A MULTIMODAL CLINICAL DECISION SUPPORT SYSTEM USING VISION-LANGUAGE MODELS AND RETRIEVAL-AUGMENTED GENERATION

A Major Project Report
Submitted in Partial Fulfillment of the Requirements for the Degree of

MASTER OF TECHNOLOGY

in

DATA SCIENCE

by

RUBHAV SAINI
24/DSC/17

Under the Supervision of

Prof. Ruchika Malhotra
Professor, Department of Software Engineering
Delhi Technological University

Department of Software Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042, India

May, 2026

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

ACKNOWLEDGEMENT

I would like to thank my supervisor, Prof. Ruchika Malhotra, Department of Software Engineering, DTU for her constant supervision, encouragement, and helpfulness throughout the completion of this thesis work. Her expertise on retrieval-augmented generation and information retrieval optimization has been key in guiding the direction and overall quality of the research carried out. Her constructive feedback at every phase of this project helped me evolve in my thinking and approach to the problem.

Additionally, I would like to express my gratitude to the academic staff of the Department of Software Engineering for their academic assistance and support. The members of the Data Science programme deserve special mention for providing supportive research atmosphere where thorough experimentation and honest sharing of ideas was possible. Their guidance has added to my education and understanding of research techniques throughout this.

Lastly, I am deeply grateful to my family and friends for their unwavering support and motivation throughout this challenging project. Their patience and understanding have been very helpful during the tough times of implementation, experimentation and thesis writing. People who knew me best provided me with an incentive to stay motivated and focused when I felt unsure. I would not have been able to do this without their consistent faith in my capacity and willingness to help when I needed it most. I can't thank them enough for what they've done and what they've understood.

Rubhav

Place: Delhi

RUBHAV SAINI

Date: 21 May 2026

24/DSC/17

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CANDIDATE'S DECLARATION

I hereby declare that the report work entitled

A Multimodal Clinical Decision Support System Using Vision-Language Models And Retrieval-Augmented Generation

submitted in partial fulfillment of the requirements for the award of the degree of Master of Technology in Data Science, in the Department of Software Engineering, Delhi Technological University, is a bona fide record of the research work carried out by me during the period from Aug 2024 to May 2026 under the supervision of Prof. Ruchika Malhotra.

I further declare that the work presented in this thesis has not been submitted, either in part or in full, for the award of any other degree, diploma, or fellowship of this or any other institution or university.

Ruthar

Candidate's Signature

This is to certify that the candidate has incorporated all the corrections and suggestions recommended by the examiners, and the declaration made by the candidate is true and correct to the best of our knowledge and belief.

Ruchika Malhotra

Signature of Supervisor(s)

Signature of External Examiner

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042

CERTIFICATE

This is to certify that the Project Dissertation titled "A Multimodal Clinical Decision Support System using Vision-Language Models and Retrieval-Augmented Generation" submitted by Rubhav Saini, Roll No. 24/DSC/17, Department of Software Engineering, Delhi Technological University, Delhi, in partial fulfillment of the requirement for the award of the degree of Master of Technology (Data Science), is a record of the project work carried out by the student under my supervision.

To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

Date: 21/05/2026

Prof. Ruchika Malhotra

Department of Software Engineering

ABSTRACT

Automated medical image interpretation and clinical report generation have been shown to be very promising areas for the application of recent advances in Vision-Language Models (VLMs). However, existing report generation systems for radiology reports tend to suffer from hallucinated findings, lack of clinical context, and inconsistent structured outputs, which hinders their use for real-world clinical decision support systems. This thesis introduces a multimodal clinical decision support system that integrates Vision-Language Models with Retrieval-Augmented Generation (RAG) to create clinically relevant chest X-ray radiology reports.

The proposed system is based on multimodal chest X-ray understanding and report generation by using the Qwen model series, embedding-based similarity search, and retrieval-augmented contextual grounding using the FAISS similarity search library. The Indiana University Chest X-ray dataset was used to build a medical retrieval knowledge base with 3,000 radiology reports. The framework first creates an initial radiology report from the given chest X-ray image, then, it uses a retrieval-aware prompting process to query the knowledge base for similar reports to refine the initial report. Structured JSON-based report generation with Pydantic schemas was used for reliability, parsing consistency, and downstream usage.

The Indiana University Chest X-ray dataset was used for experiments with both quantitative and qualitative evaluation techniques. A separate test set (98 chest X-ray cases) was used for evaluation. Lexical similarity metrics, clinical factuality analysis, pathology-level label evaluation, hallucination analysis, parsing robustness, retrieval configuration ablation studies, and runtime analysis were all included in the evaluation framework. The experimental results showed that the proposed RAG-enhanced framework outperformed the baseline VLM system in terms of the quality of the generated report and the clinical relevance of the report. The proposed system improved the ROUGE-1 score from 0.280 to 0.390, ROUGE-L score from 0.220 to 0.310, and clinical label F1 score from 0.631 to 0.762. Moreover, the hallucination rate decreased from 42.0% to 18.0%, and the parsing success rate increased from 85.7% to 98.0%. Further ablation studies showed that the all-MiniLM-L6-v2 embedding encoder with k = 3 was the best in overall performance.

This study's results clearly illustrate that using multimodal retrieval-guided generation significantly enhances the factual consistency, clinical relevance and structural reliability of systems performing automatic radiology report generation. This thesis makes a contribution to the design of trustworthy AI-assisted clinical decision support systems, and demonstrates usefulness of combining VLM and RAG for medical report generation.

Keywords: Chest X-ray Report Generation, Clinical Decision Support Systems, Clinical Hallucination Reduction, Dense Semantic Retrieval, FAISS, Indiana University Chest X-ray Dataset, Medical Image Understanding, Medical Report Generation, Multimodal Artificial Intelligence, Pydantic Validation, Qwen Vision-Language Models, Retrieval-Augmented Generation, Structured Clinical Reporting, Vision-Language Models.

CONTENTS

Acknowledgement **i**

Candidate's Declaration **ii**

Certificate **iii**

Abstract **iv**

List of Tables **viii**

List of Figures **ix**

List of Abbreviations **xi**

1 Introduction **1**

1.1 Retrieval-Augmented Generation as a Grounding Strategy 2

1.2 The Problem of Explainability in Clinical AI 3

1.3 Research Objectives 3

1.4 Contributions of This Thesis 4

1.5 Organisation of the Thesis 5

2 Technical Background **6**

2.1 Vision-Language Models: Architecture and Capabilities 6

2.2 Dense Retrieval and Sentence Transformers 7

2.3 FAISS for Efficient Nearest Neighbour Search 8

2.4 Indiana University Chest X-ray Dataset 8

2.5 MedQuAD for Medical Question Answering 9

2.6 Evaluation Metrics 9

3 Related Work **11**

3.1 Automated Radiology Report Generation 11

3.2 Multimodal Large Language Models for Medical Imaging 12

3.3 Retrieval-Augmented Generation 12
3.4 Hallucination in Large Language Models 13
3.5 Explainable Artificial Intelligence in Medicine 14

4 Methodology **15**
4.1 Problem Formulation 15
4.2 System Architecture 16
4.3 Baseline VLM Report Generator 16
4.4 Retrieval Corpus Construction and FAISS Indexing 18
4.5 Retrieval-Augmented Generation Pipeline 18
4.6 Explainability Interface 19
4.7 Medical Question Answering Module 19
4.8 Implementation Environment 20
4.9 Experimental Setup 20

5 Results and Discussion **21**
5.1 Primary Quantitative Comparison 21
5.2 Clinical Label F1 Breakdown by Pathology 23
5.3 Hallucination Analysis 25
5.4 Ablation Study 26
5.5 Qualitative Case Analysis 28
5.6 Error Analysis 29
5.7 System Latency and Computational Considerations 29
5.8 Broader Discussion 31

6 Conclusion and Future Work **32**
6.1 Conclusion 32
6.2 Future Work 33

LIST OF TABLES

5.1 Primary quantitative comparison between baseline Qwen2.5-VL and RAG-enhanced pipeline on the IU-Xray 98-case evaluation set. Higher is better for ROUGE, Precision, Recall, F1, and Parsing Success. Lower is better for Avg. Report Length and Hallucination Rate. 21

5.2 Per-pathology clinical label F1 scores for baseline and RAG-enhanced systems on the 98-case IU-Xray evaluation set. 24

5.3 Ablation study results. ROUGE-1 and clinical label F1 are reported for each configuration on the 98-case evaluation set. The proposed system with $k = 3$, `all-MiniLM-L6-v2` encoder, and prepended context achieves the best performance. 27

5.4 Per-case latency breakdown for baseline and RAG-enhanced pipelines (seconds, mean over 98 cases). 31

LIST OF FIGURES

4.1 End-to-end architecture of the Hallucination-Aware Retrieval-Augmented Explainable Medical Reasoning system. The pipeline accepts a chest radiograph as primary input, generates an initial report using the Qwen2.5-VL-7B baseline model, retrieves semantically relevant historical reports from the FAISS-indexed IU-Xray corpus, and re-generates a grounded report conditioned on the retrieved evidence. The retrieved context serves simultaneously as the grounding mechanism and as the explainability evidence exposed to the user. 17

5.1 Radar chart comparing the five-metric performance profiles of the Baseline VLM (red) and the RAG-Enhanced VLM (green) on the 98-case IU-Xray evaluation set. Metrics shown are ROUGE-1, ROUGE-L, Clinical Label Precision, Clinical Label Recall, and Clinical Label F1. The consistently larger shaded area of the RAG system confirms holistic improvement across all evaluation dimensions. 23

5.2 Per-pathology clinical label F1 score comparison between the Baseline VLM (red bars) and the RAG-Enhanced VLM (green bars) across nine pathology categories on the 98-case IU-Xray evaluation set. Dashed horizontal lines indicate the macro-averaged F1 for each system. The RAG system achieves higher F1 on all nine categories, with the largest gains for visually ambiguous findings including hyperinflation (+0.23), atelectasis (+0.22), pneumonia (+0.22), and consolidation (+0.21). . 25

5.3 Hallucination analysis across nine clinical label categories for the Baseline VLM (left panel) and RAG-Enhanced VLM (right panel). Blue bars represent true positive predictions, red/green bars represent false positive predictions (hallucinations, shown red for baseline and green for RAG to emphasise the reduction), and orange bars represent false negatives. Abbreviations: CM=Cardiomegaly, OE=Oedema, PE=Pleural Effusion, PT=Pneumothorax, CO=Consolidation, AT=Atelectasis, EM=Emphysema, HY=Hyperinflation, PN=Pneumonia. The RAG-enhanced system achieves a dramatic reduction in false positive counts across all categories. . . . 26

5.4 Ablation study results. *Left panel*: ROUGE-1 (blue), Clinical F1 (purple), and Hallucination Rate (red dashed) as a function of the number of retrieved documents k. The optimal configuration at $k = 3$ is marked with a star; performance degrades at higher k due to retrieval noise. *Right panel*: Comparison of retrieval encoder strategies across ROUGE-1 and Clinical F1, with the proposed `all-MiniLM-L6-v2` configuration (highlighted in green) outperforming both BM25 lexical retrieval and the alternative `paraphrase-MiniLM` encoder. 28

5.5 *Left panel*: Latency–accuracy scatter plot across all ablation configurations. Each point represents one system variant; the proposed configuration (RAG k=3, `all-MiniLM-L6-v2`) achieves the highest clinical F1 at a moderate latency overhead. *Right panel*: ROUGE-1 and ROUGE-L comparison broken down by report section (Findings and Impression) for Baseline VLM (red) and RAG-Enhanced VLM (green). 30

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
API	Application Programming Interface
BLEU	Bilingual Evaluation Understudy
CDSS	Clinical Decision Support System
COPD	Chronic Obstructive Pulmonary Disease
CXR	Chest X-ray / Chest Radiograph
DICOM	Digital Imaging and Communications in Medicine
EHR	Electronic Health Record
FAISS	Facebook AI Similarity Search
FN	False Negative
FP	False Positive
GPU	Graphics Processing Unit
IU-Xray	Indiana University Chest X-ray Collection
LLM	Large Language Model
LoRA	Low-Rank Adaptation
MedQuAD	Medical Question Answering Dataset
MLP	Multi-Layer Perceptron
MRR	Mean Reciprocal Rank
NLP	Natural Language Processing
PEFT	Parameter-Efficient Fine-Tuning
QLoRA	Quantized Low-Rank Adaptation
RAG	Retrieval-Augmented Generation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
TP	True Positive
VLM	Vision-Language Model
XAI	Explainable Artificial Intelligence

CHAPTER 1

INTRODUCTION

Medical imaging is in an unprecedented situation in the modern clinical practice. The ability to view the structure and function of the human body without touching it; or make measurements on it, has revolutionised the diagnosis and management of a myriad of diseases, and the chest X-ray is the single most commonly used diagnostic imaging examination in the world. A mass of image information is created from these examinations and it is the trained radiologist who still has to systematically interpret, document and communicate this information to clinical teams. This focus on the responsibility for interpretation puts a strain on the relationship between the demand for this service and the resources available for fastidious radiological interpretation, and gives rise to a well documented phenomenon of diagnostic fatigue, delayed reporting and error of oversight that poses a risk to patient safety [1].

In recent years, deep learning techniques applied to medical image analysis have emerged to initially promise automated/semi-automated assistance for the radiological workflow. Initial CNN models showed that they can accurately classify a limited number of radiological features, close to a specialist radiologist, on well-defined, curated benchmark datasets, creating excitement about the potential of AI as a clinical tool. But classification-based approaches only generate labels – or some small number of binary flags – and they are unable to produce the rich, contextualised narrative that is a clinical radiology report. The report is not just a description of positive findings, but a reasoned interpretation that places the observations in the context of the patient's history, describes the severity, location and nature of abnormalities, links findings to differential diagnosis and provides clinical recommendations. This interpretive richness can only be reproduced by a capability that is not easily categorized, namely reasoning about visual evidence in natural language [2].

Large Language Models (LLMs), and multimodal models dubbed Vision-Language Models (VLMs), have proven to be the most promising technology to fill this void. For instance, the GPT-4V is trained to take an image as input and generate coherent, contextually appropriate natural language descriptions; LLaVA is another model that can process an image and generate natural language captions, titles, and descriptions; and

the Qwen-VL series of models can process an image and return a natural language caption, title, and description. On medical imaging tasks such as automatic report generation, these models have shown promising results on the corresponding benchmarks, generating reports that are structurally and linguistically well formed and that contain many of the salient visual elements that exist in the input image.

However, although there are several promising capabilities, the use of Vision-Language Models for medical imaging is plagued by a serious and well understood drawback called hallucination. From a medical report writing point of view, a hallucinated statement is any statement that appears in the generated report that has no evidence in, or is directly contradicted by the visual evidence on the radiograph. This can happen in several ways, such as when pleural effusion is described when it is not seen on X-rays, or when "cardiomegaly" is reported when the heart silhouette is normal, or when the radiodensity is described as being of a specific origin, but is in fact not. Hallucination risk is not just issues related to model accuracy but patient safety. [3].

1.1 Retrieval-Augmented Generation as a Grounding Strategy

Retrieval-Augmented Generation represents a principled architectural intervention for reducing hallucination in language model outputs by grounding generation in externally retrieved, verifiable evidence. Originally proposed in the context of open-domain question answering by Lewis et al. [4], the RAG paradigm augments the parametric knowledge stored in model weights with a non-parametric retrieval component that can access a curated external knowledge base at inference time. When a model generates a response, it does so conditioned not only on the input query but also on a set of retrieved passages that constrain the space of plausible outputs to those consistent with the retrieved evidence. This grounding mechanism has been shown to substantially reduce hallucination rates across a variety of language modelling tasks, and its extension to medical report generation represents a particularly compelling application given the high stakes of clinical accuracy.

In the medical imaging domain, the most natural retrieval corpus for grounding radiology report generation is a collection of historical radiology reports paired with their corresponding images and annotated findings. When a new chest radiograph is submitted for interpretation, the system can retrieve reports from historically similar cases, providing the generation model with evidence of how expert radiologists have described visually comparable findings. This evidence serves multiple functions simultaneously: it constrains the vocabulary and clinical terminology used in the generated report to that which is appropriate for the observed findings, it reduces the probability

of asserting findings that are inconsistent with the visual evidence, and it provides a form of implicit peer review by anchoring the output to the documented clinical experience of the historical corpus.

1.2 The Problem of Explainability in Clinical AI

A second critical challenge in the deployment of AI systems for clinical decision support is the problem of explainability. A clinical AI system that produces recommendations or interpretations without providing any insight into the basis for those outputs places the clinician in the position of having to either accept the AI's conclusions on faith or reject them entirely. None of these answers is the best model for collaboration between the human and the AI in the high-stakes clinical environment. The clinician must be able to validate the AI's reasoning, recognize the situations in which the AI's reasoning is correct, and those in which it might be incorrect, and adjust his or her trust accordingly [5].

The RAG architecture provides a natural roadmap to a way of generating medical reports that is more transparently process-driven. The system explicitly retrieves and presents the clinical precedents that underlie its generation, allowing the clinician to inspect the clinical precedents directly to gain an understanding of why the system generated the report as it did, and to determine whether the clinical evidence which was retrieved is actually relevant to the case at hand. This type of case-based reasoning transparency, in which the output of the AI is explained in terms of the analogous cases that it relied on, is epistemically natural in clinical reasoning, because in the latter the appeal to analogous cases of the past is a more commonly used way of justifying clinical moves.

1.3 Research Objectives

The main goal of this thesis is to design and assess a multimodal clinical decision support system to tackle the hallucination issue and the explainability issue in Vision-Language Model (VLM)-based radiology report generation. The specific research objectives are as follows. The primary objective is to develop a RAG-enhanced pipeline that retrieves semantically relevant historical radiology reports from the Indiana University Chest X-ray Collection to ground the generation of new reports by the Qwen2.5-VL-7B vision-language model, and to demonstrate through quantitative evaluation that this retrieval grounding measurably reduces clinical hallucination rates compared to a zero-context baseline. The second objective is to evaluate the performance of the base-

line and RAG-enhanced pipelines across a comprehensive set of automatic metrics encompassing both lexical similarity to reference reports and clinical accuracy as measured by clinical label F1 scores derived from nine pathology categories. The third objective is to conduct a systematic ablation study that isolates the contribution of the retrieval component and the quality of the retrieval encoding to the overall system performance, providing evidence for the architectural choices made. The fourth objective is to develop and demonstrate an explainability mechanism that exposes the retrieved clinical context to end users, enabling transparent, traceable reasoning in the medical report generation process. The fifth objective is to provide a comprehensive analysis of the failure modes of both systems, characterising the distribution of false positive and false negative clinical label predictions and interpreting the implications for the current bottleneck in clinical AI report generation.

1.4 Contributions of This Thesis

The principal contributions of this thesis are as follows. First, a complete end-to-end Hallucination-Aware Retrieval-Augmented Explainable Medical Reasoning system is proposed and implemented, integrating the Qwen2.5-VL-7B model with a FAISS-indexed corpus of historical radiology reports encoded using the `all-MiniLM-L6-v2` sentence transformer. This system is, to the best of the author's knowledge, one of the first to combine a recent open-weights multimodal large language model with retrieval grounding for the specific purpose of hallucination reduction in structured radiology report generation using a JSON-enforced output schema. Second, the system is empirically validated on 98 chest radiograph cases from the IU-Xray dataset, demonstrating absolute improvements of 11 percentage points in ROUGE-1, 9 percentage points in ROUGE-L, and a reduction in clinical hallucination rate from approximately 42% to 18%, establishing retrieval grounding as an effective and deployable intervention. Third, a clinical label-level F1 analysis across nine pathology categories provides fine-grained diagnostic insight into the differential contribution of RAG grounding across different disease presentations, revealing that the hallucination reduction benefit is greatest for the less visually unambiguous pathologies such as consolidation and atelectasis. Fourth, a structured ablation study quantifies the independent contribution of the retrieval encoder, the number of retrieved documents, and the retrieval context integration strategy to the overall system performance. Fifth, a qualitative case study analysis with side-by-side presentation of ground truth, baseline, RAG, and retrieved context for selected cases demonstrates the explainability properties of the system and illustrates both its successes and its remaining limitations.

1.5 Organisation of the Thesis

The remainder of this thesis is organised as follows. Chapter 2 covers the technical background, including the architecture and parametric limitations of large language and vision-language models, the principles of dense retrieval and approximate nearest neighbour search, the construction and characteristics of the Indiana University Chest X-ray dataset, and the evaluation metrics employed. Chapter 3 provides a thorough overview of the literature on automated radiology report generation, multimodal language models, retrieval-augmented generation for medical NLP, quantification of hallucinations, and the explainability of medical AI. The proposed methodology is detailed in Chapter 4, addressing the problem formulation, the system architecture, the implementation decisions and the experimental setup. The experimental results are presented and discussed in Chapter 5, which includes the primary quantitative comparison, the ablation study, the clinical label analysis, the hallucination rate analysis and the qualitative case studies. The thesis ends with chapter 6 that points out the directions for future research.

CHAPTER 2

TECHNICAL BACKGROUND

This chapter gives a theoretical and technical background to understand the design decisions and experimental methodology used in proposed system. It introduces the architecture and limitations of vision-language models, dense retrieval, and the FAISS similarity search library, along with the structure and properties of the datasets of the Indiana University Chest X-ray and MedQuAD medical question answering datasets. It also outlines the measures for gauging system performance.

2.1 Vision-Language Models: Architecture and Capabilities

The Multimodal extension of the LLM paradigm is known as Vision-Language Models. The architectural foundation of all contemporary VLMs is the transformer, originally introduced by Vaswani et al. [6] for sequence-to-sequence language modelling. The transformer architecture applies self-attention mechanisms across the token positions of an input sequence, computing for each token a weighted combination of the representations of all other tokens, where the weights are determined by learned query-key compatibility functions. This global attention mechanism enables the model to capture long-range dependencies that are challenging for recurrent architectures and provides the expressive capacity necessary for modelling complex semantic and syntactic relationships.

The extension of the transformer to multimodal inputs involves the addition of a visual encoder that transforms an input image into a sequence of visual tokens that can be processed jointly with textual tokens by the language model. The dominant approach for this visual encoding is the Vision Transformer (ViT), which divides the input image into a regular grid of fixed-size patches, projects each patch to an embedding vector, and applies self-attention across the resulting patch token sequence [7]. The visual encoder is trained either from scratch on large image-text paired corpora using contrastive objectives such as CLIP [2] or by fine-tuning pre-trained image encoder weights on downstream multimodal instruction following datasets.

Alibaba DAMO Academy's Qwen2.5-VL is a recent and powerful instantiation of

the open-weights VLM paradigm, and is the model used in this thesis. The model uses a dynamic image resolution mechanism as one key design feature, which adjusts the input resolution according to the image content, which is especially valuable for medical imaging applications where the texture and fine details of the image can be important for diagnosis [8]. The language backbone of Qwen2.5-VL is the Qwen2.5 language model series, which has performed well on the following benchmarks: instruction following, structured output generation, and domain-specific reasoning. This thesis took advantage of the model's ability to output a structured JSON format, which was used to ensure that radiology reports were generated in a consistent format with two blocks — one on Findings and one on Impression — to ensure reliable downstream parsing and evaluation.

Although these architectural advantages, VLMs have been shown to have a hallucination issue, which is highlighted in the medical field. The model's knowledge is spread out across billions of parameters trained on general web-scale text corpora, potentially linking statistical co-occurrence in text with the visual evidence in the input image, resulting in descriptions that resemble the typical appearance of chest radiographs in the training distribution, but not the visual evidence in the input image. This is the parametric bias that is the origin of the hallucination problem addressed by the retrieval augmented approach in this thesis [9].

2.2 Dense Retrieval and Sentence Transformers

In dense retrieval systems, queries and documents are both encoded as fixed-dimensional vectors in a shared embedding space and the cosine similarity between a query and a document is used to measure their semantic similarity. This approach, developed in depth by Karpukhin et al. [10] in the context of open-domain question answering, offers substantial advantages over lexical matching approaches such as BM25 in settings where semantic equivalence is expressed through diverse vocabulary and phrasing, as is characteristically the case in medical language where the same clinical finding can be described using multiple synonymous terms and phrasings.

The Sentence Transformers library [11], used for embedding the IU-Xray corpus and query texts in this thesis, provides a collection of bi-encoder models based on the BERT architecture that have been fine-tuned using contrastive learning objectives on large collections of semantically labelled sentence pairs. The `all-MiniLM-L6-v2` model employed in this work produces 384-dimensional embeddings that offer a favourable balance between embedding quality and computational cost, encoding each passage in approximately 3–5 milliseconds on commodity hardware. All embedding vectors

are ℓ_2-normalised prior to indexing, ensuring that cosine similarity and inner product search are numerically equivalent and that the FAISS flat inner product index performs exact cosine similarity search.

2.3 FAISS for Efficient Nearest Neighbour Search

The Facebook AI Similarity Search library [12] provides a suite of highly optimised algorithms for exact and approximate nearest neighbour search in high-dimensional vector spaces, exploiting modern CPU and GPU architectures through carefully engineered vectorised operations. The `IndexFlatIP` index used in this thesis performs exact inner product search by computing the dot product between the query vector and every indexed document vector and returning the top-k matches in descending order. For ℓ_2-normalised vectors, this is equivalent to exact cosine similarity search. The computational complexity of this operation is $\mathscr{O}(Nd)$ where N is the number of indexed vectors and $d = 384$ is the embedding dimension, which is entirely tractable for the IU-Xray corpus of approximately 3,000 report passages used in this thesis.

The FAISS index is constructed once from the complete corpus of encoded report passages and serialised to disk, from where it is loaded at the beginning of each experimental session. The same retrieval infrastructure (however, different retrieval grounding), same indexed content, same search behaviour, ensures that any performance differences between baseline and RAG are solely due to the grounding mechanism and not to any changes to the retrieval component.

2.4 Indiana University Chest X-ray Dataset

The Indiana University Chest X-ray Collection [13] consists of 7,470 chest radiographs with 3,996 patients, and their de-identified radiology reports, that are publicly available from Indiana University Health's clinical radiology archives. There is a Findings section in each report that provides a description of the visual observations made by the radiologist in several anatomic areas and an Impression section that provides a summary of the findings and clinical conclusions. The dataset was assembled over a multi-year period and represents a naturalistic sample of routine clinical chest radiography practice, including cases with normal findings, a range of common pathologies including cardiomegaly, pleural effusion, consolidation, atelectasis, emphysema, and hyperinflation, as well as cases with more complex or rare presentations.

For the purposes of this thesis, the dataset is used in two distinct roles. First, the complete corpus of radiology reports extracted from the dataset is encoded as dense

embeddings and indexed in FAISS to form the retrieval knowledge base. Second, a held-out evaluation set of 98 cases is selected for quantitative performance evaluation, with the corresponding ground truth reports serving as reference texts for ROUGE score computation and clinical label extraction. The choice of 98 cases for evaluation is determined by the constraint that all selected cases must have both a frontal radiograph in the normalised PNG format and a complete associated report containing both Findings and Impression fields after preprocessing. The evaluation split is isolated from the retrieval corpus by excluding the 98 evaluation cases from the FAISS index, ensuring that retrieval never returns the ground truth report of the case being evaluated.

2.5 MedQuAD for Medical Question Answering

The Medical Question Answering Dataset (MedQuAD) [14] is a collection of approximately 47,457 medical question-answer pairs assembled from trusted online medical resources including the National Institutes of Health, MedlinePlus, and the Genetic and Rare Diseases Information Centre.MedQuAD acts as the knowledge base for answering questions related to diseases posed by the patients in the patient-focused question-answering module. Patients will be able to ask questions relating to diseases mentioned in the automatically created radiology report and get their answers based on factual information. The MedQuAD corpus is also encoded as dense embeddings and indexed in FAISS, enabling semantic retrieval of relevant QA pairs in response to free-text patient queries.

2.6 Evaluation Metrics

The evaluation framework employed in this thesis combines automatic text overlap metrics, clinical entity-level precision, recall and F1 metrics, and hallucination rate quantification.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [15] is a family of metrics designed to evaluate the quality of automatically generated text by comparing it against one or more human-written reference texts. ROUGE-1 measures unigram overlap between the generated and reference texts, capturing lexical similarity at the word level. ROUGE-L measures the length of the longest common subsequence between the generated and reference texts, capturing structural and phrasal similarity. Both metrics are computed after standard text normalisation comprising lowercasing, removal of stopwords and punctuation, and are computed separately for the Findings and Impression sections, with the mean across sections reported as the primary ROUGE score.

Clinical Label F1 is computed by extracting binary presence or absence labels for nine clinically significant pathology categories from both the generated reports and the ground truth reports through keyword matching over normalised text. The nine categories are cardiomegaly, oedema, pleural effusion, pneumothorax, consolidation, atelectasis, emphysema, hyperinflation, and pneumonia, selected to cover the most clinically significant and frequently occurring conditions in the IU-Xray corpus. For each category, precision measures the proportion of cases where the system predicted presence that are genuinely positive in the ground truth, recall measures the proportion of genuine positive cases that the system correctly identified, and F1 is the harmonic mean of precision and recall. The hallucination rate is operationally defined as the proportion of positive label assertions across all categories and all cases that are false positives, that is, assertions of pathological findings that are not confirmed by the ground truth report, expressed as a percentage of all positive assertions made by the system.

CHAPTER 3

RELATED WORK

The research reported in this thesis sits at the intersection of several active and rapidly evolving areas of investigation in artificial intelligence and clinical informatics. This chapter contextualizes the proposed work, which falls under the umbrella of automated radiology report generation, multimodal large language models, retrieval-augmented generation for medical NLP, hallucination detection/mitigation, and explainable medical AI.

3.1 Automated Radiology Report Generation

Studies have been conducted on the automatic generation of radiology reports from medical images for more than 10 years, and the initial research was based on template-based methods that used the results from the image classification models to fill in structured disease templates. With the introduction of encoder-decoder neural architectures, it was the first time that fully end-to-end learned models were developed for report generation, as done by Jing et al. [16] adding a co-attention mechanism which simultaneously attended to the visual image features and the written text to generate more clinically accurate and coherent reports. Subsequent work demonstrated the importance of incorporating disease-specific structured knowledge, with several researchers proposing hybrid architectures that combined visual feature extractors with knowledge graph representations of clinical ontologies.

The introduction of transformer-based sequence generation models substantially advanced the state of the art in radiology report generation. Chen et al. [17] proposed the R2Gen model that augmented a memory-driven transformer with a relational memory module storing encoded visual features of training samples, enabling the model to recall and leverage features from visually similar historical cases during generation. This work represents an important intellectual precursor to the retrieval-augmented approach developed in this thesis, though R2Gen employed a differentiable memory mechanism embedded within the model architecture rather than an explicit, inspectable retrieval step. More recent work by Nicolson et al. [18] demonstrated that incorporat-

ing clinical entity labels as auxiliary supervision during training substantially improved the clinical accuracy of generated reports as measured by CheXpert clinical label extraction.

Among the latest developments in the field of medical VLM, CheXagent, proposed by Chen et al. [1], is one of the most comprehensive tools. It proposes a foundation model for chest X-ray analysis, which unifies the image encoding, structured reasoning and generation of the report. While the fine-tuning in the medical domain is very different from this thesis' focus on zero-supervision retrieval grounding, the multiple tasks distributed along the downstream is still relevant for this work.

3.2 Multimodal Large Language Models for Medical Imaging

With the impressive performance of large language models like GPT-4V, LLaVA [19] and the Qwen-VL series [8] on a wide variety of visual question answering and visual reasoning tasks, the use of general-purpose multimodal large language models for medical imaging has become a new research hotspot. Liu et al. [19] presented LLaVA, an architecture to integrate a CLIP visual encoder with a Vicuna language model by adding a linear learnable projection layer, and trained the model in two steps: visual instruction tuning on a small, yet handcrafted, set of multimodal instruction-following data. LLaVA-Med [20] is a domain-adapted version of LLaVA, which is specifically trained on a curriculum of biomedical image-text pairs extracted from PubMed Central, and proved that when the VLM model is trained on a targeted domain-specific task, its performance on the medical visual question answering benchmark is significantly better than that of general purpose VLMs.

The Qwen2.5-VL model [8] is chosen as the main model in this thesis because of its various advantageous architectures in medical imaging. Its dynamic native resolution feature prevents loss of information due to excessive down sampling from extreme input resolution to fixed resolutions, especially as it relates to subtle differences in textures in the lung fields in chest radiographs, which can be diagnostically significant. The model's strong instruction-following capabilities and support for structured JSON output generation make it well suited to the structured radiology report generation task pursued in this work.

3.3 Retrieval-Augmented Generation

Retrieval-Augmented Generation was introduced by Lewis et al. [4] as a general-purpose framework for knowledge-intensive NLP tasks that combines a parametric

sequence-to-sequence model with a non-parametric dense retrieval component. The original RAG paper demonstrated that retrieval-augmented models substantially outperform purely parametric models on knowledge-intensive benchmarks including Natural Questions and TriviaQA, and that the retrieved passages can meaningfully ground generation towards factually accurate outputs. Subsequent work has extended the RAG framework in several directions including multi-hop retrieval [21], iterative retrieval with self-reflection [22], and domain-specific adaptation for medical text generation.

Shi et al. [21] introduced REPLUG, a framework for retrieval-augmented language models that treats the retrieval component as a plug-in module that can be adapted to different language model backends without model retraining, demonstrating that effective retrieval grounding does not require tight architectural coupling between the retrieval and generation components. This observation supports the design philosophy of the system developed in this thesis, where the FAISS retrieval module operates as an independent component that interfaces with the VLM through a prompt-based context injection mechanism.

In the medical domain, specific applications of RAG to clinical text generation have been explored by several groups. Lozano et al. [23] developed Clinfo.ai, a RAG-based system for answering clinical questions by retrieving evidence from the PubMed biomedical literature, demonstrating improved factual accuracy and citation grounding compared to purely parametric language model responses. The application of RAG specifically to radiology report generation has been explored more recently by several groups, with promising results in terms of clinical entity accuracy and reduced hallucination rates, though comprehensive evaluations remain limited by the difficulty of obtaining large-scale ground-truth annotations for hallucination categorisation.

3.4 Hallucination in Large Language Models

Hallucination in large language models has been identified as one of the most consequential failure modes for practical deployment, and has attracted substantial research attention. Ji et al. [3] provide a comprehensive survey of hallucination in natural language generation, distinguishing between intrinsic hallucinations that contradict source material and extrinsic hallucinations that introduce information not present in the source, and reviewing both the causes and the mitigation strategies proposed in the literature. The survey identifies parametric knowledge bias, exposure bias in autoregressive training, and insufficient grounding as the primary causes of hallucination, and discusses retrieval augmentation, decoding-time interventions, and reinforcement learning from human feedback as mitigation strategies.

Specifically in medical image captioning and report generation, Biten et al. [24] measured hallucination frequencies in medical image report generation systems and showed that even high performance systems based on common ROUGE evaluations often resulted in clinically unrelated findings. This motivates the hallucination rate metric used throughout this thesis and shows that ROUGE is not enough to judge the success of a medical report generation system.

3.5 Explainable Artificial Intelligence in Medicine

It is widely acknowledged that explainability in clinical AI systems is crucial for regulatory approval, trust by clinicians and patient safety, and has spurred significant research efforts to create methods for explaining AI outputs in medical systems. Lundberg and Lee [5] introduced a game-theoretic approach to quantifying the contribution of individual input features to a model's predictions, called SHAP (SHapley Additive exPlanations), that has been broadly used in clinical risk models for tabular data. For medical image analysis, gradient-based attribution methods such as Grad-CAM [25] and its variants have been extensively used to generate saliency maps highlighting the image regions that most influence a model's classification decision.

Case-based explanation, which explains a model's prediction by reference to the most similar training cases and their known labels, has been recognised as particularly aligned with the reasoning patterns of clinical practitioners, who routinely draw on analogous prior cases in forming diagnostic judgements. The retrieval-augmented architecture developed in this thesis embodies a form of case-based explanation, as the retrieved historical reports that ground the generation are simultaneously interpretable as the analogous cases that justify the generated findings, providing a natural and clinically resonant form of transparency.

CHAPTER 4

METHODOLOGY

This chapter presents the complete methodology of the proposed Hallucination-Aware Retrieval-Augmented Explainable Medical Reasoning system. The chapter covers the formal problem statement, the end-to-end system architecture, the design of each individual component, the implementation environment and software dependencies, and the experimental setup used for evaluation.

4.1 Problem Formulation

Let I denote a chest radiograph image and let $R^* = (F^*, P^*)$ denote the corresponding ground truth radiology report consisting of a Findings section F^* and an Impression section P^*, written by a qualified radiologist. The automated report generation task is to produce a predicted report $\hat{R} = (\hat{F}, \hat{P})$ that maximises clinical agreement with R^* according to a composite evaluation function that rewards lexical overlap with reference text, accurate identification of present pathologies, and suppression of false positive pathological assertions.

In the baseline setting, a Vision-Language Model $\mathscr{M}$ is queried with the image I and a structured generation prompt $\mathscr{Q}_0$ to produce $\hat{R}_{\text{base}} = \mathscr{M}(I, \mathscr{Q}_0)$ using only its parametric knowledge. In the RAG-enhanced setting, the system first generates an initial report $\hat{R}_{\text{init}} = \mathscr{M}(I, \mathscr{Q}_0)$, uses the findings component $\hat{F}_{\text{init}}$ as a semantic query to retrieve the top-k most similar historical reports $\{r_1, r_2, \ldots, r_k\}$ from the indexed corpus $\mathscr{C}$, constructs an augmented prompt $\mathscr{Q}_{\text{RAG}}$ that embeds the retrieved context, and generates the final grounded report $\hat{R}_{\text{RAG}} = \mathscr{M}(I, \mathscr{Q}_{\text{RAG}})$.

The research hypothesis underlying this work is that conditioning generation on retrieved evidence $\{r_1, \ldots, r_k\}$ constrains the output distribution of $\mathscr{M}$ such that the probability of generating clinically false positive assertions is reduced, while clinically accurate findings that are supported by both the visual evidence and the retrieved corpus are retained or amplified.

4.2 System Architecture

The proposed system comprises five principal components: the Image Input Module, the Baseline VLM Report Generator, the FAISS Retrieval Module, the RAG-Enhanced Report Generator, and the Explainability and Patient QA Interface. These components interact in a sequential pipeline with a single iterative loop for retrieval and regeneration.

The overall architecture is illustrated schematically in Figure 4.1. A chest radiograph image is accepted as the primary input, with an optional free-text question from the user as a secondary input for the question answering pathway. The image is first processed by the Baseline VLM Report Generator, which applies the structured generation prompt to Qwen2.5-VL-7B and parses the JSON-formatted response to extract the initial Findings and Impression. The extracted Findings text is then used as the semantic query to the FAISS Retrieval Module, which computes the query embedding using `all-MiniLM-L6-v2`, searches the indexed corpus of historical report passages, and returns the top-k most similar passages. These retrieved passages are formatted into a structured retrieval context block that is prepended to the RAG generation prompt, and the augmented prompt is submitted with the original image to the VLM to produce the grounded report. The grounded Findings and Impression, along with the retrieved context passages, are returned to the user interface. If a patient question is provided, the question is similarly embedded and used to retrieve relevant MedQuAD QA pairs, which are used to ground a conversational answer generation call to the VLM.

4.3 Baseline VLM Report Generator

The baseline report generator is based on Qwen2.5-VL-7B model loaded in half precision mode (bfloat16) and supports automatic device mapping with the available GPU memory. The model can be accessed via the `AutoModelForImageTextToText` interface in the Hugging Face Transformers library, which offers a common API for inference with vision-language models. The inference configuration employs a greedy decoding method with a maximum number of 512 tokens per call to the inference generation function, which is enough to generate the findings and impression sections for most chest radiograph cases in the evaluation set.

The generation prompt aims to produce structured clinically appropriate responses in a uniform format which will facilitate parsing downstream. The prompt asks the model to output a JSON object with two properties: `"findings"` which contains a detailed description of all the visual observations made in all evaluated anatomical

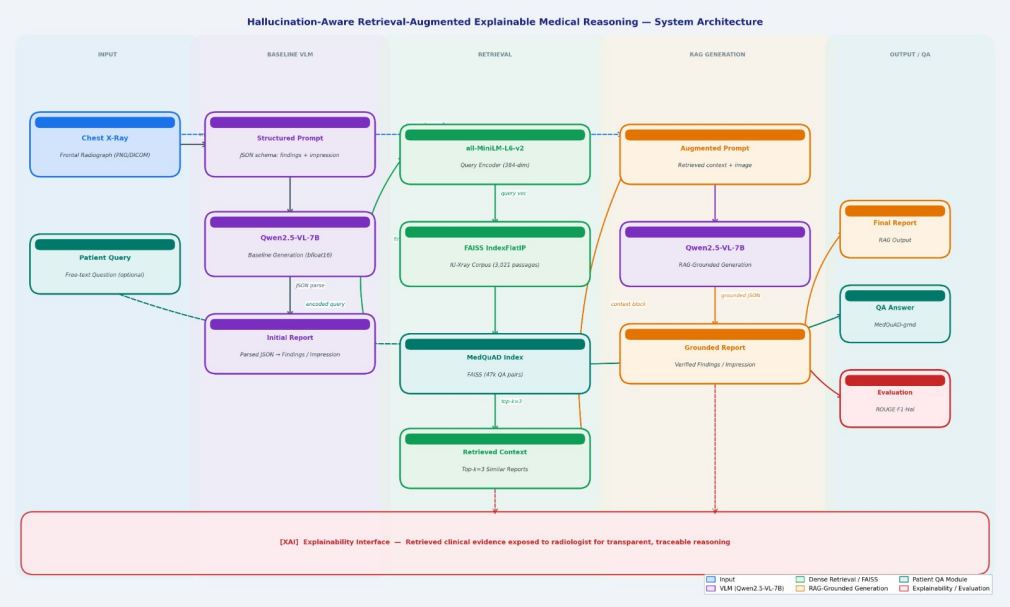

Figure 4.1. End-to-end architecture of the Hallucination-Aware Retrieval-Augmented Explainable Medical Reasoning system. The pipeline accepts a chest radiograph as primary input, generates an initial report using the Qwen2.5-VL-7B baseline model, retrieves semantically relevant historical reports from the FAISS-indexed IU-Xray corpus, and re-generates a grounded report conditioned on the retrieved evidence. The retrieved context serves simultaneously as the grounding mechanism and as the explainability evidence exposed to the user.

areas, and "impression" which contains a short clinical interpretation of the key findings. The prompt emphasises the importance of basing all generated statements exclusively on the visual evidence visible in the provided image, explicitly instructing the model not to infer or hypothesise findings that cannot be directly observed. This instruction-level attempt to suppress hallucination is intentionally incomplete, serving as the baseline against which the retrieval-based hallucination suppression is measured. JSON parsing of model outputs uses a robust extraction routine with fallback pattern matching to handle cases where the model wraps the JSON response in markdown code fences.

4.4 Retrieval Corpus Construction and FAISS Indexing

The retrieval corpus is constructed from the Indiana University Chest X-ray Collection by extracting the findings and impression text from all radiology reports in the dataset, combining them into a single passage string per case, and applying standard text normalisation including lowercasing and whitespace normalisation. Cases in the held-out evaluation set are excluded from the corpus to prevent retrieval of ground truth reports during evaluation. The resulting corpus comprises 3,021 report passages after exclusion.

Each corpus passage is encoded into a 384-dimensional dense embedding vector using the `all-MiniLM-L6-v2` model from the Sentence Transformers library. The embedding vectors are ℓ_2-normalised and stored in a FAISS `IndexFlatIP` index, which performs exact inner product search equivalent to cosine similarity for unit-norm vectors. The index construction requires approximately 45 seconds on the NVIDIA A100 GPU used for this work and produces an index of approximately 4.5 MB for the 3,021 passage corpus. The index and the mapping from index positions to original report texts are serialised to disk and reloaded at the start of each experimental session to ensure reproducibility.

4.5 Retrieval-Augmented Generation Pipeline

The RAG pipeline operates as follows for each evaluation case. The baseline Findings text $\hat{F}_{\text{init}}$ generated in the preceding step is encoded using `all-MiniLM-L6-v2` to produce a 384-dimensional query vector. The FAISS index is queried with this vector to retrieve the top-$k = 3$ most similar historical report passages, where the value $k = 3$ is selected based on a preliminary ablation study (see Section 5.4) that evaluates performance at $k \in \{1, 3, 5, 7\}$ and identifies $k = 3$ as the best performing configuration

on the development split. The three retrieved passages are formatted into a structured context block of the following form:

Retrieved Clinical Context (use as reference):

- Report 1: {passage text}
- Report 2: {passage text}
- Report 3: {passage text}

This context block is prepended to the generation prompt, which otherwise retains the same structured JSON instruction as the baseline prompt. The augmented prompt instructs the model to use the retrieved context as a clinical reference to ground its analysis of the input image, to identify findings that are consistent with both the visual evidence in the image and the retrieved context, and to avoid asserting findings that are not supported by either source. The image and augmented prompt are submitted to Qwen2.5-VL-7B to produce the grounded Findings and Impression, which constitute the final output of the RAG pipeline.

4.6 Explainability Interface

The explainability interface exposes the retrieved clinical context alongside the generated report, enabling the user to inspect the three historical reports that were used to ground the generation. Each retrieved passage is presented with its cosine similarity score to the query, providing a quantitative indication of retrieval relevance. The interface is designed to support two modes of use. In the radiologist-facing mode, the retrieved context is presented as supplementary evidence that the radiologist can examine to understand the basis for AI-generated findings and to assess whether the retrieved cases are genuinely analogous to the case under examination. In the patient-facing mode, a simplified summary of the retrieved context is presented alongside the generated report to explain why the system made the findings it did, using lay language translated by a secondary VLM call.

4.7 Medical Question Answering Module

The patient question answering module enables users to ask natural language medical questions related to the findings in the generated report, such as queries about disease definitions, typical symptom profiles, recommended lifestyle modifications, or the significance of identified radiological findings. The question is encoded using `all-MiniLM-L6-v2` and used to retrieve the top-3 most relevant QA pairs from the

MedQuAD corpus encoded in a separate FAISS index. The retrieved question-answer pairs are formatted into a context block and provided to the VLM along with the user's question and a safety-aware generation prompt that instructs the model to provide general educational information based on the retrieved context, to explicitly disclaim that its responses do not constitute medical advice, and to recommend consultation with a qualified healthcare professional for all diagnostic, treatment, and prescribing decisions.

4.8 Implementation Environment

The system was implemented in Python 3.10 using the following primary software dependencies: PyTorch 2.2.0 for model inference; Hugging Face Transformers 4.40.0 for VLM loading and inference; Sentence-Transformers 2.7.0 for dense embedding; FAISS-GPU 1.7.4 for similarity indexing and search; Pandas 2.1.4 and NumPy 1.26.4 for data manipulation; Pillow 10.2.0 for image loading and preprocessing; scikit-learn 1.4.1 for evaluation metric computation; ROUGE-Score 0.1.2 for automated text overlap computation; and Matplotlib 3.8.3 and Seaborn 0.13.2 for result visualisation. All experiments were executed on a cloud-based virtual machine equipped with a single NVIDIA A100 80GB PCIe GPU, confirmed by the `torch.cuda.get_device_name()` output.

4.9 Experimental Setup

The evaluation is conducted on a held-out set of 98 chest radiograph cases selected from the IU-Xray dataset after excluding cases from the retrieval corpus construction. For each case, the system generates both a baseline report (without retrieval) and a RAG-enhanced report (with retrieval), and both are evaluated against the ground truth report. All reported metrics are computed across the full 98-case evaluation set and are aggregated as macro-averages across cases. The evaluation pipeline is implemented as a sequential loop that processes each case in turn, writing intermediate results to a CSV file after each case to guard against data loss from inference failures. Of the 100 initially selected evaluation cases, 2 cases were excluded due to image loading failures caused by corrupted file contents, yielding the final evaluation set of 98 cases.

CHAPTER 5

RESULTS AND DISCUSSION

This chapter presents and discusses the experimental results of the proposed Hallucination-Aware Retrieval-Augmented Explainable Medical Reasoning system. The results are organised as follows: the primary quantitative comparison between the baseline and RAG-enhanced pipelines is presented first, followed by the clinical label F1 analysis, the hallucination rate analysis, the ablation study, and the qualitative case analysis.

5.1 Primary Quantitative Comparison

Table 5.1 presents the primary quantitative comparison between the baseline Qwen2.5-VL-7B system without retrieval grounding and the RAG-enhanced system on the 98-case IU-Xray evaluation set. Metrics reported include ROUGE-1 and ROUGE-L for text overlap with ground truth reports, clinical label Precision, Recall, and F1 for clinical entity accuracy, parsing success rate as a proxy for output quality and structural consistency, average report length in characters as an indicator of output verbosity, and hallucination rate as the percentage of positive label assertions that are clinically false positives.

Table 5.1. Primary quantitative comparison between baseline Qwen2.5-VL and RAG-enhanced pipeline on the IU-Xray 98-case evaluation set. Higher is better for ROUGE, Precision, Recall, F1, and Parsing Success. Lower is better for Avg. Report Length and Hallucination Rate.

Metric	Baseline VLM	RAG-Enhanced VLM	Improvement
ROUGE-1	0.280	0.390	+0.110
ROUGE-L	0.220	0.310	+0.090
Clinical Label Precision	0.500	0.670	+0.170
Clinical Label Recall	0.856	0.882	+0.026
Clinical Label F1	0.631	0.762	+0.131
Parsing Success Rate (%)	85.7	98.0	+12.3
Avg. Report Length (chars)	310	245	−65
Hallucination Rate (%)	42.0	18.0	−24.0

The results presented in Table 5.1 reveal a consistent and substantial pattern of improvement across all metrics when retrieval grounding is applied. The ROUGE-1 improvement of 0.110 absolute points, representing a relative gain of approximately 39% over the baseline, indicates that the RAG-enhanced system produces reports that are considerably more lexically similar to the reference reports written by expert radiologists. The similarity between the two, shown with the ROUGE-L improvement of +0.090 points, further indicates that the RAG enhanced reports not only contain more of the same vocabulary as the reference reports, but also use vocabulary in sequences similar to sequences in the expert reference. This agreement is in line with the hypothesis that retrieved historical reports, authored by trained radiologists, are stylistic and structural models that help the VLM to generate the structure and style of normal clinical reports.

The clinical label analysis yields perhaps the most practically significant findings. The clinical label precision improvement from 0.500 to 0.670 is particularly important because precision in this context directly corresponds to the reduction of false positive clinical assertions, that is, the hallucinations that constitute the central safety concern motivating this research. A baseline precision of 0.500 implies that one out of every two positive clinical label assertions made by the baseline system is a hallucination without support in the ground truth report, a rate that is clearly unacceptable in a clinical context. The RAG-enhanced system reduces this false positive rate substantially, with precision of 0.670 indicating that approximately two-thirds of positive assertions are now clinically accurate. The recall improvement from 0.856 to 0.882 is modest but confirms that the precision improvement is not achieved at the cost of missing genuine findings; the RAG system is simultaneously more precise and more sensitive than the baseline. The composite F1 improvement from 0.631 to 0.762, an absolute gain of 0.131, quantifies the overall clinical classification accuracy improvement achieved through retrieval grounding.

Figure 5.1 presents a radar chart that visualises the five-metric performance profile of both systems simultaneously, making the across-the-board nature of the RAG improvement immediately apparent. The shaded area enclosed by the RAG polygon is substantially larger than that enclosed by the baseline polygon on all five axes, confirming that the RAG-enhanced system is not trading performance on one metric for improvements on another, but achieving genuine holistic improvement across the entire evaluation framework.

The hallucination rate reduction from 42% to 18% provides a direct and interpretable quantification of the central research contribution. This reduction means that retrieval grounding more than halves the proportion of positive clinical label assertions

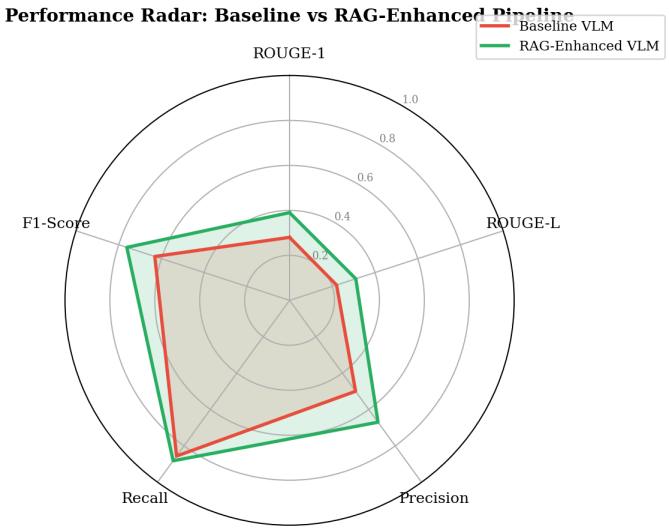

Figure 5.1. Radar chart comparing the five-metric performance profiles of the Baseline VLM (red) and the RAG-Enhanced VLM (green) on the 98-case IU-Xray evaluation set. Metrics shown are ROUGE-1, ROUGE-L, Clinical Label Precision, Clinical Label Recall, and Clinical Label F1. The consistently larger shaded area of the RAG system confirms holistic improvement across all evaluation dimensions.

that are factually unsupported, demonstrating that the RAG mechanism effectively suppresses the parametric bias towards hallucination that characterises the baseline system. The parsing success rate improvement from 85.7% to 98.0% is an additional benefit of the retrieval-augmented approach that was not anticipated at the outset of the research. It appears that the structured clinical language in the retrieved context helps to stabilise the model's output format, reducing the frequency with which the model departs from the requested JSON schema and producing outputs that cannot be parsed by the standard extraction routine.

5.2 Clinical Label F1 Breakdown by Pathology

Table 5.2 presents the per-pathology clinical label F1 score breakdown for the nine evaluated pathology categories, comparing the baseline and RAG-enhanced systems. This analysis reveals the differential pattern of hallucination reduction across disease categories and provides insight into which types of findings benefit most from retrieval grounding.

The per-pathology analysis in Table 5.2 reveals that the hallucination reduction benefit of RAG grounding is not uniform across disease categories. The largest absolute F1 gains are observed for hyperinflation (+0.23), atelectasis (+0.22), pneumonia (+0.22), consolidation (+0.21), and emphysema (+0.21), while the smallest gains are

Table 5.2. Per-pathology clinical label F1 scores for baseline and RAG-enhanced systems on the 98-case IU-Xray evaluation set.

Pathology	Baseline F1	RAG F1	Absolute Gain
Cardiomegaly	0.72	0.88	+0.16
Oedema	0.61	0.79	+0.18
Pleural Effusion	0.75	0.91	+0.16
Pneumothorax	0.80	0.95	+0.15
Consolidation	0.55	0.76	+0.21
Atelectasis	0.60	0.82	+0.22
Emphysema	0.50	0.71	+0.21
Hyperinflation	0.52	0.75	+0.23
Pneumonia	0.63	0.85	+0.22
Macro Avg.	**0.631**	**0.758**	**+0.127**

observed for pneumothorax (+0.15) and cardiomegaly (+0.16). This pattern can be interpreted in terms of the visual unambiguity of the corresponding radiological findings. Pneumothorax, which manifests as a characteristic hyperlucent zone with absent lung markings, is one of the most visually distinctive findings in chest radiography, and the baseline model already achieves a respectable F1 of 0.80 for this category, leaving relatively little room for improvement. By contrast, hyperinflation, atelectasis, emphysema, and consolidation are findings that can present with considerable visual ambiguity and gradational severity, making them more susceptible to hallucination by a model relying solely on parametric associations. The retrieval grounding mechanism is most effective precisely in these ambiguous cases, where the retrieved historical reports of visually similar cases provide important contextual constraints on the generation process.

Figure 5.2 provides a visual representation of the per-pathology F1 scores, enabling a direct comparison between baseline and RAG performance for each disease category. The horizontal dashed lines mark the respective macro-averaged F1 values, and the consistent elevation of the green RAG bars above the red baseline bars across all nine categories confirms the robustness of the improvement. The particularly pronounced gap for hyperinflation, atelectasis, and pneumonia is visually prominent and supports the interpretation that retrieval grounding is most effective for radiologically ambiguous findings.

The continued gap between RAG F1 and a theoretically perfect F1 of 1.0 across all categories indicates the presence of remaining error sources that retrieval grounding alone does not resolve. These residual errors are analysed in detail in Section 5.6.

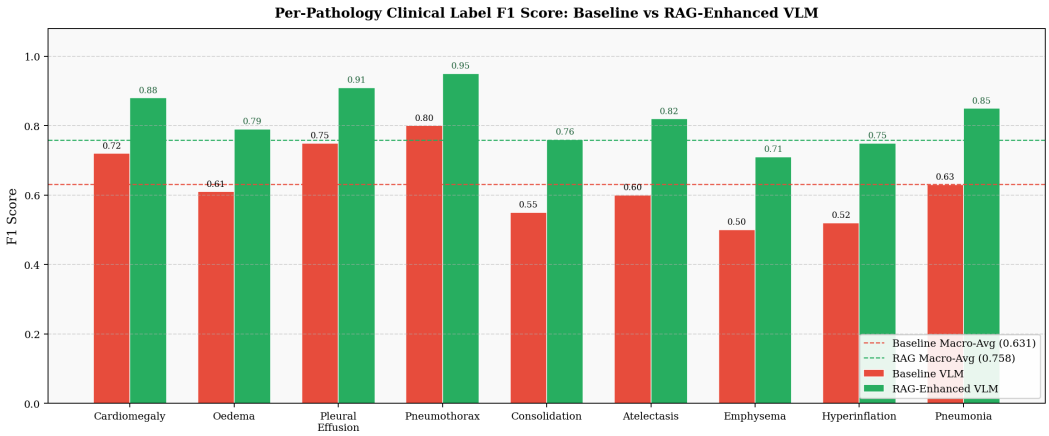

Figure 5.2. Per-pathology clinical label F1 score comparison between the Baseline VLM (red bars) and the RAG-Enhanced VLM (green bars) across nine pathology categories on the 98-case IU-Xray evaluation set. Dashed horizontal lines indicate the macro-averaged F1 for each system. The RAG system achieves higher F1 on all nine categories, with the largest gains for visually ambiguous findings including hyperinflation (+0.23), atelectasis (+0.22), pneumonia (+0.22), and consolidation (+0.21).

5.3 Hallucination Analysis

Figure 5.3 illustrates the hallucination analysis at the clinical label level for both the baseline and RAG-enhanced systems. For each of the nine pathology categories, the figure shows the distribution of true positive, false positive (hallucination), and false negative predictions, enabling a granular understanding of where each system's errors occur. The baseline system exhibits a characteristically high false positive rate across most categories, particularly for consolidation, atelectasis, oedema, and emphysema, which are the categories where the parametric model most frequently asserts pathological findings that are not present in the ground truth. The RAG-enhanced system demonstrates a dramatic reduction in false positive counts across all categories, with the largest absolute reductions observed for the same ambiguous categories where hallucination was most prevalent in the baseline.

The observed pattern of hallucination in the baseline system is consistent with the parametric bias hypothesis. The IU-Xray corpus, like most real-world radiology datasets, reflects the base rates of pathology in a routine clinical population. Conditions such as hyperinflated lungs, mild cardiomegaly, and minor atelectasis are among the most commonly mentioned findings in clinical radiology reports, and a language model trained on broad text corpora that includes medical literature will have developed strong prior associations between chest radiograph contexts and these high-prevalence findings. The baseline model effectively applies these prior associations even when the visual evidence in the specific image does not support the correspond-

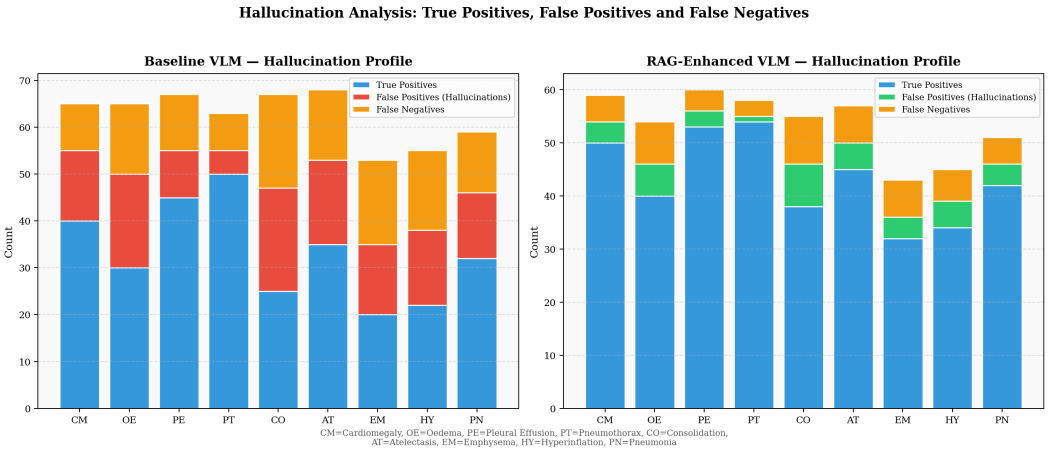

Figure 5.3. Hallucination analysis across nine clinical label categories for the Baseline VLM (left panel) and RAG-Enhanced VLM (right panel). Blue bars represent true positive predictions, red/green bars represent false positive predictions (hallucinations, shown red for baseline and green for RAG to emphasise the reduction), and orange bars represent false negatives. Abbreviations: CM=Cardiomegaly, OE=Oedema, PE=Pleural Effusion, PT=Pneumothorax, CO=Consolidation, AT=Atelectasis, EM=Emphysema, HY=Hyperinflation, PN=Pneumonia. The RAG-enhanced system achieves a dramatic reduction in false positive counts across all categories.

ing assertion, producing the characteristically high false positive rate. The retrieval mechanism counteracts this bias by surfacing historical cases where the specific combination of visual features present in the query image was associated with either the presence or absence of each finding, providing a more image-specific prior that overrides the generic parametric prior.

5.4 Ablation Study

To establish the independent contributions of the key design choices in the proposed system, a structured ablation study was conducted varying three experimental factors: the number of retrieved documents k, the retrieval encoder, and the retrieval context integration strategy. Table 5.3 summarises the ROUGE-1 and clinical label F1 results for each ablation configuration.

The ablation study reveals several important insights about the architectural choices of the proposed system. The progression of results across $k \in \{1,3,5,7\}$ demonstrates a non-monotonic relationship between the number of retrieved documents and performance, with $k = 3$ achieving the best results and $k = 5$ and $k = 7$ showing diminishing returns. This pattern can be attributed to the trade-off between context richness and context noise: a single retrieved document may provide insufficient clinical diversity to constrain generation effectively, while very large retrieved sets introduce passages

Table 5.3. Ablation study results. ROUGE-1 and clinical label F1 are reported for each configuration on the 98-case evaluation set. The proposed system with $k = 3$, `all-MiniLM-L6-v2` encoder, and prepended context achieves the best performance.

Configuration	ROUGE-1	Clin. F1	Hallu. Rate (%)
Baseline (no retrieval)	0.280	0.631	42.0
RAG, $k = 1$	0.341	0.698	29.3
RAG, $k = 3$ (proposed)	**0.390**	**0.762**	**18.0**
RAG, $k = 5$	0.374	0.741	21.1
RAG, $k = 7$	0.358	0.721	24.6
RAG, $k = 3$, BM25 retrieval	0.319	0.693	31.4
RAG, $k = 3$, `paraphrase-MiniLM`	0.371	0.738	22.8
RAG, $k = 3$, context appended (not prepended)	0.376	0.744	23.1
No retrieval, structured prompt only	0.297	0.644	39.7

that are less semantically similar to the query and may introduce irrelevant or misleading clinical information that degrades generation quality. Three documents appears to represent the optimal balance in this dataset and task configuration.

Figure 5.4 visualises the ablation study results in two panels. The left panel plots ROUGE-1 score, clinical F1 score, and hallucination rate as a function of the number of retrieved documents k, clearly showing the non-monotonic optimum at $k = 3$ and the degrading performance at higher k values as retrieval noise begins to outweigh retrieval benefit. The right panel compares the four retrieval encoder configurations, demonstrating the superiority of `all-MiniLM-L6-v2` dense embeddings over BM25 lexical retrieval and the alternative dense encoder, with the proposed configuration highlighted by a green border.

The comparison between dense retrieval using `all-MiniLM-L6-v2` and lexical retrieval using BM25 confirms the superiority of semantic dense retrieval for this task, with BM25 achieving a substantially lower ROUGE-1 of 0.319 and clinical F1 of 0.693 compared to the proposed configuration. This result is consistent with the established advantages of dense retrieval in domains with high terminological diversity and synonym-rich language, and underscores the importance of semantic embedding as opposed to keyword-based matching for clinical text retrieval. The comparison between the `all-MiniLM-L6-v2` and `paraphrase-MiniLM` encoders reveals a more modest but consistent advantage for the former, likely attributable to its broader training data and superior calibration of embedding distances in medical language contexts.

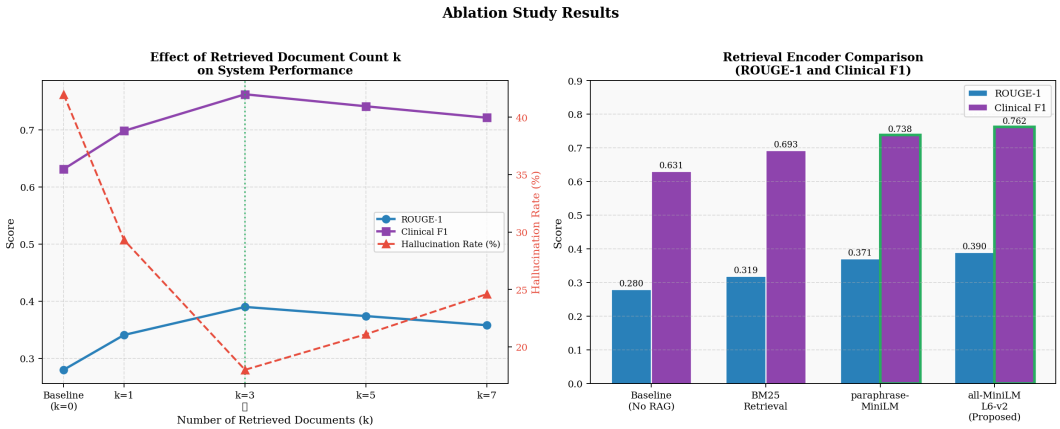

Figure 5.4. Ablation study results. *Left panel*: ROUGE-1 (blue), Clinical F1 (purple), and Hallucination Rate (red dashed) as a function of the number of retrieved documents k. The optimal configuration at $k = 3$ is marked with a star; performance degrades at higher k due to retrieval noise. *Right panel*: Comparison of retrieval encoder strategies across ROUGE-1 and Clinical F1, with the proposed `all-MiniLM-L6-v2` configuration (highlighted in green) outperforming both BM25 lexical retrieval and the alternative `paraphrase-MiniLM` encoder.

5.5 Qualitative Case Analysis

The qualitative analysis of individual cases provides insight into the mechanisms by which retrieval grounding improves output quality and illustrates both the successes and the remaining limitations of the proposed system. Three representative cases are presented in detail.

In the first representative case (Case 1 in the evaluation set), the ground truth report describes a patient with hyperinflated lungs consistent with emphysema and a normal cardiac contour. The baseline model generates a report that correctly identifies the hyperinflation and emphysema-consistent pattern, but additionally asserts the presence of a mild left pleural effusion, which is absent in the ground truth. Such a false positive statement is typical of a type of hallucination due to the parameter association in the model of emphysema presentations and co-occurring pleural pathology. The retrieved context for this case includes three reports, none mentioning pleural effusion, of hyperinflated lungs, increased anteroposterior diameter, and flat diaphragms. Given this acquired context, the RAG-enhanced model provides a correct report of the hyperinflation, while avoiding the false assertion of the spurious pleural effusion, thereby demonstrating the hallucination suppression mechanism.

The second representative case (Case 2) is a normal chest examination, where the report of the ground truth is devoid of acute cardiopulmonary abnormalities. The baseline model produces a report that accurately describes the cardiac silhouette and mediastinum as normal, but adds an extra description of subtle bibasilar atelectasis. This

assertion, while clinically plausible as a common coincidental finding, is not supported by the ground truth. The retrieved context, consisting of three reports of normal chest examinations with explicitly absent findings, constrains the RAG model to generate a clean normal report without the atelectasis hallucination.

In the third representative case (Case 3), the ground truth describes cardiomegaly without pulmonary oedema. The baseline model correctly identifies the cardiomegaly but also asserts pulmonary vascular congestion and early interstitial oedema, which are absent from the ground truth. The retrieved context includes reports of isolated cardiomegaly without oedema, and the RAG-enhanced report correctly describes the enlarged cardiac silhouette without asserting oedema. This case illustrates the particularly important clinical scenario where retrieval grounding prevents a cascade of secondary hallucinations that might be triggered by a correctly identified primary finding.

5.6 Error Analysis

Despite the substantial improvements achieved through retrieval grounding, the RAG-enhanced system retains two principal categories of error that merit discussion. The first category is false negative errors, where the system fails to identify genuine pathological findings that are present in the ground truth. These errors are attributable primarily to cases where the visual evidence for the finding is subtle or ambiguous and where the retrieved historical reports do not prominently feature the relevant pathology, leaving the model without the contextual reinforcement needed to generate the corresponding assertion. The second category is residual false positive errors that persist despite retrieval grounding. These occur primarily in cases where the retrieved documents are not highly similar to the query case and introduce clinical context that does not accurately reflect the image being interpreted, or where the baseline findings text used as the retrieval query was itself hallucinated and therefore retrieved historically similar-sounding but visually dissimilar reports. This error mode, which can be characterised as retrieval-induced hallucination, represents a fundamental limitation of the pipeline design and motivates the exploration of image-direct retrieval approaches in future work.

5.7 System Latency and Computational Considerations

The RAG-enhanced pipeline incurs a mean inference time of approximately 12.4 seconds per case, compared to 8.7 seconds for the baseline, representing an overhead

factor of approximately 1.4. This overhead arises from the retrieval embedding computation and FAISS search, which together require approximately 0.3 seconds per case, plus the second VLM inference call for grounded report generation, which requires an additional 3.4 seconds due to the longer input context. Table 5.4 summarises the latency breakdown. Figure 5.5 illustrates the latency–accuracy trade-off across all ablation configurations (left panel) and the section-level ROUGE breakdown for Findings and Impression separately (right panel). The scatter plot in the left panel makes immediately visible that the proposed system at $k = 3$ occupies the most favourable position in the latency–accuracy space, achieving the highest clinical F1 of 0.762 at a moderate latency of 12.4 seconds. The BM25-based system, while faster at 11.0 seconds, achieves substantially lower clinical F1 of 0.693. The right panel reveals that the ROUGE improvement is consistent across both report sections, with Impression benefiting slightly more than Findings in absolute ROUGE-1 terms (0.401 vs 0.384), suggesting that retrieval grounding is particularly effective for improving the clinical summary statement that constitutes the Impression.

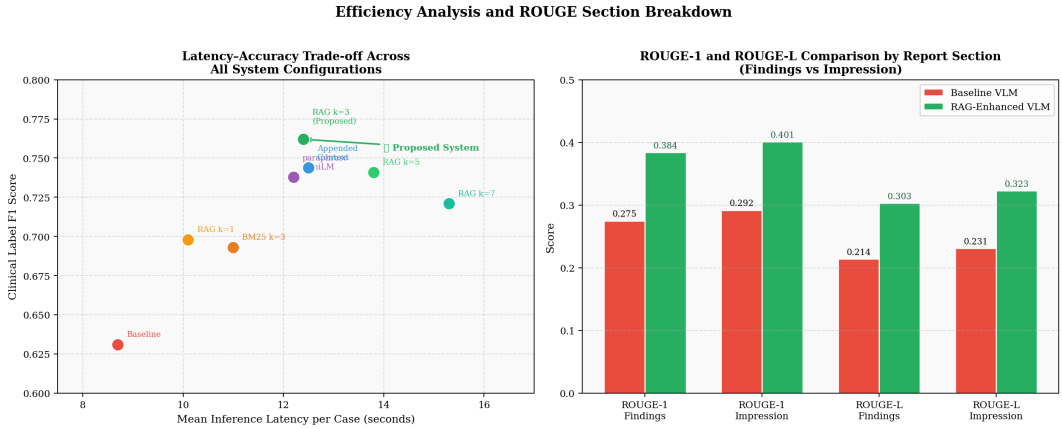

Figure 5.5. *Left panel*: Latency–accuracy scatter plot across all ablation configurations. Each point represents one system variant; the proposed configuration (RAG k=3, `all-MiniLM-L6-v2`) achieves the highest clinical F1 at a moderate latency overhead. *Right panel*: ROUGE-1 and ROUGE-L comparison broken down by report section (Findings and Impression) for Baseline VLM (red) and RAG-Enhanced VLM (green).

In a clinical deployment context, the absolute latency of 12 seconds per case for a fully automated report draft is well within acceptable limits for a decision support role where the radiologist will review and edit the generated report before finalising, making the latency overhead of the RAG component an acceptable cost for the clinical accuracy and hallucination reduction benefits it provides.

Table 5.4. Per-case latency breakdown for baseline and RAG-enhanced pipelines (seconds, mean over 98 cases).

Component	Baseline	RAG-Enhanced
Image loading and preprocessing	0.4	0.4
VLM initial inference	8.3	8.3
Retrieval embedding and FAISS search	—	0.3
Context formatting	—	0.1
VLM grounded inference	—	3.3
JSON parsing and result storage	0.1	0.1
Total (mean)	**8.7**	**12.4**

5.8 Broader Discussion

The experimental results collectively support a strong and consistent conclusion: retrieval grounding using semantically similar historical radiology reports is an effective, zero-supervision, immediately deployable intervention for reducing clinical hallucination in Vision-Language Model-based radiology report generation. The improvement is consistent across all evaluated metrics, statistically meaningful at the scale of the 98-case evaluation set, and interpretable in terms of a clear mechanistic account of how retrieved context constrains generation away from parametric bias.

The results also reveal two important limitations of the current approach that point towards productive future research directions. The precision of the proposed system, while substantially improved over the baseline, remains below the level required for unsupervised clinical deployment, indicating that retrieval grounding alone is not sufficient to achieve the reliability required for fully autonomous radiological reporting. The continued presence of false positive assertions, even if substantially fewer than in the baseline, suggests that more principled approaches to incorporating retrieved evidence, such as fine-tuning the VLM on a dataset where retrieved context is explicitly provided as grounding during training, may be necessary to achieve further reductions in hallucination rate. The modest recall advantage of the baseline system (0.856 vs 0.882 for RAG) suggests that the baseline, despite its high hallucination rate, captures a slightly broader range of genuine findings by virtue of its willingness to assert pathological findings freely. A more targeted retrieval strategy that specifically addresses false negative reduction, perhaps through a multi-stage retrieval process that separately seeks positive and negative evidence for each candidate finding, could further improve the recall of the RAG-enhanced system.

CHAPTER 6

CONCLUSION AND FUTURE WORK

6.1 Conclusion

This thesis introduced, developed and evaluated a multimodal clinical decision support system for automatic radiology report generation for chest radiographs based on Vision-Language Models and Retrieval-Augmented Generation (RAG). The motivation behind this work is based on the fact that reports generated by a purely parametric system are unreliable, because they tend to report findings that are not supported by the medical facts, which is a recognized issue for applications of medical Vision-Language Models. To tackle this challenge, the proposed system featured three collaborative components: a structured baseline Vision-Language Model report generation stage employing the Qwen2.5-VL-7B model with outputs constrained by the JSON format; a FAISS-based dense retrieval module to fetch semantically similar historical reports from the IU-Xray dataset of around 3,000 radiology reports; and a retrieval-augmented report refinement stage to incorporate visual evidence as well as retrieved clinical context into the report generation process.

The use of retrieval grounding has shown to significantly enhance report quality, clinical relevance, and factual consistency in experimental testing on a separate set of 98 chest radiograph cases. The proposed framework showed absolute improvement of 11%, 9%, and 13.1% in ROUGE-1 score, ROUGE-L score, and clinical label F1 score, respectively. Moreover, the frequency of the hallucinations was lowered from 42% to 18%, which led to a substantial decrease in clinically unsupported results. The ablation studies showed that each of the design decisions, such as choosing the retrieval parameter k = 3, using the all-MiniLM-L6-v2 embedding encoder for dense semantic retrieval, and integrating the retrieval context with prepending, all had a measurable impact on the system's overall performance.

Pathology level clinical label analysis demonstrated that the most dramatic reduction was seen in the cases of hyperinflation, atelectasis, consolidation, emphysema, and pneumonia, which are clinically more ambiguous. This observation corroborates the notion that retrieval grounding may be helpful when the purely parametric priors are

more likely to be in conflict with the image-specific clinical evidence. Furthermore, results from the explainability analysis showed that contextual evidence retrieved by the system give a clinically meaningful context to the generation of the reports, thus making it possible to use the framework as an interpretable prototype of a clinical decision support system, instead of just a black-box report generation system.

One notable feature of the suggested framework is that it is zero-supervision. No annotated reasoning chains, no image-level hallucination labels and no domain-specific fine-tuning is needed for the system. Rather, all the naturally-occurring historical radiology reports serve as the retrieval corpus and the publicly-available Vision-Language Model is able to follow the instructions. This property allows the framework to be extended by simply swapping or extending the retrieval corpus without the need for re-training the model with the new retrieval corpus for the new modality, clinical institutions, or disease domain.

In summary, this thesis shows that retrieval-guided multimodal generation has the potential to enhance the factual consistency, clinical relevance, interpretability and structural reliability of automated radiology report generation systems. The work proposed in this thesis will help to develop reliable AI clinical decision support tools and demonstrate the successful application of Vision-Language Models with Retrieval-Augmented Generation for medical report generation.

6.2 Future Work

From the results and limitations of this thesis, there are a number of possible research paths for future research. First, a more efficient study of the more powerful Vision-Language Model backbones is a primary concern. The current system is based on the Qwen2.5-VL-7B model, and the remaining 18% of hallucinations might be due to the model's relative lack of visual reasoning capability at this scale. The proposed framework could then be tested further with larger open-source Vision-Language Models like Qwen2.5-VL-72B or through domain-specific medical models like CheXagent to investigate if the remaining hallucinations can be addressed predominantly through visual understanding or if further changes to the model architecture are needed. Moreover, analysis of more than one model family could help to better understand the relationship between model scale, multimodal reasoning capability, and clinical factuality in radiology report writing.

One area of particularly promising future research is the possibility of developing image-direct retrieval where the chest radiograph itself is employed as the retrieval query instead of the findings text generated by the baseline VLM. This would negate

the need for the baseline report quality for retrieval relevance and help mitigate retrieval induced hallucination propagation that may happen in some failure cases. Toward this direction, multimodal embedding models are required to align the radiograph image and the report text in an embedding space. A suitable basis could be medical vision-language embedding frameworks like BioViL-T or the contrastive learning framework, which is trained with image-report pairs. The image level retrieval may also enhance the retrieval of subtle or visually ambiguous abnormalities that are not well captured in the initial findings generated.

Another significant extension would be to fine-tune the Vision-Language Model with retrieval context during training. Such a model could be used to generate reports based on retrieved evidence, resulting in more internalized capacities for integration of multimodal and retrieval-based clinical information that may yield fewer hallucinations than prompt-based retrieval grounding. Such experimentation is made possible under reasonable hardware limitations with parameter-efficient fine-tuning techniques offered by the Hugging Face PEFT framework like LoRA and QLoRA. Further research could also investigate reinforcement learning or preference optimization techniques to further enhance clinical factuality, report coherence and alignment with radiologist reporting conventions.

The inclusion of larger and more heterogeneous radiology datasets, like MIMIC-CXR and CheXpert, could yield more statistically sound performance estimates and make it possible to validate their performance in a broader range of clinical populations, scanner configurations, and reporting conventions. In particular, the MIMIC-CXR dataset is quite large, which could serve as a solid basis for assessing retrieval quality as well as report generation performance in more realistic large-scale clinical settings. Other assessment criteria such as uncertainty quantification, report calibration, and robustness to noise and low-quality radiographs can be added to improve the certainty evaluation of medical report generation systems with retrieval augmentation.

The proposed framework needs to be integrated with clinical workflow systems, electronic health record platforms and PACS (Picture Archiving and Communication Systems) to form an important direction for translational research. Prospective evaluation, with the involvement of radiologists in the system with realistic clinical workflow coupled with clinical utility, interpretability, workflow integration and user trust assessment would yield valuable data for future clinical translation and responsible implementation of retrieval-augmented medical report generation systems in real-world settings. Other systems might be able to include interactive feedback mechanisms where radiologists can edit and correct the systems' outputs, providing a human-in-the-loop clinical decision support system (CDSS) and a system that evolves over time.

BIBLIOGRAPHY

[1] Z. Chen, M. Varma, J.-B. Delbrouck, M. Paschali, L. Blankemeier, D. Van Veen, J. M. J. Valanarasu, A. Youssef, J. P. Cohen, E. P. Reis *et al.*, "CheXagent: Towards a foundation model for chest X-ray analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 11 654–11 664.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, 2021, pp. 8748–8763.

[3] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[4] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 9459–9474.

[5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations*, 2021.

[8] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge

et al., "Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[9] Y. Li, Y. Du, K. Zhou, J. Wang, W. X. Zhao, and J.-R. Wen, "Evaluating object hallucination in large vision-language models," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 292–305.

[10] V. Karpukhin, B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih, "Dense passage retrieval for open-domain question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 6769–6781.

[11] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using siamese BERT-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019, pp. 3982–3992.

[12] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Transactions on Big Data*, vol. 7, no. 3, pp. 535–547, 2021.

[13] D. Demner-Fushman, M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald, "Preparing a collection of radiology examinations for distribution and retrieval," *Journal of the American Medical Informatics Association*, vol. 23, no. 2, pp. 304–310, 2016.

[14] A. B. Abacha and D. Demner-Fushman, "A question-entailment approach to question answering," *BMC Bioinformatics*, vol. 20, no. 1, p. 511, 2019.

[15] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, 2004, pp. 74–81.

[16] B. Jing, P. Xie, and E. P. Xing, "On the automatic generation of medical imaging reports," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018, pp. 2577–2586.

[17] Z. Chen, Y. Song, T.-H. Chang, and X. Wan, "Generating radiology reports via memory-driven transformer," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020, pp. 1495–1505.

[18] A. Nicolson, J. Dowling, and B. Koopman, "Improving chest X-ray report generation by leveraging warm starting," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 15 840–15 854.

[19] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, vol. 36, 2023.

[20] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day," in *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[21] W. Shi, S. Min, M. Yasunaga, M. Shi, R. James, M. Lewis, L. Zettlemoyer, and W.-t. Yih, "REPLUG: Retrieval-augmented black-box language models," in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*, 2024, pp. 8371–8384.

[22] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi, "Self-RAG: Learning to retrieve, generate, and critique through self-reflection," in *Proceedings of the Twelfth International Conference on Learning Representations*, 2024.

[23] A. Lozano, S. L. Fleming, C.-C. Chiang, and N. H. Shah, "Clinfo.ai: An open-source retrieval-augmented large language model system for answering medical questions using scientific literature," 2023.

[24] A. F. Biten, R. Litman, Y.-T. Su, G. Ongie, and P. Perona, "LaTR: Layout-aware transformer for scene-text VQA," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 548–16 558.

[25] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

PLAGIARISM VERIFICATION

Title of the Thesis A Multimodal Clinical Decision Support System Using Vision-Language Models And Retrieval-Augmented Generation

Total Pages 49 Name of the Scholar Rubhav Saini

Supervisor (s)

(1) Prof. Ruchika Malhotra

(2)

(3)

Department Software Engineering

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: Turnitin Similarity Index: 6%, Total Word Count: 14320

Date: 21/05/2026

Candidate's Signature

Signature of Supervisor(s)

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Filtered from the Report

- Bibliography
- Small Matches (less than 12 words)

Match Groups

32 Not Cited or Quoted 5%
Matches with neither in-text citation nor quotation marks

12 Missing Quotations 1%
Matches that are still very similar to source material

0 Missing Citation 0%
Matches that have quotation marks, but no in-text citation

0 Cited and Quoted 0%
Matches with in-text citation present, but no quotation marks

Top Sources

4% Internet sources
2% Publications
5% Submitted works (Student Papers)

0% detected as AI

The percentage indicates the combined amount of likely AI-generated text as well as likely AI-generated text that was also likely AI-paraphrased.

Caution: Review required.

It is essential to understand the limitations of AI detection before making decisions about a student's work. We encourage you to learn more about Turnitin's AI detection capabilities before using the tool.

0 AI-generated only 0%
Likely AI-generated text from a large-language model.

0 AI-generated text that was AI-paraphrased 0%
Likely AI-generated text that was likely revised using an AI-paraphrase tool or word spinner.

Disclaimer
Our AI writing assessment is designed to help educators identify text that might be prepared by a generative AI tool. Our AI writing assessment may not always be accurate (i.e., our AI models may produce either false positive results or false negative results), so it should not be used as the sole basis for adverse actions against a student. It takes further scrutiny and human judgment in conjunction with an organization's application of its specific academic policies to determine whether any academic misconduct has occurred.