Comparative analysis of medical pseudo-report generation using LLM and external knowledge bases

Sharjil Dhanani #23269428, Ronit Loke #23265555,
School of Computing, Dublin City University, Ireland
Email: sharjil.dhanani3@mail.dcu.ie ronit.loke2@mail.dcu.ie

Abstract—**Automated radiology report generation aims to ease the burden on radiologists and boost diagnostic accuracy by leveraging artificial intelligence. In this study, we introduce an innovative method that combines data from chest X-ray images and radiology reports, utilizing pre-trained ResNet-50 and BERT models. To improve the understanding of the context, we incorporate structured knowledge from Radiopaedia articles through triplet extraction. Our system uses the ALBEF model, which aligns image and text features using a contrastive learning framework. Additionally, we employ advanced Large Language Models (LLMs) like GPT-4 to produce detailed and accurate radiology reports. We tested our approach using metrics such as BERTScore and TF-IDF score, which showed significant improvements in precision, recall, and F1 scores. Our findings suggest that our method can create context-aware radiology reports, which could be a valuable tool for clinical workflows. This approach is versatile and can be customized to fit various clinical settings, making it a promising solution for enhancing automated radiology report generation.**

I. Introduction

Automated Radiology Report Generation Systems hold significant promise for enhancing the efficiency and accuracy of radiologists by generating detailed reports for review. These systems can identify various attributes such as organ systems, pathology, abnormalities, severity, and location of findings [9]. The integration of visual and language features in large language models (LLMs) is crucial for medical imaging tasks, as accurate medical image analysis and generation require a seamless mapping between visual and textual information.

Recent advancements have focused on training adapter networks to bridge image processing and LLMs, but these approaches may limit the model's ability to fully leverage the interaction between visual and language features. A novel method for instruction-tuning LLMs has been developed, enhancing capabilities in tasks such as CXR-to-report generation and report-to-CXR generation, achieving state-of-the-art performance in understanding and generating medical images [5].

The progress in vision-language tasks has enabled the generation of high-quality radiology reports from medical images. Evaluating these generated reports necessitates robust metrics that capture the clinical relevance and accuracy of the content. The RadCliQ composite metric has been proposed to better align automated evaluations with those of radiologists, providing a more accurate measurement of report generation quality [13]. Early approaches have treated report generation similarly to image captioning, using generative models to produce descriptive reports. The CXR-RePaiR system, for example, approached report generation as a retrieval problem, leveraging a database of diagnostic details to enhance accuracy. However, these retrieval-based methods face challenges, such as irrelevant content and hallucinations in the generated reports . Recent advancements like the CXR-ReDonE [8] system have addressed some of these issues by introducing a dataset that eliminates prior references, improving the quality and relevance of generated reports [9].

To address these issues, recent work has proposed the Retrieval Augmented Generation (RAG) methodology. This approach combines the strengths of retrieval-based systems with generative capabilities, leveraging multimodally aligned embeddings from contrastively pretrained vision-language models. The RAG methodology reduces hallucinations by grounding the generated reports in relevant radiology text retrieved from a corpus, thereby enhancing clinical metrics such as BERTScore and Semb score [9].

Building on these advancements, our study proposes a novel approach that integrates multi-modal data from chest X-ray images and radiology reports using pre-trained ResNet-50 and BERT models. To enhance contextual understanding, structured knowledge from Radiopaedia articles is incorporated through triplet extraction . Our multi-modal architecture employs a contrastive learning framework using the ALBEF model to effectively align image and text features [4]. Additionally, we utilize Large Language Models (LLMs) such as GPT-4 for generating detailed and contextually accurate radiology reports. The performance of our method is evaluated using various metrics, including BERTScore and TF-IDF score, demonstrating significant improvements in precision, recall, and F1 scores [10].

This study addresses these research questions by providing a comprehensive evaluation of the proposed methodology, highlighting the importance of integrating external knowledge bases and ensuring semantic coherence in automated radiology report generation. The findings underscore the potential of AI-driven approaches to enhance clinical workflows and improve diagnostic accuracy, offering a robust tool for generating high-quality, contextually relevant radiology reports.

- **Research Questions Addressed:**

1) **To what extent does the inclusion of external knowledge bases enhance the quality and relevance of generated pseudo-reports?**

   The inclusion of structured knowledge from Radiopaedia articles significantly enhances the quality and relevance of generated pseudo-reports by providing additional contextual information. Triplet extraction and embedding generation from Radiopaedia articles add valuable context that improves the model's ability to generate accurate and contextually rich reports. The integration of these triplets with image and text embeddings ensures that the most pertinent information is used in report generation, resulting in reports that closely match the actual findings in terms of semantic content and contextual relevance. The performance improvement is demonstrated through high BERTScore [6] and TF-IDF scores, indicating the effectiveness of incorporating external knowledge.

2) **How consistently does the generated pseudo-report maintain semantic coherence within the medical context?**

   The generated pseudo-reports consistently maintain semantic coherence within the medical context due to the alignment of multi-modal embeddings using the ALBEF model and the detailed prompts that guide report generation. The ALBEF model's contrastive learning framework ensures that image and text features are effectively aligned in a common feature space, facilitating coherent integration of visual and textual information. Detailed prompts that include specific attributes such as pathology, positional, severity, and size words ensure that the generated reports use consistent and relevant medical terminology. Evaluation metrics such as BERTScore and TF-IDF cosine similarity confirm the high degree of semantic similarity and contextual accuracy of the generated reports, demonstrating the model's capability to maintain coherence within the medical context.
   These research questions provide a comprehensive evaluation of the proposed methodology, highlighting the importance of integrating external knowledge bases and ensuring semantic coherence in automated radiology report generation. The findings underscore the potential of AI-driven approaches to enhance clinical workflows and improve diagnostic accuracy, offering a robust tool for generating high-quality, contextually relevant radiology reports.

II. Literature Review

The utilization of Large Language Models (LLMs) for text generation, particularly in generating automated radiology reports, is a burgeoning area of research. This study highlights the capabilities of LLMs in producing coherent and contextually relevant text, with a focus on prompt engineering, retrieval mechanisms, and integrating medical knowledge to tailor LLMs to domains like medical radiology. The primary goal is to generate accurate and clinically relevant reports while addressing challenges related to visual and textual bias and knowledge integration. One aspect within this research is the accuracy and nuance in text generation within the medical domain. For instance, in the study on Retrieval Augmented Chest X-Ray Report Generation, multimodally aligned embeddings from a pretrained vision-language model were used for retrieval, coupled with OpenAI's LLMs for generation. This approach, evaluated against clinical metrics, demonstrated improved clinical metrics and reduced hallucinated content, with higher BERTScore, Semb score, and RadGraph F1 , indicating enhanced semantic accuracy and medical terminology precision [13]. Another study explored the standalone effectiveness of LLMs like GPT-4 in generating medically accurate reports, showing significant improvements in BERTScore, Semb score, and RadGraph F1 when evaluated against the MIMIC-CXR dataset [**?**]. Furthermore, a knowledge-enhanced approach integrating general and specific medical knowledge into an encoder-decoder framework with a novel multi-head attention mechanism was found to outperform state-of-the-art models in BLEU scores, CIDEr, and ROUGE-L metrics [11]. Evaluation metrics for AI-generated reports constitute a significant aspect of the research. One study focused on developing novel metrics like RadGraph F1 and RadCliQ, which measure clinical accuracy and relevance by calculating overlaps in clinical entities and relations between AI-generated and ground-truth reports. These metrics provide a holistic assessment of clinical quality, aligning better with radiologist evaluations [13]. Prompt engineering is also a critical element of this research. It has shown that using structured and unstructured prompts in LLMs significantly enhances the accuracy and relevance of generated radiology reports, enabling customization of report format and content [13]. Moreover, high fidelity in text generation from visual inputs has been demonstrated, as evidenced by the CXR-to-Report Generation study using the CheXpert-labeler [9]. Another major focus is knowledge injection into AI models to enhance the accuracy and clinical relevance of generated reports. Studies in this area commonly use knowledge graphs and custom retrieval methods, integrating both general and specific medical knowledge. For example, incorporating clinical information into the tokenization process significantly improves model performance in both CXR-to-report and report-to-CXR tasks [9]. Additionally, the enhanced performance of models in vision-language tasks was demonstrated through superior accuracy in answering questions about CXR images, showcasing an advanced understanding of medical imagery and text [9]. Instruction-finetuning approaches further improved vision-language integration compared to traditional adapter networks, leading to better-aligned features [9]. In summary, research on text generation with LLMs and knowledge injection highlights significant advancements in generating accurate, clinically relevant radiology reports. These studies underscore the importance of prompt engineering, evaluation metrics, and integrating medical knowledge, paving the way for enhanced clinical documentation and patient care.

III. Methodology

This study offers a thorough approach to automated radiology report generation by integrating multi-modal data from chest X-ray images and radiology reports. The methodology encompasses several key steps: data preparation, feature extraction, triplet extraction, embedding generation, report generation using GPT-4, and comparative analysis with GPT-4O.

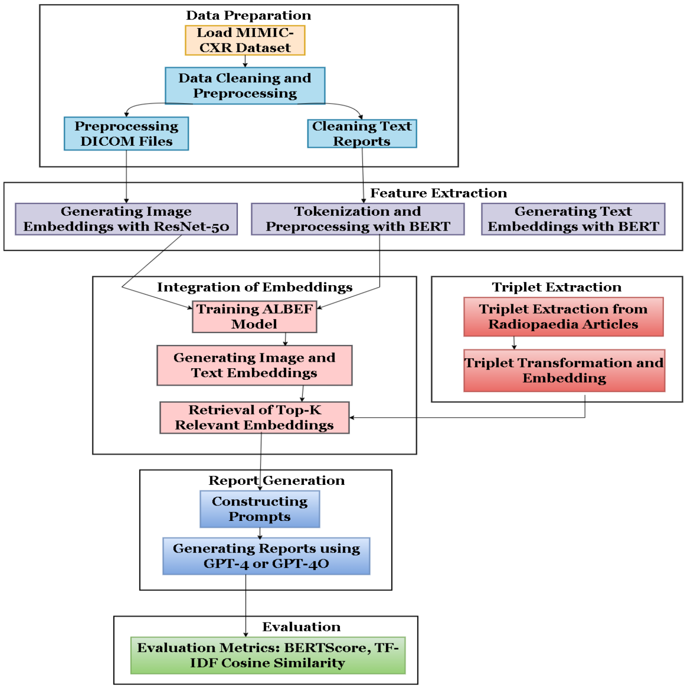
Data Preparation
Load MIMIC-CXR Dataset
Data Cleaning and Preprocessing
Preprocessing DICOM Files
Cleaning Text Reports
Feature Extraction
Generating Image Embeddings with ResNet-50
Tokenization and Preprocessing with BERT
Generating Text Embeddings with BERT
Integration of Embeddings
Training ALBEF Model
Generating Image and Text Embeddings
Retrieval of Top-K Relevant Embeddings
Triplet Extraction
Triplet Extraction from Radiopaedia Articles
Triplet Transformation and Embedding
Report Generation
Constructing Prompts
Generating Reports using GPT-4 or GPT-4O
Evaluation
Evaluation Metrics: BERTScore, TF-IDF Cosine Similarity

Figure 1. Multi-Modal Radiology Report Generation Architecture

A. *Data Preparation*

Data Sources: We utilized the MIMIC-CXR (Medical Information Mart for Intensive Care - Chest X-ray) dataset, which is a large publicly available repository of chest radiographs and corresponding free-text radiology reports. This dataset is widely used for research in medical imaging and natural language processing, providing a robust foundation for developing and evaluating algorithms related to chest radiograph interpretation.

1) Data Cleaning and Preprocessing:

Processing DICOM Files:

To process DICOM files, we used the pydicom library, which is widely used for handling DICOM files in medical imaging [1]. The pixel values in the DICOM files were normalized to a range of [0, 1] to ensure consistent intensity values for subsequent processing. After normalization, the pixel values were scaled to [0, 255] to match the input requirements of the ResNet-50 model. [3] This step involved converting the normalized pixel values into an 8-bit image format. The images were then resized to 224x224 pixels using the PIL library, ensuring all images were of a consistent size, which is necessary for batch processing in the neural network.

2) Cleaning Text Reports:

To clean the text reports, we implemented a process to remove unnecessary whitespace and newline characters, ensuring that the text data was uniform and readable. Special characters and irrelevant information were filtered out to focus on the medical content crucial for accurate report generation. This cleaning process was implemented using regular expressions with the re library.

3) Dataset Creation and Splitting:

A custom dataset class was created to handle the DICOM files and text reports, ensuring proper matching of each file type. This class facilitated the loading and preprocessing of the data for training and evaluation. The dataset was then split into training (80%), validation (12%), and test sets (8%) to facilitate robust model training and evaluation. This split was chosen to ensure that the model had sufficient data for learning and could be evaluated on unseen data to assess its generalization capability.

B. *Feature Extraction*

Image Features: To extract features from chest X-ray images, we utilized a pre-trained ResNet-50 model. This model is well-suited for extracting rich visual features from medical images due to its deep architecture and extensive training on the ImageNet dataset.

1) Preprocessing DICOM Files:

As previously described, DICOM files were read using the pydicom library, normalized, and scaled. The images were then resized to 224x224 pixels using the PIL library to match the input size required by the ResNet-50 model.

2) Generating Image Embeddings:

The preprocessed images were passed through the ResNet-50 model implemented in PyTorch. The model processes the images and generates feature vectors that capture critical visual information. The output feature vectors from ResNet-50 were then passed through a linear layer, which projects the high-dimensional image features into a lower-dimensional common embedding space. This projection aligns the dimensions of image embeddings with text embeddings, facilitating their integration in the ALBEF model.

3) Text Features:

For extracting features from the text reports, we employed a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model. BERT is renowned for its capability to capture contextual relationships between words, making it ideal for processing medical text.

4) Tokenization and Preprocessing:

Radiology reports were tokenized using the transformers library, which provides a tokenizer compatible with the BERT model. This step converts the text into input IDs and attention masks required by the model. Additionally, as described earlier, text reports were cleaned to remove unnecessary characters and whitespace.

5) Generating Text Embeddings:

The tokenized text was processed by the BERT model, implemented in PyTorch, to generate feature vectors that encapsulate the semantic content of the radiology reports. Similar to the image features, the text embeddings were then projected into the same common embedding space using a linear layer. This step ensures that both image and text features are aligned in a shared space, which is crucial for the ALBEF model's performance.

6) Align Before Fuse (ALBEF) Model:

The core of our approach is the ALBEF (Align Before Fuse) model, designed to effectively align image and text features before fusing them. This model enhances the performance of multi-modal tasks by leveraging a contrastive learning framework to align the embeddings from both modalities in a common feature space.

(i) Model Architecture:

The ALBEF model's architecture includes an image encoder, a text encoder, and projection layers. We utilized a pre-trained ResNet-50 model to extract feature vectors from chest X-ray images. These feature vectors were then projected into a common embedding space using a linear layer. Similarly, a pre-trained BERT model was employed to generate contextual embeddings from radiology reports, which were also projected into the common embedding space using a linear layer. These projection layers ensured that the dimensions of image and text embeddings matched, facilitating their integration. The projection layers were implemented in PyTorch.

(ii) Training the ALBEF Model:

The ALBEF model used a contrastive learning framework to align the image and text embeddings. The contrastive loss function encouraged the model to bring similar embeddings (from related image-text pairs) closer while pushing dissimilar embeddings apart. The contrastive loss was calculated as the average of the image-to-text and text-to-image retrieval losses. The model was trained to minimize the contrastive loss, thereby learning to effectively align the multi-modal embeddings in a shared space, which is crucial for generating accurate and contextually relevant radiology reports. During the training process, the learning rate was set to 1e-4, based on empirical experiments to balance convergence speed and stability. A batch size of 16 was used, considering memory constraints and the need for effective batch processing. The model was trained for 10 epochs, with early stopping implemented after 6 epochs to prevent overfitting. The Adam optimizer was used for its efficiency in handling sparse gradients and large parameter spaces.

(iii) Generating Embeddings:

Following training, the model was employed to generate embeddings for the chest X-ray images. Preprocessed DICOM files were passed through the trained ResNet-50 model, and the resulting feature vectors were projected into the common embedding space. Similarly, the trained model was used to generate embeddings for the text reports. The reports were tokenized, processed by the BERT model, and the resulting feature vectors were projected into the common embedding space. Additionally, structured knowledge from Radiopaedia articles was converted into triplets, transformed into text strings, and processed to generate embeddings using the BERT model. These triplet embeddings provided additional contextual knowledge.

*C. **Triplet Extraction***

Knowledge Extraction: To enhance contextual understanding, we incorporated structured knowledge from Radiopaedia articles. This involved web scraping to extract relevant information from Radiopaedia and loading the data into a pandas DataFrame. Each article was analyzed to extract information about diseases, presentations, descriptions, and conclusions. These elements were converted into structured triplets (e.g., ”Pneumonia” - ”has symptom” - ”Cough”) based on predefined relationships. The extraction process focused on identifying key entities and their relationships within the text [7].

Each triplet was transformed into a text string to be processed by the BERT model. For instance, the triplet (”Pneumonia”, ”has symptom”, ”Cough”) was converted to the text string ”Pneumonia has symptom Cough”. These text strings were then processed using the BERT model to generate embeddings. These embeddings provided additional contextual knowledge, which improved the model’s ability to generate accurate and contextually rich radiology reports [12].

D. Integration of Triplet Knowledge

To ensure the model utilized the most relevant information, we implemented a retrieval mechanism to identify the top-K relevant text embeddings and triplets for each image embedding. This process was crucial for providing comprehensive context and improving the accuracy of generated reports.

To accomplish this, we computed the cosine similarity between the image embeddings and both text and triplet embeddings. Cosine similarity measures the cosine of the angle between two vectors, indicating their similarity. A higher cosine similarity value indicates that the embeddings are more similar. The cosine similarity was calculated using the scikit-learn library’s cosine_similarity function, which efficiently computes the cosine similarity between pairs of vectors. Through experimentation, we determined that selecting the top 5 embeddings (K=5) provided a good balance between relevance and diversity of the retrieved information. For each image embedding, we retrieved the top-K text embeddings and triplets based on the highest cosine similarity scores, ensuring the model focused on the most relevant information when generating reports. In cases where multiple embeddings had identical cosine similarity scores, additional criteria such as average semantic similarity were used to break ties, ensuring the most contextually relevant embeddings were selected.

E. Integration of Embeddings:

The top-K relevant text embeddings and triplets were then integrated with the image embeddings to provide a comprehensive context for report generation. This integration was achieved by concatenating the selected embeddings with the image embeddings, creating a rich multi-modal representation. The integrated embeddings were evaluated to ensure they provided coherent and contextually accurate information, assessing the relevance and quality of the combined embeddings and their impact on the generated reports.

F. Report Generation

Construction of Detailed Prompts: To guide the report generation process using GPT-4 and GPT-4O, we constructed detailed prompts. These prompts included specific instructions and context to ensure that the generated reports were precise and relevant. Attributes such as pathology, positional, severity, and size words were specified to maintain consistent terminology. Additionally, we explored the use of few-shot prompting to improve the quality and contextual accuracy of the generated reports.

Generating Reports with GPT-4: Using the constructed prompts, GPT-4 was configured with the role of a medical AI specialized in generating radiology reports. The process involved setting up the model with the appropriate context and instructions, and then generating impression summaries based on the provided input. This setup ensured that the model produced accurate and detailed reports tailored to the medical context.

Table I
TABLE OF ATTRIBUTES AND WORDS

Attributes	Words
Pathology	atelectasis, opacities, consolidation, effusion, pneumothorax
Positional	right, left, bilateral, upper, lower, middle
Severity	mild, moderate, severe
Size	small, large

1) Direct Prompt Example [9]

pathology_words = [”atelectasis”, ”opacities”, ”consolidation”, ”effusion”, ”pneumothorax”]
positional_words = [”right”, ”left”, ”bilateral”, ”upper”, ”lower”, ”middle”]
severity_words = [”mild”, ”moderate”, ”severe”]
size_words = [”small”, ”large”]

You are an assistant designed to write impression summaries for the radiology report. Users will send a context text and you will respond with an impression summary using that context. Instructions:

- Impression should be based on the information that the user will send in the context.
- The impression should not mention anything about follow-up actions.
- Impression should not contain any mentions of prior or previous studies.
- Use the following words for attributes where applicable:
  - Pathology: atelectasis, opacities, consolidation, effusion, pneumothorax
  - Positional: right, left, bilateral, upper, lower, middle
  - Severity: mild, moderate, severe
  - Size: small, large

CONTEXT: {Top 5 radiology reports} {Top 5 relevant Radiopaedia triplets_text}

IMPRESSION:

2) Few-Shot Prompt Example

To further enhance the accuracy and relevance of the generated reports, we implemented a few-shot prompting technique.

This method provides the model with several example inputs and outputs to guide its response generation.

As an assistant designed to write impression summaries for radiology reports, you will receive a context text from users and respond with an impression summary based on that context.

Instructions:

- Impression should be based on the information that the user will send in the context.
- The impression should not mention anything about follow-up actions.
- Impression should not contain any mentions of prior or previous studies.
- Use the following words for attributes where applicable:
  - Pathology: atelectasis, opacities, consolidation, effusion, pneumothorax
  - Positional: right, left, bilateral, upper, lower, middle
  - Severity: mild, moderate, severe
  - Size: small, large
  - Example 1: CONTEXT: Frontal chest radiographs were obtained with the patient in the upright position. COMPARISON: Radiographs from ___, ___ and ___. FINDINGS: The lungs are clear of focal consolidation, pleural effusion or pneumothorax. The heart size is normal. The mediastinal contours are normal. Multiple surgical clips project over the left breast, and old left rib fractures are noted. IMPRESSION: No acute cardiopulmonary process.
  - Example 2: CONTEXT: Portable chest radiograph demonstrates unchanged mediastinal, hilar, and cardiac contours. There has been interval development of bibasilar opacities likely reflecting atelectasis, though cannot exclude developing infectious process. Additionally, there has been interval increase in small right-sided pleural effusion. IMPRESSION: Bibasilar opacities, likely atelectasis. Increased small right pleural effusion.
  - Example 3: CONTEXT: A tracheostomy is in place. Bullet fracture fragments are again noted bilaterally. Bilateral chest tubes are unchanged in positions. A right-sided pneumothorax has increased and is now small to moderate in size. There is no definite pleural effusion on the left. Vague retrocardiac opacity is similar and suggests atelectasis, but improved substantially. IMPRESSION: Increase in right-sided pneumothorax, now small to moderate, with mild recurrent leftward shift.

CONTEXT: {Top 5 radiology reports} {Top 5 relevant Radiopaedia triplets_text}

IMPRESSION:

G. *Report Evaluation*

To assess the similarity between the generated reports and actual reports, we used several metrics, including BERTScore and TF-IDF cosine similarity. BERTScore evaluated precision, recall, and F1 scores, measuring semantic similarity, while TF-IDF cosine similarity assessed the similarity based on term frequency-inverse document frequency vectors.

We used several metrics to assess the similarity between the generated reports and the actual reports. These metrics included BERTScore and TF-IDF cosine similarity, each providing unique insights into the quality of the generated text.

(i) BERTScore:

BERTScore measures the similarity between two texts based on the embeddings generated by a pre-trained BERT model. It evaluates precision, recall, and F1 scores, focusing on the semantic similarity and contextual accuracy of the text. High BERTScore values indicate that the generated reports closely match the actual reports in terms of semantic content and contextual relevance. This metric is particularly useful for assessing the accuracy of medical terminology and the coherence of the generated text.

(ii) TF-IDF Cosine Similarity:

TF-IDF cosine similarity measures the similarity between two texts based on term frequency-inverse document frequency vectors. It evaluates how well the terms in the generated reports match those in the actual reports. TF-IDF scores provide insights into the lexical similarity between the texts. Lower TF-IDF scores compared to BERTScore values suggest that while the generated reports may use different terms or phrases, they still convey the same semantic meaning. This discrepancy highlights the importance of using multiple metrics to comprehensively evaluate the quality of the generated reports.

The scores for different evaluation metrics (BERTScore Precision, Recall, F1, TF-IDF Score) were aggregated and averaged to provide an overall assessment of the model's performance. These scores indicated how well the generated reports matched the actual reports in terms of precision, recall, semantic similarity, and other relevant measures.

H. *Comparative Analysis with GPT-4 and GPT-4O*

To further evaluate our approach, we compared the performance of GPT-4 and GPT-4O in generating radiology reports. Both models were provided with the same prompts and context to ensure a fair comparison. We analyzed the generated reports using the same set of metrics and discussed the observed differences and their implications.

Results and Analysis: The comparative analysis highlighted the strengths and weaknesses of each model.

Direct Prompt Result:

Table II
AVERAGE SCORES FOR GPT-4 AND GPT-4O WITH DIRECT PROMPT

Metrics	GPT-4 (Direct Prompt)	GPT-4O (Direct Prompt)
BERTScore Precision	0.831345	0.841910
BERTScore Recall	0.827029	0.821181
BERTScore F1	0.829072	0.831197
TF-IDF Score	0.276860	0.218646

Few shot Prompt Result:

Table III
AVERAGE SCORES FOR GPT-4 AND GPT-4O WITH FEW-SHOT PROMPT

Metrics	GPT-4 (Few-Shot Prompt)	GPT-4O (Few-Shot Prompt)
BERTScore Precision	0.836496	0.857200
BERTScore Recall	0.818821	0.818081
BERTScore F1	0.827462	0.837124
TF-IDF Score	0.190207	0.183484

BERTScore Precision, Recall, and F1:

High Scores: The high BERTScore values for both GPT-4 and GPT-4O indicate that the generated reports maintain a high degree of similarity to the actual reports in terms of word choice and sentence structure [2].

Precision vs. Recall: GPT-4O has slightly higher precision, suggesting it produces text that closely matches the reference reports in word choice. On the other hand, GPT-4 has higher recall, indicating it captures a broader range of relevant information.

TF-IDF Scores:

Lower Scores: The relatively lower TF-IDF scores compared to BERTScores can be attributed to the emphasis on term frequency and the uniqueness of terms in the document. Since the generated reports may use different terms or phrases to describe the same findings, the TF-IDF score is lower.

GPT-4O Conciseness: GPT-4O, being more concise, may omit less frequent terms found in the original reports, resulting in even lower TF-IDF scores.

IV. RESULTS

The results section presents the outcomes of our automated radiology report generation system, comparing the generated reports from GPT-4 and GPT-4O models to the original cleaned reports from the dataset. This comparative analysis provides insights into the quality, coherence, and relevance of the generated reports.

A. *Comparison of Generated Reports*

GPT-4 Generated Reports: The reports generated by GPT-4 are detailed and often provide comprehensive descriptions of the radiographic findings. GPT-4 tends to include multiple observations, describe the position and severity of findings, and integrate structured medical knowledge effectively. The generated content is generally well-structured and closely aligns with the detailed nature of the original cleaned reports.

> ”No evidence of pulmonary pathology such as atelectasis, opacities, consolidation, effusion, or pneumothorax. The heart and mediastinum appear normal. No significant abnormalities detected in the bones.”

GPT-4O Generated Reports: The reports generated by GPT-4O are more concise and focused. They tend to highlight the key findings succinctly, often using fewer sentences to convey the same information. While GPT-4O's reports are precise and to the point, they may lack some of the detailed descriptions and contextual information found in GPT-4's reports.

> ”Normal chest radiograph with no evidence of pulmonary pathology. Heart and mediastinum are normal. No significant bone abnormalities.”

B. *Comparison to Original Cleaned Reports*

When comparing both GPT-4 and GPT-4O generated reports to the original cleaned reports, we observe that both models accurately capture the primary findings and important details present in the original reports. They use relevant medical terminology and provide coherent summaries of the radiographic images. However, GPT-4's reports are more verbose, offering a richer description and multiple observations that closely mimic the style and content of the original reports. In contrast, GPT-4O's reports are more streamlined, providing concise and focused impressions that may omit some of the detailed descriptions but still convey the essential findings. These differences in report length and detail level reflect the inherent design choices of the two models, with GPT-4 aiming for comprehensiveness and GPT-4O prioritizing brevity and precision.

C. *Interpretation of Results*

BERTScore Precision, Recall, and F1: These scores are high for both GPT-4 and GPT-4O, indicating that the generated reports maintain a high degree of similarity to the actual reports in terms of word choice and sentence structure. GPT-4O has slightly higher precision, indicating it produces text that closely matches the reference reports in word choice, whereas GPT-4 has higher recall, indicating it captures a broader range of relevant information.

TF-IDF Score: The TF-IDF scores are relatively low compared to the BERTScores. This can be attributed to the fact that TF-IDF emphasizes term frequency and the uniqueness of terms in the document, while BERTScore focuses on the semantic similarity and contextual accuracy. Since the generated reports may use different terms or phrases to describe the same findings, the TF-IDF score is lower. GPT-4O, being more concise, may omit less frequent terms found in the original reports, resulting in even lower TF-IDF scores.

V. Conclusions

In this study, we developed an automated radiology report generation system that integrates multi-modal data from chest X-ray images and radiology reports, incorporating structured knowledge from Radiopaedia articles. Our methodology encompassed data preparation, feature extraction, triplet extraction, embedding generation, and report generation using GPT-4 and GPT-4O models.

The system's performance was evaluated using various metrics, such as BERTScore and TF-IDF cosine similarity, which demonstrated high similarity between the generated and actual reports. Comparisons between the generated and original cleaned reports revealed that GPT-4 produced more comprehensive reports, while GPT-4O provided concise and focused summaries.

This study highlights the potential of AI-driven methods to enhance clinical workflows and improve diagnostic accuracy. By generating high-quality, contextually relevant radiology reports, our approach can reduce the workload on radiologists and support more accurate and timely diagnoses.

VI. Key Findings

GPT-4: Produces detailed and comprehensive reports that closely align with the original cleaned reports in terms of structure and content. It is well-suited for scenarios where a thorough and descriptive summary is required.

GPT-4O: Generates concise and focused reports that capture the essential findings succinctly. It is ideal for situations where brevity and precision are prioritized over detailed descriptions.

VII. Future Work

Future work could focus on several key areas to enhance the effectiveness and applicability of the automated radiology report generation system. First, consulting AI-generated reports with radiologists will be crucial. This collaboration will ensure the clinical relevance and accuracy of the reports, providing valuable feedback for refining the models and improving their practical utility in a clinical setting. Secondly, extending the current model to different parts of the body, such as CT scans, MRI, and ultrasound images, will demonstrate the model's versatility and applicability across various medical imaging modalities. This expansion can help in building a more comprehensive tool that radiologists can rely on for multiple diagnostic tasks. Improving model interpretability is another essential area of future work. Developing methods to explain the model's decisions will provide radiologists with insights into how the AI arrived at specific conclusions, thereby increasing trust and facilitating the integration of AI into clinical workflows.

If approved by radiologists, implementing the model in a clinical setting for real-time report generation could significantly enhance workflow efficiency. This real-time application will allow for the assessment of the model's performance in actual clinical environments, providing opportunities to fine-tune the system for better accuracy and speed. Additionally, incorporating personalized patient reports by leveraging external knowledge, such as the patient's medical history and previous imaging results, can tailor the AI-generated reports to individual patients. This personalization will ensure that the reports are contextually relevant and more useful for patient-specific diagnoses and treatment planning. By addressing these areas, future work can significantly enhance the impact of AI-driven approaches on clinical workflows and diagnostic accuracy, making them robust tools for radiology departments.

References

[1] Nathaniel R. Greenbaum Matthew P. Lungren Chih-ying Deng Yifan Peng Zhiyong Lu Roger G. Mark Seth J. Berkowitz Steven Horng Alistair E. W. Johnson, Tom J. Pollard. Improvingmimic-cxr-jpg, a large publicly available database of labeled chest radiographs. 2029.

[2] Eric Xing Baoyu Jing, Pengtao Xie. On the automatic generation of medical imaging reports. 2018.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, Dec 2015.

[4] Andrew Li Sina Hartung Fardad Behzadi Juan Calle David Osayande-Michael Pohlen Subathra Adithan Pranav Rajpurkar Jaehwan Jeong, Katherine Tian. Multimodal image-text matching improves retrieval-based chest x-ray report generation. Cornell University, 2023.

[5] Suhyeon Lee, Won Jun Kim, and Jong Chul Ye. Llm itself can read and generate cxr images. 2023.

[6] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation, Oct 2021.

[7] Tapas Nayak, Navonil Majumder, Pawan Goyal, and Soujanya Poria. Deep neural approaches to relation triplets extraction: a comprehensive survey. *Cognitive Computation*, 13(5):1215–1232, September 2021.

[8] Vignav Ramesh, Nathan Chi, and Pranav Rajpurkar. Improving radiology report generation systems by removing hallucinated references to non-existent priors. Sep 2022.

[9] Mercy Ranjit, Gopinath Ganapathy, Ranjit Manuel, and Tanuja Ganu. Retrieval augmented chest x-ray report generation using openai gpt models. In *Machine Learning for Healthcare Conference*, pages 650–666. PMLR, 2023.

[10] Felix Wu Kilian Q. Weinberger Yoav Artzi Tianyi Zhang, Varsha Kishore. Bertscore: Evaluating text generation with bert. 2020.

[11] Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical image analysis*, 80:102510, 2022.

[12] Emily Bao Tsai Curtis P. Langlotz Dan Jurafsky Yasuhide Miura, Yuhao Zhang. Improving factual completeness and consistency of image-to-text radiology report generation. 2021.

[13] Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y Ng, et al. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9), 2023.