describe this development as a move from static LLMs toward interactive systems that combine reasoning, planning, external capacity enhancement, and explicit action in complex environments [47]. This development is especially relevant to radiology, where many meaningful tasks are inherently sequential and embedded in workflow, involving image inspection, information retrieval, draft generation, quality control, and communication rather than a single isolated prediction [48].

One strand of work has explored collaborative multi-agent reasoning. MDAgents is representative of this approach because it dynamically chooses between solo and group collaboration structures according to problem complexity [49]. Another strand places greater emphasis on tool use. MMedAgent is a leading example in this category, showing that performance can improve when a system is trained not only to answer a question, but also to decide which external method is appropriate for the problem at hand [50]. This is particularly relevant in medicine, where many tasks are not best handled by a single universal model, but instead benefit from coordination among segmentation tools, grounding tools, report generators, classifiers, and retrieval mechanisms. Recent agentic medical systems therefore make explicit a view that a capable medical assistant may need to be a tool-using system rather than a single end-to-end predictor.

As these systems became more complex, the limitations of conventional evaluation also became more apparent. Standard medical question-answering benchmarks are often too static to reflect the stepwise character of real clinical work. AgentClinic addressed this problem by introducing a multimodal benchmark in simulated clinical environments where agents must interact with patients, gather incomplete information, and use tools across multiple stages [51]. MedAgentBench extended this direction into a realistic electronic health record setting, evaluating whether medical LLM agents can complete patient-specific tasks through interactions with structured records and standard APIs [52]. More broadly, recent agent benchmarks such as AgentBench, MMAU, SWE-bench, and medical agent benchmarks show that reasoning, tool use, and workflow completion have become important evaluation targets rather than optional extras [53].

Within radiology, recent reviews suggest that agentic systems may be relevant not only to diagnosis, but also to report generation, workflow management, reproducibility, and clinical decision support [48]. At the same time, these reviews remain cautious. They point to unresolved issues involving privacy, regulation, computational cost, and safe deployment, and they note that agentic radiology systems remain much less clinically validated than their rapid research growth might suggest [48]. The current literature is therefore marked by a clear tension. Agentic architectures offer a plausible systems-level response to fragmentation, but they also introduce new operational and governance problems that do not arise in quite the same way for narrower single-task models.

CXR interpretation has recently become one of the clearest domains in which this agentic perspective has been made concrete. MedRAX is a representative example, integrating multiple CXR analysis tools and multimodal models within a ReAct-style framework for multi-step chest radiograph reasoning [53]. Its importance lies in showing that CXR analysis can be treated as a coordinated workflow involving visual question answering, segmentation, grounding, report generation, disease classification, and iterative reasoning, rather than as a single-model prediction task.

A related issue is human-AI collaboration and workflow integration. As radiology AI matures, attention is moving beyond the question of whether an algorithm can achieve strong standalone performance. The more relevant question is increasingly how such systems should be used alongside clinicians in practice [54]. Reviews of radiology workflow AI likewise argue that meaningful benefit depends on when, where, and in what form assistance is introduced, rather than on model accuracy alone [55]. Evidence from medical

imaging more broadly suggests that human-AI collaboration can reduce workload while maintaining or modestly improving diagnostic performance, although the underlying evidence remains heterogeneous [56]. Within chest radiography, Flamingo-CXR is especially relevant because it evaluates clinician collaboration with a VLM for chest radiograph report generation, shifting attention from standalone model output to clinician-in-the-loop support [57]. More adaptive collaborative systems such as CoRaX further suggest that CXR assistance may benefit from interaction designs that respond to human behavior rather than only presenting static AI outputs [58].

2.5 Positioning of TRACE Against Prior Systems

The literature reviewed above shows that recent CXR systems are moving beyond isolated prediction tasks toward more integrated multimodal and agentic designs. However, the goals of these systems are not identical. Some works primarily target stronger visual perception, some focus on report generation, some emphasize multimodal medical reasoning, and some aim to build general-purpose tool-using agents for complex chest radiograph queries. For that reason, it is important to position the proposed framework not only against single-task baselines, but also against prior integrated systems.

Among prior systems, MedRAX is the closest reference point because it explicitly frames CXR analysis as a ReAct-style agentic workflow that coordinates multiple specialized tools [53]. MedRAX is important because it shows that chest radiograph interpretation can be organized as a multi-step reasoning process rather than a single forward pass. It also

demonstrates the value of tool orchestration for handling complex image-grounded medical questions.

At the same time, TRACE is positioned differently in both objective and scope. MedRAX is primarily oriented toward complex chest radiograph reasoning across diverse benchmark queries, with emphasis on flexible tool use, multi-step problem solving, and broad task coverage. By contrast, TRACE is designed as a transparency-centered, user-facing assistive framework. Its main objective is not to answer arbitrary benchmark-style medical queries, but to support a more specific chain of practical functions: detector-supported report generation, evidence inspection through multi-level transparency, follow-up clarification, and archive-assisted cross-study comparison.

The two systems also differ in how they handle transparency and continuity. MedRAX exposes multi-step tool use and intermediate tool outputs through an interactive agent workflow [53]. However, MedRAX describes short-term interaction memory and tool-output caching rather than a persistent patient-level archive for saving reports, retrieving prior studies, and generating interval-change summaries. TRACE therefore extends the user-facing CXR workflow by adding report-level archive-assisted comparison, in which prior case context can be retrieved and compared with the current findings within the same assistive session.

The evaluation emphasis is likewise different. MedRAX focuses on agentic reasoning capability and benchmark performance over complex CXR queries [53]. TRACE instead evaluates three narrower but practically important questions: whether a detector-supported

report-generation pipeline improves objective report quality, whether multi-level transparency improves verification-oriented follow-up preparation, and whether archive-assisted cross-study comparison improves report-level comparison efficiency and difference identification. TRACE should therefore be understood as complementary to prior agentic systems rather than as a simple variant of them.

2.6 Summary of Research Gaps

The literature reviewed above shows substantial progress in perception, report generation, multimodal modeling, grounding, and agentic medical AI. At the same time, four gaps remain especially relevant to the system developed in this thesis.

First, much of the literature remains fragmented across subtasks. Strong classifiers, localizers, report generators, and multimodal models have all emerged, but they are often studied as separate endpoints rather than as parts of a coherent CXR assistance process. This creates a mismatch between benchmark-oriented progress and realistic use, where image interpretation is often followed by evidence inspection, clarification, communication, and comparison with prior studies.

Second, multi-level transparency remains unevenly supported. Existing work has shown the value of visual overlays, grounded text, and fact-checking, but these capabilities have not yet been incorporated consistently into practical end-to-end CXR systems. A user-facing CXR assistant can therefore benefit from transparency at several levels, including

visual evidence, textual explanation, and process-level traceability.

Third, support for user-facing and workflow-oriented interaction remains limited. Many prior systems are designed to produce a prediction or a report, but not to support the natural next steps that often follow interpretation, such as follow-up questioning, history-aware comparison, or inspection of intermediate evidence. Agentic and multimodal systems suggest that such interaction is becoming technically feasible, but relatively few CXR systems are organized explicitly around a transparent and unified assistive process.

Fourth, integrated CXR systems remain weakly evaluated as transparency-centered, user-facing assistive frameworks. Prior agentic systems such as MedRAX have shown the value of tool orchestration and multi-step reasoning, but they do not directly address the narrower question investigated in this thesis: how CXR AI should be organized and evaluated as a transparency-centered framework that combines detector-supported report generation, evidence inspection, follow-up support, and archive-assisted cross-study comparison.

These gaps motivate the system developed in this thesis. The proposed work treats CXR assistance as a workflow-level problem that requires perception, reporting, follow-up consultation, archive-assisted comparison, and multi-level transparency to be integrated and evaluated within one user-facing framework.

Chapter 3. Proposed System Architecture and Workflow

3.1 Chapter Overview

TRACE is a multi-agent CXR assistance framework that integrates diagnostic analysis, follow-up consultation, and archive-assisted cross-study comparison within one user-facing workflow. The architectural objective is not merely to place a prediction model behind an interface, but to organize multiple specialized AI capabilities into a system whose outputs can be inspected at the visual, textual, and process levels.

The active runtime path begins with a unified frontend request and ends with a combined response containing report text, structured findings, visual inspection artifacts, conversational output, and execution traces. A centralized orchestration layer maintains case state and routes each request to the diagnostic, consultation, or archive workflow as needed. This chapter describes the implemented architecture and explains the main design choices behind the multi-agent workflow, detector-supported report generation, and Grad-CAM-based inspection support. These design choices are important because later evaluation chapters assess not only whether the system functions as an integrated prototype, but

also which parts of the diagnostic pipeline contribute to report quality and user-facing transparency.

3.2 System Architecture

TRACE is implemented as a centralized multi-agent framework for CXR assistance. Figure 3.1 presents the internal architecture of the TRACE reasoning core. At interaction turn t, the orchestrator observes the active case context

$$s_t = (I_t, q_t, a_t, h_t, F_t, R_t),$$

where I_t denotes the current CXR image, q_t the current user query, a_t an optional patient alias for archive retrieval, h_t the conversation history, F_t the current structured findings, and R_t the current radiology-style report. The findings and report may initially be empty and are updated as diagnostic processing proceeds. This notation provides a compact representation of the decision-relevant case context; the runtime state additionally contains implementation-specific fields described in Section 3.3.

Given the active context, the orchestrator applies a routing policy $w_t = \pi(s_t)$ and delegates the current turn to one of three specialist pathways, $w_t \in \{A_D, A_C, A_A\}$, corresponding to the diagnostic, consultation, and archive agents. These pathways have different support requirements: diagnostic analysis requires image-conditioned and detector-supported evidence, consultation requires case-grounded visual or knowledge support, and

archive-assisted comparison requires prior-report retrieval and difference analysis.

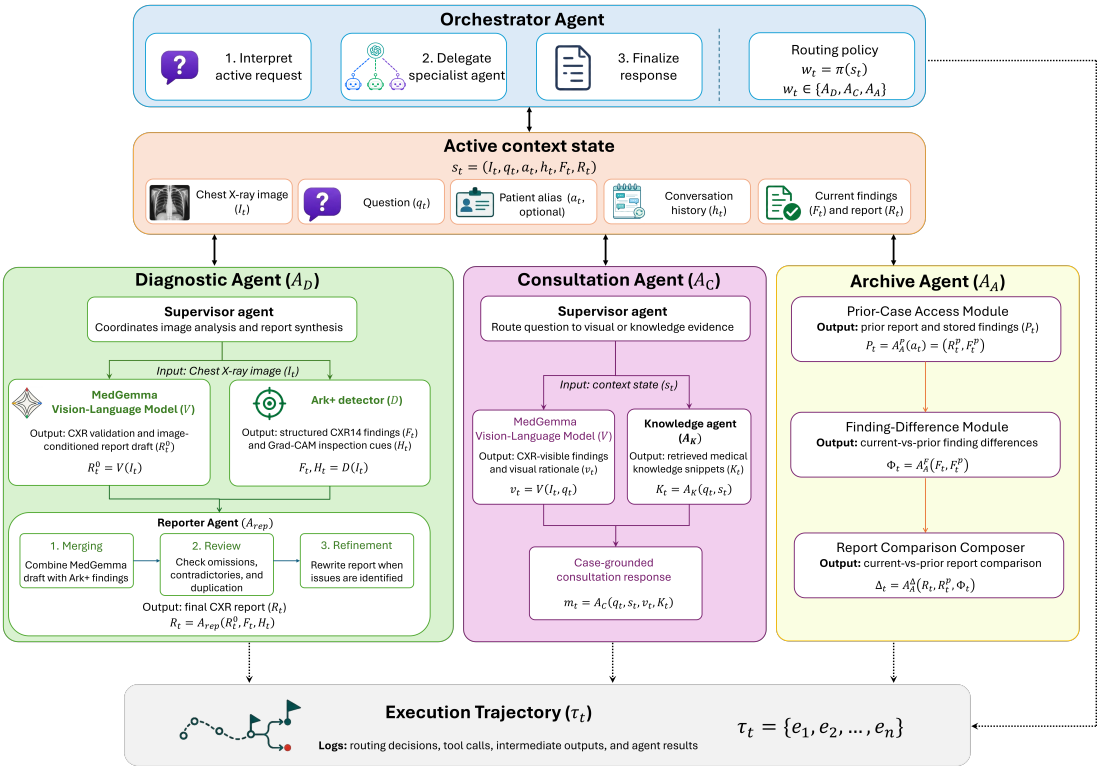


Figure 3.1: Internal multi-agent architecture of the TRACE reasoning core. The orchestrator observes the active context state s_t and delegates each interaction turn to the diagnostic agent A_D, consultation agent A_C, or archive agent A_A. The diagnostic agent combines MedGemma-based multimodal report drafting, Ark$^+$-based structured CXR detection and Grad-CAM visual cues, and reporter-agent refinement to produce the final report. The consultation agent supports follow-up questions using image-grounded visual support or retrieved medical knowledge. The archive agent retrieves prior report context and summarizes current-versus-prior differences when a patient alias is available. Routing decisions, tool calls, intermediate outputs, and specialist-agent results are recorded in the execution trajectory τ_t for process-level transparency.

Each specialist pathway contributes explicit intermediate outputs to the shared workflow. Within the diagnostic agent A_D, MedGemma generates an initial image-conditioned

report draft R_t^0, while the Ark$^+$ detector produces structured findings F_t and Grad-CAM visual inspection cues H_t. A reporter sub-agent A_{rep} then merges, reviews, and refines these materials to produce the finalized report R_t. Within the consultation agent A_C, the current question is directed either to MedGemma-based visual support, producing image-grounded consultation information v_t, or to the knowledge pathway, which provides retrieved medical knowledge K_t. These materials support generation of the case-grounded conversational response m_t. Within the archive agent A_A, the prior report and stored findings are retrieved, finding-level differences are summarized as Φ_t, and the current and prior information is composed into the report-comparison output Δ_t. Across all three pathways, routing decisions, tool calls, intermediate outputs, and agent results are logged as the execution trajectory

$$\tau_t = \{e_1, e_2, \ldots, e_n\},$$

where each e_i denotes a recorded workflow event.

The architecture is hierarchical. In the implemented runtime, an "agent" refers to an LLM-based controller that interprets the current state and decides whether and how to invoke downstream tools or bounded subworkflows. Specifically, the agent-level reasoning and control components are powered by `gpt-4.1-mini`, while domain-specific models—including the fine-tuned Ark$^+$ classification configuration used as a CXR14-style abnormality detector [25] and MedGemma, a medical VLM [59]—function as specialist model tools rather than autonomous controllers. The reporter agent is an LLM-based sub-agent within the diagnostic pathway, whereas the detector and vision-language models provide

task-specific evidence used by that agent-controlled workflow.

Alternative LLM backbones were not systematically compared in this thesis, because the primary objective was to evaluate the proposed multi-agent workflow, detector-supported reporting pipeline, and transparency-oriented interaction rather than to benchmark different general-purpose LLMs. Fixing the controller model also reduced an additional source of variation across the system and user-study experiments. The dependence of agent behavior on the selected controller LLM is nevertheless a limitation of the present implementation and is discussed in Chapter 6.

3.2.1 Agentic Control Principle

Although TRACE contains several specialized models and services, its control logic is organized around a bounded ReAct-style interaction pattern rather than a single monolithic forward pass. At runtime, the active controller receives the current user request together with the available case context, interprets the request, selects an appropriate specialist tool or sub-agent, observes the returned result, updates the working state, and then either terminates or proceeds to the next step. The core interaction cycle can therefore be summarized as observe, reason, act, and update state.

This control pattern appears at multiple levels of the implemented system. At the top level, the orchestrator decides whether the current turn should be handled by the diagnostic, consultation, or archive workflow. Within the consultation workflow, a lightweight supervisory controller determines whether the current question requires image-

grounded visual support or knowledge-grounded retrieval. Within the archive workflow, a bounded agent coordinates prior-case retrieval, finding-level difference analysis, and report-comparison composition. Within the diagnostic workflow, a bounded supervisory controller coordinates validation, structured abnormality detection, draft generation, evidence reconciliation, and report refinement. TRACE is therefore not a collection of isolated tools, but a hierarchical workflow in which intermediate outputs become decision context for later stages.

At the workflow level, the diagnostic branch combines Ark$^+$-based structured abnormality detection, MedGemma-based report drafting, and reporter-agent review and refinement. The consultation branch combines image-grounded inspection with knowledge-grounded explanation. The archive branch retrieves prior report context and supports report-level interval comparison when a patient alias is available. The shared case/session state allows these branches to operate on the same evolving case context rather than as independent applications.

To make this state-mediated control pattern more concrete, Figure 3.2 traces one representative interaction turn through the diagnostic pathway. The example illustrates how user input is incorporated into the active case context, interpreted by the orchestrator, processed by a specialist agent, written back into the shared working state, and subsequently assembled into an inspectable user-facing response.

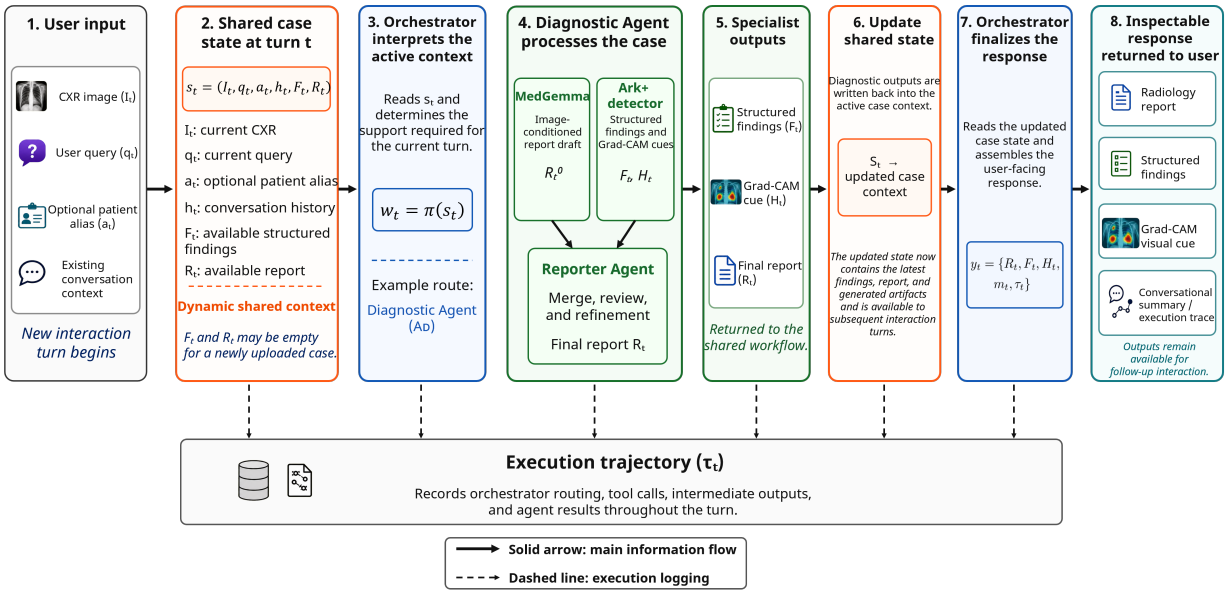


Figure 3.2: Illustrative single-turn execution through the TRACE diagnostic pathway. The active case context s_t is constructed from the current CXR, user query, optional patient alias, conversation history, and any available findings and report. The orchestrator selects the diagnostic agent A_D according to the routing policy $w_t = \pi(s_t)$. MedGemma and Ark$^+$ provide complementary report-drafting and structured visual evidence, which are reconciled by the reporter agent. The resulting findings, Grad-CAM cues, and finalized report are written back into the shared working context before the orchestrator assembles the user-facing response y_t. Dashed connections indicate events recorded in the execution trajectory τ_t. This figure shows one representative diagnostic route; consultation and archive turns follow the same state-mediated orchestration pattern with different specialist outputs.

In this example, the shared case-state update shown before response finalization represents an intra-turn working-state update: specialist outputs become available to the orchestrator before the current response is assembled. After the response is returned, these outputs remain available as context for subsequent interaction turns. Figure 3.2 therefore complements Figure 3.1: the latter shows the component-level organization of

the reasoning core, whereas the former shows the temporal information flow of one representative execution path. The diagnostic-specific stages are described in greater detail later in Section 3.4.

3.2.2 Design Rationale

The architecture was chosen not only for engineering convenience, but also to support the research questions of this thesis. Three design decisions are especially important: the use of a multi-agent workflow rather than a single end-to-end model, the use of a detector-supported report generation pipeline rather than direct report generation alone, and the use of Grad-CAM as a user-facing inspection aid.

First, TRACE adopts a multi-agent architecture because the target workflow is not limited to one output type. A user may upload an image for diagnostic assistance, ask a follow-up question about a visible abnormality, request clarification of medical meaning, or compare the current case with a prior report. These subtasks differ in evidential requirement, control logic, and response format. A single end-to-end model could in principle be prompted to handle all of them, but such a design would make routing harder to constrain, reduce task-specific controllability, and obscure workflow provenance. The multi-agent design instead separates orchestration from specialist execution. This improves modularity, makes individual pathways easier to inspect, and allows diagnostic, consultation, and archive functions to be evaluated as distinct components within one integrated system.

Second, the diagnostic workflow uses a detector-supported report-generation pipeline because direct multimodal report generation and structured abnormality detection offer complementary strengths. MedGemma provides radiology-style narrative fluency and broad image-conditioned description, but direct draft generation alone may omit findings, overstate uncertain observations, or fail to preserve a stable structured abnormality inventory. Although Ark$^+$ is introduced as a CXR foundation model with broader adaptability and extensibility, TRACE uses its fine-tuned classification configuration specifically as a structured abnormality detector. This detector provides explicit label-wise abnormality evidence, but it does not by itself generate a clinically readable radiology report. TRACE therefore combines these two roles: MedGemma supplies the draft narrative, Ark$^+$ supplies structured abnormality cues, and the reporter agent A_{rep} reconciles the two. This design also creates a meaningful progression of diagnostic configurations—draft-only, merge-only, and full review-enhanced reporting—that can later be evaluated as ablation conditions rather than being treated as an indivisible system block.

Third, Grad-CAM is included not as a claim of full model interpretability, but as a practical inspection mechanism. In this thesis, the purpose of the heatmap is to help users inspect whether a reported abnormality is associated with a plausible image region and to identify cases in which the visual evidence may warrant further confirmation. This is a narrower and more defensible role than treating heatmaps as definitive explanations of internal model reasoning. The transparency strategy of TRACE is therefore multi-level: Grad-CAM provides visual inspection cues, structured findings and report text provide textual evidence, and execution trajectories provide process-level provenance.

3.3 Orchestrator Agent

The orchestrator agent is the top-level coordinator of the system. It does not perform detailed medical reasoning by itself. Instead, it receives the current user input together with compact session state, decides which specialist pathway is required, invokes the relevant tool, and formats the returned outputs for frontend presentation.

3.3.1 LangGraph-Based Orchestration

The orchestrator is implemented as a LangGraph-based controller that follows a lightweight ReAct-style interaction pattern. For each turn, the orchestrator first observes the current user input together with compact session state, including recent message history, the current image reference when available, the latest report and findings, the patient alias, and selected archive outputs. It then reasons over the current turn to determine whether the request should be handled directly or delegated to a specialist pathway. When delegation is required, the orchestrator acts by invoking the corresponding specialist tool, receives the returned result, updates the shared session state, and either terminates the current turn or prepares the context for the next user interaction.

The active orchestrator graph contains an orchestrator node, a tool-execution node, and a handoff finalization node. This graph structure is intentionally compact. The agentic behavior lies not in a long chain of internal nodes, but in the state-aware decision process that determines whether the current turn requires diagnostic analysis, follow-up

consultation, archive retrieval, or direct non-medical response generation.

At each controller level, downstream tools are exposed through structured interfaces with task-specific inputs and outputs so that returned results can be written back into shared state in a reusable form. In practice, a turn terminates when the orchestrator can directly return a non-medical reply, when a specialist pathway has completed and produced a final response for frontend presentation, or when the system must wait for a new user turn after state update.

3.3.2 Routing Logic

The orchestrator follows an explicit routing policy. Uploaded image-analysis requests are routed to the diagnostic pathway. Medical follow-up questions are routed to the consultation pathway. Requests about previous studies or interval change are routed to the archive pathway. Non-medical casual conversation can be answered directly.

This routing policy is important because it separates coordination from specialist reasoning. The orchestrator does not answer medical questions from generic language knowledge when a grounded specialist pathway is available. The routing policy therefore improves control and maintains a clearer connection between user requests and backend evidence sources. Although the routing policy is explicit, the decision is still taken in the context of the current turn and shared session state rather than by a purely keyword-based router, which allows the orchestrator to preserve conversational continuity across upload, diagnosis, follow-up, and archive-aware turns.

3.3.3 Session State and Continuity

The orchestrator maintains a lightweight shared case/session state across turns. The stored state includes recent message history, session identifier, current image path, latest report snippet, latest findings, patient alias, selected language, and archive-related outputs. The stored state allows the user to upload an image, receive a report, ask follow-up questions, and request archive-assisted comparison within one continuous session.

In the current implementation, this session state is maintained in lightweight runtime memory rather than in a persistent conversation database. The orchestrator also trims state before each model call. This trimming step preserves the most relevant conversation turns and diagnostic outputs while avoiding unnecessary prompt growth. The trimming policy improves stability during multi-turn interaction and keeps the orchestration process focused on decision-relevant context.

3.4 Diagnostic Agent

The diagnostic agent A_D is the core image-analysis workflow of the system. Its role is to transform an uploaded image into structured findings, Grad-CAM overlays, and a radiology-style report through a staged, tool-mediated pipeline with bounded supervisory control. The workflow is designed not as a single report-generation step, but as a progressive evidence-integration process in which CXR validation, structured abnormality detection, multimodal draft generation, and report refinement play distinct roles.

Figure 3.3 summarizes this workflow. At a high level, the pipeline begins with CXR validation and diagnostic state initialization. If the uploaded image is not a chest radiograph, the system terminates the diagnostic path and prompts the user to upload an appropriate CXR. For a valid CXR, the workflow proceeds through two complementary evidence channels: MedGemma generates the initial report draft, while the fine-tuned Ark$^+$ detector performs structured abnormality detection and Grad-CAM generation. These outputs are then reconciled through merge, review, and optional rewrite stages. If prior report context is available, interval comparison can be appended after diagnostic report refinement. This staged design also defines the internal reporting configurations later used in the diagnostic-pipeline ablation.

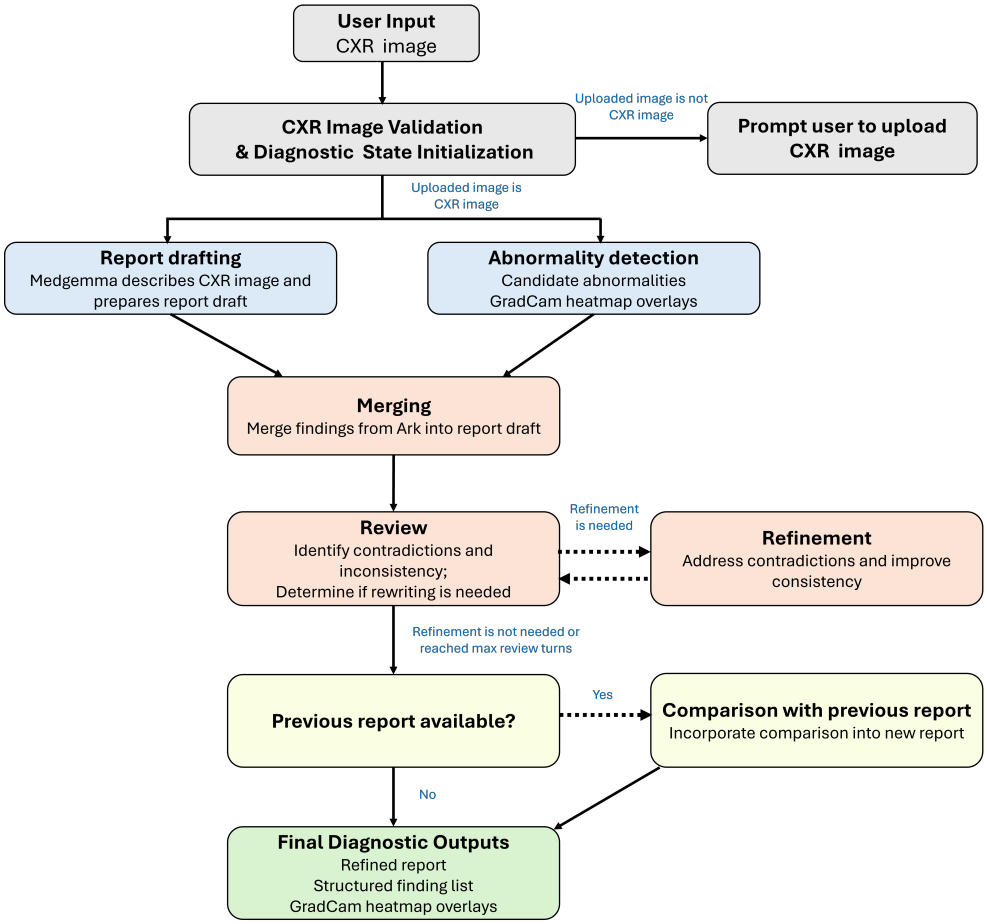


Figure 3.3: Diagnostic agent workflow.

3.4.1 Diagnostic Supervisor and Internal State

The diagnostic pathway is governed by a bounded supervisory controller that manages the internal diagnostic state and coordinates stage-to-stage handoff. The internal diagnostic state stores the input image path, output directory, language, CXR validation status, abort

flags, MedGemma draft content, Ark$^+$ findings, Grad-CAM artifacts, logging variables, and the current report text. This state is not merely a storage object; it also serves as the decision context for subsequent stages. For example, Ark$^+$ is executed only if the validation stage confirms that the uploaded image is a chest radiograph, and the reporting stage is entered only after the upstream evidence sources have been collected.

The supervisory controller does not perform unconstrained open-ended planning. Instead, it operates within a bounded action space defined by the diagnostic workflow: validate the image, run structured detection, generate or normalize a draft report, merge complementary evidence, review the merged report, and rewrite only when inconsistencies are detected. This bounded design is particularly important in the diagnostic setting, where unrestricted tool exploration would increase variability without improving the evidential structure required for report generation. The result is a diagnostic controller that preserves the benefits of agentic coordination while keeping the workflow stable, interpretable, and reproducible for evaluation and prototype operation.

3.4.2 CXR Validation and Report Drafting

The first diagnostic stage uses MedGemma to determine whether the uploaded image is a CXR. This validation step prevents the downstream diagnostic workflow from processing arbitrary uploaded images as if they were valid chest radiographs. A non-CXR input triggers an abort condition, returns a neutral description of the detected image type, and prompts the user to upload a CXR. The diagnostic workflow stops at that point.

For a valid CXR, MedGemma also generates the initial report draft. The draft covers image quality or technique when relevant, cardiomediastinal assessment, pulmonary findings, pleural findings, visible osseous or soft-tissue observations, and an impression summary. The MedGemma stage therefore serves two purposes within the workflow: scope validation and preliminary narrative report generation. The prompting also constrains laterality description to the patient perspective rather than the displayed-image perspective, which reduces orientation-related ambiguity during visual report drafting.

This stage also defines the simplest report-generation condition used in the diagnostic pipeline. When used alone, the MedGemma output functions as a direct draft-generation baseline without explicit detector fusion or subsequent report review. In the full TRACE workflow, however, this draft is not treated as the final report. Instead, it serves as the narrative starting point that will later be checked and revised against Ark^{+}-derived structured abnormality evidence.

The implementation prompts MedGemma to return structured JSON rather than unconstrained free text. The returned fields include whether the image is a CXR, a short image-type label, a brief neutral description for out-of-scope uploads, a report draft for valid CXRs, and an abort reason for invalid inputs. The diagnostic code normalizes the returned content and applies fallback parsing when the JSON is imperfect so that the MedGemma stage behaves as a stable workflow component rather than as a free-form conversational model call.

3.4.3 Ark$^+$-Based Abnormality Detection

Ark$^+$ is an open foundation model for chest radiography that was developed to support adaptation across multiple CXR-related diagnostic tasks [25]. In TRACE, however, Ark$^+$ is used in a narrower and explicitly defined role. The system uses a fine-tuned Ark$^+$ classification configuration as a CXR14-style structured abnormality detector, rather than using Ark$^+$ as a segmentation, localization, or general reasoning module. Its function is to convert a validated CXR into machine-readable pathology scores that can be consumed by downstream report synthesis, Grad-CAM generation, and transparency-oriented display.

In the implemented pipeline, the fine-tuned Ark$^+$ detector operates in the CXR14 label space and predicts 14 thoracic abnormality categories: Atelectasis, Cardiomegaly, Effusion, Infiltration, Mass, Nodule, Pneumonia, Pneumothorax, Consolidation, Edema, Emphysema, Fibrosis, Pleural Thickening, and Hernia. For each uploaded CXR, it outputs a probability for every label. These probabilities are then converted into binary abnormality decisions by label-specific operating thresholds, which determine which findings are retained in the structured output and which labels proceed to the Grad-CAM stage.

The Ark$^+$ checkpoint used in this thesis is a fine-tuned classification model for ChestX-ray14-style abnormality prediction. Before it was integrated into the full multi-agent workflow, its classification quality was examined on the test split of ChestX-ray14. Because the purpose of this evaluation was to assess its suitability as a structured evidence source, the main metric reported here is area under the receiver operating characteristic curve (AUC-ROC). AUC-ROC is appropriate in this context because it measures discriminative

ranking quality independently of any specific operating threshold. On the ChestX-ray14 train-val split, the fine-tuned Ark$^+$ detector achieved a mean AUC-ROC of 0.885 across the 14 labels. On the official test split, the mean AUC-ROC was 0.845. As shown in Table 3.1, the detector performed particularly strongly on Cardiomegaly, Edema, Emphysema, Pneumothorax, and Hernia, while labels such as Infiltration, Pneumonia, and Consolidation remained more challenging. These results indicate that the fine-tuned Ark$^+$ detector provides useful label-wise discriminative signals for structured finding extraction, while also showing that some labels require cautious downstream use.

Table 3.1: Label-wise AUC-ROC of the fine-tuned Ark$^+$ detector on the NIH ChestX-ray14 train-val and test splits. Mean values are macro-averaged across the 14 labels.

Class	Train-val AUC-ROC	Test AUC-ROC
Atelectasis	0.877	0.812
Cardiomegaly	0.937	0.905
Effusion	0.915	0.845
Infiltration	0.748	0.718
Mass	0.912	0.858
Nodule	0.860	0.827
Pneumonia	0.814	0.760
Pneumothorax	0.927	0.922
Consolidation	0.847	0.777
Edema	0.935	0.867
Emphysema	0.947	0.947
Fibrosis	0.876	0.860
Pleural_Thickening	0.857	0.816
Hernia	0.938	0.922
Mean	0.885	0.845

3.4.4 Operating Threshold Determination for Deployment

Although AUC-ROC is suitable for evaluating the discriminative ability of the fine-tuned Ark$^+$ detector, the deployed diagnostic workflow requires binary abnormality decisions rather than probabilities alone. These decisions determine which findings are included in

the structured result list, which labels trigger Grad-CAM generation, and which abnormalities are passed to the downstream report-synthesis stage. The integration problem is therefore not only whether Ark$^+$ can rank positive cases above negative cases, but also how to define a practical operating point for each label.

ChestX-ray14 is provided as a train-val partition and a separate test partition, without an additional standalone validation split. In the present work, the train-val split was therefore used as an in-domain calibration set for deployment. For each of the 14 labels, Ark$^+$ probabilities were scanned over candidate thresholds, and the threshold that maximized the Matthews correlation coefficient (MCC) was selected. MCC was chosen because it provides a balanced measure of binary classification quality under class imbalance by jointly accounting for true positives, true negatives, false positives, and false negatives. This property is particularly relevant for ChestX-ray14, where prevalence varies substantially across pathology categories.

The resulting per-label thresholds were fixed and stored in the deployment configuration used by the Ark$^+$ detector service. In other words, the threshold-selection procedure did not alter Ark$^+$ parameters or retrain the model; it only defined how continuous label probabilities were converted into binary findings inside the proposed system. This distinction is important. The AUC-ROC values in Table 3.1 demonstrate that the fine-tuned Ark$^+$ detector produces useful discriminative signals, whereas the calibrated thresholds in Table 3.2 define the operating rule that allows those signals to be consumed by the diagnostic workflow in a stable and reproducible manner.

Table 3.2: Per-label operating thresholds used by the deployed fine-tuned Ark$^+$ detector. Thresholds were selected on the NIH ChestX-ray14 train-val split by maximizing MCC.

Class	Threshold	MCC
Atelectasis	0.219	0.431
Cardiomegaly	0.137	0.393
Effusion	0.268	0.545
Infiltration	0.251	0.290
Mass	0.211	0.482
Nodule	0.164	0.379
Pneumonia	0.044	0.155
Pneumothorax	0.185	0.465
Consolidation	0.066	0.259
Edema	0.151	0.326
Emphysema	0.218	0.480
Fibrosis	0.099	0.233
Pleural_Thickening	0.137	0.266
Hernia	0.188	0.445

3.4.5 Grad-CAM-Based Visual Inspection

To support visual inspection of the retained abnormality predictions from the fine-tuned Ark$^+$ detector, the diagnostic workflow generates class-specific Grad-CAM heatmaps for labels that remain positive after thresholding. These heatmaps are not used as pixel-level lesion segmentations. Instead, they indicate image regions that contributed strongly to the

detector's prediction for a given abnormality label.

Because the fine-tuned Ark$^+$ detector uses a Swin-based vision backbone rather than a conventional convolutional architecture, the Grad-CAM procedure is applied to the final normalized token features rather than to a late convolutional feature map. For each target label, the image is first passed through Ark$^+$ to obtain multi-label logits and probabilities. Grad-CAM is then computed only for the labels retained for explanation. Let y^c denote the logit for class c, and let A_k denote the kth channel of the selected internal feature representation. The gradient of y^c with respect to these features is first computed, and the gradients are then averaged over the spatial or token dimension to obtain a class-specific importance weight for each channel:

$$\alpha_k^c = \frac{1}{Z} \sum_i \frac{\partial y^c}{\partial A_{k,i}},$$

where i indexes spatial locations or tokens and Z is the number of such positions. These weights are used to form a weighted combination of the corresponding feature channels:

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k \alpha_k^c A_k\right).$$

The ReLU operation keeps only positive evidence for the target class, so the resulting map highlights regions that support the abnormality prediction rather than regions that suppress it.

The resulting class activation map is normalized and upsampled to the image resolution

before being overlaid on the chest radiograph as a color heatmap; an example is shown in Figure 3.4. Each positive abnormality prediction is paired with a visual inspection map that can be reviewed together with the original image. Since Ark$^+$ is used here as a multi-label detector, this process is performed independently for each positive label, producing separate heatmaps for different abnormalities for one CXR. Heatmaps with a large fraction of activation concentrated outside the thoracic region can be treated as a red flag that warrants further confirmation.

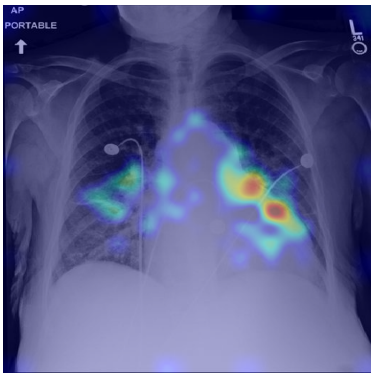


Figure 3.4: An example of a CXR with a Grad-CAM overlay.

The Grad-CAM heatmaps provide the diagnostic workflow with a visual inspection channel that is more informative than a raw probability alone while remaining consistent with the role of Ark$^+$ in the system. They are therefore used as inspection aids that complement the structured findings and final report.

3.4.6 Report Synthesis, Review, and Refinement

The final diagnostic stage is coordinated by the reporter agent A_{rep}, which reconciles the MedGemma report draft and the structured Ark$^+$ findings into a finalized radiology-style report. This stage is implemented as a bounded merge–review–rewrite loop and is central to the detector-supported report-generation pipeline proposed in this thesis. Rather than treating report generation as a single model output, the workflow explicitly separates narrative drafting, structured evidence integration, and post-merge quality control.

The merge step inserts Ark$^+$-positive findings into the MedGemma draft with minimal disruption to report structure. This merged output is itself a meaningful intermediate condition: it reflects detector-supported report generation without an explicit review stage. In other words, the merge stage defines a stronger configuration than draft-only generation, but it is not yet the full TRACE reporting pipeline.

The review step then examines the merged report for contradictions, duplication, omission of structured positives, unsupported statements, and other internal inconsistencies. The purpose of this stage is not merely stylistic polishing. Instead, it functions as a quality-control mechanism that checks whether the merged report preserves the narrative fluency of the MedGemma draft while remaining consistent with the structured abnormality evidence returned by Ark$^+$. The review output is written back into the reporter state as rewrite guidance rather than being treated as an external comment.

When the review stage identifies unresolved issues, the rewrite step revises the merged report accordingly. The cycle is bounded by a small maximum number of review rounds,

which keeps the reporting behavior predictable and prevents uncontrolled iterative rewriting. The finalize step returns the accepted report in plain text. The full TRACE report-generation pipeline therefore consists of three conceptually distinct reporting stages: draft generation, detector-supported merging, and review-guided refinement.

This staged design is important for two reasons. From a system perspective, it produces a more stable and auditable report-generation workflow than direct generation alone. From an evaluation perspective, it defines a set of progressively richer report-generation conditions that can be compared later: a draft-only condition, a detector-supported merge condition, and a full review-enhanced condition. These conditions make it possible to examine not only whether TRACE performs well as a whole, but also which parts of the reporting pipeline contribute to report-quality improvements.

The consistency audit is especially important when Ark$^+$ and MedGemma do not fully agree. The reporting stage separates overlapping findings, contradictions, and caution points. Agreed findings are written directly into the report. Disputed Ark$^+$-positive findings are retained but phrased cautiously when the evidence is not fully aligned. The reporting stage therefore combines MedGemma's narrative fluency with Ark$^+$-derived structured abnormality evidence and produces a report that is more suitable for use within the proposed assistive workflow than either source alone.

When a previous report is available through the archive workflow, interval-change information is appended after report refinement has completed. The final diagnostic outputs consist of the refined report, the structured finding list, and the Grad-CAM heatmap

overlays.

3.5 Consultation Agent

The consultation agent A_C handles follow-up questions after the diagnostic pathway has produced a report, a structured finding list, and Grad-CAM overlays. It receives the current user question together with the active case context, including the latest report, the latest findings, the current image reference when available, and selected recent conversation turns. The consultation agent then generates a case-aware response that remains grounded in the current study rather than drifting into generic medical discussion.

Figure 3.5 summarizes the internal workflow of the consultation agent. The workflow begins with the current user question together with the active case context, including the latest report, findings, and image reference when available. The consultation agent then determines whether the question requires image-grounded reinspection of the current CXR or knowledge-grounded retrieval beyond direct visual evidence. Image-grounded questions are routed to the MedGemma-based visual inspection path, whereas knowledge-grounded questions are routed to the OpenClaw-backed retrieval path. In both cases, the returned evidence is converted into a case-aware response and written back into the active session context.

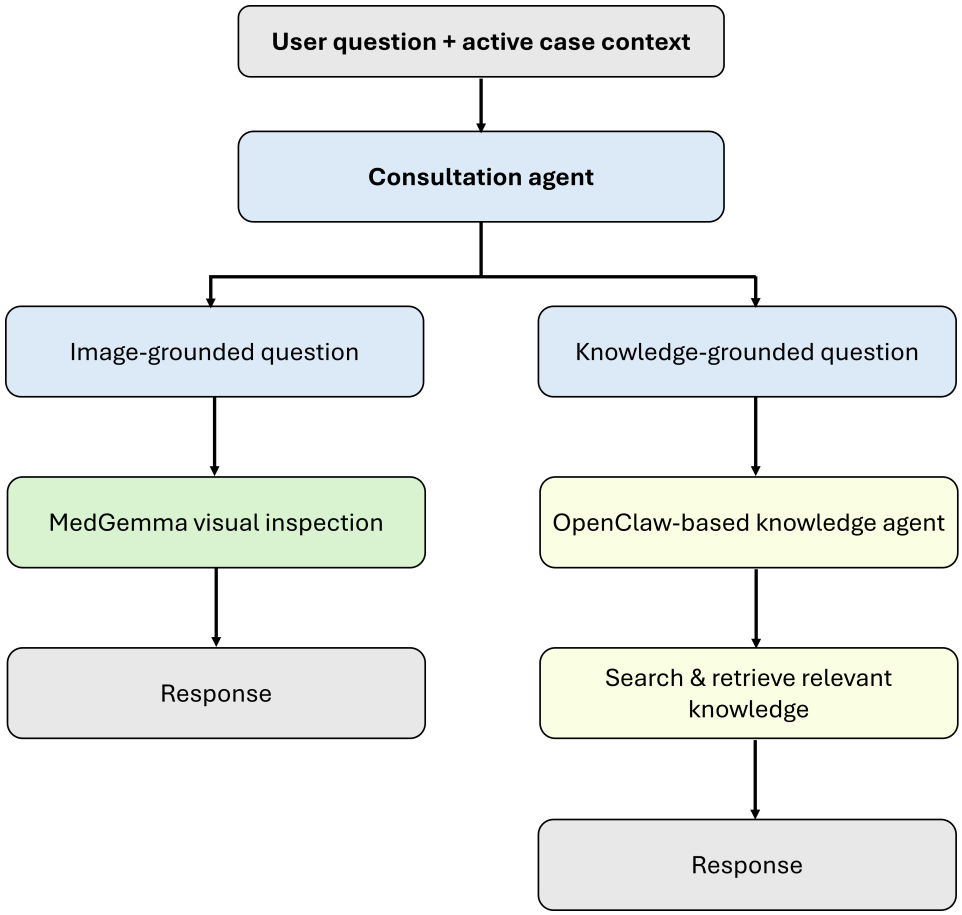


Figure 3.5: The workflow of the consultation agent.

The consultation agent is implemented as a lightweight LangGraph-based ReAct agent with one agent node and one tool node. For each follow-up turn, the agent first observes the current user question together with the active case context, including the latest report, the current findings, the image reference when available, and selected recent conversation history. It then reasons about the evidential requirement of the question: whether the answer can be grounded by reinspection of the current image or whether external medical

knowledge is required. Based on that decision, the agent acts by invoking either the MedGemma visual tool or the OpenClaw-backed knowledge retrieval pathway, receives the returned evidence, and converts that evidence into a coherent case-aware response. The resulting answer is then written back into session state so that later turns can build on the same consultation context.

The consultation pathway does not require a long internal pipeline, but it still benefits from explicit state-aware tool selection because follow-up questions are often ambiguous when taken in isolation and become answerable only in the context of the current image, report, and preceding dialogue.

The image-grounded path is used when the user asks about a specific region or visible abnormality in the uploaded image. In this path, the consultation agent invokes the MedGemma-based visual tool and reuses the current image together with the latest diagnostic context. The returned answer remains tied to the current case and supports direct inspection of image-level evidence.

The knowledge-grounded path is used when the user asks for clinical meaning or background knowledge beyond direct image inspection. In this path, the consultation agent invokes an OpenClaw-backed knowledge retrieval pathway rather than a standalone free-form answer generator. The retrieval process can draw on selected online medical sources together with configured local knowledge resources. After retrieval, OpenClaw returns a structured evidence summary that the consultation agent uses to generate the final response.

In the current implementation, retrieval is constrained to selected source ranges rather than unrestricted open-domain search. The preferred source space includes high-quality medical resources and curated CXR knowledge materials. The consultation response can therefore remain evidence-grounded while still being generated in natural language and adapted to the user's question.

After the selected path completes, the consultation agent returns the answer to the orchestrator. The orchestrator formats the answer for frontend presentation and preserves the result as part of the active session state. The consultation pathway therefore extends the system from report generation to sustained, case-specific explanation.

3.6 Archive Agent

The archive agent A_A supports archive-assisted cross-study comparison when a patient alias is available. It retrieves the most recent archived case associated with the alias, compares the archived findings against the current findings, generates a plain-text comparison summary, and passes the comparison result back to the active case workflow. The archive pathway therefore adds prior-report context to the current diagnostic session without introducing a separate archive application.

Figure 3.6 summarizes the internal workflow of the archive agent. The workflow begins with the patient alias together with the active case context. The archive agent first determines whether a prior case is available under the current alias. If no prior case is

found, the archive workflow is skipped. If a prior case is available, the system retrieves the previous report and stored findings, compares them with the current case, and generates an interval-change summary for report integration or direct history review.

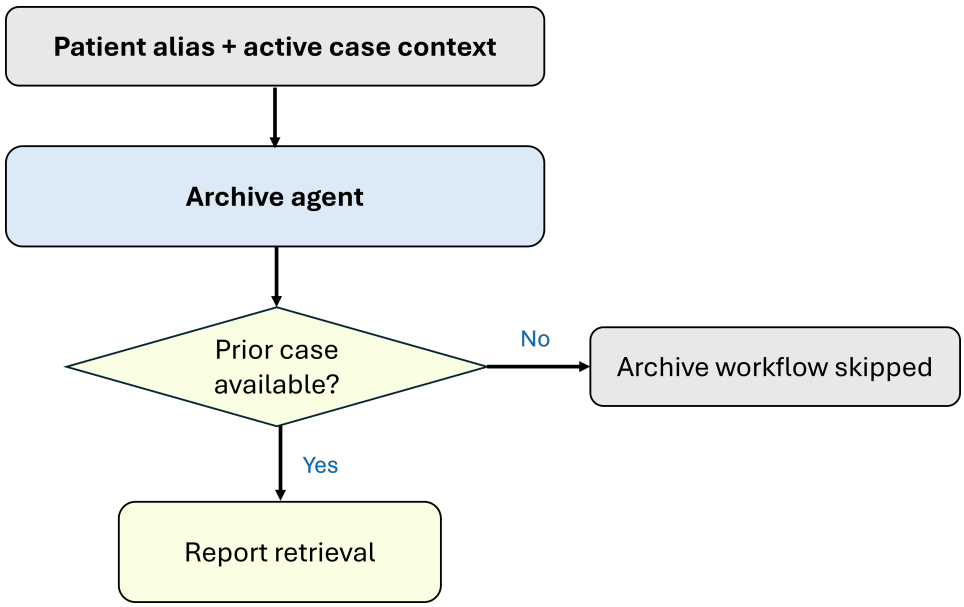


Figure 3.6: Archive agent workflow.

The archive agent is implemented as a lightweight LangGraph-based ReAct agent with one agent node and one tool node. For an archive-aware turn, the agent observes the current patient alias together with the active case findings, retrieves the most recent archived entry associated with that alias, and then reasons over the returned prior context to summarize interval change. The action space is intentionally narrow: retrieve prior case context, compare prior and current findings, and generate a plain-text summary suitable for insertion into the current report or for direct history review. This bounded design keeps the archive

pathway interpretable while still allowing the comparison summary to be generated from the retrieved case context rather than from a fixed template alone.

The comparison logic is intentionally lightweight and label-driven. The archive agent does not perform image registration or fine-grained temporal progression modeling. It summarizes whether findings persist, newly appear, or resolve across cases. This scope is sufficient for practical longitudinal comparison within the current research prototype.

Persistence is handled outside the archive agent by deterministic program logic. After the comparison summary is generated, the persistence routine creates a new alias-specific archive entry, copies the current image, saves the finalized report, and writes a metadata file containing the current labels, the comparison summary, and prior-study linkage. This separation is important because it distinguishes reasoning from record creation: interval-change interpretation remains agentic, whereas archive persistence is handled deterministically to preserve reproducibility and data consistency.

3.7 Frontend User Interface and Multi-Level Transparency

The user-facing interface provides the primary interaction environment of the proposed system. It presents a continuous consultation setting in which image upload, report review, follow-up questioning, and archive-aware interaction can occur within the same case session. This design keeps the user workflow coherent while preserving the multi-agent structure of the underlying orchestration layer.

The interface serves two related objectives. The first is usability: diagnostic review and follow-up interaction are organized within one continuous environment rather than being split across separate tools. The second is transparency: the interface is designed not merely to display final outputs, but to expose multiple forms of evidence generated by the system. Rather than relying on a single report or answer, the interface presents visual evidence, source-linked explanatory evidence, and workflow provenance in parallel so that users can inspect both what the system produced and how it produced it.

3.7.1 Multi-Level Transparency Design

The frontend implements transparency through three coordinated channels: visual transparency, knowledge-source transparency, and process transparency.

Visual transparency is provided through Grad-CAM overlays aligned with the original CXR. This layer allows the user to inspect whether a predicted abnormality is associated with a plausible image region.

Knowledge-source transparency is provided through an explicit source-reporting layer in consultation responses. When the consultation pathway relies on external knowledge, the system exposes the source block associated with the returned evidence rather than presenting the answer as unsupported free text. In the current implementation, this source space is constrained to selected high-quality medical resources and curated CXR knowledge materials, which allows users to inspect where explanatory content came from and supports source-aware verification when needed.

Process transparency is provided through the trajectory view, which exposes workflow-level provenance rather than hidden free-form internal reasoning. The trajectory view records how the current request moved through the orchestrator and specialist branches, including diagnostic processing, consultation routing, and archive-related stages when applicable. This makes it possible to inspect not only the final output but also the sequence of tool-stage decisions that produced it.

Figure 3.7 shows an example of trajectory content. In this case, the trajectory view exposes contradictions identified during the report-merging stage, illustrating how process transparency can reveal not only which components were used but also what kinds of internal consistency issues were detected during report refinement.

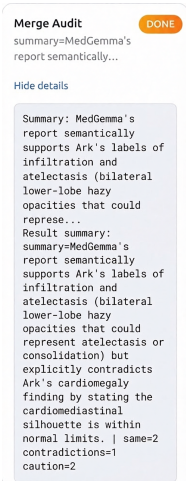


Figure 3.7: An example of trajectory content.

By presenting visual evidence, source-linked explanatory evidence, and workflow provenance within the same interface, the frontend supports inspection of both what the system

produced and how it produced it.

3.7.2 Interaction Structure

The interaction model is organized around a 3D virtual clinic scene. The application begins in an exploration view that presents a 3D clinical environment with a doctor avatar. The main operational workflow then takes place in consultation mode, where the system overlays the conversational and analytic interface on top of the 3D scene. The 3D environment provides a consistent interaction setting, whereas the consultation overlay supports image upload, report inspection, and follow-up dialogue. As shown in Figure 3.8, the consultation overlay contains two coordinated workspaces. The left workspace is the primary conversation channel. The right workspace is the analysis surface used for report reading, image inspection, and trajectory review. This arrangement separates dialogue from technical inspection while keeping both views visible within the same session. The interface also supports speech-based response presentation and avatar animation. This multimodal layer makes follow-up interaction more conversational while preserving the same underlying diagnostic, consultative, and transparency-related content.

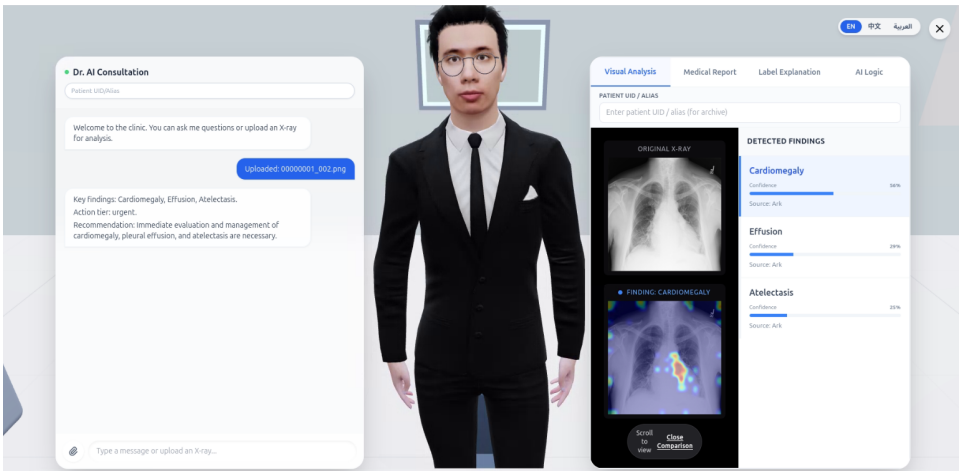


Figure 3.8: The user interface of TRACE.

The conversation side of the interface, shown in Figure 3.9, integrates text chat, image upload, and optional patient-alias input. This unified input design supports both image-centered diagnosis and text-based follow-up within the same session. The patient alias also functions as a lightweight continuity key, allowing the current case to be associated with archived prior entries when longitudinal context is available.

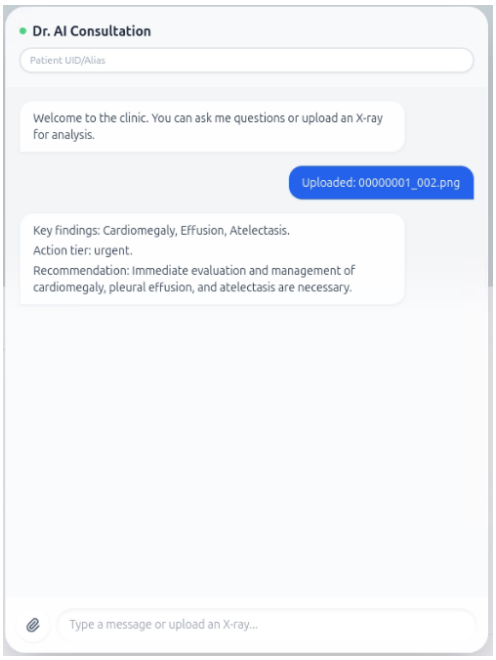


Figure 3.9: The conversation workspace of the user interface.

3.7.3 Analysis Workspace

The analysis workspace is implemented as a tabbed inspection surface that operationalizes the transparency design described above. It provides four coordinated views: visual analysis, medical report, label explanation, and AI logic.

The visual analysis view, shown in Figure 3.10, displays the uploaded CXR together with the Grad-CAM heatmap overlays generated by the diagnostic workflow. The interface supports side-by-side comparison between the original image and the selected heatmap and presents the structured finding list with confidence indicators. This view links abnormality

labels to image regions and supports direct inspection of model-linked visual evidence.

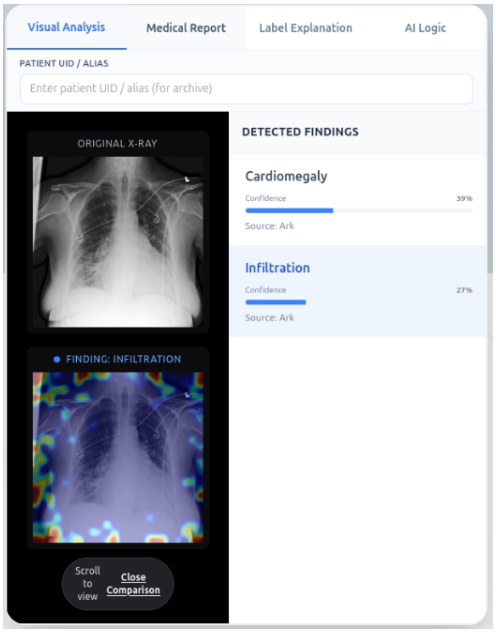


Figure 3.10: The visual analysis view of the analysis workspace.

The medical report view, shown in Figure 3.11, displays the refined radiology-style report returned by the backend. This view preserves the report as a primary diagnostic artifact and provides a focused reading surface separate from the conversational chat stream.

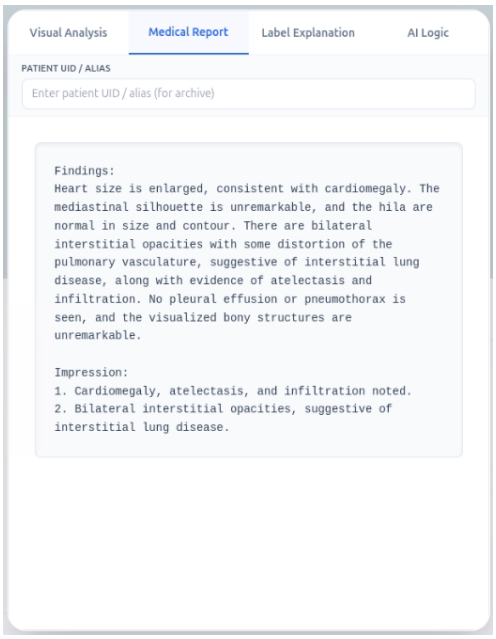


Figure 3.11: The medical report view of the analysis workspace.

The label explanation view, shown in Figure 3.12, provides an additional interpretive aid rather than a direct evidence channel. By pairing common findings with concise explanations and example images, it helps users translate model outputs into more understandable clinical concepts.

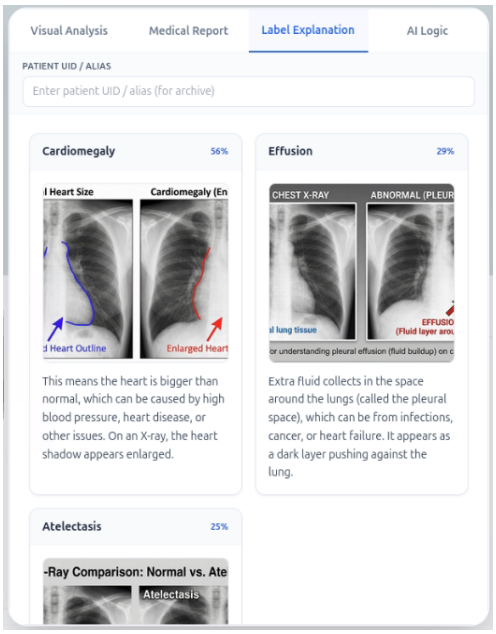


Figure 3.12: The label explanation view of the analysis workspace.

The AI logic view, shown in Figure 3.13, renders the execution trajectory collected from the backend. The trajectory component transforms raw trace events into a stepwise sequence that represents the progression of the case through the orchestrator and specialist branches. This view therefore supports process-level inspection in addition to image review and report reading.

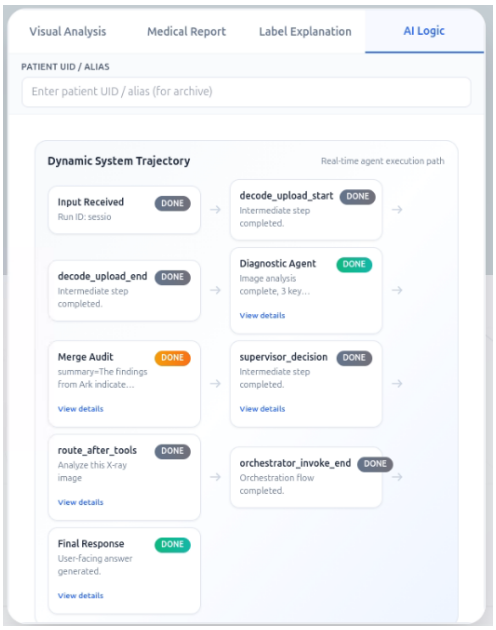


Figure 3.13: The AI logic view of the analysis workspace.

3.8 Chapter Summary

This chapter presented the implemented architecture and workflow of TRACE. The system is organized around a centralized orchestrator, a staged diagnostic workflow, a consultation pathway that combines image-grounded and knowledge-grounded follow-up support, and an archive pathway that supports report-level cross-study comparison. The chapter also explained the main design choices underlying the system, including the use of a multi-agent workflow, the detector-supported report-generation pipeline, and Grad-CAM-based visual inspection support.

The next chapter describes how these components were evaluated through objective report-quality analysis, transparency-supported verification and consultation-preparation tasks, archive-assisted comparison experiments, and questionnaire-based human-factors assessment.

Chapter 4. Evaluation

4.1 Evaluation Overview

This chapter evaluates TRACE from four complementary perspectives: objective report quality, transparency-supported verification awareness and consultation preparation, archive-assisted cross-study comparison, and subjective workload and usability. The evaluation includes one automatic report-generation experiment with an internal diagnostic-pipeline ablation analysis, two controlled user studies, and a questionnaire-based human-factors assessment.

The evaluation design follows the research questions introduced in Chapter 1. Experiment I examines whether the detector-supported report-generation pipeline improves objective report quality relative to external baselines and simpler internal reporting conditions. Experiment II examines whether the transparency-oriented interface helps users identify conclusions that warrant clinician verification and formulate more specific follow-up consultation questions. Experiment III examines whether archive-assisted comparison improves report-level comparison accuracy and efficiency. The questionnaire component assesses whether the overall *text-only* and *full TRACE* interface modes differ in perceived usability and workload.

The evaluation also examines component contribution at both pipeline and workflow levels. Experiment I provides a diagnostic-pipeline ablation by removing detector-supported finding integration and review-enhanced refinement in stages. Experiment II evaluates the transparency-oriented interface by comparing *text-only* and *full TRACE* conditions, while Experiment III evaluates archive support by comparing report-level comparison with and without archive assistance.

4.2 Study Design, Materials, and Outcome Measures

For Experiment I, a mixed report-generation benchmark was constructed by randomly sampling 1,000 frontal CXR images from the MIMIC-CXR dataset [60] and 1,000 cases from the Indiana University Chest X-rays dataset [61] that contained at least one frontal image, from which the frontal image was retained for evaluation. This yielded a 2,000-image frontal-CXR evaluation set with paired reference reports. The external comparison included LLaVA-Med, the MedRAX report-generation tool, and the full TRACE report pipeline. The internal ablation compared three TRACE reporting conditions: *report draft only*, *draft merged with detector findings without review*, and *full TRACE report pipeline*. The full TRACE report pipeline served as the shared reference condition for both analyses and was evaluated once on the same benchmark. The *report draft only* condition used the initial MedGemma draft directly as the final report. LLaVA-Med and the *report draft only* condition were given the same report-generation prompt, whereas the MedRAX tool directly accepted the CXR as input.

For Experiment II, two single-image CXR cases were prepared in advance and evaluated under a within-subject, counterbalanced design with two interface conditions: *text-only* and *full TRACE*. For Experiment III, two matched report pairs, denoted Pair X and Pair Y, were prepared for report-level comparison under a within-subject, counterbalanced design with two conditions: *without-archive* and *with-archive*. These deliberately limited stimulus sets supported a controlled and counterbalanced prototype evaluation, but were not intended to represent the full diversity of CXR findings, report styles, or longitudinal comparison scenarios.

The questionnaire-based assessment focused on two overall interface modes: a *text-only* interface and the *full TRACE* interface. The *text-only* interface provided textual interaction and textual outputs but did not expose the additional multimodal and transparency-oriented components of TRACE, such as Grad-CAM overlays, structured visual evidence, execution trajectory, or archive-assisted comparison views. The *full TRACE* interface included these additional components. Usability was assessed with the System Usability Scale (SUS), and perceived workload was assessed with Raw NASA Task Load Index (Raw TLX), including both the overall score and the six TLX dimensions.

The outcome measures were selected to match the goal of each experiment. Experiment I used SEMB, RadCliQ-v1, sample-wise F1, sample-wise positive-only F1, case-level false-positive rate (case-level FPR), case-level false-negative rate (case-level FNR), micro F1, and macro F1. In this thesis, sample-wise metrics were computed at the case level, whereas micro F1 and macro F1 were computed from aggregated CheXbert-derived positive observation labels. Experiment II used verification-worthy conclusion

identification score, risk-basis justification score, consultation question specificity score, and the composite *Consultation Preparation Quality Score*. Experiment III used common-abnormality-identification score, difference-identification score, and completion time. The questionnaire-based assessment used SUS and Raw TLX.

For the case-level FPR and FNR, the CheXbert-derived multi-label observations for each report were additionally collapsed into a binary case-level status. A report was assigned a negative case status if none of the evaluated abnormality labels was positive, and a positive case status if at least one evaluated abnormality label was positive. This rule was applied independently to both the reference report and the generated report. The case-level FPR was defined as the proportion of cases with a negative reference status for which the generated report received a positive status. Conversely, the case-level FNR was defined as the proportion of cases with a positive reference status for which the generated report received a negative status. These measures capture case-level over-calling and complete abnormality omission, respectively; they do not assess whether the specific generated abnormality labels match the reference labels, which is instead reflected by the F1-based metrics.

4.2.1 Participants and Recruitment

A total of 48 participants took part in the user-study components, including Experiment II, Experiment III, and the questionnaire-based workload and usability assessment. Participants were recruited from a general adult population rather than from radiology specialists,

because the purpose of the user studies was to evaluate how the system supports report understanding, verification-oriented follow-up questioning, and report-level comparison for non-expert users.

The participants ranged in age from 20 to 65 years. The sample included 23 male participants and 25 female participants. Participants came from diverse professional and educational backgrounds, including information technology, engineering, education, finance, media, the arts, government affairs, and retirement. The demographic distributions of participants are shown in Figure 4.1 and Figure 4.2.

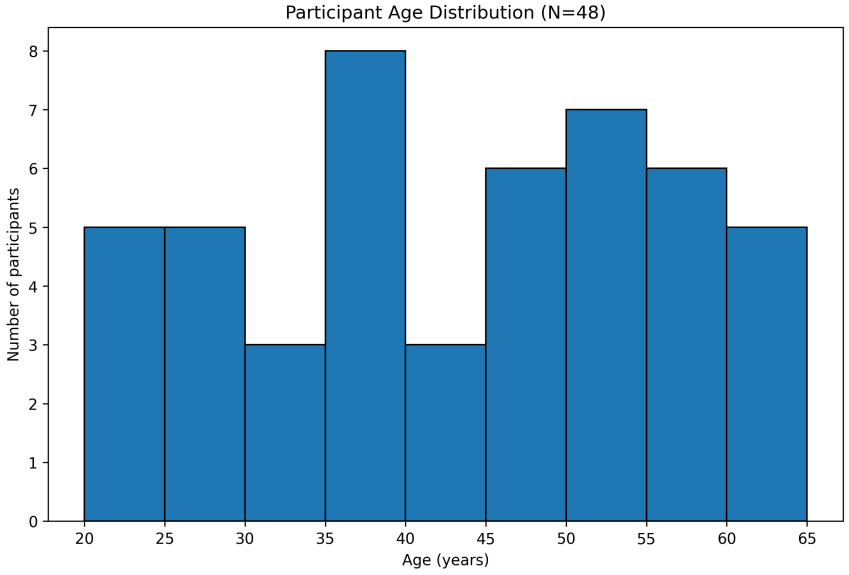


Figure 4.1: The age distribution of participants.

Before beginning the study, all participants received a brief introduction to the TRACE system and the task instructions. No specialist clinical training was required. The user-study components followed a within-subject, counterbalanced design, such that each par-

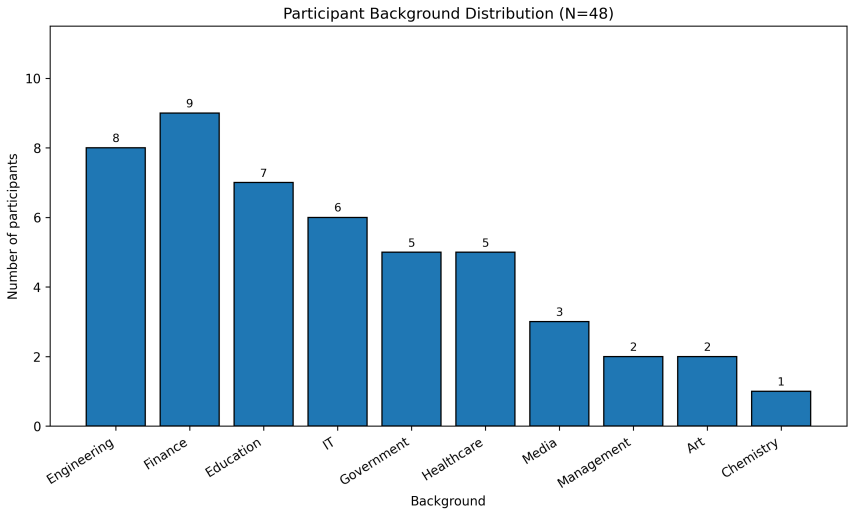


Figure 4.2: The background distribution of participants.

ticipant experienced both relevant interface conditions across the assigned tasks. Figure 4.3 summarizes the overall workflow of the user-study components. After the introduction and task briefing, participants completed Experiment II and Experiment III, followed by retrospective condition-specific SUS and Raw TLX ratings for the *text-only* and *full TRACE* interface modes.

4.3 Baseline Justification

The baselines in Experiment I were divided into external comparison systems and internal diagnostic-pipeline conditions.

The external comparison systems were LLaVA-Med and the MedRAX report-generation tool. LLaVA-Med was included as a general biomedical multimodal report-generation

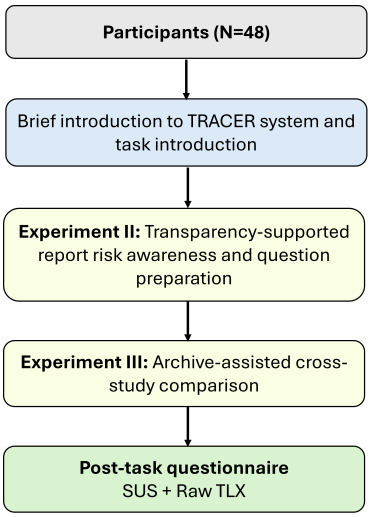


Figure 4.3: Workflow of the user study.

baseline that produces reports directly from the image prompt without detector-supported finding integration or workflow-level review. MedRAX was included as the closest prior CXR agent framework considered in this thesis. However, the present evaluation used only its report-generation tool rather than the full interactive MedRAX agent workflow. These two baselines therefore represent different external reference points: direct biomedical multimodal generation and prior CXR agent-based report generation. The results should be interpreted as a report-generation comparison, not as a full system-level comparison between TRACE and the complete MedRAX agent.

The internal diagnostic-pipeline conditions were defined directly from the implemented TRACE workflow. The *report draft only* condition isolates direct VLM-based report drafting by using the initial MedGemma draft as the final report. The *draft merged with detector findings without review* condition isolates the contribution of detector-supported

finding integration by adding Ark$^+$-derived structured findings without subsequent review or refinement. The *full TRACE report pipeline* condition adds the review-and-rewrite stage and therefore tests the contribution of full report refinement beyond draft generation and detector-supported merging alone.

MedGemma was not treated as a separate external comparison system because, in the implemented TRACE architecture, it functions as the report-draft generator within the diagnostic pipeline. Treating it as a standalone baseline would therefore duplicate the *report draft only* condition rather than add a distinct comparison category.

4.4 Statistical Methodology

The automatic report-quality evaluation and diagnostic-pipeline ablation analysis in Experiment I report benchmark metrics over the full fixed evaluation set and are interpreted descriptively rather than through inferential significance testing. This choice reflects the role of Experiment I as a benchmark-based comparative evaluation of report-generation conditions on a fixed test set, whereas inferential statistics are reserved for the within-subject user-study components.

For Experiment II and Experiment III, participant responses were scored using a GPT-5-based LLM-as-judge setup with a pre-specified rubric. This approach was selected because participant responses were expressed in natural language and could contain semantically equivalent or partially correct answers with different wording, making simple exact-

match scoring unsuitable for the predefined scoring rubrics. The GPT-5-based judge was used only for offline evaluation of participant responses and was not part of the TRACE runtime or agent-control architecture. The judge was provided with the participant response, the relevant case context, and the scoring rubric for the target task.

To constrain the scoring process and reduce condition-dependent scoring variability, the same judge model and pre-specified scoring criteria were applied across the compared conditions within each experiment. The judge was instructed to score responses according to the provided rubric and case-specific evidence, rather than to perform an independent assessment of radiological correctness. In Experiment II, participant-exposed outputs were logged and scored against the exact displayed content shown to that participant, so scoring was grounded in the actual report wording and transparency cues presented during the live interaction. In Experiment III, the judge was provided with the fixed report pair and the pre-specified gold-standard comparison context for the corresponding pair. The LLM-as-judge procedure should therefore be understood as a structured research scoring method for applying the predefined rubrics, rather than as a substitute for clinician adjudication or formal inter-rater validation.

For Experiment II and Experiment III, paired condition comparisons were analyzed using the Wilcoxon signed-rank test. For the rubric-based task scores, this non-parametric paired test was appropriate because the outcomes were bounded and ordinal or semi-ordinal in nature. For completion time in Experiment III, the same paired non-parametric approach was retained to keep the task-performance analysis consistent and to avoid relying on a normality assumption for paired time differences. All tests were two-sided with $\alpha =$

0.05.

For the questionnaire-based assessment, SUS and Raw TLX scores were treated as approximately continuous scale outcomes, which is a common approach for questionnaire-based human-factors analysis. Paired-samples t-tests were used to compare SUS total score, Raw TLX overall score, and Raw TLX dimension scores between the *text-only* and *full TRACE* interface modes. Within-subject effect sizes were reported as Cohen's d_z. Holm–Bonferroni adjustment was applied across the six Raw TLX dimension-level comparisons.

4.5 Experiment I: Objective Report Quality Evaluation

Experiment I evaluates report generation quality at two levels: external system comparison and internal diagnostic-pipeline ablation. Both analyses use the same 2,000-image mixed evaluation set described in Section 4.2 and the same automatic report-quality metrics. The full TRACE report pipeline serves as the shared reference condition for both analyses, so it is evaluated once and interpreted in two ways in the results chapter: first as the proposed system in comparison with external methods, and second as the final condition in the internal TRACE ablation.

4.5.1 External Comparison

The external comparison included LLaVA-Med, the MedRAX report-generation tool, and the full TRACE report pipeline. For each image, the generated report was paired with the

corresponding reference report and evaluated using SEMB, RadCliQ-v1, sample-wise F1, sample-wise positive-only F1, case-level FPR, case-level FNR, micro F1, and macro F1.

These metrics were chosen to provide complementary views of report quality. SEMB and RadCliQ-v1 capture semantic and clinically oriented report similarity, case-level F1 measures capture report-level label overlap, and micro F1 and macro F1 summarize aggregated label-level performance over CheXbert-derived positive observation labels. The case-level FPR and FNR use the binary case-status definitions introduced in Section 4.2 and complement the label-aware F1 metrics by capturing case-level over-calling and complete abnormality omission.

4.5.2 Diagnostic-Pipeline Ablation

The diagnostic-pipeline ablation compared three internal TRACE reporting conditions. The first condition was *report draft only*, in which the initial MedGemma draft was used directly as the final report. The second condition was *draft merged with detector findings without review*, in which Ark$^+$-derived structured findings were merged into the draft but no review or refinement stage was applied. The third condition was the *full TRACE report pipeline*, which included report drafting, detector-supported merging, and review-enhanced refinement.

The same 2,000-image evaluation set and the same automatic metrics were used for all three ablation conditions. This design allows the results chapter to examine not only whether the full TRACE reporting pipeline performs well overall, but also which parts of

the implemented diagnostic workflow contribute to those improvements.

4.6 Experiment II: Transparency-Supported Consultation Preparation

Experiment II evaluates whether the transparency-oriented TRACE interface helps users identify report conclusions that should be verified and formulate more specific consultation questions than a text-only baseline.

The study used a within-subject, counterbalanced design with two conditions: *text-only*, in which participants interacted with textual outputs without image overlays or process-level workflow information, and *full TRACE*, in which participants viewed the report together with image overlays and AI workflow information. Two single-image cases were prepared in advance, and each participant completed both cases, with case–condition assignment counterbalanced across participants to reduce case-order and condition-order effects. One case preserved a textual uncertainty cue within the report itself, whereas the *full TRACE* condition additionally exposed visual and process-level cues, including image overlays and merge-step disagreement between Ark$^+$-supported findings and MedGemma-generated report content.

Before the task began, participants were introduced to the system and the task instructions. For each case, participants were asked to identify at least one report conclusion that appeared to warrant clinician verification under the presented condition, explain why it

warranted verification, and write a consultation question.

Because the study used live system interaction, the exact report wording and transparency outputs were not guaranteed to be identical across runs. In addition, a single case could contain more than one conclusion that reasonably warranted further verification. For that reason, participant-exposed outputs were logged and all responses were scored against the exact displayed content shown to each participant rather than against a single canonical output instance. Scoring therefore followed a response-grounded rubric tied to the displayed report and transparency cues under the assigned condition. In the *text-only* condition, acceptable justifications were limited to cues visible in the report text itself, such as uncertainty, inconsistency, or weak textual support for an emphasized conclusion. In the *full TRACE* condition, acceptable justifications could additionally include transparency-related cues such as questionable image localization or disagreement between Ark$^+$-supported findings and MedGemma-generated report content. A response received credit if it identified at least one conclusion that reasonably warranted verification under the presented condition.

Three scored outcomes were used. The **Verification-worthy conclusion identification score** was scored as 0 or 1 depending on whether the participant identified at least one report conclusion that reasonably warranted clinician verification under the assigned condition. The **Risk-basis justification score** was scored as 0 or 1 depending on whether the participant provided at least one valid reason, grounded in information visible in the assigned condition, for why that conclusion should be verified. The **Consultation question specificity score** was scored from 0 to 2, where 0 indicated no meaningful con-

sultation question or an overly vague response, 1 indicated a broad but relevant question, and 2 indicated a clear and targeted verification-oriented question.

The primary outcome of Experiment II was the composite **Consultation Preparation Quality Score**, defined as the sum of the three component scores above and ranging from 0 to 4. Response scoring followed the rubric-based procedure described in Section 4.4. Representative example materials for Experiment II are provided in Appendix A.1. Because the study used live system interaction, these appendix materials should be interpreted as illustrative examples rather than identical report instances shown to all participants.

4.7 Experiment III: Archive-Assisted Cross-Study Comparison

Experiment III evaluates whether archive-assisted comparison support improves users' efficiency and accuracy in report-level comparison between a prior CXR study and a current study. Because the materials in Experiment III were fixed in advance, the shared abnormalities and key report-level difference could be annotated consistently across participants.

The study used a within-subject, counterbalanced design with two conditions: *without-archive*, in which participants reviewed the prior and current reports without archive-assisted comparison support, and *with-archive*, in which participants reviewed standardized archive-assisted outputs prepared in advance from the same cases. In the *with-archive* condition, an additional comparison summary was available to highlight common abnormalities and key report-level changes between the prior and current studies. All report

materials were fixed before the study so that completion time reflected participant comparison efficiency rather than system response time.

Participants were instructed that the task was report-level comparison. For each pair, they answered two structured questions: (1) what abnormality or abnormalities the two reports had in common, and (2) what the main difference between the two reports was. Completion time was recorded from task presentation to final submission.

Each pair was annotated in advance with a gold-standard common abnormality set and a gold-standard key difference. Four outcome measures were used. The *Common-abnormality-identification score* was scored as 0 or 1 depending on whether the participant correctly identified the shared abnormality or abnormalities. The *Difference-identification score* was scored as 0, 0.5, or 1, corresponding to incorrect, partially correct, or correct identification of the key report-level difference. The *Total comparison score* was computed as the sum of these two accuracy-related scores and ranged from 0 to 2. Completion time was recorded in seconds.

The primary outcomes of Experiment III were the *Total comparison score*, representing overall report-level comparison accuracy, and completion time, representing task efficiency. Response scoring followed the rubric-based procedure described in Section 4.4.

4.8 Questionnaire-Based Workload and Usability Assessment

In addition to the task-performance outcomes in Experiments II and III, the study assessed subjective usability and perceived workload for two overall interface modes: *text-only* and *full TRACE*. The *text-only* interface provided textual interaction and textual outputs, whereas the *full TRACE* interface additionally exposed multimodal and transparency-oriented components, including visual evidence, structured findings, execution trajectory, and archive-assisted comparison views when applicable. Because every participant experienced both interface modes, questionnaire ratings were collected as retrospective condition-specific ratings after completion of the user-study session.

Usability was measured using the System Usability Scale (SUS), reported as a total score on the standard 0–100 scale. Perceived workload was measured using Raw NASA Task Load Index (Raw TLX), including the overall score and the six TLX dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration. Questionnaire scores were analyzed as paired within-subject outcomes using the procedure described in Section 4.4.

Chapter 5. Results and Discussion

5.1 Chapter Overview

This chapter reports the results of the evaluation introduced in Chapter 4. The results are organized according to four completed evaluation components: objective report quality evaluation, transparency-supported verification awareness and consultation preparation, archive-assisted cross-study comparison, and questionnaire-based workload and usability assessment. The chapter then discusses how these findings relate to the design choices introduced in Chapter 3, with particular attention to component-level contribution, workflow-level benefit, and practical boundary conditions.

5.2 Results of Experiment I: Objective Report Quality Evaluation

Experiment I evaluated report-generation quality at two levels: external system comparison and internal diagnostic-pipeline ablation. Table 5.1 summarizes all five conditions in one consolidated view. The first two rows correspond to the external comparison systems. The last three rows correspond to the internal TRACE pipeline conditions. The full TRACE report pipeline serves as the shared reference condition for both interpretations.

Table 5.1: Objective report quality results for Experiment I. The first two rows are external comparison systems. The last three rows are internal TRACE pipeline conditions. The full TRACE report pipeline serves as the shared reference condition for both interpretations. Higher is better for SEMB, sample-wise F1, sample-wise positive-only F1, micro F1, and macro F1. Lower is better for RadCliQ-v1, case-level FPR, and case-level FNR.

Method	SEMB	RadCliQ-v1	SW-F1	SW-F1+	Case FPR	Case FNR	Micro F1	Macro F1
External comparison								
LLaVA-Med	0.220	2.050	0.281	0.138	0.533	0.532	0.150	0.126
MedRAX	0.422	1.606	0.543	0.415	0.292	**0.188**	0.468	0.334
Internal TRACE pipeline conditions								
Report draft only	0.447	1.800	0.565	0.315	**0.111**	0.345	0.411	0.236
Draft merged with detector findings	0.456	1.811	0.577	0.437	0.242	0.190	0.500	0.388
Full TRACE report pipeline	**0.512**	**1.437**	**0.584**	**0.451**	0.186	0.194	**0.505**	**0.394**

In the external comparison, the full TRACE report pipeline achieved the strongest overall metric profile. Relative to LLaVA-Med, the gains were large across all reported metrics. Relative to MedRAX, the full pipeline achieved higher SEMB (0.512 vs. 0.422), lower RadCliQ-v1 (1.437 vs. 1.606), higher sample-wise F1 (0.584 vs. 0.543), higher sample-wise positive-only F1 (0.451 vs. 0.415), lower case-level FPR (0.186 vs. 0.292), higher micro F1 (0.505 vs. 0.468), and higher macro F1 (0.394 vs. 0.334). The only exception was case-level FNR, where MedRAX remained slightly lower (0.188 vs. 0.194). These results indicate that the full TRACE report pipeline produced the strongest overall external comparison profile on this benchmark. The magnitude of these differences was not uniform across metrics. Relative to MedRAX, the improvements were more pronounced for SEMB, macro F1, and case-level FPR, whereas the differences in several other measures were more modest, and MedRAX retained a slightly lower case-level FNR. The external comparison should therefore be interpreted as a stronger overall metric profile rather than uniform superiority on every aspect of report quality.

The internal ablation shows a more specific contribution pattern. The *report draft only* condition already produced relatively strong semantic alignment and overall sample-wise F1, but it remained weaker on positive-only agreement and label-level coverage, with sample-wise positive-only F1 of 0.315, micro F1 of 0.411, and macro F1 of 0.236. It was also the most conservative condition, yielding the lowest case-level FPR (0.111). Adding detector-supported finding integration produced the largest jump in positive-finding preservation and label agreement: sample-wise positive-only F1 increased from 0.315 to 0.437, micro F1 from 0.411 to 0.500, and macro F1 from 0.236 to 0.388. At the same time, the merged-without-review condition became less conservative in no-finding cases, with case-level FPR increasing from 0.111 to 0.242.

The final refinement stage further improved the overall profile. Compared with the merged-without-review condition, the full TRACE pipeline increased SEMB from 0.456 to 0.512, reduced RadCliQ-v1 from 1.811 to 1.437, increased sample-wise F1 from 0.577 to 0.584, increased sample-wise positive-only F1 from 0.437 to 0.451, increased micro F1 from 0.500 to 0.505, increased macro F1 from 0.388 to 0.394, and reduced case-level FPR from 0.242 to 0.186. Case-level FNR changed only slightly, from 0.190 to 0.194. This pattern indicates that detector-supported merging contributed most of the gain in positive-finding preservation and label-level agreement, whereas the refinement stage improved semantic quality, overall report quality, and case-level FPR.

From a practical perspective, the ablation suggests that the two additional stages contributed different magnitudes and types of improvement. Detector-supported merging produced relatively large gains in positive-finding preservation, increasing sample-wise

positive-only F1 from 0.315 to 0.437, micro F1 from 0.411 to 0.500, and macro F1 from 0.236 to 0.388, but also increased the case-level FPR from 0.111 to 0.242. The subsequent refinement stage produced smaller changes in micro F1 (0.500 to 0.505) and macro F1 (0.388 to 0.394), while providing clearer improvements in SEMB, RadCliQ-v1, and case-level FPR. These results suggest that the practical contribution of refinement was primarily to improve report coherence and moderate over-reporting rather than to produce large additional gains across all label-oriented metrics.

Overall, the results of Experiment I show that the full TRACE report pipeline was strong both as an external comparison system and as the best-performing condition in the internal pipeline analysis. The ablation results also clarify how these gains were produced: detector-supported structured findings mainly improved positive-label coverage, whereas the final review-and-refinement stage improved overall report quality and moderated some of the over-reporting behavior introduced by naive merging alone.

5.3 Results of Experiment II: Transparency-Supported Verification Awareness and Consultation Preparation

Experiment II evaluated whether the full TRACE interface helped participants identify verification-worthy conclusions and formulate more specific consultation questions than a text-only baseline. The study included 48 participants in a within-subject, counterbalanced design. The results are presented in Table 5.2. Values are reported as mean ± standard

Table 5.2: Results of Experiment II. Values are reported as mean ± SD. The verification-worthy conclusion identification score and the risk-basis justification score were both binary (0 or 1). The consultation question specificity score ranged from 0 to 2, where 0 indicated no meaningful consultation question or an overly vague response, 1 indicated a broad but relevant question, and 2 indicated a clear and targeted verification-oriented question. The Consultation Preparation Quality Score is the sum of the three component scores and ranges from 0 to 4. Paired condition comparisons were assessed with the Wilcoxon signed-rank test.

Measure	Text-only	Full TRACE	p
Verification-worthy conclusion identification	0.625 ± 0.489	1.000 ± 0.000	2.21×10^{-5}
Risk-basis justification	0.313 ± 0.468	1.000 ± 0.000	9.22×10^{-9}
Consultation question specificity	0.625 ± 0.789	1.708 ± 0.459	3.11×10^{-7}
Consultation Preparation Quality Score	1.563 ± 1.556	3.708 ± 0.459	8.43×10^{-8}

deviation, and paired comparisons were assessed with the Wilcoxon signed-rank test.

The primary outcome, the Consultation Preparation Quality Score, was substantially higher in the full TRACE condition than in the text-only condition (3.708 ± 0.459 vs. 1.563 ± 1.556, $p = 8.43 \times 10^{-8}$). This indicates that participants were markedly better able to identify verification-worthy report concerns and prepare more targeted verification-oriented consultation questions when the transparency-oriented interface was available. The same pattern was observed across all three component measures. The verification-worthy conclusion identification score increased from 0.625 ± 0.489 in the text-only condition to 1.000 ± 0.000 in the full TRACE condition ($p = 2.21 \times 10^{-5}$), indicating that participants were more likely to identify at least one report conclusion that reasonably warranted clinician verification when transparency cues were available. The risk-basis justification score increased from 0.313 ± 0.468 to 1.000 ± 0.000 ($p = 9.22 \times 10^{-9}$), showing that participants in the full TRACE condition were far more likely to provide a valid reason for why the

selected conclusion should be verified. Consultation question specificity likewise increased from 0.625 ± 0.789 to 1.708 ± 0.459 ($p = 3.11 \times 10^{-7}$), indicating that the questions written under the full TRACE condition were more concrete, targeted, and verification-oriented.

The score distribution further clarifies this difference. In the text-only condition, 18 of 48 responses received a total score of 0, indicating that the participant failed to identify a verification-worthy conclusion, failed to provide a valid risk-based justification, or did not formulate a meaningful consultation question. By contrast, in the full TRACE condition, all responses received either 3 or 4 points: 14 responses scored 3 and 34 responses scored 4. Under the present task design, this distribution indicates not only higher average performance, but also substantially more consistent task success under the full TRACE condition.

These results are informative in two ways. First, the text-only condition was not entirely devoid of usable cues. Some participants were able to identify a verification-worthy conclusion from textual uncertainty alone, especially when the report itself contained hedging or weak textual support for an emphasized conclusion. However, many participants in the text-only condition either did not notice the textual uncertainty or could not translate it into a concrete consultation question. Second, the full TRACE condition did not simply encourage more questioning in general. Rather, it improved participants' ability to link visible uncertainty or inconsistency to a specific report concern and to formulate a more targeted follow-up question for clinician verification. Experiment II therefore supports a practical interpretation of transparency. The benefit of TRACE was not that it enabled non-expert users to determine whether the AI report was medically correct. Instead, it

Table 5.3: Results of Experiment III. Paired condition comparisons, including completion time, were analyzed using the Wilcoxon signed-rank test as a non-parametric within-subject comparison.

Measure	Without-archive	With-archive	p
Common-abnormality-identification score	0.938 ± 0.245	0.958 ± 0.202	0.317
Difference-identification score	0.760 ± 0.326	0.896 ± 0.205	0.022
Total comparison score	1.698 ± 0.446	1.854 ± 0.341	0.018
Completion time (s)	102.25 ± 12.53	68.71 ± 7.98	1.62×10^{-9}

helped them move from general uncertainty to more specific, risk-grounded consultation preparation.

5.4 Results of Experiment III: Archive-Assisted Cross-Study Comparison

Experiment III evaluated whether archive-assisted comparison improved users' ability to identify common abnormalities and key differences between a prior chest X-ray report and a current report. The study also included 48 participants under a within-subject, counterbalanced design. Table 5.3 reports the exact statistics, and Figure 5.1 provides a visual summary of the same pattern.

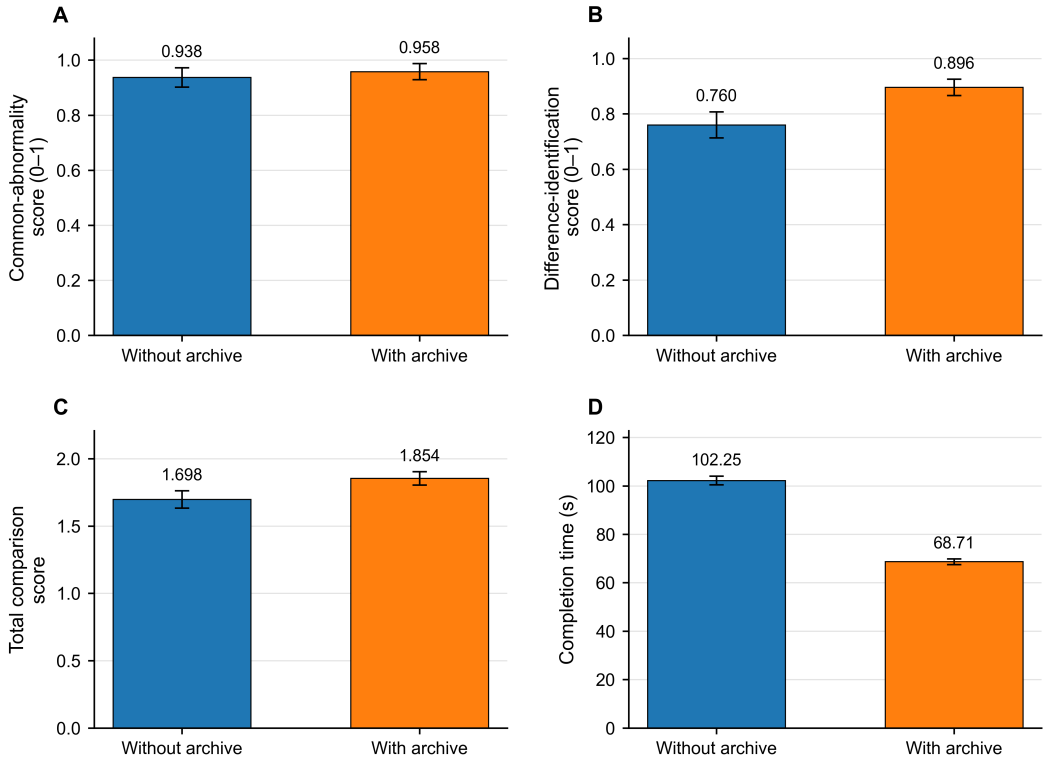


Figure 5.1: Summary of Experiment III outcomes under the *without-archive* and *with-archive* conditions. Higher scores indicate better comparison accuracy, whereas lower completion time indicates greater task efficiency. (A) Common-abnormality-identification score. (B) Difference-identification score. (C) Total comparison score. (D) Completion time. Bars show means and error bars indicate standard errors.

As shown in Figure 5.1, archive support produced its clearest effect on task efficiency. Mean completion time decreased from 102.25 seconds in the without-archive condition to 68.71 seconds in the with-archive condition, a reduction of approximately 33.54 seconds. The paired Wilcoxon comparison indicated a significant difference ($p = 1.62 \times 10^{-9}$). Archive support therefore substantially improved the efficiency of report-level comparison.

The accuracy-related measures showed a more selective pattern. The common-abnormality-identification score increased only slightly from 0.938 to 0.958, and this difference did not reach statistical significance ($p = 0.317$). By contrast, the difference-identification score increased from 0.760 to 0.896 and the total comparison score increased from 1.698 to 1.854, with both paired improvements reaching statistical significance ($p = 0.022$ and $p = 0.018$, respectively). These results indicate that archive support improved not only task speed but also users' ability to identify report-level differences across studies.

This result pattern suggests that the archive mechanism functioned as more than a simple time-saving aid. Participants in both conditions were already able to identify common abnormalities with high accuracy, which limited the room for improvement on the common-abnormality measure. In contrast, identifying changes between studies appears to have benefited more directly from the archive-assisted comparison summary, leading to significantly better difference identification and overall comparison performance.

This interpretation is consistent with the implemented role of the archive subsystem. The archive mechanism is not a dedicated longitudinal diagnosis model. Rather, it acts as a continuity-oriented reading aid that retrieves prior-case context and surfaces a report-level comparison summary. The findings of Experiment III therefore show that archive support not only made cross-study comparison substantially faster, but also significantly improved report-level difference identification and overall comparison performance.

5.5 Results of the Questionnaire-Based Workload and Usability Assessment

The questionnaire analysis compared two overall interface modes: *text-only* and *full TRACE*. The *text-only* mode provided textual interaction and textual outputs without the additional multimodal and transparency-oriented components of TRACE. The *full TRACE* mode included these components, such as visual evidence, structured findings, execution trajectory, and archive-assisted comparison views when applicable. Table 5.4 reports the exact overall questionnaire statistics, whereas Figure 5.2 provides a compact visual summary of the same usability–workload pattern. The full TRACE interface received a substantially higher SUS score than the text-only interface ($M = 81.98$, $SD = 10.88$ vs. $M = 65.99$, $SD = 14.59$). The mean paired increase was 15.99 points, and this difference was statistically significant, $t(47) = 7.93$, $p < .001$, with a large within-subject effect size ($d_z = 1.14$). This result indicates that participants perceived the full TRACE interface as more usable overall.

A complementary pattern was observed for perceived workload. Raw TLX scores were lower under full TRACE ($M = 38.87$, $SD = 7.66$) than under the text-only condition ($M = 48.63$, $SD = 8.04$), corresponding to a mean paired reduction of 9.76 points. This difference was statistically significant, $t(47) = -13.95$, $p < .001$, with a large within-subject effect size ($d_z = -2.01$). Together, these results indicate that the full TRACE interface was perceived as more usable and less demanding overall.

Table 5.4: Overall questionnaire results for usability and workload ($N = 48$, within-subject design). Mean differences are computed as full TRACE minus text-only.

Measure	Text-only Mean ± SD	Full TRACE Mean ± SD	Mean Diff.	$t(47)$	p	Cohen's d_z
SUS total (0–100)	65.99 ± 14.59	81.98 ± 10.88	15.99	7.93	3.24×10^{-10}	1.14
Raw TLX (0–100)	48.63 ± 8.04	38.87 ± 7.66	-9.76	-13.95	2.53×10^{-18}	-2.01

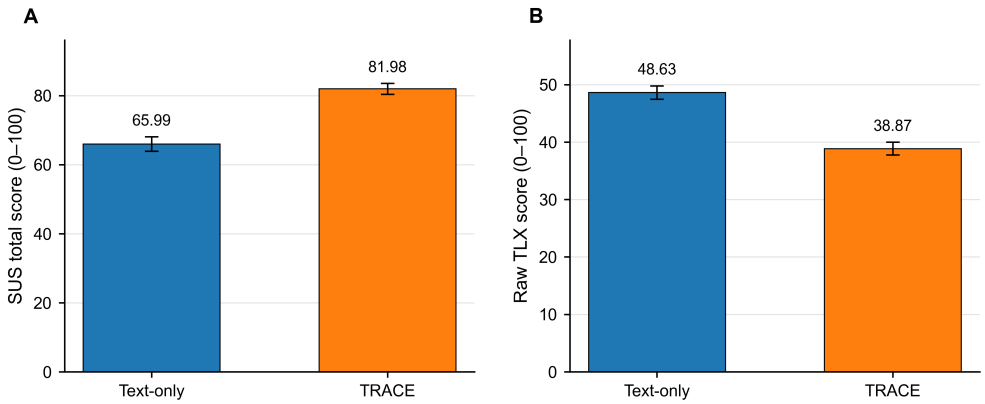


Figure 5.2: Overall questionnaire results by condition. Higher SUS indicates better usability, whereas lower Raw TLX indicates lower perceived workload. (A) Usability measured by the System Usability Scale (SUS). (B) Perceived workload measured by Raw NASA-TLX. Bars show means and error bars indicate standard errors.

A more detailed breakdown of the Raw TLX dimensions is shown in Table 5.5 and Figure 5.3. After Holm–Bonferroni adjustment across the six dimension-level comparisons, all dimensions except physical demand remained statistically significant. As also visible in Figure 5.3, the largest reductions were observed for mental demand and frustration. This pattern is consistent with the intended role of transparency in the system design. By exposing intermediate cues, visual grounding, label explanations, and explicit workflow information, the full TRACE interface reduced the cognitive burden involved in interpreting unfamiliar medical terms, understanding how the report was produced, and deciding which

findings or conclusions warranted closer scrutiny. Frustration also decreased substantially, suggesting that participants felt less uncertain and less forced to rely on text-only interaction without visual or process-level support when attempting to judge possible AI limitations or report risks.

The remaining TLX dimensions showed a more moderate pattern. Effort was lower under full TRACE, indicating that participants experienced the task as somewhat easier overall, but the reduction was smaller than the decreases observed for mental demand and frustration. Temporal demand and perceived performance also improved, although to a lesser extent, suggesting that the richer interface supported task completion without radically changing the basic pace or structure of the activity. Physical demand, by contrast, remained low in both conditions and did not differ significantly, $t(47) = -0.90$, $p_{\text{adj}} = .371$. This is unsurprising, as both interfaces required similarly minimal physical actions such as reading, clicking, and selecting answers; the primary difference between conditions lay in cognitive support rather than motor effort.

Overall, the questionnaire results provide convergent evidence that the full TRACE interface improved the subjective quality of interaction relative to the text-only interface by increasing usability and reducing cognitive workload.

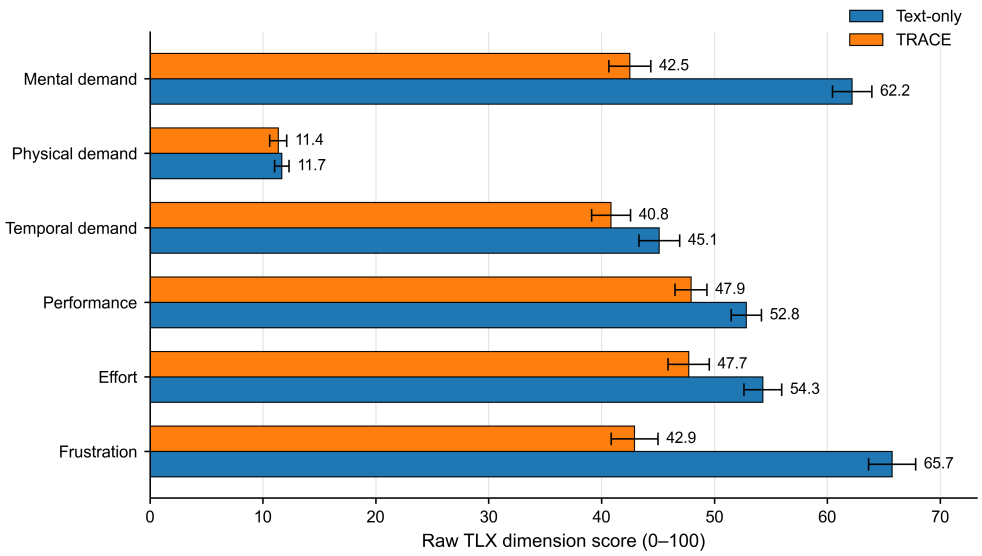


Figure 5.3: Raw TLX dimension scores by condition. Lower values indicate lower perceived workload. The largest reductions under full TRACE were observed for mental demand and frustration, whereas physical demand remained low in both conditions. Bars show means and error bars indicate standard errors.

Table 5.5: Raw TLX dimension scores by condition ($N = 48$). Lower values indicate lower perceived workload. Mean differences are computed as full TRACE minus text-only. Raw p values from the six paired dimension-level comparisons were adjusted using the Holm–Bonferroni procedure.

Dimension	Text-only Mean ± SD	Full TRACE Mean ± SD	Mean Diff.	$t(47)$	p_{adj}	Cohen's d_z
Mental demand	62.19 ± 12.11	42.50 ± 12.84	-19.69	-16.74	1.17×10^{-20}	-2.42
Physical demand	11.67 ± 4.42	11.35 ± 5.23	-0.31	-0.90	0.371	-0.13
Temporal demand	45.10 ± 12.48	40.83 ± 11.91	-4.27	-4.65	5.36×10^{-5}	-0.67
Performance	52.81 ± 9.33	47.92 ± 9.78	-4.90	-5.68	2.43×10^{-6}	-0.82
Effort	54.27 ± 11.58	47.71 ± 12.63	-6.56	-5.78	2.31×10^{-6}	-0.83
Frustration	65.73 ± 14.48	42.92 ± 14.40	-22.81	-15.18	4.81×10^{-19}	-2.19

5.6 Discussion

The results across the four evaluation components show that TRACE provided measurable benefit at both the report-generation level and the interaction level, but the improvements were not uniform across all components or all metrics. The results should therefore be interpreted not simply as overall system improvement, but as evidence that different components contributed to different types of gain.

The clearest component-level pattern appears in Experiment I. The *report draft only* condition already produced relatively strong semantic alignment, which indicates that direct VLM-based drafting can generate fluent radiology-style text with reasonable overall correspondence to the reference report. However, the same condition was substantially weaker on sample-wise positive-only F1, micro F1, and macro F1, indicating that fluent draft generation alone did not reliably preserve the set of positive abnormalities expressed in the reference report. This gap helps clarify the role of Ark$^+$ within the diagnostic pipeline.

Ark$^+$ did not improve the report by making the text more stylistically polished; its main contribution was to inject explicit positive abnormality cues that improved positive-finding preservation and label-level agreement.

The ablation also reveals that detector-supported merging alone is not sufficient. When Ark$^+$-derived findings were merged into the draft without subsequent review, positive-only F1 and label-level agreement improved sharply, but the case-level FPR increased from 0.111 to 0.242. This trade-off is important. It shows that structured detector cues are useful for recovering omitted positives, but naïve insertion of those cues can also make the report less conservative in cases where no abnormality should be emphasized. In other words, the detector-supported merge stage strengthened recall-like behavior, but it also introduced a risk of over-reporting. The final review-and-refinement stage compensated for part of this trade-off by reducing the case-level FPR to 0.186 while simultaneously improving SEMB and RadCliQ-v1. This suggests that the review stage did more than improve wording. It functioned as a consistency-control step that moderated detector insertion, resolved contradictions, and produced a more coherent final report.

At the same time, not all metric dimensions improved equally. Case-level FNR changed only slightly across the final two pipeline conditions, and MedRAX remained slightly better on this metric than the full TRACE pipeline. This indicates that the final refinement stage mainly improved report coherence, semantic quality, and case-level false-positive behavior, while having little effect on the proportion of cases in which the reference report contained at least one positive abnormality label but the generated report contained none. Thus, although the detector-supported pipeline improved overall label agreement and positive-

finding preservation, it did not eliminate complete case-level abnormality omissions.

Experiment II shows that the value of the system extends beyond report generation itself. The full TRACE interface substantially improved participants' ability to identify verification-worthy report concerns, justify why they seemed worth verifying, and formulate more targeted consultation questions. This result is especially important because it clarifies what transparency contributed in practice. The benefit of the full interface was not that it enabled non-expert users to determine whether the AI report was medically correct. Rather, it helped them recognize when a conclusion merited follow-up and translate that concern into a more specific consultation question.

This interpretation should be kept appropriately narrow. Transparency cues were useful because they made uncertainty, inconsistency, or evidential weakness more actionable, not because they served as definitive error detectors. For example, a Grad-CAM overlay that extends partly outside the thoracic region may appropriately prompt closer scrutiny, but it does not automatically imply that the associated abnormality prediction is false. In the same way, a disagreement between Ark$^+$-supported findings and MedGemma-generated report content is best interpreted as a cue for verification rather than as proof that one component is necessarily correct and the other is necessarily wrong. The main contribution of the transparency interface was therefore to support *verification awareness* rather than direct correctness judgment.

Experiment III complements this interpretation at the archive-assisted continuity level. Archive support greatly reduced task time and significantly improved difference identifica-

tion, but its effect on common-abnormality identification was small and not statistically significant. This pattern is informative. It suggests that archive assistance was most useful when the task required participants to detect what had changed between two reports, not when the task involved simply recognizing information that was already obvious in both reports. Because baseline performance on common abnormality identification was already high, the archive summary had limited room to help on that measure. The archive subsystem therefore seems most beneficial as a continuity aid for highlighting interval change rather than as a universal enhancer of all forms of cross-study reading.

The user-study findings should also be interpreted in light of the limited stimulus set. Experiment II used two single-image CXR cases, and Experiment III used two matched prior–current report pairs. Therefore, the observed improvements provide evidence of assistive benefit within the controlled tasks and selected cases, rather than broad evidence that the same effects generalize across CXR pathologies, report styles, patient populations, or clinical settings.

The questionnaire findings reinforce the interface-level results. Higher SUS scores and lower Raw TLX scores indicate that the full TRACE interface did not impose a usability penalty relative to the text-only interface, despite exposing additional multimodal and transparency-oriented components. On the contrary, the largest workload reductions occurred in mental demand and frustration, suggesting that the additional multimodal and transparency-oriented interface support functioned as cognitive support rather than as extra information clutter. This is an important point, because richer interfaces in medical AI can easily become visually dense or cognitively burdensome. In the present study, the

transparency-oriented design appears to have reduced rather than increased the subjective difficulty of the task.

Several boundary conditions and failure modes also emerge from the combined results. First, transparency support improved users' ability to prepare better follow-up questions, but the study did not test whether non-expert users could judge medical correctness or safely overrule model output. Second, archive-assisted comparison improved report-level comparison, but the archive subsystem did not perform image-level temporal reasoning and should not be interpreted as a longitudinal diagnosis model. Third, detector-supported report generation improved overall report quality, but the ablation results show that structured detector input can introduce over-reporting pressure before refinement. These observations do not weaken the contribution of TRACE; rather, they define the scope within which its benefits should be interpreted.

The results support a more specific interpretation than a simple claim of overall superiority. The diagnostic pipeline benefited from complementary roles across components: MedGemma contributed narrative drafting, Ark$^+$ contributed structured positive-finding support, and review/refinement contributed consistency control. The transparency interface then translated system-internal evidence into more actionable user-side verification behavior, while archive support improved report-level comparison efficiency and change detection. The overall contribution of TRACE is therefore not only stronger report generation, but a more inspectable and interaction-supportive assistive workflow.

5.7 Chapter Summary

This chapter reported the results of the four evaluation components and analyzed how those results relate to the implemented design of TRACE. Experiment I showed that the full TRACE report pipeline achieved the strongest overall report-quality profile, while the internal ablation clarified the distinct roles of detector-supported finding integration and review-guided refinement. Experiment II showed that the transparency-oriented interface substantially improved verification awareness and consultation preparation. Experiment III showed that archive support greatly improved comparison efficiency and significantly improved difference identification. The questionnaire results further showed that the full TRACE interface was associated with higher usability and lower cognitive workload than the text-only interface.

The next chapter concludes the thesis, summarizes the main contributions, and discusses the limitations and future directions of the current system.

Chapter 6. Conclusion and Future Work

6.1 Conclusion

This thesis presented TRACE, a multimodal multi-agent framework for transparent CXR assistance. The work addressed a practical limitation in much of the existing CXR AI literature: although recent systems have made substantial progress in report generation, multimodal reasoning, and task-specific image analysis, many remain limited as user-facing assistive systems. In particular, text-only or report-centered outputs provide limited support for evidence inspection, verification-oriented follow-up, and continuity across encounters.

TRACE was developed as a response to that gap. Rather than treating CXR AI as a single prediction step, the framework organizes detector-supported image analysis, multimodal report drafting, report refinement, transparency-oriented interaction, consultation support, and archive-assisted comparison within one assistive workflow. The contribution of the work is therefore not only architectural. More importantly, it examines how these capabilities function together when the goal is not merely to produce a report, but to support interpretation, inspection, follow-up questioning, and report-level continuity.

The evaluation supports this broader framing. At the report-generation level, the full

TRACE pipeline achieved the strongest overall profile on most reported semantic and label-oriented metrics. The internal ablation further showed that different parts of the diagnostic pipeline contributed different kinds of value: detector-supported finding integration improved positive-finding preservation and label-level agreement, whereas the review-and-refinement stage improved overall report quality, semantic alignment, and no-finding behavior. At the interaction level, the transparency-oriented interface helped participants identify verification-worthy report conclusions more reliably, justify why those conclusions warranted clinician verification, and formulate more specific consultation questions. At the continuity level, archive support reduced cross-study comparison time and improved report-level difference identification, indicating that even a lightweight archive mechanism can provide meaningful assistance across encounters.

These findings support the central claim of the thesis within the scope of the present controlled evaluations: the assistive value of CXR AI depends not only on report-generation quality, but also on whether the system supports evidence inspection, verification-oriented interaction, and continuity across encounters. The main insight is that reporting, transparency, consultation, and archive-assisted comparison should be designed and evaluated as complementary parts of an assistive workflow rather than as isolated functions. The contribution of this thesis is therefore not only the implementation of TRACE, but also the evaluation evidence showing that these components can provide measurable assistive value beyond report generation alone.

6.2 Limitations

The results should be interpreted within several important limitations. First, the report-quality evaluation in Experiment I was conducted on selected public datasets and assessed using automatic metrics. These datasets provide useful benchmark coverage, but they may reflect dataset-specific acquisition protocols, reporting styles, disease prevalence, institutional practice patterns, and demographic distributions. As a result, the evaluation may not fully capture dataset bias, domain shift, or performance variation across hospitals, scanners, patient populations, and uncommon findings. The automatic metrics capture meaningful semantic and label-oriented properties, but they do not fully represent clinical usefulness, stylistic appropriateness, downstream interpretive safety, or real clinical impact. Confidence intervals were not estimated for the Experiment I report-quality metrics; accordingly, differences between reporting conditions are interpreted descriptively in terms of their magnitude and pattern on the fixed benchmark rather than as inferential evidence of statistical superiority. The findings from Experiment I should therefore be understood as benchmark evidence rather than as complete clinical validation.

Second, the human-centered evaluations were intentionally bounded in scope. Experiment II was designed to test whether transparency helped non-expert users identify verification-worthy report concerns and prepare more targeted consultation questions; it was not intended to measure whether non-experts could judge radiological correctness directly. Similarly, the observed difference between the *text-only* and *full TRACE* interface

modes should be interpreted as the value of additional inspectable evidence within this interface comparison, not as the isolated contribution of any single transparency component.

Third, generalizability remains limited. Experiment II used only two single-image CXR cases, and Experiment III used only two matched prior–current report pairs. Although the within-subject counterbalanced design reduced case-order and condition-order effects, the small number of study stimuli limits the extent to which the observed effects can be generalized across different pathologies, report characteristics, case difficulty levels, and clinical settings. In addition, participant responses in Experiments II and III were scored using a GPT-5-based LLM-as-judge procedure grounded in pre-specified rubrics and case context. This design provided a consistent and scalable procedure for applying the pre-specified rubrics, but it does not replace clinician adjudication or formal inter-rater validation. The resulting scores should therefore be interpreted as structured evaluation outcomes within the present research design rather than as definitive clinical judgments.

Fourth, the archive mechanism was evaluated as a report-level continuity aid rather than as a full longitudinal diagnosis system. The archive pathway retrieves prior context, compares stored findings, and summarizes added, resolved, and persistent abnormalities at the report level. It does not perform image registration, temporal lesion tracking, or fine-grained progression modeling. The archive results should therefore be interpreted as evidence for continuity-oriented reading support rather than for longitudinal image reasoning.

Finally, the framework remains dependent on the behavior of its component models and

prompts. In addition, the current implementation used a single fixed general-purpose LLM, GPT-4.1-mini, as the controller for agent-level orchestration and reasoning. The sensitivity of routing, tool use, report refinement, and consultation behavior to alternative LLM backbones was not evaluated in this thesis. The fine-tuned Ark$^+$ detector, MedGemma, and the LLM-based controllers may all introduce errors, including incomplete finding coverage, false-positive abnormality cues, imperfect visual attribution, inconsistent synthesis, or misleading generated text. Transparency cues can make such behavior more inspectable, but they do not eliminate the underlying risks of model dependency, hallucinated language, or misuse. For example, a Grad-CAM overlay that extends partly outside the thoracic region may appropriately prompt closer scrutiny, but it does not automatically imply that the associated abnormality prediction is false. More generally, the system should be understood as a research prototype rather than as a deployment-ready clinical product. Although it demonstrates the feasibility of integrating report generation, transparency, consultation, and archive-assisted support within one system, it does not yet address broader requirements such as large-scale persistence, robust uncertainty calibration, prospective clinical validation, governance, and regulatory approval.

6.3 Future Work

A natural next step is to study the transparency mechanism more directly. In the current framework, visual overlays, structured findings, report text, and process-level traces are exposed together, and Experiment II showed that this combined design improved verification-

oriented consultation preparation. What remains unclear is which transparency components are most responsible for that effect, and under what task conditions they matter most. Future work should therefore move from evaluating the full interface as a bundle to comparing specific transparency configurations more systematically, such as text-only, visual-evidence-only, process-trace-only, and combined transparency conditions. This would help clarify whether the main benefit comes from visual grounding, explicit evidence disagreement, or the combination of multiple transparency cues.

The archive pathway can also be strengthened from report-level continuity toward richer longitudinal support. In the present system, archive assistance is intentionally lightweight: it retrieves the most recent prior case, compares stored findings, and summarizes added, resolved, and persistent abnormalities at the report level. The study results suggest that even this simple formulation is useful, particularly for identifying differences across studies. A more advanced version of the framework could build on this result by producing more structured interval-change summaries, aligning prior and current evidence more explicitly, and testing the system on larger collections of serial CXRs. This would allow the archive mechanism to move beyond basic reading-aid functionality toward richer longitudinal comparison support.

A further direction concerns how the framework should represent and communicate uncertainty when the available evidence is weak, incomplete, or internally inconsistent. In the current system, such situations may appear as report hedging, questionable visual support, or disagreement between transparency cues and generated report content. These signals were useful in the present study as risk cues for verification-oriented follow-up

preparation, but they were not yet analyzed as a dedicated object of study. Future work could examine these uncertainty patterns more systematically and investigate how they should be reflected in the final report, the displayed transparency cues, and the system's caution language.

Future work should also strengthen the evaluation methodology itself. In the present thesis, structured rubric-based scoring for Experiments II and III was implemented using a GPT-5-based LLM-as-judge procedure grounded in case context. A valuable next step would be to compare this approach against clinician adjudication, multi-rater human scoring, or hybrid human-plus-LLM evaluation. Such work would help determine where LLM-based scoring is sufficiently reliable for structured assistive-task evaluation and where expert review remains necessary.

A separate direction is to examine the dependence of TRACE on the underlying controller LLM. The current implementation uses GPT-4.1-mini for agent-level orchestration and reasoning, whereas alternative controller models were not systematically evaluated. Future work could compare open-source and proprietary LLM controllers to determine how strongly routing, tool use, report refinement, and consultation behavior depend on the selected reasoning model.

More broadly, the framework should be evaluated under conditions that move closer to real use while remaining methodologically controlled. This includes larger and more diverse participant groups, stronger case sampling across pathology types and report difficulty, and broader testing of how different user populations interact with the system. Such

work would not only test whether the current findings generalize, but also help establish where transparent CXR assistance is most effective, for which users, and under what task conditions.

Appendix

Chapter A.

Representative User-Study Materials

This appendix provides representative materials for Experiment II and fixed materials for Experiment III, the two task-based user-study components of the evaluation. The appendix serves two different documentation purposes. For Experiment II, it presents illustrative example outputs that reflect the kinds of reports and transparency cues shown during the study, but not a single fixed report instance viewed identically by all participants. This distinction is necessary because Experiment II used live system interaction, and the report wording could vary across runs due to the multimodal language-model component. For Experiment III, by contrast, the appendix reproduces the fixed report pairs and standardized archive-assisted comparison summaries used in the study. These materials are included here to document the nature of the user-study stimuli without overloading the main evaluation chapter.

A.1 Experiment II Materials

Experiment II examined whether transparency cues helped participants identify a verification-worthy report conclusion, justify why it should be checked, and formulate a more specific consultation question. For the representative Experiment II tasks, the *text-only* condition exposed the textual report output only, including any textual uncertainty cues embedded in that output. The *full TRACE* condition additionally exposed transparency cues such as questionable Grad-CAM visual attribution and disagreement between Ark$^+$-supported findings and MedGemma-generated report content. Because the reports were generated during live system use, the exact wording was not guaranteed to be identical across runs. The underlying case images and task structure were fixed in advance, but the exact generated wording and transparency details could vary slightly across runs. The cases below should therefore be understood as representative examples of the types of reports and transparency cues used in the study rather than as exact fixed report instances viewed identically by all participants. In these materials, expressions such as "possible" were treated as valid textual uncertainty cues for participant interpretation and response scoring.

A.1.1 Representative Case A

Representative transparency cue in the full TRACE condition: The effusion heatmap included substantial activation outside the thoracic region (Figure A.1), which

was treated as a cue for closer verification rather than as a definitive error signal, and the MedGemma draft did not flag atelectasis, effusion, or cardiomegaly.

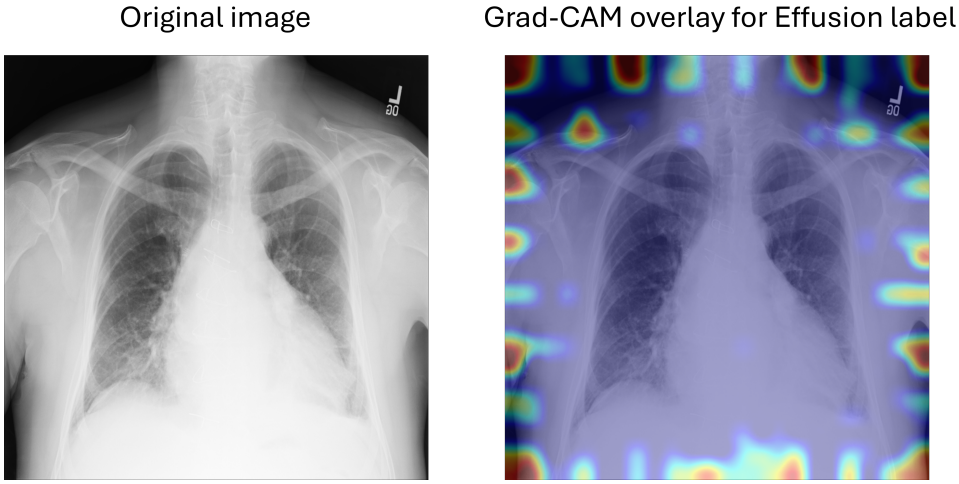


Figure A.1: Representative original image and heatmap overlay for Case A.

Representative report example:

Findings: Cardiomediastinal silhouette is enlarged. The aorta is mildly tortuous. Pulmonary vasculature otherwise unremarkable. No focal airspace consolidation or pneumothorax identified. Possible small pleural effusion. Mild atelectatic change or subsegmental volume loss may be present. Visualized osseous structures intact; no lines or tubes.

Impression:

1. Cardiomegaly.

2. Possible small pleural effusion and possible mild atelectasis.

A.1.2 Representative Case B

Representative transparency cue in the full TRACE condition: The infiltration heatmap included substantial activation outside the thoracic region (Figure A.2), which was treated as a cue for closer verification rather than as a definitive error signal, and the MedGemma draft did not flag infiltration.

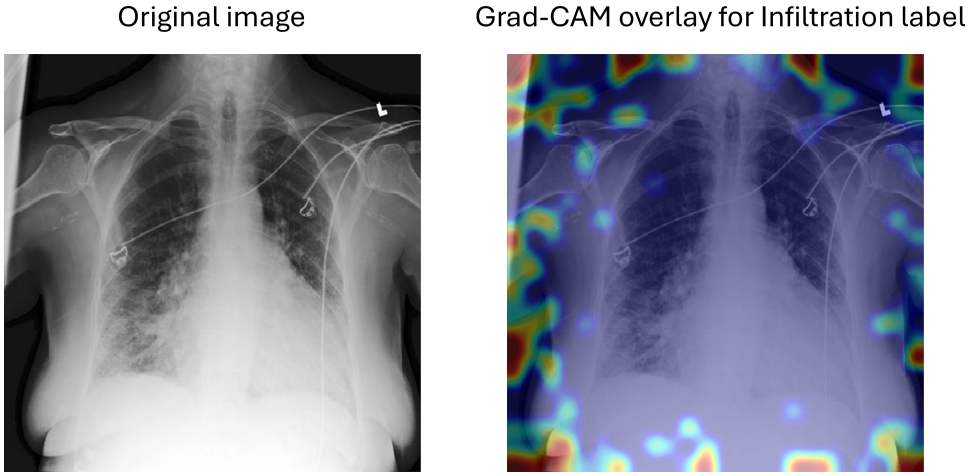


Figure A.2: Representative original image and heatmap overlay for Case B.

Representative report example:

Findings: There is increased opacity in the right lung field, greatest in the right upper lobe, suggestive of consolidation or infiltrate. The cardiomediastinal silhouette is overall within normal size and contour for the frontal technique, though mild cardiomegaly cannot be excluded on this single view. The left lung is clear. No pleural effusion or pneumothorax identified. Visualized bony structures appear intact. No lines or devices.

Impression:

1. Right upper-lobe opacity, concerning for consolidation/infiltrative process such as pneumonia.

2. Possible mild cardiomegaly (cardiomediastinal silhouette approximates normal on this frontal view; correlate clinically or with additional views if needed).

The examples above are intended to show the kinds of uncertainty cues, report-level risks, and transparency-supported inconsistencies that participants could encounter during Experiment II. They are representative rather than exhaustive, and they should not be interpreted as implying that every participant saw identical report wording in the live condition. They should also not be interpreted as implying that heatmap activation extending outside the thoracic region automatically indicates a false abnormality prediction; in the present study, such cues were treated as prompts for verification rather than as definitive correctness judgments.

A.2 Experiment III Materials

Experiment III examined whether archive-assisted comparison support improved report-level cross-study comparison. In the *without-archive* condition, participants viewed only the prior-study and current-study reports. In the *with-archive* condition, they additionally viewed a standardized comparison summary generated in advance from the same report

pair. The materials below reproduce the actual fixed texts used in the study under the standardized report-pair design.

A.2.1 Pair X

First study report:

Findings: A portable upright chest radiograph demonstrates a right-sided chest tube in position. A moderate right pneumothorax is present, evidenced by a visceral pleural line superiorly and laterally, with associated right-sided atelectasis. The cardiomediastinal silhouette is midline. The left lung is clear. No acute osseous abnormality identified.

Impression:

1. Moderate right pneumothorax with associated right-sided atelectasis, present despite chest tube placement.
2. Right-sided chest tube in position.

Second study report:

Findings: Frontal radiograph demonstrates the cardiomediastinal silhouette within expected limits for technique, though mild cardiomegaly cannot be entirely excluded on a single frontal view — correlation with prior imaging or a lateral radiograph is recommended. There are bilateral, lower-predominant hazy air-space opacities consistent with atelectasis and/or pulmonary infiltration (infection or edema). No pleural effusion or pneumothorax identified. Visualized osseous structures intact.

Impression:

1. Cardiomediastinal silhouette within expected limits on this frontal view; possible mild cardiomegaly cannot be excluded — correlate with prior imaging or obtain a lateral radiograph if concern. Bilateral lower-predominant hazy air-space opacities consistent with atelectasis and/or pulmonary infiltration (infection or edema); correlate clinically and consider follow-up imaging.

Additional text shown only in the with-archive condition:

Comparison With Previous Study: Persistent findings: Lower-predominant air-space opacities consistent with atelectasis are present (previously reported as right-sided, now described as bilateral lower-predominant). New or more conspicuous findings: Possible mild cardiomegaly and bilateral lower-predominant air-space infiltration (infection or edema) are newly described. Previously reported findings not clearly described on the current study: Moderate right pneumothorax and the right-sided chest tube reported previously are not described now, and no pneumothorax is identified on the current frontal radiograph.

A.2.2 Pair Y

First study report:

Findings: The cardiomediastinal silhouette is enlarged. Pulmonary vascular congestion with indistinct vessels and Kerley B lines is present, suggesting pulmonary edema/infiltration. There are bilateral pleural effusions, larger on the right. The lung

bases demonstrate opacities, likely atelectasis or consolidation. An endotracheal tube and nasogastric tube are in expected position. No pneumothorax or acute osseous abnormality identified.

Impression:

1. Cardiomegaly with pulmonary vascular congestion/pulmonary edema (infiltration) and bilateral pleural effusions, greater on the right.

2. Basilar air-space opacities, possibly atelectasis or consolidation; endotracheal and nasogastric tubes in place.

Second study report:

Findings: Endotracheal tube terminates approximately 4 cm above the carina. Cardiomediastinal silhouette is enlarged. There is increased opacity in the right lower lung field consistent with atelectasis; superimposed consolidation cannot be excluded. The left lung appears relatively clear. No large pleural effusion or pneumothorax identified. Visualized bony structures appear intact.

Impression:

1. Right lower-lung atelectasis; superimposed consolidation is possible.

2. Enlarged cardiac silhouette. Endotracheal tube projects about 4 cm above the carina.

Additional text shown only in the with-archive condition:

Comparison With Previous Study: Persistent findings: Right lower-lung atelectasis (superimposed consolidation cannot be excluded), enlarged cardiomediastinal silhouette (cardiomegaly), and endotracheal tube projecting about 4 cm above the carina. Previously reported findings not clearly described on the current study: Prior pulmonary vascular congestion/pulmonary edema and bilateral pleural effusions (greater on the right) are not described on the current radiograph.

References

1. United Nations Scientific Committee on the Effects of Atomic Radiation. *Sources, Effects and Risks of Ionizing Radiation: UNSCEAR 2020/2021 Report, Volume I* (United Nations, New York, NY, 2022).
2. NHS England. *Diagnostic Imaging Dataset Annual Statistical Release 2024/25* Official Statistics (NHS England, Nov. 2025).
3. Calli, E., Sogancioglu, E., van Ginneken, B., van Leeuwen, K. G. & Murphy, K. Deep Learning for Chest X-ray Analysis: A Survey. *Medical Image Analysis* **72,** 102125 (2021).
4. Bahl, S., Ramzan, T. & Maraj, R. Interpretation and Documentation of Chest X-rays in the Acute Medical Unit. *Clinical Medicine* **20,** s73 (2020).
5. Huang, J. *et al.* Generative Artificial Intelligence for Chest Radiograph Interpretation in the Emergency Department. *JAMA Network Open* **6,** e2336100 (2023).
6. Tanno, R., Barrett, D. G. T., Sellergren, A., *et al.* Collaboration Between Clinicians and Vision-Language Models in Radiology Report Generation. *Nature Medicine* **31,** 599–608 (2025).
7. Ouis, M. Y. & Akhloufi, M. A. ChestBioX-Gen: Contextual Biomedical Report Generation from Chest X-ray Images Using BioGPT and Co-attention Mechanism. *Frontiers in Imaging* **3,** 1373420 (2024).
8. Baltruschat, I. *et al.* Smart Chest X-ray Worklist Prioritization Using Artificial Intelligence: A Clinical Workflow Simulation. *European Radiology* **31,** 3837–3845 (2021).
9. Ahn, J. S., Ebrahimian, S., McDermott, S., *et al.* Association of Artificial Intelligence-Aided Chest Radiograph Interpretation With Reader Performance and Efficiency. *JAMA Network Open* **5,** e2229289 (2022).
10. Pham, H. H., Nguyen, H. Q., Nguyen, H. T., Le, L. T. & Khanh, L. An Accurate and Explainable Deep Learning System Improves Interobserver Agreement in the Interpretation of Chest Radiograph. *IEEE Access* **10,** 104512–104531 (2022).
11. Shin, H. J., Han, K., Ryu, L. & Kim, E.-K. The Impact of Artificial Intelligence on the Reading Times of Radiologists for Chest Radiographs. *npj Digital Medicine* **6,** 82 (2023).
12. Nori, H., King, N., McKinney, S. M., Carignan, D. & Horvitz, E. *Capabilities of GPT-4 on Medical Challenge Problems* 2023. arXiv: `2303.13375 [cs.CL]`.
13. Yan, Z. *et al. Multimodal ChatGPT for Medical Applications: An Experimental Study of GPT-4V* 2023. arXiv: `2310.19061 [cs.CV]`.

14. Javan, R., Kim, T. & Mostaghni, N. GPT-4 Vision: Multi-Modal Evolution of ChatGPT and Potential Role in Radiology. *Cureus* **16,** e68298 (2024).

15. Li, C. *et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day* 2023. arXiv: `2306.00890 [cs.CV]`.

16. Chen, Z. *et al. CheXagent: Towards a Foundation Model for Chest X-ray Interpretation* 2024. arXiv: `2401.12208 [cs.CV]`.

17. Fallahpour, A., Ma, J., Munim, A., Lyu, H. & Wang, B. *MedRAX: Medical Reasoning Agent for Chest X-ray* in *Proceedings of the 42nd International Conference on Machine Learning* **267** (PMLR, 2025), 15661–15676.

18. Wang, X. *et al. ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases* in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), 2097–2106.

19. Irvin, J. *et al. CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison* in *Proceedings of the AAAI Conference on Artificial Intelligence* **33** (2019), 590–597.

20. Nguyen, H. Q. *et al.* VinDr-CXR: An Open Dataset of Chest X-Rays with Radiologist's Annotations. *Scientific Data* **9,** 429 (2022).

21. Liu, J., Lian, J. & Yu, Y. *ChestX-Det10: Chest X-ray Dataset on Detection of Thoracic Abnormalities* 2020. arXiv: `2006.10550 [eess.IV]`.

22. Gaggion, N. *et al.* CheXmask: a Large-Scale Dataset of Anatomical Segmentation Masks for Multi-Center Chest X-Ray Images. *Scientific Data* **11,** 511 (2024).

23. Rajpurkar, P. *et al. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning* 2017. arXiv: `1711.05225 [cs.CV]`.

24. Islam, N. U. *et al. Foundation X: Integrating Classification Localization and Segmentation through Lock-Release Pretraining Strategy for Chest X-ray Analysis* in *Proceedings of the Winter Conference on Applications of Computer Vision (WACV)* (2025), 3647–3656.

25. Ma, D., Pang, J., Gotway, M. B. & Liang, J. A fully open AI foundation model applied to chest radiography. *Nature* **643,** 488–498 (2025).

26. Wang, X. *et al.* A Survey of Deep-Learning-Based Radiology Report Generation Using Multimodal Inputs. *Medical Image Analysis* **103,** 103627 (2025).

27. Chen, Z., Song, Y., Chang, T.-H. & Wan, X. *Generating Radiology Reports via Memory-Driven Transformer* in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2020), 1439–1449.

28. Zhang, Y. *et al. When Radiology Report Generation Meets Knowledge Graph* in *Proceedings of the AAAI Conference on Artificial Intelligence* **34** (2020), 12910–12917.

29. Jin, H., Che, H., Lin, Y. & Chen, H. *Diagnosis-Driven Prompts for Medical Report Generation* in *Proceedings of the AAAI Conference on Artificial Intelligence* **38** (2024), 2607–2615.

30. Wang, Z., Liu, L., Wang, L. & Zhou, L. R2GenGPT: Radiology Report Generation with Frozen Large Language Models. *Intelligence-Based Medicine* **7,** 100099 (2024).

31. Liu, C., Tian, Y., Chen, W., Song, Y. & Zhang, Y. *Bootstrapping Large Language Models for Radiology Report Generation* in *Proceedings of the AAAI Conference on Artificial Intelligence* **38** (2024), 18635–18643.

32. Paschali, M. *et al.* Foundation Models in Radiology: What, How, Why, and Why Not. *Radiology* **314,** e240597 (2025).

33. Nam, Y. *et al.* Multimodal Large Language Models in Medical Imaging: Current State and Future Directions. *Korean Journal of Radiology* **26,** 900–923 (2025).

34. Lee, S., Youn, J., Kim, H., Kim, M. & Yoon, S. H. CXR-LLaVA: a Multimodal Large Language Model for Interpreting Chest X-ray Images. *European Radiology* **35,** 4374–4386 (2025).

35. Zambrano Chaves, J. M. *et al.* A Clinically Accessible Small Multimodal Radiology Model and Evaluation Metric for Chest X-ray Findings. *Nature Communications* **16,** 3108 (2025).

36. Ahmed, F. *et al.* Explainable Artificial Intelligence (XAI) in Medical Imaging: A Systematic Review of Techniques, Applications, and Challenges. *BMC Medical Imaging* **26,** 37 (2026).

37. Bluethgen, C. *et al.* Best Practices for Large Language Models in Radiology. *Radiology* **315,** e240528 (2025).

38. Selvaraju, R. R. *et al. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization* in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2017), 618–626.

39. Houssein, E. H. *et al.* Explainable Artificial Intelligence for Medical Imaging Systems Using Deep Learning: A Comprehensive Review. *Cluster Computing* **28,** 469 (2025).

40. Bigolin Lanfredi, R. *et al.* REFLACX, a Dataset of Reports and Eye-Tracking Data for Localization of Abnormalities in Chest X-Rays. *Scientific Data* **9,** 350 (2022).

41. Ghelichkhan, E. & Tasdizen, T. *A Comparison of Object Detection and Phrase Grounding Models in Chest X-Ray Abnormality Localization Using Eye-Tracking Data* in *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)* (2025), 1–5.

42. Ghelichkhan, E. & Tasdizen, T. *LATTE-CXR: Locally Aligned TexT and imagE, Explainable Dataset for Chest X-Rays* PhysioNet dataset. 2025.

43. Bannur, S. *et al. MAIRA-2: Grounded Radiology Report Generation* 2024. arXiv: `2406.04449 [cs.CV]`.

44. Castro, D. C. *et al.* PadChest-GR: A Bilingual Chest X-Ray Dataset for Grounded Radiology Report Generation. *NEJM AI* **2,** AIdbp2401120 (2025).

45. Mahmood, R. *et al. Phrase-grounded Fact-checking for Automatically Generated Chest X-ray Reports* in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025* (2025).

46. Sharma, N. *et al. CXR-Agent: Vision-Language Models for Chest X-Ray Interpretation with Uncertainty Aware Radiology Reporting* 2024. arXiv: `2407.08811 [cs.CV]`.

47. Wang, W. *et al. A Survey of LLM-based Agents in Medicine: How Far Are We from Baymax?* in *Findings of the Association for Computational Linguistics: ACL 2025* (Association for Computational Linguistics, 2025), 10345–10359.

48. Tzanis, E. *et al.* Agentic Systems in Radiology: Principles, Opportunities, Privacy Risks, Regulation, and Sustainability Concerns. *Diagnostic and Interventional Imaging* (2025).

49. Kim, Y. *et al. MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making* 2024. arXiv: `2404.15155 [cs.CL]`.

50. Li, B. *et al. MMedAgent: Learning to Use Medical Tools with Multi-modal Agent* 2024. arXiv: `2407.02483 [cs.CL]`.

51. Schmidgall, S. *et al. AgentClinic: a Multimodal Agent Benchmark to Evaluate AI in Simulated Clinical Environments* 2024. arXiv: `2405.07960 [cs.HC]`.

52. Jiang, Y. *et al. MedAgentBench: A Realistic Virtual EHR Environment to Benchmark Medical LLM Agents* 2025. arXiv: `2501.14654 [cs.CL]`.

53. Fallahpour, A., Ma, J., Munim, A., Lyu, H. & Wang, B. *MedRAX: Medical Reasoning Agent for Chest X-ray* in *Proceedings of the 42nd International Conference on Machine Learning* **267** (PMLR, 2025), 15661–15676.

54. Kocak, B. & Cuocolo, R. Human–AI Interaction and Collaboration in Radiology: From Conceptual Frameworks to Responsible Implementation. *Diagnostic and Interventional Radiology.* Published online 26 January 2026 (2026).

55. Gu, Z., Dogra, S., Siriruchatanon, M., Kneifati-Hayek, J. & Kang, S. K. Radiology Workflow Assistance With Artificial Intelligence: Establishing the Link to Outcomes. *Journal of the American College of Radiology.* Available online 15 October 2025 (2025).

56. Chen, M. *et al.* Impact of Human and Artificial Intelligence Collaboration on Workload Reduction in Medical Image Interpretation. *npj Digital Medicine* **7,** 349 (2024).

57. Tanno, R. *et al.* Collaboration between clinicians and vision-language models in radiology report generation. *Nature Medicine* **31,** 599–608 (2025).

58. Awasthi, A., Le, N., Deng, Z., Wu, C. C. & Nguyen, H. V. Collaborative Integration of AI and Human Expertise to Improve Detection of Chest Radiograph Abnormalities. *Radiology: Artificial Intelligence* **7,** e240277 (2025).

59. Sellergren, A. *et al.* Medgemma technical report. *arXiv preprint arXiv:2507.05201* (2025).

60. Gupta, A. *itsanmolgupta/mimic-cxr-dataset: MIMIC-CXR Dataset on Hugging Face* Hugging Face Datasets. https://huggingface.co/datasets/itsanmolgupta/mimic-cxr-dataset. 2025.

61. raddar. *Chest X-rays (Indiana University)* Kaggle. `https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university`. 2020.