Air University

Department of Creative Technologies

Faculty of Computing and Artificial Intelligence

Final Term Examination, Fall 2025

Evidence-Grounded Multi-Modal RAG for Automated Medical Report Generation: A Research Protocol

Course Code:	AI-831
Course Title:	Special Topics in NLP
Program:	PhD Artificial Intelligence
Instructor:	Dr. Imran Ihsan

Submitted by:

Muhammad Kashif

Registration No: 2504465

January 17, 2026

Word Count Summary

Section	**Word Count**
Abstract	318
Introduction	485
Literature Review	756
Research Questions	248
Methodology	628
Data Analysis and Presentation	287
Practicalities and Contingency Plan	268
Resources Required	245
Timeline	178
Total (excluding Abstract & References)	**3,095**

Table 1: Word count by section

Abstract

Background: Large Language Models have generated significant interest in medical education, particularly for radiology report generation. Vision-language models such as GPT-4V demonstrate impressive capabilities but suffer from factual hallucinations and lack domain-specific knowledge essential for clinical safety. Retrieval-Augmented Generation offers a promising solution by grounding outputs in external evidence. However, existing multi-modal RAG systems employ fixed retrieval parameters regardless of case complexity and lack evaluation of human factors such as radiologist trust. This protocol proposes Evidence-Grounded Multi-Modal RAG (EG-MM-RAG), a framework designed to optimize multi-modal retrieval and enable evidence-transparent generation for chest radiograph reporting.

Methods: This study will employ the CheXpert Plus dataset containing chest radiographs with paired radiology reports. The proposed system will combine visual similarity using BiomedCLIP and textual similarity using Bio-ClinicalBERT through optimized fusion weights. We will systematically explore retrieval parameters to identify suitable configurations for different case complexities. The system will enforce evidence-grounded generation with structured citations using GPT-4V, enabling radiologist verification. Evaluation will encompass clinical accuracy, linguistic quality, retrieval performance, and human factors including radiologist trust ratings.

Discussion: Expected outcomes include improved clinical accuracy compared to zero-shot baselines, reduced hallucination rates through evidence grounding, and enhanced radiologist trust via transparent citations. This research will contribute best practices for retrieval parameter selection in medical imaging RAG systems and provide insights into multi-modal fusion for clinical applications. The findings will inform development of AI-assisted medical education tools while emphasizing the balance between technical performance and human factors in medical domains.

Keywords: Large Language Models, Retrieval-Augmented Generation, Medical Education, Multi-modal Learning, Clinical Decision Support, Medical Report Generation, Radiology AI, Vision-Language Models

1. Introduction

1.1 Background and Motivation

Medical education faces growing demands as students must master extensive clinical knowledge while developing diagnostic reasoning skills. Large Language Models have emerged as potentially transformative tools, demonstrating notable capabilities across medical domains [7]. Recent studies show ChatGPT can pass written steps of the United States Medical Licensing Examination [5], reflecting knowledge typically gained throughout medical training.

However, these capabilities come with critical limitations. Vision-language models like GPT-4V frequently generate factual hallucinations and lack domain-specific medical knowledge required for patient safety [8]. In medical education contexts, such errors could reinforce incorrect clinical reasoning in trainees.

Chest radiography—the most commonly performed radiological examination worldwide—presents an ideal domain for investigating AI-assisted report generation. Automated systems could support medical education and reduce radiologist workload, but only if they demonstrate both clinical accuracy and trustworthiness.

1.2 The Promise of Retrieval-Augmented Generation

Retrieval-Augmented Generation addresses limitations of purely parametric models by combining generative capabilities with retrieval from external knowledge sources [6]. In medical imaging contexts, RAG enables grounding of generated reports in similar historical cases, providing both improved accuracy and interpretability.

Recent systems such as MMed-RAG have shown improvements in factual accuracy over baselines [10]. However, current approaches exhibit limitations: fixed retrieval parameters regardless of case complexity, limited evaluation scope, and insufficient validation of human factors such as radiologist trust.

1.3 Research Objectives

This protocol proposes Evidence-Grounded Multi-Modal RAG (EG-MM-RAG) to address these gaps. The objectives are:

1. Design an evidence-grounded multi-modal RAG system with optimized retrieval parameters
2. Conduct multi-dimensional evaluation addressing clinical accuracy, linguistic quality, and human factors
3. Establish guidelines for retrieval parameter selection in medical imaging applications
4. Validate automated metrics against radiologist assessments

2. Literature Review

2.1 Large Language Models in Medical Education

Lucas et al. [7] conducted a systematic review of 40 studies examining LLM applications in medical education. Key benefits include personalized learning support, instant access to medical knowledge, and interactive clinical simulations. However, challenges persist: hallucination, impact on critical thinking development, and academic integrity concerns.

2.2 Vision-Language Models for Medical Imaging

GPT-4V represents a breakthrough in multi-modal AI, processing both images and text [8]. In medical contexts, it can identify pathologies in chest X-rays and suggest diagnoses. However, evaluations reveal limitations including hallucination and oversight of subtle pathologies.

Domain-specific models address some limitations. BiomedCLIP achieves higher retrieval precision than generic CLIP on medical images [11]. BioClinicalBERT improves performance on clinical NLP tasks [1]. These models highlight the value of domain-specific pre-training.

2.3 Medical Report Generation Approaches

Early approaches employed CNN-RNN architectures [4], producing generic outputs. Memory-driven transformers such as R2Gen [3] improved results but remain limited—they cannot cite evidence or explain reasoning.

2.4 Retrieval-Augmented Generation in Medical Domains

Lewis et al. [6] introduced RAG, combining retrieval with generation. RAG offers advantages: access to updated information, ability to cite sources, and reduced hallucination through grounding.

MMed-RAG [10] represents current state-of-the-art in multi-modal medical RAG. However, it uses fixed retrieval parameters, focuses on aggregate metrics, and lacks human evaluation of trust and interpretability.

2.5 Research Gaps

Three gaps motivate this research:

Gap 1: Existing systems employ fixed retrieval configurations without exploring how optimal parameters might vary with case complexity.

Gap 2: Current evaluations focus on technical metrics without assessing clinical accuracy, linguistic quality, and human factors together.

Gap 3: Despite evidence that many radiologists distrust AI recommendations [9], no multi-modal RAG study has evaluated trust and interpretability.

3. Research Questions

This research will investigate three primary questions:

3.1 RQ1: Clinical Effectiveness

Does evidence-grounded multi-modal RAG improve factual grounding and clinical accuracy compared to zero-shot vision-language models and existing baselines?

We expect EG-MM-RAG to show improvements in clinical accuracy metrics (CheXbert F1, RadGraph F1) and reduced hallucination rates compared to GPT-4V zero-shot and existing baselines.

3.2 RQ2: Retrieval Parameter Optimization

How do retrieval parameters (depth K and fusion weight α) influence factual consistency and report quality?

We will explore whether optimal retrieval depth varies with case complexity, whether multi-modal retrieval outperforms single-modality approaches, and whether different pathologies benefit from different parameter settings.

3.3 RQ3: Human-Centered Impact

How does explicit presentation of retrieved evidence affect radiologists' trust and interpretability perceptions?

We will investigate whether evidence presentation increases trust ratings, whether reports achieve acceptable interpretability, and how automated faithfulness metrics correlate with radiologist assessments.

4. Methodology

4.1 Dataset

We will utilize CheXpert Plus [2], containing chest radiographs with paired radiology reports. All data is de-identified and available through PhysioNet. We will employ patient-level splitting (80% training, 10% validation, 10% test) with a stratified test subset ensuring representation across pathology categories.

4.2 System Architecture

The EG-MM-RAG system comprises four components:

Image Preprocessing: Chest radiographs will undergo contrast enhancement (CLAHE), resizing, and normalization.

Multi-Modal Retrieval: Visual encoding uses BiomedCLIP; textual encoding uses BioClinicalBERT. Both embeddings are normalized and combined via weighted fusion:

$$e_q = \alpha \cdot \hat{v}_q + (1 - \alpha) \cdot \hat{t}_q \tag{1}$$

where α controls the visual-textual balance. Retrieval uses vector database indexing for efficient search.

Evidence-Aware Generation: Report generation employs GPT-4V with prompting that requires inline citations linking findings to retrieved cases.

Faithfulness Verification: An LLM-as-judge approach classifies generated claims as Supported, Contradicted, or Unsupported based on retrieved evidence.

4.3 Baseline Methods

Five baselines will contextualize EG-MM-RAG performance: GPT-4V Zero-Shot, R2Gen [3], Text-Only RAG, Visual-Only RAG, and MMed-RAG [10].

4.4 Experimental Design

Validation Phase: Explore retrieval depth K and fusion weight α on validation cases to identify suitable configurations.

Test Phase: Apply selected configuration to stratified test cases alongside all baselines.

Human Evaluation: Two radiologists or senior residents will independently evaluate a case subset, rating faithfulness, trust, interpretability, and clinical utility.

5. Data Analysis and Presentation

5.1 Evaluation Framework

Evaluation will span four dimensions:

Clinical Accuracy: CheXbert F1 and RadGraph F1 with per-pathology breakdown.

Linguistic Quality: ROUGE-L and BERTScore for fluency assessment.

Retrieval Performance: Precision@K and similarity score analysis.

Human Factors: Trust ratings, interpretability scores, and inter-rater reliability.

5.2 Statistical Analysis

We will report mean differences with confidence intervals and use appropriate statistical tests for paired comparisons across methods.

5.3 Results Presentation

Results will include comparison tables, per-pathology visualizations, parameter sensitivity plots, and representative case studies illustrating successes and limitations.

6. Practicalities and Contingency Plan

6.1 Anticipated Challenges

API Costs: GPT-4V calls involve costs. *Mitigation:* Apply for researcher credits; cache results; use cheaper models for development. *Contingency:* Reduce parameter search scope if needed.

Evaluator Availability: Radiologists have limited time. *Mitigation:* Appropriate local compensation; efficient interface; flexible scheduling. *Contingency:* Use senior residents or reduce evaluation subset.

LLM-as-Judge Reliability: Using GPT-4 to verify GPT-4V may introduce bias. *Mitigation:* Validate against human assessments. *Contingency:* Expand human evaluation if correlation is poor.

Computational Resources: Grid search requires GPU compute. *Mitigation:* Use university cluster; apply for cloud credits. *Contingency:* Use coarser parameter grid.

6.2 Ethical Considerations

All data is pre-de-identified. Ethics approval will be obtained for human evaluation. We will clearly communicate that EG-MM-RAG is a research prototype requiring expert review before any clinical application.

7. Resources Required

7.1 Software and Data

All software is open-source (Python, PyTorch, Transformers, ChromaDB). Pre-trained models (BiomedCLIP, Bio-ClinicalBERT, CheXbert) are freely available via Hugging Face. CheXpert Plus is accessible through PhysioNet with appropriate data use agreement. Application development will be student-led academic effort.

7.2 Budget Summary (12 Months)

Note: Application development and system integration treated as in-kind academic contribution.

8. Timeline

This research will be conducted over 12 months:

Phase 1: Preparation (Months 1–2) — Data access, ethics approval, preprocessing pipeline, retrieval index setup.

Phase 2: Development (Months 3–5) — Generation pipeline, evaluation framework, parameter exploration, baseline implementation.

Phase 3: Evaluation (Months 6–9) — Test evaluation, statistical analysis, human evaluation, failure analysis.

Category	**Description**	**Cost (USD)**
Compute (GPU)	Intermittent GPU usage for preprocessing, embedding generation (∼22K X-rays), re-indexing, ablations, and inference (shared/on-demand, not 24/7)	1,800
API Usage (VLM)	GPT-4V API for iterative evaluation, prompt refinement, ablations, and final testing with capped tokens	900
Storage	Persistent storage for images, reports, embeddings, logs, and indices (∼500 GB, 12 months)	150
Vector DB / Hosting	VM or server for ChromaDB, retrieval services, and experiment tracking (12 months)	600
Human Evaluation	Radiology experts (2 evaluators × 10 hrs × $20/hr) across pilot and final rounds	400
Monitoring & Backups	Logging, backups, and contingency overhead	150
Total		**≈4,000**

Table 2: Estimated research costs over 12 months

Phase 4: Dissemination (Months 10–12) — Manuscript preparation, submission, presentation materials.

Phase	**M1–2**	**M3–5**	**M6–9**	**M10–12**
Preparation	■			
Development		■		
Evaluation			■	
Dissemination				■

Table 3: Research timeline overview

References

[1] Alsentzer, E., Murphy, J., Boag, W., et al. (2019). Publicly available clinical BERT embeddings. *Proceedings of the Clinical NLP Workshop*, 72–78.

[2] Chambon, P., Delbrouck, J.B., Sounack, T., et al. (2024). CheXpert Plus: Augmenting a large chest X-ray dataset with text radiology reports. *arXiv:2405.19538*.

[3] Chen, Z., Song, Y., Chang, T.H., & Wan, X. (2020). Generating radiology reports via memory-driven transformer. *Proceedings of EMNLP*, 1439–1449.

[4] Jing, B., Xie, P., & Xing, E. (2018). On the automatic generation of medical imaging reports. *Proceedings of ACL*, 2577–2586.

[5] Kung, T.H., Cheatham, M., Medenilla, A., et al. (2023). Performance of ChatGPT on USMLE. *PLOS Digital Health*, 2(2), e0000198.

[6] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of NeurIPS*, 9459–9474.

[7] Lucas, H.C., Upperman, J.S., & Robinson, J.R. (2024). A systematic review of large language models and their implications in medical education. *Medical Education*, 58(11), 1276–1285.

[8] OpenAI. (2023). GPT-4V(ision) system card. *OpenAI Technical Report*.

[9] Shen, Y., Heacock, L., Elias, J., et al. (2023). Evaluating physician trust in AI-driven medical systems. *npj Digital Medicine*, 6, 89.

[10] Xia, P., Zhu, K., Li, H., et al. (2024). MMed-RAG: Versatile multimodal RAG system for medical vision language models. *arXiv:2410.13085*.

[11] Zhang, S., Xu, Y., Usuyama, N., et al. (2023). BiomedCLIP: A multimodal biomedical foundation model. *arXiv:2303.00915*.

END OF RESEARCH PROTOCOL