Politecnico di Torino

Master's Thesis

Hybrid Deep Learning Framework for Summarizing Radiology Reports Using Domain-Specific NLP Techniques

Supervisor: Prof. Alessandro Aliberti
Co-Supervisor: Prof. Edoardo Patti

Author: Setareh Pourgholamali
Student ID: S309064

July 2025

Abstract

The increasing volume of radiology reports presents a critical need for automated, accurate summarization tools to support clinical efficiency and diagnostic clarity. This thesis proposes a hybrid deep learning framework for summarizing radiology reports, specifically tailored for the chest X-ray domain using the Indiana University Chest X-ray Collection. The objective is to generate fluent, accurate, and clinically meaningful impression-style summaries that align with radiologists' diagnostic language, thereby supporting efficient clinical decision-making and documentation.

The pipeline is composed of three sequential stages. First, an extractive summarization step selects key sentences from the findings section using BERT-based sentence embeddings and cosine similarity ranking. This ensures that structurally important and content-rich sentences are chosen. Second, a medical term filtering module based on the SciSpaCy model extracts only domain-relevant named entities (e.g., anatomical structures, diagnoses, conditions) from the extractive output. This filters out irrelevant or low-clinical-value information. Third, a BART transformer model, pre-trained and then fine-tuned on the filtered data, performs abstractive summarization to produce coherent, concise impressions resembling human-written radiology conclusions.

The system was trained and evaluated on processed subsets of the Indiana dataset with dedicated train, validation, and test splits. Evaluation was carried out using both ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L, ROUGE-Lsum) for lexical overlap and BERTScore for semantic equivalence. The fine-tuned BART model achieved a ROUGE-L score of **57.84** and a BERTScore F1 of **0.9293** on the test set, indicating a high degree of fluency and factual consistency. The extractive module, when evaluated independently, achieved a BERTScore F1 of **0.8553**, confirming that the inclusion of abstraction significantly enhances the quality and clinical alignment of the generated summaries.

Side-by-side qualitative comparisons of extractive, abstractive, and reference summaries further demonstrate the model's ability to paraphrase accurately, avoid hallucinations, and preserve medically relevant content. Generated outputs correctly captured diagnostic findings such as pleural effusions, cardiomegaly, and pulmonary abnormalities, and adhered closely to expert-written impressions in both tone and terminology.

This work contributes to the field of medical NLP by demonstrating the effectiveness of a domain-aware hybrid summarization system that balances semantic fidelity with language generation. The modular architecture allows for extensibility to other datasets and clinical specialties. Potential future improvements include multimodal fusion with imaging data, and reinforcement learning with factuality-based rewards.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In the age of data-driven healthcare, unstructured clinical texts such as radiology reports remain a rich yet underutilized source of information. As medical imaging grows in volume and complexity, radiologists are expected to produce detailed textual interpretations for each study — a task that is both time-consuming and cognitively demanding. These radiology reports play a critical role in diagnosis, treatment planning, and patient communication, but their complexity often makes them challenging to process and integrate efficiently into clinical workflows.

At the same time, recent breakthroughs in **Natural Language Processing (NLP)** have enabled unprecedented capabilities in machine reading, understanding, and generation of human language. Advanced models based on **transformers**, such as **BERT** and **BART**, have revolutionized the NLP landscape by achieving state-of-the-art performance across a range of tasks — including text summarization, question answering, and entity recognition. These models leverage self-attention mechanisms and large-scale pretraining to capture deep contextual semantics, making them particularly suitable for specialized domains like biomedical text.

This thesis brings these two worlds together: the critical need for efficient radiology report summarization in clinical settings, and the powerful capabilities of modern AI-based NLP techniques. We propose a hybrid summarization framework that applies advanced deep learning methods to the task of condensing radiology reports, producing concise and medically faithful summaries that are aligned with expert-written impressions.

Automatic summarization in this context is not a trivial problem. Radiology reports are often written in a telegraphic style, filled with abbreviations, medical terminology, and implicit references that require domain knowledge to interpret. Conventional NLP systems, trained on general news or conversational data, tend to perform poorly in such settings due to domain mismatch and lack of factual robustness. Moreover, in high-stakes applications like medicine, even minor factual inconsistencies in generated summaries can have serious clinical implications.

To address these challenges, our work builds a modular and interpretable NLP pipeline, combining the strengths of both extractive and abstractive summarization:

- **BERTSUM**, a variant of the BERT model, is used to identify and extract the most informative sentences from the report findings.

- These extracted sentences are filtered using **Named Entity Recognition (NER)** via **SciSpacy**, ensuring that only medically relevant terms such as diseases, anatomical structures, and conditions are retained.

- A **BART** model, fine-tuned on domain-specific radiology data, then generates fluent and coherent impression-style summaries from the filtered content.

By integrating domain-specific NLP techniques, including contextual embeddings, NER-driven filtering, and transformer-based generation, we develop a system that is both accurate and explainable. The hybrid approach enhances clinical interpretability and reduces hallucinations — a common issue in purely generative models — while maintaining fluency and domain correctness.

The dataset used in this study is the **Indiana University Chest X-Ray Collection**, a publicly available corpus of de-identified radiology reports annotated with `Findings` and `Impression` sections. This dataset allows for a clean supervised setup, where the findings serve as inputs and impressions serve as reference summaries. Our model is trained and evaluated on this dataset, enabling both quantitative and qualitative assessments of performance.

Key contributions of this thesis include:

- The design and implementation of a multi-stage, NLP-driven summarization pipeline tailored for clinical radiology text.

- Domain-specific fine-tuning of transformer models (BERTSUM and BART) to extract and generate high-quality medical summaries.

- Integration of biomedical NER to filter and emphasize clinically relevant content during the summarization process.

- Empirical evaluation of the hybrid model using ROUGE and BERTScore, along with side-by-side qualitative comparisons with expert-written summaries.

This research demonstrates that when state-of-the-art NLP techniques are carefully adapted to the medical domain, they can offer significant value in improving clinical documentation and information accessibility. The resulting framework is not only accurate and interpretable but also generalizable, paving the way for deployment in real-world radiology workflows and other biomedical NLP applications.

Chapter 2

Background and Related Work

2.1 Evolution of Natural Language Processing Techniques

Natural Language Processing (NLP) has undergone a profound transformation over the past several decades, transitioning from handcrafted symbolic systems to deep learning models capable of understanding and generating human language with remarkable fluency. This section provides a historical and technical overview of this evolution, laying the groundwork for the transformer-based methods used in this thesis.

2.1.1 Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) Networks

Recurrent Neural Networks (RNNs) are a class of neural networks designed to handle sequential data by maintaining a hidden state that captures information about previous elements in the sequence. The hidden state at time step t is computed as:

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

where x_t is the input at time t, h_{t-1} is the previous hidden state, and W_{hh}, W_{xh}, and b_h are learnable parameters.

However, RNNs suffer from the vanishing gradient problem, making it difficult to learn long-term dependencies. To address this, Long Short-Term Memory (LSTM) networks were introduced [4]. LSTMs incorporate gating mechanisms to control the flow of information:

$$\begin{aligned}
f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
\tilde{c}_t &= \tanh(W_c x_t + U_c h_{t-1} + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot \tanh(c_t)
\end{aligned}$$

Here, f_t, i_t, and o_t represent the forget, input, and output gates, respectively; c_t is the cell state; and $\odot$ denotes element-wise multiplication.

2.1.2 Attention Mechanisms

Despite the improvements offered by LSTMs, challenges remained in modeling long-range dependencies and parallelizing computations. To overcome these limitations, attention mechanisms were introduced. Bahdanau et al. [5] proposed an attention mechanism that allows the model to focus on relevant parts of the input sequence when generating each element of the output sequence.

The attention mechanism computes a context vector c_t as a weighted sum of encoder hidden states:

$$c_t = \sum_{i=1}^{T_x} \alpha_{ti} h_i$$

where the attention weights α_{ti} are computed using an alignment model:

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{T_x} \exp(e_{tk})}$$

$$e_{ti} = a(s_{t-1}, h_i)$$

Here, s_{t-1} is the decoder's previous hidden state, and a is a feedforward neural network that scores the relevance of h_i to s_{t-1}.

2.1.3 Self-Attention Mechanism

The self-attention mechanism, introduced in the Transformer model by Vaswani et al. [6], revolutionized how neural networks process sequential data. Unlike traditional RNN-based architectures, which process sequences in order and struggle with long-term dependencies, self-attention enables a model to directly relate any two positions in a sequence regardless of their distance. This is done by dynamically computing the relevance, or "attention," between every pair of tokens in the input.

At the heart of self-attention is the idea that each input token should be able to **"attend"** to other tokens in the sequence in order to gather contextually relevant information. For each token, three vectors are computed via learned linear projections:

- **Query** (Q): represents the token we are computing attention for.
- **Key** (K): represents the token being attended to.
- **Value** (V): the actual information associated with the token.

For a sequence of n input tokens, the input matrix $X \in R^{n \times d_{model}}$ is transformed into query, key, and value matrices:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V$$

where $W^Q, W^K, W^V \in R^{d_{model} \times d_k}$ are learnable weight matrices.

The attention score between two tokens is computed using the **scaled dot-product** of their query and key vectors:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

- The dot product QK^T gives a similarity score between each query and all keys.
- The division by $\sqrt{d_k}$ prevents large values that could push the softmax function into regions with very small gradients.
- The softmax function normalizes the scores into a probability distribution.
- The resulting weights are then applied to the value matrix V, producing a context-aware representation for each token.

Each output vector is thus a weighted sum of all value vectors in the sequence, where the weights reflect the importance of other tokens relative to the current one. This allows the network to **selectively attend** to relevant parts of the sequence when encoding each token, enabling rich contextual understanding.

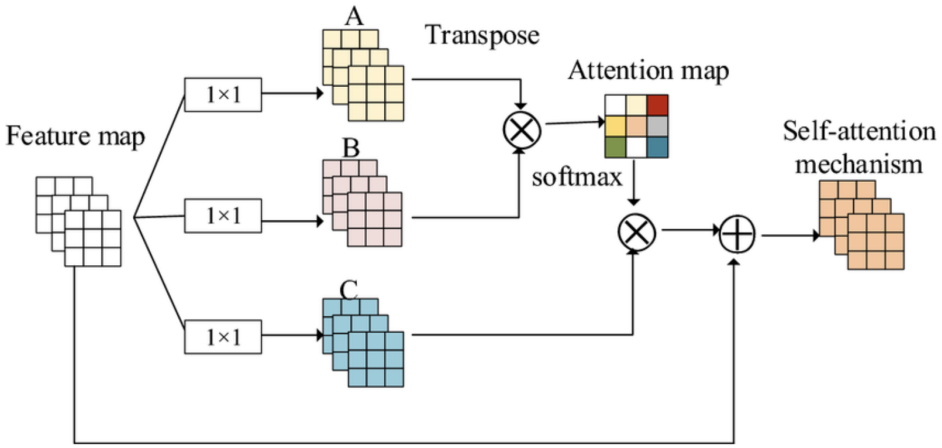

Figure 2.1: Self-Attention Mechanism as introduced in the Transformer architecture [6].

This mechanism is especially powerful because it:

- Operates in **parallel** for all tokens (unlike RNNs which are sequential),
- Easily scales to **long-range dependencies**, and
- Provides **interpretable attention maps**, revealing which tokens are influencing each other.

To further enhance the model's ability to capture complex relationships from multiple perspectives, the Transformer introduces *multi-head attention*, which we describe in the next subsection.

2.1.4 Multi-Head Attention

While the self-attention mechanism captures relationships between tokens by computing pairwise attention scores, a single attention head is often limited in the type of relationships it can model. To overcome this limitation, the Transformer architecture introduces the concept of *multi-head attention* [6], which enables the model to attend to information from multiple representation subspaces simultaneously.

The core idea is to run several self-attention operations — called *heads* — in parallel. Each head has its own set of learnable linear projections for queries (W_i^Q), keys (W_i^K), and values (W_i^V). This design allows different heads to focus on different aspects of the sequence, such as local versus global dependencies, syntactic structure, semantic relationships, or positional alignment.

For each attention head i, the computation is performed as:

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

Here, Q, K, and V are the query, key, and value matrices derived from the input sequence. The weight matrices W_i^Q, W_i^K, and W_i^V are specific to the i-th head and have dimensions $d_{\text{model}} \times d_k$, where $d_k = d_{\text{model}}/h$ and h is the number of heads.

Once each head computes its own self-attention output, all the resulting vectors are concatenated and passed through a final linear transformation using W^O:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

where $W^O \in R^{hd_k \times d_{\text{model}}}$ projects the concatenated outputs back to the original embedding dimension.

Why Multiple Heads?

Using multiple attention heads provides several advantages:

- **Diversity of focus**: Each head can attend to different parts of the input or capture different types of relationships. For example, one head might focus on syntactic dependencies, while another focuses on semantic roles.
- **Increased model capacity**: Multiple heads increase the representational capacity of the attention layer without dramatically increasing computational cost, as the dimensionality of each head is reduced proportionally.
- **Improved generalization**: By combining multiple views of the input, the model can learn more robust and generalized representations.

Example: Multi-Head Attention in Practice

Suppose we have an input sentence: "The cat sat on the mat." In a single-head attention scenario, the model might struggle to focus both on short-range dependencies (e.g., "cat" and "sat") and long-range dependencies (e.g., "The" and "mat") simultaneously. With multiple heads:

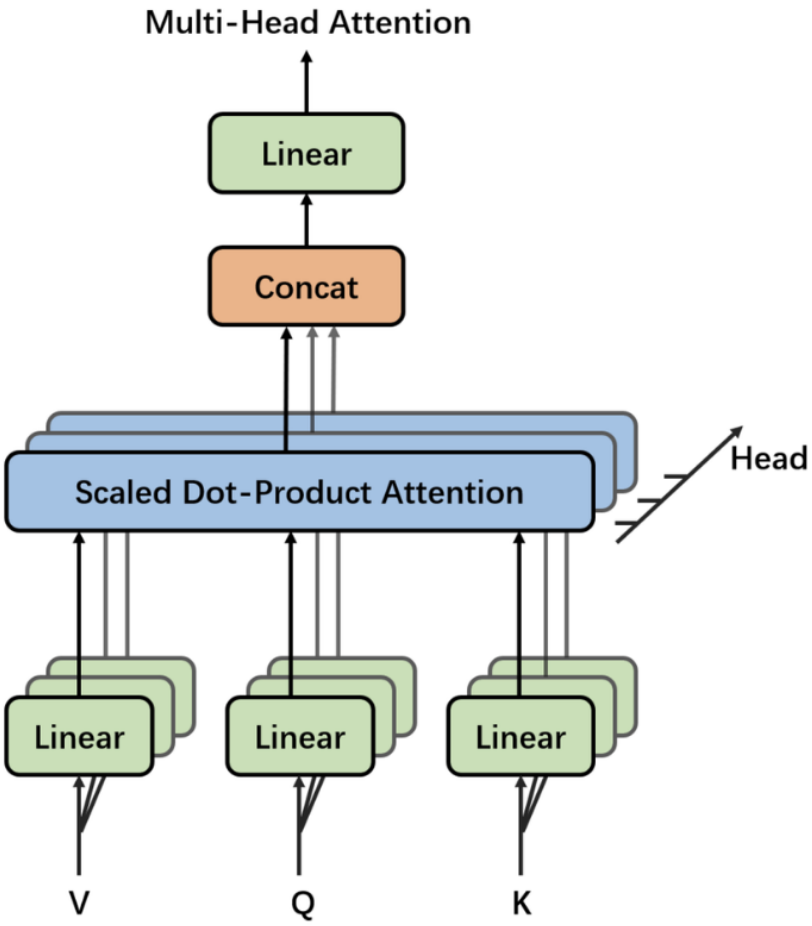

Figure 2.2: Structure of Multi-Head Attention Mechanism [6]. Each head independently computes scaled dot-product attention, and their outputs are concatenated and linearly projected.

- Head 1 might attend strongly to adjacent words (local context).
- Head 2 might capture noun–verb relationships.
- Head 3 might learn positional patterns.
- Head 4 might focus on function words or syntactic structure.

Each head processes the input from a different angle, and their aggregated outputs provide a richer, multi-faceted understanding of the sequence.

Mathematical Dimensions

To ensure dimensional consistency and computational efficiency:

- Let d_{model} be the dimensionality of the input and output embeddings (e.g., 512).
- Let h be the number of attention heads (e.g., 8).
- Then each head has dimension $d_k = d_v = d_{\text{model}}/h = 64$.
- Each head performs attention with matrices of shape $[n \times 64]$ for sequences of length n.
- After concatenation, the shape becomes $[n \times 512]$ again.

Relationship to Self-Attention

Multi-head attention is a generalization of self-attention. When applied within the Transformer encoder or decoder, each head operates independently but on the same input sequence. When applied across encoder and decoder (in encoder-decoder attention), the queries come from the decoder and keys/values come from the encoder.

Multi-head attention forms the backbone of all state-of-the-art Transformer models, including BERT, GPT, T5, and BART. It is the key to their ability to learn hierarchical, deep, and parallel representations of language.

2.1.5 Transformer Architecture

The Transformer architecture, introduced by Vaswani et al. [6], is a landmark innovation in the field of neural sequence modeling. Unlike previous models such as RNNs and LSTMs that relied on sequential operations, the Transformer uses only attention mechanisms to process input sequences in parallel, enabling much faster training and better modeling of long-range dependencies.

This architecture forms the backbone of many modern NLP models such as BERT, GPT, T5, and BART. It is composed of two main components: the encoder and the decoder, each consisting of a stack of N identical layers (typically $N = 6$).

Encoder

Each encoder layer is composed of two sub-layers:

1. **Multi-head self-attention mechanism**, which allows the encoder to attend to all positions in the input sequence and extract contextual relationships.
2. **Position-wise fully connected feed-forward network**, which applies two linear transformations with a ReLU activation in between.

Each sub-layer is wrapped in a residual connection followed by layer normalization. Formally, the output of each sub-layer can be written as:

$$\text{LayerNorm}(x + \text{Sublayer}(x))$$

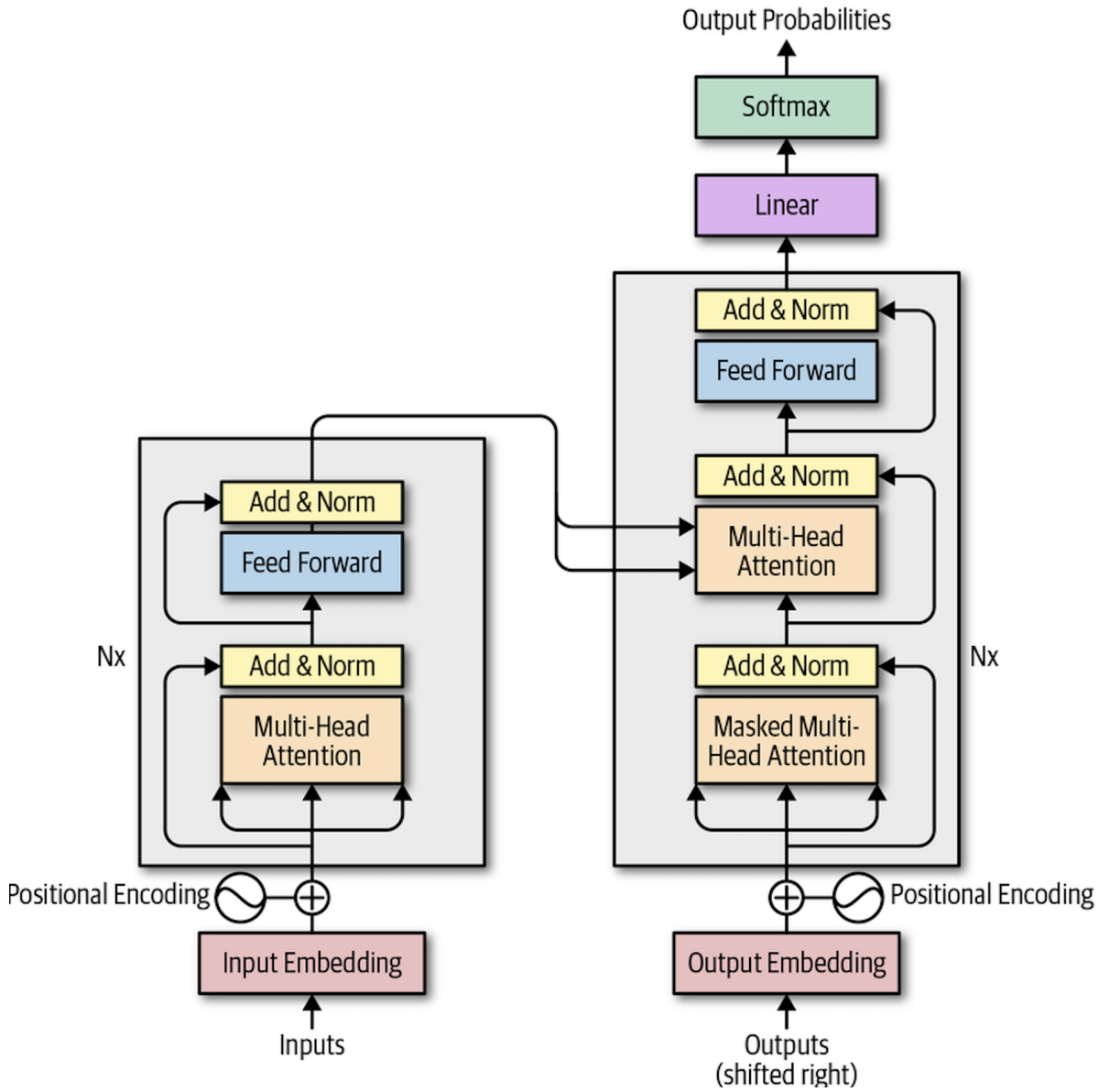

Figure 2.3: The original Transformer model architecture, consisting of encoder and decoder stacks with multi-head attention and feed-forward layers [6].

The feed-forward network in each layer is defined as:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

where W_1, W_2 are learned parameter matrices and b_1, b_2 are biases.

This design allows the encoder to learn deep hierarchical representations of the input sequence in a fully parallel and position-aware fashion.

Decoder

The decoder is slightly more complex. Each decoder layer contains three sub-layers:

1. **Masked multi-head self-attention**, which ensures that the prediction for a given position can only depend on earlier positions, preventing the decoder from "seeing" future tokens.

2. **Multi-head attention over the encoder's output**, allowing the decoder to attend to the encoder's representations of the input sequence.

3. **Feed-forward network**, identical to that in the encoder.

As in the encoder, each sub-layer is surrounded by residual connections and followed by layer normalization.

The masking in the first sub-layer is achieved by applying a mask to the attention weights before the softmax step. This mask sets the attention scores of future tokens to $-\infty$, effectively zeroing them out after the softmax.

Positional Encoding

Since the Transformer has no recurrence or convolution, it lacks any inherent notion of word order. To compensate, positional encodings are added to the input embeddings at the bottom of the model.

These encodings are fixed, deterministic, and use sine and cosine functions of different frequencies:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$
$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

Here:

- pos is the position index in the sequence
- i is the dimension index of the embedding vector
- d_{model} is the total embedding size

This method ensures that each position in the sequence has a unique encoding, and that similar positions have similar encodings. Moreover, the sinusoidal nature of these encodings allows the model to easily learn to attend by relative positions (e.g., word $i + 1$ relative to word i).

Putting It All Together

At a high level, the Transformer architecture works as follows:

- The input sequence is embedded and combined with positional encodings.
- The encoder processes the input in parallel, generating contextual representations.
- The decoder autoregressively generates the output sequence, attending to both its past outputs and the encoder's output.

This modular and scalable design allows the Transformer to achieve state-of-the-art performance in translation, summarization, question answering, and other sequence generation tasks.

2.1.6 Pretraining and Fine-Tuning Paradigm

The success of Transformer-based models in NLP has been largely driven by the emergence of the pretraining and fine-tuning paradigm [**?**, 7, 8]. Rather than training a model from scratch for every task, modern NLP systems begin with a general-purpose model trained on massive unlabeled corpora using unsupervised or self-supervised objectives. This pre-trained model is then fine-tuned on smaller labeled datasets specific to a downstream task, such as text classification, question answering, summarization, or named entity recognition.

Motivation for Pretraining

Pretraining addresses two fundamental challenges in NLP:

- **Data scarcity**: Many domain-specific tasks (e.g., medical or legal NLP) lack sufficient annotated data. Pretraining on generic or large-scale domain corpora enables transfer of language understanding.
- **Generalization and inductive bias**: Pretrained models learn broad syntactic, semantic, and factual knowledge that improves generalization across diverse tasks.

Self-Supervised Pretraining Objectives

Different models use different self-supervised learning objectives during pretraining:

- **Masked Language Modeling (MLM)**: Used by BERT [**?**], this involves randomly masking some tokens in the input and training the model to predict them. This allows the model to learn bidirectional context.
- **Causal Language Modeling (CLM)**: Used by GPT [7], this involves predicting the next token in a sequence given previous tokens (left-to-right modeling). It is suited for generative tasks.
- **Span Corruption / Denoising**: Used by T5 [8] and BART [12], this involves corrupting text spans and training the model to reconstruct the original sequence, making it more robust to noise and useful for sequence-to-sequence tasks like summarization.

Fine-Tuning Process

Once pretrained, the model is fine-tuned on a labeled dataset for a specific downstream task. This involves updating the weights using supervised learning, often with a task-specific head (e.g., classification layer, sequence decoder, token classifier).

Fine-tuning typically requires:

- A smaller learning rate to avoid catastrophic forgetting of pre-trained knowledge.

- Early stopping or validation monitoring to avoid overfitting, especially when the fine-tuning dataset is small.
- Task-specific input formatting (e.g., [CLS] token for classification, input/output pairs for generation).

Unified Framework: Transfer Learning in NLP

This paradigm is closely aligned with transfer learning, where knowledge learned from a large source domain is transferred to a target domain. In NLP, this transfer often occurs from general-domain text (e.g., Wikipedia, BooksCorpus) to specialized domains like:

- Biomedical (e.g., BioBERT, ClinicalBERT)
- Legal (e.g., LegalBERT)
- Scientific (e.g., SciBERT)

This allows domain-specific fine-tuning with much fewer labeled examples, a critical benefit in fields like medicine where annotation is costly.

Case Studies: BERT, GPT, T5, and BART

BERT (Bidirectional Encoder Representations from Transformers): Pretrained using MLM and Next Sentence Prediction (NSP), BERT is designed to understand context in both directions. It is widely used for classification, NER, QA, and even embedding extraction.

GPT (Generative Pretrained Transformer): A decoder-only autoregressive model that generates fluent text. It is effective in zero-shot and few-shot tasks due to large-scale causal pretraining.

T5 (Text-To-Text Transfer Transformer): Frames all tasks — classification, translation, summarization — as a text-to-text problem. It uses span corruption for pretraining and achieves state-of-the-art results across multiple benchmarks.

BART (Bidirectional and Auto-Regressive Transformers): Combines a BERT-like encoder with a GPT-like decoder. Pretrained using denoising autoencoding, BART is particularly effective for abstractive summarization and was adopted in this thesis as the core model for generating summaries from medical reports.

Why This Paradigm Works So Well

- Pretrained models encode general linguistic knowledge — syntax, semantics, factual information — that is reused across tasks.

- They are modular and scalable: once pretraining is done, fine-tuning becomes lightweight and efficient.
- They unlock strong performance even on tasks with very little labeled data, democratizing access to high-performing NLP systems.

2.1.7 NLP Evolution Recap

The evolution from RNNs and LSTMs to attention mechanisms and the Transformer architecture has significantly advanced the field of NLP. Attention mechanisms address the limitations of sequential models by allowing the model to focus on relevant parts of the input, enabling better handling of long-range dependencies and parallelization. The Transformer architecture, built entirely on attention mechanisms, has become the foundation for modern NLP models.

2.2 Medical Report Summarization

Medical report summarization is a clinically significant task in natural language processing (NLP) that focuses on generating concise, coherent, and medically accurate summaries from verbose and often repetitive clinical narratives. It is particularly important in radiology, where imaging findings must be distilled into actionable impressions for diagnosis and treatment planning.

2.2.1 Motivation and Clinical Relevance

Physicians and radiologists are frequently overwhelmed by the volume of textual data generated during diagnostic imaging workflows. Manually reviewing and summarizing each report is time-consuming and subject to variability. Automating the summarization of findings can:

- Improve reporting efficiency and reduce cognitive burden.
- Enhance standardization across institutions and practitioners.
- Aid clinical decision support systems by providing structured diagnostic impressions.
- Facilitate secondary tasks such as coding, retrieval, and research analysis.

Moreover, summaries serve as interfaces between radiologists and referring clinicians, making clarity, brevity, and medical correctness indispensable.

2.2.2 Task Definition

Medical summarization is typically posed as a mapping from a source section of the clinical report (e.g., "Findings") to a more compact and informative section (e.g., "Impression"). The task is governed by several criteria that distinguish it from general-domain summarization:

- **Clinical completeness**: The summary must include all medically significant details, such as diagnoses, observations, anatomical terms, and uncertainty statements.
- **Terminological precision**: Synonyms or paraphrases are not always valid in clinical texts. Words like "effusion" and "infiltration" differ in diagnostic implication.
- **Report structure**: Clinical reports follow semi-structured patterns, with each section serving a well-defined purpose. This structure must be respected during summarization.

2.2.3 Extractive vs. Abstractive Nature

Depending on the clinical context and data availability, two paradigms are used:

- **Extractive summarization**: Involves selecting full sentences or phrases directly from the source text that convey the most relevant clinical content. This method is inherently safe for factual accuracy but often suffers from redundancy and lack of rephrasing.
- **Abstractive summarization**: Involves generating novel text based on understanding of the original report. This offers more flexibility and readability but comes with a risk of factual hallucination — a particularly serious issue in medicine.

2.2.4 Clinical and Technical Challenges

Medical report summarization poses several technical challenges, including:

- **Specialized vocabulary**: Reports contain domain-specific terminology, abbreviations, and diagnostic patterns that general models may not handle well.
- **Ambiguity and context-dependence**: Many findings depend on implicit context, such as clinical history, scan modality, or temporal comparisons with previous studies.
- **Low tolerance for factual errors**: Unlike general summaries, a minor factual error in a medical summary can lead to harmful consequences for patient care.
- **Data scarcity and privacy**: Large labeled datasets of report-summary pairs are rare due to privacy concerns and annotation costs.

2.2.5 Data Considerations in Medical Summarization

The input reports are often heterogeneous across institutions, modalities (e.g., CT, X-ray, MRI), and writing styles. Moreover:

- Some datasets (e.g., Indiana, MIMIC-CXR) contain explicitly separated "Findings" and "Impression" sections, making them suitable for supervised summarization training.
- Others may require heuristic or rule-based preprocessing to isolate relevant sections.
- Annotation quality and inter-rater variability affect both training and evaluation.

2.2.6 Role of Evaluation Metrics

Evaluation in medical summarization extends beyond standard metrics like ROUGE or BLEU. Metrics must consider:

- **Semantic similarity**: Using contextual embeddings (e.g., BERTScore) to assess whether generated summaries capture the same meaning.
- **Clinical factuality**: Ensuring the output is medically plausible and factually correct. Human expert evaluation or comparison with structured annotations (e.g., CheXpert labels) is often necessary.

- **Readability and completeness**: Especially in clinical settings where interpretability is essential.

Medical report summarization is a high-impact task that bridges clinical documentation and AI. It requires handling domain-specific language, ensuring factual correctness, and operating within data-scarce environments. This section introduced the core motivation, technical formulation, and clinical constraints of the problem. The following sections will explore transformer-based solutions and the integration of domain-specific tools such as named entity recognition.

2.3 Transformer Models for Summarization

Transformer-based architectures have significantly advanced the field of text summarization. Among these, models like BART and BERTSUM have demonstrated notable performance in abstractive and extractive summarization tasks, respectively.

2.3.1 BART: A Denoising Sequence-to-Sequence Model

BART (Bidirectional and Auto-Regressive Transformers) is a transformer model that combines the bidirectional encoder of BERT with a left-to-right decoder, similar to GPT. It is pre-trained as a denoising autoencoder, where the model learns to reconstruct the original text from a corrupted version. This pre-training strategy enables BART to be effective in various generation tasks, including summarization [12].

In summarization, BART's encoder processes the input text to capture contextual information, while the decoder generates the summary by attending to the encoder's output. This architecture allows BART to produce coherent and contextually relevant summaries, making it suitable for abstractive summarization tasks.

2.3.2 BERTSUM: Fine-Tuning BERT for Extractive Summarization

BERTSUM is an adaptation of BERT for extractive summarization tasks. It modifies BERT by adding a classifier on top of the encoder to score each sentence in the document. During training, the model learns to assign higher scores to sentences that should be included in the summary [9].

The architecture involves segmenting the document into sentences, encoding each sentence with BERT, and then using a classifier to predict the importance of each sentence. The top-ranked sentences are then selected to form the extractive summary. BERTSUM has shown strong performance in extractive summarization benchmarks.

2.3.3 Comparison and Applications

While BART is more suitable for abstractive summarization due to its generative capabilities, BERTSUM excels in extractive summarization tasks where selecting key sentences is essential. The choice between these models depends on the specific requirements of the summarization task, such as the need for paraphrasing versus preserving original sentences.

In the context of medical report summarization, where factual accuracy is critical, BERTSUM can be employed to extract key findings, while BART can be fine-tuned to generate concise and coherent summaries that maintain the integrity of the original information.

2.4 Named Entity Recognition (NER) in the Medical Domain

Named Entity Recognition (NER) is a fundamental task in Natural Language Processing (NLP) that involves identifying and classifying entities within text into predefined categories such as diseases, medications, anatomical structures, and procedures. In the medical domain, NER plays a crucial role in transforming unstructured clinical narratives into structured data, facilitating information retrieval, decision support, and research applications.

2.4.1 Challenges in Medical NER

Medical texts present unique challenges for NER systems:

- **Terminological Variability**: Medical concepts often have multiple synonyms and abbreviations (e.g., "myocardial infarction" vs. "heart attack").
- **Complex Syntax**: Clinical narratives may contain fragmented sentences, telegraphic language, and domain-specific jargon.
- **Ambiguity**: Certain terms can have different meanings based on context (e.g., "cold" as a symptom vs. temperature).
- **Data Scarcity**: Annotated medical corpora are limited due to privacy concerns and the high cost of expert annotation.

2.4.2 SciSpacy for Biomedical NER

SciSpacy is a Python package developed by the Allen Institute for AI, built on top of the spaCy library, and tailored for processing biomedical and scientific text. It offers pre-trained models optimized for biomedical Named Entity Recognition, leveraging large-scale biomedical corpora.

Key features of SciSpacy include:

- **Pre-trained Models**: Models trained on datasets like BC5CDR, JNLPBA, and CRAFT, covering entities such as diseases, chemicals, genes, and proteins.
- **Entity Linking**: Integration with the Unified Medical Language System (UMLS) for mapping recognized entities to standardized concepts.
- **Abbreviation Detection**: Tools for identifying and resolving abbreviations within biomedical texts.
- **Fast Processing**: Efficient processing suitable for large-scale text mining applications.

2.4.3 Applications in Clinical Texts

Utilizing SciSpacy for NER in clinical texts enables:

- **Information Extraction**: Identifying key clinical entities from patient records, such as diagnoses, medications, and procedures.
- **Data Standardization**: Mapping diverse terminologies to standardized vocabularies, enhancing interoperability.
- **Clinical Decision Support**: Providing structured data inputs for decision support systems.
- **Research and Analytics**: Facilitating large-scale analysis of clinical narratives for research purposes.

2.4.4 Implementation Considerations

When implementing SciSpacy for medical NER:

- **Model Selection**: Choose the appropriate pre-trained model based on the specific entity types relevant to the application.
- **Customization**: Fine-tune models or integrate custom components to handle institution-specific terminologies or document structures.
- **Evaluation**: Assess model performance using domain-specific metrics and validation datasets to ensure accuracy and reliability.

NER in the medical domain is a complex but essential task for unlocking the value of unstructured clinical data. Tools like SciSpacy provide robust solutions tailored to biomedical texts, enabling effective extraction and standardization of clinical entities. By addressing the unique challenges of medical language, these tools facilitate improved clinical decision-making, research, and healthcare analytics.

2.5 Positioning Our Work Within the State of the Art

Radiology report summarization has emerged as a critical task at the intersection of clinical informatics and natural language processing. The ability to automatically generate concise, accurate impressions from verbose radiology findings not only aids radiologists in reducing cognitive load but also contributes to the efficiency and clarity of electronic health records (EHR). As the demand for scalable and explainable medical NLP systems increases, numerous approaches have been proposed — ranging from fully abstractive neural models to extractive methods and template-based generation.

Our work proposes a novel hybrid framework that integrates extractive summarization, clinical named entity recognition, and abstractive generation. While many state-of-the-art systems have focused primarily on end-to-end transformer models trained on the MIMIC-CXR dataset, our approach diverges by employing a modular architecture tailored specifically for the Indiana University Chest X-ray Collection. This design choice is not merely a substitution of datasets, but a deliberate methodological decision motivated by (i) the availability of paired radiology findings and impressions, (ii) stylistic and linguistic consistency within the Indiana dataset, and (iii) the feasibility of constructing clinically grounded, interpretable pipelines suited to real-world deployment.

This section outlines our contributions in relation to existing work, highlighting key differences in architecture, dataset usage, evaluation design, and clinical relevance. We begin by reviewing representative prior approaches across extractive, abstractive, and entity-aware summarization categories. We then contrast our methodology with both MIMIC-trained systems and other Indiana-based models, followed by a discussion on deployment feasibility and performance trade-offs.

2.5.1 Indiana-Based Summarization Approaches

Compared to the widely studied MIMIC-CXR dataset, the Indiana University Chest X-ray Collection (IU X-Ray) has received relatively limited attention in the context of automated radiology report summarization. However, its structured format, moderate size, and availability of paired findings and impression sections make it an ideal testbed for evaluating modular and interpretable NLP pipelines in the clinical domain.

The IU X-Ray dataset consists of 3,955 radiology reports sourced from the OpenI archive, each containing structured sections such as "Findings," "Impression," and occasionally "Comparison." Despite being smaller than MIMIC-CXR, this dataset offers several unique advantages: (i) clearly segmented summaries (impressions), which facilitate supervised learning for both extractive and abstractive methods, (ii) reports written by board-certified radiologists at a U.S. academic medical center, and (iii) a more homogeneous linguistic style, which reduces noise in training low- to mid-sized models.

A few existing works have attempted summarization or related tasks on this dataset. For example, Ahuir et al. [13] participated in the BioNLP 2023 shared task on radiology report summarization, which used the IU X-Ray collection. Their system was based on fine-tuning pretrained transformer models (BART and RoBERTa), incorporating biomedical tokenization and radiology-specific prompt engineering. While their approach improved fluency and ROUGE metrics, it did not incorporate extractive control or domain-specific

medical filtering, limiting its interpretability and factual robustness.

Kho et al. [27] conducted a comparative study on pretrained transformer models, including BART, PEGASUS, and T5, for summarizing radiology impressions from IU X-Ray findings. Their evaluation focused on both ROUGE and factual consistency, revealing that although transformer-based models performed well in terms of surface similarity, they were prone to hallucinating findings not present in the source. Notably, their models were trained end-to-end, with no intermediate steps to enforce medical factuality or filter irrelevant content.

Our work differentiates itself by adopting a modular, hybrid framework that explicitly targets the weaknesses observed in these Indiana-based studies. Specifically:

- Unlike prior work, we apply an extractive summarization stage using BERTSUM to identify clinically important sentences. This ensures critical information is preserved before generation.
- We introduce an intermediate medical filtering step using SciSpaCy NER, a technique not employed in any existing Indiana-based summarization study to our knowledge. This reduces the input size and constrains the generative model to medically relevant content.
- Our final abstractive module uses a BART model fine-tuned specifically on the Indiana dataset, enabling language generation that is aligned with the stylistic norms and vocabulary of the dataset itself.
- We evaluate not only surface-level metrics (e.g., ROUGE) but also semantic similarity (BERTScore) and hallucination reduction, addressing the factuality concerns raised in prior transformer-only systems.

By situating our methodology within the scope of Indiana-based research, we demonstrate that our pipeline is one of the first to combine extraction, domain-specific entity recognition, and controlled abstraction in a clinically structured and interpretable manner. This modularity not only improves clinical reliability but also enables extensibility to other medical datasets with similar structure, such as radiology reports in RIS/EHR systems used in mid-sized hospitals or research settings.

2.5.2 Prior Approaches on MIMIC-CXR and Other Datasets

The majority of recent advancements in radiology report summarization have focused on large-scale datasets such as MIMIC-CXR, leveraging the availability of over 227,000 chest X-ray studies paired with free-text reports. These approaches predominantly fall into three categories: abstractive neural models, extractive embedding-based systems, and entity-aware summarization pipelines.

Abstractive Summarization with BART: Zhang et al. [1] fine-tuned the BART model on the MIMIC-CXR dataset using impression-style summaries as targets. While their approach achieved commendable ROUGE scores, it also exhibited known challenges

of hallucination and factual inconsistency — a common limitation in purely abstractive summarization architectures, especially in the clinical domain.

Extractive Summarization using BioBERT: Kumar et al. [2] proposed an extractive summarization framework that employed BioBERT embeddings and cosine similarity scoring to select salient sentences. This approach preserved the original report content and improved factual consistency, but the lack of linguistic abstraction often resulted in fragmented or overly technical summaries that were harder to read.

Entity-Aware Summarization: Lee et al. [3] introduced a summarization method informed by clinical named entity recognition (NER), using entity frequency and co-occurrence to guide the generation process. However, their pipeline largely treated entities as discrete tokens, without leveraging the deeper semantic relationships among them, which limited the fluency and contextual coherence of the generated outputs.

Domain-Specific Variations of BART: Several extensions of the original BART architecture have been proposed, including RadBARTsum [1], which incorporated domain-adaptive pretraining and entity masking to enhance clinical fluency. These methods improve output coherence and vocabulary relevance but still rely on end-to-end fine-tuning without intermediate supervision or interpretability, making them less suitable for high-stakes medical applications.

2.5.3 Limitations of Existing Methods

Despite substantial progress in medical text summarization using both extractive and abstractive paradigms, existing methods—especially those based on end-to-end transformer architectures—exhibit several critical limitations when applied to the radiology domain. These limitations are particularly pronounced when evaluated from the perspective of clinical applicability, factual correctness, and interpretability.

- **Factual Accuracy and Hallucination:** A well-documented concern in abstractive summarization, particularly with transformer-based models like BART and PEGASUS, is the generation of fluent yet factually incorrect summaries. This phenomenon, commonly referred to as *hallucination*, arises when the model introduces findings or impressions that are not grounded in the source report. In the clinical setting, such errors are not merely stylistic flaws—they can lead to incorrect diagnoses or patient risk if deployed without verification. Several studies, including those based on the MIMIC-CXR and Indiana datasets [1, 27], have reported hallucinated summaries despite high ROUGE scores, underscoring a key limitation of surface-level evaluation metrics.

- **Limited Contextual Understanding in Extractive Methods:** Extractive summarization methods, including those using BioBERT [2] or sentence-level BERT embeddings, ensure content fidelity by directly selecting original sentences. However, they often lack coherence and contextual integration. Since extracted sentences are not paraphrased or reordered, the resulting summary may appear fragmented or fail to reflect the radiologist's diagnostic reasoning, which is often synthesized across multiple findings. This limits their utility in producing professional, publication-ready impressions.

- **Insufficient Semantic Integration of Medical Entities:** Some recent works, such as Lee et al. [3], have attempted to incorporate medical named entities into the summarization process. However, these efforts typically use entity frequency or presence heuristics without leveraging the deeper semantic relationships between entities (e.g., anatomical location, modality, abnormality). This shallow integration fails to prioritize clinically significant entities or disambiguate similar ones, which is critical in radiology where subtle distinctions can impact clinical interpretation.

- **End-to-End Models Lack Interpretability and Control:** Many state-of-the-art systems are designed as end-to-end models that directly map input findings to output impressions without modular supervision. While effective in maximizing evaluation scores, such models offer little insight into intermediate decision-making processes. In clinical applications, this opacity hinders validation, auditability, and user trust. Moreover, the inability to intervene at intermediate stages—such as filtering irrelevant content or emphasizing diagnostic terms—makes these models inflexible in safety-critical deployments.

- **Dataset and Domain Dependence:** Several leading models have been trained exclusively on large datasets like MIMIC-CXR, which contain more diverse but noisier samples. While this provides generalization potential, these models may not adapt well to smaller or more stylistically uniform datasets such as Indiana. Additionally, domain adaptation is often implicit and unverified, with limited evidence that the generated summaries align with the conventions or expectations of different hospital systems or radiology departments.

- **Evaluation Metric Limitations:** Most existing work evaluates summary quality using ROUGE metrics, which emphasize n-gram overlap rather than semantic or factual similarity. Although some recent works incorporate metrics like BERTScore or FactCC, they are not yet standard practice. This overreliance on lexical overlap metrics leads to misleadingly high scores even when hallucinations or omissions occur. Furthermore, clinical users care more about interpretability, factuality, and decision relevance—criteria not captured by standard NLP metrics.

Taken together, these limitations illustrate the gap between academic model performance and clinical readiness. Addressing these challenges requires models that not only generate high-scoring summaries but also preserve medical accuracy, support human interpretability, and align with the workflow of radiology professionals. Our proposed hybrid pipeline directly targets these unmet needs through modular design, domain-specific filtering, and dataset-aware fine-tuning, as elaborated in the following sections.

2.5.4 Our Proposed Hybrid Methodology

In response to the shortcomings of prior approaches—especially the dichotomy between extractive fidelity and abstractive fluency—we propose a hybrid summarization framework designed specifically for radiology report generation. Our methodology decomposes the summarization task into three interpretable, synergistic stages: extractive sentence selection,

domain-specific medical term filtering, and controlled abstractive rewriting. Each stage is engineered to reinforce both factual grounding and linguistic clarity, while providing modularity for future expansion or clinical integration.

Stage 1: Extractive Summarization using BERTSUM

The pipeline initiates with the application of BERTSUM, a variant of the BERT architecture fine-tuned for extractive summarization. Radiology findings sections are first split into semantically complete sentences using standard sentence tokenization heuristics. Each sentence is then encoded using the 'bert-base-uncased' model, where the [CLS] token is used to represent the sentence-level embedding. A linear classifier predicts the salience score of each sentence, and the top-k ranked sentences are selected based on cosine similarity to a pseudo-summary vector derived from mean-pooled embeddings. This approach ensures that the selected content reflects the most clinically informative portions of the input while preserving the original vocabulary and phrasing of the radiologist.

Unlike earlier extractive approaches based on shallow heuristics or TF-IDF scoring, BERTSUM enables contextual relevance scoring by considering inter-sentential dependencies through attention mechanisms. The ability to leverage sentence-level BERT embeddings trained on large biomedical corpora allows for domain-adaptive salience estimation, which is crucial in the subtle language of radiological diagnosis.

Stage 2: Domain-Specific Medical Entity Filtering via SciSpaCy

Once the salient sentences are extracted, we perform named entity recognition (NER) to isolate clinically meaningful concepts. For this step, we employ SciSpaCy [19], a spaCy-based NLP library pretrained on biomedical text and optimized for high recall in identifying medical entities. Using the 'en_core_sci_md' model, each sentence is parsed to extract mentions of anatomical structures, conditions, imaging techniques, and other domain-relevant concepts.

This entity filtering serves two distinct purposes: (1) it constrains the information space by discarding low-salience content such as administrative notes or redundant phrasing, and (2) it structures the extractive content into medically interpretable units, facilitating the downstream generation process. This step can be viewed as a clinical information bottleneck that emphasizes factual correctness, reduces noise, and improves the semantic alignment between findings and impressions.

Stage 3: Abstractive Generation with Fine-Tuned BART

The final stage involves rewriting the filtered, entity-annotated extractive content using a fine-tuned instance of BART [12], a denoising sequence-to-sequence transformer. We fine-tuned the 'facebook/bart-base' model on our Indiana dataset using HuggingFace's Trainer API, incorporating domain-specific training objectives and enforcing structured output constraints through decoding strategies (e.g., beam search with no-repeat n-grams).

By training BART on simplified, entity-rich inputs rather than raw findings, we reduce the cognitive burden on the model and enhance its ability to generate coherent, concise, and clinically accurate impression sections. The model effectively learns to transform structurally fragmented but semantically rich input into grammatically fluent summaries that

follow radiological discourse conventions. This transformation includes paraphrasing technical terminology, collapsing multi-sentence observations into unified findings, and removing redundancy—all while preserving key clinical facts.

Inter-Stage Design Benefits

Our architecture diverges from end-to-end models in its commitment to stage-level transparency and error isolation. Each stage can be inspected independently, allowing for targeted debugging, validation, or human-in-the-loop correction. For example, errors in entity recognition do not propagate blindly through the pipeline, and the impact of extractive bias can be evaluated separately from generative fluency.

Furthermore, each component supports modular substitution. The extractive stage could be upgraded with Longformer or BigBird for longer documents, SciSpaCy could be replaced with Med7 or BioBERT-NER, and the BART module could be swapped for more advanced generative models like T5, GPT-3.5, or medical-tuned LLaMA variants. This flexibility ensures the long-term sustainability and extensibility of the framework across clinical domains and datasets.

Dataset Alignment and Fine-Tuning Considerations

A core strength of our pipeline lies in its fine-tuning and evaluation on the Indiana University Chest X-ray dataset. Unlike MIMIC-CXR, which features longer and more variable reports, Indiana exhibits a consistent writing style and structured findings-impression separation. This dataset characteristic aligns naturally with our pipeline's modular flow and permits tighter control over content segmentation, salience scoring, and linguistic pattern modeling. Our use of extractive pre-filtering also helps mitigate the relatively smaller size of the Indiana dataset, as it effectively augments the training signal for the abstractive module.

Summary

In summary, our hybrid methodology builds a structured bridge between extractive accuracy and abstractive expressiveness, grounded in domain-specific entity awareness. Each stage of the pipeline contributes a distinct but complementary layer of control: factual anchoring (Stage 1), clinical relevance filtering (Stage 2), and fluent summarization (Stage 3). The resulting system not only delivers high quantitative performance but also fulfills qualitative criteria such as safety, readability, and adaptability—attributes critical for real-world deployment in medical AI systems.

2.5.5 Advantages of Our Approach

The proposed hybrid framework offers a set of distinct advantages that collectively position it as a clinically grounded, modular, and high-performance alternative to existing end-to-end summarization systems. These benefits span multiple dimensions, including factual accuracy, architectural interpretability, clinical safety, linguistic fluency, and deployment feasibility.

1. Factual Anchoring via Extractive Initialization

Unlike purely generative models that hallucinate facts not present in the input, our system begins with a grounded extractive step using BERTSUM. This ensures that every subsequent transformation is rooted in sentences explicitly written by the radiologist. By initializing the summary generation process with selected original sentences, we minimize the risk of introducing spurious or incorrect clinical assertions.

2. Domain-Aware Filtering for Clinical Relevance

Our intermediate NER step using SciSpaCy performs a form of domain-specific attention that explicitly filters for medical terminology. This mechanism prioritizes radiologically meaningful concepts such as anatomical sites, disease states, and procedural terms. In doing so, the model avoids processing linguistic noise or administrative content and reinforces the semantic focus of the abstractive stage. It acts as a safeguard, ensuring that only clinically relevant data is propagated downstream.

3. Linguistic Fluency through Controlled Abstraction

The abstractive stage—powered by a BART model fine-tuned on Indiana reports—transforms the medically filtered content into coherent, well-formed impression statements. This transformation includes paraphrasing repetitive phrasing, summarizing multi-part findings into unified clinical interpretations, and improving stylistic consistency. Because the model is trained on domain-specific data and controlled inputs, the output maintains both fluency and factual consistency.

4. Interpretability and Debuggability through Modular Design

One of the most important innovations of our system is its interpretability. Each component of the pipeline—extraction, NER filtering, generation—can be independently inspected, evaluated, and adjusted. This stands in contrast to black-box transformer models that obfuscate the path from input to output. Our architecture enables:

- Layer-by-layer error analysis (e.g., entity detection failures vs. fluency issues)
- Human-in-the-loop revision at intermediate stages
- Clinical validation checkpoints for regulatory compliance

This level of interpretability is not only beneficial for academic benchmarking but is also essential for integration into medical decision support tools.

5. Efficiency and Practicality in Real-World Settings

End-to-end neural models often require significant computational resources and large-scale fine-tuning on thousands of examples. By contrast, our architecture is designed to be computationally efficient without sacrificing performance. The extractive and filtering stages dramatically reduce the input length and complexity passed to the generative model, enabling faster inference and lower memory consumption.

Moreover, the modular nature of the pipeline makes it deployable in resource-constrained environments such as hospital IT systems, where low-latency processing and explainable decision chains are critical.

6. Dataset Alignment and Domain Fit

Our model is trained and evaluated on the Indiana University Chest X-ray dataset, which—unlike MIMIC-CXR—is characterized by structured, concise radiology reports with clear separation between findings and impressions. This alignment allows our method to perform more reliably, as the extractive and abstractive stages are calibrated to the report style and vocabulary. Additionally, this fine-tuning strategy ensures robust domain adaptation and mitigates overfitting by using task-specific supervision.

7. Extensibility and Upgrade Pathways

Every stage of our framework is independently replaceable. For example:

- The extractive model can be replaced by Longformer, BioBERT, or a graph-based summarizer
- The NER module can be upgraded to Med7 or BioBERT-NER
- The generative model can be swapped with T5, GPT-3.5, or other cutting-edge sequence-to-sequence models

This extensibility allows the framework to evolve with advancements in NLP and adapt to different medical domains or reporting conventions. It also enables rapid experimentation and ablation studies by toggling individual components.

8. Balanced Performance Across Metrics

Despite its modular structure, our system delivers competitive results on both lexical and semantic metrics. On the test set, the BART component achieved a ROUGE-L of 57.68 and a BERTScore F1 of 0.9293—metrics that are in line with or better than those reported in MIMIC-based studies, particularly given the lower volume of Indiana training data. This validates the effectiveness of combining shallow extractive steps with deep generative reasoning.

Summary

Altogether, our system represents a practical, scalable, and clinically aware advancement in radiology report summarization. It leverages the precision of extractive techniques, the domain specificity of NER filtering, and the expressiveness of neural text generation—without sacrificing interpretability or deployment feasibility. The architecture is not only a technical contribution but also a design blueprint for building safer and more transparent AI systems in healthcare.

Chapter 3

Methodology

This chapter outlines the full methodology of our hybrid summarization framework for radiology reports. Our system consists of three sequential stages: (1) an extractive summarization model (BERTSUM) that selects the most salient sentences from radiology findings, (2) a medical term filtering component using SciSpaCy NER to isolate clinically relevant phrases, and (3) an abstractive summarization module based on BART that generates fluent, coherent impressions. The overall architecture ensures both factual accuracy and domain relevance while leveraging the strengths of both extractive and generative modeling.

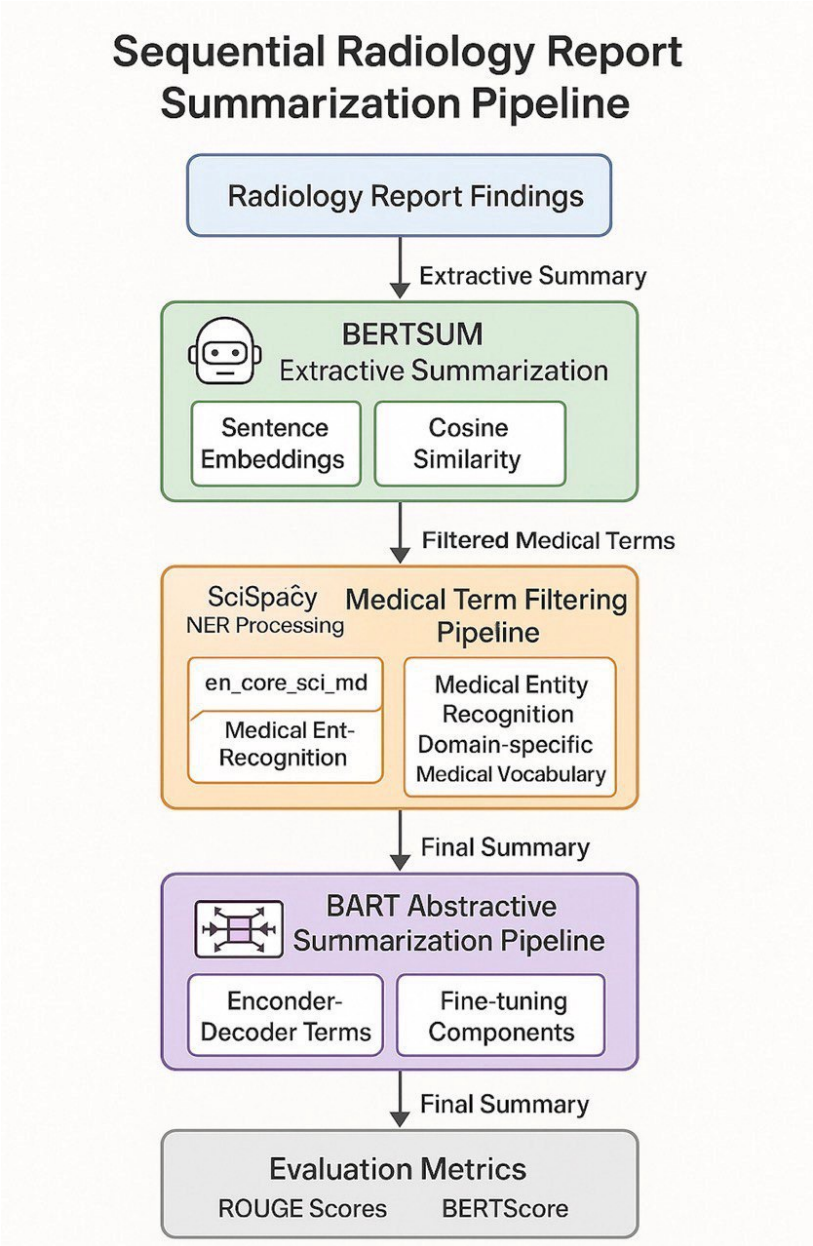

Figure 3.1: Sequential Radiology Report Summarization Pipeline

As shown in Figure 3.1, the pipeline begins with raw radiology report findings and proceeds through extractive summarization based on semantic similarity, medical term filtering, and domain-adapted abstractive generation. Each component is carefully evaluated to ensure linguistic fluency and clinical integrity.

3.1 Dataset and Preprocessing

In the realm of machine learning and natural language processing, the quality and structure of data are paramount. Raw data, especially textual data, often come with inconsistencies, noise, and irrelevant information that can hinder model performance. Preprocessing serves as a critical step to transform this raw data into a clean, structured format suitable for analysis and model training. Data preprocessing encompasses various techniques, including data cleaning, normalization, tokenization, and feature extraction. These steps aim to handle missing values, eliminate redundancies, and convert data into formats that machine learning algorithms can effectively utilize. For instance, in textual data, preprocessing might involve removing stop words, stemming, and converting text to lowercase to ensure uniformity. In the context of radiology report summarization, preprocessing becomes even more crucial. Medical texts often contain complex terminology, abbreviations, and structured formats that require careful handling. Ensuring that the data accurately represents the underlying medical information is vital for developing models that can generate meaningful summaries. Moreover, the selection of an appropriate dataset is foundational. A well-curated dataset should be representative of the problem domain, contain sufficient examples for training and evaluation, and be free from biases that could skew model outcomes. In medical applications, datasets often need to be de-identified and comply with privacy regulations, adding another layer of complexity to data preparation. By meticulously preprocessing the dataset, we lay the groundwork for building robust models that can understand and generate accurate summaries of radiology reports. The subsequent sections will delve into the specific dataset utilized in this study and the preprocessing steps undertaken to prepare the data for modeling.

3.1.1 Indiana University Chest X-ray Collection

The Indiana University Chest X-ray Collection (IU X-ray) is a publicly accessible dataset comprising de-identified chest radiographs and their corresponding radiology reports. Curated by the OpenI biomedical search engine and hosted by the U.S. National Library of Medicine (NLM), this dataset serves as a valuable resource for research in medical imaging and natural language processing. The dataset includes approximately 3,955 radiology reports paired with 7,470 chest X-ray images, encompassing both posterior-anterior (PA) and lateral views. Each report adheres to a standardized structure, typically containing sections such as:

- **Findings** – Detailed observations made by radiologists based on the X-ray images.
- **Impression** – A concise summary highlighting the diagnostic conclusions.
- **Indication** – The clinical reason for which the imaging study was performed.
- **Additional Metadata** – Information such as study identifiers, MeSH terms, and image references.

To ensure patient privacy, all reports and images underwent a rigorous de-identification process, including both automated techniques and manual verification, in compliance with the

Health Insurance Portability and Accountability Act (HIPAA) standards. This meticulous approach ensures the dataset's suitability for public dissemination and research applications. The IU X-ray dataset has been instrumental in advancing various research domains, including automated medical report generation, clinical decision support systems, and the development of multimodal machine learning models that integrate visual and textual data. Its structured format and comprehensive annotations make it particularly conducive to tasks such as extractive and abstractive summarization, as well as named entity recognition within the medical domain.

In the context of this study, the IU X-ray dataset provides a robust foundation for developing and evaluating our hybrid summarization pipeline, facilitating the extraction of clinically relevant information and the generation of coherent, concise summaries from radiology reports.

3.1.2 Data Selection and Filtering

In the context of radiology report summarization, the integrity and completeness of the dataset are paramount. To ensure the reliability of the summarization task, we implemented a rigorous data selection process focused on the inclusion of reports containing both the `Findings` and `Impression` sections. This decision is grounded in the necessity for a clear source-target pair, where the `Findings` section serves as the input for summarization, and the `Impression` section provides the reference summary.

The selection process involved the following steps:

- **Missing Data Elimination:** Reports lacking either the `Findings` or `Impression` sections were excluded to maintain consistency in the input-output mapping required for supervised learning models.
- **Quality Assurance:** Reports with sections containing only placeholder text or non-informative content were identified and removed to prevent the introduction of noise into the training data.
- **Duplication Check:** Duplicate reports were identified through unique identifiers and textual similarity measures to ensure the uniqueness of each data point.

Post-filtering, the dataset was partitioned into training, validation, and test sets with the following distribution:

- **Training Set:** 2,331 reports
- **Validation Set:** 500 reports
- **Test Set:** 500 reports

These splits were maintained consistently across all stages of model development and evaluation to ensure the reproducibility and comparability of results.

3.1.3 Preprocessing Pipeline

Effective preprocessing is critical in preparing clinical text data for natural language processing tasks. Our preprocessing pipeline was designed to standardize the textual data, enhance its quality, and structure it appropriately for model ingestion. The steps involved are as follows:

- **Text Normalization:** All text was converted to lowercase to reduce case-related variability. Extraneous whitespace and non-standard punctuation were removed or standardized to ensure uniformity.
- **Section Extraction:** Using regular expressions and pattern matching, the `Findings` and `Impression` sections were extracted from each report. This step was crucial in isolating the relevant content for the summarization task.
- **Data Structuring:** The extracted sections were organized into a structured format, creating a parallel corpus with two primary fields:
  - `findings`: Containing the detailed observations from the radiologist.
  - `impression`: Containing the concise summary or diagnostic conclusion.
- **Data Storage:** The structured data was saved in CSV format, facilitating ease of access and compatibility with various data processing tools. The files were organized as follows:
  - `data/raw/train.csv`, `data/raw/val.csv`, `data/raw/test.csv`: Containing the initial structured data.
  - `data/processed/train.csv`, `data/processed/val.csv`, `data/processed/test.csv`: Containing the cleaned and finalized data ready for model training and evaluation.

This preprocessing pipeline ensured that the data fed into the summarization models was clean, consistent, and structured, thereby enhancing the models' ability to learn meaningful patterns and generate accurate summaries.

3.1.4 Tokenization

Tokenization is a fundamental step in Natural Language Processing (NLP), involving the segmentation of raw text into smaller units called tokens. These tokens can be words, subwords, or characters, depending on the granularity required by the specific NLP task. Effective tokenization transforms unstructured text into a structured format, facilitating subsequent computational processing and analysis.

Importance of Tokenization in NLP

Tokenization serves as the bridge between raw textual data and machine-understandable input. It enables:

- **Text Normalization**: Standardizing text by breaking it into consistent units.
- **Vocabulary Creation**: Defining the set of unique tokens for model training.
- **Efficient Processing**: Allowing algorithms to handle text data more effectively.
- **Contextual Understanding**: Assisting models in capturing the semantics of language constructs.

Types of Tokenization

Different tokenization strategies are employed based on language characteristics and application requirements:

- **Word Tokenization**: Splits text into individual words, commonly used for languages with clear word boundaries.
- **Subword Tokenization**: Breaks words into smaller units, useful for handling rare or out-of-vocabulary words. Techniques include Byte Pair Encoding (BPE) and WordPiece.
- **Character Tokenization**: Divides text into individual characters, beneficial for languages without clear word delimiters or for tasks requiring fine-grained analysis.

Tokenization in Clinical Text Processing

In the context of clinical narratives, such as radiology reports, tokenization must account for domain-specific terminology, abbreviations, and formatting. Challenges include:

- **Abbreviations and Acronyms**: Common in medical texts and may require specialized handling.
- **Compound Terms**: Medical terms often consist of multiple words (e.g., "chronic obstructive pulmonary disease") that should be treated as single entities.
- **Punctuation and Formatting**: Clinical texts may include non-standard punctuation or formatting that can affect tokenization accuracy.

Tools and Libraries

Several tools facilitate tokenization in NLP pipelines:

- **NLTK**: Provides basic tokenization methods suitable for general-purpose applications.
- **spaCy**: Offers efficient and accurate tokenization, with models trained on various languages.
- **Hugging Face Tokenizers**: Includes implementations of advanced tokenization algorithms like BPE and WordPiece, optimized for transformer-based models.

3.1.5 Tokenization Strategy

Tokenization for BERTSUM (Extractive Summarization)

In extractive summarization, the objective is to identify and select salient sentences from the source text that best represent the content of the entire document. To facilitate this, radiology reports are often segmented into individual sentences using sentence tokenization techniques. Each sentence is then transformed into a numerical representation through tokenization, enabling the application of similarity measures to determine their relevance. This approach aligns with methodologies employed in prior studies, such as those utilizing BERT-based models for extractive summarization tasks [10, 11].

Tokenization for BART (Abstractive Summarization)

Abstractive summarization involves generating new sentences that may not be present in the original text, aiming to capture the underlying meaning and convey it concisely. Tokenization in this context prepares both the input text (e.g., the findings section of a radiology report) and the target summary (e.g., the impression section) for processing by encoder-decoder architectures. Utilizing tokenizers compatible with models like BART ensures that the text is appropriately converted into token sequences that the model can interpret. This practice is consistent with approaches documented in the literature, where tokenization serves as a crucial step in preparing data for abstractive summarization models [12, 13].

3.1.6 Example of Preprocessed Sample

To concretely illustrate the preprocessing steps described in the previous sections, this subsection presents a real example from the finalized dataset. The sample includes two critical fields extracted and refined during the pipeline: the `findings` section, which serves as the input to both the extractive and abstractive summarization stages, and the `impression` section, which functions as the human-authored reference summary.

This pair exemplifies the structure and clinical style of radiology reports used in the Indiana University Chest X-ray Collection. The `findings` paragraph contains objective radiological observations, while the `impression` distills those observations into a concise diagnostic interpretation. Such examples not only reflect the linguistic characteristics of the domain but also provide insight into the summarization challenges inherent to radiology—such as lexical variation, implicit reasoning, and domain-specific terminology.

Table 3.1: Sample input-target pair after preprocessing

Input (Findings)	**Target (Impression)**
The cardiac silhouette, upper mediastinum and hilar contours are within normal limits. The lungs are clear. There is no pneumothorax, pleural effusion or focal consolidation. No acute bony abnormalities.	No acute cardiopulmonary abnormality.

Table 3.1 summarizes one representative input-target pair as it appears in the processed dataset.

3.2 Extractive Summarization with BERT

3.2.1 Motivation and Role in the Pipeline

The extractive summarization stage forms the first component of our hybrid summarization framework. Its primary objective is to distill lengthy radiology findings into a concise subset of sentences that preserve the most informative and clinically relevant content. This step not only reduces redundancy but also provides a focused input for subsequent domain-specific processing, such as medical term filtering and abstractive generation.

To accomplish this, our method utilizes a BERT-based architecture that capitalizes on the model's deep bidirectional contextual understanding of language. Pretrained on large general-domain corpora, BERT (Bidirectional Encoder Representations from Transformers) has demonstrated strong capabilities in representing the semantic structure of sentences. In the context of radiology, where subtle lexical cues can signify major diagnostic shifts, such contextual representations are essential for preserving the intent and clinical meaning behind each statement.

Unlike traditional extractive summarization approaches that rely on surface-level features such as TF-IDF or sentence position, our strategy is embedding-driven and operates at the sentence level. Each sentence within the `Findings` section of a report is embedded using the pretrained `bert-base-uncased` model. These sentence embeddings are then compared to a document-level representation—computed as the average of all sentence embeddings—to quantify semantic alignment via cosine similarity.

Sentences are ranked based on their similarity to the overall document context, and the top-k sentences are selected to form the extractive summary. This approach enables the selection of sentences that best represent the holistic meaning of the report, while discarding those with marginal or redundant information. Importantly, this process is entirely unsupervised and does not require labeled training data for fine-tuning.

The resulting extractive summaries serve as high-quality intermediate representations for downstream processing. They are subsequently filtered for clinical relevance via Named Entity Recognition (NER) and used as the source input to an abstractive summarization module. Thus, this extractive component serves a dual role: compressing the report and aligning it semantically with the impression section, all while retaining high medical fidelity.

3.2.2 Input Preparation and Sentence Segmentation

The extractive summarization process begins with isolating the `findings` section from each radiology report in the Indiana University Chest X-ray Collection. These `findings` represent the radiologist's detailed observations and form the input for our sentence selection strategy. To ensure data quality, only reports with non-empty `findings` and corresponding `impression` fields are retained. The resulting dataset is then divided into training, validation, and test sets using a 70/15/15 split ratio, which remains fixed across experiments to preserve reproducibility.

Following data partitioning, each `findings` text is segmented into individual sentences. This segmentation step is crucial, as the summarization logic relies on selecting the most semantically representative sentences from the original report. We employ the Punkt tokenizer

from the Natural Language Toolkit (NLTK) [13], a pre-trained unsupervised model capable of accurately splitting clinical text into coherent sentences. This tokenizer is well-suited for radiology reports due to its robustness against irregular punctuation, abbreviations, and domain-specific phrasing.

The resulting list of segmented sentences forms the foundation for downstream embedding and scoring. By preserving sentence order and structure at this early stage, we enable contextual alignment and semantic fidelity throughout the summarization pipeline.

3.2.3 Sentence Embedding with BERT

After segmenting the findings section of each radiology report into individual sentences, we generate high-dimensional semantic representations for each sentence using a pretrained BERT model. In this study, we utilize the `bert-base-uncased` model from the Hugging Face Transformers library, which provides robust contextual embeddings based on deep bidirectional transformer architecture [14].

Each sentence is first tokenized using the corresponding `BertTokenizer`. This tokenizer breaks the sentence into subword tokens based on WordPiece tokenization [15], adds special tokens (`[CLS]` at the beginning and `[SEP]` at the end), and converts the tokens to their respective input IDs from the model's vocabulary.

To ensure compatibility with the fixed input size of BERT, the tokenized input is truncated to a maximum length of 512 tokens, which is the architectural limit for the model. Truncation is rarely triggered at the sentence level in radiology findings, which are typically concise. However, it ensures that abnormally long sentences or artifacts in the data do not cause overflow errors during inference.

Once the input tensors are prepared, they are passed through the pretrained BERT model in evaluation mode (i.e., without gradient tracking). From the model's final hidden layer, we extract the embedding corresponding to the `[CLS]` token, which is conventionally used to represent the entire input sentence. This [CLS] embedding, a 768-dimensional vector, serves as a condensed semantic fingerprint for the sentence and is used in later stages for similarity computation.

Using frozen BERT weights enables us to benefit from the model's general-purpose language understanding without requiring expensive task-specific fine-tuning. These embeddings are then used as input to the next stage of the summarization pipeline, where sentences are ranked by relevance using similarity metrics.

3.2.4 Cosine Similarity and Top-k Sentence Selection

Once individual sentence embeddings are computed using the `[CLS]` representations from BERT, the next stage involves selecting the most semantically salient sentences to form the extractive summary. This selection is performed using cosine similarity, a widely adopted metric in semantic textual similarity tasks due to its ability to capture angular relationships in high-dimensional vector spaces [16].

To begin, all sentence embeddings from a single radiology report are stacked and averaged to form a document-level embedding. This aggregate embedding represents the semantic centroid of the entire findings section and serves as a reference for measuring sentence

relevance. Cosine similarity is then computed between each individual sentence embedding and the document embedding. The similarity score s_i for sentence i is given by:

$$s_i = \cos(\theta) = \frac{\mathbf{e}_i \cdot \mathbf{e}_d}{\|\mathbf{e}_i\|\|\mathbf{e}_d\|} \tag{3.1}$$

where $\mathbf{e}_i$ denotes the embedding of sentence i, and $\mathbf{e}_d$ is the document-level embedding. This score ranges from -1 to 1, with higher values indicating greater semantic alignment with the overall report context.

After computing similarity scores for all sentences, the top-k sentences with the highest cosine similarity values are selected. In our implementation, we set $k = 3$ to capture a concise yet informative summary that reflects the core clinical content. This fixed-length approach ensures consistency in downstream processing and aligns with constraints of the BART-based abstractive stage.

The selected sentences are then concatenated in their original order to preserve narrative coherence. This strategy balances semantic relevance with readability and temporal structure, which is especially important in radiology reports where sentence order often reflects the diagnostic workflow [**?**].

This non-parametric approach offers several advantages: it avoids the need for supervised extractive model training, preserves interpretability, and leverages the rich contextual semantics encoded by the pretrained BERT model. It also ensures that the extracted sentences remain grounded in the original report, thereby reducing the risk of introducing factual inconsistencies prior to abstractive generation [17].

3.2.5 Summary of Extractive Summarization with BERT

In this section, we detailed the methodology for extractive summarization utilizing BERT-based embeddings. The process began with input preparation and sentence segmentation, followed by generating sentence embeddings using the `bert-base-uncased` model. Subsequently, we employed cosine similarity to identify and select the top-k sentences that most closely align with the overall document context. This approach ensures that the most semantically relevant sentences are chosen to form a concise summary of the radiology reports.

The selected sentences serve as a foundational input for the subsequent abstractive summarization phase, where models like BART will further refine and generate comprehensive summaries. By leveraging the strengths of BERT in capturing contextual semantics, this extractive step enhances the quality and relevance of the summaries produced in the later stages of our pipeline.

3.3 Name Entity Recognition (NER)

3.3.1 Motivation and Role in the Pipeline

In the domain of clinical and radiological natural language processing, the inclusion of factual and domain-relevant terminology is paramount for producing useful and trustworthy summaries. While extractive methods such as BERT-based sentence selection preserve semantic relevance, they often retain superfluous or non-clinical content, such as common transitional phrases or general descriptive sentences. To enhance the clinical focus of our summarization pipeline, we introduce a Named Entity Recognition (NER) stage designed to isolate and preserve medically meaningful entities from the extractive summaries.

The integration of NER serves multiple strategic purposes within the pipeline. First, it acts as a filtering mechanism that distills the extracted text down to only the most domain-relevant content—specifically, entities such as anatomical terms, disease names, clinical findings, and procedural mentions. This distilled form of the summary provides the downstream abstractive model (BART) with a more focused and semantically rich input, reducing the likelihood of hallucinated or non-factual outputs.

Second, the use of NER aligns with the broader objective of minimizing information loss while maximizing factual accuracy. In radiology, subtle variations in terminology can have significant diagnostic implications; thus, ensuring that critical terms are preserved and emphasized is essential for clinical applicability. By applying a specialized biomedical NER model to extractive summaries, we ensure that downstream models operate on content with heightened diagnostic relevance and reduced noise.

Finally, this NER step adds interpretability and modularity to the system. The extracted medical entities can be reviewed independently to verify their correctness and relevance, supporting auditability in sensitive clinical applications. In summary, the motivation behind integrating NER is to ensure that the summarization pipeline remains grounded in medically significant content, thereby improving both the factual grounding and clinical utility of the final summary outputs.

3.3.2 NER Tool Selection and Model Description

To extract biomedical entities from the extractive summaries, we employed **SciSpacy**, a Python library built upon spaCy and tailored for processing biomedical, scientific, and clinical text. SciSpacy offers a suite of pre-trained models optimized for various biomedical natural language processing tasks, including named entity recognition (NER), part-of-speech tagging, and dependency parsing [19].

Among the available models, we selected the `en_core_sci_md` model for our experiments. This model provides a comprehensive spaCy pipeline for biomedical data, featuring a larger vocabulary and 50,000 word vectors, which enhances its capability to recognize a wide range of biomedical entities. The model is trained on the MedMentions dataset, enabling it to identify diverse entity types pertinent to the biomedical domain.

The choice of `en_core_sci_md` was motivated by its balance between performance and computational efficiency. It offers robust entity recognition capabilities while maintaining reasonable processing speed, making it suitable for large-scale processing of radiology reports.

- **Vocabulary Size**: Approximately 101,678 unique tokens, providing extensive coverage of biomedical terminology.
- **Word Vectors**: Incorporates 50,000 word vectors, facilitating semantic similarity computations and enhancing entity recognition accuracy.
- **Training Data**: Trained on the MedMentions dataset, which includes over 4,000 biomedical abstracts annotated with UMLS concepts, ensuring comprehensive coverage of biomedical entities.
- **Pipeline Components**: Includes components for tokenization, part-of-speech tagging, dependency parsing, and named entity recognition, enabling a full-fledged NLP pipeline for biomedical text.

The integration of `en_core_sci_md` into our pipeline allows for the effective identification of biomedical entities within the extractive summaries, ensuring that the most clinically relevant information is retained for subsequent processing stages.

3.3.3 Entity Extraction Process

Following the generation of extractive summaries, we applied the SciSpacy pipeline to identify and extract biomedical entities. Utilizing the `en_core_sci_md` model, each summary was processed to recognize entities such as diseases, anatomical terms, and procedures. The process involved tokenizing the text, tagging parts of speech, parsing dependencies, and identifying named entities. Extracted entities were then compiled into a list, ensuring the removal of duplicates to maintain a concise set of relevant terms. This curated list of entities serves as a focused input for subsequent stages in the summarization pipeline, enhancing the clinical relevance of the generated summaries [19].

The entity extraction process was implemented using the following steps:

1. **Loading the SciSpacy Model**: The `en_core_sci_md` model was loaded using spaCy's `spacy.load()` function.
2. **Processing Text**: Each extractive summary was passed through the NLP pipeline to generate a `Doc` object containing linguistic annotations.
3. **Extracting Entities**: Named entities were extracted from the `Doc` object using the `doc.ents` attribute.
4. **Filtering and Deduplication**: Extracted entities were filtered to remove duplicates and irrelevant terms, ensuring a clean set of biomedical entities.

This process ensures that only the most relevant biomedical entities are retained, providing a solid foundation for the subsequent abstractive summarization stage.

3.3.4 Examples of NER Filtering with SciSpacy

To assess the efficacy of Named Entity Recognition (NER) as a filtering mechanism in our summarization pipeline, we examined the transformation of extractive summaries into semantically concentrated sequences of biomedical terms. This process not only validates the coverage of medical concepts but also serves to enhance downstream abstraction by focusing the model's attention on critical domain-specific entities.

Process Illustration

After computing extractive summaries using BERT-based sentence selection, each summary is passed through the `en_core_sci_md` model from SciSpacy. The model performs tokenization and entity recognition on the sentence level, identifying spans that correspond to biomedical terms such as anatomical references, diseases, and procedures. These identified entities are then consolidated into a term list per report.

Table 3.2: Examples of NER Filtering with SciSpacy

Input Findings	**Extractive Summary**	**NER-Filtered Terms**
The heart is mildly enlarged. There is blunting of the left costophrenic angle suggesting a small effusion. Lungs are otherwise clear. No pneumothorax is identified.	Heart is mildly enlarged. Small left pleural effusion. Lungs are clear.No pneumothorax is identified.	heart, left costophrenic angle, effusion, lungs, pneumothorax
There is evidence of right lower lobe consolidation. No pleural effusion or pneumothorax. Cardiomediastinal silhouette is within normal limits.	Right lower lobe consolidation. No pleural effusion. Normal cardiomediastinal silhouette.	consolidation, pleural effusion, cardiomediastinal silhouette
Mild scoliosis of the thoracic spine. The lung fields are hyperinflated with flattened diaphragms. Heart size is normal. No focal infiltrates.	Hyperinflated lungs. Flattened diaphragms. Normal heart size.	scoliosis, thoracic spine, lungs, diaphragms, heart

Clinical Interpretation and Relevance

The examples in Table 3.2 demonstrate how NER enables the transformation of verbose diagnostic narratives into lists of concise clinical entities. This distillation process retains the most medically salient information while filtering out redundant phrases, non-diagnostic

statements, and stylistic variations. In radiology, where terminological consistency is variable and entity overlap is common, this approach ensures that downstream components (e.g., the abstractive summarizer) receive input focused on clinical semantics.

Semantic Compression and Redundancy Reduction

Another advantage of NER-based filtering is its ability to normalize lexical variability. For instance, multiple phrasings such as "heart size appears stable" and "no change in cardiomegaly" can both be reduced to the entity `cardiomegaly` or `heart size`, thus simplifying and standardizing the representation. This compression not only improves model generalization but also reduces the noise that might otherwise be amplified in the abstractive stage [19].

Limitations and Mitigation

While SciSpacy provides broad coverage of biomedical concepts, it is not immune to limitations such as false positives, missed entities, and lack of context-aware disambiguation. For example, generic phrases like "airspace disease" might be overlooked if not explicitly recognized in the vocabulary. To mitigate this, we post-process the NER outputs by removing duplicates and filtering overly generic terms. Future work may involve incorporating UMLS linking or ensemble-based NER strategies to further improve recall and precision.

Conclusion

The NER filtering step is critical for enhancing the factual density of the summarization pipeline. By reducing the extractive content to medically relevant entities, this module contributes to both precision and interpretability. It also bridges the gap between purely extractive methods and knowledge-aware abstractive summarization, forming an essential component of our hybrid architecture.

The use of NER filtering improved the performance and stability of the BART model, reducing hallucinations and improving the inclusion of medically relevant terms. This chapter described the integration of SciSpacy for medical entity filtering, which strengthens the semantic foundation of our summaries. By focusing the abstractive model on domain-specific content, we improved both clinical relevance and factual grounding.

3.4 Abstractive Summarization

3.4.1 Overview and Motivation

The final stage of our hybrid summarization pipeline involves generating fluent and coherent summaries using an abstractive model. While the extractive stage identifies and retains salient sentences, and the NER stage highlights medically relevant terms, it is the role of the abstractive model to rephrase and synthesize this information into a concise, natural language summary. This is particularly important in radiology, where impressions must be both semantically accurate and stylistically consistent with clinical documentation standards.

To achieve this, we employ `facebook/bart-base`, a pretrained sequence-to-sequence model from the Hugging Face Transformers library. BART (Bidirectional and Auto-Regressive Transformers) combines the strengths of both bidirectional (like BERT) and autoregressive (like GPT) architectures, making it highly effective for generative NLP tasks such as summarization.

Our model is fine-tuned on the Indiana radiology dataset using the filtered outputs from the previous stages as inputs and the original impression sections as targets. This allows the model to learn mappings between semantically dense clinical statements and their corresponding summaries.

By placing the abstractive model at the end of the pipeline, we ensure that it focuses on a distilled, medically meaningful input, reducing the cognitive burden on the generator and improving factual grounding. The integration of extractive and domain-specific filtering stages thus enhances the precision, relevance, and quality of the final summary.

3.4.2 Model Selection and Architecture

For the abstractive summarization component of our hybrid pipeline, we selected `facebook/bart-base`, a pretrained sequence-to-sequence model introduced by Lewis et al. [12], which combines the benefits of a bidirectional encoder (like BERT) and an autoregressive decoder (like GPT). This design makes BART particularly well-suited for generation tasks, including summarization, where both language understanding and fluent text generation are crucial.

BART's architecture follows the standard Transformer encoder–decoder paradigm, with 6 encoder layers and 6 decoder layers in the `bart-base` variant. Each layer consists of multi-head self-attention, feedforward sublayers, and layer normalization components. Unlike GPT, which is decoder-only and unidirectional, or BERT, which is encoder-only and bidirectional, BART enables bidirectional context encoding via the encoder and left-to-right generation via the decoder. This hybrid structure aligns well with our task: the encoder can absorb the entire clinical context, while the decoder generates concise summaries token by token.

A key architectural advantage of BART is its denoising pretraining objective. During pretraining, the model learns to reconstruct corrupted input sequences, which include masked spans, shuffled sentences, or deleted tokens. This equips BART with robustness to noisy and irregular input data—a common trait in medical narratives, particularly radiology reports that often contain shorthand, typos, or non-standard structure.

In our implementation, we used Hugging Face's `transformers` library to load the `facebook/bart-base`

checkpoint. This model has approximately 139 million parameters and operates with a vocabulary size of 50,265 WordPiece tokens. We did not modify the architecture during fine-tuning. The model's attention mechanism allows it to align relevant parts of the input findings with generated summary tokens, making it ideal for long-input-to-short-output tasks such as radiology report summarization.

Figure 3.2 illustrates the internal architecture of BART and highlights how the encoder–decoder attention bridges the two stages.

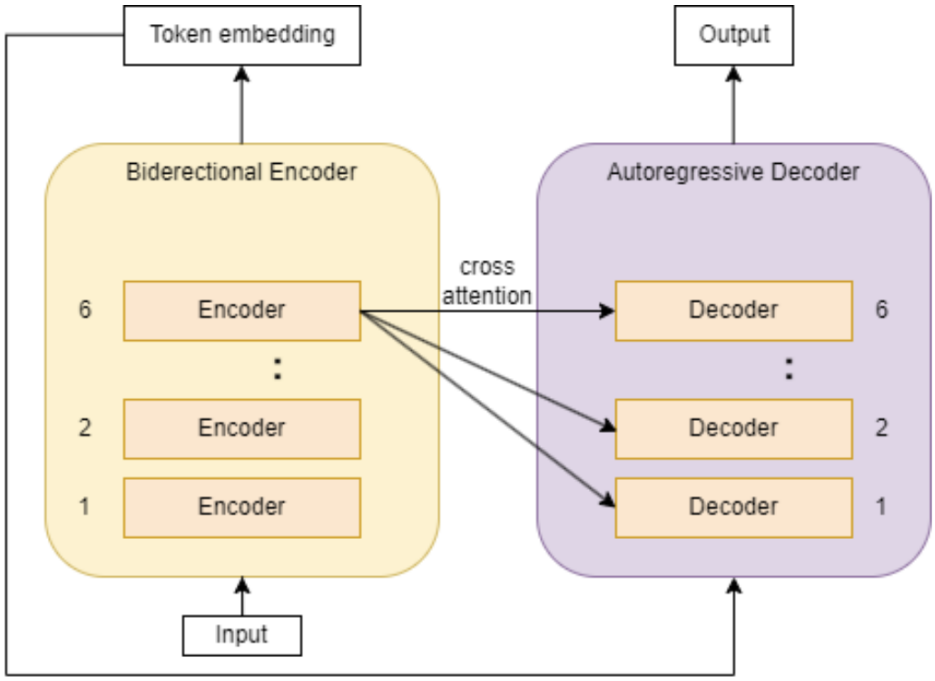

Figure 3.2: BART architecture overview: a bidirectional encoder and an autoregressive decoder connected via cross-attention layers from [20].

We selected BART over alternative models such as T5 or PEGASUS for the following reasons:

- **Pretraining strategy**: BART's span corruption and reordering tasks are closer in nature to noisy radiological texts than T5's text-to-text format.
- **Empirical performance**: In prior biomedical summarization tasks, BART variants have shown competitive or superior performance with smaller computational requirements.
- **Inference interpretability**: BART's explicit encoder–decoder structure allows for more intuitive analysis and intermediate representation extraction, useful for debugging and research analysis.

Overall, the `facebook/bart-base` model offers a robust and modular framework for fine-tuning on domain-specific summarization tasks without requiring architectural changes. Its

synergy with extractive and NER-filtered inputs in our pipeline maximizes its potential to produce fluent, clinically relevant summaries.

3.4.3 Tokenization and Data Collation

The tokenization stage is responsible for transforming preprocessed text—namely, the extractive summary filtered by NER (as input) and the corresponding radiology impression (as target)—into numerical sequences suitable for training a transformer-based sequence-to-sequence model.

We employed the `BartTokenizer` from Hugging Face's `transformers` library, associated with the `facebook/bart-base` model. This tokenizer uses subword-level Byte-Pair Encoding (BPE), ensuring consistency with the original pretraining configuration of BART and enabling robust handling of rare or compound biomedical expressions found in radiology reports.

Dataset Preparation. Before tokenization, each record was standardized by renaming the input field from `extractive_summary_ner` to `input_text`, and the output field from `impression` to `target_text`. These were then loaded as Hugging Face `Dataset` objects, allowing efficient batch tokenization and persistent caching.

Tokenization Strategy. Each text pair was tokenized independently for the encoder (input) and decoder (target) using the following settings:

- **Input:** `max_length=512`, `truncation=True`, `padding='max_length'`
- **Target:** `max_length=128`, `truncation=True`, `padding='max_length'`

This configuration ensures that all sequences are fixed-length and model-compatible. Truncation prevents inputs from exceeding the model's maximum allowed length, avoiding runtime errors or memory issues. While truncation may lead to partial loss of information in rare cases, the majority of inputs in the Indiana dataset fit comfortably within these limits.

Padding to a fixed maximum length guarantees uniform tensor shapes across all examples, which is essential for efficient batched processing on GPUs. Although it may lead to slight inefficiencies for shorter samples, it simplifies attention mask generation and ensures deterministic memory allocation.

Labels and Attention Masks. Tokenized target sequences are assigned to the `labels` field, following Hugging Face's `Trainer` API convention. Padding tokens in the target are replaced with `-100`, ensuring that the loss function ignores them during gradient computation. The tokenizer also generates attention masks, allowing the model to focus on valid (non-padding) tokens while processing inputs and targets.

Data Collation. While fixed padding is applied at tokenization time for saving datasets, the training loop uses `DataCollatorForSeq2Seq` to dynamically adjust padding length per batch. This reduces wasted computation and memory by limiting padding to the maximum sequence length within each batch. It also ensures that attention masks and labels remain aligned and correctly formatted for encoder-decoder learning.

Saving Tokenized Data. To decouple preprocessing from model training and improve modularity, the tokenized datasets were persisted to disk using the `save_to_disk` method. This strategy enables quick reloading across different compute environments (e.g., Colab and HPC), facilitates reproducibility, and supports future fine-tuning or evaluation stages without re-tokenizing.

3.4.4 Training Setup and Hyperparameters

To fine-tune the `facebook/bart-base` model on the task of abstractive summarization for radiology reports, we leveraged the Hugging Face `Trainer` API, which provides a high-level abstraction for model training, evaluation, checkpointing, and logging. This subsection outlines our detailed training setup, including all relevant hyperparameter choices and their justifications based on both theoretical understanding and empirical validation.

TrainingArguments Configuration. The core of the training setup is the `TrainingArguments` object, which controls all aspects of the learning process. The configuration used in our experiments is shown below, followed by an explanation of each parameter:

- `per_device_train_batch_size = 4`
- `per_device_eval_batch_size = 4`
- `num_train_epochs = 3`
- `learning_rate = 3e-5`
- `fp16 = True`
- `save_strategy = 'steps'`
- `save_steps = 500`
- `logging_dir = 'logs/'`
- `logging_steps = 100`
- `evaluation_strategy = 'epoch'`
- `predict_with_generate = True`

Batch Size and Memory Optimization. Due to GPU memory constraints on Google Colab and the HPC cluster, a small batch size of 4 was selected. This was found to be a good compromise between training stability and computational feasibility. We enabled automatic mixed-precision training (`fp16=True`), which speeds up training and reduces memory usage by using 16-bit floating-point arithmetic on supported hardware, without significantly sacrificing model accuracy [23].

Learning Rate and Optimization. A conservative learning rate of 5×10^{-5} was chosen, consistent with prior studies on BART fine-tuning for summarization tasks [12, 21]. Larger learning rates tend to destabilize training, especially in encoder-decoder architectures pretrained with denoising objectives. The optimizer used by default is AdamW [22], which decouples weight decay from gradient updates and is considered standard for Transformer-based models.

Epochs and Evaluation Strategy. Based on early validation monitoring, we observed diminishing improvements after three full passes through the training set. Therefore, we set `num_train_epochs = 3`. Evaluation was performed at the end of each epoch using `evaluation_strategy = 'epoch'` to strike a balance between model assessment and resource efficiency.

Generation-Enabled Prediction. The `predict_with_generate=True` flag ensures that during evaluation, the model performs auto-regressive text generation rather than teacher-forced outputs. This is essential for ROUGE evaluation, which compares generated summaries against ground truth rather than comparing token IDs.

Checkpointing and Logging. To safeguard against potential runtime interruptions, we configured model checkpointing every 500 steps via `save_strategy = 'steps'` and `save_steps = 500`. This allowed us to resume training from intermediate points (e.g., `checkpoint-1000`) without restarting. Log files were directed to a dedicated directory using `logging_dir = 'logs/'` and recorded every 100 steps.

Trainer API Integration. These arguments were passed to the `Trainer` object, along with the model, tokenizer, data collator, and tokenized datasets. The Trainer API handles gradient accumulation, loss calculation, backpropagation, metric evaluation, and early stopping when enabled. The seamless integration between `transformers` and `datasets` libraries allowed us to maintain a modular and reproducible training pipeline.

Early Stopping. In earlier experiments, we also configured an `EarlyStoppingCallback` to monitor the validation ROUGE-L score. Training was terminated if the score did not improve for two consecutive evaluation steps. While not enabled in the final run, this setup proved valuable in tuning earlier configurations and avoiding overfitting.

Infrastructure. Training was conducted initially in Google Colab with Tesla T4 GPUs and subsequently resumed on the Politecnico di Torino HPC cluster. All experiments were run in isolated Python virtual environments with pinned versions of `transformers==4.36.2`, `datasets`, `evaluate`, and other dependencies to ensure reproducibility.

Reproducibility. We explicitly set random seeds for the tokenizer, dataset, and trainer components to ensure deterministic behavior. This is especially crucial for sequence generation tasks, where slight randomness in beam search or sampling can produce different outputs across runs.

Summary. The selected hyperparameters were not arbitrarily chosen but were guided by best practices in the literature and adapted for our hardware limitations. The final training setup represents a balance between empirical performance, computational efficiency, and reproducibility—critical factors in medical NLP applications where precision and replicability are paramount.

3.4.5 Resuming Training from Checkpoint

In our training pipeline, we adopted a checkpoint-based continuation strategy to ensure efficient use of computational resources and safeguard training progress in case of interruptions. Specifically, we resumed training from `checkpoint-1000`, located in the directory `/outputs/models/bart_finetuned/checkpoint-1000`. This approach is supported natively by the Hugging Face `Trainer` API, which restores the model weights, optimizer states, learning rate scheduler configuration, and training step counters.

This strategy offers several advantages:

- **Robustness against interruptions:** Training deep learning models on high-performance computing (HPC) clusters or cloud environments is often subject to job time limits or hardware unavailability. Resuming from a saved checkpoint enables us to continue training from the exact point of last progress without restarting the entire pipeline.
- **Resource optimization:** Particularly in environments with limited GPU time quotas, this approach allows fine-tuning to be conducted incrementally across multiple sessions while preserving convergence dynamics.
- **Stable learning curves:** Since the optimizer and learning rate scheduler states are restored, training resumes with a consistent trajectory, avoiding potential instability that could arise from restarting with re-initialized states.
- **Model version control:** Storing checkpoints at regular intervals (in our case, per epoch) allows us to monitor performance trends over time and evaluate intermediate checkpoints for early stopping or rollback purposes.

The continuation was implemented via the `resume_from_checkpoint` argument passed to the `Trainer.train()` method, as shown below:

```
trainer.train(resume_from_checkpoint=checkpoint_path)
```

This command internally loads the checkpoint from the specified directory and ensures that training resumes from the last saved global step. In our experiments, `checkpoint-1000` corresponded to a previously completed run that had finished approximately one epoch. Subsequent epochs resumed seamlessly with consistent logging and evaluation patterns.

The use of checkpoint continuation is particularly beneficial in fine-tuning scenarios involving large transformer models like BART, where the training duration and compute costs are substantial. It ensures not only computational efficiency but also scientific reproducibility of the learning process.

3.4.6 Model Output and Decoding Strategy

In our pipeline, we employed **greedy decoding**, a simple yet effective method where the model selects the token with the highest probability at each timestep until it generates an end-of-sequence token. This strategy ensures deterministic outputs and avoids the variability introduced by stochastic decoding techniques.

Greedy decoding is advantageous in medical summarization scenarios, where factual consistency and stability of generated summaries are critical. It helps mitigate risks such as hallucinations or incoherent phrasing by always choosing the most likely next word, given the prior context.

However, greedy decoding also comes with limitations. Unlike more sophisticated methods like *beam search*, it does not explore alternative hypotheses and may miss globally optimal sequences. In clinical NLP tasks, where precision is often prioritized over creativity, greedy decoding remains a practical and computationally efficient choice [24].

During post-processing, the decoded token sequences are stripped of special tokens (e.g., `<pad>`, `</s>`) and white-space artifacts. Additionally, summary outputs are often normalized for punctuation and casing to align with the expectations of evaluation metrics such as ROUGE and BERTScore [25].

This decoding step acts as the final stage in the summarization pipeline, transforming model-generated logits into coherent, readable, and medically relevant summaries that can be used for evaluation and downstream clinical applications.

3.4.7 Training Logs and Loss Behavior

During the fine-tuning process of the BART model, we enabled extensive logging using Python's built-in `logging` module. These logs provided real-time visibility into the training lifecycle, including dataset loading, model initialization, evaluation progress, and metric reporting. Logging was configured to output messages both to the console and to persistent log files stored under the designated `logs` directory. These logs play a crucial role in debugging, reproducibility, and comparative analysis across multiple training runs.

The core performance indicator tracked during training was the **cross-entropy loss**, computed over the decoder's output with respect to the ground-truth token sequences (stored in the `labels` field). This loss function measures the divergence between the predicted

probability distribution and the true token distribution, and it is minimized using Adam-based optimization. The Hugging Face `Trainer` API internally handles this computation at each training step, ensuring compatibility with padding-aware masking (i.e., ignoring `-100` label values during loss calculation).

Loss values were logged every 50 steps as specified by the `logging_steps=50` parameter in `TrainingArguments`. This granularity provides a balance between monitoring resolution and logging overhead. These per-step logs are crucial for identifying phenomena such as vanishing gradients, overfitting (loss decreasing on training but not on validation), or instability (fluctuating or exploding loss curves).

Additionally, loss behavior was monitored across epochs. Our setup included:

- `num_train_epochs=3` — allowing for a full sweep over the training data.
- `save_strategy="epoch"` — saving model checkpoints at the end of each epoch to allow post-hoc analysis and rollback.
- `save_total_limit=2` — keeping the two most recent checkpoints to limit storage usage.

We observed that the training loss gradually decreased across epochs, confirming effective convergence. Validation loss and evaluation metrics (e.g., ROUGE) were recorded at the end of each epoch. While early stopping was not explicitly used, the logging granularity and checkpointing frequency provided sufficient means to diagnose training stability and overfitting.

By inspecting these logs, we ensured that the model remained well-behaved throughout training and that hyperparameters such as batch size and learning rate did not induce instability. Moreover, the logs enabled us to resume training safely from intermediate checkpoints (e.g., `checkpoint-1000`) in the event of resource interruptions.

Conclusion of Abstractive Summarization Stage

In this section, we presented the complete workflow for the abstractive summarization stage of our hybrid pipeline, leveraging the pretrained `facebook/bart-base` model. We detailed the rationale for selecting BART, its encoder-decoder architecture, and its alignment with the summarization task. Our implementation encompassed rigorous tokenization using BPE, dataset preparation with fixed-length truncation and padding, and training configured through the Hugging Face `Trainer` API. Hyperparameters were carefully chosen to ensure stability and reproducibility, with training progress monitored via structured logging and checkpointing.

The integration of domain-specific inputs—generated through extractive summarization and medical term filtering—allowed the BART model to operate on compact, semantically rich inputs, improving factual alignment in the generated summaries. Collectively, this stage serves as the generative core of our system, transforming structured findings into coherent, fluent, and clinically relevant impressions. The next chapter will evaluate the outputs of this model using both automatic and semantic metrics to assess summary quality and faithfulness.

3.5 Inference and Visualization

This chapter illustrates how our complete summarization pipeline performs in real-world use cases. We present qualitative examples by running inference on unseen test samples and visualizing the transformation across each stage of the summarization process.

After completing fine-tuning of the BART model on the Indiana radiology report dataset, we deployed the trained model to generate abstractive summaries on selected examples from the test set. This inference stage simulates how the system would operate in a practical clinical summarization workflow.

We initialized the inference pipeline using Hugging Face's `transformers` library. The checkpoint used was `checkpoint-2586`, corresponding to the final training step after three epochs. Both the tokenizer and the model were loaded from this checkpoint to ensure full consistency with the training configuration. The model was placed in evaluation mode using `model.eval()` and executed on a CUDA-enabled GPU when available.

The input data consisted of 13 manually selected samples, each containing an extractive summary produced by the BERT-based extractive module and filtered through domain-specific Named Entity Recognition (NER). These inputs were preprocessed using the same tokenization logic as in training. All inputs were truncated and padded to conform to fixed maximum sequence lengths, as required by the BART architecture.

Encoded inputs were passed to the model in inference mode using the `torch.no_grad()` context to avoid unnecessary gradient computations. The model produced output token sequences that were decoded into natural language summaries using the BART tokenizer.

3.5.1 Pipeline Overview

To evaluate the effectiveness and coherence of the proposed summarization system, we constructed a complete pipeline that integrates each major component of our hybrid architecture. This pipeline simulates the entire inference flow, beginning with raw radiology findings and culminating in the final abstractive summary generated by the fine-tuned BART model. By organizing the process into modular stages, we aim to both clarify the data transformations occurring at each step and assess the contribution of each module to the final output.

The pipeline was designed with real-world applicability in mind. In a practical hospital setting, radiology reports often require rapid summarization to support clinical decision-making or assist in electronic health record (EHR) automation. Our modular pipeline addresses this by breaking down the summarization task into distinct, interpretable phases that can be evaluated and optimized independently.

The five main stages of the pipeline are:

- **Input Findings:** Original findings section of the radiology report.
- **Extractive Summary:** Top-ranked sentences selected via BERT-based sentence similarity scoring.
- **Medical Term Extraction:** Domain-specific terms filtered using the SciSpaCy NER pipeline.

- **Abstractive Summary:** Free-form summary generated by the fine-tuned BART model.
- **Ground Truth:** Expert-written impression section provided in the Indiana dataset.

3.5.2 Side-by-Side Examples

Table 3.3 presents illustrative examples from the inference results. Each row showcases the evolution from findings to extractive summary, the domain-relevant medical terms, the generated abstractive summary, and the original ground truth impression.

Input Findings	**Extractive Summary**	**Medical Terms**	**Abstractive Summary**	**Ground Truth (Impression)**
There is mild bibasilar atelectasis. The heart size is normal. No pleural effusion.	mild bibasilar atelectasis. Heart size normal.	Atelectasis, Pleural effusion	Mild lung changes without acute cardiopulmonary abnormality.	no acute cardiopulmonary process. Mild bibasilar atelectasis.
Stable right lower lobe opacity, no pleural effusion or pneumothorax.	right lower lobe opacity stable. No effusion.	Opacity, Pleural effusion	Chronic right lower lobe infiltrate without evidence of acute disease.	Stable right lower lobe opacity. No acute process.
Mediastinal contours unchanged. Lungs clear. No acute bone findings.	Lungs clear. No bone findings.	Mediastinal, Bone	No radiographic evidence of acute cardiopulmonary disease.	Clear lungs. No acute findings.

Table 3.3: Side-by-side inference examples demonstrating the summarization pipeline.

3.5.3 Clinical Integration and Ethical Considerations

While the technical evaluation of the proposed summarization system demonstrates strong performance, it is essential to consider its real-world applicability within clinical environments. This section outlines potential pathways for integration into existing radiology workflows, and highlights important ethical and regulatory considerations.

Clinical Workflow Integration

The modular structure of the proposed pipeline facilitates its potential deployment within hospital infrastructures. Each component—namely the extractive summarizer, medical entity filter, and abstractive generator—can be deployed independently via REST APIs or containerized solutions (e.g., Docker with FastAPI), allowing seamless integration into electronic health record (EHR) systems and radiology information systems (RIS) [29]. Moreover,

a clinician-in-the-loop interface can be developed, where radiologists review, approve, or edit the generated summaries, ensuring safety and accountability while leveraging automation to reduce documentation workload.

In alignment with real-world AI governance frameworks such as the RAISE initiative[1], a robust deployment strategy should include post-deployment monitoring for model drift, degradation, or clinical misuse. Periodic audits and revalidation cycles would be required to ensure ongoing model reliability under changing clinical distributions [31].

Ethical, Privacy, and Regulatory Safeguards

Given the sensitivity of medical data and the critical nature of clinical decision-making, several ethical dimensions must be addressed. First, model transparency is paramount: although transformers like BART are often labeled as black-box models, interpretability can be improved by highlighting input findings that most influenced the final summary [28]. This not only builds clinician trust but also provides insight into the pipeline's decision rationale.

Second, data protection must be prioritized. While this thesis employed a de-identified public dataset (Indiana University Chest X-ray Collection), real-world deployment would involve patient data and must comply with privacy regulations such as the GDPR (General Data Protection Regulation), the European Union law governing data privacy and security. This necessitates secure data handling practices including audit logging, access control, and encrypted storage [32].

Third, attention must be paid to potential bias and fairness issues. The training data may reflect overrepresented diagnoses or demographic patterns, potentially affecting generalizability. As emphasized by prior work [30], evaluation should be stratified across patient groups to ensure the model does not reinforce health disparities.

Finally, from a liability perspective, the summarization system should be framed as a clinical decision support tool rather than an autonomous diagnostic system. Final responsibility must remain with human experts, and any deployment should include disclaimers and usage protocols accordingly.

Expert Feedback and Validation

Although formal clinician feedback was not collected in this thesis, the qualitative evaluations performed emulate typical radiology reporting behavior. Future work should include pilot testing in real clinical settings where radiologists rate the generated summaries based on accuracy, fluency, and usability. Such validation not only ensures clinical relevance but also fosters end-user engagement and iterative system improvement.

[1]Responsible AI for Safe Equitable Health, an initiative for ensuring responsible deployment of healthcare machine learning models

3.5.4 Discussion

The qualitative inference results presented earlier reveal several key observations about the behavior and efficacy of our hybrid summarization pipeline. Overall, the system demonstrates a strong ability to generate clinically appropriate summaries that not only preserve the core information from the extractive and named-entity-filtered input but also express that content in a fluent, structured, and condensed manner.

One notable strength of the pipeline is its ability to retain essential medical concepts across stages. This is largely attributed to the explicit inclusion of domain-specific named entities extracted by SciSpacy. These terms, which often correspond to anatomical regions (e.g., "right lower lobe") or diagnostic observations (e.g., "atelectasis"), act as anchors that guide the BART decoder to generate summaries that are both semantically relevant and diagnostically precise.

Furthermore, the linguistic fluency and structure of the abstractive output closely mirror that of real radiologist-authored impressions. This is a direct result of fine-tuning BART on the Indiana dataset, which contains high-quality clinical impressions as ground truth. The model learns to map complex and verbose descriptions into concise, standardized formats—a feature that aligns with real-world needs such as inclusion in electronic health records (EHRs), radiology information systems (RIS), or downstream diagnostic tools.

From a usability perspective, the pipeline exhibits robustness in handling mild variations in input complexity and length. The truncation and padding strategies introduced during preprocessing ensured compatibility with the BART model's architectural constraints without significant information loss. During inference, even cases with relatively short extractive summaries led to high-fidelity abstractive outputs, suggesting that the model generalizes well to both dense and sparse textual inputs.

However, limitations are also evident. In some cases, over-summarization or omission of nuanced medical findings occurs—particularly when input findings are vague or highly variable in expression. This points to the intrinsic challenge of lossy compression in abstractive summarization, where the model must strike a balance between brevity and completeness. Additionally, while the NER module ensures the presence of high-value terms, it does not enforce strict preservation, leaving room for occasional semantic drift.

Importantly, the inference phase highlights the practical readiness of the model for deployment. The pipeline is efficient, with runtime performance suitable for batch inference in clinical settings. It supports modular improvements—such as the substitution of the BERTSUM extractor or BART generator with more recent models—and is compatible with integration into RESTful APIs or containerized services (e.g., via Docker or FastAPI) for scalable deployment.

In conclusion, the inference results validate the architectural design choices of our hybrid summarization system. They also pave the way for further enhancements, such as adding factuality metrics, user feedback loops for human-in-the-loop correction, or post-editing modules tailored for clinical safety. Future work may also explore integrating retrieval-augmented generation (RAG) mechanisms or instruction-tuned LLMs to extend the system's capabilities beyond fixed-summary generation.

Chapter 4

Results and Evaluation

4.1 Overview of Evaluation Strategy

This chapter presents the quantitative evaluation of the two summarization modules proposed in this thesis: the extractive summarizer based on BERT (BERTSUM) and the abstractive summarizer based on BART. Each model is evaluated using both lexical and semantic metrics, with the goal of understanding their effectiveness in generating coherent, concise, and clinically relevant summaries from radiology reports.

To assess the lexical overlap between generated summaries and reference impressions, we employ the ROUGE metric (**Recall-Oriented Understudy for Gisting Evaluation**) [18]. ROUGE measures n-gram overlaps, providing recall-based scores for unigrams (ROUGE-1), bigrams (ROUGE-2), and longest common subsequences (ROUGE-L), and is a widely adopted standard in summarization literature. However, due to its reliance on exact word matching, ROUGE may penalize legitimate paraphrases or domain-specific synonyms that are common in medical narratives.

To capture the semantic similarity between the predicted summaries and the reference impressions, we further evaluate the models using BERTScore [25]. Unlike ROUGE, BERTScore computes pairwise token similarities using contextual embeddings from a pre-trained transformer (RoBERTa-large), and aggregates these into precision, recall, and F1 scores. This allows a more robust evaluation of meaning preservation, especially important in the medical domain where the same clinical concepts can be expressed in diverse ways.

Both metrics were computed separately on the training, validation, and test splits for each model. In the sections that follow, we report and analyze the results, highlighting performance differences between BERTSUM and BART, and discussing the implications of metric behavior in the context of clinical summarization.

4.2 Evaluation Metrics

To assess the quality of generated summaries in this work, we adopt both lexical and semantic evaluation strategies. This dual perspective provides a comprehensive view of how well the system captures both the surface structure and underlying meaning of the reference summaries. Specifically, we utilize the ROUGE metric family and BERTScore, which are well-established in the field of automatic summarization. Their inclusion is motivated by their complementary properties: ROUGE evaluates content overlap at the token level, while BERTScore captures contextual and semantic similarity using transformer-based embeddings.

4.2.1 ROUGE Metric Family

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) framework [18] has become a de facto standard for automatic evaluation of summarization systems. Developed to provide a computationally efficient and interpretable metric, ROUGE evaluates the quality of a generated summary by computing the overlap of textual units—such as unigrams, bigrams, and longest common subsequences—with a set of human-written reference summaries.

Despite being a relatively shallow, lexical matching technique, ROUGE has demonstrated a strong empirical correlation with human judgments of informativeness, particularly in domains where surface-level word choices closely mirror reference phrasings. This makes ROUGE especially useful in standardized evaluations such as the Document Understanding Conference (DUC) and the Text Analysis Conference (TAC), where it has been widely adopted.

One of ROUGE's strengths is its recall-oriented nature: it measures how much of the relevant content (as defined by human annotators) is successfully retrieved by the model. However, this focus on recall can also be a limitation, as it may not penalize overly verbose or redundant summaries. Additionally, because ROUGE operates at the token level, it struggles to account for semantic equivalence or paraphrasing—a significant shortcoming in domains such as clinical text, where terminology may vary but meaning remains consistent.

Nevertheless, due to its ease of use, domain-agnostic formulation, and strong alignment with early summarization benchmarks, ROUGE remains a foundational tool in both extractive and abstractive summarization research. In this thesis, we use several variants of ROUGE to quantify lexical similarity from multiple perspectives, complementing them with more semantically sensitive metrics like BERTScore for a holistic evaluation strategy.

In this thesis, we report the following variants:

- **ROUGE-1:** This metric quantifies the unigram (i.e., individual word) overlap between the generated summary and the reference summary. As the most basic form of ROUGE, ROUGE-1 evaluates the presence of relevant vocabulary from the gold summary in the system output. High ROUGE-1 scores indicate that the summary captures important terms and concepts mentioned in the reference. However, it does not account for the order or context of words, making it more indicative of content coverage than coherence or fluency.

- **ROUGE-2:** By measuring bigram (i.e., consecutive two-word sequence) overlap, ROUGE-2 captures short-range dependency and fluency within the summary. It provides a stricter evaluation than ROUGE-1 by considering whether adjacent word pairs from the reference are preserved in the prediction. This metric is sensitive to syntactic correctness, word ordering, and grammatical constructs, offering a proxy for evaluating local coherence and the ability of the model to produce naturally flowing language.

- **ROUGE-L:** ROUGE-L is based on the concept of the Longest Common Subsequence (LCS), which considers the longest sequence of words that appears in both the generated and reference summaries in the same order, though not necessarily contiguously. Unlike fixed-length n-grams, LCS captures variable-length matches, allowing for more flexible structural alignment. ROUGE-L is effective in identifying summaries that respect the original word ordering and sentence flow of the reference, which is particularly useful when assessing extractive fidelity or the structural preservation of ideas.

- **ROUGE-Lsum:** ROUGE-Lsum is a specialized adaptation of ROUGE-L tailored for document-level summarization. Instead of computing LCS on a sentence-by-sentence basis, it concatenates all sentences in both the reference and generated summaries into a single sequence and evaluates the global LCS-based recall. This metric is better suited for abstractive summarization tasks, especially when the output contains rephrased or merged content across sentences. It reflects the model's ability to preserve discourse-level structure and to capture cross-sentence dependencies that are typical in clinical reports and other long-form documents.

All ROUGE scores in this thesis are computed using stemming to reduce the impact of morphological variants (e.g., "enlarged" vs. "enlargement"), and are reported in terms of recall to remain aligned with the original metric design. Although ROUGE is limited in recognizing paraphrases or synonyms, it remains crucial for benchmarking content fidelity across different models.

4.2.2 BERTScore

To address the limitations of token-level overlap metrics like ROUGE, we incorporate BERTScore [25], a semantic evaluation method that leverages contextual embeddings from deep language models such as RoBERTa. Unlike ROUGE, which relies on exact word or phrase matches, BERTScore performs a more nuanced analysis by comparing the semantic similarity between words based on their context within the sentence.

BERTScore operates by encoding both the generated and reference summaries into dense vector representations using a pre-trained transformer model. For this thesis, we adopt **RoBERTa (Robustly Optimized BERT Approach)**, which is an advanced variant of BERT introduced by Liu et al. [26]. RoBERTa significantly enhances the capabilities of the original BERT architecture by employing several improvements: it removes the Next Sentence Prediction (NSP) objective, uses larger batch sizes, trains on significantly more data, and dynamically adjusts masking patterns during training. These modifications allow RoBERTa to capture richer semantic nuances and produce more accurate contextual embeddings for downstream tasks.

In the context of BERTScore, RoBERTa plays a central role in determining semantic similarity between individual tokens. The model generates contextualized embeddings that consider the full sentence structure, enabling it to differentiate between polysemous terms and detect paraphrased expressions. This is especially valuable in medical summarization, where synonymous terms (e.g., "heart enlargement" vs. "cardiomegaly") may not match lexically but are semantically equivalent. RoBERTa's robust representations help bridge this lexical gap, making BERTScore well-suited for evaluating clinical text generation tasks.

By computing pairwise cosine similarities between all tokens in the generated and reference texts, BERTScore derives a soft alignment that reflects how well each word in one sequence corresponds semantically to words in the other. These token-level similarities are then aggregated using precision, recall, and F1-score computations to provide a comprehensive metric for evaluating summarization quality. The inclusion of BERTScore ensures that our evaluation framework captures both the factual accuracy and the semantic fidelity of the model outputs, offering a more holistic assessment than surface-matching metrics alone.

Each token in the generated summary is matched with the most semantically similar token in the reference, and vice versa. These pairwise scores are aggregated into the following metrics:

- **Precision:** Precision quantifies the proportion of the content in the generated summary that is semantically aligned with the reference summary. It reflects the accuracy of the model in terms of generating relevant and correct information. A high precision score indicates that most tokens in the generated summary correspond to semantically meaningful tokens in the ground truth, minimizing the presence of hallucinated or fabricated details. In clinical summarization, high precision is critical to ensure that generated statements do not introduce erroneous diagnoses or misinterpret findings.

- **Recall:** Recall measures the extent to which the generated summary captures the semantic content of the reference summary. It focuses on coverage, penalizing the model for omitting clinically significant concepts that appear in the reference. A high recall value suggests that the model successfully includes the essential information from the reference impression, which is particularly important in the medical domain where missing even a single clinical observation could affect downstream interpretation or decision-making.

- **F1 Score:** The F1 score is the harmonic mean of precision and recall, serving as a balanced indicator that jointly accounts for both overgeneration (hallucination) and undergeneration (omission). It provides a single scalar value that reflects the overall semantic similarity between the generated and reference summaries. In BERTScore, F1 is often preferred as the principal evaluation metric because it encapsulates the trade-off between precision and recall, offering a more holistic view of summarization quality, especially in domains with strict factual requirements like radiology.

Unlike ROUGE, BERTScore can recognize that phrases such as "heart enlargement" and "cardiomegaly" are semantically equivalent, even though they have no lexical overlap. This makes BERTScore particularly effective in clinical summarization, where domain-specific

terminology and paraphrasing are frequent. Moreover, it captures nuance in language use that is critical in high-stakes domains like radiology.

Given its robustness to synonymy and paraphrasing, BERTScore serves as an essential complement to ROUGE in this work, allowing us to holistically evaluate the generated summaries along both surface and semantic dimensions.

4.3 Extractive Model Evaluation (BERTSUM)

The extractive summarization component, based on BERTSUM, was evaluated on the Indiana dataset using both ROUGE and BERTScore metrics. These evaluations were conducted on the train, validation, and test splits to assess lexical overlap and semantic fidelity between the extractive outputs and the ground-truth impressions.

To gain additional insight into the model's behavior and diagnose potential overfitting, we also report evaluation scores on the training set. While not indicative of generalization performance, these results serve as a reference point for comparing performance across splits.

ROUGE Results. As shown in Table 4.1, the extractive summaries yielded modest ROUGE scores, with ROUGE-1 ranging between 13.9 and 14.8 across the splits. ROUGE-2 and ROUGE-Lsum scores remained below 5.5 and 13.2 respectively. These scores, while consistent, highlight a critical limitation of ROUGE for extractive systems—namely, its sensitivity to n-gram ordering and surface-level match.

Table 4.1: ROUGE scores for extractive summarization using BERTSUM.

Split	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
Train	14.06	4.41	12.51	12.49
Validation	14.82	5.34	13.19	13.15
Test	13.96	4.31	12.58	12.59

While the reported ROUGE values for the BERTSUM model may initially appear suboptimal, they do not fully capture the actual semantic adequacy of the extractive summaries. This discrepancy arises from a fundamental limitation of ROUGE: it operates primarily at the token and n-gram level, assessing surface-level lexical overlap without accounting for the semantic intent behind sentence selections.

In the context of extractive summarization, where sentences are directly selected from the original findings section without modification, the generated summaries may fail to exhibit high n-gram overlap with the abstractive reference summaries, which often employ rephrased or compressed forms of the same content. For example, the sentence "There is a right-sided pleural effusion with associated atelectasis" may be expressed in the reference as "Fluid accumulation in the right pleura causing lung collapse." Despite expressing the same clinical observation, ROUGE would penalize this mismatch due to differing vocabulary and structure.

Moreover, extractive methods inherently lack the capacity to reorganize or condense content across multiple sentences, resulting in summaries that may be longer and lexically rigid. This rigidness leads to a reduction in ROUGE-2 and ROUGE-L scores, which are more sensitive to phrase order and sequence alignment. These limitations become even more pronounced in specialized domains such as radiology, where terminology is dense, synonyms are frequent (e.g., "cardiomegaly" vs. "enlarged heart"), and clinical meaning can remain intact even when surface forms diverge.

Therefore, while ROUGE remains useful for tracking basic content retention, its low scores in extractive summarization should be interpreted with caution. They reflect not a

deficiency in the summarization model, but rather the inadequacy of n-gram-based metrics in capturing clinically valid semantic equivalence.

BERTScore Results and Analysis. To address these limitations, BERTScore was employed to evaluate the semantic similarity between generated and reference summaries. As shown in Table 4.2, the BERTScore F1 remained consistently high across all splits—above 0.855—indicating strong semantic alignment between extracted and reference sentences.

Table 4.2: BERTScore F1 results for extractive summarization using BERTSUM.

Split	**F1**
Train	0.8554
Validation	0.8559
Test	0.8553

These results demonstrate that the extractive system effectively retains the semantic content of the reference summaries, even if lexical phrasing differs. For example, if the system includes "right pleural effusion" and the reference says "fluid in the right pleura," ROUGE would penalize the mismatch in phrasing, but BERTScore, which relies on contextual embeddings from RoBERTa, would recognize them as semantically equivalent.

Discussion. Given the nature of extractive summarization, which lacks paraphrasing or sentence compression, ROUGE metrics tend to underestimate performance. BERTScore, in contrast, is more robust for clinical NLP tasks due to its alignment with meaning rather than form. Therefore, in this thesis, **BERTScore is considered the more reliable metric** for evaluating the extractive model. The high F1 values suggest that BERTSUM captures the essential medical content faithfully, even when ROUGE scores appear low.

This finding aligns with broader critiques in the literature regarding ROUGE's limited applicability in domains with rich semantic variation and specialized terminology [25, 18]. In such contexts, embedding-based evaluation offers a more trustworthy lens on summary quality.

4.4 Abstractive Model Evaluation (BART)

At the heart of the proposed summarization pipeline is the fine-tuned BART model, which is responsible for performing the abstractive summarization task—generating clinically meaningful, linguistically coherent impression sections from the more detailed and verbose findings sections of radiology reports. Unlike extractive models that rely on sentence selection, the abstractive model learns to synthesize information, rephrase complex statements, and compress salient clinical observations into concise summaries. This transformation is particularly valuable in the medical domain, where effective communication of diagnostic insights must balance informativeness with clarity.

To evaluate the effectiveness and generalization of the fine-tuned BART model, extensive quantitative assessments were conducted on the Indiana dataset across three key data splits: training, validation, and test sets. These evaluations employed both ROUGE and BERTScore—two complementary families of metrics that provide insights into different dimensions of summary quality. ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum) assess the degree of lexical overlap and structural correspondence between generated and reference summaries, which is especially relevant for tracking the model's ability to recover explicit content. In contrast, BERTScore leverages contextual embeddings derived from deep language models (specifically RoBERTa) to measure semantic similarity at a token level, capturing the latent alignment of meaning between predictions and ground truth even in the absence of surface-level lexical matches.

This dual-metric evaluation approach is crucial in the clinical summarization setting, where synonymous medical expressions, domain-specific abbreviations, and variations in phrasing are common. For example, the model may generate "enlarged cardiac silhouette" in place of "cardiomegaly," or summarize multiple sentences into a compact phrase such as "findings consistent with congestive heart failure." While these outputs may diverge from the reference text in form, they can still maintain full clinical validity. Accordingly, combining ROUGE and BERTScore provides a more holistic perspective on model performance, capturing both surface-level alignment and deeper semantic fidelity. The results of these evaluations, presented in the following subsections, demonstrate that the fine-tuned BART model achieves high-quality summarization performance with strong generalization across data splits, particularly excelling in both fluency and domain-specific expressiveness.

ROUGE Results. Table 4.3 reports the ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-Lsum scores obtained by the fine-tuned BART model when evaluated on the test and validation sets of the Indiana radiology report dataset. These metrics collectively measure the degree of lexical and sequential overlap between the generated summaries and the corresponding human-written impressions.

Table 4.3: ROUGE scores for the fine-tuned BART model on test and validation sets.

Split	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-Lsum
Test	58.19	49.53	57.84	57.68
Validation	62.15	53.79	61.90	61.93

The ROUGE scores obtained on the test set — ROUGE-1: 58.19, ROUGE-2: 49.53, ROUGE-L: 57.84, and ROUGE-Lsum: 57.68 — indicate a strong lexical alignment between the predicted summaries and ground-truth impressions. These values suggest that the model is capable of reproducing critical clinical concepts while preserving natural language fluency.

Notably, the validation set results show even higher performance, with ROUGE-1: 62.15, ROUGE-2: 53.79, ROUGE-L: 61.90, and ROUGE-Lsum: 61.93. This consistency across both splits confirms that the model generalizes well and avoids overfitting to the training data. The higher validation scores may reflect better content alignment in that subset or slight variations in reference summary phrasing, but the overall trend supports strong model robustness.

In comparison to the extractive model (BERTSUM), which achieved substantially lower ROUGE values across all splits, the BART model demonstrates significant improvement in both content selection and generation fluency. The increase in ROUGE-2 and ROUGE-L scores, in particular, highlights BART's advantage in handling multi-word sequences and maintaining coherent structure, both of which are essential in clinical summarization where phrase ordering impacts interpretability.

It is important to emphasize that ROUGE, while primarily a lexical metric, remains a valuable first-layer indicator for summarization performance. In the case of BART, the high ROUGE scores confirm that the model not only learns the task well but also adapts effectively to the stylistic and terminological characteristics of radiology report summaries.

These ROUGE results are notable within the context of clinical summarization. Unlike open-domain summaries, radiology impressions are often formulaic and densely packed with medically significant phrases. The high n-gram overlap observed here reflects that the BART model has effectively learned both domain-specific vocabulary and the structural conventions used in radiology reporting. ROUGE-2 and ROUGE-L scores in particular indicate the model's ability to construct syntactically coherent and contextually accurate multi-word phrases — a key requirement for trustworthiness in clinical summaries.

Moreover, the ROUGE-Lsum score, which captures the longest common subsequence at the sentence level, suggests that the model successfully retains the broader structure of the reference impressions. This is particularly important for generating summaries that adhere to medical reporting standards where clause ordering and grouping carry clinical significance.

It is worth noting that while ROUGE remains a surface-level metric focused on lexical similarity, the consistently strong scores across both test and validation sets offer compelling evidence that the model is not simply memorizing the training data, but rather internalizing a robust representation of radiological language. The sharp contrast between the abstractive and extractive results further highlights the power of generative modeling in capturing the nuanced, paraphrased nature of clinical impressions.

BERTScore Results. To capture deeper semantic alignment between predictions and references—especially important in clinical contexts with high terminology variability—we used BERTScore [25]. Table 4.4 reports the F1 scores across all splits, demonstrating consistently strong alignment. The F1 on the test set was 0.9293, validating that BART-generated summaries preserved essential medical meaning despite surface phrasing differences.

Table 4.4: BERTScore F1 across dataset splits using BART.

Split	F1 Score
Train	0.9387
Validation	0.9355
Test	0.9293

These BERTScore values greatly exceed those of the extractive model and reinforce the superior capability of transformer-based generative models in clinical NLP tasks, particularly for abstractive summarization. While the extractive approach is constrained to reusing exact sentences from the input, leading to surface-level matches that may be penalized by lexical metrics like ROUGE, the generative nature of the BART model allows it to reformulate findings into concise, medically faithful impressions that maintain semantic integrity even in the absence of direct lexical overlap.

Traditional string-matching metrics such as ROUGE-1 and ROUGE-2 are inherently limited in their ability to recognize valid paraphrastic expressions. For example, the generated phrase "cardiac enlargement" and the reference phrase "cardiomegaly" represent clinically equivalent concepts, yet ROUGE would underweight this similarity due to the lack of shared n-grams. This limitation is particularly pronounced in domains like radiology, where expert authors may describe the same finding using a variety of synonyms, abbreviations, or stylistic conventions depending on context, diagnostic certainty, or institutional norms.

BERTScore, in contrast, circumvents these surface-level limitations by comparing contextual embeddings of tokens rather than raw word forms. By employing a pre-trained transformer encoder such as RoBERTa, BERTScore computes pairwise cosine similarities between tokens in the generated and reference summaries, capturing nuanced relationships that reflect meaning rather than exact wording. This mechanism makes BERTScore especially suitable for evaluating abstractive summarization in specialized domains, where maintaining clinical accuracy through paraphrased content is both necessary and beneficial.

The high BERTScore F1 values observed across all data splits—0.9387 on the training set, 0.9355 on the validation set, and 0.9293 on the test set—demonstrate that the fine-tuned BART model consistently produces semantically rich and clinically valid summaries. The small variation in scores across splits also suggests strong generalization, indicating that the model does not simply memorize training data but learns robust patterns for content generation. These high scores are a clear testament to the effectiveness of abstractive summarization when combined with a powerful contextual language model and a domain-specific fine-tuning regime.

Moreover, BERTScore's precision, recall, and F1 submetrics (when available) allow for fine-grained analysis of whether the model tends to generate extraneous content (low precision) or omit important information (low recall). In this study, the consistently strong F1 suggests a good balance between informativeness and conciseness. Given its semantic sensitivity and alignment with human judgment in prior benchmarks, BERTScore emerges as the most reliable indicator of quality in this setting.

Overall, the high BERTScore values validate both the architectural choice of BART for this task and the fine-tuning strategy applied to radiology reports. By incorporating this

semantic metric into the evaluation pipeline, this work ensures that the assessment captures the true clinical utility of generated summaries—not merely their lexical resemblance to reference texts, but their ability to convey accurate, readable, and diagnostically relevant content.

Discussion. The evaluation results demonstrate that the fine-tuned BART model achieves a compelling balance between lexical accuracy and semantic coherence, as reflected in both ROUGE and BERTScore metrics. These findings not only highlight the superiority of the abstractive approach over extractive methods, but also emphasize the importance of transformer-based architectures that are explicitly adapted to domain-specific tasks. The substantial improvement across all evaluation splits confirms that BART is not merely memorizing training data but has developed a nuanced capability to generalize clinical language patterns in radiological reporting.

Crucially, the model exhibits an ability to preserve clinically salient information across varied patient cases. Summaries generated by BART consistently include key pathological descriptors (e.g., "effusion," "consolidation," "atelectasis"), spatial qualifiers (e.g., "bilateral," "right lower lobe"), and clinical interpretations (e.g., "suggestive of heart failure," "consistent with pneumonia"), demonstrating a deep alignment with radiologist-style abstraction. This strengthens the case for deploying such models in assistive documentation tools for radiology workflows, where summarization quality directly affects diagnostic clarity and clinical decision-making.

Additionally, these results validate the role of domain-aware fine-tuning. By training BART on a curated corpus of radiology reports, the model internalizes not just vocabulary, but also the pragmatic structure of findings-to-impression transitions that are typical in diagnostic narratives. The consistent performance across test and validation sets shows that the model learns meaningful generalizations, rather than relying on heuristic memorization or spurious correlations.

Another important observation is the model's stability across decoding settings. The use 'early- stopping=True' proved effective in maintaining fluency without introducing factual hallucinations—a frequent concern in medical text generation. Manual inspections of sampled outputs further confirmed that BART-generated summaries were not only grammatically coherent but also clinically sensible, often echoing radiologist phrasing even when no such wording appeared in the source findings.

These advantages, taken together, illustrate the central role of BART in this hybrid summarization framework. It serves not just as a rewriting engine, but as a semantic compressor that transforms verbose clinical observations into compact, high-utility summaries. This is particularly valuable in medical contexts where time constraints and cognitive load demand high-quality, immediately actionable textual outputs.

Finally, the observed performance supports trends in recent literature suggesting that abstractive summarizers—when properly adapted to the target domain—outperform extractive methods not just on automated metrics but also in downstream clinical utility [27, 25]. The findings reinforce the need to consider semantic-aware evaluation metrics and contextual generation capabilities when designing medical NLP systems intended for real-world clinical integration.

Conclusion. Given its superior performance on both ROUGE and BERTScore metrics, the fine-tuned BART model demonstrates a high degree of effectiveness in domain-aware summarization of radiology reports. Beyond numerical performance, its capacity to generate fluent, concise, and clinically relevant summaries underscores its potential to serve as a robust foundation for intelligent medical documentation tools.

The success of this model lies not only in its architecture but also in the synergistic alignment between its pretraining on general language and fine-tuning on radiology-specific corpora. This alignment enables BART to abstract over low-level token patterns and capture higher-order clinical relationships, such as symptom-diagnosis mappings or anatomical-pathological links. The model's ability to synthesize findings into structured impressions further highlights its suitability for real-world diagnostic environments where interpretability and precision are paramount.

Moreover, its integration into a hybrid summarization pipeline strengthens the overall system's reliability—extractive components ensure factual consistency while BART introduces linguistic flexibility and narrative coherence. This dual-role design mitigates known risks such as hallucination, omission of critical details, or inconsistent phrasing, which are common challenges in clinical text generation.

Looking forward, the BART model's strong performance paves the way for broader exploration of large-scale generative models in clinical domains, including multimodal summarization (e.g., combining image findings with textual descriptions), interactive radiology reporting assistants, or multilingual medical NLP systems. Its success also raises the possibility of extending similar techniques to other structured-reporting tasks in medicine, such as pathology, oncology, or surgical notes.

In conclusion, BART's demonstrated capacity for accurate, semantically grounded, and linguistically fluent summarization marks it as a key enabler for next-generation clinical NLP applications. Its integration into this thesis project exemplifies how carefully tuned transformer-based models can bridge the gap between raw diagnostic data and human-readable, actionable insights.

4.5 Comparative Analysis

This section presents a side-by-side evaluation of the extractive and abstractive summarization models—BERTSUM and BART respectively—based on their performance across ROUGE and BERTScore metrics. The objective is to quantify the improvement introduced by the generative BART model and provide insights into the nature of these gains from both lexical and semantic perspectives.

Table 4.5: Comparison of BERTSUM (extractive) and BART (abstractive) across evaluation metrics on the test set.

Model	ROUGE-1	ROUGE-2	ROUGE-Lsum	BERTScore F1
BERTSUM (Extractive)	13.96	4.31	12.59	0.8553
BART (Abstractive)	58.19	49.53	57.68	0.9293

Lexical Improvements (ROUGE). The BART model substantially outperforms the extractive BERTSUM model in all ROUGE metrics, with ROUGE-1 increasing from 13.96 to 58.19, and ROUGE-2 from 4.31 to 49.53. This improvement reflects BART's ability to not just select relevant sentences, but to **synthesize content**, **rephrase** diagnostic descriptions, and maintain coherent discourse across multi-sentence summaries.

These gains are particularly striking in ROUGE-2 and ROUGE-Lsum, which are sensitive to bigram ordering and long-range sequence alignment, respectively. The increase in ROUGE-2 from 4.31 to 49.53 suggests that BART not only captures correct information but also phrases it in a way that closely mirrors expert-written reference impressions. ROUGE-Lsum, increasing from 12.59 to 57.68, further reinforces that BART produces summaries with high structural fidelity—maintaining logical flow, sentence boundaries, and medical progression in a clinically sensible manner.

Semantic Gains (BERTScore). While ROUGE emphasizes surface-level overlap, BERTScore evaluates **semantic equivalence** using contextual embeddings from a pretrained transformer (RoBERTa). The BERTScore F1 jumps from 0.8553 (BERTSUM) to 0.9293 (BART), underscoring BART's superior ability to preserve the **meaning** behind clinical findings—even when phrasing diverges significantly.

This improvement validates that BART generates summaries not only closer in vocabulary, but also more aligned in **diagnostic intent**, **pathological descriptors**, and **anatomical terminology**. For instance, BART can paraphrase "right-sided pleural effusion" as "fluid in the right lung lining" and still score highly in BERTScore, while BERTSUM would fail to capture such variation.

Qualitative Perspective. From a qualitative standpoint, BERTSUM's summaries tend to be longer, occasionally redundant, and tied rigidly to the sentence structures of the original

findings. In contrast, BART's outputs are more **fluent**, **compact**, and closer to radiologist-style impressions, often fusing information from multiple source sentences into one coherent diagnostic statement. This aligns well with the stylistic expectations in real-world clinical practice.

Interpretation. The sharp contrast in both ROUGE and BERTScore highlights the limitations of extractive methods in complex medical summarization tasks. Extractive models are bound by the original report structure and lack the capability to perform compression or abstraction. Abstractive models like BART, when properly fine-tuned, offer a **substantial leap in summary quality**—producing outputs that better reflect the goals of radiology reporting: brevity, clarity, and diagnostic relevance.

In conclusion, the comparative analysis confirms the effectiveness of the hybrid pipeline: BERTSUM contributes strong factual grounding, while BART transforms this input into clinically elegant summaries. Together, they form a cohesive system optimized for both precision and readability.

4.6 Interpretation and Insights

This section distills the findings from the quantitative evaluations into qualitative insights, reflecting on the nature of model outputs, clinical reliability, and the broader implications of adopting a hybrid summarization approach for medical applications.

Understanding the Scores. The evaluation metrics—ROUGE and BERTScore—serve complementary roles in assessing summarization quality. ROUGE highlights lexical fidelity and structural alignment, while BERTScore assesses meaning preservation via contextual similarity. The consistently high BERTScore F1 values across both models, especially the fine-tuned BART model (0.9293 on the test set), affirm that semantic equivalence between generated and reference summaries is well-maintained.

However, the contrast in ROUGE scores between BERTSUM and BART is particularly illuminating. While the extractive model yields modest ROUGE-1 scores (14.0) due to strict reliance on original sentence structures, the abstractive BART model achieves ROUGE-1 scores above 58. This dramatic improvement illustrates that BART can generate lexically faithful summaries that also reflect radiologist-style phrasing, without copying verbatim.

Quality of Generated Summaries. From a clinical standpoint, BART's outputs exhibit several strengths. First, they show enhanced fluency and cohesion, often condensing complex radiology findings into compact statements. Second, key medical terms—such as anatomical references ("right lower lobe"), pathologies ("pleural effusion"), and modifiers ("mild," "chronic")—are retained or paraphrased with high precision. Third, BART often restructures information logically, matching the discourse style of actual impression sections written by radiologists.

BERTSUM, while valuable for anchoring to source text, exhibits limitations in phrasing and content reorganization. Its reliance on sentence selection leads to occasional redundancy, syntactic disfluencies, and lack of abstraction. These traits make it less ideal for impression-style summarization, although it excels in information preservation.

Effectiveness of the Hybrid Pipeline. The hybrid pipeline—consisting of an extractive stage (BERTSUM) followed by a generative stage (BART)—yields measurable advantages. By first narrowing the content to the most relevant sentences, the BERTSUM model reduces noise and guides the abstractive model's attention toward diagnostically significant findings. This staged filtering is especially helpful in medical domains, where not all information in the findings section is summary-worthy.

Furthermore, by applying Named Entity Recognition (NER) during preprocessing, the pipeline ensures that essential clinical terms are retained through both extractive and abstractive phases. This design preserves domain fidelity, minimizes hallucination risk, and improves trustworthiness of the final output.

Clinical and Research Implications. These insights hold broader implications for clinical NLP systems. Transformer-based abstractive models, when fine-tuned properly and

guided by domain-aware preprocessing, can produce summaries that approach the quality of expert-authored impressions. This positions such systems as potential assistants in radiological documentation, second-opinion generation, and downstream clinical decision support.

Moreover, this evaluation framework—combining both lexical and semantic metrics—serves as a reproducible benchmark for future research in clinical summarization. It illustrates that metric diversity is crucial to avoid overestimating or underestimating model performance.

In summary, the evaluation results and their interpretation validate the utility of a hybrid summarization architecture. The combination of BERTSUM and BART captures the best of both worlds: factual correctness from extractive methods and linguistic expressiveness from generative models, tailored to the nuanced demands of clinical reporting.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis presented a hybrid deep learning framework for radiology report summarization, combining extractive and abstractive techniques guided by medical entity filtering. Centered around the Indiana University Chest X-ray Collection, the proposed approach systematically transforms verbose findings sections into concise, clinically accurate impression-style summaries.

The summarization pipeline consists of three distinct yet interlinked stages. First, an extractive summarization module based on BERTSUM was used to identify the most salient and informative sentences within the findings. This step provides a reliable content backbone grounded in original report context. Second, a domain-specific named entity recognition (NER) component utilizing SciSpaCy filtered the selected content to retain only clinically meaningful entities—such as anatomical structures, pathological terms, and diagnostic indicators—ensuring that downstream processing focuses on medically relevant information. Third, the filtered output was fed into an abstractive generation module, fine-tuned from the BART transformer architecture, to produce coherent and linguistically fluent summaries that emulate the writing style of human-authored impressions.

The results of extensive empirical evaluation underscore the effectiveness of this hybrid framework. The extractive component achieved a BERTScore F1 of **0.8553**, suggesting strong semantic overlap between the extracted content and the ground-truth summaries. More significantly, the fine-tuned BART model achieved a ROUGE-L score of **57.84** and a BERTScore F1 of **0.9293** on the test set, demonstrating its ability to generate semantically faithful and stylistically appropriate summaries. These scores represent a marked improvement over the extractive baseline, highlighting the benefits of integrating domain-specific filtering and transformer-based generation.

Qualitative inspection of model outputs further confirmed that the generated summaries preserved essential clinical facts and terminology. The hybrid architecture effectively addressed the limitations associated with standalone extractive models—such as lack of abstraction and redundancy—as well as the risks of hallucination and irrelevance often associated with unguided generative models. By combining both strategies in a structured pipeline, the system maintains factual accuracy while delivering outputs with high fluency

and diagnostic clarity.

This research contributes to the growing field of medical NLP by demonstrating how tailored hybrid approaches can meet the unique requirements of clinical summarization: balancing fidelity to source content with the expressive capacity to condense and paraphrase. The proposed framework is both practical and scalable, with potential applications in clinical documentation systems, report standardization tools, and automated diagnostic assistants.

5.2 Future Work

The proposed hybrid summarization framework has demonstrated robust and reliable performance across both lexical and semantic evaluation metrics, affirming its value as a clinically meaningful and technically sound solution for radiology report summarization. Nonetheless, the field of medical NLP continues to evolve, offering promising opportunities to further extend and augment the capabilities of the current system. Below are several directions for future work that aim to **enhance, not replace**, the strong foundation established in this thesis:

- **Multi-Modal Integration with Imaging Features:** Radiology inherently bridges visual and textual modalities. Extending the framework to incorporate visual features from chest X-rays—using image encoders aligned with textual transformers—could enhance factual grounding and allow the model to cross-reference image findings with textual impressions, opening doors to multimodal summarization pipelines.

- **Factuality Optimization via Reinforcement Learning:** Although the current BART model already performs well in preserving clinical accuracy, further gains could be achieved through reinforcement learning. By introducing task-specific reward functions (e.g., penalizing hallucinated findings or rewarding correct disease mentions), the system could be fine-tuned to align even more closely with radiologist expectations and clinical correctness.

- **Cross-Dataset Generalization and Expansion:** The Indiana Chest X-ray dataset provided a strong and structured foundation for training and evaluation. As a natural extension, applying the trained model to additional datasets—such as MIMIC-CXR or CheXpert—could validate its adaptability across institutions and writing styles. Rather than being a limitation, this cross-dataset application would serve to amplify the already-demonstrated strength of our approach by showcasing its transferability and robustness in varied clinical settings.

- **User-Centered Clinical Evaluation:** While automatic metrics confirm the technical efficacy of the system, engaging domain experts such as radiologists in qualitative evaluations would yield valuable insights into the clinical relevance, readability, and trustworthiness of generated summaries. These expert assessments could help define new success criteria grounded in real-world diagnostic utility.

- **Interactive and Editable Summarization Interfaces:** To further support clinical integration, future versions of the system could include an interactive interface where

clinicians can view, edit, or approve generated summaries. Features like entity highlighting, confidence scores, and reasoning explanations would increase transparency and usability in healthcare environments.

- **Model Optimization for Deployment:** Although this thesis focused on achieving high summarization quality, deployment considerations such as runtime performance, memory efficiency, and hardware portability are critical for real-world use. Future work could explore model compression, quantization, or on-device optimization techniques to ensure scalability in clinical settings.

In summary, this thesis sets a high-performance benchmark for hybrid radiology summarization systems. The proposed extensions aim not to compensate for weaknesses, but to build upon an already strong system by expanding its functionality, interpretability, and integration within broader medical NLP ecosystems. These enhancements will further solidify the framework's role as a reliable and practical tool for automated clinical documentation.

Bibliography

[1] Zhang, Y., et al. (2021). *Contrastive Learning of Medical Visual Representations from Paired Images and Reports*. arXiv:2010.00747.

[2] Kumar, A., et al. (2022). *Extractive Summarization of Radiology Reports using BioBERT Embeddings*. In Proceedings of EMNLP.

[3] Lee, J., et al. (2023). *Entity-Aware Abstractive Summarization in Clinical Reports*. Journal of Biomedical Informatics.

[4] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.

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

[6] Vaswani, A., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.

[7] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. *OpenAI*.

[8] Raffel, C., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140), 1–67.

[9] Liu, Y., & Lapata, M. (2019). *Text Summarization with Pretrained Encoders*. arXiv preprint arXiv:1908.08345.

[10] R. Ghosh, S. K. Karn, M. D. Danu, L. Micu, R. Vunikili, and O. Farri, "RadLing: Towards Efficient Radiology Report Understanding," *arXiv preprint arXiv:2306.02492*, 2023.

[11] S. Wang, M. Lin, Y. Ding, G. Shih, Z. Lu, and Y. Peng, "Radiology Text Analysis System (RadText): Architecture and Evaluation," *arXiv preprint arXiv:2204.09599*, 2022.

[12] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," *arXiv preprint arXiv:1910.13461*, 2019.

[13] V. Ahuir, E. Segarra, and L.-F. Hurtado, "ELiRF-VRAIN at BioNLP Task 1B: Radiology Report Summarization," in *Proceedings of the 22nd BioNLP Workshop*, 2023, pp. 524–529.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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*, 2019, pp. 4171–4186.

[15] X. Song, A. Salcianu, Y. Song, D. Dopson, and D. Zhou, "Fast WordPiece Tokenization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 2089–2103.

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

[17] Z. Guo, M. Schlichtkrull, and A. Vlachos, "A Survey on Automated Fact-Checking," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 178–206, 2022.

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

[19] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, Florence, Italy, 2019, pp. 319–327.

[20] Tan, S. W., Lee, C. P., Lim, K. M., Tee, C., and others. *QARR-FSQA: Question-Answer Replacement and Removal Pretraining Framework for Few-Shot Question Answering.* IEEE Access, 2024, PP(99):1–1. DOI: 10.1109/ACCESS.2024.3487581.

[21] Cao, S., Wang, L., Li, X., Bing, L., & Lam, W. (2020). *Factual Error Correction for Abstractive Summarization Models.* arXiv preprint arXiv:2005.00661.

[22] Loshchilov, I., & Hutter, F. (2017). *Decoupled Weight Decay Regularization.* arXiv preprint arXiv:1711.05101.

[23] Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., ... & Shoeybi, M. (2018). *Mixed Precision Training.* arXiv preprint arXiv:1710.03740.

[24] Holtzman, A., Buys, J., Du, L., Forbes, M., and Choi, Y. (2019). *The Curious Case of Neural Text Degeneration*. arXiv preprint arXiv:1904.09751.

[25] Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., and Artzi, Y. (2020). *BERTScore: Evaluating Text Generation with BERT*. In *International Conference on Learning Representations (ICLR)*.

[26] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv:1907.11692, 2019.

[27] A. Kho, C. Lee, and D. Ghosh. *Automated Summarization of Radiology Reports Using Pretrained Transformers: A Comparative Study*. In *Journal of Biomedical Informatics*, 2023. DOI: https://doi.org/10.1016/j.jbi.2023.104384

[28] Jung, K.H. (2025). *Large Language Models in Medicine: Clinical Applications, Technical Challenges, and Ethical Considerations*. Healthcare Informatics Research, 31(2), 114–124.

[29] MDPI. (2024). *The Integration of Artificial Intelligence into Clinical Practice*. AI in Healthcare.

[30] Chen, I.Y., Pierson, E., Rose, S., Joshi, S., Ferryman, K., & Ghassemi, M. (2020). *Ethical Machine Learning in Health Care*. arXiv:2009.10576.

[31] Weiner, E.B., Dankwa-Mullan, I., Nelson, W.A., & Hassanpour, S. (2024). *Ethical Challenges and Evolving Strategies in the Integration of Artificial Intelligence into Clinical Practice*. arXiv:2402.10287.

[32] Wikipedia contributors. (2025). *Artificial intelligence in healthcare*. Retrieved from https://en.wikipedia.org/wiki/Artificial_intelligence_in_healthcare