EFFECTIVE RETRIEVAL APPROACHES FOR RADIOLOGY SUMMARY GENERATION WITH OPEN-SOURCE LARGE LANGUAGE MODELS

A Thesis Submitted to
the Graduate School of Engineering and Sciences of
İzmir Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

in Computer Engineering

by
Serhat CANER

July 2026
İZMİR

We approve the thesis of **Serhat CANER**

Examining Committee Members:

Prof. Dr. Yalın BAŞTANLAR
Department of Computer Engineering, Izmir Institute of Technology

Prof. Dr. Oğuz DİKENELLİ
Department of Computer Engineering, Ege University

Assoc. Prof. Dr. Selma TEKİR
Department of Computer Engineering, Izmir Institute of Technology

Prof. Dr. Aytuğ ONAN
Department of Computer Engineering, Izmir Institute of Technology

Prof. Dr. Senem KUMOVA METİN
Department of Software Engineering, Izmir University of Economics

8 July 2026

Prof. Dr. Yalın BAŞTANLAR
Department of Computer Engineering
Izmir Institute of Technology

Assoc. Prof. Dr. Emrah İNAN
Department of Computer Engineering
Izmir Institute of Technology

Prof. Dr. Onur DEMİRÖRS
Head of the Department of Computer Engineering

Prof. Dr. Mehtap EANES
Dean of the Graduate School of Engineering and Sciences

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my thesis advisor, Prof. Dr. Yalın BAŞTANLAR, for his invaluable guidance, patience, understanding, and continuous support throughout this thesis. His academic perspective, constructive feedback, and willingness to help whenever needed have been essential in shaping this work. I am truly grateful for his encouragement and for the insight he provided during every stage of my doctoral journey.

I would also like to sincerely thank my co-advisor, Assoc. Prof. Dr. Emrah İNAN, for his guidance, support, and valuable contributions to this thesis. His thoughtful suggestions, technical perspective, and generous help greatly contributed to the development and refinement of this study. I deeply appreciate his patience, encouragement, and constant willingness to support my work.

I am especially grateful to Prof. Dr. Oğuz DİKENELLİ for his significant contributions, insightful comments, and valuable guidance throughout this process. His feedback helped improve the direction and quality of the thesis in many important ways. I would also like to thank Assoc. Prof. Dr. Selma TEKİR for her kind support, helpful suggestions, and valuable assistance during my studies.

I would like to extend my sincere thanks to my thesis jury members, Prof. Dr. Aytuğ ONAN and Prof. Dr. Senem KUMOVA METİN, for their valuable comments, constructive feedback, and careful evaluation of this thesis.

I would like to express my heartfelt gratitude to my parents, Yalçın CANER and Seyhan CANER, for their unconditional love, trust, and support throughout my life. Their encouragement and sacrifices have always been a source of strength for me, and I am deeply grateful for everything they have done.

Most importantly, I would like to thank my beloved wife, Güneş TOK CANER. Her endless love, patience, understanding, and support carried me through the most difficult moments of this journey. I am grateful for her encouragement, for standing by me with kindness and strength, and for making this long process more meaningful. This thesis would have been much harder without her presence, support, and belief in me.

ABSTRACT

EFFECTIVE RETRIEVAL APPROACHES FOR RADIOLOGY SUMMARY GENERATION WITH OPEN-SOURCE LARGE LANGUAGE MODELS

Radiology impression generation is often studied using proprietary LLMs, multimodal architectures, or label-driven retrieval mechanisms that limit transparency, reproducibility, and comparability. This thesis investigates retrieval-guided impression generation using compact open-source language models. The proposed framework generates concise impressions from radiology findings without proprietary APIs, medical image encoders, or language model fine-tuning. Two retrieval strategies are evaluated. The first combines semantic similarity with lexical refinement to select clinically aligned findings-impression examples. The second extends this design with projection-guided multi-signal matching, combining target-oriented summary estimation, findings-to-findings matching, findings-to-impression matching, key concept carry-over, and structured clinical matching. Across OpenI and MIMIC-CXR experiments, zero-shot and random-shot prompting are shown to be insufficient, while retrieval-guided prompting provides consistent improvements. Few-shot prompting, example ordering, hard negative examples, and iterative prompt feedback further improve generation quality. On OpenI, the proposed methods achieve strong performance compared with graph-based, multimodal, and proprietary LLM-based systems. On MIMIC-CXR, the projection-guided strategy remains competitive, especially in clinical label agreement. An evaluation on six CT and MR domains from the MIMIC-III RadSum23 benchmark examines transfer beyond chest X-ray reporting. The results show that retrieval and prompting make open-source text-only models effective for radiology impression generation and provide a transparent, reproducible alternative.

ÖZET

AÇIK KAYNAKLI BÜYÜK DİL MODELLERİ İLE RADYOLOJİ ÖZETLERİ OLUŞTURMA İÇİN ETKİN GERİ GETİRME YAKLAŞIMLARI

Radyoloji izlenim üretimi çoğunlukla şeffaflığı, yeniden üretilebilirliği ve karşılaştırılabilirliği sınırlayan kapalı kaynak büyük dil modelleri, çok modlu mimariler veya etiket güdümlü geri getirme mekanizmaları kullanılarak incelenmektedir. Bu tezde, kompakt açık kaynak dil modelleri kullanılarak geri getirme yönlendirmeli izlenim üretimi incelenmektedir. Önerilen çerçeve, kapalı kaynak API'ler, tıbbi görüntü kodlayıcıları veya dil modeli ince ayarı kullanmadan radyoloji bulgularından kısa izlenimler üretmektedir. Çalışmada iki geri getirme stratejisi değerlendirilmiştir. İlk strateji, klinik olarak uyumlu bulgu-izlenim örneklerini seçmek için anlamsal benzerliği sözcüksel iyileştirme ile birleştirir. İkinci strateji, hedef odaklı özet kestirimi, bulgudan bulguya eşleştirme, bulgudan izlenime eşleştirme, ana kavram aktarımı ve yapılandırılmış klinik eşleştirmeyi birleştiren projeksiyon yönlendirmeli çok sinyalli bir yapı sunar. OpenI ve MIMIC-CXR deneylerinde örneksiz ve rastgele örnekli istemlemenin yetersiz kaldığı, geri getirme yönlendirmeli istemlemenin ise tutarlı kazanımlar sağladığı gösterilmiştir. Az örnekli istemleme, örnek sıralaması, zor negatif örnekler ve yinelemeli istem geri bildirimi üretim kalitesini daha da artırmaktadır. OpenI üzerinde önerilen yöntemler grafik tabanlı, çok modlu ve kapalı kaynak büyük dil modeli tabanlı sistemlerle karşılaştırıldığında güçlü performans göstermektedir. MIMIC-CXR üzerinde projeksiyon yönlendirmeli strateji özellikle klinik etiket uyumu açısından rekabetçi kalmaktadır. MIMIC-III RadSum23 değerlendirme ortamındaki altı BT ve MR alanında yapılan değerlendirme, çerçevenin göğüs röntgeni raporlamasının ötesine aktarımını incelemektedir. Sonuçlar, geri getirme ve istemlemenin açık kaynaklı, yalnızca metne dayalı modelleri radyoloji izlenim üretiminde etkili hâle getirdiğini ve şeffaf, yeniden üretilebilir bir alternatif sunduğunu göstermektedir.

To my beloved wife, Güneş

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

API	Application Programming Interface
BART	Bidirectional and Auto-Regressive Transformers
BERT	Bidirectional Encoder Representations from Transformers
BERTScore	BERT-based semantic similarity score
BioBERT	Biomedical Bidirectional Encoder Representations from Transformers
CCL	ChatGPT-based Contrastive Learning
CE	Clinical Efficacy
CNN	Convolutional Neural Network
CT	Computed Tomography
CXR	Chest X-ray
EMAI	Experience-Guided Multi-Agent Interpretable framework
F1	F1-score
FC	Factual Consistency
FN	False Negative
FP	False Positive
GAT	Graph Attention Network
GPT	Generative Pre-trained Transformer
IDF	Inverse Document Frequency
LCS	Longest Common Subsequence
LLM	Large Language Model
Llama	Large Language Model Meta AI
LLaVA	Large Language and Vision Assistant
MIMIC-CXR	Medical Information Mart for Intensive Care Chest X-ray
MIMIC-III	Medical Information Mart for Intensive Care III
MR	Magnetic Resonance
NLP	Natural Language Processing

OpenI	OpenI Chest X-ray Dataset
PEGASUS	Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence-to-sequence models
RAG	Retrieval-Augmented Generation
REALM	Retrieval-Augmented Language Model Pre-training
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
R1/R2/RL	ROUGE-1, ROUGE-2, and ROUGE-L
Seq2Seq	Sequence-to-sequence
SVD	Singular Value Decomposition
T5	Text-to-Text Transfer Transformer
TF-IDF	Term Frequency-Inverse Document Frequency
TP	True Positive

CHAPTER 1

INTRODUCTION

Radiology reports are one of the primary means through which radiologists communicate diagnostic information to referring physicians and other clinical decision makers ((ESR) 2011; Larson et al. 2013). A typical radiology report includes several sections, among which the *Findings* and *Impression* sections are particularly important. The findings section contains the detailed observations made from the radiological examination, including anatomical descriptions, normal appearances, and abnormal findings. The impression section provides a concise diagnostic summary derived from these observations and is often the part of the report most directly used in clinical decision-making (Zhang et al. 2018; Hartung et al. 2020; Gershanik, Lacson, and Khorasani 2011). Therefore, the automatic generation of a reliable impression from the findings section has become an important problem in clinical natural language processing and medical text summarization.

Radiology summary generation centers on converting detailed clinical findings into an impression that is concise and diagnostically informative. Unlike general-domain summarization, this task cannot be understood merely as shortening a source text. A generated impression must preserve clinically important abnormalities, avoid unsupported statements, maintain anatomical and diagnostic specificity, and follow the style of radiological reporting. Small omissions or inaccurate additions may substantially alter the clinical meaning of the report, a concern that is consistent with broader findings on factual inconsistency and hallucination in abstractive summarization (Maynez et al. 2020; Kryściński et al. 2020). For instance, omitting findings such as pneumothorax, pleural effusion, consolidation, or cardiomegaly may reduce the usefulness of the generated impression, while introducing a finding that is not supported by the findings section may lead to clinically misleading information (Miura et al. 2021; Hartung et al. 2020).

Despite recent advancements in large language models, many current radiology summarization methods depend on proprietary systems, task-specific training procedures, fixed clinical schemas, multimodal encoders, or complex architectural components. Such dependencies may improve performance, but they can also restrict transparency, reproducibility, and the feasibility of deployment in real-world clinical NLP settings (Sun et al.

2023; Ma et al. 2024; Artsi et al. 2025). These concerns are particularly important in medical domains, where system behavior should be inspectable, experimental conditions should be reproducible, and data flow should remain compatible with privacy and local deployment requirements (Price, and Cohen 2019; Char, Shah, and Magnus 2018; Kocak et al. 2023). This situation creates a need for summarization frameworks that are lightweight, controllable, and openly reproducible while still producing clinically coherent impressions (Chen et al. 2023).

Clinical radiology reports introduce distinct difficulties for automatic summarization. They contain specialized medical vocabulary, implicit diagnostic reasoning, negation, uncertainty, and dense anatomical descriptions (Luo, and Arase 2025). A summarization model must preserve clinically relevant information while avoiding unsupported additions, omissions, and medically unsound generalizations. This requirement is more demanding than producing a fluent general-domain summary, because a small difference in wording may alter the diagnostic meaning of the report. For example, omitting findings such as pneumothorax, pleural effusion, consolidation, or cardiomegaly may reduce the usefulness of the generated impression, whereas introducing a finding not supported by the findings section may produce clinically misleading information (Miura et al. 2021; Hartung et al. 2020). Prior studies also show that general-purpose NLP systems may struggle with factual consistency in medical contexts, which is problematic in high-stakes clinical settings where patient safety is central (Asgari et al. 2025; Zhu et al. 2025).

Another difficulty arises from the linguistic variability of radiology reports. Similar clinical conditions may be expressed through different phrases, abbreviations, or reporting styles. Conversely, reports that appear lexically similar may describe clinically different cases. This creates a challenge for systems that rely only on surface-level word overlap. At the same time, purely semantic representations may overlook clinically important terminology, negation, anatomical location, or severity. A useful impression generation system must therefore balance semantic flexibility with lexical precision. This observation motivates the hybrid semantic-lexical retrieval component of this thesis, where dense representations are used to identify semantically related cases and lexical similarity is used to preserve clinically important wording (Lee et al. 2023).

However, direct similarity between findings sections is not the only possible basis for retrieval. In a findings-to-impression task, the most useful reference example may

not always be the report whose findings section is closest to the candidate findings at the source-text level. The impression section is shorter, more selective, and diagnostically focused. It reflects not only what is mentioned in the findings but also how the clinically relevant observations are condensed into a final summary. Therefore, two reports with similar findings may still require different impressions, while reports with different surface forms may lead to similar diagnostic summaries. This motivates a second retrieval strategy in this thesis: impression-space projection retrieval.

The proposed framework therefore considers retrieval from two complementary perspectives. The first strategy performs direct findings-based retrieval. It compares the candidate findings section with reference findings sections, retrieves a broad candidate pool, and refines the selected examples using lexical similarity. The second strategy performs impression-space projection retrieval. Instead of retrieving examples only by comparing findings to findings, it estimates an impression-oriented representation from the candidate findings and compares this projected representation with reference impression representations. In this way, retrieval is not limited to source-side similarity; it also considers a target-oriented representation space that is closer to the expected output of the summarization task.

More recent work has shifted toward generation methods that combine structural constraints with task-specific guidance, especially for radiology report summarization. These methods include graph-based representations, multimodal pipelines that use radiographic images, and retrieval or feedback procedures guided by predefined clinical labels (Hu et al. 2021; 2022; 2023; Ma et al. 2024). Similar forms of structured supervision have also been widely used in chest radiograph research, where predefined observation categories are extracted from free-text reports to support dataset construction and report labeling (Irvin et al. 2019; Smit et al. 2020). These approaches have demonstrated strong performance, but they often depend on handcrafted structures, fixed taxonomies, image encoders, proprietary APIs, or task-specific training procedures (Delbrouck, Zhang, and Rubin 2021; Delbrouck et al. 2022; 2023). Recent LLM-based methods, including experience-guided multi-agent frameworks (Li et al. 2025) and ChatGPT-driven contrastive learning pipelines (Luo et al. 2025), have also shown improved summarization performance. However, these approaches rely on large proprietary models or additional training components, which can restrict transparency, reproducibility, and practical de-

ployment.

In contrast, this thesis is intentionally designed as a fully text-based and open-source framework for radiology impression generation. It operates over free-text clinical report sections rather than medical images, fixed disease-anatomy schemas, or proprietary model outputs. Retrieved examples are integrated into compact prompts for open-source decoder-only language models. The framework is not based on fine-tuning the language model; instead, it studies how retrieval design, representation-based example selection, prompt construction, example ordering, hard negative examples, and prompt-level refinement affect generation quality. This makes the system suitable for controlled experimentation and allows the contribution of each component to be analyzed separately.

The central idea of this thesis is that compact open-source language models can be substantially improved not only through careful prompt construction, but also through the design of the retrieval space from which in-context examples are selected. Rather than treating the language model as a standalone generator, the proposed framework supplies it with clinically relevant examples selected from the training set. These examples act as in-context demonstrations and guide the model toward appropriate radiological phrasing. In this sense, the framework shifts part of the burden from model scale to retrieval quality, representation design, prompt organization, and feedback design.

In addition to positive example selection, this thesis investigates the role of hard negative examples in retrieval-guided prompting. Hard negatives are examples that remain close to the target case according to one similarity signal but differ in clinically important ways according to another signal. Such examples represent plausible but potentially misleading mappings between findings and impressions. By incorporating a small number of hard negatives into the prompt, the framework provides the model not only with examples to imitate, but also with examples whose patterns should be avoided. This design follows the broader intuition of contrastive in-context learning and prompt retrieval studies, where the distinction between positive and negative examples can improve the model's ability to discriminate between appropriate and inappropriate outputs (Gao, and Das 2024; Rubin, Herzig, and Berant 2022).

A major contribution of this work is the comprehensive analysis of how compact open-source language models perform under different retrieval and prompting strategies. Rather than assuming that larger models are inherently superior, the analysis shows that

small open-source models with 7–8B parameters can exhibit distinct and controllable behavioral patterns depending on prompt structure, example ordering, retrieval configuration, and retrieval space design. This result suggests that careful system design may be as important as model scale in the context of clinical text generation.

This thesis therefore focuses not only on obtaining high automatic evaluation scores, but also on understanding how different components of a retrieval-guided generation pipeline affect model behavior. In particular, the study investigates zero-shot prompting, random-shot prompting, retrieval-guided one-shot prompting, retrieval-guided few-shot prompting, example ordering strategies, direct semantic retrieval, hybrid semantic-lexical retrieval, impression-space projection retrieval, lexical metric choices, hard negative prompting, and iterative refinement. Through this analysis, the thesis aims to provide a clearer understanding of how compact decoder-only models can be used effectively in a specialized medical summarization task.

1.1. Contributions of the Thesis

The contributions of this thesis can be summarized as follows:

- A fully open-source and text-only framework is developed for radiology impression generation. The proposed framework operates over free-text report sections and avoids dependencies on proprietary APIs, multimodal encoders, predefined disease-anatomy label schemas, and language model fine-tuning.

- Two retrieval strategies are proposed for selecting clinically relevant in-context examples. The first strategy combines sentence embedding based semantic retrieval with lexical refinement to preserve clinically important terminology. The second strategy extends retrieval toward the impression space by using projection-guided multi-signal matching, allowing reference examples to be selected according to both findings-side similarity and target-oriented summary relevance. The contribution of the individual retrieval signals is further examined through systematic weight ablations.

- The effect of example-based prompting is systematically analyzed through zero-shot, random-shot, one-shot, and few-shot settings. The analysis also examines

example ordering, hard negative examples, and iterative prompt feedback, showing that prompt composition is a critical factor for compact open-source language models in radiology impression generation.

- Multiple compact open-source models and embedding configurations are evaluated on OpenI (Demner-Fushman et al. 2016) and MIMIC-CXR (Johnson et al. 2019), and the proposed approach is compared with prior graph-based, multimodal, label-guided, and proprietary LLM-based systems. An additional evaluation on six CT and MR domains from the MIMIC-III RadSum23 benchmark (Delbrouck et al. 2023) examines generalization beyond chest X-ray reporting. GPT-5.4 is also included as an external controlled generator comparison under fixed retrieval and prompting conditions. Together, these experiments provide a controlled analysis of how language model choice, embedding representation, retrieval strategy, lexical refinement, and prompt design interact across different clinical summarization settings.

1.2. Organization of the Thesis

The thesis is organized as follows. Chapter 1 introduces the problem of radiology impression generation, explains the motivation behind open-source and retrieval-guided summarization, and summarizes the main contributions of the thesis. Chapter 2 provides the necessary background on radiology reports, text summarization, transformer-based language models, sentence embeddings, prompting, retrieval-guided generation, and ROUGE-based evaluation. Chapter 3 reviews the relevant literature on clinical text summarization, radiology report summarization, graph-based and multimodal methods, label-guided approaches, large language models, retrieval-based prompting, semantic representation methods, prompt ordering, hard negative example selection, and cross-modality radiology summarization. Chapter 4 presents the proposed retrieval-guided impression generation framework, including direct findings-based retrieval, lexical refinement, hard negative selection, impression-space projection retrieval, prompt construction, example ordering, and iterative refinement. Chapter 5 describes the datasets, preprocessing protocols, evaluated models, experimental settings, inference environment, retrieval parameters, and evaluation metrics. Chapter 6 presents the experimental results

and discusses the effects of zero-shot prompting, random-shot prompting, one-shot retrieval, few-shot prompting, ordering strategies, hard negative examples, retrieval design choices, projection-based retrieval, multi-signal weight ablations, and computational cost. It also presents a controlled comparison with GPT-5.4 and an additional cross-modality evaluation on the MIMIC-III RadSum23 benchmark. Chapter 7 discusses the broader interpretation of the results, the role of text-only design, comparison with structured and proprietary systems, generalization across radiology modalities, limitations, and clinical implications. Finally, Chapter 8 concludes the thesis by summarizing the main findings and final implications of the study.

CHAPTER 2

BACKGROUND

This chapter presents the conceptual background required to understand the proposed retrieval-guided framework for radiology impression generation. The chapter does not aim to review individual prior systems in detail; that task is reserved for Chapter 3. Instead, the purpose of this chapter is to introduce the main clinical and technical concepts used throughout the thesis, including the structure of radiology reports, abstractive summarization, transformer-based language models, in-context learning, sentence embeddings, sparse and latent text representations, term weighting, n-gram features, target-oriented retrieval, representation projection, feature-level retrieval, multi-signal ranking, lexical similarity, retrieval-guided generation, hard negative examples, ROUGE-based evaluation, and clinical label-based evaluation. These concepts provide the foundation for understanding why this thesis studies not only language model prompting, but also the retrieval space from which in-context examples are selected.

2.1. Radiology Reports and the Findings-to-Impression Task

Radiology reports are structured clinical documents that translate imaging observations into textual diagnostic communication. Although report structure can vary across institutions, chest X-ray reports commonly include sections such as indication, comparison, findings, and impression. The findings section describes radiological observations in greater detail, while the impression section summarizes the clinically most important conclusions. The impression therefore performs a synthesis function: it condenses detailed observations into a form that can be quickly used by clinicians ((ESR) 2011; Hartung et al. 2020; Gershanik, Lacson, and Khorasani 2011).

The importance of the impression section is also reflected in radiology reporting guidelines and educational materials. The European Society of Radiology describes the impression or conclusion section as the interpretive part of the report in which imaging features are synthesized with relevant clinical information to formulate an overall impression ((ESR) 2011). Structured reporting studies similarly suggest that consistent report organi-

zation can improve completeness, clarity, and communication quality (Larson et al. 2013; Nobel, van Geel, and Robben 2022). Radiology communication literature also highlights that the impression section captures critical findings and supports effective communication between radiologists and referring clinicians (Gershanik, Lacson, and Khorasani 2011; Hartung et al. 2020).

The findings-to-impression generation task can be viewed as a specialized form of clinical summarization. The input is a findings section, and the output is the corresponding impression. However, this task differs from general summarization because the output must preserve diagnostic correctness, negation, anatomical location, severity, and uncertainty. A generic or fluent summary is not sufficient if it omits clinically relevant abnormalities or introduces unsupported diagnoses. For this reason, impression generation requires careful control over both content selection and wording (Zhang et al. 2018; Miura et al. 2021).

The findings and impression sections also differ in communicative role. The findings section is descriptive and often includes both normal and abnormal observations. It may mention anatomical regions, comparison with prior studies, uncertainty, and detailed visual descriptions. The impression section is more selective and usually focuses on the most clinically relevant conclusions. Therefore, generating an impression from findings requires more than sentence compression. It requires selecting the most important observations, suppressing less relevant details, preserving clinically important negative findings when needed, and producing a concise diagnostic summary in the style expected by radiologists and clinicians.

This distinction is central to the framework proposed in this thesis. Since the impression is not a simple copy of the findings section, example selection for prompting must consider not only whether two findings sections look similar, but also whether the corresponding impressions provide useful demonstrations. This motivates the use of retrieval-guided generation, hybrid semantic-lexical matching, and impression-oriented retrieval strategies.

2.2. Abstractive Text Summarization

Text summarization is commonly divided into extractive and abstractive approaches. Extractive summarization selects and combines existing fragments from the source text,

whereas abstractive summarization generates new text that may paraphrase, compress, or reorganize the input (Nenkova, and McKeown 2011). Radiology impression generation is naturally closer to abstractive summarization because the impression often rephrases and compresses findings rather than copying them verbatim (Zhang et al. 2018).

Neural abstractive summarization systems traditionally relied on encoder-decoder architectures. The encoder maps the source text into a hidden representation, and the decoder generates the target summary token by token. This general sequence-to-sequence formulation was established in neural machine translation and later adapted across generation tasks (Sutskever, Vinyals, and Le 2014). Attention mechanisms improved these systems by allowing the decoder to focus on different parts of the source text during generation (Bahdanau, Cho, and Bengio 2014). Early neural attention-based summarization systems showed that abstractive summarization could be approached with attention-based neural models rather than purely extractive pipelines (Rush, Chopra, and Weston 2015). Later pointer-generator models improved abstractive summarization by combining generation with copying from the source text, helping reduce repetition and out-of-vocabulary problems (See, Liu, and Manning 2017).

Transformer-based architectures further changed summarization research by replacing recurrent computation with self-attention, enabling more effective modeling of long-range dependencies and greater parallelization during training (Vaswani et al. 2017). Subsequent pretrained encoder-decoder models such as T5 and BART demonstrated that large-scale pretraining can substantially improve sequence-to-sequence generation and summarization performance (Raffel et al. 2020; Lewis et al. 2020). These developments provide the general neural summarization background on which radiology impression generation methods build.

A major limitation of abstractive summarization is factual inconsistency. Because abstractive models generate new text rather than selecting only source spans, they may introduce information that is unsupported by the input. Prior work has shown that neural abstractive summarization models can produce hallucinated or unfaithful content, and that common overlap-based metrics do not fully capture factual consistency (Maynez et al. 2020; Kryściński et al. 2020). This issue is especially important in radiology impression generation, where unsupported additions or omitted abnormalities may change the clinical meaning of the report.

In medical summarization, fluency and lexical overlap are not sufficient by themselves. A generated impression may be grammatically well formed and still be clinically problematic if it changes the polarity of a finding, omits an important abnormality, or introduces a diagnosis not supported by the findings. This thesis therefore treats radiology impression generation as a controlled summarization problem in which retrieval, example selection, prompt construction, and evaluation must be designed with the clinical nature of the text in mind.

2.3. Transformer-Based Language Models

Transformer models are built around self-attention, which allows each token in a sequence to attend to other tokens. This mechanism makes transformers suitable for modeling long textual contexts and has become the foundation of modern language modeling (Vaswani et al. 2017). Transformer-based models can be broadly grouped into encoder-only, encoder-decoder, and decoder-only architectures.

Encoder-only models, such as BERT-style models, produce contextual representations of input text and are commonly used for classification, retrieval, and similarity tasks (Devlin et al. 2019). Encoder-decoder models, such as T5 and BART, are designed for sequence-to-sequence generation and have been widely used in summarization (Raffel et al. 2020; Lewis et al. 2020). Decoder-only models generate text autoregressively and have become especially important in instruction following, in-context learning, and prompt-based generation (Brown et al. 2020).

In this thesis, decoder-only models are used for impression generation, while embedding models are used for retrieval. This separation reflects the modular design of the proposed framework: representation models identify useful examples, and decoder-only language models generate the final impression from the constructed prompt. This design also makes it possible to change the retrieval model and the generation model independently, which supports the controlled analysis of retrieval and prompting strategies.

This modular structure is important for lightweight clinical NLP. Instead of modifying language model weights, the framework changes the information supplied to the model through the prompt. The same decoder-only model can therefore be evaluated under different retrieval strategies, prompt structures, example orderings, and refinement

settings. This allows the thesis to examine how much of the final performance comes from retrieval and prompt design rather than from the scale or fine-tuning of the language model itself.

2.4. In-Context Learning and Prompting

In-context learning refers to the ability of a language model to perform a task by conditioning on examples provided in the prompt, without updating its parameters. Brown et al. (Brown et al. 2020) showed that large autoregressive language models can perform many tasks in zero-shot, one-shot, and few-shot settings by using natural language instructions and demonstrations. In a findings-to-impression setting, demonstrations consist of reference findings-impression pairs. These examples show the model how a detailed findings section should be transformed into a concise impression.

The quality of in-context learning depends on several factors: the relevance of the selected examples, the number of examples, the order in which examples are presented, the wording of the instruction, and the position of the query within the prompt. Prior work has shown that few-shot prompting can be unstable and sensitive to prompt format and example choice (Zhao et al. 2021). Other studies further demonstrate that the ordering of demonstrations can substantially affect performance (Lu et al. 2022). Beyond the existence of demonstrations, prior work also shows that the content and structure of demonstrations strongly affect in-context learning. Min et al. (Min et al. 2022) show that demonstrations can provide useful information about input distribution, output space, and formatting, even when the exact input-output mapping is not always the only driver of performance. Other work formulates demonstration selection as an active or learned selection problem, further supporting the importance of choosing examples rather than sampling them randomly (Zhang, Feng, and Tan 2022). Therefore, prompt construction is not simply a formatting step; it is a central part of the method.

This thesis systematically studies these factors by comparing zero-shot, random-shot, one-shot, and few-shot settings, as well as different example ordering strategies. The aim is to understand how compact decoder-only models behave when the prompt is structured with clinically relevant retrieved examples.

2.5. Sentence Embeddings and Semantic Similarity

Sentence embeddings map textual units into dense vector representations. In this representation space, texts with similar meanings are expected to have vectors that are close to one another. Sentence-BERT introduced a practical way to derive semantically meaningful sentence embeddings using siamese and triplet network structures, enabling efficient semantic similarity comparison with measures such as cosine similarity (Reimers, and Gurevych 2019). This makes sentence embeddings useful for retrieval, clustering, semantic search, and example selection.

In biomedical and clinical settings, domain-specific language can make semantic representation more difficult. Medical reports contain abbreviations, negation, uncertainty, anatomical references, and specialized terms. For this reason, biomedical and medical-domain embedding models may be useful for representing clinical text. In this thesis, the evaluated embedding models include MiniLM, PubMedBERT, and MedEmbed, representing different trade-offs between efficiency and domain specificity (Wang et al. 2020; Gu et al. 2021; Balachandran 2024). MiniLM provides an efficient general-domain representation, PubMedBERT provides a biomedical language representation, and MedEmbed provides a medically oriented embedding model. Comparing these representations helps determine whether domain-specific embeddings improve retrieval for radiology impression generation.

Cosine similarity is used to compare embedding vectors. Given two vectors, cosine similarity measures the angle between them rather than their raw magnitude. This makes it suitable for comparing textual representations produced by embedding models, since the direction of the vector often carries more useful semantic information than the absolute vector length (Reimers, and Gurevych 2019). A higher cosine similarity indicates that two findings sections are closer in semantic content.

Let $\mathbf{u}_c$ denote the embedding vector of the candidate findings section, and let $\mathbf{u}_r$ denote the embedding vector of a reference findings section. The cosine similarity between these two vectors is computed as follows:

$$\text{CosSim}(\mathbf{u}_c, \mathbf{u}_r) = \frac{\mathbf{u}_c \cdot \mathbf{u}_r}{\|\mathbf{u}_c\|_2 \|\mathbf{u}_r\|_2}$$

The resulting score ranges from -1 to 1, where larger values indicate greater

semantic similarity. In practice, reference findings sections are ranked according to this score, and the highest-scoring reports form the initial semantic candidate pool.

Semantic retrieval is useful because radiology reports can express the same clinical meaning in different ways. For example, two reports may describe similar cardiopulmonary findings while using different wording or sentence structure. However, semantic similarity alone may not preserve exact clinical terminology, especially when subtle lexical differences correspond to important diagnostic distinctions. Therefore, the proposed method combines sentence embedding retrieval with lexical refinement.

2.6. Sparse and Latent Text Representations

Text retrieval can be based on different types of representations. Dense sentence embeddings represent a text as a continuous vector produced by a neural model. Sparse lexical representations, in contrast, represent a text according to the words, terms, or symbolic features that occur in it. In a bag-of-words representation, the order of the full sentence is not modeled directly; instead, a document is represented by the features that appear in it and their counts. This type of representation has been widely used in text retrieval and text classification because textual documents often contain many potentially informative features (Joachims 1998).

A classical sparse weighting scheme is term-frequency inverse-document-frequency, commonly known as TF-IDF. Term frequency reflects how often a term occurs in a document, while inverse document frequency reduces the influence of terms that occur frequently across the corpus. The intuition is that rare and distinctive terms are usually more informative than common terms. Salton and Buckley describe term weighting as a central component of automatic text retrieval, where different weighting schemes affect how documents are compared and ranked (Salton, and Buckley 1988). In radiology reports, this distinction is important because common words such as ‘study’, ‘view’, or ‘chest’ are less discriminative than clinical terms such as ‘pneumothorax’, ‘effusion’, ‘opacity’, ‘edema’, or ‘consolidation’.

A simplified inverse document frequency score for a feature f can be written as follows:

$$\text{idf}(f) = \log\left(\frac{N+1}{\text{df}(f)+1}\right) + 1$$

where N is the number of documents in the reference corpus, and $df(f)$ is the number of documents in which feature f appears. The smoothing terms prevent division by zero and reduce instability for rare features. A feature with low document frequency receives a higher IDF value, while a feature that appears in many documents receives a lower value.

Sparse representations can also include short phrases rather than only individual words. N-gram models represent sequences of adjacent tokens and have a long history in statistical language modeling (Brown et al. 1992). In retrieval, unigram features capture individual terms, while bigram features can preserve short phrase-level patterns. This is useful in clinical text because phrases may carry more specific meaning than isolated words. For example, "pleural effusion", "right opacity", and "mild edema" are more informative than their individual tokens alone. Bigram features can therefore help retain local clinical context while keeping the representation simple and interpretable.

Sparse lexical representations are useful in medical text because many clinically important observations are expressed through specific terms. A representation that preserves lexical evidence can help maintain clinically important distinctions, including differences in abnormality type, anatomical location, and severity. However, sparse representations are also limited because they depend heavily on exact feature overlap. Two reports may describe similar clinical conditions using different wording, abbreviations, or phrase structures, which may reduce lexical similarity even when the underlying clinical meaning is close.

Another important issue in radiology text is negation. The difference between "pleural effusion" and "no pleural effusion" is clinically fundamental. If a retrieval representation treats both phrases as similar simply because they share the word "effusion", it may select misleading examples. For this reason, feature extraction in clinical retrieval often needs to preserve polarity or assertion information. A simple way to do this is to encode negated concepts as distinct features, so that an absent finding and a present finding do not match each other as the same lexical event.

Structured clinical features can also be represented as sparse tokens. For example, an extracted observation may be encoded differently depending on whether it is present,

absent, or uncertain. Anatomical mentions can also be represented as symbolic features. Such structured tokens allow symbolic retrieval signals to capture clinical distinctions that may be missed by plain word overlap. This is particularly useful when a retrieval mechanism combines free-text features with structured representations derived from clinical information extraction tools.

RadGraph provides one example of such structured clinical information extraction for radiology reports. It defines an information extraction schema for chest X-ray reports in which clinically relevant entities and relations are extracted from free text (Jain et al. 2021). In this schema, entities can represent observations and anatomical mentions, while relations can connect observations to anatomical locations or modify their clinical status. This makes it possible to convert parts of an unstructured radiology report into symbolic clinical features. For retrieval, such features can complement ordinary word overlap because they preserve clinically meaningful distinctions such as observation type, anatomical grounding, and assertion status. Therefore, RadGraph-style representations provide a useful conceptual basis for structured feature-level retrieval in radiology impression generation.

Latent semantic representations address some limitations of sparse lexical representations by projecting sparse term-document representations into lower-dimensional spaces. Latent semantic analysis uses matrix factorization to identify lower-dimensional structure in text data, allowing texts to be compared according to broader patterns of term co-occurrence rather than only exact word overlap (Deerwester et al. 1990). In practice, singular value decomposition can be used to reduce a high-dimensional sparse text matrix into a smaller latent representation. This lower-dimensional representation may capture broader semantic regularities while reducing noise from rare or highly specific terms.

In the context of radiology impression generation, sparse and latent representations provide a useful complement to neural sentence embeddings. Neural embeddings can capture semantic similarity between findings sections, but they may not always preserve exact clinical terminology or assertion status. Sparse representations preserve terminology, n-gram patterns, and symbolic features, while latent representations can smooth surface-level variation. Multiple representation views can therefore support different retrieval strategies in radiology impression generation.

This distinction is especially relevant for impression-space projection retrieval.

The goal of this strategy is not only to compare findings sections directly, but also to estimate a representation that is closer to the expected impression. For this purpose, the findings and impression sections can be represented in vector spaces, and a mapping can be estimated between these spaces using reference report pairs. The resulting projected representation can then be used to retrieve reference examples whose impressions are close to the estimated target representation.

2.7. Source-Side and Target-Oriented Retrieval

Retrieval is often formulated as a source-side similarity problem. In this formulation, a query is compared with candidate inputs, and the most similar candidate inputs are retrieved. In the context of this thesis, the query is the candidate findings section, and the reference inputs are the findings sections in the reference set. A direct findings-based retrieval strategy therefore ranks reference reports according to how similar their findings sections are to the candidate findings section.

This formulation is intuitive because the findings section is the input available during generation. If two findings sections describe similar observations, then their corresponding impressions may also be useful as demonstrations. However, findings-to-impression generation is not a simple nearest-neighbor copying task. The findings section and the impression section have different communicative roles. The findings section describes imaging observations in a more detailed and descriptive way, whereas the impression section is shorter, more selective, and diagnostically oriented. As a result, similarity in the source findings space does not always imply similarity in the target impression space.

This distinction is important for retrieval-guided prompting. The goal of retrieval is not only to find reports that look similar to the candidate findings, but also to find examples that show how a clinically relevant impression should be written. A report whose findings section is very similar to the query may not always provide the most useful demonstration if its impression differs in diagnostic focus, wording, or level of abstraction. Conversely, a report with different surface wording in the findings section may still contain an impression that is close to the expected output. Therefore, retrieval for impression generation can be considered from two perspectives: source-side retrieval

and target-oriented retrieval.

Source-side retrieval compares the candidate findings with reference findings. This approach is useful for identifying reports that are close to the input case. Target-oriented retrieval, in contrast, attempts to retrieve examples according to the expected output space. Since the true target impression is not available during generation, the target representation must be estimated from the candidate findings. This is the motivation behind the impression-space projection retrieval strategy used in this thesis. The strategy estimates an impression-oriented representation from the candidate findings and retrieves reference examples by comparing this projected representation with reference impression representations.

This idea does not replace direct semantic retrieval. Instead, it provides a complementary retrieval view. Direct retrieval asks which reference findings sections are similar to the candidate findings. Projection-based retrieval asks which reference impressions are close to the estimated impression representation of the candidate case. The first view emphasizes input similarity, while the second emphasizes expected output similarity. Comparing these two retrieval strategies allows the thesis to study whether source-side or target-oriented example selection provides more useful demonstrations for compact decoder-only language models.

2.8. Representation Projection for Retrieval

Representation projection refers to transforming a vector from one representation space into another. In the simplest case, a projection function maps an input vector to an output vector space. In this thesis, this idea is used to connect findings representations with impression representations. The findings section and the impression section are different textual objects: the findings section provides detailed observations, while the impression section summarizes the clinically most important conclusions. Therefore, the representation space of findings and the representation space of impressions may encode different information.

Let $\mathbf{x}_i$ denote the vector representation of a reference findings section F_i, and let $\mathbf{y}_i$ denote the vector representation of its corresponding impression I_i. A projection function $g(\cdot)$ can be estimated from reference report pairs so that $g(\mathbf{x}_i)$ approximates $\mathbf{y}_i$. For a

candidate findings section F_c, the representation $\mathbf{x}_c$ is first computed, and the projection function is applied to obtain an estimated impression-space representation:

$$\hat{\mathbf{y}}_c = g(\mathbf{x}_c)$$

The estimated vector $\hat{\mathbf{y}}_c$ does not generate the final impression by itself. Instead, it is used for retrieval. Reference reports can be ranked by comparing $\hat{\mathbf{y}}_c$ with the stored impression representations $\mathbf{y}_i$. This changes the retrieval criterion from source-side similarity to target-oriented similarity. In other words, the system retrieves examples whose impressions are close to the estimated impression representation of the candidate case.

A simple and effective way to estimate such a mapping is regularized linear regression. Ridge regression estimates a linear relationship between input and output vectors while adding an L_2 penalty to reduce overfitting and stabilize the solution when input dimensions are numerous or correlated (Hoerl, and Kennard 1970). In a text representation setting, this is useful because sparse or latent text vectors can have many correlated features. The ridge objective can be written as follows:

$$\hat{W} = \underset{W}{\operatorname{argmin}} \left(\|XW - Y\|_2^2 + \lambda \|W\|_2^2 \right)$$

Here, X contains findings representations, Y contains the corresponding impression representations, W is the projection matrix, and λ controls the strength of regularization. After estimating $\hat{W}$ from reference report pairs, a candidate findings representation can be projected as follows:

$$\hat{\mathbf{y}}_c = \mathbf{x}_c \hat{W}$$

This formulation is useful because it keeps the generation model unchanged. The projection is part of the retrieval mechanism, not the language model. The language model is still prompted with selected examples and generates the final impression autoregressively. Therefore, the projection-based retrieval strategy changes how examples are selected, while the prompt construction and generation phase remain modular.

The motivation for this projection is specific to the findings-to-impression task. Direct findings-based retrieval assumes that input similarity is the best criterion for selecting demonstrations. Impression-space projection retrieval instead assumes that the most

useful demonstrations are those whose target impressions are close to the expected summary representation. This distinction is important because impression generation requires content selection, compression, and diagnostic abstraction. A projection-based retrieval strategy can therefore be interpreted as a way of aligning example selection with the target side of the summarization task.

2.9. Feature-Level Retrieval and Multi-Signal Ranking

Retrieval for clinical summarization can use multiple similarity signals rather than a single retrieval score. A dense semantic signal may identify examples that are conceptually related to the candidate report, while lexical and symbolic signals may preserve exact clinical terminology, phrase-level overlap, assertion status, and anatomical information. These signals are complementary because each representation captures a different aspect of report similarity.

Feature-level retrieval treats each report as a collection of discrete features. These features may include unigrams, bigrams, negation-aware tokens, or structured clinical tokens. Similarity can then be computed by comparing feature bags. Unlike dense embedding similarity, which compares continuous vectors, feature-level overlap is symbolic: two features match only if they are represented by the same string. This makes feature-level retrieval stricter and more interpretable. It is especially useful in radiology because small lexical differences can correspond to important clinical distinctions.

A simple overlap score counts how many features are shared between a query and a candidate. However, not all features should contribute equally. Matching a rare clinical concept is usually more informative than matching a common word. For this reason, feature overlap can be weighted by IDF. Let q denote the query feature bag, let c denote the candidate feature bag, and let $\mathrm{idf}(f)$ denote the inverse document frequency weight of feature f. An IDF-weighted overlap can be written as follows:

$$\mathrm{Overlap}(q, c) = \sum_{f \in q \cap c} \mathrm{idf}(f) \times \min(q_f, c_f)$$

where q_f and c_f denote the counts of feature f in the query and candidate feature bags. The use of $\min(q_f, c_f)$ limits the matched count to the amount shared by both texts.

This overlap can be normalized into precision, recall, and F1-style scores. Let

$T(q)$ and $T(c)$ denote the IDF-weighted totals of the query and candidate feature bags:

$$T(q) = \sum_{f \in q} \text{idf}(f) \times q_f$$

$$T(c) = \sum_{f \in c} \text{idf}(f) \times c_f$$

The weighted precision, recall, and F1-score can then be written as follows:

$$P_{\text{w}} = \frac{\text{Overlap}(q,c)}{T(c)}$$

$$R_{\text{w}} = \frac{\text{Overlap}(q,c)}{T(q)}$$

$$F1_{\text{w}} = 2 \times \frac{P_{\text{w}} \times R_{\text{w}}}{P_{\text{w}} + R_{\text{w}}}$$

This type of weighted F1 score provides an interpretable lexical similarity signal. It rewards candidates that share important features with the query, while reducing the influence of frequent and less informative terms. Since the features can include negation-aware tokens or structured observation tokens, the same formulation can be used for both free-text and symbolic clinical overlap.

In a findings-to-impression setting, similarity can also be computed across different report sections. Direct findings-based overlap compares the candidate findings with reference findings. Cross-section overlap compares the candidate findings with a reference impression. The first asks whether two reports describe similar source observations, while the second asks whether the candidate findings already contain concepts that appear in the reference impression. These two comparisons are related but not identical, because findings and impressions differ in length, selectivity, and clinical function.

Another useful idea is a carry-over signal. Some findings concepts are likely to appear in the impression, while others are less likely to be included in the final summary. For example, a clinically important abnormality may often be carried from findings into impression, whereas routine descriptors or less salient normal observations may not be repeated. A carry-over score can therefore estimate how reliably a source-side feature tends to reappear on the target side. Conceptually, this produces a bridge between findings

features and impression features. Such a signal is not the same as ordinary lexical overlap because it asks whether the candidate impression contains the query concepts that are expected to carry into an impression.

Multi-signal ranking combines these different views into a single candidate ranking. A retrieval score may include dense projection similarity, findings-to-findings lexical overlap, findings-to-impression lexical overlap, structured clinical overlap, and carry-over evidence. Dense projection similarity can reward semantic closeness in a target-oriented latent space, while lexical and structured signals can preserve surface precision, assertion status, and clinical terminology. The goal is not to make all signals identical, but to let them compensate for one another.

This multi-signal view is important for retrieval-guided impression generation. A candidate example should be similar enough to the input findings to be relevant, but its impression should also be useful as a demonstration of the expected summary. Dense target-oriented projection helps identify examples with similar impression-space meaning, while feature-level signals help prevent the retrieval process from ignoring clinically important lexical and symbolic distinctions. This perspective provides the conceptual basis for retrieval mechanisms that combine semantic, lexical, and structured evidence before selected examples are passed to a language model.

2.10. Lexical Similarity and ROUGE

Lexical similarity measures the overlap between two texts at the word or phrase level. In this thesis, ROUGE is used not only as an evaluation metric but also as a lexical signal during retrieval. ROUGE was originally introduced as a package for automatic evaluation of summaries by comparing generated summaries with reference summaries using overlapping units such as n-grams and word sequences (Lin 2004). ROUGE-1 captures unigram overlap, ROUGE-2 captures bigram overlap, and ROUGE-L measures similarity based on the longest common subsequence.

ROUGE-N measures the overlap of n-grams between a candidate text and a reference text. Let C denote the candidate text, and let R denote the reference text. Let $M_N(C, R)$ denote the number of matched n-grams between C and R. Let $T_N(C)$ denote the total number of n-grams in C, and let $T_N(R)$ denote the total number of n-grams in R.

ROUGE-N precision, recall, and F1-score can be written as follows:

$$P_{\text{ROUGE-N}} = \frac{\sum_{gram_n \in C} \text{Count}_{match}(gram_n)}{\sum_{gram_n \in C} \text{Count}(gram_n)}$$

$$R_{\text{ROUGE-N}} = \frac{\sum_{gram_n \in R} \text{Count}_{match}(gram_n)}{\sum_{gram_n \in R} \text{Count}(gram_n)}$$

$$F1_{\text{ROUGE-N}} = 2 \times \frac{P_{\text{ROUGE-N}} \times R_{\text{ROUGE-N}}}{P_{\text{ROUGE-N}} + R_{\text{ROUGE-N}}}$$

ROUGE-1 is obtained when $N = 1$, and ROUGE-2 is obtained when $N = 2$. In the proposed framework, ROUGE-1 is particularly useful during positive example selection because unigram overlap helps preserve clinically important terms.

ROUGE-L is based on the longest common subsequence between the reference text and the candidate text. Unlike ROUGE-N, which uses fixed-length n-gram overlap, ROUGE-L captures sequence-level similarity by measuring the longest ordered sequence of tokens shared by the two texts. Let $L(C, R)$ denote the length of the longest common subsequence between the candidate text C and the reference text R. Let $|C|$ and $|R|$ denote the token lengths of the candidate and reference texts, respectively. ROUGE-L precision, recall, and F1-score can be written as follows:

$$P_{\text{ROUGE-L}} = \frac{LCS(R,C)}{|C|}$$

$$R_{\text{ROUGE-L}} = \frac{LCS(R,C)}{|R|}$$

$$F1_{\text{ROUGE-L}} = 2 \times \frac{P_{\text{ROUGE-L}} \times R_{\text{ROUGE-L}}}{P_{\text{ROUGE-L}} + R_{\text{ROUGE-L}}}$$

Since ROUGE-L considers ordered subsequence similarity, it can capture broader structural overlap between two texts. In this thesis, ROUGE-L is especially useful for analyzing differences between semantically similar reports whose wording or phrase structure diverges.

For positive example selection, lexical overlap is useful because important clinical terms should be preserved. For hard negative selection, lexical divergence is also useful because semantically similar reports with low lexical overlap may represent clinically

different cases. Thus, ROUGE-based scores serve two roles in the thesis: they are used for evaluation and also for designing the retrieval-guided prompting strategy.

The use of both semantic and lexical signals is motivated by the complementarity between dense and lexical retrieval. Dense representations can capture conceptual similarity, while lexical signals can preserve exact term overlap. Hybrid retrieval studies show that these two forms of matching can provide complementary information (Lee et al. 2023). In the proposed framework, semantic retrieval first selects a clinically related candidate pool, and ROUGE-based lexical refinement then helps prioritize examples with stronger terminology alignment.

2.11. Clinical Label-Based Evaluation

ROUGE-based metrics are useful for measuring lexical overlap between generated and reference impressions, but they do not directly measure whether the generated text preserves the same clinical observations. In radiology report summarization, two impressions may use different wording while expressing similar clinical findings. Conversely, two impressions may have high lexical overlap but differ in the presence, absence, or uncertainty of a clinically important observation. For this reason, several radiology summarization studies also use clinical label-based evaluation in addition to ROUGE.

Clinical label-based evaluation typically follows a two-step procedure. First, an automatic labeler is applied to both the generated impression and the reference impression. The labeler converts each text into a set of clinical observation labels. Second, the predicted labels are compared with the reference labels using precision, recall, and F1-score. In this setting, precision measures how many of the observations predicted by the generated impression are also present in the reference impression, recall measures how many reference observations are recovered by the generated impression, and F1-score combines these two quantities.

A common label schema in chest X-ray research comes from CheXpert, which defines a set of thoracic observations and uncertainty labels derived from radiology reports (Irvin et al. 2019). The original CheXpert labeler is a rule-based system designed to extract observation labels from free-text reports. It is often associated with a 14-observation schema that includes findings such as atelectasis, cardiomegaly, consolidation, edema,

pleural effusion, pneumothorax, and support devices. In some image-classification contexts, a smaller subset of CheXpert observations is also used, which can create ambiguity when studies report only a single "CheXpert" score without specifying the exact label subset.

CheXbert is a related but distinct report labeler. Instead of using only rule-based extraction, CheXbert uses a BERT-based model to assign CheXpert-style observation labels to radiology reports (Smit et al. 2020). Therefore, CheXpert and CheXbert should not be treated as identical tools. CheXpert usually refers to the original dataset, label schema, or rule-based labeler, whereas CheXbert refers to a neural report labeler trained for the same general observation-labeling task. Both can produce labels over similar clinical categories, but their implementations and outputs may differ.

In the radiology summarization literature, clinical label-based metrics are reported under several related names, including factual consistency (FC), clinical efficacy (CE), CheXpert score, and CheXbert-F1. These terms are not always used consistently across papers. In many cases, the underlying idea is similar: generated and reference reports are converted into clinical labels, and agreement between these label sets is measured. However, the exact labeler, label subset, averaging method, and treatment of uncertain or negative labels may vary across studies. As a result, these metrics should be interpreted as clinical label agreement scores rather than as a single perfectly standardized metric.

For a set of clinical labels, precision, recall, and F1-score can be written as follows:

$$P_{\text{clin}} = \frac{TP}{TP+FP}$$

$$R_{\text{clin}} = \frac{TP}{TP+FN}$$

$$F1_{\text{clin}} = 2 \times \frac{P_{\text{clin}} \times R_{\text{clin}}}{P_{\text{clin}} + R_{\text{clin}}}$$

where TP denotes clinical observations that are present in both the generated and reference impressions, FP denotes observations predicted by the generated impression but not present in the reference impression, and FN denotes reference observations missed by the generated impression. These scores may be computed using micro-averaging or macro-averaging. Micro-averaging pools decisions across all labels before computing the

score, while macro-averaging computes scores separately for each label and then averages them. Therefore, two studies may report different values even when both use clinical label-based evaluation, depending on the labeler and averaging scheme.

Clinical label-based evaluation complements lexical metrics. ROUGE emphasizes word and phrase overlap, while CheXpert- or CheXbert-based scores focus on agreement in extracted clinical observations. A model may obtain lower ROUGE scores if it paraphrases the reference impression, but it may still preserve the same clinical observations. Conversely, high lexical overlap does not guarantee clinical correctness if an important abnormality, negation, or uncertainty cue is altered. For this reason, this thesis treats clinical label-based evaluation as a complementary perspective rather than a replacement for ROUGE.

2.12. Retrieval-Guided Generation

Retrieval-guided generation combines information retrieval with text generation. Instead of asking a language model to generate an output from the query alone, the system first retrieves relevant examples or documents and then conditions the model on them. Dense retrieval methods such as Dense Passage Retrieval showed that learned dense representations can retrieve relevant passages effectively, providing a foundation for retrieval-based NLP pipelines (Karpukhin et al. 2020). Retrieval-augmented language models such as REALM and RAG further demonstrated that prediction or generation can benefit from non-parametric retrieved information (Guu et al. 2020; Lewis et al. 2020). Fusion-in-Decoder similarly showed that retrieved passages can be integrated into generative models for knowledge-intensive tasks (Izacard, and Grave 2021).

The framework in this thesis follows the same broad principle but uses retrieval differently. Retrieved examples are not external knowledge passages but previous findings-impression pairs from the reference set. These examples provide demonstrations of clinically appropriate phrasing and allow the prompt to adapt to each candidate findings section. The framework therefore treats retrieval quality as a central component of generation quality. If the retrieved examples are clinically relevant and structurally appropriate, they can guide the decoder-only language model toward more suitable impressions. If the retrieved examples are mismatched, overly generic, or clinically misleading, they can negatively

affect the generated output.

Retrieval-guided generation is especially suitable for the setting of this thesis because it allows the language model to remain unchanged across experimental configurations. The decoder-only model is not fine-tuned for each retrieval strategy. Instead, performance is influenced by selecting useful examples, arranging them in the prompt, and applying prompt-level refinement. This makes the framework modular, interpretable, and easier to reproduce. It also allows the effects of retrieval design and prompt design to be analyzed separately from changes in language model parameters.

In the proposed framework, retrieval can be performed in different representation spaces. Direct findings-based retrieval selects examples according to similarity between the candidate findings and reference findings. Hybrid semantic-lexical retrieval first identifies semantically related candidates and then refines them using lexical overlap. Impression-space projection retrieval changes the retrieval criterion by estimating an impression-oriented representation from the candidate findings and comparing it with reference impression representations. These alternatives share the same generation stage but differ in how the demonstrations are selected.

This modular structure is important for the thesis. The generation model receives a prompt containing the candidate findings and selected reference examples. The retrieval module determines which examples enter the prompt, while the decoder-only model determines how the final impression is generated from that prompt. By separating these components, the thesis can examine how much improvement comes from the retrieval strategy rather than from modifying the language model itself.

2.13. Hard Negatives and Contrastive Prompting

Hard negative examples are examples that appear relevant under one similarity signal but differ from the target case in important ways. In retrieval settings, hard negatives are often used to make models better at distinguishing truly relevant examples from superficially similar but incorrect ones (Rubin, Herzig, and Berant 2022). In this thesis, hard negatives are selected from semantically similar candidates that have lower lexical similarity to the candidate findings. This means that they may be broadly related to the target case but differ in clinically important wording.

The use of hard negatives in prompting is related to contrastive in-context learning. Contrastive prompting studies show that providing both positive and negative examples can help language models distinguish desired outputs from undesired ones (Gao, and Das 2024). In the proposed framework, positive examples show the model how to map findings to a clinically appropriate impression, while hard negative examples expose misleading mappings that should not be imitated.

This contrastive structure is particularly useful in radiology impression generation because reports may be semantically close while differing in critical clinical details. For example, two chest X-ray reports may both discuss lung opacity, but the clinical interpretation may differ depending on location, extent, chronicity, associated effusion, or the presence of comparison studies. By including a small number of hard negatives, the framework aims to make the prompt structure more discriminative and to reduce the chance that the model imitates an inappropriate reference pattern.

Hard negative prompting is also relevant to the broader objective of retrieval-guided generation. Retrieval is not only about finding examples that are close to the query; it is also about controlling which examples influence the model. Positive examples provide patterns to follow, while hard negative examples provide contrastive evidence about patterns that should be avoided. This distinction supports a more structured form of in-context learning in which the prompt contains both supportive and cautionary demonstrations.

2.14. Scope of the Background Chapter

The concepts introduced in this chapter provide the foundation for the rest of the thesis. Radiology report structure and abstractive summarization define the clinical and textual nature of the findings-to-impression task. Transformer-based language models and in-context learning explain why compact decoder-only models can be used through prompting. Sentence embeddings, sparse representations, latent spaces, lexical similarity, representation projection, ROUGE-based evaluation, and clinical label-based evaluation provide the basis for the retrieval and evaluation strategies studied in the thesis. Retrieval-guided generation and hard negative prompting explain how selected examples are incorporated into the prompt and how the model can be guided through demonstrations.

Chapter 3 situates these concepts within the existing literature on clinical summa-

rization, radiology report generation, retrieval-based prompting, structured supervision, and open-source language models. Chapter 4 then explains how these concepts are combined into the proposed framework for open-source, text-only radiology impression generation.

CHAPTER 3

RELATED WORK

3.1. General Text Summarization and Pretrained Language Models

Text summarization has long been studied as a core natural language generation task, with early work distinguishing between extractive and abstractive approaches and later neural work formulating summarization as sequence-to-sequence generation (Nenkova, and McKeown 2011; Sutskever, Vinyals, and Le 2014; Rush, Chopra, and Weston 2015). Earlier neural summarization systems were commonly based on encoder-decoder architectures, in which an encoder represents the source document and a decoder generates a shorter target sequence. With the introduction of attention-based sequence-to-sequence models and later transformer architectures, abstractive summarization systems became increasingly effective at producing fluent summaries that are not limited to copying source sentences directly (Bahdanau, Cho, and Bengio 2014; Rush, Chopra, and Weston 2015; See, Liu, and Manning 2017; Vaswani et al. 2017).

A major shift in summarization research came with large-scale pretraining. Models such as BERT (Devlin et al. 2019) demonstrated the effectiveness of bidirectional contextual representation learning for many language understanding tasks. Although BERT itself is an encoder-only model and is not directly designed for free-form generation, it strongly influenced later pretrained architectures and domain-specific language models. Encoder-decoder models such as T5 (Raffel et al. 2020), BART (Lewis et al. 2020), and PEGASUS (Zhang et al. 2020) further established pretraining as a central approach for abstractive summarization. T5 formulates many NLP tasks within a unified text-to-text framework, BART combines denoising pretraining with a sequence-to-sequence architecture, and PEGASUS introduces a pretraining objective specifically designed for abstractive summarization.

These models are highly relevant to medical summarization because they provide strong general-purpose foundations. However, clinical summarization differs from general-domain summarization in several important ways. It requires the preservation of domain-specific terminology, accurate handling of negation and uncertainty, and avoidance

of unsupported inference. A fluent but clinically inaccurate summary may be unacceptable even if it achieves strong lexical overlap with a reference summary. Therefore, clinical summarization systems require not only language fluency but also factual consistency and domain sensitivity.

3.2. Clinical and Biomedical Text Summarization

Clinical summarization constitutes a more constrained and higher-risk form of text summarization. In clinical documents, the summary must retain medically important concepts while filtering redundant or less relevant information. This is particularly challenging because clinical notes and reports frequently contain abbreviations, implicit reasoning, institution-specific phrasing, and dense terminology. As a result, direct application of general-domain summarization models may be insufficient without domain adaptation or additional clinical guidance.

Domain-specific pretrained language models were introduced to address these limitations. BioBERT (Lee et al. 2020) adapts BERT to biomedical corpora and shows that pretraining on biomedical text improves performance on biomedical text mining tasks. PubMedBERT (Gu et al. 2021) further demonstrates the value of pretraining from scratch on biomedical literature rather than merely continuing pretraining from a general-domain model. These works support the broader observation that biomedical and clinical language often requires representations tuned to domain-specific terminology and distributional patterns.

In the context of summarization, pretrained encoder-decoder models such as T5 and BART have been adapted to clinical documents, including radiology reports. Nishio et al. (Nishio et al. 2024) constructed language models for automatic summarization of radiology reports and evaluated them on datasets including MIMIC-CXR. Such studies show that modern language models can support radiology summarization, but they also highlight that model adaptation, data characteristics, and evaluation design strongly influence performance. This thesis follows the same broad motivation of improving radiology summarization, but instead of fine-tuning large summarization models, it investigates retrieval-guided prompting with compact open-source decoder-only models.

3.3. Factual Consistency in Abstractive Summarization

A central challenge in abstractive summarization is that generated summaries may be fluent but not faithful to the source document. Maynez et al. (Maynez et al. 2020) show that neural abstractive summarization models are prone to hallucinating content that is not supported by the input. Kryscinski et al. (Kryściński et al. 2020) further argue that common summarization metrics do not directly evaluate factual consistency and propose model-based factual consistency evaluation. These findings are especially relevant for medical summarization, where factual errors may have clinical consequences. In radiology impression generation, hallucinated abnormalities, omitted findings, or incorrect negation may change the diagnostic meaning of the report. Therefore, methods for radiology summarization must consider not only fluency and lexical overlap but also whether the generated impression remains grounded in the findings section.

3.4. Radiology Impression Generation

Radiology impression generation is a specialized form of clinical summarization in which the goal is to generate the *Impression* section from the *Findings* section of a radiology report. In this setting, the findings section usually contains detailed observations, while the impression section condenses the clinically most important information into a shorter diagnostic summary. Table 3.1 shows representative findings-impression pairs that illustrate this transformation.

Table 3.1. Representative findings-impression pairs in radiology impression generation.

Findings	**Impression**
The lungs appear clear. The heart and pulmonary vessels appear normal. The pleural spaces are clear. Mediastinal contours are normal.	No acute cardiopulmonary disease.
There is retrocardiac soft tissue density suggesting a hiatal hernia. There is interval development of bandlike opacity in the left lung base, possibly representing atelectasis. No pneumothorax or pleural effusion is seen.	Retrocardiac soft tissue density suggesting hiatal hernia. Left base bandlike opacity suggesting atelectasis.

The impression section summarizes the most clinically important observations and is central to communication between radiologists and referring physicians. Zhang et al. (Zhang et al. 2018) formulated this task as neural sequence-to-sequence summarization and showed that impression generation can be approached as an abstractive summarization problem. Their work is one of the foundational studies for automatic radiology findings summarization.

Subsequent work improved content selection by incorporating clinical and ontological information. Gharebagh et al. (Gharebagh, Goharian, and Filice 2020) investigated the use of medical ontologies for content selection in clinical abstractive summarization. This line of work is important because radiology summarization requires selecting clinically salient information rather than merely compressing text. A system must decide which observations should appear in the final impression and which details can be omitted without changing the diagnostic meaning.

Graph-based methods were later introduced to better represent relationships among clinically important words and concepts. WordGraph (Hu et al. 2021) uses word-level graph structures to guide summarization of radiology findings. GraphContrastive (Hu et al. 2022) extends this direction by incorporating graph-enhanced contrastive learning to improve representation alignment. These methods show that structural information can improve radiology summarization, especially when the model must preserve relations among findings. However, they also introduce additional processing steps, graph construction procedures, or specialized model components.

The task has also been studied in broader radiology report generation settings, where the input may include radiographic images rather than only text. Early work by Jing et al. (Jing, Xie, and Xing 2018) proposed a multi-task framework for generating medical imaging reports with tags and paragraph-level reports. TieNet (Wang et al. 2018) jointly used image and text representations for chest X-ray disease classification and reporting. Later, R2Gen (Chen et al. 2020) introduced a memory-driven transformer for radiology report generation, and Progressive Transformer-based generation (Nooralahzadeh et al. 2021) divided the generation process into concept generation and report generation stages. These studies are important because they demonstrate the value of visual and multimodal information in report generation. Nevertheless, they address a broader image-to-report setting, whereas this thesis focuses specifically on text-based findings-to-impression gen-

eration.

3.5. Graph-Based, Multimodal, and Label-Guided Approaches

Radiology report summarization has increasingly benefited from structured and multimodal signals. Hu et al. proposed a sequence of methods that progressively incorporated graph-based representations and radiograph-derived information. WordGraph (Hu et al. 2021) models word relations within findings, GraphContrastive (Hu et al. 2022) introduces graph-enhanced contrastive learning, and VisualPrompt (Hu et al. 2023) uses radiograph and anatomy prompts to improve impression generation. These approaches achieve strong performance and demonstrate that external structure can help models preserve clinically meaningful content.

However, the use of graphs, image encoders, and anatomical prompts also increases system complexity. Such systems often require additional annotations, pretrained visual backbones, graph construction modules, or task-specific training procedures. This can make reproduction and deployment more difficult, especially when the objective is to design lightweight clinical NLP pipelines. Moreover, multimodal methods require access to radiographic images and corresponding preprocessing pipelines, while in many practical settings only the text report may be available for summarization.

Label-guided retrieval approaches provide another way of introducing clinical structure. In chest X-ray research, predefined observation categories have been widely used to transform free-text radiology reports into structured labels. CheXpert introduced a large chest radiograph dataset and an automatic labeler for extracting 14 observations with uncertainty labels from radiology reports (Irvin et al. 2019). Smit et al. later combined automatic labelers and expert annotations using a BERT-based model to improve radiology report labeling (Smit et al. 2020). These studies demonstrate the usefulness of structured labels in radiology NLP and medical imaging. However, they also illustrate a limitation relevant to this thesis: fixed label spaces may not fully capture the variability and detail of free-text findings.

Structured information extraction from radiology reports provides another way to introduce clinical structure without relying only on raw word overlap. RadGraph (Jain et al. 2021) defines a schema for extracting clinical entities and relations from chest X-

ray reports, including observations, anatomical mentions, and relations between them. Such structured representations are relevant to radiology summarization because they can capture clinical concepts and their anatomical grounding more explicitly than plain lexical features. However, using structured extraction as an auxiliary retrieval signal is different from depending on a fixed disease-label prediction pipeline, since the extracted entities are derived from the free-text report itself.

ImpressionGPT (Ma et al. 2024) retrieves examples using predefined disease-anatomy label pairs and applies iterative self-correction with ChatGPT. This method is particularly relevant to the present thesis because it demonstrates the importance of retrieval and iterative prompting for impression generation. However, it also relies on fixed label schemas and proprietary LLMs. The use of predefined label spaces may limit flexibility when reports contain findings outside the covered categories or when clinical concepts are expressed in non-standard language. In contrast, the present thesis avoids predefined disease-anatomy label pairs for retrieval and operates primarily over free-text findings, while using structured clinical tokens only as auxiliary matching signals.

3.6. LLM-Based Approaches for Radiology Report Summarization

Large language models have recently been explored as flexible generators for radiology report summarization, particularly in text-based settings where impressions are generated directly from free-text findings. EMAI (Li et al. 2025) proposes an experience-guided multi-agent framework in which multiple LLM components collaboratively refine radiological reasoning and improve interpretability through iterative interactions. A similar strategy is found in CCL (Luo et al. 2025), which uses a ChatGPT-based contrastive learning approach to strengthen radiology-specific representations through auxiliary training targets.

While these LLM-centered pipelines demonstrate clear gains in summarization, their architecture is fundamentally separate from the lightweight retrieval-guided setting studied in this thesis. Methods such as EMAI and CCL prioritize multi-agent coordination, contrastive training, or proprietary model assistance. These design choices can improve performance, but they also complicate reproducibility and may limit local deployment in clinical environments. In particular, closed-source models make it difficult to inspect

model behavior, verify system changes over time, or separate the contribution of the model from the contribution of prompting, retrieval, or hidden system components (Haibe-Kains et al. 2020; Kocak et al. 2023).

Recent automatic radiology summarization studies using large language models further suggest that strong language models can improve summarization workflows (Nishio et al. 2024). However, many such approaches depend on either fine-tuning, large model access, or external computational services. The present thesis takes a different direction by investigating whether compact open-source decoder-only models can achieve strong performance through retrieval-guided prompting alone. This enables a reproducible framework without language model fine-tuning, making it more suitable for settings where transparency, privacy, and computational accessibility are important.

3.7. Retrieval-Augmented Generation and In-Context Learning

Retrieval-augmented generation has become an influential strategy for improving language generation by providing models with external contextual information. Dense retrieval is an important foundation for retrieval-guided generation. Karpukhin et al. (Karpukhin et al. 2020) show that dense dual-encoder retrieval can outperform traditional sparse retrieval for open-domain question answering. REALM (Guu et al. 2020) incorporates retrieval into language model pretraining, while RAG (Lewis et al. 2020) combines pretrained sequence-to-sequence models with retrieved non-parametric memory. Fusion-in-Decoder (Izacard, and Grave 2021) further demonstrates that retrieved passages can be integrated into generative models by processing retrieved evidence in the encoder and fusing information during decoding. Although these methods primarily target open-domain question answering or knowledge-intensive generation, they motivate the broader idea that generation can be improved by conditioning the model on retrieved relevant context.

In-context learning provides another foundation for the proposed framework. Brown et al. (Brown et al. 2020) showed that large autoregressive language models can perform tasks from a small number of examples included in the prompt, without gradient-based parameter updates. This observation motivated a large body of work on prompt design, demonstration selection, and example ordering. In the context of this thesis, retrieved findings-impression pairs act as in-context demonstrations that guide

compact decoder-only models toward appropriate radiology impression generation.

The selection of demonstrations is crucial. Rubin et al. (Rubin, Herzig, and Berant 2022) proposed learning to retrieve prompts for in-context learning and showed that performance depends strongly on which examples are selected. This directly motivates the retrieval component of the proposed method. Rather than using random demonstrations, the thesis retrieves clinically relevant examples using retrieval strategies that combine semantic, lexical, projection-based, and structured clinical signals. This design is intended to provide examples that are both semantically appropriate and terminologically aligned with the target findings.

3.8. Semantic Embeddings and Hybrid Retrieval

Dense sentence embeddings are widely used for semantic similarity search because they map text into continuous vector spaces where semantically related texts can be compared using measures such as cosine similarity. Sentence-BERT (Reimers, and Gurevych 2019) introduced a siamese and triplet network modification of BERT to produce sentence embeddings suitable for semantic similarity search. This is relevant to the present thesis because the retrieval stage requires efficient comparison between a target findings section and a large set of candidate reports.

In biomedical and clinical NLP, domain-specific embedding models can provide additional benefits by representing medical terminology more effectively. PubMedBERT (Gu et al. 2021) is trained on biomedical literature and is designed to capture domain-specific biomedical language patterns. MedEmbed (Balachandran 2024) similarly aims to provide medical text embeddings for healthcare-related retrieval and similarity tasks. In this thesis, MiniLM, PubMedBERT, and MedEmbed are evaluated as retrieval models in order to understand whether lightweight general-purpose embeddings or domain-specific embeddings provide better conditioning examples for radiology impression generation.

However, semantic similarity alone may not be sufficient for clinical summarization. Two radiology findings may be semantically close while differing in key diagnostic terms, severity, anatomical location, or negation. Conversely, lexical overlap may preserve important terminology but fail when clinically equivalent findings are expressed through different phrases. Hybrid retrieval addresses this issue by combining dense semantic re-

trieval with sparse or lexical matching. Lee et al. (Lee et al. 2023) show that dense and sparse retrieval signals can be complementary, motivating hybrid approaches that balance semantic coverage with exact lexical matching. The proposed framework follows this logic by first retrieving semantically similar candidates and then applying ROUGE-based lexical refinement.

Beyond direct semantic and lexical matching, retrieval can also be guided by representations that are closer to the target output space. In findings-to-impression generation, the most semantically similar findings section may not always provide the most useful impression example, because the impression depends on which findings are clinically salient and likely to be summarized. This motivates target-oriented retrieval strategies that compare candidate reports not only at the input level but also through representations related to the expected output. In this thesis, this idea is explored through a projection-guided multi-signal matching strategy, where candidate findings are mapped toward an impression-oriented representation and combined with lexical, carry-over, and structured clinical matching signals.

3.9. Prompt Design, Ordering, and Hard Negative Examples

Prompt design has a major impact on the behavior of decoder-only language models. In few-shot prompting, the model does not update its parameters; instead, it conditions generation on the examples and instructions provided in the prompt. Zhao et al. (Zhao et al. 2021) show that few-shot performance can be unstable and sensitive to prompt format, example choice, and example order. Lu et al. (Lu et al. 2022) further demonstrate that the order of demonstrations can cause large performance differences, with some orderings leading to near state-of-the-art performance and others approaching random guessing. These findings motivate the ordering analysis conducted in this thesis.

In addition to ordering, the content of demonstrations is also important. Min et al. (Min et al. 2022) show that demonstrations provide information about formatting, input distribution, and output space, helping explain why even imperfect demonstrations may influence model behavior. Zhang et al. (Zhang, Feng, and Tan 2022) formulate example selection for in-context learning as an active selection problem, further supporting the view that demonstration choice is a key component of prompt-based learning. These

studies motivate the retrieval-based example selection strategy used in this thesis.

The position of information in long prompts is also important. Liu et al. (Liu et al. 2024) show that language models may use long contexts unevenly and can struggle when relevant information appears in the middle of the input. This is relevant to retrieval-guided few-shot summarization because prompts containing many examples may place the most useful evidence at different positions. The present thesis therefore compares normal and reverse example orderings to examine how compact models respond to the placement of relevant demonstrations.

Hard negative examples provide another way to control model behavior. In retrieval-guided prompting, hard negatives are examples that appear relevant according to one similarity signal but differ from the target in clinically important ways. Contrastive in-context learning studies show that positive and negative examples can help guide language model outputs by clarifying both desired and undesired patterns (Gao, and Das 2024). Similarly, prompt retrieval work has shown that contrastive training with positive and hard negative examples can improve the selection of useful prompts (Rubin, Herzig, and Berant 2022). In this thesis, hard negatives are incorporated into prompts to expose the model to plausible but incorrect mappings between findings and impressions, helping the model distinguish clinically meaningful differences during generation.

3.10. Cross-Modality Radiology Summarization and the RadSum23 Benchmark

Most radiology impression generation studies focus on chest X-ray reports. Although this setting provides established datasets and enables comparison with prior work, it represents only a limited part of radiology reporting. Reports produced for computed tomography and magnetic resonance imaging may differ substantially in anatomical coverage, report length, terminology, and summarization style. Evaluating impression generation across multiple modalities and anatomical regions is therefore important for determining whether a method captures general findings-to-impression summarization patterns or mainly benefits from the repetitive structure of chest X-ray reports.

Chen et al. (Chen et al. 2023) introduced MIMIC-RRS to expand radiology report summarization beyond chest X-rays. The dataset was constructed from MIMIC-III and

MIMIC-CXR and covers three imaging modalities and seven anatomical regions. Their experiments evaluated summarization models both within individual modality-anatomy pairs and across different pairs. The results showed that performance can vary considerably across radiology domains, indicating that a system that performs well on chest X-ray reports does not necessarily transfer equally well to reports involving other modalities or anatomical regions. The study also used RadGraph-based evaluation to examine clinical information preservation beyond lexical overlap.

RadSum23 extended this direction through a shared-task setting for multi-modal and multi-anatomical radiology report summarization (Delbrouck et al. 2023). The shared task included a text-based dataset derived from MIMIC-III with eleven modality-anatomy pairs, together with a multimodal dataset based on MIMIC-CXR and an additional test set from Stanford Hospital. By providing common data partitions and evaluation protocols, RadSum23 enabled more reproducible comparison of systems across a broader range of radiology reporting conditions. The task received 112 submissions from 11 participating teams, demonstrating increasing research interest in radiology summarization beyond the conventional chest X-ray setting.

These studies provide the basis for the cross-modality evaluation conducted in this thesis. While OpenI and MIMIC-CXR remain the main datasets for method development, ablation analysis, and comparison with established findings-to-impression systems, the official RadSum23 MIMIC-III setting is additionally used to examine whether retrieval-guided generation transfers to computed tomography and magnetic resonance imaging reports. This evaluation does not change the text-only findings-to-impression formulation of the thesis; instead, it tests the same formulation under more diverse anatomical and reporting conditions.

3.11. Positioning of This Thesis

The studies reviewed above show that radiology impression generation can benefit from several types of guidance: structured graphs, radiographic images, ontology-aware content selection, predefined labels, proprietary LLMs, dense retrieval, and prompt engineering. Each direction contributes useful insights, but many existing systems rely on complex architectures, task-specific training, fixed label schemas, or closed-source mod-

els. These dependencies can complicate reproducibility and limit deployment in clinical NLP settings.

This thesis is positioned as a lightweight and reproducible alternative. It focuses on text-only findings-to-impression generation using compact open-source decoder-only models. The proposed framework combines semantic retrieval, lexical refinement, projection-guided multi-signal matching, structured clinical overlap, hard negative prompting, example ordering analysis, and iterative feedback without requiring multimodal encoders, handcrafted disease-anatomy label pairs, proprietary APIs, or language model fine-tuning. In this respect, the thesis contributes not only a competitive impression generation system but also a controlled analysis of how retrieval and prompt design affect compact language models in a clinically specialized summarization task.

CHAPTER 4

METHOD

Radiology impression generation is formulated as a text based generation task in which a detailed *Findings* section is converted into a concise and clinically coherent *Impression*. In this thesis, the task is addressed through a retrieval-guided prompting framework. Given a candidate findings section, the system retrieves clinically relevant reference examples from a reference report collection and uses these examples as in context demonstrations for a compact open-source decoder only language model. The objective is not to update the language model parameters, but to improve generation quality through example selection, prompt construction, example ordering, and prompt level refinement.

The thesis studies two standalone retrieval strategies within the same prompt based generation framework. The first strategy is *semantic-lexical similarity matching*. It first retrieves a broad pool of semantically related reference reports using sentence embeddings and then applies lexical refinement to select positive examples and hard negative examples. The second strategy is *projection-guided multi-signal matching*. It estimates an impression-oriented representation from the candidate findings and combines this projection signal with lexical, carry-over, and structured clinical matching signals. Both strategies return reference findings and impression pairs that are inserted into the prompt as demonstrations.

Figure 4.1 presents the overall retrieval-guided generation framework. It illustrates the general two phase structure of the method, where reference examples are first selected through retrieval and then inserted into a structured prompt for impression generation. In this figure, the retrieval phase is shown through the semantic-lexical similarity matching strategy. Figure 4.2 presents the second retrieval strategy, projection-guided multi-signal matching, in greater detail. Although the two retrieval strategies differ in how they rank and select reference examples, both operate over the same reference report collection and feed their selected examples into the same prompt based generation stage.

The framework is designed to be open-source, text only, and compatible with compact decoder only language models. It does not require proprietary APIs, visual encoders, handcrafted disease-anatomy label pairs, or language model fine-tuning. Structured clin-

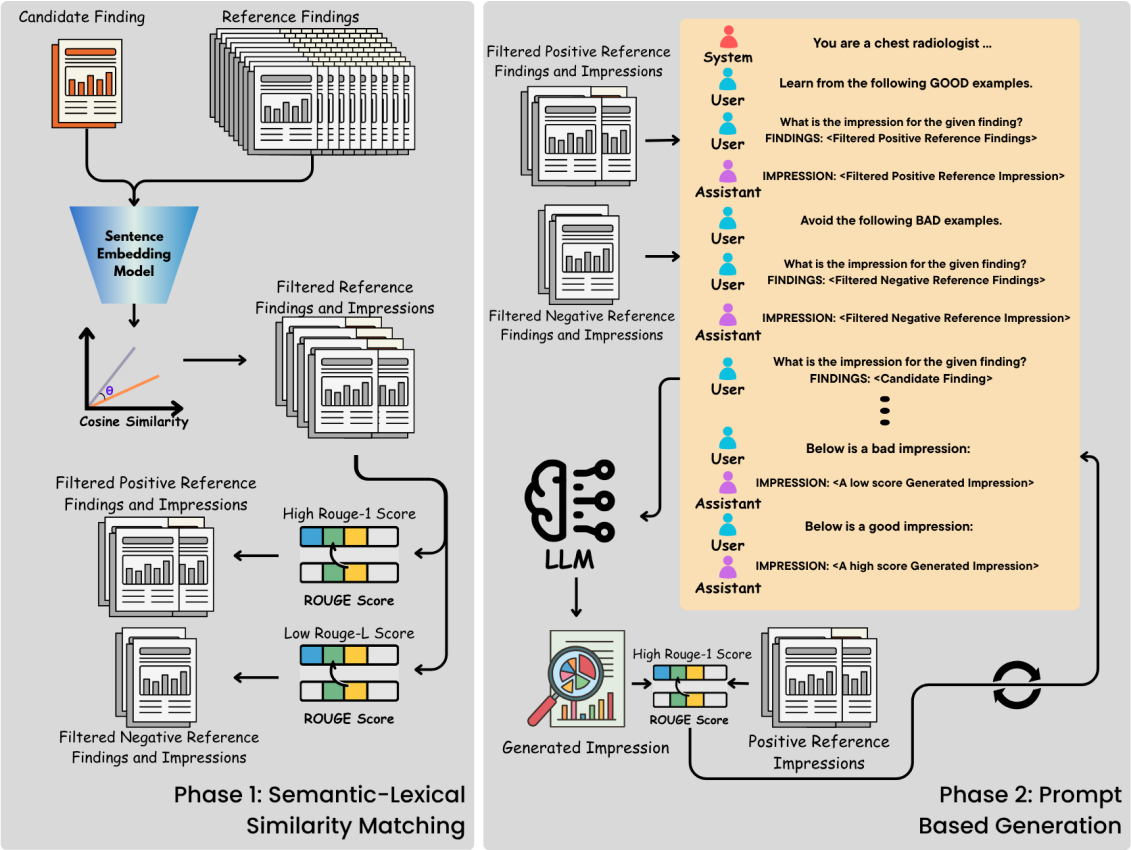

Figure 4.1. Overview of the retrieval-guided generation framework. In Phase 1, semantically similar reference reports are retrieved using sentence embeddings and refined via lexical similarity. From this pool, two sets are constructed: positive examples with strong semantic and lexical alignment, and hard negative examples with high semantic similarity but weaker lexical alignment. In Phase 2, both sets are incorporated into a structured prompt, and the model iteratively refines its output using similarity based feedback.

ical information can be used as an auxiliary matching signal, but the retrieval procedure is not based on predefined disease-anatomy label pairs. This design allows the effects of retrieval quality, lexical refinement, hard negatives, structured matching, and example ordering to be studied in a controlled manner.

In addition to positive examples, the framework uses a small number of hard negative examples. These examples are selected from reports that are related enough to be plausible but weakly aligned according to lexical similarity. Their purpose is to expose the model to plausible but potentially misleading mappings and to help it distinguish clinically appropriate impressions from inappropriate ones. This design is related to contrastive in context learning and prompt retrieval studies, where the use of positive and negative examples can improve discrimination between desired and undesired outputs (Gao, and

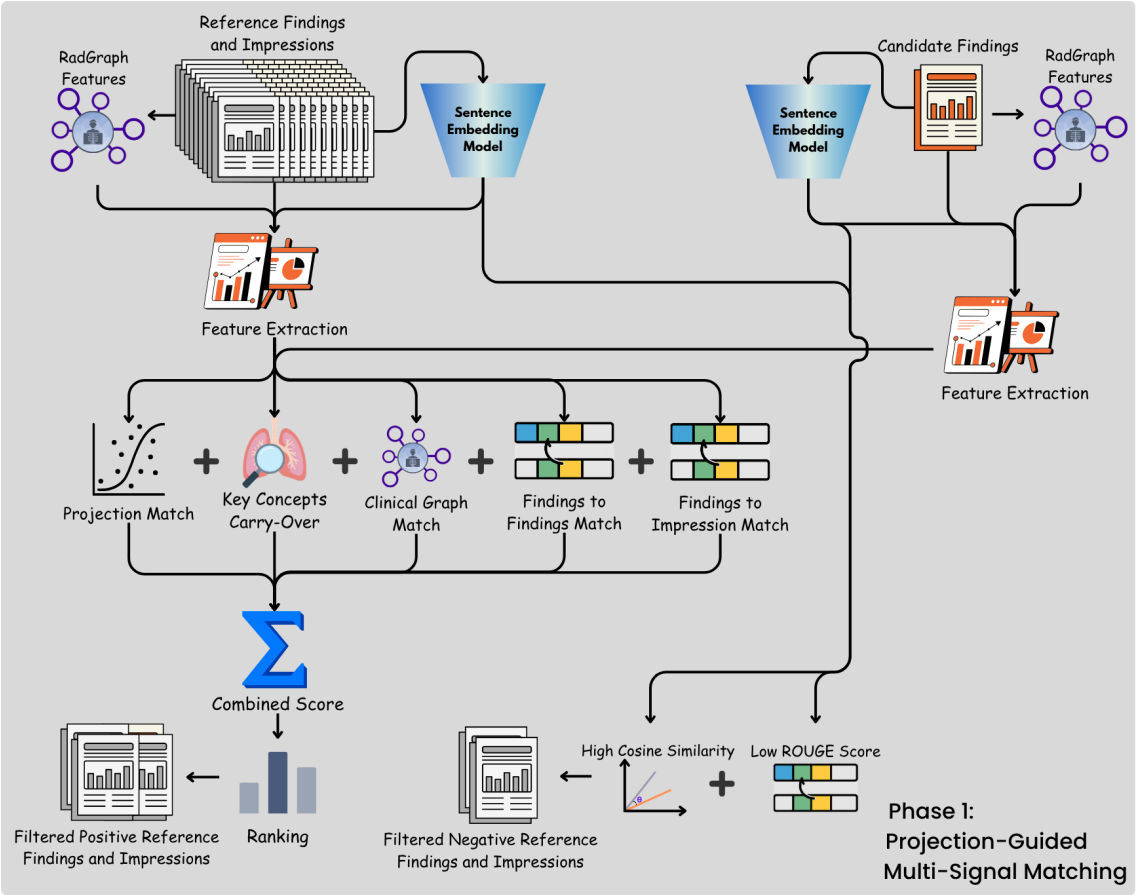

Figure 4.2. Overview of the projection-guided multi-signal matching strategy. Candidate findings and reference report pairs are represented through dense, lexical, and structured clinical features. The retrieval score combines projection based similarity, findings-to-findings matching, findings-to-impression matching, key concept carry-over, and clinical graph matching. The combined ranking is used to select positive reference examples, while hard negatives are selected from related examples with weaker lexical alignment.

Das 2024; Rubin, Herzig, and Berant 2022).

4.1. Problem Formulation and Notation

Let F_c denote the candidate findings section for which an impression must be generated. Let G_c denote the structured clinical representation of the candidate report when such information is available. Let $\mathcal{D} = \{(F_i, I_i, G_i)\}_{i=1}^{N}$ denote the reference report collection, where each F_i is a findings section, each I_i is the corresponding human written impression, and each G_i is the optional structured representation extracted from the report. When structured information is not used, $\mathcal{D}$ can be treated as paired findings-impression

examples $\{(F_i, I_i)\}_{i=1}^{N}$. The goal is to generate an impression $\hat{I}_c$ for F_c by conditioning a language model on a carefully selected subset of reference examples from $\mathcal{D}$.

Each retrieval strategy constructs two sets of reference examples. The positive example set is denoted by $\mathcal{P}$, and the hard negative example set is denoted by $\mathcal{N}$. The positive set contains paired findings and impression examples that are expected to guide the model toward an appropriate impression for the candidate case. The hard negative set contains reports that are related enough to be plausible but lexically divergent in ways that may represent clinically misleading associations.

Formally, the two sets are defined as follows:

- $\mathcal{P} = \{(F_p, I_p)\}$ denotes the positive example set.
- $\mathcal{N} = \{(F_n, I_n)\}$ denotes the hard negative example set.

The final prompt is constructed using the candidate findings F_c, the positive examples $\mathcal{P}$, and the hard negative examples $\mathcal{N}$. The language model then generates an impression conditioned on this prompt. Since the language model weights are not updated, the same retrieval and prompting framework can be applied to different open-source decoder only models without modifying their parameters.

4.2. Retrieval Strategy 1: Semantic-Lexical Similarity Matching

The first retrieval strategy follows a two stage design. The first stage performs broad semantic filtering using sentence embeddings. The second stage applies lexical refinement inside the semantic candidate pool. This strategy is intended to balance semantic flexibility with lexical precision. Semantic similarity captures conceptual relatedness between reports, while lexical similarity helps preserve clinically important terminology. This combination is useful in radiology because two reports may describe similar clinical conditions using different wording, while small lexical differences may also reflect important diagnostic distinctions (Lee et al. 2023).

4.2.1. Semantic Candidate Retrieval

The first step is to identify reference reports whose findings sections are semantically related to the candidate findings. Each candidate findings section and each reference findings section are encoded into dense vector representations using a sentence embedding model. These vectors allow reports to be compared in a continuous semantic space rather than only through exact word overlap.

In this thesis, three embedding models are evaluated: MiniLM-L6 (Wang et al. 2020), PubMedBERT (Gu et al. 2021), and MedEmbed (Balachandran 2024). These models represent different trade-offs between computational efficiency and domain specificity. MiniLM-L6 is a lightweight general purpose sentence embedding model, PubMedBERT is trained on biomedical literature and is therefore more specialized for biomedical language, and MedEmbed is designed for medical similarity and retrieval tasks. Evaluating these models allows the thesis to examine whether domain specific embeddings provide consistent advantages over lightweight general purpose representations in radiology impression generation.

For each candidate findings section F_c, an embedding vector $\mathbf{u}_c$ is computed. Similarly, each reference findings section F_i is encoded as $\mathbf{u}_i$. Semantic similarity is computed using cosine similarity, as defined in Chapter 2. After computing similarities between F_c and all reference findings sections, the top M candidates with the highest semantic similarity scores are retained as the semantic candidate pool:

$$C_M(F_c) = \text{TopM}_{(F_i,I_i)\in\mathcal{D}} \left(\text{CosSim}(\mathbf{u}_c, \mathbf{u}_i)\right)$$

where $C_M(F_c)$ denotes the top M semantically retrieved candidates for F_c. This step acts as a broad semantic filtering stage. It reduces the search space from the full reference collection to a smaller pool of reports that are likely to be clinically related to the candidate case.

4.2.2. Lexical Refinement for Positive Example Selection

Although semantic retrieval is effective for capturing conceptual similarity, it may not always preserve clinically important terminology. In radiology reports, small lexical differences can correspond to important clinical distinctions. Reports that are semantically close may differ in negation, anatomical location, severity, or the presence of a key abnormality. Therefore, after semantic filtering, the framework applies lexical refinement to the semantic candidate pool.

For each report in $C_M(F_c)$, lexical similarity between the candidate findings F_c and the reference findings F_i is computed using a ROUGE based score (Lin 2004). In the main positive selection setting, ROUGE-1 is used because unigram overlap helps preserve important clinical terms. This refinement step favors examples that are not only semantically related but also share clinically relevant wording with the target findings.

Let $\rho_P(F_c, F_i)$ denote the lexical score used for positive example selection. The positive example set $\mathcal{P}$ is constructed by selecting the top k candidates from the semantic candidate pool according to ρ_P:

$$\mathcal{P} = \text{TopK}_{(F_i, I_i) \in C_M(F_c)} \left(\rho_P(F_c, F_i)\right)$$

This two stage selection procedure can be interpreted as follows. Semantic retrieval first identifies reports that are conceptually relevant, and lexical refinement then prioritizes candidates that best preserve important terminology. This design aims to balance broad semantic coverage with precise clinical phrasing.

4.2.3. Hard Negative Example Selection

Hard negative examples are drawn from the same semantic candidate pool used for positive example selection. However, unlike positive examples, they are intentionally selected to have weaker lexical alignment with the candidate findings. In other words, hard negatives are reports that appear related under semantic similarity but differ in important surface level clinical content.

Let $\rho_N(F_c, F_i)$ denote the lexical score used for hard negative selection. The hard

negative set $\mathcal{N}$ is selected from $C_M(F_c)$ by choosing candidates with low lexical similarity according to ρ_N, while excluding the examples already selected as positives:

$$\mathcal{N} = \text{BottomH}_{(F_i,I_i)\in C_M(F_c)\setminus\mathcal{P}}\ (\rho_N(F_c, F_i))$$

where H denotes the number of hard negative examples. This produces examples that are close enough to be plausible but different enough to be potentially misleading. Such examples can help the language model avoid overgeneralizing from semantically related but clinically distinct reports.

For example, two findings sections may both describe abnormalities in the lungs, but one may involve pleural effusion while another involves consolidation. If the model treats these cases as interchangeable, it may generate an incorrect impression. Hard negative examples are included to reduce this risk by showing the model that semantically related findings should not always lead to the same impression pattern.

In the prompt, hard negatives function as contrastive examples. They provide information about mappings that should not be imitated for the candidate case. In practice, only a small number of hard negative examples is used. Adding too many hard negatives may distract the model from the positive demonstrations and weaken the influence of the most relevant references. Therefore, the number of hard negatives is treated as an experimental design parameter and evaluated in the results chapter.

Algorithm 1 summarizes semantic retrieval, positive example selection, and hard negative selection for the semantic-lexical strategy.

4.3. Retrieval Strategy 2: Projection-Guided Multi-Signal Matching

The second retrieval strategy is designed as a standalone alternative to semantic-lexical similarity matching. Instead of relying only on direct similarity between the candidate findings and reference findings, it also estimates what the candidate impression representation may look like and retrieves reference examples that match this estimated target-oriented representation. This strategy is referred to as *projection-guided multi-signal matching*.

The motivation is that the most similar findings section is not always the most useful in context example for impression generation. The impression is a compressed

Algorithm 1 Semantic-Lexical Similarity Matching with Positive and Hard Negative Selection

Require: Candidate findings F_c, reference collection $\mathcal{D} = \{(F_i, I_i)\}_{i=1}^{N}$, semantic pool size M, number of positives k, number of hard negatives H
Ensure: Positive set $\mathcal{P}$ and hard negative set $\mathcal{N}$

```
Encode F_c as u_c
for all (F_i, I_i) ∈ D do
    Encode F_i as u_i
    Compute semantic similarity s_i = CosSim(u_c, u_i)
end for
C_M ← select the top M examples with the highest s_i
for all (F_i, I_i) ∈ C_M do
    Compute positive lexical score p_i = ρ_P(F_c, F_i)
    Compute hard negative lexical score n_i = ρ_N(F_c, F_i)
end for
P ← select the top k examples from C_M according to p_i
N ← select the bottom H examples from C_M \ P according to n_i
return P, N
```

clinical summary, and only some findings are usually carried into it. Therefore, a retrieval mechanism should consider not only whether two findings sections are similar, but also whether a reference impression contains clinically relevant conclusions that are likely to be appropriate for the candidate case. The second strategy addresses this by combining one projection based signal with four feature level matching signals.

4.3.1. Reference Feature Construction

Before ranking, each reference report is converted into several reusable representations. The findings section and the impression section are represented as feature bags. These feature bags are constructed from lowercased lexical tokens, clinically meaningful n-grams, and negation-aware variants of terms. Negation-aware features allow expressions such as 'effusion' and 'no effusion' to be represented differently. This is important because the same clinical term may indicate opposite meanings depending on its assertion status.

When structured clinical information is available, RadGraph-style entities are also converted into structured clinical tokens (Jain et al. 2021). These tokens represent observations, anatomical mentions, and assertion-aware clinical concepts. For example, a present observation and an absent observation are encoded as different structured tokens, so

that positive and negative mentions do not incorrectly match each other. In this thesis, such tokens are used as auxiliary retrieval features rather than as predefined disease-anatomy labels.

Let $\phi_F(F_i)$ denote the feature bag extracted from a reference findings section, $\phi_I(I_i)$ denote the feature bag extracted from a reference impression, and $\phi_G(G_i)$ denote the structured clinical feature bag extracted from the report. The candidate report is represented analogously by $\phi_F(F_c)$ and $\phi_G(G_c)$. The same feature extraction rules are applied to candidate and reference reports.

IDF weights are estimated once from the reference report collection. Let $\mathrm{df}(f)$ denote the number of reference reports in which feature f occurs, and let N denote the number of reference reports. The IDF value of a feature is computed as:

$$\mathrm{idf}(f) = \log\left(\frac{N+1}{\mathrm{df}(f)+1}\right) + 1$$

These weights assign larger values to rare and clinically distinctive features, while reducing the effect of very common words.

4.3.2. Projection Mapping into Impression Space

The first signal in projection-guided matching is a dense projection signal. The goal is to estimate an impression-oriented representation from the candidate findings without using the true candidate impression. To do this, a projection mapping is estimated once from the reference report collection.

For each reference report, an input representation $\mathbf{x}_i$ is constructed from the findings side using complementary representation views. The source representation includes TF-IDF features reduced with SVD, structured clinical tokens derived from the report, and normalized sentence embeddings from multiple encoders. In this study, MiniLM, MedEmbed, and PubMedBERT embeddings are concatenated with the reduced sparse representation to form the final findings-side vector. A target representation $\mathbf{y}_i$ is constructed from the corresponding reference impression using an impression-space text representation. The projection mapping learns a linear transformation from this fused findings-side representation to the impression-side representation.

Let X be the matrix of reference findings-side representations and let Y be the

matrix of reference impression-side representations. The projection matrix $\hat{W}$ is estimated using ridge regression:

$$\hat{W} = \underset{W}{\operatorname{argmin}} \left(\|XW - Y\|_2^2 + \lambda \|W\|_2^2 \right)$$

where λ is the regularization coefficient. The projection is estimated only from reference reports and is then kept fixed. It does not update the language model and does not use the candidate impression.

For a candidate findings section, the same findings-side representation is computed as $\mathbf{x}_c$. The estimated impression representation is then obtained as:

$$\hat{\mathbf{y}}_c = \mathbf{x}_c \hat{W}$$

This vector represents the predicted location of the candidate report in the impression representation space. It is compared with the stored impression representations of the reference reports to compute the projection based matching signal.

4.3.3. Feature Bag Construction

Before defining the individual matching signals, it is useful to describe how feature bags are constructed. In the projection-guided retrieval strategy, a feature bag is a count-based representation extracted from a report section. It is not a dense embedding. Instead, it is a multiset of discrete clinical features, where each feature can occur once or multiple times.

For findings and impression text, the extraction procedure first normalizes the text by lowercasing it and extracting alphanumeric tokens. Very short tokens, general stopwords, and highly frequent radiology terms that are not sufficiently discriminative are removed. Negation words such as 'no', 'not', 'without', and 'absent' are preserved because they change the clinical meaning of a finding. When a negation cue is detected, nearby clinical tokens are marked with a negation prefix. This allows the representation to distinguish findings such as 'pneumothorax' from 'no pneumothorax'.

After filtering and negation marking, the remaining tokens are used as unigram features. However, unigram features alone lose local word associations. This can create

misleading matches when the same words appear in different clinical contexts. For example, if the candidate findings contain 'mild pleural effusion', a unigram representation may match another report containing 'mild atelectasis' and 'large effusion', because both reports contain the words 'mild' and 'effusion'. To reduce this problem, adjacent surviving tokens are also combined into bigram features. A bigram such as 'mild_effusion' preserves the local association between the modifier and the finding, and therefore gives a more specific signal than the two separate unigram features. In this way, bigrams help the findings-to-findings and findings-to-impression matching signals capture short clinical phrase patterns rather than only isolated word overlap. The resulting feature bag records both unigram and bigram features together with their counts.

The same text-based feature extraction procedure is applied to reference findings and reference impressions. The notation $\phi_F(F_i)$ denotes the feature bag extracted from the findings section of the i-th reference report, while $\phi_I(I_i)$ denotes the feature bag extracted from its impression section. For the candidate report, $\phi_F(F_c)$ denotes the feature bag extracted from the candidate findings section. These feature bags are then compared using IDF-weighted overlap scores.

Structured clinical feature bags are constructed separately from RadGraph-style entity annotations when available. These features encode clinical observations, anatomical mentions, and assertion status. For example, present, absent, and uncertain observations are represented with different feature prefixes, and anatomical entities are represented as anatomy features. The notation $\phi_G(G_i)$ denotes the structured clinical feature bag of the i-th reference report, and $\phi_G(G_c)$ denotes the structured clinical feature bag of the candidate report.

Thus, the feature-level signals do not operate directly on raw text. They operate on extracted feature bags. Findings-to-findings matching compares $\phi_F(F_c)$ with $\phi_F(F_i)$, findings-to-impression matching compares $\phi_F(F_c)$ with $\phi_I(I_i)$, key concept carry-over uses feature-level statistics learned from the reference report collection, and clinical graph matching compares $\phi_G(G_c)$ with $\phi_G(G_i)$.

4.3.4. Feature-Level Matching Signals

Projection-guided multi-signal matching ranks each reference report using five complementary signals. These signals are designed to capture different notions of usefulness for an in-context example. Some signals compare the candidate report with the reference findings, some compare the candidate report with the reference impression, and one signal compares the candidate with an estimated impression-space representation. In Figure 4.2, simplified names are used for readability. In this section, each signal is defined more explicitly.

The first signal is *Projected Summary Match*. This signal does not compare the candidate findings directly with the reference findings. Instead, it first maps the candidate findings representation into the impression representation space. The resulting vector $\hat{\mathbf{y}}_c$ can be interpreted as the estimated impression-side representation of the candidate report. This estimated representation is then compared with the stored impression representation $\mathbf{y}_i$ of each reference report.

$$S_{\text{proj}}(c, i) = \max\left(0, \text{CosSim}(\hat{\mathbf{y}}_c, \mathbf{y}_i)\right)$$

A high projected summary match means that the reference impression is close to the impression that the projection mapping expects for the candidate findings. This signal is useful because the best demonstration is not always the report with the most similar findings text. In impression generation, the selected example should also contain a summary pattern that is compatible with the expected output. The true candidate impression is never used in this step.

The remaining four signals are computed over discrete feature bags. These feature bags contain lexical terms, clinically meaningful n-grams, negation-aware variants, and structured clinical tokens when available. To compare two feature bags, the method uses an IDF-weighted overlap score. Given two feature bags q and r, the weighted overlap is defined as follows.

$$\text{Overlap}(q, r) = \sum_{f \in q \cap r} \text{idf}(f) \times \min(q_f, r_f)$$

The total IDF-weighted mass of each feature bag is computed separately.

$$T(q) = \sum_{f \in q} \mathrm{idf}(f) \times q_f$$

$$T(r) = \sum_{f \in r} \mathrm{idf}(f) \times r_f$$

Weighted precision measures how much of the reference-side feature mass is covered by the candidate-side features.

$$P_{\mathrm{w}} = \frac{\mathrm{Overlap}(q,r)}{T(r)}$$

Weighted recall measures how much of the candidate-side feature mass is covered by the reference-side features.

$$R_{\mathrm{w}} = \frac{\mathrm{Overlap}(q,r)}{T(q)}$$

The final feature overlap score is the weighted F1 score.

$$F1_{\mathrm{w}}(q,r) = 2 \times \frac{P_{\mathrm{w}} \times R_{\mathrm{w}}}{P_{\mathrm{w}} + R_{\mathrm{w}}}$$

The second signal is *Findings-to-Findings Match*. This signal compares the candidate findings section with the findings section of a reference report.

$$S_{\mathrm{ff}}(c,i) = F1_{\mathrm{w}}(\phi_F(F_c), \phi_F(F_i))$$

Here, $\phi_F(F_c)$ denotes the feature bag extracted from the candidate findings section, and $\phi_F(F_i)$ denotes the feature bag extracted from the findings section of the i-th reference report. A high findings-to-findings match means that the candidate report and the reference report describe similar clinical content at the input level. For example, if both findings sections mention similar abnormalities, locations, or negated observations, this score increases. This signal preserves direct case similarity and prevents the retrieval process from selecting examples whose impressions look useful but whose original findings are clinically unrelated to the candidate case.

The third signal is *Findings-to-Impression Match*. This signal compares the candidate findings section with the impression section of a reference report.

$$S_{\mathrm{fi}}(c, i) = F1_{\mathrm{w}}(\phi_F(F_c), \phi_I(I_i))$$

Here, $\phi_I(I_i)$ denotes the feature bag extracted from the impression section of the i-th reference report. This signal asks a different question from findings-to-findings matching. Instead of asking whether two findings sections are similar, it asks whether the reference impression already contains terms or concepts that are present in the candidate findings. A high findings-to-impression match means that the conclusion written in the reference impression is lexically connected to the candidate input. This is useful because an in-context example should not only have a similar input report, but should also demonstrate how relevant findings are compressed into an impression.

The fourth signal is *Key Concept Carry-Over*. This signal is designed to model a different type of evidence from ordinary lexical overlap. Findings sections often contain many words and short clinical phrases, but not all of them are equally important for the final impression. Some concepts are routinely summarized in the impression because they represent clinically salient conclusions, while others remain descriptive details in the findings section. Therefore, the carry-over signal estimates which candidate findings features are likely to be carried into an impression.

The carry-over statistics are estimated from the reference report collection. For each feature f, the method counts how often f appears in the findings-side feature bag of a reference report and how often the same feature also appears in the corresponding impression-side feature bag. Let $\mathrm{df}_F(f)$ denote the number of reference reports in which f appears in the findings-side feature bag. Let $\mathrm{hit}(f)$ denote the number of reference reports in which f appears both in the findings-side feature bag and in the corresponding impression-side feature bag. A feature that frequently appears in findings but rarely appears in impressions receives a low carry-over value. A feature that often survives from findings into impressions receives a higher carry-over value.

$$\mathrm{carry}(f) = \frac{\mathrm{hit}(f)+0.5}{\mathrm{df}_F(f)+1.0}$$

This probability is then multiplied by the IDF value of the feature. This prevents very common and weakly informative features from dominating the score. As a result, the final carry-over weight becomes high only when a feature is both relatively informative

and likely to appear in impressions when it appears in findings.

$$w_{\text{carry}}(f) = \text{carry}(f) \times \text{idf}(f)$$

For a candidate report, the method first identifies the candidate findings features that have a learned carry-over weight. This set is denoted by Q_c. The score of a reference report is then computed by checking how much of this weighted candidate carry-over mass is covered by the reference impression. In other words, the signal asks whether the reference impression contains the candidate concepts that are usually important enough to be summarized.

$$S_{\text{carry}}(c, i) = \frac{\sum_{f \in Q_c \cap \phi_I(I_i)} w_{\text{carry}}(f)}{\sum_{f \in Q_c} w_{\text{carry}}(f)}$$

This signal is asymmetric. It starts from the candidate findings and evaluates whether the reference impression contains the candidate-side concepts that are expected to carry over into impressions. Therefore, it is not the same as findings-to-impression lexical overlap. Findings-to-impression matching treats all overlapping features according to the weighted feature overlap formula, whereas key concept carry-over gives special importance to features that have historically moved from findings into impressions in the reference collection.

For example, if a feature such as 'pneumothorax' frequently appears in impressions whenever it appears in findings, it receives a relatively strong carry-over weight. If the candidate findings contain this feature, then reference impressions that also contain it receive a higher carry-over score. In contrast, a general or descriptive feature that often appears in findings but rarely appears in impressions receives a weaker carry-over weight, even if it overlaps lexically. If the candidate report does not contain any feature with a carry-over weight, the carry-over score is set to zero.

The fifth signal is *Clinical Graph Match*. This signal compares structured clinical tokens extracted from the candidate report and the reference report.

$$S_{\text{graph}}(c, i) = F1_{\text{w}}(\phi_G(G_c), \phi_G(G_i))$$

Here, $\phi_G(G_c)$ denotes the structured clinical feature bag extracted from the candidate report, and $\phi_G(G_i)$ denotes the structured clinical feature bag extracted from the

i-th reference report. This signal captures agreement at the level of structured clinical facts rather than plain word overlap. The structured tokens can represent observations, anatomical mentions, and assertion status. For example, a present observation and an absent observation are encoded differently, so the method does not treat 'effusion' and 'no effusion' as equivalent only because they share the same surface term. A high clinical graph match means that the candidate and reference reports contain similar structured clinical content. This signal is used as an auxiliary refinement that helps distinguish clinically similar reports from reports that merely share isolated words.

4.3.5. Combined Ranking and Example Selection

For each reference candidate, the final projection-guided retrieval score is computed as a weighted combination of the five signals:

$$S_{\text{pg}}(c,i) = 0.30S_{\text{proj}}(c,i) + 0.45S_{\text{ff}}(c,i) + 0.05S_{\text{fi}}(c,i) + 0.15S_{\text{carry}}(c,i) + 0.05S_{\text{graph}}(c,i)$$

The coefficients in this equation are nominal weights. Their effect on the final ranking depends not only on the coefficient value, but also on the scale and variation of the corresponding signal. Therefore, a signal with a smaller nominal weight can still have a substantial effect if its raw values are consistently high or if it varies strongly among top candidates. In this design, findings-to-findings matching and projected summary match are intended to act as the main ranking drivers. Findings-to-findings matching preserves direct clinical similarity between the candidate findings and reference findings, while projected summary match provides a target-oriented signal by comparing the estimated impression representation with reference impression representations.

The remaining signals are used as refinements. Key concept carry-over emphasizes candidate concepts that are likely to appear in impressions. Findings-to-impression matching checks whether the reference impression already reflects concepts observed in the candidate findings. Clinical graph matching adds structured clinical agreement based on observation and anatomy features. Although these auxiliary signals have smaller nominal or effective contributions, they can still affect the ordering of near-equal candidates

and serve as tie-breaking evidence in clinically specific cases. Thus, the final ranking is not a simple unweighted summation, but a weighted fusion of dense, lexical, carry-over, and structured clinical evidence.

The positive example set $\mathcal{P}$ is selected by ranking reference reports according to $S_{\text{pg}}(c, i)$ and choosing the top k examples. During selection, duplicate or near-duplicate impressions can be skipped to avoid filling the prompt with repeated boilerplate summaries. This produces a diverse set of reference findings-impression pairs for in context learning.

Hard negative selection follows the same contrastive principle used in semantic-lexical similarity matching. In the projection-guided strategy, positive examples are selected from the highest ranked reference reports according to the combined score. Hard negatives are then selected from the remaining high-ranked non-positive reference reports. Within this pool, examples with low lexical alignment to the candidate findings are preferred. This keeps the role of hard negatives consistent across both retrieval strategies: they remain plausible enough to be relevant, but weakly aligned enough to serve as misleading examples that should not be imitated.

Algorithm 2 summarizes the projection-guided multi-signal matching strategy.

4.4. Prompt Construction

After constructing the positive set $\mathcal{P}$ and hard negative set $\mathcal{N}$ using one of the retrieval strategies, the next step is to build the prompt given to the language model. The prompt contains four main components:

- a system level instruction defining the role of the model as a chest radiology assistant,
- retrieved positive paired findings and impression examples,
- selected hard negative examples or feedback examples,
- the candidate findings section for which the final impression must be generated.

The positive examples are presented as demonstrations of correct findings to impression mappings. Each demonstration contains a reference findings section followed by its corresponding impression. These examples are intended to guide the model toward the expected style, level of detail, and clinical phrasing of the impression section.

Algorithm 2 Projection-Guided Multi-Signal Matching

Require: Candidate findings F_c, optional candidate structured features G_c, reference collection $\mathcal{D} = \{(F_i, I_i, G_i)\}_{i=1}^{N}$, projection matrix $\hat{W}$, IDF weights, carry-over weights, number of positives k, hard negative candidate pool size B, number of hard negatives H

Ensure: Positive set $\mathcal{P}$ and hard negative set $\mathcal{N}$

Extract candidate feature bags $\phi_F(F_c)$ and $\phi_G(G_c)$
Build candidate findings-side representation $\mathbf{x}_c$
Compute estimated impression representation $\hat{\mathbf{y}}_c \leftarrow \mathbf{x}_c \hat{W}$
for all $(F_i, I_i, G_i) \in \mathcal{D}$ **do**
    Obtain stored feature bags $\phi_F(F_i)$, $\phi_I(I_i)$, and $\phi_G(G_i)$
    Obtain stored reference impression representation $\mathbf{y}_i$
    Compute projected summary match $S_{\text{proj}}(c, i)$
    Compute findings-to-findings match $S_{\text{ff}}(c, i)$
    Compute findings-to-impression match $S_{\text{fi}}(c, i)$
    Compute key concept carry-over $S_{\text{carry}}(c, i)$
    Compute clinical graph match $S_{\text{graph}}(c, i)$
    Compute final score $S_{\text{pg}}(c, i)$
end for
$\mathcal{P} \leftarrow$ select the top k diverse examples according to $S_{\text{pg}}(c, i)$
$C_H \leftarrow$ select the top B high-ranked non-positive examples after excluding $\mathcal{P}$
for all $(F_i, I_i, G_i) \in C_H$ **do**
    Compute hard negative lexical score $n_i = \rho_N(F_c, F_i)$
end for
$\mathcal{N} \leftarrow$ select the bottom H examples from C_H according to n_i
return $\mathcal{P}, \mathcal{N}$

Although the retrieval stage returns a ranked positive set, the generation stage does not necessarily need to use all retrieved positive examples. In iterative configurations, only the highest ranked subset of the retrieved positive set is used for prompt construction and feedback computation. This prompt level filtering keeps the context focused on the most reliable demonstrations and reduces the influence of lower ranked examples that may provide weaker or noisier guidance. The exact values of the retrieved positive set size and the iterative subset size are specified in the experimental setup.

Hard negative examples are incorporated as examples that should be avoided or treated as misleading. Depending on the prompt configuration, they can be explicitly marked as negative examples or included in a feedback style format. This allows the model to compare desirable and undesirable impression patterns within the same context.

The prompt templates follow the structure introduced by ImpressionGPT (Ma et al. 2024), but the implementation differs in two important ways. First, the proposed frame-

work uses compact open-source decoder only models rather than proprietary ChatGPT based models. Second, retrieval operates primarily over free text report sections and may use semantic, lexical, projection based, and structured clinical matching signals rather than predefined disease-anatomy label pairs. The complete prompt templates used in the experiments are provided in Appendix A.

4.5. Example Ordering Strategies

The order of retrieved examples can influence the behavior of decoder only language models. In few-shot prompting, the model conditions its output on a sequence of demonstrations, and the position of each example may affect how strongly it influences the final generation. Therefore, this thesis explicitly investigates example ordering as part of the retrieval-guided generation framework.

Two ordering strategies are considered:

- **Normal order**: examples are arranged from most similar to least similar.
- **Reverse order**: examples are arranged from least similar to most similar.

In normal order, the model first observes the most relevant examples. This may be beneficial if the model places strong emphasis on early context or uses the first examples to establish the expected pattern. In reverse order, the most relevant examples appear closer to the candidate query. This may be beneficial if the model exhibits recency bias and relies more heavily on information near the end of the prompt.

The effect of ordering is evaluated separately for each model. This is important because different decoder only models may not use context in the same way. By comparing normal and reverse ordering, the thesis examines whether model-specific prompt sensitivity affects radiology impression generation.

4.6. Prompt Based Generation with Iterative Refinement

The second phase of the framework performs impression generation using the selected examples and prompt structure. The language model receives the constructed prompt and generates an initial impression for the candidate findings section. Instead of

directly accepting this first output, the framework applies an iterative refinement procedure based on similarity feedback.

The overall generation process is outlined in Algorithm 3. The retrieval stage first returns a ranked positive set $\mathcal{P}_k$. For iterative refinement, the framework uses only the highest ranked subset of this set, denoted as $\mathcal{P}_m$, where $m \leq k$. The initial prompt P_0 is constructed using the system instruction, $\mathcal{P}_m$, hard negative examples, and the candidate findings section. The model then generates an initial impression I_{gen}. This generated impression is compared against the impressions in $\mathcal{P}_m$ using ROUGE based similarity. The comparison produces a feedback signal that determines whether the generated output should be treated as a GOOD or BAD example in the next prompt.

Let $\mathrm{R}(I_{\text{gen}}, I_p)$ denote the ROUGE based similarity between the generated impression and a positive reference impression I_p. The average similarity between the generated impression and the positive set is computed as:

$$q(I_{\text{gen}}, \mathcal{P}_m) = \frac{1}{|\mathcal{P}_m|} \sum_{(F_p, I_p) \in \mathcal{P}_m} \mathrm{R}(I_{\text{gen}}, I_p)$$

where $q(I_{\text{gen}}, \mathcal{P}_m)$ is the average similarity score used to assign the generated output as GOOD or BAD. Outputs with stronger similarity to the highest ranked positive impressions are treated as GOOD examples, whereas outputs with weaker similarity are treated as BAD examples. This feedback is then incorporated into the next prompt.

The purpose of iterative refinement is to guide the model using feedback without changing its parameters. A GOOD example reinforces generation patterns that are close to the retrieved positive impressions, while a BAD example signals content or phrasing that should be avoided. This feedback is incorporated only at the prompt level. No gradient updates, language model fine-tuning, reinforcement learning, or additional correction model is used.

The feedback examples used in iterative refinement are generated dynamically. The first prompt contains the highest ranked positive subset $\mathcal{P}_m$ and the hard negative examples. After the model generates an impression, the generated output is evaluated against the impressions in $\mathcal{P}_m$. Depending on the resulting similarity score, it is added to the feedback context as a GOOD or BAD generation. In subsequent iterations, the model receives not only retrieved examples but also feedback derived from its own previous generations.

Algorithm 3 Prompt Based Generation with Iterative Refinement

Require: Candidate findings F_c, ranked positive set $\mathcal{P}_k$, hard negative set $\mathcal{N}$, language model L, iteration limit T, iterative positive subset size m
Ensure: Final impression I_{final}
$\mathcal{P}_m \leftarrow$ select the top m examples from $\mathcal{P}_k$
Construct initial prompt P_0 using $\mathcal{P}_m$, $\mathcal{N}$, and F_c
Generate initial impression $I_{\text{gen}} \leftarrow L(P_0)$
Compute average ROUGE similarity $q(I_{\text{gen}}, \mathcal{P}_m)$
Assign I_{gen} as GOOD or BAD according to $q(I_{\text{gen}}, \mathcal{P}_m)$
for $t = 1$ to T **do**
    Construct prompt P_t using $\mathcal{P}_m$, $\mathcal{N}$, GOOD/BAD generations, and F_c
    Generate refined impression $I_{\text{gen}} \leftarrow L(P_t)$
    Recompute average ROUGE similarity $q(I_{\text{gen}}, \mathcal{P}_m)$
    Update GOOD/BAD generations according to $q(I_{\text{gen}}, \mathcal{P}_m)$
end for
$I_{\text{final}} \leftarrow I_{\text{gen}}$
return I_{final}

The iterative process is repeated up to a fixed iteration limit T. The final impression I_{final} is the last generated output after refinement. Since the entire process is performed through prompt updates, the method remains lightweight and compatible with different open-source language models.

4.7. Text Only and No Fine-Tuning Design

A central design choice of this thesis is to avoid language model fine-tuning, proprietary APIs, multimodal encoders, and predefined disease-anatomy label pairs. This decision is motivated by the goal of building a framework that is transparent, reproducible, and practical for environments with limited computational resources. The proposed method therefore relies on retrieval, prompt construction, and prompt level feedback rather than updating model parameters.

The framework is also text-only. It uses the findings section as the input and generates the impression section as the output. No medical image features are used in the final pipeline. This does not imply that visual information is unimportant for radiology. Rather, the purpose of this thesis is to examine how far text only generation can be improved through retrieval-guided prompting. By focusing on the findings to impression setting,

the framework remains applicable even when only reports are available or when image processing pipelines are not accessible.

The projection-guided retrieval strategy estimates auxiliary retrieval components from the reference report collection, including IDF weights, carry-over weights, and a projection mapping from findings-side representations to impression-side representations. These components are used only for retrieval and do not modify the language model. Therefore, the generation model remains unchanged across experiments, and improvements are attributed to retrieval design, prompt construction, example ordering, and prompt level feedback.

This design makes the method modular. The embedding model, language model, lexical similarity metric, retrieval strategy, number of positive examples, number of hard negatives, and ordering strategy can each be changed independently. This modularity enables systematic analysis of how different components affect summarization performance, which is a central objective of the thesis.

4.8. Summary of Methodological Differences from Prior Work

The proposed method differs from prior radiology impression generation frameworks in several important ways. First, it evaluates two standalone retrieval strategies for selecting in context examples. The first strategy, semantic-lexical similarity matching, retrieves examples through direct semantic and lexical similarity between findings sections. The second strategy, projection-guided multi-signal matching, retrieves examples through a target-oriented projection signal combined with lexical, carry-over, and structured clinical matching signals. This allows the thesis to study both direct input-side retrieval and target-oriented multi-signal retrieval within the same prompt based generation framework.

Second, the framework avoids predefined disease-anatomy label pairs for retrieval. Instead, it operates over free-text report sections and uses structured clinical tokens only as auxiliary matching signals. Third, it does not use medical image features or multimodal encoders, making the framework text-only and easier to reproduce. Fourth, it relies on compact open-source decoder only models rather than proprietary LLM APIs. Fifth, it improves generation through retrieval design, prompt construction, hard negative examples, ordering strategies, and iterative prompt feedback rather than language model fine-tuning.

Although the second retrieval strategy estimates a lightweight projection mapping for retrieval, the language model itself is not fine tuned and its parameters are not updated.

These design choices make the framework suitable for studying how retrieval and prompting affect compact open-source language models in a specialized clinical summarization task. The following chapters describe the experimental setup and evaluate how each component influences radiology impression generation performance.

CHAPTER 5

EXPERIMENTAL SETUP

This chapter describes the experimental setting used to evaluate the proposed retrieval-guided impression generation framework. The experiments are designed to examine not only final summarization performance, but also the effect of retrieval strategy, embedding model choice, prompt composition, hard negative examples, example ordering, and iterative refinement. OpenI and MIMIC-CXR are used as the primary chest X-ray evaluation datasets, while an additional evaluation based on the RadSum23 MIMIC-III benchmark examines generalization across computed tomography and magnetic resonance imaging reports. Since reproducibility is one of the main motivations of the thesis, this chapter explains the dataset protocols, preprocessing procedures, model configurations, retrieval parameters, inference settings, and evaluation metrics used throughout the study.

5.1. Dataset Description

The experiments in this thesis use two primary chest X-ray report datasets, OpenI and MIMIC-CXR, together with an additional cross-modality evaluation setting derived from MIMIC-III. OpenI is used as the main public benchmark for comparison with prior findings-to-impression generation studies, while MIMIC-CXR is used to examine whether the retrieval-guided framework remains effective on a larger and more heterogeneous chest X-ray report collection. The RadSum23 MIMIC-III benchmark is used separately to evaluate whether the same text-only findings-to-impression formulation transfers to computed tomography and magnetic resonance imaging reports covering different anatomical regions.

The OpenI Chest X-ray Dataset is a publicly available collection provided by the National Library of Medicine (Demner-Fushman et al. 2016). The dataset contains 7,470 chest X-ray images associated with radiology reports collected from 3,955 clinical studies. Each report includes metadata such as patient age, gender, and imaging modality, together with textual sections such as comparison, indication, findings, and impression. The findings section provides detailed radiological observations, while the impression

section presents a concise diagnostic summary derived from those observations.

MIMIC-CXR is a large-scale chest radiography dataset collected in a real clinical environment (Johnson et al. 2019). The full database contains 227,835 radiographic studies and 377,110 chest radiograph images associated with free-text radiology reports. Compared with OpenI, MIMIC-CXR provides a larger and more heterogeneous reporting setting, with greater variation in report style, length, terminology, and clinical complexity. After section extraction and filtering, the MIMIC-CXR subset used in this thesis contains more than 114,000 usable findings-impression report pairs.

Table 5.1. Descriptive statistics of the processed dataset splits. Word counts are computed over the extracted findings and impression sections.

Dataset	**Split**	**Reports**	**Avg. Findings Words**	**Avg. Impression Words**	**Findings-to-Impression Length Ratio**
OpenI	Reference collection	2,843	31.83	8.22	3.87
OpenI	Evaluation set	576	31.49	8.16	3.86
MIMIC-CXR	Reference collection	114,689	48.84	13.39	3.65
MIMIC-CXR	Evaluation set	1,606	62.12	17.09	3.63

Table 5.1 summarizes the processed splits used in the experiments. The two datasets differ not only in scale, but also in report length and evaluation difficulty. In OpenI, the reference collection and evaluation set have very similar average report lengths. The average findings length is 31.83 words in the reference collection and 31.49 words in the evaluation set, while the average impression length is 8.22 and 8.16 words, respectively. This indicates that the OpenI evaluation set is close to the reference collection in terms of report length. In contrast, the MIMIC-CXR evaluation set is systematically longer than its reference collection. The average findings length increases from 48.84 words in the reference collection to 62.12 words in the evaluation set, and the average impression length increases from 13.39 to 17.09 words. This corresponds to an increase of approximately 27% in both findings and impression length. Therefore, MIMIC-CXR represents a more challenging evaluation setting in which the model is tested on longer and more complex

reports than those typically observed in the reference collection.

Table 5.2. Dataset difficulty indicators for OpenI and MIMIC-CXR. Exact match and frequency statistics are computed over evaluation impressions with respect to the corresponding reference collection.

Indicator	**OpenI**	**MIMIC-CXR**
Average nearest-neighbor findings similarity	0.915	0.879
Evaluation impressions found exactly in reference collection	63.0%	21.3%
Top-5 most frequent impressions in evaluation set	22.6%	10.2%
Unique evaluation impressions	50.9%	87.2%
Evaluation findings length shift relative to reference collection	-1.1%	+27.2%
Evaluation impression length shift relative to reference collection	-0.7%	+27.7%

The two datasets also differ in how easily evaluation reports can be matched to similar reports in the reference collection. Table 5.2 summarizes several retrieval and repetition indicators that help characterize this difference. In OpenI, evaluation impressions are more repetitive and template-like. A large portion of the OpenI evaluation impressions have exact matches in the reference collection, and the nearest-neighbor findings similarity is high. This makes OpenI a useful benchmark for reproducibility and comparison with prior work, but it also means that retrieval can often find highly similar examples. MIMIC-CXR is less repetitive and has a lower degree of direct impression reuse. Its evaluation impressions are more diverse, and its nearest-neighbor retrieval ceiling is lower. These properties make MIMIC-CXR a stricter setting for evaluating whether a retrieval-guided generation method can move beyond copying similar reference impressions.

5.1.1. MIMIC-III RadSum23 Evaluation Setting

An additional cross-modality evaluation is conducted using the text-based MIMIC-III component of the RadSum23 benchmark (Delbrouck et al. 2023). This benchmark extends radiology report summarization beyond chest X-rays by including reports from different imaging modalities and anatomical regions. The dataset is derived from the `NOTEEVENTS` table of MIMIC-III and follows the official RadSum23 data construction

and evaluation setting. The reconstructed evaluation data were verified against the official RadSum23 files using the provided checksums.

The additional experiment uses six in-distribution computed tomography and magnetic resonance imaging domains. The same findings-to-impression task formulation is retained: the findings section is provided as input, and the system generates the corresponding impression section. Table 5.3 presents the evaluation domains and report counts. In total, the cross-modality evaluation contains 7,420 reports.

Table 5.3. RadSum23 MIMIC-III domains used in the cross-modality evaluation.

Modality	**Anatomical Region**	**Evaluation Reports**
CT	Head	3,141
CT	Spine	553
CT	Chest	1,280
MR	Head	732
CT	Neck	115
CT	Abdomen/Pelvis	1,599
	Total	**7,420**

Unlike the primary OpenI and MIMIC-CXR experiments, this setting is not used for method development or the main retrieval ablations. It is used as an additional generalization experiment to determine whether the proposed retrieval-guided generation strategy remains applicable when report structure, anatomical coverage, and imaging modality differ from chest X-ray reporting.

The three datasets play complementary roles in the thesis, as summarized in Table 5.4. OpenI enables comparison with prior radiology summarization studies that report findings-to-impression generation results on this dataset. MIMIC-CXR provides a larger chest X-ray evaluation setting and is used to analyze robustness, clinical label agreement, and comparison with recent radiology summarization systems. The RadSum23 MIMIC-III benchmark provides an additional evaluation of cross-modality and cross-anatomy generalization.

These differences are important for interpreting the experimental results. OpenI provides a relatively compact and reproducible benchmark where retrieval can often

Table 5.4. Summary of the datasets used in this thesis.

Dataset	Main Role	Reason for Inclusion
OpenI	Public benchmark	Enables comparison with prior findings-to-impression generation studies and supports reproducible evaluation.
MIMIC-CXR	Larger chest X-ray evaluation setting	Provides a more heterogeneous report collection for evaluating robustness and clinical agreement.
MIMIC-III RadSum23	Cross-modality evaluation setting	Examines generalization across six CT and MR domains covering different anatomical regions.

identify highly similar reference examples. MIMIC-CXR provides a more difficult chest X-ray evaluation setting because reports are longer, impressions are more diverse, and the evaluation set differs more strongly from the reference collection. MIMIC-III introduces a different source of difficulty by extending the task to CT and MR reports from multiple anatomical domains. Therefore, OpenI and MIMIC-CXR are used for the primary method development and comparison experiments, while MIMIC-III is used as an additional evaluation of whether the resulting framework transfers beyond chest X-ray reporting.

5.2. Preprocessing Protocol

Data preprocessing for OpenI follows the standardized protocol used in prior radiology findings summarization studies. Following Zhang et al. (Zhang et al. 2018) and Gharebagh et al. (Gharebagh, Goharian, and Filice 2020), reports were excluded if they lacked either a *Findings* or an *Impression* section, if the findings section contained fewer than 10 words, or if the impression section contained fewer than 2 words. After this filtering procedure, 2,843 cases remained in the reference collection and 576 cases were retained for evaluation.

For MIMIC-CXR, the same task formulation is used: the findings section is treated as the input and the impression section is treated as the target output. Reports without usable findings or impression sections are excluded. After section extraction and filtering, more than 114,000 usable findings-impression report pairs remain for the MIMIC-CXR experiments. The official dataset separation is preserved so that evaluation reports are not

included in the reference collection used for retrieval. The MIMIC-CXR evaluation set used in the experiments contains 1,606 reports.

The OpenI preprocessing protocol is compatible with the methodology adopted by later studies, including the Hu et al. series and ImpressionGPT (Hu et al. 2021; 2022; 2023; Ma et al. 2024). This supports comparison with previous findings-to-impression generation studies under a consistent dataset setting.

For OpenI and MIMIC-CXR, the corresponding reference collection is used for example retrieval. For each evaluation case, the system retrieves reference findings-impression pairs and constructs a prompt using the selected examples. The evaluation split is used only for generation and scoring, and it is not included in the retrieval collection during generation. This separation ensures that generated impressions are evaluated on unseen reports.

The MIMIC-III cross-modality experiment follows the official text-based RadSum23 evaluation setting (Delbrouck et al. 2023). The required reports were reconstructed from the MIMIC-III `NOTEEVENTS` table according to the official benchmark organization. The reconstructed files were verified against the official RadSum23 data using the provided checksums. This verification ensures that the evaluation reports correspond exactly to the reports defined by the shared-task protocol.

The additional evaluation includes six in-distribution modality-anatomy domains: CT Head, CT Spine, CT Chest, MR Head, CT Neck, and CT Abdomen/Pelvis. The official domain-specific separation is preserved for each domain. Reference examples are retrieved from the corresponding reference partition, while the official test reports are used only for generation and evaluation. Across the six domains, the MIMIC-III evaluation contains 7,420 reports.

The same findings-to-impression formulation is maintained across all three datasets. However, the MIMIC-III setting is used only as an additional cross-modality generalization experiment. It is not used for selecting the main retrieval configuration or for conducting the primary ablation studies reported on OpenI and MIMIC-CXR.

Table 5.5. Preprocessed dataset settings used in the experiments.

Dataset	Reference Collection	Evaluation Cases	Use in Experiments
OpenI	2,843 reports	576	Main public benchmark for findings-to-impression generation and primary ablation analysis.
MIMIC-CXR	More than 114,000 usable pairs	1,606	Larger chest X-ray evaluation setting used to examine robustness and clinical agreement.
MIMIC-III RadSum23	Official domain-specific reference partitions	7,420	Additional evaluation across six CT and MR modality-anatomy domains.

5.3. Evaluated Language Models

The generation experiments focus on compact open-source language models. The main evaluated models are Llama3 (Grattafiori et al. 2024), LLaVA (Liu et al. 2023), and Meditron (Chen et al. 2023). These models represent different model backgrounds and allow the thesis to evaluate whether retrieval-guided prompting behaves similarly across general purpose, multimodal instruction tuned, and medically adapted open-source models.

Llama3 is used as a general purpose decoder only model. LLaVA is included to examine whether a model developed in a multimodal instruction following setting behaves differently when used only with text prompts. Meditron is included because it is adapted to biomedical and clinical language. DeepSeek (Guo et al. 2025) is also evaluated in one shot experiments to examine broader applicability, but it is not included in the full few-shot analysis because of its considerably higher inference latency.

5.4. Embedding Models

The retrieval stage uses sentence embeddings to represent findings sections in a dense semantic space. Three embedding models are evaluated: MiniLM-L6 (Wang et al. 2020), PubMedBERT (Gu et al. 2021), and MedEmbed (Balachandran 2024). MiniLM-L6 is a lightweight general purpose model, PubMedBERT provides biomedical domain adap-

Table 5.6. Overview of the language models evaluated in the thesis.

Model	Role in Experiments	Reason for Inclusion
Llama3	Main generator	General purpose compact decoder only model used to evaluate open-source prompting behavior.
LLaVA	Main generator	Included to examine whether a model developed in a multimodal instruction following setting behaves differently under text only prompting.
Meditron	Main generator	Included as a medically adapted open-source model for comparison with general purpose models.
DeepSeek	One shot analysis	Included for additional comparison but excluded from later few-shot experiments because of high inference latency.

tation, and MedEmbed is designed for medical similarity and retrieval settings. Comparing these embedding models allows the thesis to analyze whether domain specific embedding models are necessary, or whether a lightweight general purpose representation can be effective when combined with lexical refinement.

Table 5.7. Embedding models used for semantic retrieval.

Embedding Model	Domain	Use in Framework
MiniLM-L6	General purpose	Efficient semantic retrieval and main embedding model in several ablation experiments.
PubMedBERT	Biomedical	Biomedical domain representation for clinical and biomedical language.
MedEmbed	Medical and clinical	Medical similarity representation for fine grained clinical retrieval.

5.5. Retrieval and Prompting Parameters

The experiments vary several retrieval and prompting parameters. For semantic-lexical similarity matching, the semantic retrieval pool size is denoted as Top-M. After this initial semantic filtering step, lexical similarity is applied to select positive examples and hard negative examples. For projection-guided multi-signal matching, reference examples are ranked using a weighted combination of projection-based, lexical, carry-over, and structured clinical signals. The number of positive examples, number of hard negatives, lexical similarity metric, ordering strategy, retrieval strategy, and iterative subset size used during prompt refinement are treated as experimental factors.

The purpose of varying these parameters is to understand how each component affects generation quality. For example, changing Top-M examines whether the initial semantic candidate pool should be narrow or broad. Varying the number of positive examples tests how much retrieved context is beneficial before the prompt becomes too long or distracting. Varying the number of hard negatives tests whether contrastive guidance helps the model avoid misleading associations. Comparing ordering strategies evaluates whether model behavior is affected by the placement of more relevant examples inside the prompt.

Table 5.8. Main retrieval and prompting parameters evaluated in the experiments.

Parameter	**Description**
Top-M	Number of semantically retrieved candidates before lexical refinement.
Positive examples	Retrieved findings and impression pairs used as demonstrations of correct mappings.
Hard negatives	Semantically related but lexically divergent examples used to expose misleading mappings.
Lexical metric	ROUGE based score used for positive or hard negative selection.
Ordering strategy	Arrangement of few-shot examples in normal or reverse order.
Iterative subset size	Number of top-ranked retrieved examples used during iterative refinement and feedback computation.
Retrieval strategy	Choice between semantic-lexical similarity matching and projection-guided multi-signal matching.

5.6. Experimental Conditions

The experiments are organized to isolate the effect of different components of the framework. Zero shot prompting is used to evaluate the behavior of each language model without examples. Random shot prompting is used to test whether adding examples without semantic selection is sufficient. Retrieval-guided one shot and few-shot prompting are then used to evaluate the effect of semantically and lexically selected examples.

Additional experiments examine the effect of embedding model choice, semantic pool size, positive example count, hard negative count, lexical metric selection, retrieval strategy, example ordering, and iterative refinement. This design allows the thesis to evaluate the proposed framework not only as a final system, but also as a collection of design choices that interact with model behavior.

Table 5.9. Main experimental conditions considered in the thesis.

Condition	**Purpose**
Zero shot	Evaluate generation without examples.
Random shot	Evaluate whether randomly selected examples improve generation.
Retrieval-guided one shot	Measure the effect of a single retrieved example.
Retrieval-guided few-shot	Measure the effect of multiple retrieved positive examples.
Hard negative prompting	Test whether contrastive examples improve clinical discrimination.
Ordering analysis	Evaluate sensitivity to normal and reverse ordering.
Iterative refinement	Assess whether ROUGE based feedback improves generation quality.
Semantic-lexical retrieval	Evaluate direct retrieval based on semantic similarity followed by lexical refinement.
Projection-guided retrieval	Evaluate target-oriented multi-signal retrieval using projection, lexical, carry-over, and structured clinical signals.

5.7. Inference Environment and Computational Considerations

The framework is designed to avoid language model fine-tuning. No parameter updates are applied to the language models or embedding models during the experiments. The projection-guided retrieval strategy estimates a lightweight mapping for retrieval, but the generator itself remains unchanged. All improvements are obtained through retrieval, prompt construction, hard negative selection, example ordering, and prompt level iterative refinement. This design reduces computational requirements and allows the method to be applied to multiple open-source models without additional training.

Computational cost is evaluated through inference time measurements. Since clinical summarization systems must be practical as well as accurate, the runtime of each model is considered alongside ROUGE performance. DeepSeek, for example, is included in one shot evaluation but excluded from later few-shot experiments because it increases inference time substantially without outperforming the best compact medical model.

Because inference time can be affected by hardware, implementation details, prompt length, and decoding configuration, runtime comparisons are interpreted as practical indicators rather than universal model speed estimates. The main purpose of the computational analysis is to show whether the proposed prompting strategy remains feasible for compact open-source models under the experimental setting used in this thesis.

5.8. Evaluation Metrics

The thesis evaluates generated impressions using lexical, semantic, clinical label-based, and structured clinical metrics. ROUGE is used as the primary lexical evaluation metric because it is widely reported in prior radiology summarization studies and enables direct comparison with earlier work (Lin 2004). The thesis reports ROUGE-1, ROUGE-2, and ROUGE-L F1-scores. ROUGE-1 measures unigram overlap, ROUGE-2 measures bigram overlap, and ROUGE-L measures similarity based on the longest common subsequence. The mathematical definitions of ROUGE-N and ROUGE-L are provided in Chapter 2; this chapter uses them only as evaluation measures.

In addition to ROUGE, BERTScore is used as a semantic similarity metric. BERTScore compares generated and reference impressions using contextual token repre-

sentations rather than exact surface overlap (Zhang et al. 2019). This is useful for radiology impression generation because clinically similar statements may be expressed with different wording. When comparing with prior work, the reporting convention of BERTScore is kept consistent with the comparison setting, since raw and rescaled BERTScore values are not directly interchangeable.

Clinical label agreement for chest X-ray reports is evaluated using CheXbert-based F1. CheXbert maps generated and reference impressions into a set of chest radiology observation labels, and the generated labels are compared with the reference labels (Smit et al. 2020). This metric provides a complementary view of clinical agreement because it evaluates whether clinically relevant observations are preserved, omitted, or added. In this thesis, CheXbert-F1 is reported for OpenI and MIMIC-CXR. Since CheXbert is designed for chest X-ray observations, it is not used to interpret the CT and MR results in the MIMIC-III evaluation.

Structured clinical agreement is additionally evaluated using RadGraph-F1 (Jain et al. 2021). RadGraph extracts clinical entities and relations from generated and reference reports and compares the resulting structured representations. This metric is particularly important for the MIMIC-III cross-modality evaluation, where chest X-ray specific labelers such as CheXbert are not appropriate. Simple, partial, and complete RadGraph-F1 values are reported according to the matching configuration used in the corresponding experiment.

Different prior studies use slightly different names for clinical evaluation, including factual consistency, clinical efficacy, CheXpert Score, CheXbert-F1, and RadGraph-F1. These names do not necessarily refer to identical evaluation procedures. Therefore, comparisons involving clinical metrics are interpreted carefully, with attention to the labeler or graph extractor, label subset, matching configuration, and averaging strategy when this information is available.

Although these automatic metrics provide useful quantitative evidence, none of them fully captures clinical correctness. A generated impression can obtain high lexical overlap while still containing an unsupported statement, and a clinically reasonable paraphrase can receive a lower ROUGE score. For this reason, the results chapter also discusses qualitative examples and failure modes where appropriate.

5.9. Reproducibility Notes

All experiments use the dataset-specific preprocessing and separation protocols described above. For OpenI, the standardized filtering protocol used in prior findings-to-impression studies is followed. For MIMIC-CXR, reports without usable findings or impression sections are excluded, and the evaluation set is kept separate from the reference collection used for retrieval. For the MIMIC-III cross-modality evaluation, the official RadSum23 report organization is reconstructed from the MIMIC-III `NOTEEVENTS` table, and the resulting evaluation files are verified against the provided checksums.

The retrieval collection, evaluation set, metric implementation, and prompting settings are kept consistent when comparing configurations within the same experiment. The main retrieval parameters, example counts, hard negative settings, ordering strategies, and iterative feedback settings are reported together with the corresponding experiments. Prompt templates are provided in Appendix A. The modular structure of the framework allows the embedding model, lexical metric, retrieval strategy, number of positive examples, number of hard negatives, ordering strategy, and language model to be varied independently.

The OpenI and MIMIC-CXR experiments are used for the primary configuration analysis and comparison with earlier findings-to-impression systems. The MIMIC-III RadSum23 experiment is maintained as a separate cross-modality evaluation and is not used to select the main retrieval configuration. This distinction prevents the additional CT and MR evaluation from influencing the configurations established on the primary chest X-ray datasets.

The framework does not require language model fine-tuning, proprietary inference APIs, or external visual encoders. The projection-guided strategy estimates a lightweight mapping for retrieval and may use automatically extracted structured clinical features as auxiliary ranking signals, but the generator parameters remain unchanged. This makes the experimental design reproducible in settings where the same dataset partitions, retrieval procedures, prompt templates, metric implementations, and open-source models are available.

CHAPTER 6

RESULTS

6.1. Zero-shot and Random-shot Baselines

We first evaluate each model in a zero-shot setting on the OpenI test set consisting of 576 reports, where the LLM receives only a system instruction and the target Findings section. This setting reflects the model's intrinsic ability to summarize clinical text without any external guidance. As shown in Table 6.1, zero-shot performance is extremely low across all models, indicating that unguided generation is insufficient for producing impressions that closely match the reference summaries.

To further assess the sensitivity of decoder-only LLMs to demonstration examples, we introduce a *random-shot baseline*, in which a single randomly selected findings and impression pair is provided in the prompt. This experiment does not provide target-specific clinical guidance, but tests whether the presence of an example, regardless of relevance, improves model behavior.

Interestingly, all models exhibit substantial performance gains when conditioned on a random example, despite the example not being selected according to clinical or semantic relevance. This indicates that decoder-only models strongly rely on prompt structure and few-shot patterns, even when the content is not aligned with the target findings. However, random-shot performance remains far below the retrieval-guided one-shot and few-shot configurations presented later. These results emphasize that models benefit from example-driven prompting, but that semantically and lexically relevant retrieval is required for high-quality summarization.

Zero-shot performance demonstrates that none of the evaluated models produces outputs with strong lexical agreement to the reference impressions without external guidance. Random-shot conditioning provides a substantial improvement, confirming the structural dependence of decoder-only models on few-shot exemplars. However, because these examples are semantically unrelated to the target case, the gains remain far below retrieval-guided results. This indicates that both semantic relevance and lexical alignment are critical for high-quality impression generation. Motivated by these limitations, we

Table 6.1. ROUGE performance of all models under zero-shot and random-shot prompting conditions.

Model	Zero-shot			Random-shot		
	R-1	R-2	R-L	R-1	R-2	R-L
Llama3	0.1449	0.0457	0.1227	0.3205	0.1601	0.3026
LLaVA	0.1308	0.0385	0.1119	0.1859	0.0678	0.1644
DeepSeek	0.1620	0.0545	0.1387	0.3123	0.1540	0.2924
Meditron	0.0829	0.0237	0.0684	0.2264	0.0890	0.2060

next evaluate retrieval-guided one-shot prompting, where each model is conditioned on a single semantically selected reference example.

6.2. One-shot Performance with Similarity Filtering

Building on the zero-shot and random-shot baselines, we next analyze how retrieval-guided one-shot prompting improves summarization performance by changing how conditioning examples are chosen. This approach uses a similarity-based retrieval mechanism to identify clinically relevant reference reports that are semantically aligned with the target findings, avoiding reliance on random or predetermined selections. Specifically, cosine similarity is computed between the sentence embeddings of the target findings and those of candidate reference findings, and the most similar samples are selected to condition impression generation.

6.2.1. Non-iterative One-shot Results

Tables 6.2 and 6.3 report one-shot results obtained without iterative prompt refinement. In this setting, we compare two Phase 1 retrieval strategies: selecting reference examples solely based on sentence embedding similarity, and using a hybrid strategy that applies ROUGE-1 based lexical filtering after an initial semantic retrieval step.

The results show that retrieval-guided one-shot prompting substantially improves over the zero-shot and random-shot baselines. Even without iterative refinement, all retrieval-guided configurations produce much stronger ROUGE scores. The comparison

Table 6.2. ROUGE scores for one-shot prompting without iterative refinement, using an embedding-only reference retrieval strategy where both initial filtering and final selection rely on sentence embeddings.

Model (Embedding)	**R-1**	**R-2**	**R-L**
Llama3 (MiniLML6)	0.4520	0.3411	0.4348
Llama3 (PubMedBert)	0.4327	0.3265	0.4179
Llama3 (MedEmbed)	0.4434	0.3300	0.4268
LLaVA (MiniLML6)	0.4072	0.2906	0.3884
LLaVA (PubMedBert)	0.4011	0.2880	0.3818
LLaVA (MedEmbed)	0.4304	0.3084	0.4095
DeepSeek (MiniLML6)	0.5176	0.4078	0.4993
DeepSeek (PubMedBert)	0.5055	0.3976	0.4900
DeepSeek (MedEmbed)	0.5000	0.3844	0.4800
Meditron (MiniLML6)	0.4847	0.3892	0.4648
Meditron (PubMedBert)	0.5049	0.4098	0.4896
Meditron (MedEmbed)	0.4879	0.3926	0.4700

between embedding-only retrieval and the hybrid strategy shows that lexical refinement is especially beneficial for Meditron, where MiniLM-based hybrid retrieval improves ROUGE-1 from 0.4847 to 0.5095. For some models, embedding-only retrieval remains competitive, indicating that semantic similarity alone can retrieve useful examples. However, the hybrid strategy provides a more controlled mechanism for preserving clinically important terminology.

For reference, ImpressionGPT (Ma et al. 2024) reports ROUGE-1, ROUGE-2, and ROUGE-L scores of 0.4207, 0.2702, and 0.3978 in its non-iterative setting. Compared to these results, the retrieval-guided one-shot configurations in this thesis achieve substantially higher performance even without iterative refinement. It is also important to note that ImpressionGPT relies on up to fifteen conditioning examples in this setting, whereas the one-shot results reported here are obtained using a single retrieved example. This highlights the importance of precise semantic and lexical retrieval.

Although semantic embeddings are effective at capturing conceptual similarity, they can sometimes miss clinically critical terms that differ subtly in wording. This limitation is especially important in medical summarization, where diagnostic accuracy depends heavily on precise terminology. The ROUGE-1 refinement step helps address this issue by enforcing word-level overlap between candidate and reference findings, thereby

Table 6.3. ROUGE scores for one-shot prompting without iterative refinement, using a hybrid reference retrieval strategy with initial semantic filtering via sentence embeddings followed by ROUGE-1 based lexical selection.

Model (Embedding)	**R-1**	**R-2**	**R-L**
Llama3 (MiniLML6)	0.4329	0.3270	0.4155
Llama3 (PubMedBert)	0.4479	0.3393	0.4329
Llama3 (MedEmbed)	0.4350	0.3275	0.4195
LLaVA (MiniLML6)	0.4162	0.2990	0.3948
LLaVA (PubMedBert)	0.4109	0.2990	0.3897
LLaVA (MedEmbed)	0.4067	0.2917	0.3889
DeepSeek (MiniLML6)	0.5172	0.4011	0.4943
DeepSeek (PubMedBert)	0.5087	0.3943	0.4896
DeepSeek (MedEmbed)	0.5054	0.3932	0.4836
Meditron (MiniLML6)	0.5095	0.4214	0.4946
Meditron (PubMedBert)	0.5031	0.4181	0.4881
Meditron (MedEmbed)	0.5058	0.4202	0.4897

encouraging the preservation of key clinical expressions.

6.2.2. Iterative One-shot Results

Tables 6.4 and 6.5 present the corresponding one-shot results when iterative prompt refinement is enabled. This setting evaluates whether similarity-based feedback can further improve generation after the initial retrieved example has been selected.

Iterative refinement produces a substantial improvement over the corresponding non-iterative results. For example, Meditron with MiniLM hybrid retrieval improves from 0.5095 to 0.6662 in ROUGE-1. Llama3 and LLaVA also show clear gains after refinement. These results indicate that similarity-based feedback contributes more than simple prompt formatting; it provides an additional signal that guides the model toward outputs that more closely match the retrieved positive impressions.

The comparison between embedding-only retrieval and hybrid retrieval also becomes clearer under iterative refinement. The hybrid strategy generally provides the strongest results, especially for Meditron, which reaches the highest one-shot ROUGE-1 score of 0.6662 with MiniLM. By first performing broad semantic retrieval and then enforcing lexical consistency through ROUGE-1 based filtering, the hybrid approach supports

Table 6.4. ROUGE scores for one-shot prompting with iterative refinement, using an embedding-only reference retrieval strategy where both initial filtering and final selection rely on sentence embeddings.

Model (Embedding)	**R-1**	**R-2**	**R-L**
Llama3 (MiniLML6)	0.5863	0.4745	0.5704
Llama3 (PubMedBert)	0.5936	0.4722	0.5796
Llama3 (MedEmbed)	0.5920	0.4674	0.5777
LLaVA (MiniLML6)	0.5927	0.4774	0.5730
LLaVA (PubMedBert)	0.5916	0.4769	0.5728
LLaVA (MedEmbed)	0.5835	0.4689	0.5660
DeepSeek (MiniLML6)	0.6182	0.5074	0.5977
DeepSeek (PubMedBert)	0.6063	0.4926	0.5879
DeepSeek (MedEmbed)	0.5970	0.4835	0.5773
Meditron (MiniLML6)	0.6387	0.5503	0.6273
Meditron (PubMedBert)	0.6268	0.5286	0.6139
Meditron (MedEmbed)	0.6301	0.5358	0.6183

both semantic relevance and terminology preservation.

Sentence embeddings are well suited for capturing semantic equivalence and accommodating paraphrasing, which improves recall by enabling the retrieval of clinically relevant examples even when surface-level wording varies (Ng, and Abrecht 2015; Kuzi et al. 2020). However, this semantic flexibility may occasionally reduce attention to specific diagnostic terms. The ROUGE-1 refinement step helps mitigate this limitation by placing greater emphasis on lexical overlap and clinically important wording. Earlier studies show that human evaluators often prefer summaries that cover key clinical concepts and terminology in a clear and consistent manner (Ng, and Abrecht 2015; Takeshita, Ponzetto, and Eckert 2024), and the observed improvements are consistent with this preference.

Taken together, these results support the adoption of a hybrid retrieval strategy in which broad semantic filtering is complemented by lexical refinement for retrieval-guided clinical summarization (Kuzi et al. 2020). The gains achieved through iterative refinement suggest that lexical feedback contributes more than straightforward error correction, instead guiding the model toward generating impressions that are both more stable and more consistent with the retrieved clinical examples.

Table 6.5. ROUGE scores for one-shot prompting with iterative refinement, using a hybrid reference retrieval strategy with initial semantic filtering via sentence embeddings followed by ROUGE-1 based lexical selection.

Model (Embedding)	**R-1**	**R-2**	**R-L**
Llama3 (MiniLML6)	0.5997	0.4773	0.5818
Llama3 (PubMedBert)	0.6026	0.4826	0.5818
Llama3 (MedEmbed)	0.5959	0.4751	0.5782
LLaVA (MiniLML6)	0.6042	0.4939	0.5861
LLaVA (PubMedBert)	0.6055	0.4947	0.5867
LLaVA (MedEmbed)	0.5985	0.4836	0.5793
DeepSeek (MiniLML6)	0.6141	0.5200	0.5972
DeepSeek (PubMedBert)	0.6113	0.5176	0.5928
DeepSeek (MedEmbed)	0.6022	0.5073	0.5861
Meditron (MiniLML6)	0.6662	0.5823	0.6533
Meditron (PubMedBert)	0.6647	0.5826	0.6545
Meditron (MedEmbed)	0.6618	0.5799	0.6501

6.2.3. Efficiency Considerations in One-shot Experiments

To assess the broader applicability of the proposed retrieval and generation framework, we extended the one-shot experiments to include DeepSeek-R1, a large-scale open model known for strong general-domain reasoning and text generation capabilities (Guo et al. 2025). DeepSeek was evaluated under the same embedding-based and hybrid retrieval configurations as the other models, and its results are incorporated into Tables 6.4 and 6.5.

Although DeepSeek achieved reasonable one-shot performance, it did not surpass Meditron and required considerably higher inference time. Due to its substantially slower generation speed, DeepSeek was not included in the subsequent few-shot analyses. This decision keeps the later ablation experiments focused on models that provide a more practical balance between efficiency and summarization performance.

The timing results in Table 6.6 show that DeepSeek requires substantially longer inference time than Llama3, LLaVA, and Meditron. The remaining compact models therefore provide a more favorable balance between accuracy and efficiency for the repeated retrieval and prompting experiments required in this thesis. Llama3, LLaVA, and Meditron

Table 6.6. Query times for the models using a hybrid reference retrieval approach with sentence embeddings and ROUGE scores.

Model	**Single Query Time**	**Time for One Finding**
DeepSeek	7.9826 sec	366.2924 sec
Llama3	1.2501 sec	63.9205 sec
LLaVA	1.2775 sec	57.8840 sec
Meditron	1.5011 sec	69.5777 sec

can be evaluated across multiple retrieval and prompting settings while keeping runtime manageable. This observation supports the central objective of the thesis: improving clinical summarization through retrieval and prompt design rather than relying only on larger or slower models.

Because inference time can be affected by hardware, implementation details, prompt length, and decoding configuration, runtime comparisons are interpreted as practical indicators rather than universal model speed estimates. The main purpose of the computational analysis is to show whether the proposed prompting strategy remains feasible for compact open-source models under the experimental setting used in this thesis.

6.3. Embedding-Guided Few-Shot Selection

After evaluating one-shot prompting, we next examine whether increasing the number of retrieved demonstrations further improves impression generation performance. In this setting, the models are conditioned on multiple examples selected through the hybrid retrieval strategy. This experiment tests whether compact open-source models benefit from additional retrieved context, and whether the ordering of examples affects the final generated impression.

To analyze prompt ordering, the top fifteen retrieved examples are arranged in two ways. In the normal order, examples are placed from most similar to least similar. In the reverse order, examples are placed from least similar to most similar, so that the most relevant examples appear closer to the target findings. Performance is evaluated using different numbers of examples, ranging from 3 to 15. Figures 6.1, 6.2, and 6.3 show that ordering has a clear effect on model performance.

LLaVA and Meditron achieve their best results with reverse ordering, whereas Llama3 performs best when examples are presented in normal order. This suggests that different decoder-only models use prompt context differently. For LLaVA and Meditron, placing the most relevant examples closer to the candidate findings may increase their influence during generation. In contrast, Llama3 appears to benefit from observing the strongest examples at the beginning of the demonstration sequence. These findings support the view that few-shot performance depends not only on which examples are selected, but also on how they are arranged in the prompt.

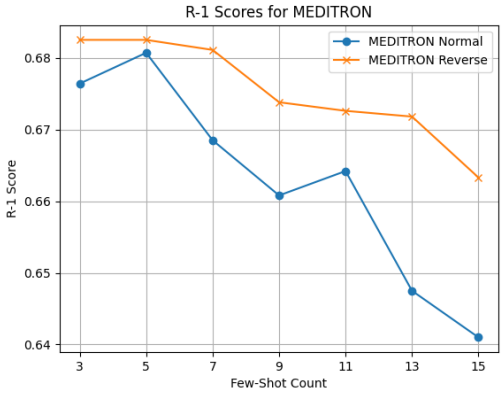

Figure 6.1. Few-shot performance from 3 to 15 examples using normal and reverse ordering for Meditron.

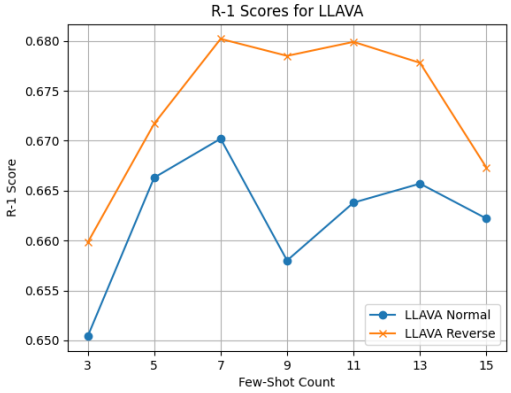

Figure 6.2. Few-shot performance from 3 to 15 examples using normal and reverse ordering for LLaVA.

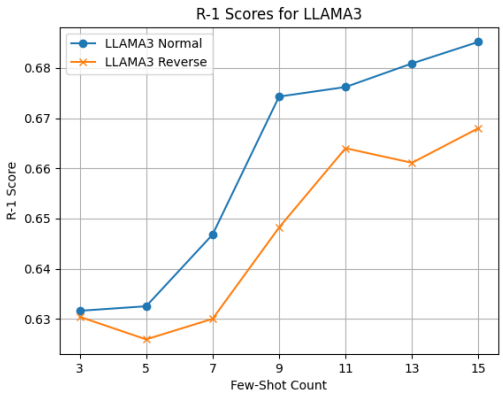

Figure 6.3. Few-shot performance from 3 to 15 examples using normal and reverse ordering for Llama3.

These ordering effects are consistent with prior work showing that few-shot prompting can be sensitive to demonstration order (Lu et al. 2022; Zhao et al. 2021). The difference between Llama3 and the models derived from the Llama2 line may be related to differences in model training, tokenizer design, and context handling, although a controlled architectural analysis is beyond the scope of this thesis (Grattafiori et al. 2024).

Table 6.7. Best few-shot ROUGE scores for the models using a hybrid reference retrieval approach.

Model and Configuration	**R-1**	**R-2**	**R-L**
Llama3, normal order, 15 shots	**0.6852**	0.5901	0.6732
LLaVA, reverse order, 7 shots	0.6802	0.5832	0.6633
Meditron, reverse order, 5 shots	0.6825	**0.6084**	**0.6745**

Table 6.7 summarizes the best few-shot configurations for all models using hybrid reference retrieval. Llama3 obtains the highest ROUGE-1 score with 0.6852 using normal ordering and 15 examples. Meditron obtains the highest ROUGE-2 and ROUGE-L scores with reverse ordering and 5 examples, indicating strong retention of phrasal content and sequence-level structure. LLaVA also reaches a comparable performance range with reverse ordering and 7 examples.

These results demonstrate that compact open-source models can produce high-

quality impressions when combined with carefully selected few-shot examples. The findings also confirm that prompt configuration is a critical factor in retrieval-guided summarization. More examples do not automatically guarantee better performance; rather, the number of demonstrations and their ordering must be tuned jointly with the model.

6.4. Impact of Hard Negatives and Semantic-Lexical Design Choices

We further analyze the effect of incorporating hard negative examples into the retrieval and generation framework. Unlike the earlier experiments, which relied only on positive demonstrations, this setting augments the prompt with semantically similar but lexically divergent examples in order to expose the model to misleading clinical patterns. To isolate these effects, all experiments in this section use Llama3 for generation and MiniLM embeddings for retrieval.

The effect of prompt composition was first examined by varying the number of positive and hard negative examples. Table 6.8 shows that adding a small number of hard negatives consistently improves performance, with the strongest overall results achieved using two negative examples. Beyond this point, adding further hard negatives does not provide additional gains and slightly reduces performance, suggesting that excessive negative context may weaken the influence of highly relevant positive demonstrations.

The number of positive examples in the prompt also affects generation quality. Increasing the number of positive examples improves performance initially, but very long prompts eventually become less effective. Configurations with more than 15 positive examples tend to perform worse, indicating that overly long prompts may dilute the influence of the most relevant demonstrations.

After introducing hard negatives, we next examine how different lexical similarity metrics affect positive and negative example selection. Table 6.9 shows that different metrics serve different roles within the retrieval pipeline. ROUGE-1 yields the strongest positive examples, likely because unigram overlap helps preserve clinically important terminology. In contrast, ROUGE-L performs well for hard negative selection because it captures sequence-level divergence between reports that are semantically similar. The best overall setup uses ROUGE-1 for positive selection and ROUGE-L for hard negative selection. This configuration achieves the best ROUGE-1 and ROUGE-L scores, while

Table 6.8. Effect of positive and hard negative example counts using ROUGE-1 positive selection and ROUGE-L hard negative selection with MiniLM embeddings.

Pos	Neg	R-1	R-2	R-L
15	1	0.6970	0.6123	0.6875
15	2	**0.6973**	**0.6148**	**0.6889**
15	3	0.6955	0.6113	0.6854
15	4	0.6959	0.6118	0.6852
15	5	0.6971	0.6111	0.6867
17	2	0.6903	0.6025	0.6802
19	2	0.6905	0.6001	0.6800
21	2	0.6882	0.5979	0.6782

ROUGE-L positive selection produces only a negligibly higher ROUGE-2 score.

Table 6.9. Effect of lexical similarity metrics on positive and hard negative selection.

Pos	Neg	R-1	R-2	R-L
R1	R1	0.6920	0.6087	0.6834
R1	R2	0.6922	0.6082	0.6819
R1	RL	**0.6973**	0.6148	**0.6889**
RL	R1	0.6904	0.6073	0.6812
RL	R2	0.6943	0.6106	0.6855
RL	RL	0.6952	**0.6150**	0.6858
R2	R1	0.6901	0.6071	0.6802
R2	R2	0.6868	0.6040	0.6776
R2	RL	0.6880	0.6026	0.6785

Table 6.10 evaluates the size of the initial semantic filtering pool used before lexical refinement. Increasing the pool of candidates from 150 to 200 and 250 improves retrieval quality, suggesting that a broader semantic candidate pool helps the system identify better examples before lexical refinement. The setting with 250 candidates achieves the best ROUGE-1 score, whereas the 200-candidate setting achieves the best ROUGE-2 and ROUGE-L scores. Conversely, performance decreases when the candidate pool is expanded to 300 reports, showing that overly large retrieval pools may introduce less relevant examples and reduce the precision of the final selected context. In summary, these results indicate that moderate retrieval pool sizes offer a good balance between semantic

diversity and candidate relevance.

Table 6.10. Effect of semantic filtering pool size (Top-M) using ROUGE-1 positive selection and ROUGE-L hard negative selection with MiniLM embeddings (15 positives, 2 negatives).

Top-M	R-1	R-2	R-L
150	0.6920	0.6062	0.6813
200	0.6973	**0.6148**	**0.6889**
250	**0.6984**	0.6131	0.6859
300	0.6916	0.6056	0.6814

Finally, we analyze the effect of the number of comparison samples used during iterative refinement. In this experiment, the retrieval stage preserves the full set of 15 retrieved positive examples, while the refinement feedback score is computed using only the top-ranked subset of these examples. As shown in Table 6.11, using five comparison samples produces the strongest overall performance. Using all 15 retrieved positives is not optimal, even though the same 15-example positive pool is available. This suggests that lower-ranked positives may introduce weaker or less specific references into the feedback signal. The top-five setting provides a more focused comparison set while still avoiding reliance on a single retrieved example.

Table 6.11. Effect of the number of comparison samples used during iterative refinement. All configurations retrieve 15 positive examples, while the Samples column indicates how many of the highest-ranked positives are used to compute the refinement feedback score. All configurations use ROUGE-1 positive selection, ROUGE-L hard negative selection, 2 hard negatives, Top-M=200, and MiniLM embeddings.

Samples	R-1	R-2	R-L
1	0.6923	0.6074	0.6827
3	0.6947	0.6119	0.6848
5	**0.7002**	**0.6174**	**0.6906**
7	0.6947	0.6132	0.6858
15	0.6973	0.6148	0.6889

These findings show that retrieval quality, example selection strategy, and prompt structure are collectively important for clinical summarization performance. Notably, MiniLM exhibits competitive retrieval quality without explicit medical pretraining, suggesting that lightweight general-purpose embeddings can still be effective when combined with lexical refinement and carefully designed prompting strategies. The experiments in this section identify the strongest semantic-lexical configuration. The next section extends the analysis by comparing this retrieval strategy with projection-guided multi-signal matching across both OpenI and MIMIC-CXR.

6.5. Cross-Dataset Comparison of Retrieval Configurations

The previous sections examine how retrieval and prompting choices affect performance on OpenI. Although the semantic-lexical strategy provides the strongest configuration among the first group of experiments, it still relies mainly on direct findings-side similarity. The projection-guided strategy is therefore evaluated to test whether target-oriented and feature-level matching signals can further improve retrieval beyond direct semantic and lexical alignment. This section compares the main retrieval configurations across both OpenI and MIMIC-CXR. The comparison includes the semantic-lexical retrieval strategy and two versions of the projection-guided multi-signal retrieval strategy. The first projection-guided configuration uses a simpler findings representation based on a single sentence encoder and sparse text features. The second projection-guided configuration uses a richer findings representation that combines multiple sentence encoders with structured clinical features. Both projection-guided configurations use the same multi-signal ranking framework described in Chapter 4.

Table 6.12. OpenI results for the main retrieval configurations.

Configuration	**R-1**	**R-2**	**R-L**	**BERT**	**CheXbert**
Semantic-lexical retrieval	0.7002	0.6174	0.6906	0.9462	0.7857
Projection-guided, single-encoder input	0.7092	0.6223	0.6960	0.9460	0.8103
Projection-guided, multi-encoder input	**0.7126**	**0.6289**	**0.6983**	**0.9466**	**0.8236**

Table 6.13. MIMIC-CXR results for the main retrieval configurations.

Configuration	R-1	R-2	R-L	BERT	CheXbert
Semantic-lexical retrieval	0.4484	0.2847	0.4004	0.9002	0.6777
Projection-guided, single-encoder input	0.4783	0.3034	0.4227	0.9030	0.6972
Projection-guided, multi-encoder input	0.4873	**0.3163**	**0.4326**	**0.9049**	0.6989
Multi-encoder input, full-reference check	**0.4878**	0.3158	0.4315	0.9048	**0.7025**

The OpenI results show that the projection-guided retrieval mechanism improves over the semantic-lexical retrieval strategy. ROUGE-1 increases from 0.7002 to 0.7092 with the single-encoder projection-guided configuration and reaches 0.7126 with the multi-encoder configuration. ROUGE-2 and ROUGE-L follow the same trend. The clinical label based score also improves, with CheXbert-F1 increasing from 0.7857 to 0.8236 in the multi-encoder configuration. These results indicate that the second retrieval mechanism provides additional benefit beyond direct semantic and lexical matching.

The MIMIC-CXR results show a similar pattern in a larger and more heterogeneous reporting setting. The semantic-lexical retrieval strategy obtains 0.4484 ROUGE-1, 0.2847 ROUGE-2, and 0.4004 ROUGE-L. The projection-guided single-encoder configuration improves these scores to 0.4783, 0.3034, and 0.4227, respectively. The multi-encoder configuration further improves the results to 0.4873 ROUGE-1, 0.3163 ROUGE-2, and 0.4326 ROUGE-L. This suggests that projection-guided multi-signal matching is especially useful when direct semantic and lexical matching alone is insufficient for selecting strong in-context examples. These MIMIC-CXR results should also be interpreted together with the dataset characteristics described in Table 5.1 and Table 5.2. Compared with OpenI, MIMIC-CXR contains longer evaluation reports, more diverse impressions, and a lower degree of direct impression reuse, which makes exact lexical matching and retrieval-based copying more difficult.

The comparison between the single-encoder and multi-encoder projection-guided configurations shows the effect of representation richness. The single-encoder version already improves over semantic-lexical retrieval by adding the projection-guided ranking mechanism. The multi-encoder version further improves performance by combining complementary sentence representations with structured clinical features. This effect is visible on both datasets, especially in ROUGE-2, ROUGE-L, and CheXbert-F1.

The full-reference check on MIMIC-CXR ranks the complete reference collection directly, rather than using a preliminary candidate reduction stage before reranking. Its scores are very close to the recall-based multi-encoder configuration. This indicates that the candidate reduction stage does not create a meaningful retrieval bottleneck in this setting. Instead, it functions mainly as an efficiency mechanism for large-scale retrieval, since full-reference ranking produces only marginal changes while requiring substantially more preparation and scoring time.

6.6. Ablation of the Multi-Signal Weights

The projection-guided retrieval strategy combines five signals with different coefficients. This ablation examines how the balance among these signals affects generation performance. All experiments are conducted on the 576-report OpenI evaluation set using Llama3. The generation, prompting, and iterative selection settings are kept unchanged, while only the coefficients used in the retrieval score are varied.

In the following tables, P denotes projected summary match, F denotes findings-to-findings match, B denotes key concept carry-over, C denotes findings-to-impression match, and R denotes clinical graph match. The systematic weight search contains 23 alternative configurations. The original configuration used in the main projection-guided experiments is additionally included as a reference.

6.6.1. Balance Between the Primary Retrieval Signals

The first analysis examines the balance between projected summary match and findings-to-findings match without using the three auxiliary signals. Table 6.14 presents the complete progression from projection-only retrieval to findings-only retrieval.

The projection-only configuration produces the weakest result, with 0.6785 ROUGE-L. Findings-to-findings matching provides a considerably stronger individual baseline, reaching 0.6920 ROUGE-L. This shows that the projected summary representation is not sufficiently reliable when used as the only retrieval criterion.

Combining the two signals generally improves over either extreme. Among the configurations using only the two primary signals, the strongest ROUGE-L score is ob-

Table 6.14. Ablation of the two primary retrieval signals on OpenI. All auxiliary signal coefficients are set to zero.

Configuration	P	F	**R-1**	**R-2**	**R-L**
Projection only	1.00	0.00	0.6915	0.6001	0.6785
Projection dominant	0.70	0.30	0.7034	0.6207	0.6922
Projection dominant	0.65	0.35	0.7056	**0.6244**	0.6943
Projection dominant	0.60	0.40	**0.7065**	0.6242	0.6942
Projection dominant	0.55	0.45	0.7042	0.6201	0.6924
Equal weighting	0.50	0.50	0.7023	0.6209	0.6906
Findings dominant	0.45	0.55	0.7051	0.6210	0.6918
Findings dominant	0.40	0.60	0.7064	0.6237	**0.6955**
Findings dominant	0.35	0.65	0.7042	0.6194	0.6923
Findings dominant	0.30	0.70	0.7026	0.6241	0.6914
Findings only	0.00	1.00	0.7030	0.6236	0.6920

tained with $P = 0.40$ and $F = 0.60$. However, performance does not change monotonically as the balance shifts toward either signal. The results instead indicate that projected summary match is most useful when it complements direct findings similarity rather than replacing it.

6.6.2. Contribution of Key Concept Carry-Over

The second analysis fixes the two primary coefficients at $P = 0.30$ and $F = 0.45$ and progressively adds key concept carry-over. The two remaining auxiliary signals are kept at zero. Table 6.15 reports the complete carry-over coefficient sweep.

Table 6.15. Ablation of the key concept carry-over coefficient on OpenI. The primary coefficients are fixed at $P = 0.30$ and $F = 0.45$, while C and R are set to zero.

Configuration	B	**R-1**	**R-2**	**R-L**
Low carry-over	0.05	0.7072	0.6227	0.6946
Carry-over	0.10	0.7083	0.6243	0.6963
Moderate carry-over	0.15	**0.7097**	**0.6269**	**0.6965**
Carry-over	0.20	0.7088	0.6218	0.6953
Carry-over	0.25	0.7090	0.6252	0.6961
High carry-over	0.30	0.7078	0.6194	0.6933

The results identify $B = 0.15$ as the most effective carry-over coefficient among the tested settings. Increasing the coefficient from 0.05 to 0.15 improves all three ROUGE scores. Further increases do not provide a consistent benefit, and setting $B = 0.30$ reduces ROUGE-2 to 0.6194 and ROUGE-L to 0.6933.

This pattern indicates that key concept carry-over is useful as a refinement signal, but should not dominate the retrieval score. A moderate coefficient helps prioritize reference examples whose findings contain concepts that are likely to be preserved in the impression, while a larger coefficient may overemphasize individual terms at the expense of broader report similarity.

6.6.3. Contribution of Findings-to-Impression and Clinical Graph Matching

The final analysis retains $P = 0.30$, $F = 0.45$, and $B = 0.15$, and evaluates the two remaining auxiliary signals. Table 6.16 first shows the configuration without either auxiliary signal, then introduces findings-to-impression matching and clinical graph matching separately, examines larger auxiliary coefficients, and finally reports the original configuration in which both signals receive a coefficient of 0.05.

Table 6.16. Ablation of findings-to-impression and clinical graph matching on OpenI. The coefficients $P = 0.30$, $F = 0.45$, and $B = 0.15$ are fixed. The original configuration is included as the final reference row.

Configuration	C	R	**R-1**	**R-2**	**R-L**
No additional signal	0.00	0.00	0.7097	0.6269	0.6965
Findings-to-impression only	0.05	0.00	0.7097	0.6277	0.6974
Clinical graph only	0.00	0.05	0.7119	**0.6322**	**0.6993**
Clinical graph only	0.00	0.10	0.7088	0.6282	0.6971
Clinical graph only	0.00	0.15	0.7103	0.6310	0.6989
Higher findings-to-impression	0.10	0.05	0.7104	0.6262	0.6967
Higher findings-to-impression	0.15	0.05	0.7102	0.6245	0.6967
Original configuration	0.05	0.05	**0.7126**	0.6289	0.6983

Adding findings-to-impression matching alone provides a small improvement over

the configuration without either auxiliary signal. ROUGE-L increases from 0.6965 to 0.6974, while ROUGE-1 remains unchanged. Increasing the findings-to-impression coefficient beyond 0.05 does not improve performance when clinical graph matching is also present. The configurations with $C = 0.10$ and $C = 0.15$ both produce lower ROUGE-2 and ROUGE-L scores.

Clinical graph matching produces a clearer improvement. Assigning $R = 0.05$ while keeping $C = 0$ achieves the highest ROUGE-2 and ROUGE-L scores in the ablation, reaching 0.6322 and 0.6993, respectively. Increasing the graph coefficient to 0.10 or 0.15 does not improve over the smaller value, although the $R = 0.15$ configuration remains competitive.

The original configuration assigns 0.05 to both findings-to-impression matching and clinical graph matching. It achieves the highest ROUGE-1 score, 0.7126, together with 0.6289 ROUGE-2 and 0.6983 ROUGE-L. Removing findings-to-impression matching while retaining clinical graph matching therefore improves ROUGE-2 and ROUGE-L slightly, but reduces ROUGE-1 by 0.0007. The two configurations are very close overall and emphasize different aspects of lexical agreement.

Overall, the complete ablation supports the multi-signal retrieval design. Findings-to-findings matching provides the strongest standalone foundation, while projected summary match becomes useful when combined with direct findings similarity. Key concept carry-over provides its strongest contribution at a moderate coefficient of 0.15. Among the two lower-weighted auxiliary signals, clinical graph matching produces the clearer independent improvement, whereas findings-to-impression matching provides a smaller refinement. The original configuration remains highly competitive, while the alternative configuration with $C = 0$ and $R = 0.05$ obtains the strongest ROUGE-2 and ROUGE-L results.

6.7. Controlled Generator Comparison with GPT-5.4

The previous experiments evaluate the proposed retrieval framework using compact open-source language models. This section introduces GPT-5.4 as an external controlled comparator in order to examine how much the generator affects performance when retrieval and prompting conditions are kept fixed. GPT-5.4 is not a component of the proposed

framework and does not alter its open-source design.

For each dataset, the retrieved examples, prompt construction, iterative generation protocol, and output selection procedure are preserved. Only the generator is changed between Llama3 and GPT-5.4. On OpenI, both generators use the same multi-encoder projection-guided retrieval configuration. On MIMIC-CXR, both generators similarly use the same fixed projection-guided retrieval setting. This design isolates generator behavior from changes in example selection.

Table 6.17 reports lexical and clinical agreement results. CheXbert denotes 14-category micro-F1, while RG-C denotes complete RadGraph-F1. All clinical metrics are calculated using the same evaluation implementations for both generators.

Table 6.17. Controlled comparison between Llama3 and GPT-5.4 under fixed retrieval and prompting conditions. CheXbert denotes 14-category micro-F1, and RG-C denotes complete RadGraph-F1.

Dataset	**Generator**	**R-1**	**R-2**	**R-L**	**CheXbert**	**RG-C**
OpenI	Llama3	0.7126	**0.6289**	0.6983	0.8236	0.6165
OpenI	GPT-5.4	**0.7235**	0.6262	**0.7074**	**0.8371**	**0.6207**
MIMIC-CXR	Llama3	**0.4873**	**0.3163**	**0.4326**	0.6989	**0.3459**
MIMIC-CXR	GPT-5.4	0.4872	0.3067	0.4295	**0.7051**	0.3385

On OpenI, GPT-5.4 achieves higher ROUGE-1 and ROUGE-L scores than Llama3, increasing ROUGE-1 from 0.7126 to 0.7235 and ROUGE-L from 0.6983 to 0.7074. GPT-5.4 also obtains higher CheXbert and complete RadGraph agreement. Llama3, however, produces the higher ROUGE-2 score, with 0.6289 compared with 0.6262. Therefore, GPT-5.4 provides the stronger overall OpenI result, but it does not improve every reported metric.

The MIMIC-CXR results show a different pattern. ROUGE-1 is effectively identical for the two generators, with 0.4873 for Llama3 and 0.4872 for GPT-5.4. Llama3 obtains higher ROUGE-2, ROUGE-L, and complete RadGraph-F1 scores. GPT-5.4 obtains higher CheXbert agreement, increasing 14-category micro-F1 from 0.6989 to 0.7051. The clinical metrics therefore do not identify one generator as uniformly stronger: GPT-5.4 provides higher observation-label agreement, while Llama3 provides stronger structured

clinical and sequence-level agreement.

The contrast between the two datasets suggests that generator advantages are dataset-dependent. GPT-5.4 provides a clearer improvement on the shorter and more repetitive OpenI reports, whereas the compact open-source generator remains approximately equal or slightly stronger on several metrics in the longer and more heterogeneous MIMIC-CXR setting. Changing to a proprietary generator therefore does not automatically improve every evaluation dimension when retrieval evidence is held constant.

Overall, the controlled comparison supports the importance of retrieval design. Generator choice affects the final output, but the same retrieval framework enables Llama3 to remain close to GPT-5.4 and to outperform it on several MIMIC-CXR metrics. This result does not imply that the two generators are equivalent in general. It shows that, under the fixed retrieval and prompting conditions examined in this thesis, a compact open-source model can remain competitive with a proprietary generator.

6.8. Comparative Analysis with Prior Studies

The final configurations are compared with prior radiology impression generation and report summarization studies on OpenI and MIMIC-CXR. The same set of comparison methods is used across the two datasets when directly comparable results are available. This allows the proposed semantic-lexical retrieval strategy and the projection-guided multi-signal retrieval strategy to be compared with graph-based, multimodal, proprietary LLM-based, and multi-agent approaches under a common presentation. For MIMIC-CXR, the projection-guided row reports the final projection-guided variant selected for external comparison based on the cross-dataset analysis in the previous section.

The OpenI comparison shows that both proposed retrieval strategies are competitive with prior work. The semantic-lexical retrieval strategy already obtains higher ROUGE scores than the listed prior studies, with 0.7002 ROUGE-1, 0.6174 ROUGE-2, and 0.6906 ROUGE-L. The projection-guided multi-signal retrieval strategy further improves these scores to 0.7126 ROUGE-1, 0.6289 ROUGE-2, and 0.6983 ROUGE-L.

The additional metrics show the same general trend where they are available. The semantic-lexical retrieval strategy obtains 0.9462 raw BERTScore-F1 and 0.7857 CheXbert-F1, while the projection-guided multi-signal retrieval strategy increases these

Table 6.18. Comparison with previous studies on OpenI. BERT denotes raw BERTScore-F1, and CheXbert denotes clinical label based F1 when reported.

Model / Method	**R-1**	**R-2**	**R-L**	**BERT**	**CheXbert**
WordGraph (Hu et al. 2021) Seq2Seq + Word Graph (GAT)	0.6432	0.5548	0.6397	–	–
GraphContrastive (Hu et al. 2022) Transformer + Graph encoder	0.6497	0.5559	0.6445	–	–
VisualPrompt (Hu et al. 2023) CNN image encoder + Transformer	0.6800	0.5989	0.6787	–	–
ImpressionGPT (Ma et al. 2024) ChatGPT (GPT-3.5 / GPT-4)	0.6637	0.5493	0.6547	0.9383	0.6562
EMAI (Li et al. 2025) GPT-4 multi-agent	0.6650	0.5764	0.6603	–	0.7578
CCL (Luo et al. 2025) T5 + ChatGPT (contrastive)	0.6835	0.5948	0.6810	–	–
Ours Semantic-lexical retrieval	0.7002	0.6174	0.6906	0.9462	0.7857
Ours Projection-guided multi-signal retrieval	**0.7126**	**0.6289**	**0.6983**	**0.9466**	**0.8236**

values to 0.9466 and 0.8236, respectively. This suggests that the improvement is not limited to lexical overlap, but is also reflected in semantic similarity and clinical label agreement.

The MIMIC-CXR comparison presents a more challenging setting. The semantic-lexical retrieval strategy obtains 0.4484 ROUGE-1, 0.2847 ROUGE-2, and 0.4004 ROUGE-L. The projection-guided multi-signal retrieval strategy improves these scores to 0.4878 ROUGE-1, 0.3158 ROUGE-2, and 0.4315 ROUGE-L. This demonstrates that the second retrieval mechanism provides a clear gain over the base retrieval strategy on the larger and more heterogeneous MIMIC-CXR dataset.

Compared with prior studies, the proposed projection-guided strategy does not obtain the highest ROUGE scores on MIMIC-CXR. Among the directly comparable methods in the table, EMAI reports stronger ROUGE-1, ROUGE-2, and ROUGE-L scores. However, the proposed method obtains a strong CheXbert-F1 score of 0.7025, which is higher than the reported CheXbert values for WordGraph, GraphContrastive, VisualPrompt, and EMAI. This indicates that the proposed retrieval-guided framework

Table 6.19. Comparison with previous studies on MIMIC-CXR. BERT denotes raw BERTScore-F1, and CheXbert denotes 14-category micro-F1 when reported. Results with substantial comparability concerns are discussed in the text but excluded from the main table.

Model / Method	**R-1**	**R-2**	**R-L**	**BERT**	**CheXbert**
WordGraph (Hu et al. 2021) Transformer variant	0.4837	0.3334	0.4668	–	0.5388
GraphContrastive (Hu et al. 2022) Graph contrastive learning	0.4913	0.3376	0.4712	–	0.5452
VisualPrompt (Hu et al. 2023) Radiograph and anatomy prompts	0.4763	0.3203	0.4613	–	0.5455
EMAI (Li et al. 2025) GPT-4 multi-agent	**0.5064**	**0.3647**	**0.4933**	–	0.6421
Ours Semantic-lexical retrieval	0.4484	0.2847	0.4004	0.9002	0.6777
Ours Projection-guided multi-signal retrieval	0.4878	0.3158	0.4315	**0.9048**	**0.7025**

ImpressionGPT (Ma et al. 2024) and CCL (Luo et al. 2025) are discussed in the text but excluded from the main MIMIC-CXR comparison table because of substantial comparability concerns related to target-side label usage and split construction.

is particularly competitive in terms of clinical label agreement, even when exact n-gram overlap and sequence-level alignment are not the highest.

The MIMIC-CXR results reported for ImpressionGPT (Ma et al. 2024) and CCL (Luo et al. 2025) are not included in Table 6.19 because they introduce comparability concerns that make direct tabular ranking difficult. ImpressionGPT reports 0.5445 ROUGE-1, 0.3450 ROUGE-2, 0.4793 ROUGE-L, 0.9141 BERTScore-F1, and 0.8009 CheXbert-F1 on MIMIC-CXR. However, its example selection procedure appears to use CheXpert label information associated with the target case. Under a findings-only generation setting, this suggests a high likelihood of target-side information leakage and makes the result unsuitable as a directly comparable no-leak retrieval baseline. CCL reports 0.5138 ROUGE-1, 0.4234 ROUGE-2, 0.5124 ROUGE-L, and 0.7548 CheXbert-F1, but these scores are obtained on a custom split of 12,392 cases rather than the 1,606-report official evaluation setting used by the remaining MIMIC-CXR comparisons. This split difference is important because the official evaluation setting is more constrained and challenging, while a separately constructed split can change the difficulty and distribution of the test cases.

Therefore, both results are discussed for context, but they are excluded from the main MIMIC-CXR comparison table.

Overall, the comparison supports two conclusions. First, the semantic-lexical retrieval strategy provides a strong open-source text-only baseline, especially on OpenI. Second, the projection-guided multi-signal retrieval strategy improves over that baseline and provides stronger semantic and clinical label agreement across datasets. Unlike multimodal systems that use radiographic image features and proprietary systems that depend on ChatGPT or GPT-4, the proposed framework relies on text-only retrieval-guided prompting without language model fine-tuning or target-side clinical label retrieval.

6.9. Cross-Modality Results on MIMIC-III

The additional MIMIC-III experiment evaluates whether the retrieval-guided framework transfers beyond chest X-ray reporting. The experiment uses Llama3 under a fixed projection-guided retrieval configuration applied consistently across all six domains, together with the official RadSum23 test reports reconstructed from MIMIC-III. The same findings-to-impression formulation is retained across all domains. In total, the evaluation covers 7,420 reports from six computed tomography and magnetic resonance imaging domains.

6.9.1. Domain-Level ROUGE Performance

Table 6.20 presents the domain-level ROUGE results. Performance varies substantially across modality-anatomy pairs. CT Spine obtains the highest ROUGE-L score of 0.3941, followed by CT Head with 0.3721. CT Abdomen/Pelvis is the most difficult domain, with a ROUGE-L score of 0.2021.

The test-size-weighted scores across the six domains are 0.4183 ROUGE-1, 0.2071 ROUGE-2, and 0.3106 ROUGE-L. These values are lower than the corresponding OpenI and MIMIC-CXR results, indicating that cross-modality radiology summarization presents a substantially more difficult retrieval and generation setting. Unlike chest X-ray reports, the MIMIC-III domains cover different anatomical regions and exhibit greater variation in report structure, terminology, and summary content.

Table 6.20. Domain-level ROUGE results on the RadSum23 MIMIC-III evaluation set.

Domain	Reports	R-1	R-2	R-L
CT Head	3,141	0.4577	0.2612	0.3721
CT Spine	553	**0.4712**	**0.2816**	**0.3941**
CT Chest	1,280	0.4011	0.1666	0.2624
MR Head	732	0.4324	0.2131	0.3108
CT Neck	115	0.3966	0.1741	0.2739
CT Abdomen/Pelvis	1,599	0.3315	0.1071	0.2021

A notable pattern is the relatively wide difference between the weighted ROUGE-1 and ROUGE-L scores. The gap of 0.1077 may indicate that the generated impressions preserve a considerable amount of relevant unigram-level terminology, while matching the ordered phrasing and structural organization of the reference impressions less closely. This pattern is consistent with the greater variation in impression structure across CT and MR domains. However, the gap cannot be compared directly with the participating RadSum23 systems, since the official shared-task results do not report ROUGE-1 or ROUGE-2.

The differences among domains also show that imaging modality alone does not determine task difficulty. CT Spine and CT Head produce the strongest results, whereas CT Abdomen/Pelvis and CT Chest are considerably more difficult. This suggests that anatomical scope, report diversity, and the availability of reusable findings-to-impression patterns also affect retrieval-guided generation performance.

6.9.2. Structured Clinical Agreement

Because CheXbert is designed for chest X-ray observations, it is not used to interpret the CT and MR results. Structured clinical agreement is instead evaluated using RadGraph-XL (Delbrouck et al. 2024). Table 6.21 reports simple, partial, and complete RadGraph-F1 for each domain.

The RadGraph-XL results show a domain pattern broadly consistent with the ROUGE analysis. CT Head obtains the highest structured clinical agreement, while CT Abdomen/Pelvis obtains the lowest scores under all three matching settings. The weighted simple, partial, and complete RadGraph-F1 scores across the six domains are approximately 0.338, 0.293, and 0.230, respectively. The results indicate that the difficulty

Table 6.21. Domain-level RadGraph-XL results on the RadSum23 MIMIC-III evaluation set.

Domain	Simple F1	Partial F1	Complete F1
CT Head	**0.4438**	**0.3925**	**0.3221**
CT Spine	0.3585	0.2844	0.2437
CT Chest	0.2514	0.2220	0.1630
MR Head	0.3313	0.2745	0.2014
CT Neck	0.2874	0.2323	0.1565
CT Abdomen/Pelvis	0.2009	0.1712	0.1182

of CT Abdomen/Pelvis is not limited to lexical variation; the generated impressions also preserve less of the structured clinical content found in the reference impressions.

CT Spine ranks first according to ROUGE but not according to RadGraph-XL. This difference demonstrates that lexical agreement and structured clinical agreement capture related but distinct properties. A generated impression may reproduce the expected phrasing closely while achieving lower agreement over extracted entities and relations, or it may preserve important clinical content despite using different wording.

6.9.3. Contextual Comparison with RadSum23 Systems

The aggregate results are compared contextually with the leading systems reported for the public RadSum23 MIMIC-III test set (Delbrouck et al. 2023). The official public leaderboard contains 6,526 reports across eleven modality-anatomy pairs, whereas the evaluation conducted in this thesis contains 7,420 reports from six in-distribution CT and MR domains. Therefore, the aggregate values are not based on identical evaluation subsets and should not be interpreted as a direct head-to-head ranking.

The official shared-task evaluation reports ROUGE-L and F1-RadGraph, but does not provide ROUGE-1 or ROUGE-2. The F1-RadGraph metric used in RadSum23 corresponds to the RG_ER matching level, implemented as the partial reward level. The partial F1 obtained with the classical RadGraph implementation is therefore reported for the proposed framework.

The proposed framework obtains numerically lower aggregate ROUGE-L and F1-RadGraph scores than the two leading RadSum23 systems. The difference is approximately

Table 6.22. Contextual comparison with leading systems reported for the RadSum23 MIMIC-III task. The reported systems and the proposed framework are evaluated on different aggregate subsets. F1-RadGraph corresponds to the RG_ER, or partial, matching level.

System	Reports	R-L	F1-RadGraph	Method Setting
utsa-nlp	6,526	0.3407	0.3525	Retrieval-based few-shot prompting and ensemble
shs-nlp	6,526	0.3393	0.3493	Domain-adaptive pre-training
Ours	7,420	0.3106	0.3298	Retrieval-guided generation without language model fine-tuning

0.0301 in ROUGE-L and 0.0227 in F1-RadGraph relative to utsa-nlp. However, these differences cannot be attributed solely to the generation methods because the reported values are calculated over different evaluation subsets.

The comparison nevertheless provides a useful indication of the performance range obtained by the proposed framework. The results remain within approximately two to three points of the leading shared-task scores, despite keeping the language model parameters unchanged. The compared systems also use different forms of inference support or model adaptation: utsa-nlp combines retrieval-based prompting with few-shot chain-of-thought and ensemble techniques, while shs-nlp uses domain-adaptive pre-training. Therefore, the comparison should be interpreted as contextual evidence rather than a direct leaderboard ranking.

CHAPTER 7

DISCUSSION

This chapter interprets the experimental findings beyond the numerical scores reported in the previous chapter. The results show that retrieval-guided prompting substantially improves radiology impression generation, but the reasons for this improvement are distributed across several design choices: semantic retrieval, lexical refinement, projection-guided ranking, prompt structure, example ordering, hard negative selection, and iterative feedback. The purpose of this discussion is to explain how these components interact and why they are important for compact open-source text-only language models.

The discussion also considers the differences among OpenI, MIMIC-CXR, and the MIMIC-III RadSum23 evaluation setting. These datasets do not provide the same retrieval environment. OpenI is more compact, repetitive, and template-like, whereas MIMIC-CXR contains longer reports, more diverse impressions, and a lower degree of direct impression reuse. MIMIC-III extends the evaluation beyond chest X-ray reporting to CT and MR reports from multiple anatomical domains. These differences affect how ROUGE, BERTScore, CheXbert, and RadGraph results should be interpreted.

Finally, the chapter examines qualitative behavior observed in representative test cases. Although automatic metrics provide a useful basis for comparison with previous studies, they do not fully explain whether a generated impression preserves subtle abnormalities, avoids unsupported findings, or follows the concise style expected in radiology reporting. Therefore, this chapter also discusses typical success cases, recurring error patterns, and finding types that remain difficult for the proposed framework.

7.1. Dataset Difficulty and Metric Interpretation

The experimental results should be interpreted together with the dataset characteristics described in Chapter 5. OpenI and MIMIC-CXR differ not only in scale, but also in repetition, report length, and retrieval difficulty. These differences help explain why OpenI produces higher absolute ROUGE scores, while MIMIC-CXR provides a stricter evaluation setting.

OpenI contains a higher degree of impression repetition. A substantial portion of evaluation impressions can be found exactly in the reference collection, and the nearest-neighbor findings similarity is high. This means that retrieval can often identify examples whose findings and impressions are very close to the target case. As a result, ROUGE scores on OpenI can be strongly influenced by the ability to retrieve and reuse common impression templates.

MIMIC-CXR is less repetitive and more heterogeneous. Its evaluation set contains longer findings and impression sections, more unique impressions, and a lower degree of direct impression reuse. This reduces the effectiveness of simple retrieval-based copying and makes exact lexical overlap harder to achieve. Therefore, lower ROUGE scores on MIMIC-CXR do not necessarily indicate a proportional loss of clinical content preservation. They also reflect the greater diversity and difficulty of the evaluation setting.

This distinction is important for comparing ROUGE and clinical label based metrics. ROUGE rewards exact n-gram overlap and local sequence similarity. It is therefore sensitive to phrasing, ordering, and template reuse. CheXbert-F1, in contrast, evaluates agreement over clinical observation labels and can remain strong even when the generated impression uses different wording. The MIMIC-CXR results should therefore be read as evidence that the proposed framework is stronger in clinical content agreement than in exact phrase-level reconstruction.

7.2. Generalization Across Radiology Modalities

The MIMIC-III experiment extends the evaluation beyond the chest X-ray setting used for the main method development and ablation studies. The same findings-to-impression formulation and a fixed projection-guided retrieval configuration are applied across six CT and MR domains. The resulting performance shows that the framework can be transferred to reports involving different imaging modalities and anatomical regions without language model fine-tuning. However, the lower aggregate scores relative to OpenI and MIMIC-CXR also show that this transfer is substantially more difficult than evaluation within the chest X-ray domain.

The domain-level results demonstrate that cross-modality difficulty is not uniform. CT Spine obtains the strongest ROUGE results, while CT Head obtains the strongest

RadGraph-XL results. In contrast, CT Abdomen/Pelvis produces the lowest scores under both lexical and structured clinical evaluation. CT Chest also performs below the head and spine domains. These differences suggest that imaging modality alone does not determine performance. Anatomical scope, reporting style, terminology, summary structure, and the availability of closely aligned reference examples may all affect the usefulness of retrieval-guided prompting.

The contrast between CT Spine and CT Head is particularly informative. CT Spine ranks first according to ROUGE, indicating stronger lexical and sequential agreement with the reference impressions. CT Head ranks first according to all three RadGraph-XL matching settings, indicating stronger agreement over extracted clinical entities and relations. The different rankings show that a system can reproduce reference wording relatively well without achieving the same level of structured clinical agreement. Conversely, clinically relevant entities and relations may be preserved even when the generated impression does not closely reproduce the reference phrasing.

Another notable result is the difference between the aggregate ROUGE-1 and ROUGE-L scores. The weighted ROUGE-1 score is 0.4183, whereas the weighted ROUGE-L score is 0.3106. This difference may indicate that the generated impressions preserve a meaningful portion of the terminology found in the reference impressions, but follow their ordered phrasing and structural organization less closely. CT and MR reports cover more varied anatomical regions and reporting patterns than the chest X-ray datasets, making exact sequence-level reconstruction more difficult. This interpretation should nevertheless remain cautious, because the official RadSum23 results do not report ROUGE-1 or ROUGE-2 for the participating systems.

The RadGraph-XL results provide additional evidence that the lower MIMIC-III performance cannot be explained only by variation in surface wording. CT Abdomen/Pelvis remains the weakest domain under simple, partial, and complete RadGraph matching. This indicates that the generated impressions also preserve less of the structured clinical content in this domain. Reports covering the abdomen and pelvis may contain multiple organs, findings, anatomical relationships, and clinically relevant priorities, creating a more complex summarization setting than narrower anatomical domains.

The comparison with the leading RadSum23 systems should be interpreted contextually rather than as a direct leaderboard ranking. The proposed framework is evaluated on

7,420 reports from six in-distribution CT and MR domains, whereas the public RadSum23 results are aggregated over a different test subset and a broader set of modality-anatomy pairs. The compared systems also use different adaptation strategies. Therefore, the numerical differences do not isolate the effect of the generation method alone. Nevertheless, the results show that retrieval-guided generation without language model fine-tuning remains within a relatively close performance range of systems using model adaptation, domain-adaptive pre-training, or more elaborate inference procedures.

Overall, the MIMIC-III results support a qualified generalization claim. The proposed framework is applicable beyond chest X-ray reporting and can generate impressions for multiple CT and MR domains under a common retrieval-guided formulation. At the same time, the substantial differences among domains show that performance remains sensitive to anatomical scope, report characteristics, and the quality of the available retrieval evidence. Cross-modality transfer is therefore possible, but it should not be interpreted as uniform performance across radiology domains.

7.3. Why Retrieval-Guided Prompting Works

The zero-shot and random-shot results show that compact decoder-only models are not sufficient as standalone clinical summarizers. Without relevant examples, the models either produce weak summaries or rely on generic impression patterns. Random-shot prompting improves performance compared with zero-shot prompting, which indicates that these models are sensitive to the presence of demonstrations. However, random examples do not provide target-specific clinical guidance, and their performance remains far below retrieval-guided configurations.

Retrieval-guided prompting improves performance because it provides case-specific demonstrations that are close to the target findings. These demonstrations serve two main functions. First, they provide clinical content patterns that are relevant to the current case. Second, they provide stylistic guidance by showing how detailed findings sections are typically compressed into concise impressions. In radiology impression generation, both functions are important. The model must determine which findings are clinically important, but it must also express the output in the compact style expected from an impression section.

The supporting retrieval analysis further confirms that retrieval quality is closely related to generation quality. Table 7.1 groups MIMIC-CXR cases by retrieval quality. Cases are first sorted according to the ROUGE-1 score of the best retrieved impression among the top-15 retrieved examples, and then divided into three equal-sized groups. The generated output quality increases consistently from the lowest retrieval-quality group to the highest retrieval-quality group.

Table 7.1. Relationship between retrieval quality and generation quality on MIMIC-CXR. Cases are sorted by the ROUGE-1 score of the best retrieved impression among the top-15 retrieved examples and divided into three equal-sized groups. Values report group averages.

Retrieval Quality Group	**Avg. Best Retrieved R-1**	**Avg. Generated R-1**
Lowest third	32.88	28.00
Middle third	53.08	46.65
Highest third	85.82	68.89

The same analysis shows a strong correlation between best-of-15 retrieval quality and generated output quality, with a correlation coefficient of 0.747. This indicates that retrieval is not merely an auxiliary prompt formatting step; it directly affects the clinical and lexical content available to the model. When the retrieved examples are closer to the target case, the generated impression is more likely to match the reference impression.

This result supports the broader claim that retrieval quality can compensate for limited model scale. Instead of requiring a larger language model or language model fine-tuning, the framework improves the input context supplied to the model. This is especially important for clinical NLP, where the use of proprietary or very large models may be limited by privacy, cost, reproducibility, or local deployment constraints.

7.4. Generation Contribution Beyond Retrieval Copying

The dataset analysis also helps distinguish retrieval support from simple retrieval copying. On OpenI, the nearest retrieved examples are often very close to the target report. In contrast, MIMIC-CXR provides weaker direct copy opportunities because the

evaluation impressions are more diverse and less repetitive.

Table 7.2 compares direct retrieval copying with the generated outputs. The top-1 copy row represents the result of directly copying the impression of the highest-ranked retrieved example. The best-of-15 row represents an oracle retrieval ceiling in which the best impression among the retrieved examples is selected according to the reference. These rows do not represent deployable generation systems; they are diagnostic baselines used to understand how much of the final performance can be explained by retrieval alone.

Table 7.2. Retrieval-copy baselines and generated output performance. Retrieval-copy rows are used as diagnostic baselines, while the generated row reports the best corresponding results reported in the experiments.

Condition	**MIMIC R-1**	**MIMIC R-L**	**OpenI R-1**	**OpenI R-L**
Top-1 retrieved impression copied	40.98	36.47	69.27	68.26
Best retrieved impression among top-15	57.25	52.81	76.76	75.57
Generated impression	48.78	43.15	71.26	69.83
Gain over top-1 copy	+7.80	+6.68	+1.99	+1.57

The results in Table 7.2 show that the role of generation differs across datasets. On OpenI, copying the top-1 retrieved impression already produces a very high ROUGE-1 score, and generation improves over this copy baseline only modestly. This supports the interpretation that OpenI provides a strong retrieval ceiling because many target impressions follow recurring templates.

The pattern is different on MIMIC-CXR. Directly copying the top-1 retrieved impression produces a substantially lower score, while the generated impressions improve over this copy baseline by a larger margin. This suggests that the model contributes more actively on MIMIC-CXR by combining information from retrieved examples and adapting it to the candidate findings, rather than simply reproducing the closest retrieved impression.

This distinction is important for interpreting the contribution of the proposed framework. Strong OpenI performance demonstrates that the retrieval and prompting system can exploit highly relevant examples when they exist. MIMIC-CXR performance

demonstrates a different property: the system can still improve generation when exact retrieval matches are weaker and when the model must synthesize a more case-specific impression from less directly reusable examples.

7.5. From Semantic-Lexical to Projection-Guided Retrieval

The comparison between embedding-only retrieval and semantic-lexical retrieval shows that semantic similarity and lexical overlap provide complementary information. Sentence embeddings are useful for identifying conceptually related reports even when phrasing differs. This is important because radiology reports often describe similar clinical conditions with different wording. A retrieval strategy that depends only on exact word overlap may fail to identify clinically relevant examples when terminology varies across reports.

However, semantic similarity alone is not sufficient. Radiology reports contain many clinically important terms whose presence, absence, or negation can change the diagnostic meaning. Lexical refinement helps preserve these terms by favoring examples that share important wording with the target findings. This is particularly useful for findings such as pleural effusion, pneumothorax, atelectasis, consolidation, cardiomegaly, or costophrenic angle blunting, where small lexical differences may correspond to meaningful clinical distinctions.

The semantic-lexical strategy therefore provides a strong and relatively simple first retrieval mechanism. Its strength is that it is transparent, reproducible, and directly tied to findings-side similarity. This makes it a useful baseline and also explains why it performs well on OpenI, where many evaluation impressions are repetitive and where similar findings often lead to similar impression templates. However, this simplicity also defines its limitation. The method mainly asks whether two findings sections are similar. It does not explicitly ask whether the reference impression is useful for summarizing the candidate findings.

The projection-guided multi-signal strategy was introduced to address this limitation. Its design is not intended to add complexity for its own sake. Each signal targets a different failure mode observed in direct semantic-lexical retrieval. The projected summary match adds a target-oriented signal by estimating the impression-space

representation that is likely to correspond to the candidate findings. Findings-to-findings matching preserves direct source-side clinical similarity. Findings-to-impression matching checks whether the reference impression already contains concepts that are present in the candidate findings. Key concept carry-over emphasizes findings features that are likely to survive into impressions. Structured clinical matching adds agreement over clinical observations, anatomy, and assertion status when such information is available.

Using multiple signals is therefore motivated by the structure of the findings-to-impression task. A good in-context example should not only have similar findings; it should also provide an impression that is clinically useful as a summary pattern. A smaller set of criteria would make the retrieval mechanism simpler, but it would also remove one of these complementary views. For example, using only projected summary match could weaken direct findings similarity, while using only findings-to-findings matching would ignore whether the paired reference impression is useful. Similarly, lexical overlap alone may miss paraphrases, while dense similarity alone may blur clinically important negation or anatomical details.

The multi-signal design should be interpreted as a controlled fusion of complementary retrieval evidence rather than as a claim that the exact weighting scheme is universally optimal. The main ranking drivers preserve direct clinical similarity and target-oriented similarity, while the lower-weighted auxiliary signals refine the ordering of near-equal candidates. The improvement obtained by the projection-guided strategy suggests that effective retrieval for impression generation should consider both the similarity of findings sections and the usefulness of the corresponding reference impressions.

7.6. Interpreting the Multi-Signal Weighting Strategy

The projection-guided retrieval strategy uses five signals with nominal weights of 0.30, 0.45, 0.15, 0.05, and 0.05. The weight ablations reported in Tables 6.14–6.16 provide an empirical basis for interpreting this configuration. The results show that the final ranking should not be dominated by a single signal. Instead, the strongest configurations combine direct findings similarity with target-oriented projection and smaller auxiliary contributions.

The primary-signal ablation shows that findings-to-findings matching provides the

stronger standalone foundation. It achieves 0.6920 ROUGE-L when used alone, whereas projected summary matching alone obtains 0.6785. Nevertheless, several combined configurations improve over both standalone settings. This indicates that projected summary matching is useful not as a replacement for direct findings similarity, but as a complementary signal that introduces target-oriented information into example selection.

The carry-over ablation provides a similar result for key concept carry-over. Performance improves as its coefficient increases from 0.05 to 0.15, but declines when the coefficient reaches 0.30. The signal therefore contributes most effectively as a moderate refinement rather than as a primary ranking criterion. Among the two lower-weighted auxiliary signals, clinical graph matching provides the clearer independent improvement. A coefficient of 0.05 produces the strongest ROUGE-2 and ROUGE-L scores in the weight search, while larger graph coefficients do not provide further gains.

The original configuration, which assigns 0.05 to both findings-to-impression and clinical graph matching, remains highly competitive. It achieves the highest ROUGE-1 score in the reported ablation, while the alternative configuration that removes findings-to-impression matching and retains clinical graph matching achieves slightly stronger ROUGE-2 and ROUGE-L scores. The small differences between these configurations indicate that the method has a stable high-performing region around the primary coefficients of 0.30 and 0.45, a carry-over coefficient of 0.15, and small auxiliary coefficients.

The nominal weights should not be interpreted as direct percentages of influence. A signal's actual effect depends not only on its coefficient, but also on the scale, variation, and ranking behavior of the raw signal. Therefore, a lower nominal weight can still have a meaningful effect if the corresponding signal has relatively large or discriminative values among top candidates.

Table 7.3 reports the effective contribution of each signal among the selected top-15 candidates. The results show that the final score is mainly driven by two signals: projected summary match and findings-to-findings match. Although projected summary match has a nominal weight of 0.30, its raw values are high among selected candidates, giving it an effective score share of 45.5%. Findings-to-findings match has a nominal weight of 0.45 and an effective share of 45.3%. Thus, the final ranking is not equally distributed across all five signals; it is primarily controlled by a target-oriented semantic signal and a direct findings-side lexical signal.

Table 7.3. Effective contribution of the projection-guided retrieval signals among selected top-15 candidates. Mean raw signal denotes the average raw signal value before multiplication by the nominal weight. Effective share denotes the average percentage contribution to the final weighted score.

Signal	**Nominal Weight**	**Mean Raw Signal**	**Effective Share**	**Role**
Projected Summary Match	0.30	0.7494	45.5%	Main driver
Findings-to-Findings Match	0.45	0.4979	45.3%	Main driver
Key Concept Carry-Over	0.15	0.0923	2.8%	Specialist refinement
Findings-to-Impression Match	0.05	0.0361	0.4%	Tie-breaking refinement
Clinical Graph Match	0.05	0.5954	6.0%	Structured refinement

This distribution explains why the multi-signal strategy remains interpretable despite using several components. The two main signals carry approximately 90% of the effective score. Projected summary match provides a target-oriented estimate of which reference impressions are likely to be useful, while findings-to-findings match preserves direct clinical similarity between the candidate findings and reference findings. The remaining signals do not dominate the ranking. Instead, they provide additional evidence in cases where top candidates are close to each other or where specific clinical features need to be distinguished.

Table 7.4 shows a complementary leave-one-signal-out analysis. Each signal is removed from the scoring function, the candidates are re-ranked, and the resulting ranking is compared with the original top-15 set. This analysis measures ranking influence rather than score share. Findings-to-findings and projected summary match again behave as the main ranking drivers: removing either signal changes the top-ranked candidate in more than half of the cases and preserves only about 64% of the original top-15 set. In contrast, removing any of the auxiliary signals preserves approximately 87–90% of the top-15 candidates.

Table 7.5 provides a case-level view of how the same weighting strategy behaves for a clinically normal report and for a disease-positive report. The percentages in the table are computed relative to the total weighted retrieval score used for example selection. These values describe retrieval signal behavior and should not be interpreted as evaluation metrics such as ROUGE, BERTScore, or CheXbert-F1.

The following two examples show the target reports and the top-ranked retrieved

Table 7.4. Leave-one-signal-out ranking influence analysis for the projection-guided retrieval strategy. Top-1 changed indicates how often the highest-ranked candidate changes after removing the signal. Top-15 preserved indicates the proportion of the original top-15 candidates that remain after re-ranking.

Removed Signal	**Top-1 Changed**	**Top-15 Preserved**	**Interpretation**
Findings-to-Findings Match	63.4%	63.6%	Main ranking driver
Projected Summary Match	54.7%	63.8%	Main ranking driver
Key Concept Carry-Over	35.2%	87.2%	Refinement signal
Findings-to-Impression Match	33.2%	90.0%	Tie-breaking signal
Clinical Graph Match	35.9%	87.8%	Structured refinement signal

reports used in the case-level signal analysis. The first example is a normal case where the retrieved report is almost identical to the target report. The second example is a disease-positive case where the retrieved report shares some relevant concepts but is only partially aligned with the target report.

Normal case:

The target findings are: 'The lungs appear clear. The heart and pulmonary XXXX appear normal. The pleural spaces are clear. Mediastinal contours are normal.'

The gold impression is: 'No acute cardiopulmonary disease.'

The top-ranked retrieved findings are: 'Lungs appear clear. Heart and pulmonary XXXX appear normal. Pleural spaces are clear. Mediastinal contours are normal. No pneumothorax.'

The impression paired with the top-ranked retrieved reference report is: 'No acute cardiopulmonary disease.'

Disease-positive case:

The target findings are: 'Heart size and pulmonary vascularity appear within normal limits. Retrocardiac soft tissue density is present. There appears to be air within this which could suggest that this represents a hiatal hernia. Vascular calcification is noted. Calcified granuloma is seen. There has been interval development of bandlike opacity in the left lung base. This may represent atelectasis. No pneumothorax or pleural effusion is seen. Osteopenia is present in the spine.'

The gold impression is: '1. Retrocardiac soft tissue density. The appearance suggests hiatal hernia. 2. XXXX left base bandlike opacity. The appearance suggests atelectasis.'

> The top-ranked retrieved findings are: 'The heart size and pulmonary vascularity appear within normal limits. Left pleural effusion is present. A mass density is present in the left midlung zone. This measures approximately 3.2 cm in diameter. Air-fluid level is present behind the heart which probably represents a hiatal hernia. Some XXXX'
> The impression paired with the top-ranked retrieved reference report is: '1. Left midlung mass. 2. Left base effusion. 3. Probable hiatal hernia.'

In the normal case, the retrieved report provides an almost exact findings-to-impression pattern. This produces high projected summary, findings-to-findings, and clinical graph scores. In the disease-positive case, the retrieved report shares clinically relevant concepts such as hiatal hernia and left-sided abnormality, but it also introduces mismatched concepts such as left midlung mass and pleural effusion. This makes the case harder: the main signals remain active but are much lower than in the normal case, while key concept carry-over and findings-to-impression matching become nonzero. This example should therefore not be interpreted as a perfect retrieval match. Instead, it illustrates a difficult retrieval case in which the best available reference report is only partially aligned with the target findings. The lower raw signal values reflect this weaker match. In the full generation setting, this retrieved impression is only one reference example used for prompting, not the generated output itself.

The contrast between the two cases illustrates why the auxiliary signals are useful even though they have smaller average effective contributions. In the normal case, direct similarity is sufficient: projected summary match and findings-to-findings match nearly saturate, and the selected retrieved reference impression exactly matches the gold impression. In the disease-positive case, direct similarity is weaker and the retrieved reference example is only partially aligned with the target report. The auxiliary signals then provide additional evidence about which candidate concepts are likely to be summarized in the impression. Therefore, the multi-signal design is most useful in cases where simple normal-template matching is insufficient.

These results also explain why a smaller set of criteria was not used. If only findings-to-findings matching were used, the retrieval mechanism would reduce to a direct source-side similarity method and would lose the target-oriented information provided by projected summary match. If only projected summary match were used, the retrieval

Table 7.5. Case-level behavior of retrieval signals for a normal report and a disease-positive report. Each cell reports raw signal value / weighted contribution / share of the total weighted retrieval score.

Signal	**Normal Case**	**Disease-Positive Case**	**Interpretation**
Projected Summary Match	0.9791 / 0.2937 / 38.1%	0.5588 / 0.1676 / 47.9%	High in the normal template case; lower but still dominant in the disease-positive case
Findings-to-Findings Match	0.9521 / 0.4284 / 55.6%	0.2955 / 0.1330 / 38.0%	Strong normal lexical match; weaker alignment in the multi-finding disease case
Key Concept Carry-Over	0.0000 / 0.0000 / 0.0%	0.1779 / 0.0267 / 7.6%	Activated mainly when disease-specific concepts are likely to carry into the impression
Findings-to-Impression Match	0.0000 / 0.0000 / 0.0%	0.1292 / 0.0065 / 1.8%	Small but nonzero refinement in the disease-positive case
Clinical Graph Match	0.9585 / 0.0479 / 6.2%	0.3205 / 0.0160 / 4.6%	High structured agreement in the normal case; lower agreement in the complex disease case

mechanism could select reference impressions that are semantically plausible but insufficiently aligned with the candidate findings. The two main signals therefore address complementary parts of the task.

The auxiliary signals are included for a different reason. They are not intended to dominate the score. Key concept carry-over emphasizes findings concepts that tend to survive into impressions. Findings-to-impression matching checks whether the reference impression already contains candidate-side concepts. Clinical graph matching adds structured agreement over clinical observations, anatomy, and assertion status. Their lower effective contributions indicate that they act mainly as refinements, but the leave-one-signal-out analysis shows that they can still affect the top candidate in near-tie situations.

The weight configuration should therefore be interpreted as a controlled fusion of complementary evidence rather than as a claim that one exact coefficient tuple is universally optimal. The weight ablations, effective contribution analysis, and leave-one-signal-out analysis provide consistent but distinct forms of evidence. The ablations identify a stable

high-performing coefficient region, the effective contribution analysis shows that the two primary signals account for approximately 90% of the weighted score, and the removal analysis confirms that these signals have the greatest influence on candidate ranking.

The auxiliary signals do not need to dominate the score to be useful. Key concept carry-over performs best at a moderate coefficient, clinical graph matching provides a measurable refinement at a low coefficient, and findings-to-impression matching has a smaller and less consistent effect. The original configuration remains balanced and highly competitive, even though removing findings-to-impression matching produces a small improvement in ROUGE-2 and ROUGE-L. Overall, the results support the structure of the multi-signal design while also showing that its main contribution comes from the complementary use of projected summary and findings-to-findings matching rather than from precise optimization of every coefficient.

7.7. Model-Specific Prompt Sensitivity

The few-shot ordering experiments demonstrate that compact language models do not use prompt context identically. LLaVA and Meditron benefit from reverse ordering, where the most relevant examples are placed closer to the query. This suggests a recency effect in which examples appearing later in the prompt exert stronger influence on generation. Llama3, in contrast, performs best when the most relevant examples appear first. This difference indicates that example ordering should be treated as a model-specific design parameter rather than a fixed preprocessing choice.

This finding is important because many retrieval-guided systems assume that selecting the correct examples is sufficient. The results in this thesis show that the placement of those examples is also important. A highly relevant demonstration may be less useful if it is positioned in a part of the prompt that the model underuses. Therefore, retrieval quality and prompt organization should be considered together.

The different ordering behavior also suggests that prompt design cannot be fully generalized across model families. Even when models have similar parameter scales, their training data, instruction tuning procedure, tokenizer, and context handling may influence how they use retrieved examples. For this reason, prompt ordering should be evaluated empirically for each model rather than assumed to transfer directly from one model to

another.

7.8. Iterative Refinement as Candidate Search

The iterative refinement experiments show that refinement should not be interpreted as a strictly monotonic improvement process. The model does not necessarily produce a slightly better impression at each step. Instead, the iterative procedure functions more like a controlled candidate search process. Each iteration gives the model another opportunity to generate a plausible impression under retrieved positive examples, hard negative examples, and feedback from previous generations.

Table 7.6 summarizes this behavior on MIMIC-CXR. Individual iterations remain in a similar average performance range, but selecting the best generated candidate across iterations improves the final score. The best output is also frequently not produced in the first iteration, which supports the interpretation that the iterative process explores alternative summaries rather than performing simple monotonic correction.

Table 7.6. Summary of iterative refinement behavior on MIMIC-CXR. The selected candidate score refers to the iterative analysis setting rather than the final best MIMIC-CXR configuration reported in the main results.

Measure	**Value**
Average ROUGE-1 of individual iterations	~44.5
Selected candidate ROUGE-1 in this analysis	47.84
Approximate gain from iteration selection	+3.3
Median best iteration index	8
Cases where the best output is not the first iteration	84.1%

This interpretation also clarifies the role of automatic feedback. The GOOD and BAD labels do not provide expert clinical correction. Instead, they guide the model toward outputs that are more similar to the retrieved positive impressions and away from weaker generated candidates. The method therefore acts as a lightweight prompt-level search mechanism rather than a true learning process. The language model parameters remain unchanged, and the improvement comes from candidate generation, feedback-

guided prompting, and final output selection.

The ablation on comparison sample size further supports this interpretation. Although the retrieval stage keeps a broader set of 15 positive examples, the strongest refinement results are obtained when feedback is computed using only the top five positives. This suggests that iterative feedback benefits from multiple reference comparisons, but that the comparison set must remain focused. A single retrieved example can make the feedback too narrow, while using all retrieved positives can dilute the signal with lower-ranked or less specific examples. Therefore, the top-five setting provides a practical balance between diversity and relevance during feedback computation.

7.9. Role of Hard Negative Examples

Hard negative examples improve performance when used in small numbers. Their benefit appears to come from contrastive guidance: the model is shown not only what a good findings-to-impression mapping looks like, but also what kinds of semantically plausible mappings should be avoided. This is particularly useful in radiology, where clinically different cases may still share many surface-level similarities.

The supporting retrieval analysis indicates that the selected positive examples and hard negative examples are meaningfully separated. Table 7.7 shows that positive examples have substantially higher average similarity to the target reference than hard negatives. This confirms that the hard negative selection process is not arbitrary. The hard negatives remain plausible enough to be useful as contrastive context, but they are less aligned with the target impression than the selected positives.

Table 7.7. Average ROUGE-based similarity of selected positive examples and hard negative examples on MIMIC-CXR.

Example Type	**Average ROUGE-Based Similarity**
Selected positive examples	45.37
Selected hard negative examples	21.55
Difference	23.82

This separation is important because a purely positive demonstration set may not sufficiently teach the model what to avoid. For example, two reports may both describe abnormal lung findings, but one may involve pleural effusion while another involves atelectasis or scarring. Hard negatives help by introducing examples that are close enough to be plausible but different enough to be clinically misleading if imitated.

However, the experiments also show that adding too many hard negatives does not continue to improve performance. Excessive negative context can distract the model from the positive demonstrations and reduce the influence of the most relevant examples. Therefore, hard negative prompting is useful but must be controlled carefully. In the experiments, a small number of hard negatives provides the best balance between contrastive guidance and preservation of the positive context.

7.10. Qualitative Interpretation of Model Behavior

Automatic metrics provide useful quantitative comparisons, but they do not fully explain how the proposed framework behaves on individual radiology reports. ROUGE, BERTScore, and CheXbert-F1 summarize different aspects of output quality, yet they do not directly show which types of findings are preserved, omitted, generalized, or incorrectly emphasized. Therefore, representative generations were qualitatively inspected to better understand the strengths and limitations of retrieval-guided prompting.

The qualitative examples in this section are drawn from OpenI. This choice is appropriate because OpenI is the main public benchmark used for direct comparison with prior findings-to-impression generation studies, and because its more template-like structure makes it easier to inspect how retrieval-guided prompting behaves in common reporting patterns. MIMIC-CXR is used in this thesis primarily for larger-scale robustness, cross-dataset comparison, and clinical agreement analysis.

The qualitative inspection shows that the framework performs particularly well when the target findings correspond to frequent reporting patterns, such as normal chest examinations or cases where the final impression is primarily expressed as the absence of acute cardiopulmonary abnormality. In these cases, retrieval often provides highly similar examples, and the generated impression tends to preserve the concise style of radiology reporting.

The framework is less reliable when the reference impression contains subtle abnormalities, uncertainty expressions, incidental findings, or chronic conditions that must be distinguished from acute disease. These cases are important because they reveal limitations that may not be fully captured by aggregate automatic scores.

Table 7.8 summarizes the main qualitative behavior patterns observed across representative test cases. The table is not intended as a manual error count, but as a structured interpretation of recurring model behaviors that help explain the numerical results. These patterns show that the framework is strongest when the target case follows frequent reporting templates, and less reliable when the clinically relevant content is subtle, uncertain, or outside common cardiopulmonary phrasing.

Table 7.8. Qualitative difficulty patterns observed in representative OpenI test cases.

Pattern	**Observed Behavior**	**Interpretation**
Common negative cases	Outputs often match reference impressions such as "No acute cardiopulmonary disease."	Frequent templates are well supported by retrieved examples.
Subtle opacity or atelectasis	The model may omit mild basilar opacity, scarring, or atelectasis.	Negative context can dominate small positive findings.
Uncertain findings	Expressions such as "question of," "may represent," or "versus" may be simplified or omitted.	Uncertainty is difficult to preserve through lexical prompting alone.
Nonpulmonary findings	Possible rib fracture or bony findings may be ignored.	The model tends to prioritize cardiopulmonary templates.
Chronic or incidental findings	Cardiomegaly, hyperinflation, hiatal hernia, or granulomatous disease may be underemphasized or overemphasized.	The system can confuse acute summary style with chronic finding preservation.

7.10.1. Cases That Are Handled Well

The model performs well when the findings contain a common negative pattern and the expected impression is a concise statement of no acute cardiopulmonary abnormality. In such cases, retrieval often provides highly similar examples with nearly identical impression phrasing, allowing the model to produce the expected concise summary.

Table 7.9 shows a representative successful case. The findings describe clear lungs, normal mediastinal contours, and clear pleural spaces. The generated output closely matches the reference impression.

Table 7.9. Representative successful case in which retrieval-guided prompting produces the expected concise impression.

Field	**Text**
Findings	The lungs appear clear. The heart and pulmonary XXXX appear normal. The pleural spaces are clear. Mediastinal contours are normal.
Reference Impression	No acute cardiopulmonary disease.
Generated Impression	No acute cardiopulmonary disease.
Commentary	The generated output preserves the expected negative impression and follows the concise style of radiology reporting.

A similar pattern is observed in cases where the findings contain no focal consolidation, pleural effusion, pneumothorax, or acute bony abnormality. When the reference impression states that there is no radiographic evidence of acute cardiopulmonary disease, the model often reproduces this structure accurately. These cases are relatively easy because the retrieval pool contains many similar examples, and the target impression follows a frequent reporting template.

7.10.2. Omission of Subtle Abnormalities

A recurring error pattern is the omission of subtle abnormalities. The model sometimes collapses a report into a generic negative impression even when the reference impression mentions a mild but relevant finding. This occurs particularly when the findings contain many normal statements but also include a small abnormality such as mild basilar opacity, subsegmental atelectasis, scarring, or a possible rib fracture.

Table 7.10 illustrates this behavior. The findings mention minimal left basilar opacities, likely representing subsegmental atelectasis or scarring. The reference impression preserves this abnormality, whereas the generated impression reduces the case to a generic negative statement.

Table 7.10. Representative omission error involving a subtle abnormality.

Field	**Text**
Findings	There are minimal XXXX left basilar opacities, XXXX subsegmental atelectasis or scarring. There is no focal airspace consolidation to suggest pneumonia. No pleural effusion or pneumothorax. Heart size is at the upper limits of normal.
Reference Impression	Minimal left basilar subsegmental atelectasis or scarring. No acute findings.
Generated Impression	No acute cardiopulmonary disease identified.
Commentary	The output captures the absence of acute cardiopulmonary disease but omits the mild left basilar atelectasis or scarring retained in the reference impression.

This example shows an important limitation of retrieval-guided prompting. When most of the findings section consists of negative statements, the model may overprioritize the common “no acute disease” pattern and suppress a subtle abnormality. This behavior is clinically relevant because mild abnormalities may still be important for follow up, comparison, or diagnostic interpretation.

7.10.3. Failure to Preserve Nonpulmonary Findings

Another difficult pattern involves findings that are clinically relevant but outside the most common cardiopulmonary impression templates. In chest X-ray reports, some impressions may focus on bony abnormalities, device positioning, hiatal hernia, or chronic incidental findings. If the retrieved examples are dominated by common cardiopulmonary patterns, the model may ignore these less frequent findings.

Table 7.11 shows an example involving a possible posterior rib fracture. The findings mention a lucency crossing the left posterior rib, and the reference impression recommends clinical correlation. The generated impression instead states that there is no acute cardiopulmonary abnormality.

Table 7.11. Representative error involving omission of a nonpulmonary finding.

Field	**Text**
Findings	Lucency crosses the 10th left posterior rib. Visualized portions of the thoracic spine are unremarkable. Mediastinal contours are normal. Lungs are clear. There is no pneumothorax or large pleural effusion.
Reference Impression	Question of posterior 10 rib fracture, correlate with XXXX tenderness.
Generated Impression	No acute cardiopulmonary abnormality.
Commentary	The generated output correctly reflects the absence of acute cardiopulmonary disease but misses the possible rib fracture, which is the main abnormality in the reference impression.

This type of error suggests that the system is strongest when the relevant impression belongs to frequent chest disease patterns. It is less reliable when the clinically important information is less common, located outside the cardiopulmonary template, or expressed as a recommendation for correlation.

7.10.4. Partial Improvement Through Refinement

Iterative refinement can improve the generated impression by reducing unnecessary details and moving the output closer to the expected impression style. Table 7.12 shows an example in which the initial output contains several details, including calcified granuloma and vascular calcification. The selected output is more concise and better aligned with the reference impression, focusing on probable hiatal hernia and interval left base opacity suggesting atelectasis.

Table 7.12. Representative case in which iterative refinement improves conciseness and relevance.

Field	**Text**
Findings	Heart size and pulmonary vascularity appear within normal limits. Retrocardiac soft tissue density is present. There appears to be air within this which could suggest that this represents a hiatal hernia. Vascular calcification is noted. Calcified granuloma is seen. There has been interval development of bandlike opacity in the left lung base. This may represent atelectasis. No pneumothorax or pleural effusion is seen.
Reference Impression	1. Retrocardiac soft tissue density. The appearance suggests hiatal hernia. 2. XXXX left base bandlike opacity. The appearance suggests atelectasis.
Initial Output	1. Hiatal hernia with air within the retrocardiac soft tissue density. 2. Calcified granuloma and vascular calcification. 3. Left lung base atelectasis may be present.
Selected Output	1. Probable hiatal hernia. 2. Interval development of left base opacities which may represent atelectasis.
Commentary	The selected output removes less central details and better matches the concise structure of the reference impression.

This case illustrates the value of iterative prompt feedback. The model does not simply generate a longer description of all findings; instead, refinement can help select a more impression like summary that emphasizes the clinically relevant conclusions.

7.10.5. Overgeneralization and Template Bias

The qualitative inspection also shows that the model can overgeneralize toward common impression templates. The phrase “No acute cardiopulmonary disease” appears frequently in chest X-ray impressions, and the model sometimes uses this type of output even when the reference impression contains a specific abnormality. This template bias is most visible when the findings contain many normal statements around one subtle positive finding.

This behavior helps explain why retrieval quality is critical. If retrieved examples are highly similar and preserve the key abnormality, they can guide the model away from overly generic outputs. If retrieved examples are only broadly similar or dominated by common negative templates, the generated impression may become too generic. The results therefore show that retrieval quality depends not only on overall semantic similarity, but also on whether rare, uncertain, or subtle findings are represented in the selected examples.

7.10.6. Implications for Retrieval Design

The qualitative examples suggest that retrieval quality should not be interpreted only as high overall similarity. In several difficult cases, retrieved examples may be broadly similar to the target report but still fail to preserve the clinically decisive abnormality. This is especially important when the target findings contain a small positive finding surrounded by many normal statements. In such cases, a retrieved example with a frequent negative impression template may guide the model toward an overly generic output.

This observation has two implications for retrieval design. First, the retrieval stage should preserve not only global semantic similarity but also abnormality coverage. A report that shares the same normal background but omits the target abnormality may be less useful than a report that is slightly less similar overall but contains the same clinically important finding. Second, uncertainty expressions and low-frequency findings require special attention. Terms such as ‘possible,’ ‘question of,’ ‘may represent,’ and ‘versus’ often carry important diagnostic meaning, but they may be weakened when retrieval is dominated by common lexical or semantic patterns.

This interpretation also explains the motivation for the projection-guided multi-signal retrieval mechanism. The additional retrieval signals are designed to preserve target-oriented summary usefulness, clinical salience, and structured observation agreement without abandoning the simplicity of text-based retrieval. In this sense, the second retrieval mechanism is a response to the observed limitations of direct semantic-lexical matching rather than a separate or unrelated extension.

7.11. Summary of Qualitative Findings

The qualitative cases can be summarized along a clear difficulty pattern. The framework handles common normal or no acute cardiopulmonary cases relatively well. These reports often contain standard negative findings and have reference impressions that follow common phrasing such as 'No acute cardiopulmonary disease' or 'No acute abnormality.' Because many similar examples exist in the retrieval pool, the model can often reproduce the expected impression style.

More difficult cases tend to involve subtle abnormalities, uncertainty expressions, nonpulmonary findings, or chronic conditions that must be distinguished from acute disease. Examples include mild atelectasis versus scarring, costophrenic angle blunting without definite effusion, possible rib fracture, cardiomegaly without acute disease, and hyperinflation with air trapping. These cases are challenging because the generated impression must preserve a specific abnormality while avoiding unnecessary or unsupported additions.

This pattern suggests that retrieval-guided prompting is effective when the target case has close lexical and semantic neighbors in the reference collection, but remains less reliable when the clinically important information is subtle, uncertain, or expressed through low-frequency phrasing. This limitation is consistent with the dataset-level analysis: when reference examples are repetitive and close to the target report, retrieval-guided generation is easier; when findings are rare, ambiguous, or weakly represented in the reference collection, the model has less reliable evidence to follow.

7.12. Why Compact Open-Source Models Can Compete

The results indicate that compact open-source models can achieve strong performance when the surrounding system is carefully designed. This does not imply that model scale is unimportant. Instead, it shows that retrieval, prompt construction, and feedback can substantially improve smaller models. In clinical summarization, this is especially valuable because compact open-source models are easier to reproduce, inspect, and deploy locally than large proprietary systems.

The strongest results are achieved not by changing model parameters but by changing the information supplied to the model. This supports the thesis argument that system-level design can be as important as model size for specialized clinical generation tasks. The improvements obtained through semantic-lexical retrieval, projection-guided multi-signal ranking, hard negative examples, ordering analysis, and iterative refinement show that careful control over retrieval and prompt construction can produce large gains without language model fine-tuning.

The controlled comparison with GPT-5.4 provides additional evidence for this interpretation. When retrieval, prompt construction, iterative generation, and output selection are held constant, GPT-5.4 produces the stronger overall result on OpenI. It achieves higher ROUGE-1, ROUGE-L, CheXbert-F1, and complete RadGraph-F1, although Llama3 retains a slightly higher ROUGE-2 score. This shows that generator capability continues to affect performance even when both models receive the same retrieval evidence.

The MIMIC-CXR comparison presents a more balanced pattern. ROUGE-1 is nearly identical for the two generators, while Llama3 obtains higher ROUGE-2, ROUGE-L, and complete RadGraph-F1. GPT-5.4 obtains the higher CheXbert-F1 score. Therefore, the proprietary generator does not uniformly outperform the compact open-source model across lexical and clinical evaluation dimensions. The relative advantage depends on both the dataset and the metric used for evaluation.

These findings do not imply that Llama3 and GPT-5.4 are generally equivalent. The comparison is limited to the datasets, retrieval configurations, and prompting procedures examined in this thesis. However, it shows that carefully selected in-context evidence can substantially reduce the practical performance gap between compact open-source

and proprietary generators. This is particularly visible on MIMIC-CXR, where Llama3 remains equal or stronger on several metrics despite its smaller and fully open-source design.

This finding has practical significance for research environments with limited computational resources. Instead of requiring language model fine-tuning or large-scale model adaptation, the proposed framework uses the reference collection as a retrieval pool and dynamically constructs prompts for each evaluation case. This makes the approach more accessible and easier to reproduce than systems that depend on proprietary APIs or large-scale model adaptation.

7.13. Why a Text-Only Framework without Language Model Fine-Tuning?

The thesis intentionally focuses on a text-only framework without language model fine-tuning. This choice does not deny the value of visual information in radiology. Multimodal systems can be highly effective when medical images and robust image encoders are available. However, multimodal pipelines also introduce additional complexity, such as image preprocessing, visual feature extraction, model alignment, and higher computational cost. In contrast, the findings-to-impression setting is directly applicable when textual findings sections are available.

Avoiding language model fine-tuning also supports reproducibility. Fine-tuning can improve performance, but it introduces additional hyperparameters, computational cost, and sensitivity to the fine-tuning data. The framework developed in this thesis avoids updating the language model parameters and instead focuses on retrieval, ranking, prompt construction, and candidate selection. This makes the approach easier to evaluate across different compact open-source language models.

The text-only design also clarifies the contribution of the thesis. Since no medical image features or proprietary APIs are used, the improvements can be attributed more directly to retrieval design and prompt construction. The second retrieval mechanism does estimate a lightweight projection mapping for retrieval, and it can use structured clinical features as auxiliary retrieval signals. However, these components do not fine-tune the language model and do not use target-side clinical labels during inference.

7.14. Design Choices Excluded from the Final Generation Framework

Several design choices were intentionally kept outside the final generation framework in order to preserve the findings-to-impression setting and maintain comparability. These include medical image features, language model fine-tuning, and target-side clinical label retrieval.

Medical image features were not included because the task studied in this thesis is findings-to-impression generation rather than image-to-report generation. Adding visual encoders would change the task setting and make it harder to isolate the effect of retrieval and prompting over report text.

Language model fine-tuning was also not used. Although fine-tuning can adapt a model to radiology language, it introduces additional training-related variables, computational cost, and sensitivity to the fine-tuning data. The final framework instead emphasizes prompt-level adaptation through retrieval and candidate selection.

Target-side clinical label retrieval was avoided because it can create comparability concerns under the findings-only generation setting. If labels derived from the target case are used during example selection, the retrieval stage may indirectly access information that would not be available from the findings text alone. The framework in this thesis therefore selects examples from the available findings-side information and auxiliary retrieval features, without using the target impression or target-side labels at inference time.

Structured clinical features are used only as auxiliary retrieval signals in the projection-guided strategy. They do not replace text-based retrieval, and they are not used as privileged target-side labels. This keeps the framework aligned with the findings-to-impression task while still allowing structured clinical information to improve retrieval ranking.

7.15. Limitations

The study has several limitations. First, evaluation is still largely based on automatic metrics. ROUGE captures lexical overlap, BERTScore captures contextual semantic similarity, CheXbert-F1 captures agreement over clinical observation labels, and

RadGraph-F1 captures agreement over extracted clinical entities and relations. These metrics provide complementary information, but none of them fully measures clinical correctness, factual consistency, or usefulness in real reporting workflows. Two generated impressions may receive similar automatic scores while differing in safety, completeness, or clinical acceptability.

Second, the main method development and ablation experiments are conducted on chest X-ray reports from OpenI and MIMIC-CXR. The additional MIMIC-III evaluation extends the analysis to six CT and MR domains, but these domains are derived from a single clinical database and do not represent the full diversity of radiology reporting. Therefore, the findings may not directly generalize to other institutions, languages, imaging modalities, anatomical regions, or reporting styles.

Third, the framework does not incorporate medical images. As a result, it cannot correct or verify information that is absent, ambiguous, or incorrectly described in the findings section. The method assumes that the findings section already contains the relevant clinical observations. This is appropriate for findings-to-impression generation, but it limits the system's ability to function as a full image-to-report generation pipeline.

Fourth, the quality of the generated impression still depends on the reference collection. When the target case has close semantic, lexical, or projection-guided neighbors, the system often performs well. When the target case involves rare findings, uncertain language, unusual report structure, or weakly represented clinical patterns, retrieval may provide less helpful examples. In such cases, iterative refinement may improve phrasing but cannot fully compensate for weak retrieval evidence.

Fifth, the structured clinical features used in the projection-guided strategy are automatically extracted and may contain errors. These features are useful as auxiliary retrieval signals, but they should not be interpreted as gold-standard clinical annotations. Extraction errors, missed entities, or incorrect assertion status may affect retrieval ranking.

Finally, the method uses automatic prompt-based feedback rather than expert clinical judgment. Although this keeps the system lightweight and reproducible, it means that the refinement process is guided by similarity signals rather than direct clinical validation. Expert review would be necessary to assess safety, factuality, and clinical usefulness before any practical deployment. These limitations define the intended scope of the present study: a reproducible findings-to-impression generation framework based

on text-only retrieval-guided prompting, rather than a full clinical deployment system or an image-to-report generation pipeline.

7.16. Clinical and Practical Implications

The framework should be understood as a research system for radiology impression generation rather than a clinical decision support tool. Its main practical implication is that compact open-source language models can be substantially improved through transparent retrieval design. In settings where proprietary APIs or large-scale fine-tuning are not practical, retrieval-guided prompting offers a reproducible alternative.

The results suggest that clinical summarization systems do not necessarily require increasingly larger models to improve performance. Instead, retrieval quality, prompt design, example ordering, projection-guided ranking, and contrastive guidance can have substantial effects. This is important for research groups and institutions with limited computational resources, because the proposed framework can be implemented without language model fine-tuning or proprietary infrastructure.

However, any clinical deployment would require expert validation, safety evaluation, and integration with institutional workflows. In particular, omission of subtle findings and overgeneralization toward common negative templates must be carefully evaluated before practical use. The system may be useful as an assistive drafting or research tool, but final clinical interpretation must remain under the responsibility of qualified radiology professionals.

CHAPTER 8

CONCLUSION

In this thesis, radiology impression generation was investigated as a retrieval-guided text generation problem. The study primarily focused on generating concise and clinically meaningful impressions from the findings sections of chest X-ray reports using compact open-source language models. Rather than depending on proprietary systems, multimodal encoders, or language model fine-tuning, the proposed framework combines open-source decoder-only models with retrieval design, prompt construction, hard negative prompting, example ordering, and iterative prompt feedback.

The experiments show that zero-shot prompting is insufficient for this task, and that random-shot prompting improves performance only to a limited extent. These findings indicate that compact decoder-only models are sensitive to prompt structure, but also require examples that are relevant to the target case. Retrieval-guided prompting provides this relevance by selecting reference findings-impression pairs from the reference collection. As a result, the model receives both content guidance and stylistic guidance, which helps it generate impressions that are more closely aligned with reference summaries.

A central finding of the thesis is that semantic retrieval and lexical refinement provide complementary forms of guidance. Sentence embeddings help identify reports that are conceptually similar to the candidate findings, even when the wording differs. However, radiology reports often contain clinically important terms whose presence, absence, or negation can change the meaning of the impression. Lexical refinement helps preserve these terms by selecting examples that share important wording with the target findings. The semantic-lexical retrieval strategy therefore provides a strong and relatively transparent baseline for retrieval-guided clinical summarization.

The thesis further shows that retrieval can be improved by moving beyond direct findings-side similarity. The projection-guided multi-signal retrieval strategy adds target-oriented and feature-level matching signals to the retrieval process. It combines projected summary match, findings-to-findings match, findings-to-impression match, key concept carry-over, and structured clinical matching. The results show that this second retrieval mechanism improves over the semantic-lexical strategy on both OpenI and MIMIC-CXR.

The weight ablations further show that findings-to-findings matching provides the strongest standalone foundation, while projected summary matching, moderate key concept carry-over, and low-weight structured clinical evidence provide complementary improvements. This indicates that effective example selection for impression generation should consider not only which findings sections are similar, but also which reference impressions are likely to provide useful summary patterns.

The experimental results also demonstrate that prompt composition strongly affects generation quality. Few-shot prompting improves performance when examples are carefully selected, but the number and ordering of examples must be configured deliberately. The experiments show that example ordering is model-specific: different compact open-source models respond differently to the position of the most relevant examples in the prompt. This confirms that example ordering is not a minor formatting choice, but an important design factor in retrieval-guided prompting.

In addition to positive examples, the framework evaluates the use of hard negative examples. The results show that a small number of hard negatives can improve performance by exposing the model to semantically plausible but clinically misleading examples. This contrastive prompting strategy helps the model distinguish between appropriate and inappropriate findings-to-impression mappings. However, adding too many hard negatives reduces the benefit, suggesting that negative context must be used carefully so that it does not weaken the influence of highly relevant positive examples.

Iterative refinement also contributes to the final performance, but its role is best understood as prompt-level candidate search rather than monotonic correction. The model does not necessarily improve step by step in every iteration. Instead, repeated prompting produces multiple plausible candidates, and final selection identifies the strongest output among them. The experiments further show that feedback computation benefits from a focused subset of high-ranked positive examples. Using the top five retrieved positives for refinement feedback gives stronger results than relying on a single example or using the full positive set, which suggests that feedback should balance relevance and diversity.

The cross-dataset results provide an important perspective on the difficulty and generalizability of the task. OpenI is a useful public benchmark for comparison with prior findings-to-impression generation studies, but it is relatively compact and repetitive. MIMIC-CXR provides a larger and more heterogeneous chest X-ray evaluation setting,

with longer reports, more diverse impressions, and lower direct impression reuse. The additional MIMIC-III evaluation extends the analysis to six CT and MR domains covering different anatomical regions. The proposed framework achieves strong results on OpenI, remains competitive on MIMIC-CXR, especially in clinical label agreement, and can be applied to the evaluated CT and MR domains without language model fine-tuning. However, the substantial performance differences among the MIMIC-III domains show that cross-modality generalization remains sensitive to anatomical scope, reporting style, and the availability of relevant reference examples.

The comparative analysis shows that the proposed approach is competitive with prior graph-based, multimodal, and proprietary LLM-based systems. On OpenI, both retrieval strategies achieve strong ROUGE scores, and the projection-guided strategy gives the best overall performance. On MIMIC-CXR, the proposed method does not achieve the highest ROUGE scores, but it obtains strong CheXbert-F1 compared with directly comparable systems. This suggests that the framework is particularly effective at preserving clinical observation-level content, even when the generated wording differs from the reference impression. Comparisons with systems that use target-side clinical labels or substantially different dataset splits should be interpreted with caution.

The controlled generator comparison further shows that the advantage of a proprietary model is dataset-dependent. Under the same retrieval and prompting conditions, GPT-5.4 produces the stronger overall result on OpenI, while Llama3 remains nearly identical in ROUGE-1 and achieves higher ROUGE-2, ROUGE-L, and complete RadGraph-F1 on MIMIC-CXR. GPT-5.4 obtains higher CheXbert-F1 on both datasets. These results show that generator choice affects the balance among lexical, label-based, and structured clinical agreement, but changing to a proprietary generator does not uniformly improve all evaluation dimensions.

The qualitative analysis clarifies both the strengths and limitations of the framework. The system performs well on common negative cases where the expected impression follows frequent templates such as the absence of acute cardiopulmonary disease. It also benefits from retrieval and iterative refinement when the generated output needs to be made more concise. However, the model may still omit subtle abnormalities, uncertain findings, incidental findings, nonpulmonary findings, or chronic conditions when these are embedded within otherwise common report patterns. These observations show that

retrieval-guided prompting is effective, but it does not fully solve the problem of clinical completeness.

The study has several limitations. Evaluation is based primarily on automatic metrics, including ROUGE, BERTScore, CheXbert-F1, and RadGraph-F1. These metrics provide complementary information, but they cannot fully measure clinical safety, factual consistency, or practical usefulness in real reporting workflows. The main method development and ablation experiments focus on chest X-ray reports, while the additional MIMIC-III evaluation covers only six CT and MR domains from a single clinical database. The framework also assumes that the findings section already contains the relevant clinical observations. Since medical images are not used, the system cannot verify or correct information that is missing or incorrect in the findings text. In addition, the quality of the generated impression remains dependent on the quality and coverage of the reference collection used for retrieval. These limitations define the scope of the present study as a text-only findings-to-impression generation framework rather than a clinical deployment system.

In summary, this thesis shows that compact open-source language models can be substantially improved for radiology impression generation through retrieval-guided prompting. By combining semantic-lexical retrieval, projection-guided multi-signal matching, hard negative examples, example ordering, and iterative prompt feedback, the proposed framework achieves strong performance without proprietary APIs, multimodal encoders, or language model fine-tuning. The results indicate that transparent and reproducible system design can provide a practical alternative to larger and less accessible clinical summarization pipelines.

BIBLIOGRAPHY

Artsi, Y., E. Klang, J. D. Collins, B. S. Glicksberg, P. Korfiatis, G. N. Nadkarni, and V. Sorin (2025). Large language models in radiology reporting—a systematic review of performance, limitations, and clinical implications. *medRxiv*, 2025–03.

Asgari, E., N. Montaña-Brown, M. Dubois, S. Khalil, J. Balloch, J. A. Yeung, and D. Pimenta (2025). A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine 8*(1), 274.

Bahdanau, D., K. Cho, and Y. Bengio (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Balachandran, A. (2024). Medembed: Medical-focused embedding models. https://github.com/abhinand5/MedEmbed.

Brown, P. F., V. J. Della Pietra, P. V. Desouza, J. C. Lai, and R. L. Mercer (1992). Class-based n-gram models of natural language. *Computational linguistics 18*(4), 467–480.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems 33*, 1877–1901.

Char, D. S., N. H. Shah, and D. Magnus (2018). Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine 378*(11), 981.

Chen, Z., A. H. Cano, A. Romanou, A. Bonnet, K. Matoba, F. Salvi, M. Pagliardini, S. Fan, A. Köpf, A. Mohtashami, et al. (2023). Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Chen, Z., Y. Song, T.-H. Chang, and X. Wan (2020). Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1439–1449.

Chen, Z., M. Varma, X. Wan, C. Langlotz, and J.-B. Delbrouck (2023). Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 469–484.

Deerwester, S., S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman (1990). Indexing by latent semantic analysis. *Journal of the American society for information science 41*(6), 391–407.

Delbrouck, J.-B., P. Chambon, C. Bluethgen, E. Tsai, O. Almusa, and C. Langlotz (2022). Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4348–4360.

Delbrouck, J.-B., P. Chambon, Z. Chen, M. Varma, A. Johnston, L. Blankemeier, D. Van Veen, T. Bui, S. Truong, and C. Langlotz (2024). Radgraph-xl: A large-scale expert-annotated dataset for entity and relation extraction from radiology reports. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 12902–12915.

Delbrouck, J.-B., M. Varma, P. Chambon, and C. Langlotz (2023). Overview of the radsum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *Proceedings of the 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pp. 478–482.

Delbrouck, J.-B., C. Zhang, and D. Rubin (2021). Qiai at mediqa 2021: Multimodal radiology report summarization. In *Proceedings of the 20th workshop on biomedical language processing*, pp. 285–290.

Demner-Fushman, D., M. D. Kohli, M. B. Rosenman, S. E. Shooshan, L. Rodriguez, S. Antani, G. R. Thoma, and C. J. McDonald (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association 23*(2), 304–310.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019*

conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), pp. 4171–4186.

(ESR), E. S. O. R. (2011). Good practice for radiological reporting. guidelines from the european society of radiology (esr). *Insights into imaging 2*(2), 93–96.

Gao, X. and K. Das (2024). Customizing language model responses with contrastive in-context learning. In *Proceedings of the aaai conference on artificial intelligence*, Volume 38, pp. 18039–18046.

Gershanik, E. F., R. Lacson, and R. Khorasani (2011). Critical finding capture in the impression section of radiology reports. In *AMIA Annual Symposium Proceedings*, Volume 2011, pp. 465.

Gharebagh, S. S., N. Goharian, and R. Filice (2020). Attend to medical ontologies: Content selection for clinical abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1899–1905.

Grattafiori, A., A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. (2024). The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Gu, Y., R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon (2021). Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH) 3*(1), 1–23.

Guo, D., D. Yang, H. Zhang, J. Song, P. Wang, Q. Zhu, R. Xu, R. Zhang, S. Ma, X. Bi, et al. (2025). Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature 645*(8081), 633–638.

Guu, K., K. Lee, Z. Tung, P. Pasupat, and M. Chang (2020). Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR.

Haibe-Kains, B., G. A. Adam, A. Hosny, F. Khodakarami, M. A. Q. C. M. S. B. of Directors Shraddha Thakkar 35 Kusko Rebecca 36 Sansone Susanna-Assunta 37

Tong Weida 35 Wolfinger Russ D. 38 Mason Christopher E. 39 Jones Wendell 40 Dopazo Joaquin 41 Furlanello Cesare 42, L. Waldron, B. Wang, C. McIntosh, A. Goldenberg, A. Kundaje, et al. (2020). Transparency and reproducibility in artificial intelligence. *Nature 586*(7829), E14–E16.

Hartung, M. P., I. C. Bickle, F. Gaillard, and J. P. Kanne (2020). How to create a great radiology report. *Radiographics 40*(6), 1658–1670.

Hoerl, A. E. and R. W. Kennard (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics 12*(1), 55–67.

Hu, J., Z. Chen, Y. Liu, X. Wan, and T.-H. Chang (2023). Improving radiology summarization with radiograph and anatomy prompts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 12066–12080.

Hu, J., J. Li, Z. Chen, Y. Shen, Y. Song, X. Wan, and T.-H. Chang (2021). Word graph guided summarization for radiology findings. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 4980–4990.

Hu, J., Z. Li, Z. Chen, Z. Li, X. Wan, and T.-H. Chang (2022). Graph enhanced contrastive learning for radiology findings summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4677–4688.

Irvin, J., P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, Volume 33, pp. 590–597.

Izacard, G. and E. Grave (2021). Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pp. 874–880.

Jain, S., A. Agrawal, A. Saporta, S. Q. Truong, D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. P. Lungren, A. Y. Ng, et al. (2021). Radgraph: Extracting clinical entities and relations from radiology reports. *arXiv preprint arXiv:2106.14463*.

Jing, B., P. Xie, and E. Xing (2018). On the automatic generation of medical imaging reports. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 2577–2586.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pp. 137–142. Springer.

Johnson, A. E., T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-y. Deng, R. G. Mark, and S. Horng (2019). Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data 6*(1), 317.

Karpukhin, V., B. Oguz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih (2020). Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781.

Kocak, B., A. H. Yardimci, S. Yuzkan, A. Keles, O. Altun, E. Bulut, O. N. Bayrak, and A. A. Okumus (2023). Transparency in artificial intelligence research: a systematic review of availability items related to open science in radiology and nuclear medicine. *Academic Radiology 30*(10), 2254–2266.

Kryściński, W., B. McCann, C. Xiong, and R. Socher (2020). Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 9332–9346.

Kuzi, S., M. Zhang, C. Li, M. Bendersky, and M. Najork (2020). Leveraging semantic and lexical matching to improve the recall of document retrieval systems: A hybrid approach. *arXiv preprint arXiv:2010.01195*.

Larson, D. B., A. J. Towbin, R. M. Pryor, and L. F. Donnelly (2013). Improving consistency in radiology reporting through the use of department-wide standardized structured reporting. *Radiology 267*(1), 240–250.

Lee, D., S.-w. Hwang, K. Lee, S. Choi, and S. Park (2023). On complementarity objectives for hybrid retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13357–13368.

Lee, J., W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics 36*(4), 1234–1240.

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer (2020). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 7871–7880.

Lewis, P., E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems 33*, 9459–9474.

Li, J., T. Zhou, Z. Zhou, X. Wei, H. Song, Z. Wang, Y. Chen, and H. Lv (2025). Experience-guided multi-agent interpretable framework for radiology report summarization. *Computer Methods and Programs in Biomedicine*, 109078.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81.

Liu, H., C. Li, Q. Wu, and Y. J. Lee (2023). Visual instruction tuning. *Advances in neural information processing systems 36*, 34892–34916.

Liu, N. F., K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang (2024). Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics 12*, 157–173.

Lu, X., M. Bartolo, S. Riedel, and P. Stenetorp (2022). Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Luo, G. and Y. Arase (2025). Medsummrag: Domain-specific retrieval for medical summarization. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pp. 27–33.

Luo, Z., Z. Jiang, M. Wang, X. Cai, D. Gao, and L. Yang (2025). Chatgpt based contrastive learning for radiology report summarization. *Expert Systems with Applications 267*, 125827.

Ma, C., Z. Wu, J. Wang, S. Xu, Y. Wei, Z. Liu, F. Zeng, X. Jiang, L. Guo, X. Cai, et al. (2024). An iterative optimizing framework for radiology report summarization with chatgpt. *IEEE Transactions on Artificial Intelligence 5*(8), 4163–4175.

Maynez, J., S. Narayan, B. Bohnet, and R. McDonald (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 1906–1919.

Min, S., X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer (2022). Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 11048–11064.

Miura, Y., Y. Zhang, E. Tsai, C. Langlotz, and D. Jurafsky (2021). Improving factual completeness and consistency of image-to-text radiology report generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5288–5304.

Nenkova, A. and K. McKeown (2011). Automatic summarization. *Foundations and Trends® in Information Retrieval 5*(2–3), 103–233.

Ng, J.-P. and V. Abrecht (2015). Better summarization evaluation with word embeddings for rouge. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 1925–1930.

Nishio, M., T. Matsunaga, H. Matsuo, M. Nogami, Y. Kurata, K. Fujimoto, O. Sugiyama, T. Akashi, S. Aoki, and T. Murakami (2024). Fully automatic summarization of radiology reports using natural language processing with large language models. *Informatics in Medicine Unlocked 46*, 101465.

Nobel, J. M., K. van Geel, and S. G. Robben (2022). Structured reporting in radiology: a systematic review to explore its potential. *European radiology 32*(4), 2837–2854.

Nooralahzadeh, F., N. P. Gonzalez, T. Frauenfelder, K. Fujimoto, and M. Krauthammer (2021). Progressive transformer-based generation of radiology reports. In *Findings of the association for computational linguistics: EMNLP 2021*, pp. 2824–2832.

Price, W. N. and I. G. Cohen (2019). Privacy in the age of medical big data. *Nature medicine 25*(1), 37–43.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research 21*(140), 1–67.

Reimers, N. and I. Gurevych (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992.

Rubin, O., J. Herzig, and J. Berant (2022). Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 conference of the North American chapter of the association for computational linguistics: human language technologies*, pp. 2655–2671.

Rush, A. M., S. Chopra, and J. Weston (2015). A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 379–389.

Salton, G. and C. Buckley (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management 24*(5), 513–523.

See, A., P. J. Liu, and C. D. Manning (2017). Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1073–1083.

Smit, A., S. Jain, P. Rajpurkar, A. Pareek, A. Y. Ng, and M. Lungren (2020). Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 1500–1519.

Sun, Z., H. Ong, P. Kennedy, L. Tang, S. Chen, J. Elias, E. Lucas, G. Shih, and Y. Peng (2023). Evaluating gpt-4 on impressions generation in radiology reports. *Radiology 307*(5), e231259.

Sutskever, I., O. Vinyals, and Q. V. Le (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems 27*.

Takeshita, S., S. P. Ponzetto, and K. Eckert (2024). Rouge-k: Do your summaries have keywords? In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pp. 69–79.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). Attention is all you need. *Advances in neural information processing systems 30*.

Wang, W., F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems 33*, 5776–5788.

Wang, X., Y. Peng, L. Lu, Z. Lu, and R. M. Summers (2018). Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9049–9058.

Zhang, J., Y. Zhao, M. Saleh, and P. Liu (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pp. 11328–11339. PMLR.

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi (2019). Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhang, Y., D. Y. Ding, T. Qian, C. D. Manning, and C. P. Langlotz (2018). Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pp. 204–213.

Zhang, Y., S. Feng, and C. Tan (2022). Active example selection for in-context learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9134–9148.

Zhao, Z., E. Wallace, S. Feng, D. Klein, and S. Singh (2021). Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pp. 12697–12706. Pmlr.

Zhu, Z., Y. Zhang, X. Zhuang, F. Zhang, Z. Wan, Y. Chen, Q. QingqingLong, Y. Zheng, and X. Wu (2025). Can we trust ai doctors? a survey of medical hallucination in large language and large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6748–6769.

APPENDIX. PROMPT TEMPLATES

This appendix reports the prompt templates used in the experiments. The overall prompt structure follows the iterative in-context prompting strategy introduced in ImpressionGPT (Ma et al. 2024) and is implemented using system, user, and assistant messages. Each prompt consists of a system instruction, retrieved GOOD demonstrations, retrieved BAD demonstrations, an initial generation request, and optional iterative refinement messages.

Two wording variants are used. OpenI and MIMIC-CXR use a chest X-ray-specific template, whereas the MIMIC-III CT and MR evaluation uses a modality-neutral radiology template. Both variants preserve the same overall message structure, the same GOOD and BAD demonstration format, and the same iterative feedback logic. However, modality references and some regeneration instructions differ slightly. The prompt text below is reproduced as used in the experiments.

A.1. System Instructions

A.1.1. Chest X-ray System Instruction

The following system message is used for OpenI and MIMIC-CXR.

Role: system

> You are a chest radiologist that identify the main findings and diagnosis or impression based on the given FINDINGS section of the chest X-ray report, which details the radiologists' assessment of the chest X-ray image. Please ensure that your response is concise and does not exceed the length of the FINDINGS.

A.1.2. Modality-Neutral System Instruction

The following modality-neutral system message is used for the MIMIC-III CT and MR evaluation.

Role: system

> You are a radiologist who writes the IMPRESSION section from the given FINDINGS section of a radiology report. Please keep the response concise, clinically faithful, and not longer than the FINDINGS.

A.2. GOOD Example Conditioning

The retrieved positive examples are introduced as GOOD demonstrations. The main retrieval-guided configuration uses 15 positive examples. The header message is shared by the chest X-ray and modality-neutral prompt variants.

Role: user

> Here are some GOOD examples. Please learn how to write IMPRESSION from these examples, and pay particular attention to the consistent use of phrases in these below examples.

A.2.1. Chest X-ray GOOD Example Format

The following user and assistant messages are repeated for each retrieved positive example in the OpenI and MIMIC-CXR experiments.

Role: user

What are the main findings and diagnosis or impression based on the given Finding in chest X-ray report:
FINDINGS:
{Reference Finding}

Role: assistant

IMPRESSION:
{Reference Impression}

A.2.2. Modality-Neutral GOOD Example Format

The following user and assistant messages are repeated for each retrieved positive example in the MIMIC-III CT and MR evaluation.

Role: user

What are the main findings and diagnosis or impression based on the given FINDINGS section of this radiology report:
FINDINGS:
{Reference Finding}

Role: assistant

IMPRESSION:
{Reference Impression}

A.3. BAD Example Conditioning

The GOOD demonstrations are followed by two retrieved hard negative examples in the main retrieval-guided configuration. These examples are presented as BAD demonstrations that contain similar wording but a different clinical context. The same header and example format are used in the chest X-ray and modality-neutral settings.

Role: user

Here are some BAD examples that you should AVOID. These examples have different clinical findings despite similar text:

The following user and assistant messages are repeated for each retrieved hard negative example.

Role: user

FINDINGS:
{Reference Finding}

This IMPRESSION is INCORRECT for these findings:

Role: assistant

IMPRESSION:
{Reference Impression}

(This is a BAD example - different clinical context)

A.4. Initial Generation Prompts

Initial generation begins after the GOOD and BAD demonstrations have been added to the message sequence. The same initial generation stage is also used before iterative refinement begins.

A.4.1. Chest X-ray Initial Generation Prompt

The following transition message is used for OpenI and MIMIC-CXR.

Role: user

Now, based on the GOOD examples above (and avoiding the patterns shown in BAD examples if any), please write an impression for these findings:

The candidate findings are then provided using the following message.

Role: user

What are the main findings and diagnosis or impression based on the given Finding in chest X-ray report:
FINDINGS:
{Candidate Finding}

A.4.2. Modality-Neutral Initial Generation Prompt

The following transition message is used for the MIMIC-III CT and MR evaluation.

Role: user

Now, based on the GOOD examples above, please write an impression for these findings:

The candidate findings are then provided using the following message.

Role: user

What are the main findings and diagnosis or impression based on the given FINDINGS section of this radiology report:
FINDINGS:
{Candidate Finding}

A.5. Iterative Refinement Prompts

During iterative refinement, the candidate findings are presented again, followed by previously generated impressions selected as GOOD or BAD feedback. The regeneration instruction depends on whether the refinement step contains only a BAD impression, only a GOOD impression, or both.

A.5.1. Chest X-ray Candidate Query

The following candidate query is used during each refinement round for OpenI and MIMIC-CXR.

Role: user

> What are the main findings and diagnosis or impression based on the given Finding in chest X-ray report:
> FINDINGS:
> {Candidate Finding}

A.5.2. Modality-Neutral Candidate Query

The following candidate query is used during each refinement round for the MIMIC-III CT and MR evaluation.

Role: user

> What are the main findings and diagnosis or impression based on the given FINDINGS section of this radiology report:
> FINDINGS:
> {Candidate Finding}

A.5.3. BAD Impression Feedback

When a previous candidate is used as negative feedback, it is added through the following user and assistant messages. This format is shared by both prompt variants.

Role: user

> Below is a bad impression of the FINDINGS above:

Role: assistant

> IMPRESSION:
> {Bad Impression}

A.5.4. GOOD Impression Feedback

When a previous candidate is used as positive feedback, it is added through the following user and assistant messages. This format is shared by both prompt variants.

Role: user

> Below is an excellent impression of the FINDINGS above:

Role: assistant

> IMPRESSION:
> {Good Impression}

A.5.5. Chest X-ray Regeneration Instructions

A.5.5.1. BAD-Only Feedback

When only a BAD impression is available, the following regeneration request is used for OpenI and MIMIC-CXR.
Role: user

> Please give another impression of the FINDINGS above. It is important to note that your answer should avoid being consistent with the bad impression above. Please pay particular attention to the consistent use of phrases in the above examples

A.5.5.2. GOOD-Only Feedback

When only a GOOD impression is available, the following regeneration request is used.
Role: user

> This is an excellent impression, please give another impression in this format, and do not exceed the length of this impression. Please pay particular attention to the consistent use of phrases in the above examples

A.5.5.3. Combined GOOD and BAD Feedback

When both GOOD and BAD impressions are available, the following regeneration request is used.
Role: user

> Please give another impression of the FINDINGS above. It is important to note that your answer should avoid being consistent with the bad impression above, but should be consistent with the excellent impression above, and do not exceed the length of the excellent impression. And please pay particular attention to the consistent use of phrases in the above examples

A.5.6. Modality-Neutral Regeneration Instructions

A.5.6.1. BAD-Only Feedback

When only a BAD impression is available, the following regeneration request is used for the MIMIC-III CT and MR evaluation.
Role: user

> Please give another impression of the FINDINGS above. It is important to avoid being consistent with the bad impression above. Please pay particular attention to the consistent use of phrases in the good examples.

A.5.6.2. GOOD-Only Feedback

When only a GOOD impression is available, the following regeneration request is used.

Role: user

> This is an excellent impression. Please give another impression in this format, and do not exceed the length of this impression. Please pay particular attention to the consistent use of phrases in the good examples.

A.5.6.3. Combined GOOD and BAD Feedback

When both GOOD and BAD impressions are available, the following regeneration request is used.

Role: user

> Please give another impression of the FINDINGS above. It should avoid the bad impression above, stay consistent with the excellent impression above, and not exceed the length of the excellent impression.

A.6. Prompt Configuration and Reproducibility

The chest X-ray and modality-neutral variants use the same high-level prompting structure. Each complete retrieval-guided prompt contains 15 GOOD demonstrations and two BAD demonstrations before the candidate findings are presented. The iterative procedure uses previously generated candidates as GOOD or BAD feedback and updates the message sequence without modifying the language model parameters.

The prompt templates are applied through system, user, and assistant roles. OpenI and MIMIC-CXR use the chest X-ray-specific wording, whereas the six MIMIC-III CT and MR domains use the modality-neutral wording. The MIMIC-III variant preserves the same demonstration structure and feedback logic while removing chest X-ray-specific references and slightly simplifying the regeneration instructions. No additional language model fine-tuning is performed during initial generation or iterative refinement.

VITA

SERHAT CANER

Teaching and Professional Experience

- Research and Teaching Assistant, İzmir Institute of Technology, 2019–2026.
- Software Engineer, Accenture Turkey, 2016-2019.

Higher Education

- MSc in Computer Engineering, İzmir Institute of Technology, 2020.
- BSc in Computer Engineering, İzmir Institute of Technology, 2016.

Selected Publications

- Caner S, Erdoğmuş N, Erten Y.M. Performance analysis and feature selection for network-based intrusion detection with deep learning. Turkish Journal of Electrical Engineering and Computer Sciences. 2022; 30(3): 629–643. https://doi.org/10.55730/1300-0632.3802.
- Caner S. Accuracy and efficiency analysis for deep learning based intrusion detection systems. İzmir Institute of Technology, 2020. (MSc thesis)
- Yaşar D, Caner S, Yıldız A, Demirörs O. Çeviklik Değerlendirme (AgilityMod) ve Süreç Yetenek Değerlendirme (SPICE) Modelleri Saha Uygulamaları. UYMS. 2019.