Analyzing Multimodal Machine Learning Model Performance and Evaluation Metrics for Medical Report Generation

Ankit Gupta

CMU-CS-24-155

December 2024

Computer Science Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Min Xu, Chair
Martin Zhang
Bryan Wilder

Submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer Science.

Keywords: Multimodal Learning, Vision-Language Models, Medical Report Generation

Abstract

As a result of recent advancements in foundation models, including large vision-language models, several researchers have explored methods of combining multiple modalities of data as inputs for visual question answering. One key application of visual question answering in the context of the healthcare domain is automated medical report generation, where chest X-ray images and text-based symptom data for a patient might be provided as inputs, with the intention of generating a relevant medical report as an output. However, very few studies analyze the performance of these models alongside unimodal fine-tuned LLMs, and even fewer compare the performance of these multimodal models depending on whether they are provided symptom information as an input. Furthermore, past studies often use simple evaluation metrics that look at n-gram overlaps, such as BLEU and ROUGE scores, which are not effective for generative foundation models that can generate different sentences with the same semantic meaning.

In this paper, we present two main contributions. First, we compare the performance of a variety of approaches for generating medical reports on a dataset of chest X-Ray medical reports, including a unimodal fine-tuned medical LLM, a multimodal model without symptom data, and a multimodal model with symptom data. Second, we introduce four new metrics for evaluating the similarity between generated and reference medical reports, which we term Word Pairs, Sentence Average, Sentence Pairs, and Sentence Pairs (Bio). Our results show that multimodal approaches to medical report generation far outperform unimodal approaches, and providing symptom data slightly improves accuracy for generated medical reports. We also find that our newly introduced Sentence Pairs evaluation metric more closely measures similarity between generated and reference medical reports than all prior metrics, as evidenced by thorough quantitative and qualitative case study comparisons.

This research fundamentally pushes the frontier of medical report generation by further reinforcing the accuracy benefits of using multimodal models with symptom inputs and introducing several more comprehensive, customized scoring metrics for evaluating generated medical reports.

Acknowledgments

First, I'd like to thank Prof. Ruben Martins, Prof. Dave Eckhardt, and Angy Malloy for their support throughout my time in the MSCS program within Carnegie Mellon's Computer Science Department. Prof. Martins gave me extremely valuable advice throughout the semester as I worked on my thesis, Prof. Eckhardt supported me throughout my time in the MSCS program as my academic advisor, and Angy helped me with setting up my thesis defense and many other things throughout the program. I'd also like to thank Catherine Copetas for her support throughout my time here and for helping with publishing my thesis as a technical report.

Next, I'm incredibly grateful to Prof. Min Xu for serving as the chair of my thesis committee, and Prof. Martin Zhang and Prof. Bryan Wilder for serving on my thesis committee during the Fall 2024 semester. All of their advice has been incredibly useful throughout writing and revising my thesis, so I'm incredibly grateful for their support. I'd also like to thank Prof. Russ Salakhutdinov and Martin Ma for advising me during Summer 2024 with my independent study focused on multimodal machine learning. I'm very grateful to Prof. Salakhutdinov for serving as my independent study advisor, and to Martin for guiding me throughout the summer.

In terms of my friends, first and foremost, I am beyond grateful to two of my best friends, Arnav Bansal and Kunal Sharma, who have both pushed me to grow and become the best version of myself since the end of high school. They've supported me since far before I started my research thesis, and even before I got to undergrad. They've seen me throughout all of my ups and downs, and have constantly supported me. I couldn't be more grateful to have them by my side. It's safe to say that I'd be nowhere close to where I am today without their ardent support.

Next, I'd like to express my highest gratitude to Daniel Li, Philip Pan, and Nitin Maddi. Daniel taught me the importance of being well-rounded and more lighthearted in the way I approach life, as one of the most relaxed, yet hard working and likable people I know. Philip served as a mentor to me for years by giving me helpful advice throughout middle school and high school, all which helped serve as a catalyst for my personal and professional growth. Nitin pushed me to think critically about mistakes I was making and greatly supported me by consistently giving me valuable advice. I'm grateful that I had the opportunity to visit Philip at UPenn in Fall 2023, Daniel at UChicago in Fall 2024, and Nitin in Chicago in Fall 2024, all during my time in graduate school.

From undergrad, I'd like to thank my close friends within the Rodman Scholars program, including Shiva Manandhar, Alexi Gladstone, Abhir Karande, and Daniel Xue. All four of these friends were incredibly supportive during my time at UVA, and I couldn't be more appreciative for the time that I spent with them. I'd also like to give my greatest thanks to two of my best friends and Rodman alums, Vincent Lin and Paul Lee, who've been great mentors over the past 3 years. Lastly, I'd like to thank Prof. Rich Nguyen, Prof. Nada Basit, Prof. Jim Cheng, and Prof. Nathan Brunelle for supporting me during my time in undergrad.

From my internship at IMC Trading, I'd like to give my highest gratitude to Iris Xia, James Guo, Catherine Huang, Kathryn Wang, and Jimmy Zhang. Iris remains one of my closest and most supportive friends from my time at IMC, and is extremely kind, understanding, and hard-working. James continues to be one of the most well-rounded, creative, and intelligent people I know. Catherine is one of my most bubbly and compassionate friends, and it's been great getting to see her both in Boston and at ICML 2024. Kathryn is one of the most social and well-balanced people I know, but is also extremely hard-working, driven, and ambitious. Jimmy is by far one of the most loyal and supportive people I've ever met in my entire life, and he's gone above and beyond to support me in my personal growth, both at IMC and at Carnegie Mellon. I couldn't be more grateful to know all of these wonderful people.

I'd also like to thank some other people at IMC that have been supportive during my time in grad school, including Janet Qian, Erica Wang, Mike Li, Anni Pan, Ian McKibben, Uche Ochuba, Brian Xiang, Ethan Zhang, Sean Park, Benny Sun, Soham Bose, Yunpeng Liu, Ziyu Zhu, Allen Zheng, Will Fan, Arya Majjiga, Aditya Mehta, Aadeesh Shastry, Jefferson Yu, Nicole Stiles, Janvi Shah, Agnim Agarwal, Andrew Nguyen, Tailai Wang, Forrest Fan, Stephen Buck, Troy Feng, Shiva Ganapathy, Trent Stauffer, and Prayaag Gupta. It's been great keeping in touch with all of these people, and I look forward to continuing to visit them in the future.

In terms of my friends in Chicago, I'd like to give my highest recognition to Allison Zhuang, Ryan Grayson, and Patrick Au. Allison and I met through a mutual friend in Chicago, and she's continued to be one of the kindest, most mature people I've ever met. Ryan and I met towards the end of my time at UVA, and he's constantly impressed me with his resilience, hard work, and drive, all of which have pushed me to improve holistically. I met Patrick at ICML in Austria, and he continues to be one of the most mindful, humble, and sociable people I know. I'm truly grateful to all of these people for their role in my growth during graduate school.

From graduate school at Carnegie Mellon's CS Department, I've met tons of incredible people that I'd like to thank.

First and foremost, I'd like to give my absolute highest thanks to Patrick Wang, Saaketh Medepalli, Lawrence Jang, Vincent Tombari, and Maxwell Jones. Each one of these individuals has gone leagues above and beyond in their day-to-day life, teaching me about things both inside and outside of academics. Patrick was one of my most supportive friends during my time in MSCS, especially as we organized events together for MSCS, and he constantly supported me during my setbacks. Saaketh is one of the most reliable, honest, and dependable people I've ever met, and he was tremendously helpful as we built social events across MSCS and MSML. Larry is one of the hardest working people I know, with an incredible balance between his social life and drive to make a strong research impact. Vinnie is one of the kindest, most down-to-earth people I've ever met in my entire life. Maxwell is one of the most supportive people I've ever met, is incredibly hard working, and extremely well-versed. In many ways, I grew and learned a ton from these 5 friends — both socially and intellectually — and I couldn't be more appreciative for the time I spent with them.

From the SCS Class of 2024, I'd like to give my greatest gratitude to Jimmy Zhang, Deep Patel, and Bharat Narayanasamy. I met Jimmy during Summer 2023, and he supported me fervently, helping me meet undergrads, giving me advice about my time here, and being a kind and loyal friend. I met Deep and Bharat at SCS Day and both of them shocked me with how incredibly open, kind, and supportive they were during my time as a graduate student here — they're easily two of the kindest people I've ever met in my entire life. I'd also like to thank other friends that supported me during my time here, including Sophie Liu, Abby Li, Yoseph Mak, Raehash Shah, David Luo, Shreeya Khurana, Raaid Tanveer, Arya Shah, Alex Xu, Meher Mankikar, and Nikhil Patel. I'm going to miss all of you, and I look forward to keeping in touch.

From the SCS Class of 2025, I'd like to give my greatest apppreciation to Archan Das, Claire Jin, and Hugo Contant. I met Archan and Claire during my 2nd semester here, and I quickly found that they were some of the kindest, most down-to-earth people to hang out with. Both of them are incredibly ambitious, accomplished, and likable people, and I had an amazing time getting to know them outside of class. I met Hugo my 2nd semester through events that I organized, and he was always kind, fun, and supportive. I'd also like to thank Michael Zhou, David Krajewski, Lucy Mo, and Anoushka Shrivastava for being kind and supportive friends during my time here.

From the MSCS Class of 2024, I'd like to give my greatest thanks to Ines Vignal, Alexis Schlomer, Karthik Balaji Ganesh, and Jayesh Singla. Ines helped push me to grow across the board, especially with personal development, and was always compassionate, understanding, and empathetic. Alexis was one of the most open, fun people to talk to, especially because of his humor and incredible drive. Karthik is one of the kindest, funniest, and most intelligent people I've ever met in my entire life. I roomed with Jayesh at ICML 2024 in Vienna, Austria and got to know him well throughout the program — he's easily one of the kindest people I've ever met, driven, ambitious, and fun to spend time with. I'd also like to recognize some other people who've been very supportive during my time here, including Raphael Della Vecchia, Logan Nye, Jerry Ji, Devin Qu, Raghu Radhakrishnan, Bobby Pare, Katrina Van Laan, Aidan Wagner, Charlie Ruan, Mixalis Dontas, Shivang Dalal, and Ashwin Rao. All of these individuals went above and beyond to support me both inside and outside of the classroom, and I greatly appreciate all of them.

From the MSCS Class of 2025, I'd like to give my highest recognition to Max Pandolpho, Kandasamy Chokkalingam, Saransh Malik, and Yassine Kachrad. These four individuals are the new generation of people that will be organizing social events for MSCS, and it has been great showing them the ropes. I'm confident they'll do a great job. I'd also like to thank Param Damle, Akash Nayar, Emily Song, Rosy Chen, Pratyush Gupta, Anton Efremov, and Terrance Chen. It's been great getting to spend time with you all through MSCS events and classes, and I wish you the absolute best with your remaining time at Carnegie Mellon.

From the MSML Class of 2024, I'd like to give my highest gratitude to Edoardo Botta, Utkarsh Priyam, Taeyoun Kim, Aman Gupta, Vineeth Kada, Arun Chintala-

pati, Ritvik Gupta, Shrey Gupta, Sumukh Aithal, Hemit Shah, Bao Nguyen, Richa Gadgil, Kartik Khandelwal, Michael Mu, Animesh Bohara, Kaushal Gumpula, James Zhao, Vishwajeet Agrawal, Shao Xuan Seah, and Arya Shah. I'm incredibly grateful that I've gotten the opportunity to talk to all of them during my time here, especially with our similar interests in ML research, companies, and more.

From the MSML Class of 2025, I'd like to give my highest recognition to Neil Kale, Mark Levin, and Eileen Xiao. Neil is an incredibly well-balanced individual that is hard-working, well-liked, and kind. Mark is one of the most talented people I've ever met, and one of the kindest and most reliable. Eileen was easily one of the most considerate and supportive friends I had during my last semester. I'd also like to thank Bhargav Hadya, Alessandro Marmi, Saket Durbha, Matthew Yang, Tanmay Garg, Katy Chu, Aditya Agarwal, Manan Sharma, Annanya Chauhan, Kartik Srinivas, Snigdha Saha, Shreya Sridhar, Harsh Shah, Duncan Soiffer, Ayudh Saxena, Visisht Rao, Rishi Shah, and Udit Arora.

I'd also like to thank all of the PhD students in the ML Department and the CS Department. I had the pleasure of meeting many of them during my 3 semesters here, and learned a ton about the day-to-day of being a PhD student. I had the opportunity to interact with them in classes, at events, and even internationally at ML conferences. In particular, I'd like to give my highest gratitude to Chris Ki, Gaurav Ghosal, Jennifer Hsia, June Hwang, Kanad Pardeshi, Ellie Haber, Kelly He, Yuxiao Qu, and Stephan Xie for being supportive during my time here.

Lastly, I'd like to give my utmost gratitude to my family for their continued support throughout my life, including graduate school. To my parents, thank you for raising me and supporting me throughout everything I've done. To my brother, thank you for always supporting me and pushing me to improve holistically. I'm incredibly grateful for all of your support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

One important responsibility for doctors today is writing medical reports for patients [14]. Since every patient is different and doctors often see many patients, doctors often spend hours writing medical reports, when this time could be better spent in other ways. This key problem is shown in Figure 1.1. In addition, the content in radiology medical reports is often predictable, especially for clearly diagnosable diseases based on X-Rays [21].

One way to benefit doctors is to automate the process of generating these medical reports. Doing this would give doctors more time to spend on other tasks, like spending more time with patients. In addition, automated methods are less likely to make errors, and can be given more past data to look at, which could potentially make them have more knowledge than any one given doctor.

In order to generate medical reports, one method we can use involves machine learning, which involves giving a model a series of input and output examples of inputs for generating a medical report. In the context of chest X-Rays, inputs could look like frontal/lateral images of a chest X-Ray, symptoms that the patient has, and medical history for the given patient. Out-

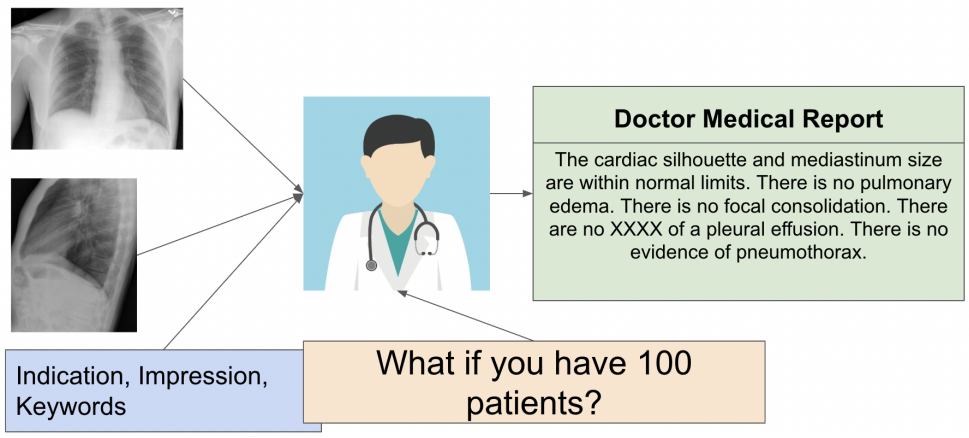

Figure 1.1: Problem: Manually Writing Medical Reports Takes Time

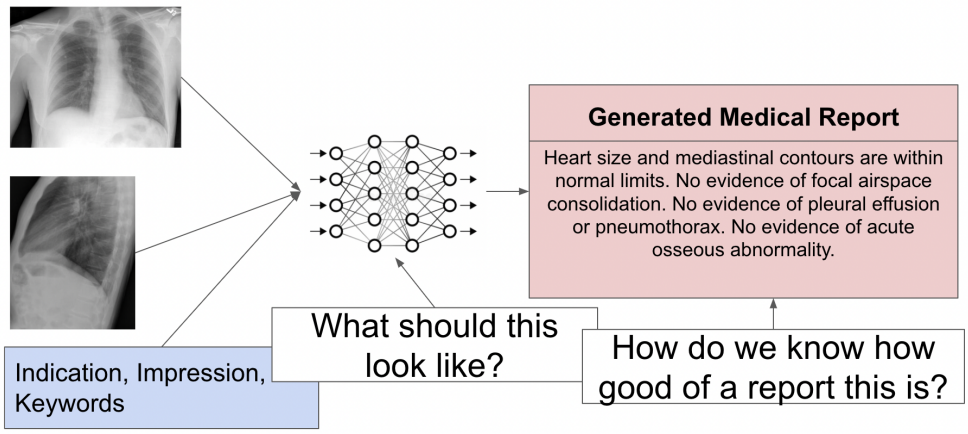

Figure 1.2: Solution: Automating Writing Medical Reports Saves Time

puts could be a text-based representation of a medical report that best describes the given user's disease, if applicable, or whether the user is normal. This is shown in Figure 1.2.

Within the scope of machine learning, there are several approaches that can be used, namely unimodal and multimodal models. As the name implies, unimodal models focus on one input modality of data, like just text-based input or just image-based input. Similarly, multimodal models focus on combining several modalities of data, such as an image along with text. In the context of chest X-Rays, we can see that simply passing in the image of a chest X-ray into a model would be unimodal, while passing in the image of a chest X-ray along with the given patient's symptoms would be multimodal.

1.2 Motivation

One key application of multimodal machine learning is precision health, with applications including neurology and oncology. One such application is medical report generation, which involves taking in some form of an input, such as an image or text, then generating a relevant medical report. For example, in the context of Chest X-rays, one such input could be an image of a Chest X-ray image, while an output could be a report that includes what the X-ray image indicates, any findings based off of the image, and any impressions the image might have. In this example, the inputs are some given image of a chest X-ray, along with some text that asks the user to answer a question about the image, and the output is some text that represents the report for the given input image. This falls within the domain of visual question answering, specifically image question answering.

One major limitation of previous work in the domain of medical report generation is that it focuses only on the image data. However, this isn't representative of the real world, where radiologists have access to multiple modalities of data for a given patient, including clinical notes, symptoms, and a given X-ray. For example, many previous projects use an encoder-decoder based architecture, where the input image gets passed into an encoder, and the decoder uses a transformer. Recent papers have looked into contextual biomedical report imaging, but

they still fundamentally use images as the main modality. For example, the ChestBioX-Gen paper used BioGPT to get the contextual understanding of the task, while also using co-attention to relate certain parts of the image with text-based descriptions.

Another limitation of previous work is that the generated and reference medical reports are not compared in the most effective way. As an example, currently many studies refer to BLEU and ROUGE scores as metrics for comparing generated and reference medical reports. These methods focus entirely on word overlap, which isn't relevant in the case of medical reports, where there are often multiple ways to convey the same diagnostic for a given patient, and where there are also different types of medical terms used. The generated medical report by some model could easily be classified as not being similar to a reference medical report, just because the generated medical report uses synonyms of words in the reference medical report.

1.3 Overview

In this paper, we focus on answering two key questions. First, how effective are multimodal models (with and without symptom data) in comparison to the standard uni-modal models for medical report generation? Second, how can we design a better evaluation technique for medical report generation to best capture the similarities between generated and reference medical reports?

For the first question, we will focus on comparing the accuracy of several approaches on a dataset of Chest X-Ray medical reports, using both old metrics and new metrics. By systematically comparing how similar generated and reference reports are across a series of metrics with 500 reports, we'll be able to tell which models are able to generate reports that are most similar to reference reports.

For the second question, we will try out 4 new techniques beyond BLEU and ROUGE scores, namely word pairs, sentence average, sentence pairs, and sentence pairs (bio). We will look at a subset of 100 generated/reference medical reports, then manually label the comparison of the two reports, and measure how similar both the prior and new metrics are to the manually scored similarity for these 100 medical reports.

Chapter 2

Related Work

In order to better contextualize our contributions, we need to look at the current state of the art. For our first research question, focused on comparing unimodal and multimodal models, we can look at existing types of medical report generation models. For our second question, focused on creating better evaluation metrics for generated medical reports, we can look at current metrics for evaluating the similarity between generated and reference medical reports.

2.1 Medical Report Generation Models

2.1.1 Unimodal Models

2.1.1.1 LLMs

Recently, large language models, also known as LLMs, have emerged as an effective tool for generating chat-like responses [24]. LLMs are often trained on a large amount of text-based data, with the end goal being to generate new text.

One important concept within the LLM space is called fine-tuning, which refers to taking a pre-trained LLM, then passing in a series of inputs/outputs for a given space, such that the LLM can effectively learn the same patterns [23]. For example, if an LLM was fine-tuned with a series of biomedical data question and answer pairs, it would become a fine-tuned medical LLM, with the ability to generate new answers to a given question in a medical context [23].

Fine-tuned medical LLMs have been used for several biomedical visual question answering tasks. For example, Yuan et al. looked into creating a continual pretrained method for automatic medical report generation using an LLM [22]. Similarly, Jung et al. looked into using an LLM for generating medical notes [10]. Specifically, they used a supervised fine-tuning approach to finetune the LLM to be able to generate discharge notes given progress notes, then prompted the finetuned LLM to generate discharge notes.

For our unimodal baseline, we decided to use a fine-tuned medical LLM. There were several reasons why we chose to do this. To begin with, unlike encoder-decoder models, which we describe below, LLMs are much larger, like LLaMA [19]. In addition, encoder-decoder models involve converting an image to text, which has two different modalities of data, even if the input

is just an image. Since we wanted to comparte a completely unimodal baseline and LLMs are purely text-based, we thought that using a fine-tuned medical LLM would be a good choice.

2.1.1.2 Encoder-Decoder Models

Encoder-decoder models are models that involve both an encoder and decoder component for medical report generation [11]. As an example, the encoder might be a CNN to extract features from an image, while a decoder might be an LSTM to create a sequence of words. One key challenge with these models is that generated medical reports are often very similar to each other, when the pictures are relatively similar to each other.

There have been several studies that used encoder-decoder models for medical report generation. For example, Li et al. focused on an auxiliary signal-guided knowledge encoder-decoder [11]. Similarly, Babar et al. looked at an encoder-decoder model that involves using a CNN as an encoder and an LSTM as a decoder [2]. The CNN extracts features from the image, which is then passed to a decoder, which can generate a sequence of words.

As another example, Sirshar et al created an encoder-decoder based framework that also uses attention for medical report generation, where they used a CNN encoder, attention mechanism, and LSTM decoder to generate medical reports [17].

2.1.2 Multimodal Models

Multimodal models involve multiple modalities of data, like images, text, audio, and video. As opposed to encoder-decoder models, which only involve one type of input data, specifically images, multimodal models can take multiple modalities of data as inputs, including any combination of text, images, audio, and videos.

As an example of a multimodal model for the radiology report generation, Thawkar et al. created XRay-GPT. XRay-GPT passes a given chest X-Ray image into a frozen medical vision encoder to get relevant features, then a learnable linear transformation layer, and this output is passed along with a give question to a medical LLM [18]. In this case, XRay-GPT is multimodal, because there are two main inputs involved, namely an input chest X-ray image and an input text prompt.

As another example, Wu et al. created MRCL, which stands for multimodal model with recursive contrastive learning [20]. In this model, contrastive pre-training gets used to generate more expressive text-based and visual-representations. This model involves pre-training an image encoder and sentence encoder, then has two modules, one which generates an impression, and one which generates the findings for a given medical report.

MAIRA-2 is another multimodal model, but specifically for grounded radiology report generation [3]. MAIRA-2 was created by a team of researchers at Microsoft, and takes a series of multimodal inputs, including a frontal image, lateral image, prior frontal image, prior report, task instruction, and indication/technique/comparison. The system message, prior report, task instruction, and indication/technique/comparison all get converted as tokens/embedding and passed into a language model. The frontal image, lateral image, and prior frontal image get passed into a frozen vision encoder, then passed into an adapter to get a representation of visual tokens, which then get passed into the language model.

For this research, we chose to MAIRA-2 as our main multimodal baseline. There were several reasons we chose to do this. To begin with, MAIRA-2 was released in September 2024, which means that it was one of the most recent models in the radiology report generation domain. In addition, MAIRA-2 was publicly available on HuggingFace, which made it easier to evaluate, since it was easier to load. MAIRA-2 is also extremely flexible, since users can choose how many inputs they want to include. For example, users can choose whether to input symptom information, and regardless of whether the user passes in an empty string for the symptom information or lots of symptom information, MAIRA-2 is able to make predictions. This made MAIRA-2 a strong choice for this research specifically, since one of the aspects of our first research question was how the performance of the multimodal model would change when it was given and when it wasn't given symptom information.

2.1.3 Other Models

In addition to the medical report generation models described earlier, there are several other types of models, like retrieval-based models and reinforcement learning-based models.

As an example of a retrieval-based model, Endo et al. created CXR-RePaiR, which generates medical reports with just an image as input [6]. The method involves first storing a large number of reports, then using a pre-trained encoder to encode each of the reports to get a text embedding. Next, every time an image is given as an input to the model, the input image gets passed into a pre-trained image encoder to create an image embedding. The image embedding gets compared to all of the text embeddings to find the report that is most similar to the image embedding, then the corresponding report gets returned as the predicted report.

One interesting approach to medical report generation involves reinforcement learning. As an example, Hou et al. did this with their paper, where they used adversarial reinforcement learning [7]. In this paper, the main architecture that they used involved a CNN encoder for the image and sentence decoder, along with adversarial training between the decoder component and the reward module. In this case, the decoder component creates a report, while the reward module determines how accurate each report is using a diagnostic accuracy measurement component. Thus, this method depends on having an accurate method for measuring how accurate a generated medical report is. We discuss these metrics in the next section.

2.2 Evaluation Metrics

2.2.1 Overview

One important factor in determining how accurate generated medical reports are is the evaluation metric [15]. Ouis et al split their analysis of evaluation metrics into two parts, specifically quantitative metrics and qualitative metrics. As examples of quantitative metrics, they mentioned BLEU, ROUGE, CIDEr, and METEOR. As examples of qualitative metrics, they mentioned MeSH, MIRQI, and Keyword Accuracy.

In this research paper, we focus on creating a series of more effective quantitative metrics for evaluating the quality of generated medical reports. Thus, our key research question, as shown

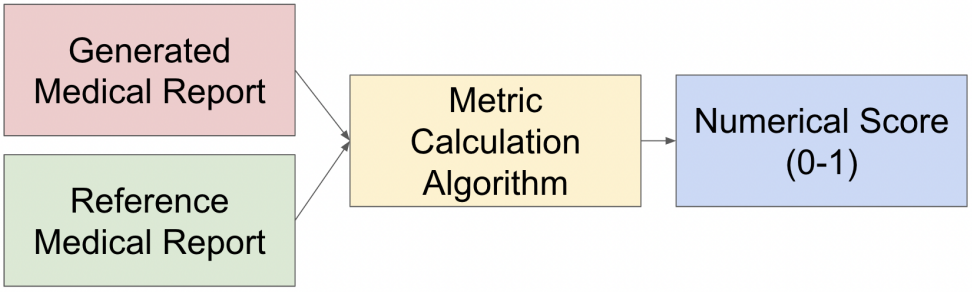

Figure 2.1: Metric Calculation Overview

in Figure 2.1, is how to come up with a metric calculation algorithm that will give us an accurate numerical score.

2.2.2 BLEU Score

BLEU score is a very common evaluation metric for machine translation [16]. The score gets calculated by comparing n-grams of the generated and reference sentences. In this research, we used BLEU-1 and BLEU-2 as 2 of our baseline metrics.

In order to calculate the BLEU score, we first imported nltk, then we split the predicted and reference text into two arrays with their sentences. Next, we used the nltk.translate.bleu_score.sentence_bleu function and passed in the reference text along with the predicted text and the weights. For BLEU-1, we set the weights to be [1,0,0,0], and for BLEU-2, we set the weights to be [0,1,0,0]. Lastly, we returned the output from the sentence_bleu function with the [1,0,0,0] weights input as the BLEU-1 score and the output from the sentence_bleu function with the [0,1,0,0] weights input as the BLEU-2 score.

2.2.3 ROUGE Score

ROUGE score is another common evaluation metric [12]. In this research, we used ROUGE-1, ROUGE-2, and ROUGE-L as 3 of our baseline metrics.

Rouge-N is a metric that looks at the overlap of n-grams between two pieces of text. Rouge-1 focuses on the overlap of unigrams, meaning each word. Rouge-2 looks at the overlap of bigrams. Rouge-L looks at the longest common subsequence.

In order to load the ROUGE score, we used the rouge_score library from Python. Specifically, we used pip3 install rouge_score, then imported rouge_scorer from rouge_score, then created a RougeScorer object with ”rouge1”, ”rouge2”, and ”rougeL” passed in as input metrics. Lastly, we used the created object’s score function and passed in the target and predicted report as inputs. This returned a dictionary of ROUGE values for ROUGE-1, ROUGE-2, and ROUGE-L. For each key in the dictionary, there was a Score object with a precision, recall, and f-measure value. We chose to use the f-measure value as our ROUGE score metric value, since the f-measure value uses both precision and recall.

One key reason why we chose to introduce new metrics for evaluating generated medical reports is because BLEU and ROUGE scores have several problems, especially in the context of

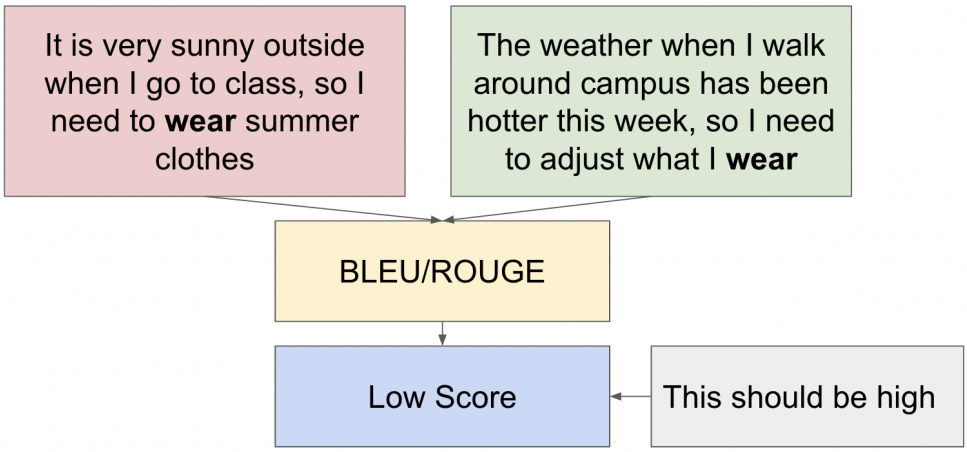

Figure 2.2: Problem with BLEU Score and ROUGE Score

generative models. As shown in Figure 2.2, there are cases where two pieces of text have the same meaning, but use different phrases or two words that are synonyms to express the same idea. In these cases, the BLEU/ROUGE scores are low, because these scores are looking at exact word overlap.

2.2.4 RaTE Score

As an example of a more recent approach, Zhao et al. created the RaTE Score, which is a metric for radiology report generation [25]. The RaTE score paper mentions that they handle cases with medical synonyms and cases with negation values.

The RaTE score is computed by first getting the medical entity and the corresponding entity type, then computing the entity embedding, and getting the cosine vector similarity with the maximum value. The RaTE score uses medical entity recognition to generate a fine-grained medical entity, which gets passed into a synonym disambiguation function. In addition, the RaTE score uses medical entity recognition to generate a contextual entity type, which gets passed into type-aware parameters. The synonym disambiguation component finds the maximum cosine similarity, which gets passed in as an input to calculate the final RaTE score. The contextual entity type gets combined with the type set, affinity matrix, and the negative penalty factor to create a weighted score, which is also passed in as an input to calculate the final RaTE score.

In order to load the RaTE score, we used the RaTEScore library from pip, then created a RaTEScore object. Each time we compared the predicted and reference text, we split the predicted text into an array of sentences, and we split the reference text into an array of sentences. One limitation of the RaTEScore method is that the number of sentences in the reference array and the predicted array need to be the same, which means that the arrays needed to have the same length. However, since the lengths of these two arrays was different in several cases, we chose to find the array with the lower length, then take the same number of sentences from each array. For example, if the reference text has 4 sentences and the predicted text has 5 sentences, we found that the minimum number of sentences across the two was 4 sentences, then we took the first

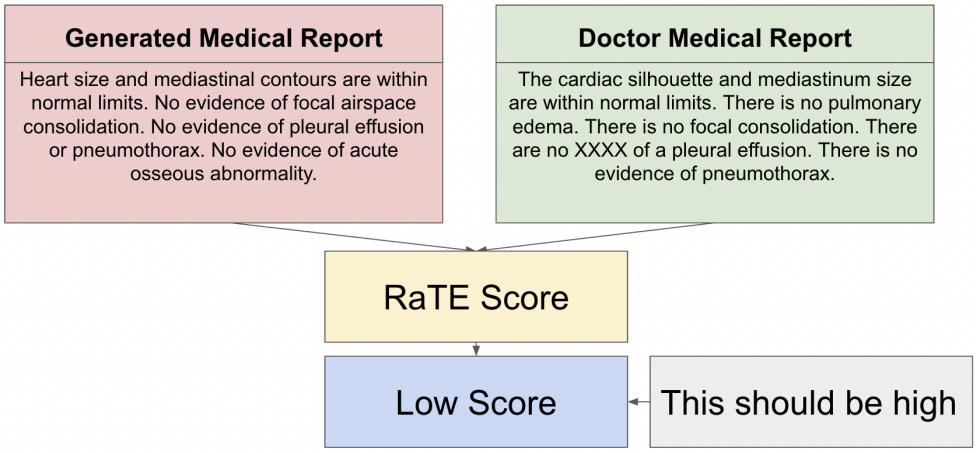

Figure 2.3: Problem with RaTE Score

4 sentences in the predicted text and passed in that array, so that both the predicted and target arrays have 4 elements. Next, we passed these two arrays as inputs into the RaTEScore object that we created earlier, using the RaTEScore.compute_score function. This gave us an array of all of the scores. Lastly, we took the average of all of these scores to get the final average score.

As shown in Figure 2.3, the RaTE score also has a similar problem to the BLEU/ROUGE scores. Specifically, the RaTE score underestimates how similar two pieces of text are. In Figure 2.3, the generated medical report and doctor medical report are extremely similar, but the RaTE score gives a very low score, even though the score should be high. This shows us that the RaTE Score metric isn't the most accurate metric.

Given that the more classical scores, like BLEU and ROUGE have problems with understanding text that is phrased differently, while RaTE Score consistently gives low scores, even when two medical reports are similar to each other, we decided to create our own series of medical report evaluation metrics.

Chapter 3

Methods

There are two main types of methods that we used. The first method is for comparing model performance, which involved loading different types of models and evaluating them on the same dataset to generate a series of medical reports. The second method is for creating and evaluating metrics, where we describe each of the 4 new metrics from this research, along with how we evaluated how effective each of these metrics are, relative to the prior metrics.

3.1 Comparing Model Performance

We chose specifically to compare two techniques to medical report generation, namely uni-modal text-based models and multimodal text and image-based models. Specifically, we picked a fine-tuned medical LLM as our uni-modal model and the MAIRA-2 radiology report generation model as our multimodal model. In order to standardize comparison, we evaluated both of these models on the same 500 samples from the same dataset, which was the Indiana University Chest X-Ray dataset. This process is shown in Figure 3.1.

3.1.1 Dataset

For our dataset, we used the Indiana University Chest X-Ray dataset. This dataset contains 7,470 X-ray images, along with 3,955 corresponding reports [13]. The IU-XRay dataset consists of several columns of data, including the frontal/lateral chest X-ray images, the MeSH value, problems, information about the images provided, indication, comparison, findings, and impression. We chose to set the "findings" column in the IU-XRay dataset to be the generated medical report. We also used a HuggingFace version of this dataset, titled "NLMCXR", to make accessing the Chest X-Ray images easier, since we could load them more easily use HuggingFace. The specific link that we used for our Kaggle Dataset is https://www.kaggle.com/datasets/raddar/chest-xrays-indiana-university, and the specific link that we used for our HuggingFace dataset is here: https://huggingface.co/datasets/Fakhraddin/NLMCXR.

3.1.1.1 Pre-Processing

There was a large amount of pre-processing needed for this dataset. There were 3 main input files that we used. The first was a CSV file consisting of the text-based data, like the MeSH value, problems, information about the images, indication, comparison, findings, and impression. The second was a CSV file with the two columns, where the first column was the path to each Chest X-Ray image and the second column was whether the given image represented a frontal or lateral Chest X-Ray. The third was a HuggingFace dataset based on the IU-XRay dataset, which had a series of 7,400 images in the dataset, split with 5.93k rows of data in the train split of the dataset and 1.51k rows of data in the validation split. This HuggingFace dataset had 3 main categories, including one for the text of the reference report, another with the path to the image, and a third with the corresponding image.

First, we loaded the HuggingFace dataset with images for each of the Chest X-Rays, and used the CSV file that mapped from the path of the image to whether the image was a frontal or lateral chest X-Ray image to identify which images from the HuggingFace dataset were frontal and lateral images. We stored each image in a dictionary, where the key was the ID for the patient along with whether the image was frontal or lateral, and the values were the reference report, the image, and the filename. We repeated this process for both train split and the validation split, so the final dictionary had around 7,400 keys.

Next, we went through the first CSV, which has all of the text-based data, and used Pandas to read the CSV file as a Pandas Dataframe. We iterated through this Pandas Dataframe and created a dictionary for all of the indications, where the key was the User ID, and the value was a string of the format "The indication is <indication>, the problems are <problems>, and the impression is <impression>". Within this format, the <indication> value was the text in the indication field for the row corresponding to the given User ID, the <problems> value was the text in the problems field, and the <impression> value was the text in the impression field. Since some of these values have de-identified characters, we removed these by replacing all occurrences of "XXXX", which represents de-identified information, with the empty string. We also stored the reference report using the "findings" section of the current row for the same User ID. We repeated this process for all rows in the dataset, and the final dictionary had 3,851 keys.

We iterated through the dataset starting with the first User ID, then looked for the frontal and lateral keys in the information dictionary, checked to make sure that the given User ID had a value in the indication dictionary, retrieved the indication string and the reference report string from the indication dictionary, then checked to make sure that frontal and lateral keys were the information dictionary, and the reference report was at least 20 characters long.

For each report that met these conditions, we got the frontal and lateral chest x-ray images, sent it as an input to the MAIRA-2 model without the indication, then got the indication, and sent it as an input, along with the frontal and lateral chest x-ray images to the MAIRA-2 model with the indication. We repeated this process until we generated 500 reports with 500 rows of data that met these conditions. Between each step, we used the torch.cuda.empty_cache method to minimize the amount of GPU RAM that we used.

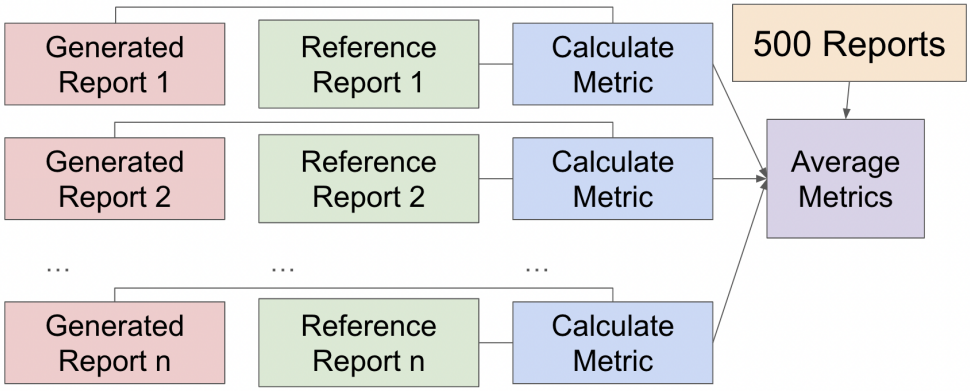

Figure 3.1: Model Evaluation Method Overview

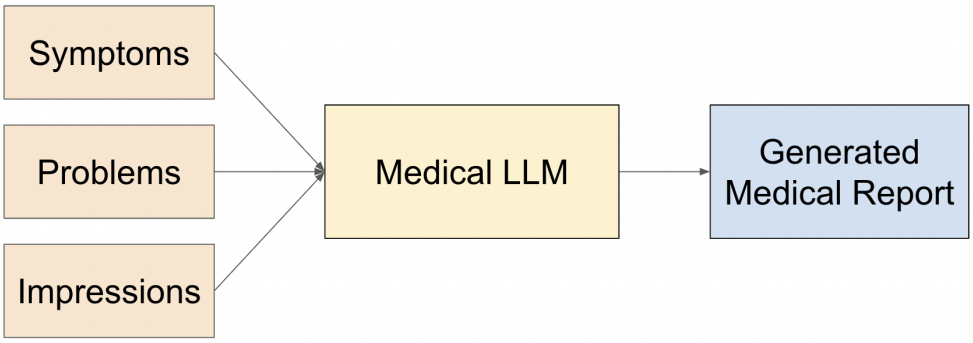

Figure 3.2: Medical LLM Overview

3.1.2 Models

There were two main types of models that we used, specifically the Medical LLM and the MAIRA-2 model. For both models, we ran loading and evaluating the model using Google Colab Notebooks with 1 A100 GPU. The maximum GPU RAM was 40GB for the A100 GPU, which is why we chose to run 500 samples, since we were very close to the maximum GPU RAM limit when we ran the MAIRA-2 model.

3.1.2.1 Medical LLM Model

As shown in Figure 3.2, the purpose of the medical LLM was to convert a series of text-based inputs, like the symptoms, problems, and impressions from a given patient to a generated medical report.

In order to do this, we loaded an already fine-tuned Medical LLM model, called ”Bio-Medical-Llama-3-8B”, from HuggingFace [1]. The model was developed by a company called ”ContactDoctor”, and the model was created by fine-tuning the Llama-3-8B-Instruct base model. As mentioned in the HuggingFace documentation, the model was fine-tuned on over 500,000 entries of biomedical data from a custom dataset that covers several biomedical topics.

The process of fine-tuning this model is shown in Figure 3.3. As shown in the diagram, a

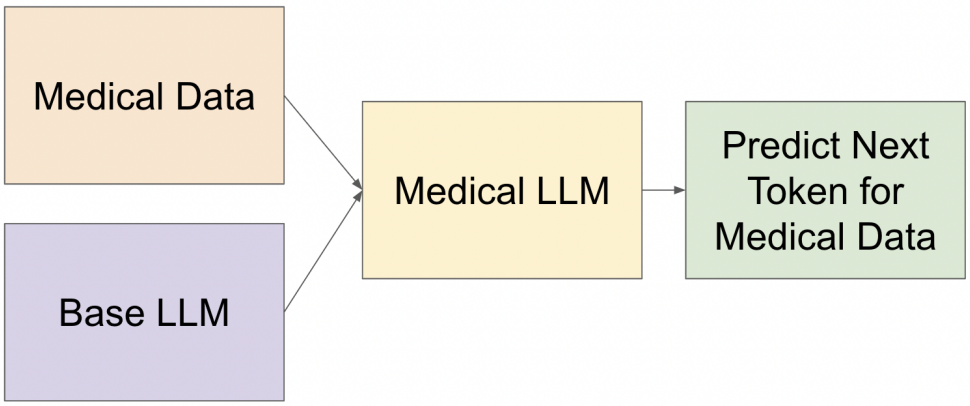

Figure 3.3: Fine-tuning Medical LLM Method

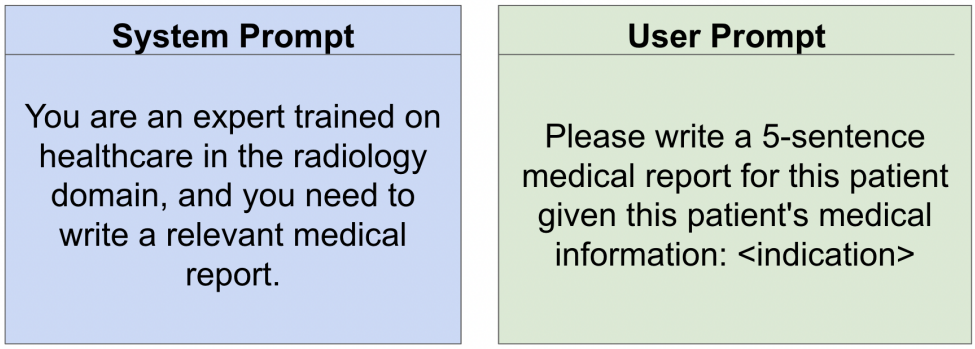

Figure 3.4: Prompt for Medical LLM

series of biomedical questions and answers are provided to the base Llama-3-8B-Instruct model, which produces a fine-tuned medical LLM, like ”Bio-Medical-Llama-3-8B”. This fne-tuned medical LLM has the ability to predict the next token for medical data, which allows it to generate medical reports.

The creators of the ”Bio-Medical-Llama-3-8B” model mention that some of the key applications of this model are helping researchers analyze biomedical articles, helping with making decisions in a clinical setting, and helping as an educational tool for medical students.

We loaded the model from HuggingFace, then used the transformer text-generation pipeline with the Torch float 16 datatype. The model was given the system prompt that ”You are an expert trained on healthcare in the radiology domain, and you need to write a relevant medical report.”. The user was given the prompt ”Please write a 5-sentence medical report for this patient given this patient’s medical information:”, followed by the indication information. The indication information was in the format ”The indication is <indication>, the problems are <problems>, and the impression is <impression>”. These were all retrieved from the IU-XRay dataset, and were formatted as one sentence with all 3 pieces of information. Thus, the final prompt was ”Please write a 5-sentence medical report for this patient given this patient’s medical information: The indication is <indication>, the problems are <problems>, and the impression is <impression>”. These prompts are shown below, in Figure 3.4.

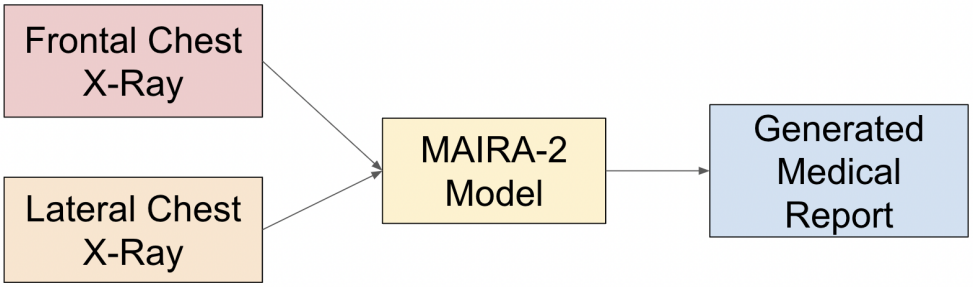

Figure 3.5: Multimodal Model without Indication Flowchart

Once the system prompt and the user prompt were defined, we used the transformer text-generation pipeline to apply the chat template, then created a series of end-of-sentence tokens to add to the end of the prompt. We passed this as an input to the pipeline, with the parameters of 256 max new tokens, do sample set to true, the temperature set to 0.6, and the top p-value set to 0.9. We used these as input parameters because they were the default input parameter values on the HuggingFace page for this model. After we ran the pipeline, we got the final result by accessing the first output's generated text category, then found the remaining words after the input prompt, which became the medical LLM's generated output.

We repeated this process of prompting the LLM to generate responses for all 500 samples. For each sample, we calculated all 10 metrics, then averaged the values across all 500 samples for each of the metrics, and stored these values as results.

3.1.2.2 MAIRA-2 Model

We loaded the MAIRA-2 model from HuggingFace [3]. The MAIRA-2 model allows users to input a few different things, including the frontal X-Ray image, the lateral X-Ray image, the indication, the technique, and prior reports. As shown in Figure 3.5, for the MAIRA-2 model without the indication, we passed in the frontal and lateral X-Ray images as inputs, along with the technique as "PA and lateral views of the chest". As shown in Figure 3.6, for the MAIRA-2 model with the indication, we passed in the frontal and lateral X-Ray images as inputs, along with the indication and the same technique. Since the IU-XRay dataset has a different patient for each row, we didn't input anything in the prior reports category.

In order to load the MAIRA-2 model, we loaded the model using the AutoModelForCausalLM library and we loaded the processor using the AutoProcessor lobrary. We converted the model to be run in eval mode and converted it to be run with CUDA.

In order to run the MAIRA-2 model, we used the processor to format and pre-process the input, then used the model.generate() function with 300 max new tokens and the use cache field set to true. Next, we got the prompt length from the shape of the processed input, and got the output by taking the generated text and finding the rest of the text after the length of the input, while skipping special tokens. We used the processor.decode() function to decode the output, removed any leading spaces, then used the processor to convert the output to plaintext, which we then returned as the final generated text. The only difference between the MAIRA-2 model without the indication and with the indication is that we passed in the indication text of the

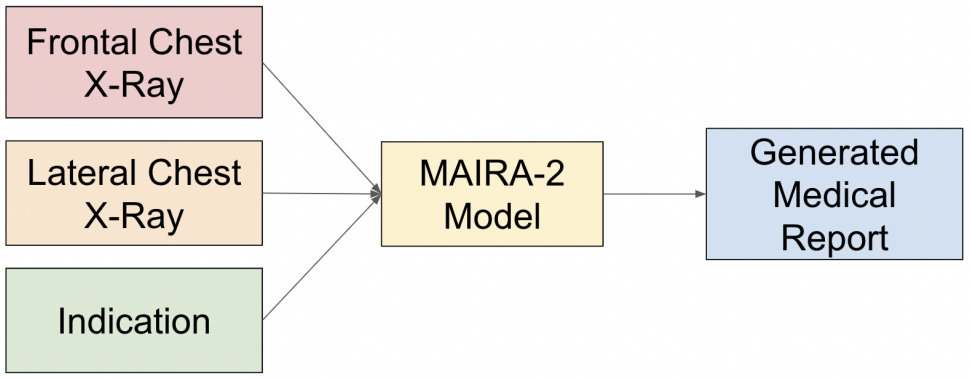

Figure 3.6: Multimodal Model with Indication Flowchart

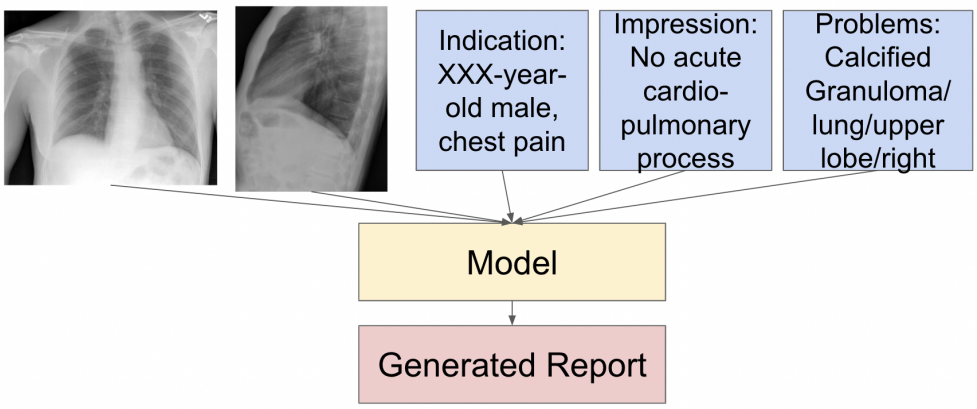

Figure 3.7: Inputs to Multimodal Model with Indication

format ”The indication is <indication>, the problems are <problems>, and the impression is <impression>” as an input in the ”indication” field of the processor’s format and pre-process input function. Figure 3.7 shows the series of inputs to the multimodal model with the indication is shown.

We repeated this process for all 500 samples, then calculated the metrics for all 500 samples, and found the average for each. These values are in the results section.

3.2 Analyzing Evaluation Metrics

3.2.1 Creating Metrics

There are four new metrics that we designed in this research. These metrics are called Word Pairs, Sentence Average, Sentence Pairs, and Sentence Pairs (Bio).

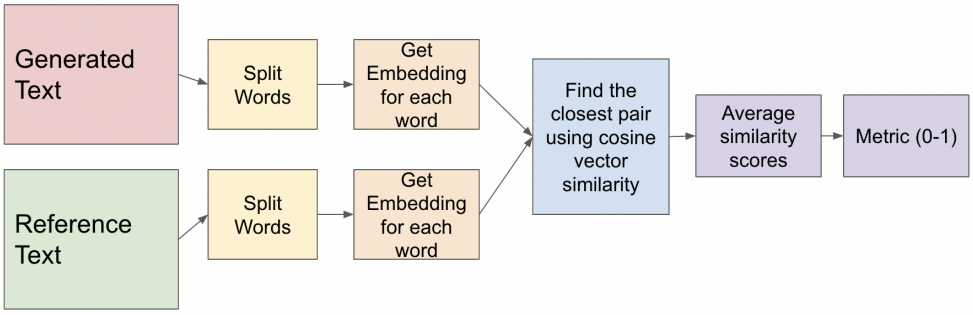

Figure 3.8: Word Pairs Flowchart

3.2.1.1 Word Pairs

The Word Pairs method is shown in Figure 3.8. First, we start by pre-processing the predicted and reference reports by splitting each piece of text into an array of all of the words, then removing all filler words from the array. Some examples of filler words that we remove are ”and”, ”or”, ”for”, ”not”, ”is”, ”a”, ”the”, ”to”, ”there”, and others.

Once we have these two arrays representing the words in the predicted and target reports, we go through each word in the target array, then iterate through each word in the reference array. When we start with the first word in the target array, we compare the first word in the target array with every word in the reference array, and find the one that is the most similar.

In order to measure how similar two given words are, we first encode each word using a word vector embedding model, specifically Word2Vec [5]. We loaded the ”Word2Vec-Google-News-300” vector embedding model using the Gensim Downloader API. This model was trained on Google News with 100 billion words, and has 3 million words across 300 dimensions. After we load the word vector for each word, we use the PyTorch cosine similarity function with the two vectors as input to calculate how similar the word vectors are. This gives us a number from 0 to 1 that represents how similar the two word vectors are.

We repeat this process of comparing the first word in the target array to all of the other words in the predicted array, and we find the word in the predicted array with the highest word vector cosine similarity. We add this number to a total score, as shown in Figure 3.9. Next, we go to the second word in the target array, and repeat this process for each word in the predicted array, then add the highest cosine similarity value to the total score. Once we have gone through all of the words in the target array, we calculate the average score by dividing the total score by the number of times we added a similarity value to the total score. This gives us a final average similarity score.

We call this metric ”Word Pairs”, because each time we add a new similarity value, that given value represents how similar a pair of words are, where the first word is from the target array, and the second word is from the predicted array.

3.2.1.2 Sentence Average

The Sentence Average method is shown in Figure 3.10. First, we split the predicted and reference text into arrays by splitting on the period character in each report, which gives us an array of

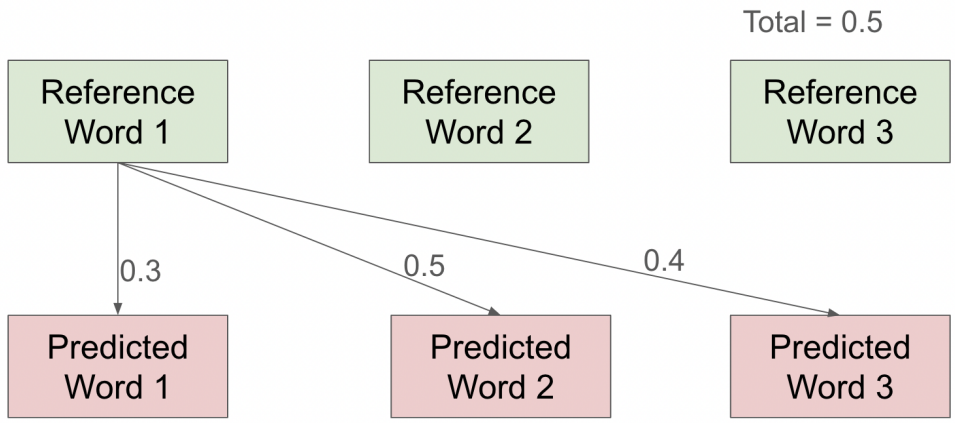

Figure 3.9: Comparing Words using the Word Pairs Method

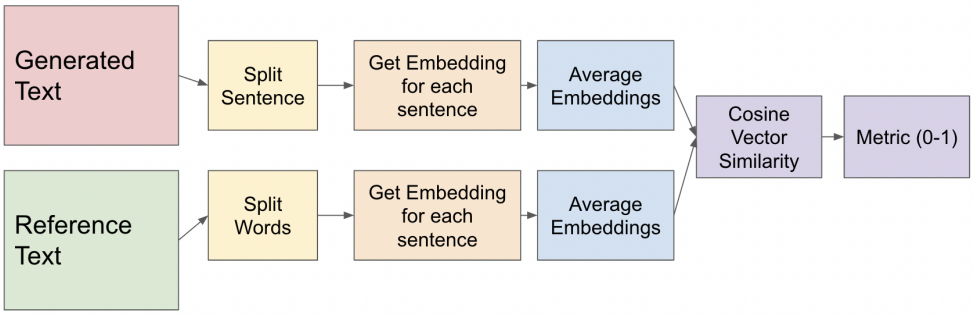

Figure 3.10: Sentence Average Overview

sentence for each report. Next, we go through each sentence in each report, and check that the sentence is at least 10 characters long, then we replace the de-identified characters ”XXXX” with the empty string so that these filler characters don’t get encoded later on.

As shown in Figure 3.11, once we have processed each of the sentences in the predicted and target arrays, we first go through each sentence in the predicted array, then get the sentence embedding for each sentence, add each embedding to a total embedding, then divide by the number of sentences to get the average embedding for the predicted array. In order to get the sentence embedding, we use a sentence transformer, specifically using the SentenceTransformer library, with the ”sentence-transformers/all-MiniLM-L6-v2” model. In order to encode the sentence, we first create a model instance using the SentenceTransformer library, then we use the model.encode() method, where we pass in each sentence as an input, with the convert to tensor method set to be true.

Similarly, as shown in Figure 3.12, we repeat this process for the reference array, where we get the target embedding for each sentence in the reference array, then average these embeddings to a final embedding for the reference array.

Once we have these two averaged sentence embeddings, we use the PyTorch cosine vector similarity function to measure how similar the averaged sentence embeddings are for the predicted and target reports. Lastly, we return the cosine vector similarity value, which is between

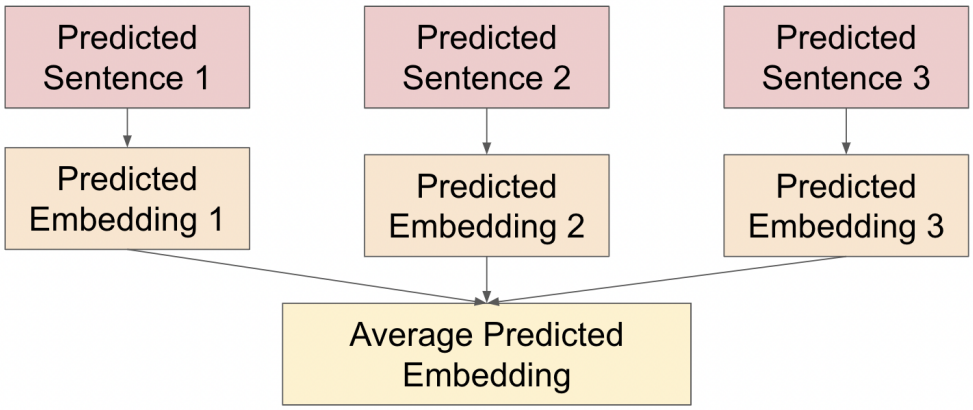

Figure 3.11: Sentence Average Method, Averaging Predicted Embeddings

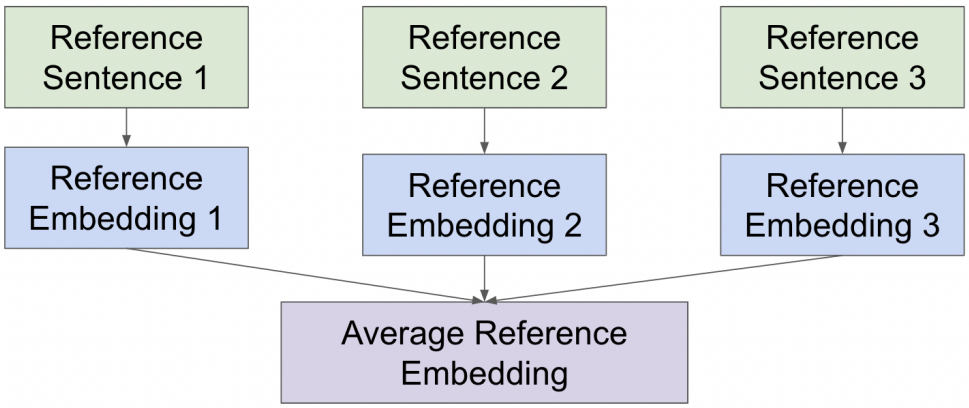

Figure 3.12: Sentence Average Method, Averaging Reference Embeddings

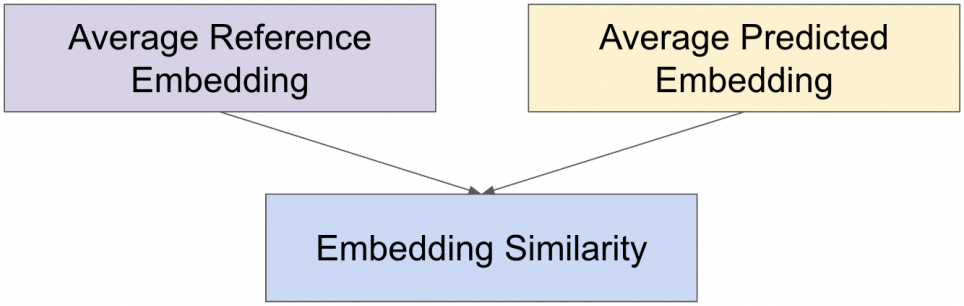

Figure 3.13: Sentence Average Method, Calculating Final Similarity

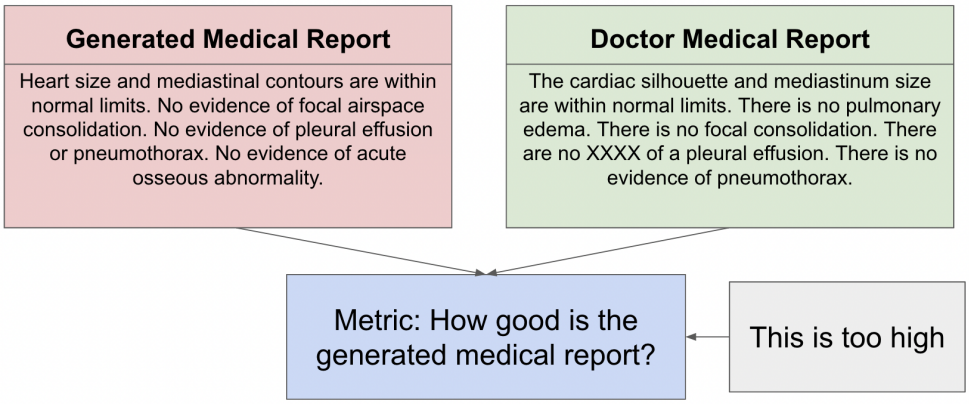

Figure 3.14: Problem with Sentence Average Score

0 and 1. This is shown in Figure 3.13.

As shown in Figure 3.14, the main problem with the Sentence Average score is that the values are all too high. This results in a call to action to design a better metric, which brings us to the next metric we created.

3.2.1.3 Sentence Pairs

The Sentence Pairs method is very similar to combining the Word Pairs and Sentence Average methods. First, we split the target and reference report into two arrays, where each array has all of the sentences in each report, just like in the Sentence Average method. Next, we go through each sentence, find the ones that have at least 10 characters, and replace the de-identified ”XXXX” characters with an empty string so that the don’t get encoded, just like we did in the Sentence Average method. The flowchart for this method is show in Figure 3.15.

Next, we go through each sentence in the target array of sentences. For the first sentence in the target, we compare that sentence with each of the sentences in the predicted array using cosine vector similarity with the same sentence embedding model from the Sentence Average method, then we add the value representing the highest cosine vector similarity for the first target sentence to a total similarity variable. We repeat this process for every sentence in the target array, then average the similarity values to get a final similarity value, which we return as the

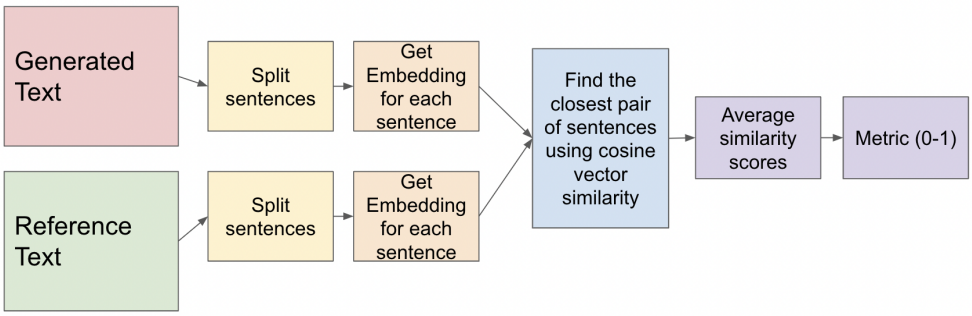

Figure 3.15: Sentence Pairs Flowchart

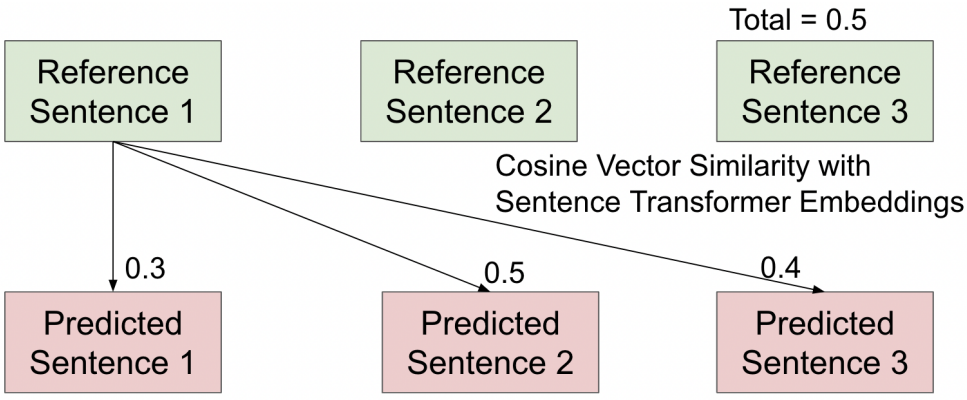

Figure 3.16: Sentence Pairs Method

final metric. This process is shown in Figure 3.16.

In other words, the Sentence Pairs method is essentially just the Word Pairs method, but instead of using the pre-processing method for words, we use the pre-processing method for sentences, instead of comparing words, we compare sentences, and instead of using Word2Vec to encode each word, we use the sentence transformer to encode each sentence.

3.2.1.4 Sentence Pairs (Bio)

The Sentence Pairs (Bio) method is a variation of the Sentence Pairs method, where instead of using a more general sentence transformer to encode each sentence, then using cosine vector similarity to compare the encoded sentence, we use radiology embeddings. Thus, the main difference is in how the similarity between the reference and predicted sentences is calculated, as shown in Figure 3.17.

First, we load the Microsoft BioMedVLP-CXR-BERT-Specialized model from HuggingFace, then get the tokenizer and the model [4]. When we compare a target sentence and a predicted sentence, we first put both of the sentences into an array with two elements, then pass this array in as two text prompts using the loaded tokenizer's batch encode plus method. We pass in the text prompts as input, and specify that we want to add special tokens, we want the longest padding, and we want to return tensors as a PyTorch tensor.

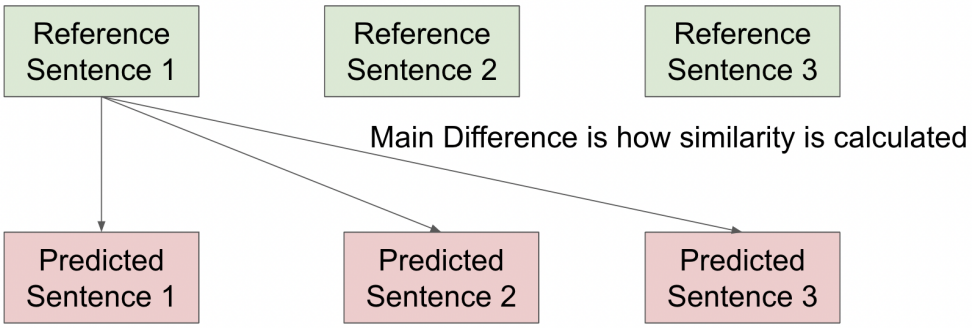

Figure 3.17: Sentence Pairs (Bio) Difference

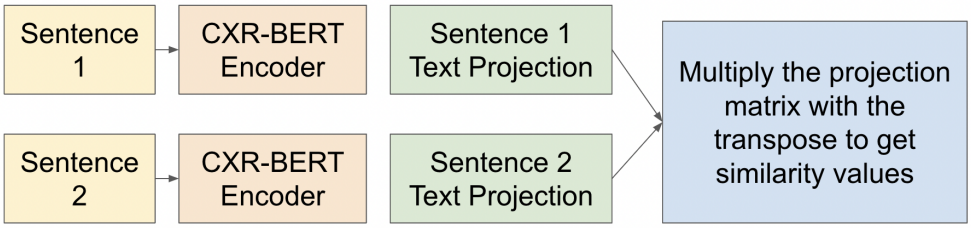

Figure 3.18: Sentence Pairs (Bio) Similarity Calculation Method

Next, we use the loaded CXR-BERT model's get_projected_text_embeddings function with the input IDs from the tokenizer output and the attention mask from the tokenizer output. Lastly, we use the torch.mm() method to multiply these embeddings with the transposed embeddings, which gives us a 2x2 similarity matrix for the predicted and target sentence. We return the value in the first row and second column, which corresponds to how similar the target sentence is to the predicted sentence, then convert the value from a tensor to a float, then return the value as the similarity score between the two sentences. Aside from this difference in calculating the sentence similarity score, the rest of the method is the exact same as the regular Sentence Pairs method. This similarity calculation method is shown in Figure 3.18.

The key intuition behind this method is that we wanted to use radiology-focused embeddings. The CXR-BERT-specialized model from Microsoft was trained on chest X-Ray information, by taking the CXR-BERT-general model, then using continual pretraining to make the model be even more specialized for Chest X-Ray information. The CXR-BERT-specialized model is also trained using contrastive learning at the end, in order to align the text and image embeddings for Chest X-Ray information.

The goal of using these radiology embeddings instead of the sentence transformer is that they might be able to better distinguish between two sentences that use keywords that are more relevant to Chest X-Ray information, since the sentences that are being passed into the CXR-BERT-specialized model to be encoded are coming from generated and reference Chest X-Ray reports. As we discuss in the results section, the Sentence Pairs (Bio) method actually performs worse than the regular Sentence Pairs method.

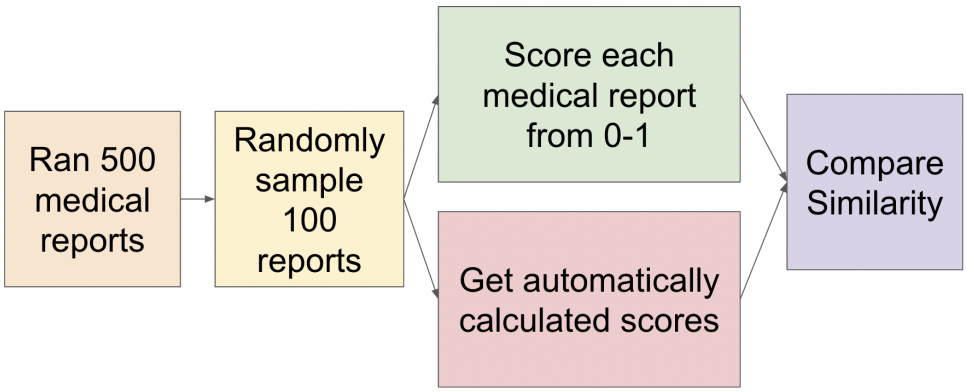

Figure 3.19: Quantitative Evaluation Method

3.2.2 Analyzing Metrics

Once we had defined both the prior metrics mentioned in the related work section and the new metrics mentioned earlier, we needed to come up with an effective method for evaluating how accurate our metrics were. In order to most effectively analyze these metrics, we used both quantitative analysis and qualitative analysis.

3.2.2.1 Quantitative Analysis

First, we randomly sampled 100 generated reports from the 500 reports that we ran our method on. For this random sample of 100 reports, we manually scored each generated report from a scale of 0 to 10. We repeated this process for all 100 pairs of generated and reference reports, which gave us a ground truth value as a manually-labeled score for how similar the generated report was to the reference report. This method is shown in Figure 3.19. The rubric that we used to manually score each generated report is shown below in Table 3.1.

Once we had these 100 manual labels for each report, we got the values for all 10 automatic metrics, including the 6 prior metrics and the 4 prior metrics, by getting the relevant metrics for each generated report. Next, we compared each of these 10 metrics with the manual score. In order to measure how tightly correlated each metric was to the manual score, we created scatter plots, then plotted the trendline and recorded the R-squared value. In order to measure how close each metric was to the manual score, we calculated the RMSE between each metric and the manual score, across all 100 generated reports. All 10 of the plots and the RMSE chart are in the results section.

3.2.2.2 Qualitative Analysis

In order to better understand the performance of our 4 new metrics on individual examples of generated and reference medical reports, we sampled 10 reports from the larger sample of 100 reports, and recorded each of these reports, along with the manual score and the values from each of the 4 new metrics for how similar each pair of generated and reference report were. We made sure that each report had a different value for the manual score, to make sure that we could see

how the metrics performed across different levels of how good the report was. The table with these 10 reports, along with the manual score for each generated report and the values of the 4 new metrics for each generated report is in the results section.

Manual Score Rubric	
Similarity of the Generated Report to the Reference Report	**Manual Score**
The generated report is focused on a different topic from the reference report with no keywords in common	0
All major important details missing, but there is at least one relatively important keyword mentioned	1
Almost all major important details missing, except for one or two important keywords	2
At least two very important pieces of information are missing, or the generated report has a relatively different meaning from the reference report	3
Half of the report is the same as the reference report, but multiple important pieces of information are not included	4
Most important information is included, but there are multiple important keywords not included	5
Mostly similar report, except missing one very important piece of information or two pieces of relatively important information	6
Mostly similar report, except missing one important piece of information	7
Similar report, except for a few keywords, at least one of which is relatively important to include	8
Extremely similar report, except for a few keywords that aren't that important to include	9
Exact same report, with the exception of 1 or 2 terms	10

Table 3.1: Manual Score Rubric for Evaluating Metrics

Chapter 4

Results

4.1 Model Comparison

There were three main types of models that we evaluated, specifically the fine-tuned Medical LLM, the MAIRA-2 multimodal model without symptom information as an input, and the MAIRA-2 multimodal model with symptom information as an input.

In order to compare the performance of each of these 3 models, we ran each of them separately on the same 500 samples, compared the predicted report with the reference report for each sample, calculate all of the metrics for each sample, then averaged the metrics across all 500 samples.

As mentioned, there were a total of 10 metrics that we measured. The first 5 are classical prior work, specifically BLEU-1, BLEU-2, ROUGE-1, ROUGE-2, and ROUGE-L. The 6th metric is more recent prior work, called the RaTE Score, which was developed in 2024 by a group of researchers as a metric specifically for radiology report generation. The last four metrics are all new metrics presented in this research paper, specifically the Word Pairs score, Sentence Average score, Sentence Pairs score, and the Sentence Pairs (Bio) score.

4.1.1 BLEU score

Table 4.1, below, shows the BLEU score across all three models. As the table shows, the BLEU-1 score for the multimodal models is significantly higher than that of the medical LLM, with 0.149 and 0.207 for MAIRA-2, compared to 0.105 for the medical LLM. The BLEU-2 score shows a similar pattern, where the BLEU-2 score for the multimodal models is 0.053 and 0.067, compared to the medical LLM, which has a score of 0.020. The BLEU scores also show that, within the context of the MAIRA-2 model, introducing the symptoms as an input helps improve the model's performance, with the BLEU scores for MAIRA-2 with the indication being 0.207 and 0.067, compared to 0.149 and 0.053 without the indication.

4.1.2 ROUGE Score

Table 4.2, below, shows the ROUGE-1, ROUGE-2, and ROUGE-L scores for all 3 models. Similar to the bLEU scores, we can see that there is a significant benefit that MAIRA-2 has

BLEU Score Across Methods		
Model	**BLEU-1**	**BLEU-2**
Medical LLM	0.105	0.020
MAIRA-2 (no indication)	0.149	0.053
MAIRA-2 (indication)	0.207	0.067

Table 4.1: Comparison of BLEU Metric Values for Each Model

in comparison the medical LLM. We can also see that there is a slight benefit from adding the indication for the MAIRA-2 model with the indication, since the values are 0.367, 0.126, and 0.257, as compared to 0.341, 0.123, and 0.246 for the model without the indication.

ROUGE Score Across Methods			
Model	**ROUGE-1**	**ROUGE-2**	**ROUGE-L**
Medical LLM	0.190	0.040	0.126
MAIRA-2 (no indication)	0.341	0.123	0.246
MAIRA-2 (indication)	0.367	0.126	0.257

Table 4.2: Comparison of ROUGE Metric Values for Each Model

4.1.3 RaTE Score

As shown by Table 4.3, below, the RaTE score shows a similar pattern as the BLEU and ROUGE scores. The RaTE score for MAIRA-2 is much higher than the RaTE score from the medical LLM, with the MAIRA-2 model without the indication having a RaTE score of 0.265, while the medical LLM has a RaTE score of 0.207. Similarly, we can see that the MAIRA-2 model with the indication has slightly better performance than the MAIRA-2 model without the indication.

RaTE Score Across Methods	
Model	**RaTE Score**
Medical LLM	0.207
MAIRA-2 (no indication)	0.265
MAIRA-2 (indication)	0.276

Table 4.3: Comparison of RaTE Metric ValuesValues for Each Model

4.1.4 Word Pairs and Sentence Average

Table 4.4, below shows the scores from the Word Pairs and Sentence Average methods. First, we can see that the Word Pairs score gives the medical LLM an extremely high score, at 0.462. This makes sense, because the medical LLM is prompted with the indication information, which consists of relevant keywords for the problems that the given patient is having, along with relevant symptoms. When the fine-tuned medical LLM is prompted with this information, it is

more likely that the model will include these keywords in the generated report. The Word Pairs method, as mentioned earlier, tries to find relevant keywords from the reference text in the generated text, which means that the generated text is more likely to have these words, since it has been prompted with these keywords. By contrast, the MAIRA-2 model without the indication has not been given any information about relevant symptoms or problems, which means that the MAIRA-2 model without the indication does not have the ability to mention these keywords. Thus, we can see that the Word Pairs score for the medical LLM is higher than that of MAIRA-2 without the indication. Similarly, we can see that the MAIRA-2 model with the indication information performs better than the MAIRA-2 model without the indication, with the model having a Word Pairs score of 0.514 with the indication, compared to 0.443 without the indication.

In terms of the Sentence Average score, we can see two key observations. First, we notice the same pattern of the MAIRA-2 model being more effective than the medical LLM. This makes sense, because instead of searching for specific keywords like the Word Pairs metric, the Sentence Average metric calculates sentence embeddings for each sentence, averages them, then compares the two vector cosine similarity values. Thus, the medical LLM doesn't get a much larger advantage from being given relevant keywords, since it still needs to generate sentences that are similar to that of the reference report. The other key observation is that all of the Sentence Average metric values are extremely high in comparison to other metrics. For example, the value for the medical LLM is 0.575, the value for MAIRA-2 without the indication is 0.748, and the value for MAIRA-2 with the indication is 0.745. Although MAIRA-2 is relatively effective at medical report generation, an average score around 0.75 is higher than the manually-graded average from the sample of 100 reports, which was around 0.62. This shows us that the Sentence Average metric might not be the best metric to use.

Word Pairs and Sentence Average Values for Each Model		
Model	**Word Pair**	**Sentence Average**
Medical LLM	0.462	0.575
MAIRA-2 (no indication)	0.443	0.748
MAIRA-2 (indication)	0.514	0.745

Table 4.4: Comparison of Word Pairs and Sentence Average for Each Model

Putting together the last two metrics, we can introduce two new metrics, specifically Sentence Pairs and Sentence Pairs (Bio). Table 4.5, below, shows these values across all of the models.

As shown below, the Sentence Pairs method shows a significant benefit from using MAIRA-2 compared to the medical LLM. We can see that the Sentence Pairs metric has a value of 0.5583 for MAIRA-2 without the indication and 0.5792 for MAIRA-2 with the indication, which is much higher than the medical LLM value of 0.417.

Based off of the intuition of the Sentence Pairs method, we would expect that using the CXR-BERT embeddings would improve the performance of the metric. However, Table 6 shows that this isn't the case. As shown by the metrics for the Sentence Pairs (Bio) metric, the metric is extremely high across all 3 models. We see the same pattern of the MAIRA-2 model performing better than the medical LLM, but all of the models have a value greater than 0.7, which means that all of the metric values are higher than expected. Similar to the other metrics, the Sentence

Pairs (Bio) method shows a significantly higher value for MAIRA-2 than the medical LLM, but the value of MAIRA-2 with the indication is only slightly higher than MAIRA-2 without the indication.

Sentence Pair and Sentence Pairs (Bio) Values for Each Model		
Model	**Sentence Pair**	**Sentence Pairs (Bio)**
Medical LLM	0.417	0.707
MAIRA-2 (no indication)	0.5583	0.797
MAIRA-2 (indication)	0.5792	0.804

Table 4.5: Comparison of Sentence Pair and Sentence Pairs (Bio) for Each Model

4.2 Evaluation Metric Comparison

As mentioned earlier, we measured 6 prior metrics and created 4 new metrics in this paper. In order to best measure how effectively these metrics are able to grade medical reports, we decided to run a small study where we manually graded a subset of medical reports and compared these manual grades with the evaluation metric values from all 10 of these evaluation metric methods.

As mentioned earlier, we ran the prediction algorithm on 500 reports from the IU-XRay dataset. In order to get a random sample, we randomly sampled 100 reports form this subset, and manually graded the generated report for the MAIRA-2 model with the indication. For each report, we compared the generated report with the reference report to check how similar the reports were. In particular, we checked to see if there were any important details that were not in the generated report when compared to the reference report. We graded each of these reports from a scale of 0 to 10, where 0 means that there were no similarities across the predicted and reference reports, and 10 means that they were exactly the same, with the exception of a few words that are synonyms of each other.

After manually grading these reports on a scale from 0 to 10, we converted the score to a score from 0 to 1 by dividing each of the grades for the reports by a factor of 10. Once we had these scores from 0 to 1, we created a series of scatterplots showing these values for each report, compared to the automatically generated metric value. We repeated this process for all 10 metrics in order to compare how similar the metric's value was to a manually scored value for how similar the generated and reference text are.

4.2.1 Plot Comparison

4.2.1.1 BLEU Score Plot Comparison

Figures 4.1 and 4.2 show that the BLEU-1 score is not an effective metric compared to the manual score, mainly because the BLEU-1 score underestimates the value compared to the BLEU-1 score. For example, in Figure 4.1, for the manual scores between 0.8 and 1, the majority of the BLEU-1 scores are below 0.5, which shows that BLEU-1 scores are too low in comparison to the actual values needed. Figure 1 also shows the equation for the trendline, which is y=0.148x +

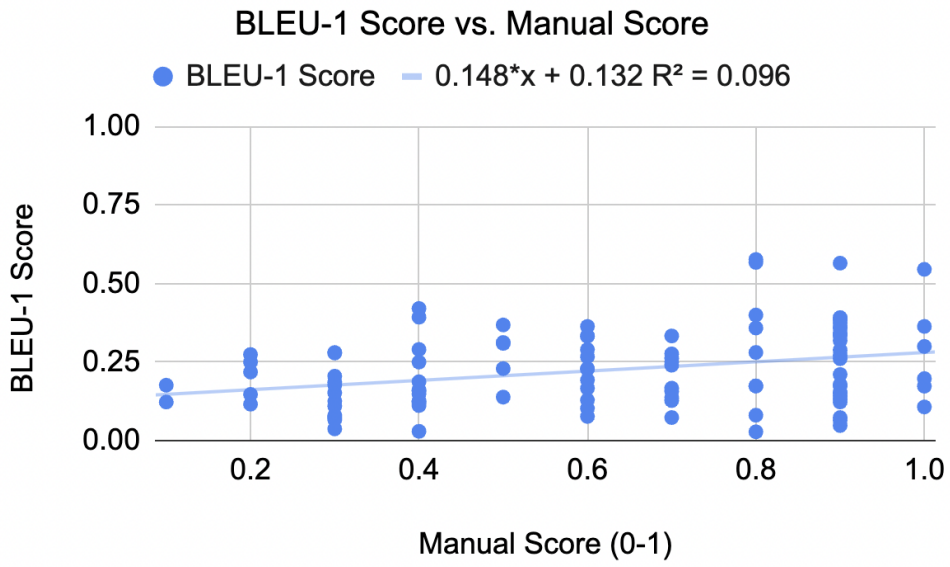

Figure 4.1: BLEU-1 Score vs. Manual Score Scatter Plot

0.132, and the R-squared value, which is 0.096. This shows us that the BLEU-1 score is not very tightly correlated with the manual score, which means that the BLEU-1 score is not the most effective metric. In Figure 4.2, the standardized version of Figure 4.1, we see the same pattern, where there is high variance, with the points that have a high manual score still sometimes being low, even in the standardized version of the BLEU-1 score. This shows us that BLEU-1 is not the most effective metric, even with standardization. Figures 4.3 and 4.4 show a similar pattern of BLEU scores being too low. However, since BLEU-2 looks for similarities in pairs of words, we can see that the plot has values that are even lower than BLEU-1. In Figure 4.3, the majority of the BLEU-2 scores are under 0.25, even for manual scores that are between 0.8 and 1. Similar to the BLEU-1 metric, the equation for the trendline for the BLEU-2 score plot is y=0.0971*x + 0.0118, while the R-squared value is 0.104. The R-squared value is extremely low, which shows us that the BLEU-2 score is not the most effective metric. As shown in Figure 4.4, when we standardize the values, we see that the BLEU-2 score values has several outliers that are significantly above the manual score values, which shows that the BLEU-2 score might overestimate values compared to the manual score, when it is standardized.

4.2.1.2 ROUGE Score Plot Comparison

As shown in Figures 4.5 and 4.6, the ROUGE-1 score is much more effective at being more correlated with the manual score. The trendline in Figure 4.5 has an equation of y=0.233*x+0.242, which has a higher coefficient than the BLEU-1 and BLEU-2 score coefficients. Similarly, the R-squared value is much higher for ROUGE-1, with a value of R-squared = 0.237. Thus, ROUGE-1 seems like a more effective metric than BLEU-1 and BLEU-2. Similarly, Figure 4.6 shows that most points are relatively close to the trendline, which means that the ROUGE-1 score more closely matches the manual score than any of the BLEU metrics. We can also see that the stan-

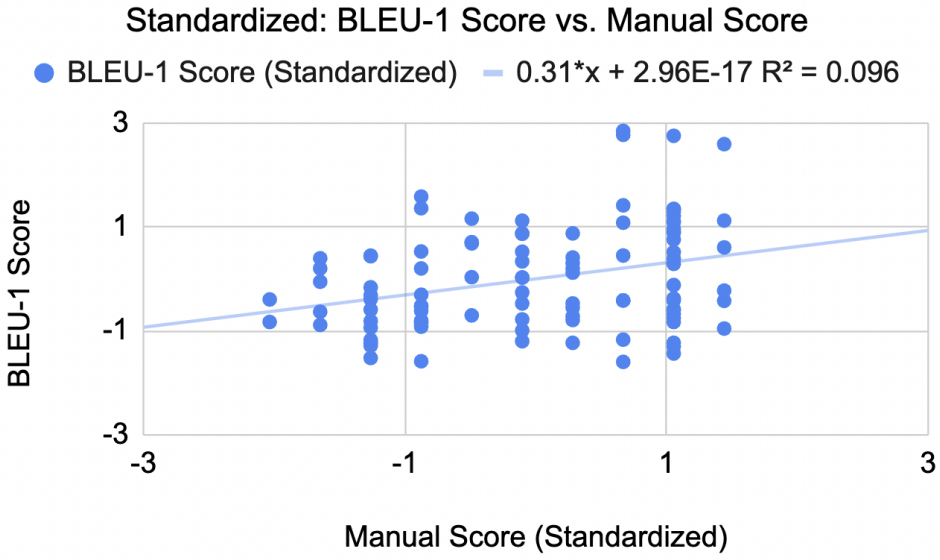

Figure 4.2: Standardized: BLEU-1 Score vs. Manual Score Scatter Plot

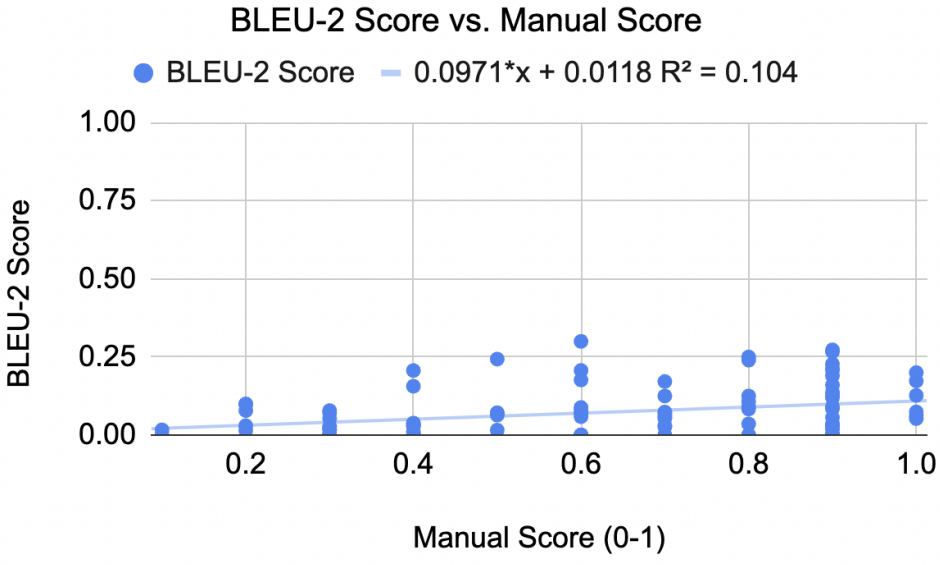

Figure 4.3: BLEU-2 Score vs. Manual Score Scatter Plot

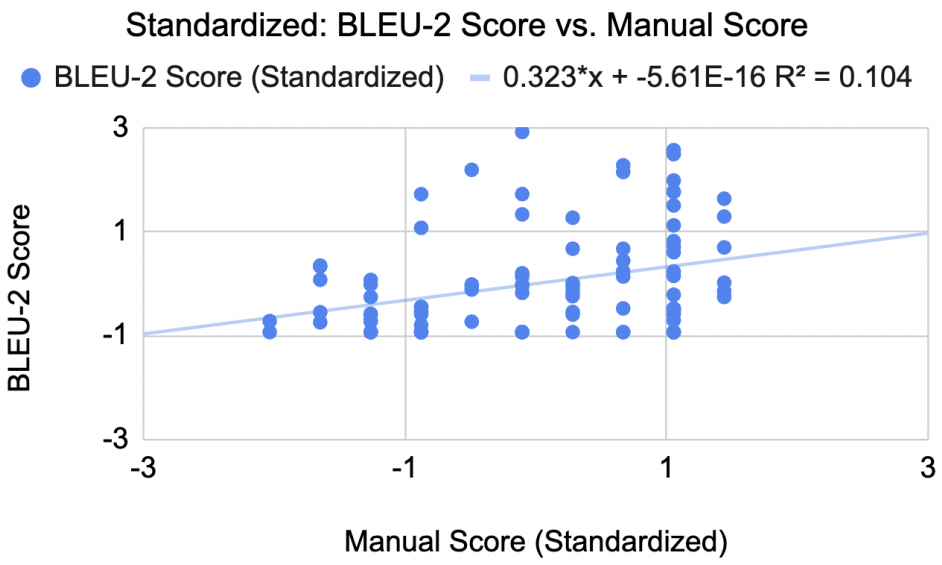

Figure 4.4: Standardized: BLEU-2 Score vs. Manual Score Scatter Plot

dardized slope is 0.487, which is significantly higher than the standardized slope of the BLEU-1 and BLEU-2 metrics, which were 0.31 and 0.323.

As shown in Figures 4.7 and 4.8, ROUGE-2 seems more effective than the BLEU score metrics, but also seems to be much lower than the manual scores. As shown in Figure 4.7, we can see that all of the points with a manual score between 0.8 and 1.0 have a ROUGE-2 score below 0.5, which shows us that the ROUGE-2 score is too low. The R-squared value is higher than the BLEU-1 and BLEU-2 metrics, with an R-squared value of 0.179, but the R-squared value is lower than the R-squared value for the ROUGE-1 metric, which was 0.237. Figure 4.8 shows a similar pattern, with there being high variance in the ROUGE-2 score when the standardized manual score is 1, and a standardized slope of 0.423, which is lower than the ROUGE-1 standardized slope of 0.487. Thus, we can see that ROUGE-2 performs worse than ROUGE-1.

As shown in Figures 4.9 and 4.10, ROUGE-L has the highest R-squared value out of all of the BLEU and ROUGE metrics, with an R-squared value of 0.251. This makes sense, since ROUGE-L measures the longest common subsequence between the generated and reference reports, and reports with the same subsequence of words are more likely to be much more similar to each other. As shown in Figure 4.9, similar to ROUGE-2, ROUGE-L values are on the lower end. This makes sense, because in order for two reports to have a high ROUGE-L value, they would have to use the exact same words in the same order, which is very rare. Figure 4.10 shows this very clearly, with most points being towards the center and there being relatively few outliers. In addition, Figure 4.10 has a standardized slope of 0.501, which is higher than the standardized slope from BLEU-1, BLEU-2, ROUGE-1, and ROUGE-2.

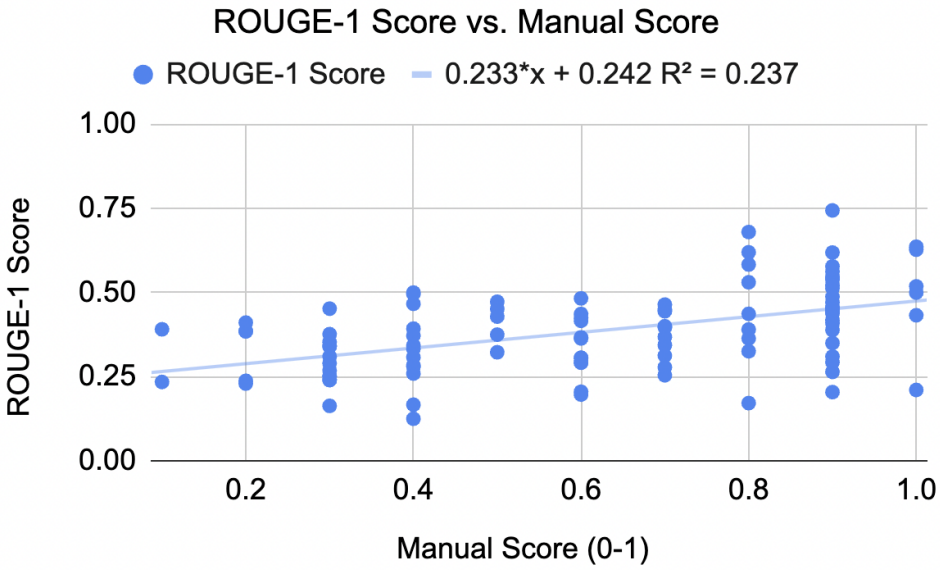

Figure 4.5: ROUGE-1 Score vs. Manual Score Scatter Plot

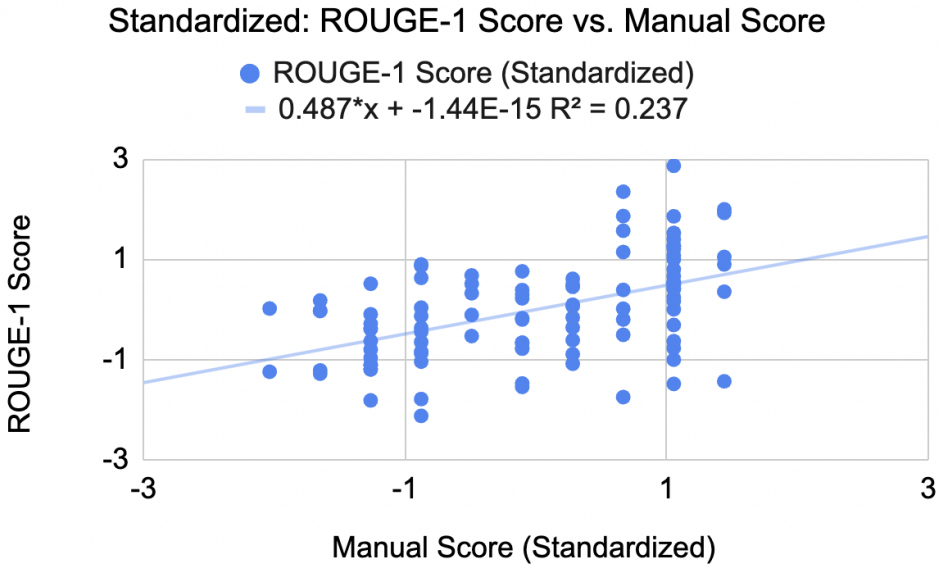

Figure 4.6: Standardized: ROUGE-1 Score vs. Manual Score Scatter Plot

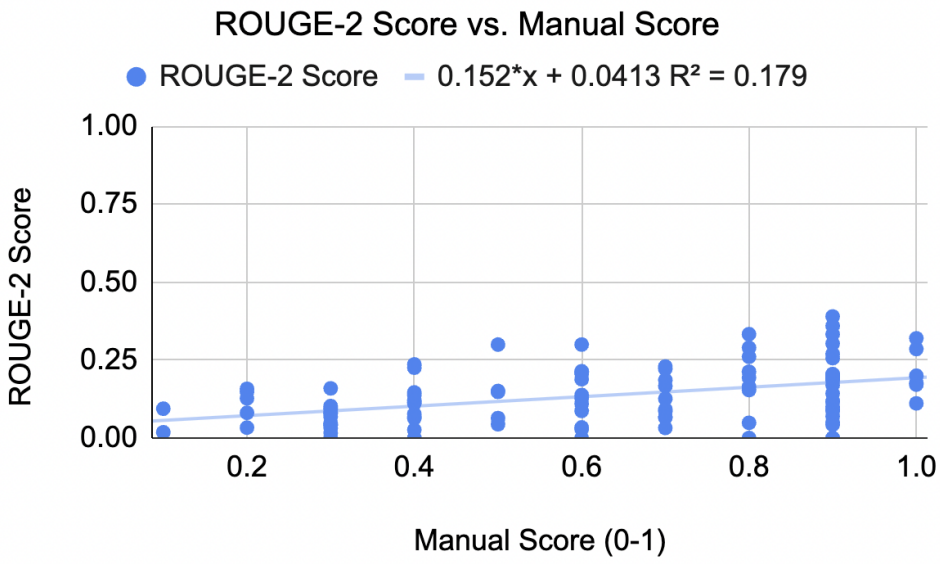

Figure 4.7: ROUGE-2 Score vs. Manual Score Scatter Plot

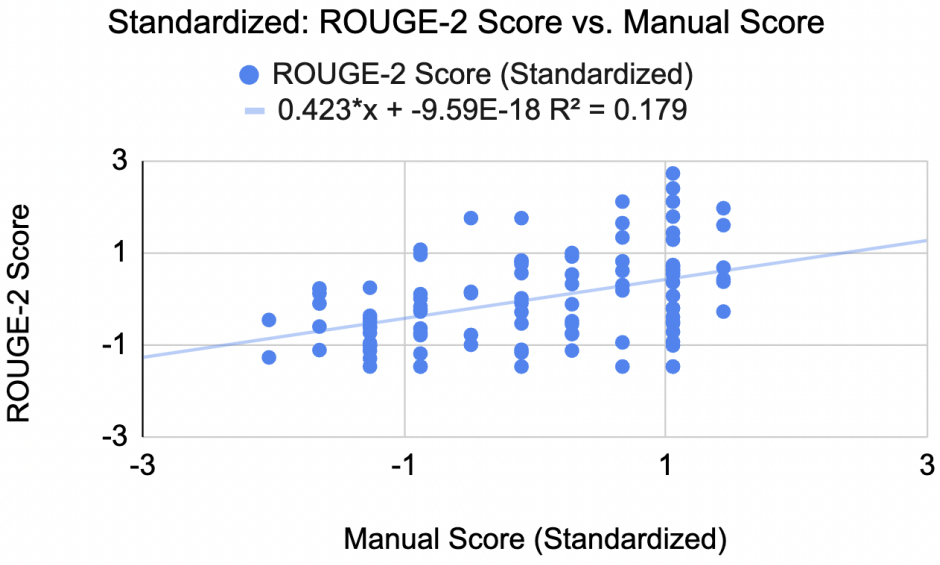

Figure 4.8: Standardized: ROUGE-2 Score vs. Manual Score Scatter Plot

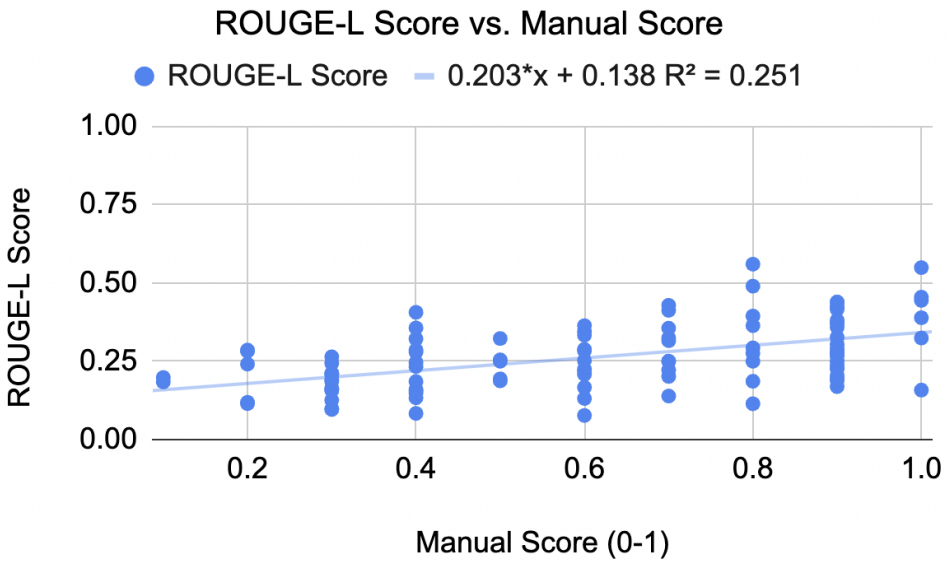

Figure 4.9: ROUGE-L Score vs. Manual Score Scatter Plot

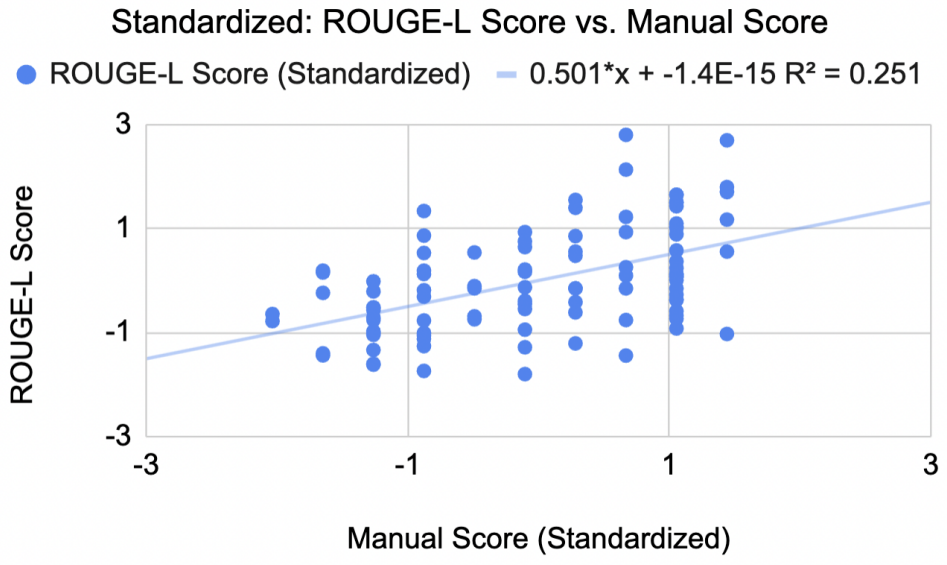

Figure 4.10: Standardized: ROUGE-L Score vs. Manual Score Scatter Plot

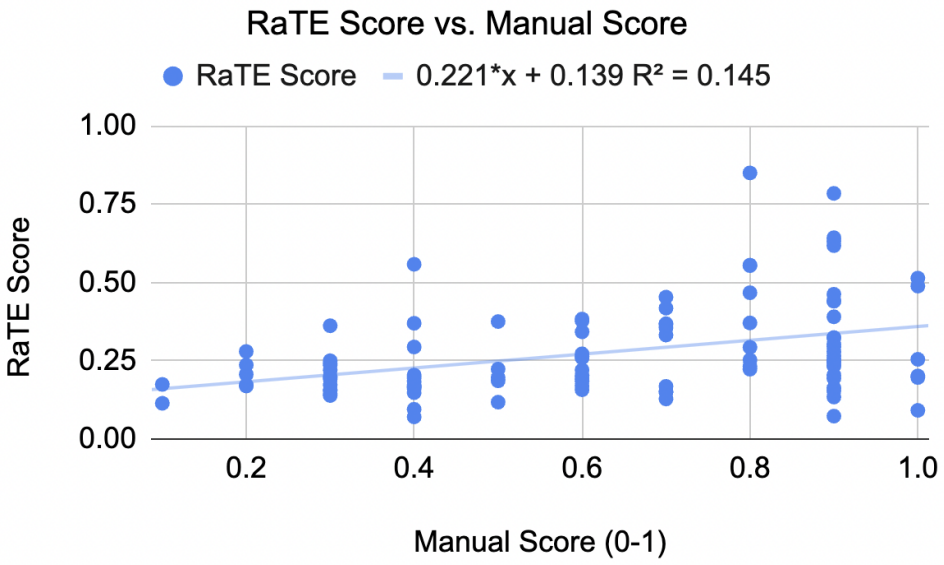

Figure 4.11: RaTE Score vs. Manual Score Scatter Plot

4.2.1.3 RaTE Score Plot Comparison

As shown in Figures 4.11 and 4.12, the RaTE score also has a low R-squared value of 0.145. As shown in Figure 4.11, unlike the BLEU and ROUGE metrics, the RaTE score has some points that have a high score, especially for points with a manual score between 0.8 and 1. In addition, the trendline equation is y=0.221x+0.139, which has a relatively high coefficient. However, the majority of the RaTE score values are under 0.5, which shows that the RaTE score gives values that are too low, in comparison to the manual score. In addition, the R-squared value for the ROUGE-L metric is much higher, which shows that the RaTE score is not the most effective metric, even when compared to past metrics. Figure 4.12 shows a similar trend, where when the standardized manual score is 1, the majority of the standardized RaTe score values are under 1, with only a few points above 1.

4.2.1.4 Word Pairs and Sentence Average Plot Comparison

As mentioned earlier, there are 4 new metrics that we introduce in this paper, which are Word Pairs Sentence Average, Sentence Pair, and Sentence Pairs (Bio). In this section, we'll look at how effective the Word Pairs and Sentence Average metrics are.

As shown in Figures 4.13 and 4.14, the Word Pairs score has values closer to 0.5 on average, but the R-squared value is 0.133, which is relatively low. The trendline for the points in Figure 4.13 is y=0.159x + 0.43, which is relatively low as well. This shows us that, although the values for the Word Pairs metric are closer to the expected values on average, the Word Pairs values are not as tightly correlated with the manual score as other methods. Figure 4.14 shows a similar pattern, where there is high variance in the standardized Word Pairs score when the standardized manual score is slightly above 1, but significantly lower variance when the manual score is less

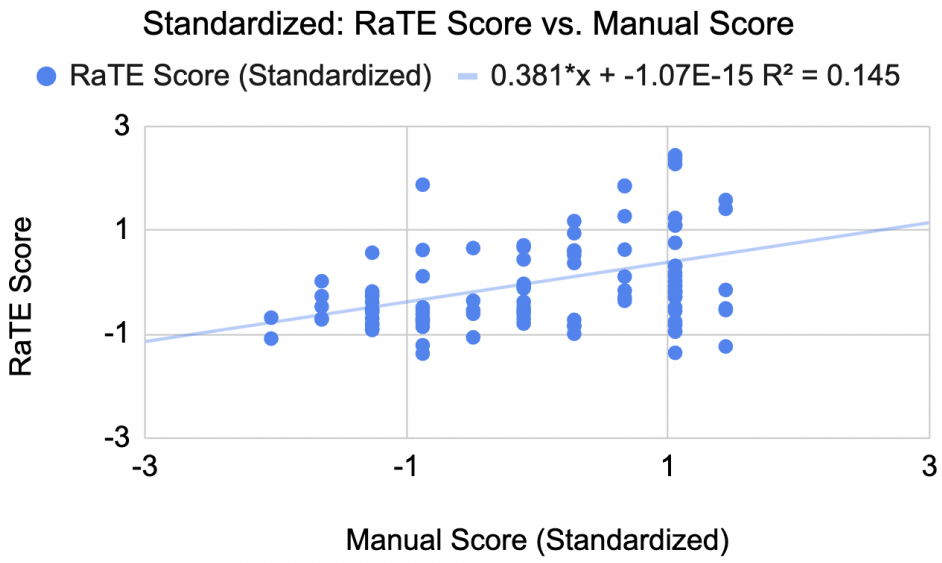

Figure 4.12: Standardized: RaTE Score vs. Manual Score Scatter Plot

than 0.

As shown in Figures 4.15 and 4.16, the Sentence Average score has extremely weak performance in terms of being correlated with the manual score. The trendline equation for the Sentence Average score in Figure 4.15 is y=0.0612x+0.722, which has a very low coefficient. The R-squared value is 0.054, which is also extremely low, thus showing that the Sentence Average score does not have a strong correlation with the manual score. As shown in the plot, we can see that the majority of values for the Sentence Average are around 0.75, and all of the values are above 0.5, even for generated reports that have a manual score between 0 and 0.2. Given this, we can see that the Sentence Average metric has values that are too high and is also very weakly correlated with the manual score. This shows us that using the average method is not as effective as the pairing method used in the Word Pairs method. This pattern is also shown in Figure 4.16, where the standardized Sentence Average score points are far from the trendline across each of the standardized manual score points. In addition, the corresponding trendline slope for the Sentence Average score is 0.232, which is much lower than any of the other standardized trendline slope values.

4.2.1.5 Sentence Pair and Sentence Pairs (Bio) Plot Comparison

Building upon the intuition from the Word Pairs metric and the Sentence Average metric, we can look at how effective the Sentence Pairs metric is. As shown in Figures 4.17 and 4.18, this metric has the most effective correlation with the manual score. The R-squared value is 0.283, which is the highest out of all of the metrics measured. In addition, the trendline equation in Figure 4.17 is y=0.208x + 0.461, which has a much higher slope than both the Word Pairs method and Sentence Average methods. One potential limitation that the Sentence Pairs score has is that almost all of the score value are above 0.5, even for values with a manual score between 0 and

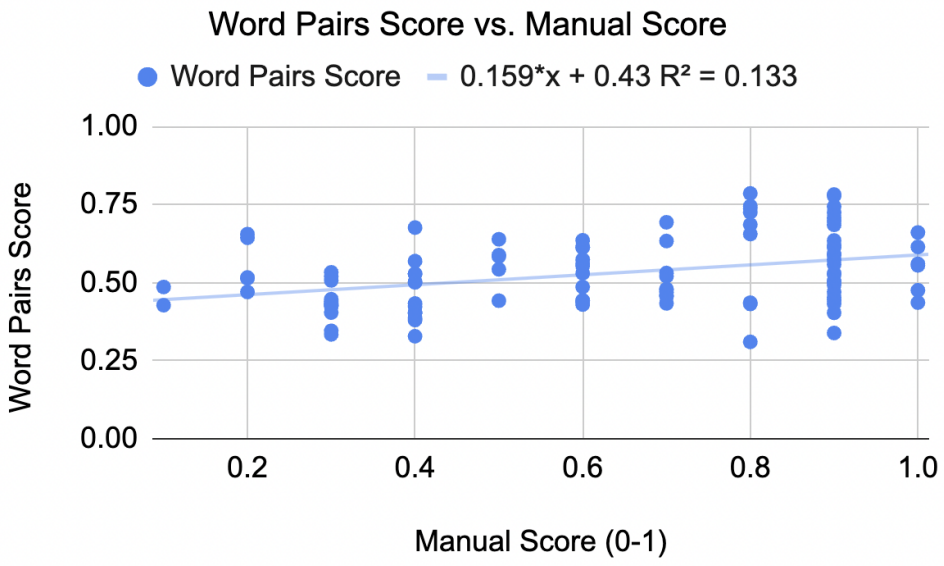

Figure 4.13: Word Pairs Score vs. Manual Score Scatter Plot

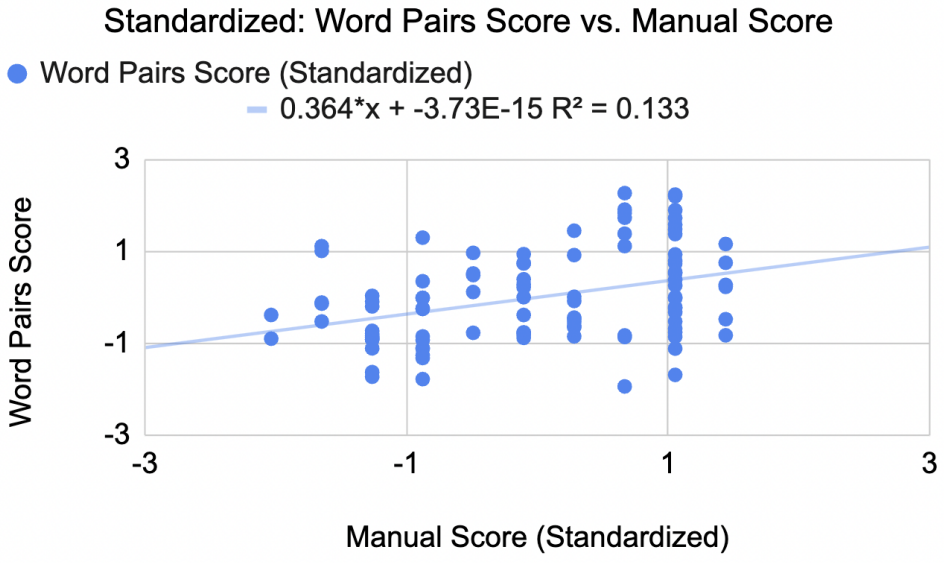

Figure 4.14: Standardized: Word Pairs Score vs. Manual Score Scatter Plot

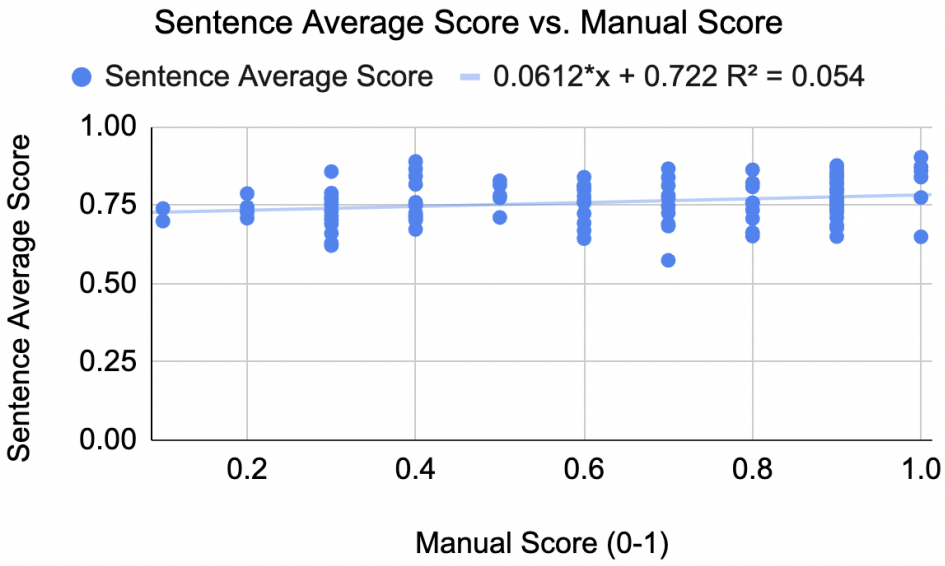

Figure 4.15: Sentence Average Score vs. Manual Score Scatter Plot

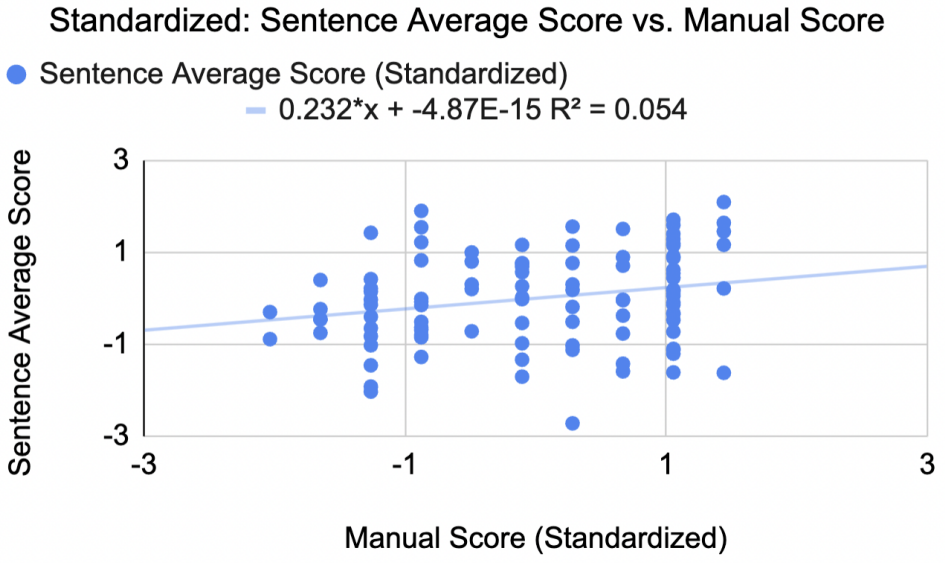

Figure 4.16: Standardized: Sentence Average Score vs. Manual Score Scatter Plot

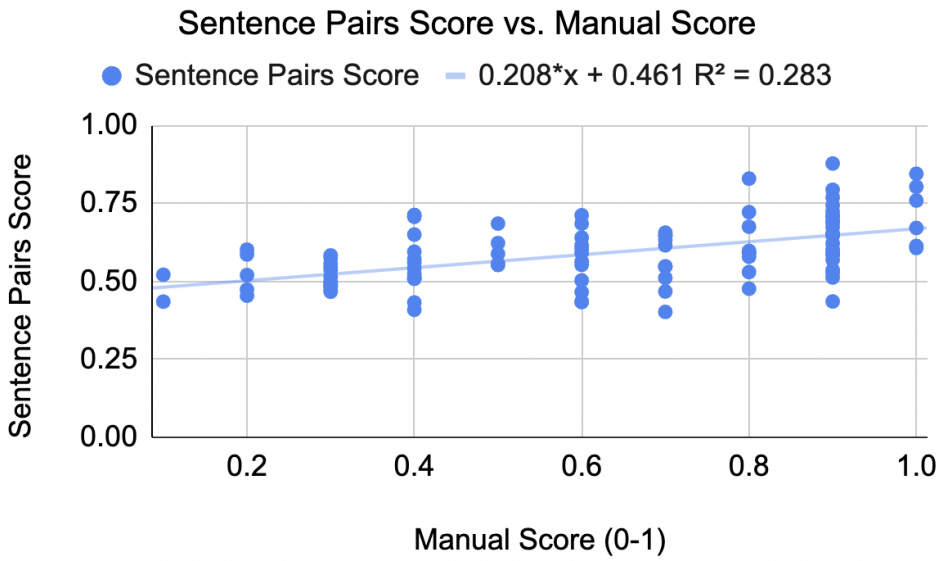

Figure 4.17: Sentence Pairs Score vs. Manual Score Scatter Plot

0.2, but even considering this, the Sentence Pairs metric has the best performance out of all of the metrics measured. This pattern is also shown in Figure 4.18, where all of the points are relatively close to the trendline, and the slope is higher than all past metrics, at 0.532. Thus, we can see that the Sentence Pairs metric is the most effective metric.

Similar to the Sentence Pairs metric, the Sentence Pairs (Bio) metric uses the same concept, but with radiology sentence embeddings instead of sentence transformer embeddings. As shown in Figure 4.19 and 4.20, the Sentence Pairs (Bio) method has the majority of points having a score over 0.75, which means that the values are much higher than the manual score. In addition, the equation for the trendline in Figure 4.19 is y=0.189x+0.7, which has a lower slope than the Sentence Pairs metric without the radiology embeddings. The R-squared value for this metric is 0.211, which is much lower than the Sentence Pairs score without the CXR-BERT embeddings, which had a value of 0.283. Thus, we can see that the Sentence Pairs metric without the CXR-BERT embeddings has the highest performance. This pattern is also shown in Figure 4.20, where there are several outliers with significantly lower standardized Sentence Pairs (Bio) scores than the standardized manual scores, and the points are spread out. In addition, the standardized trendline slope is 0.459, which is lower than both the Sentence Pairs metric and ROUGE-L metric. This shows that Sentence Pairs (Bio) is not the most effective metric.

4.2.2 RMSE Comparison

Although using scatterplots is an effective way of measuring how closely the metrics correlate with the manual score, another important method for determining how effective metrics are is measuring their numerical similarity to the manual score. One technique that is often used to compare a series of predicted and actual values is RMSE, or root mean squared error. In order to compute this, we went through each of the 100 manual scores from 0 to 1 and calculated the

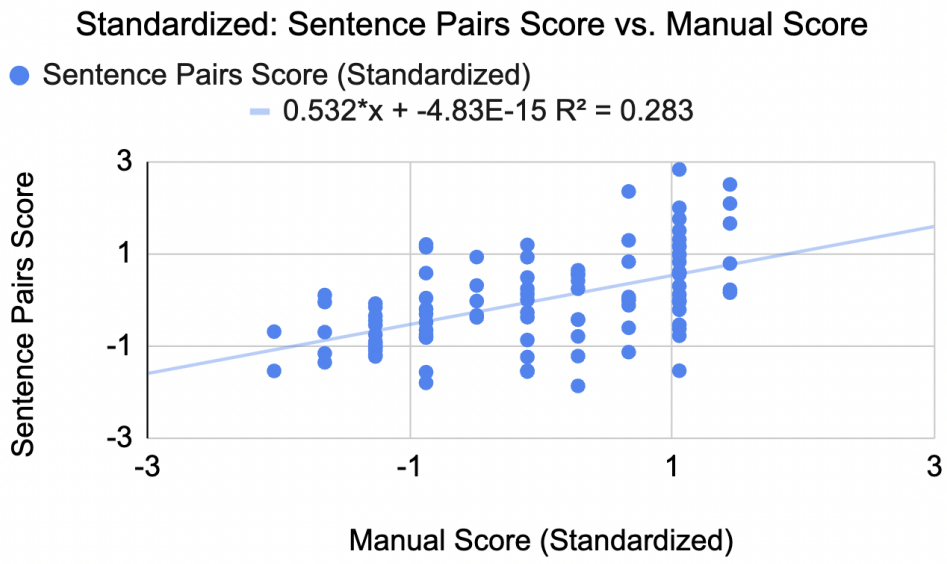

Figure 4.18: Standardized: Sentence Pairs Score vs. Manual Score Scatter Plot

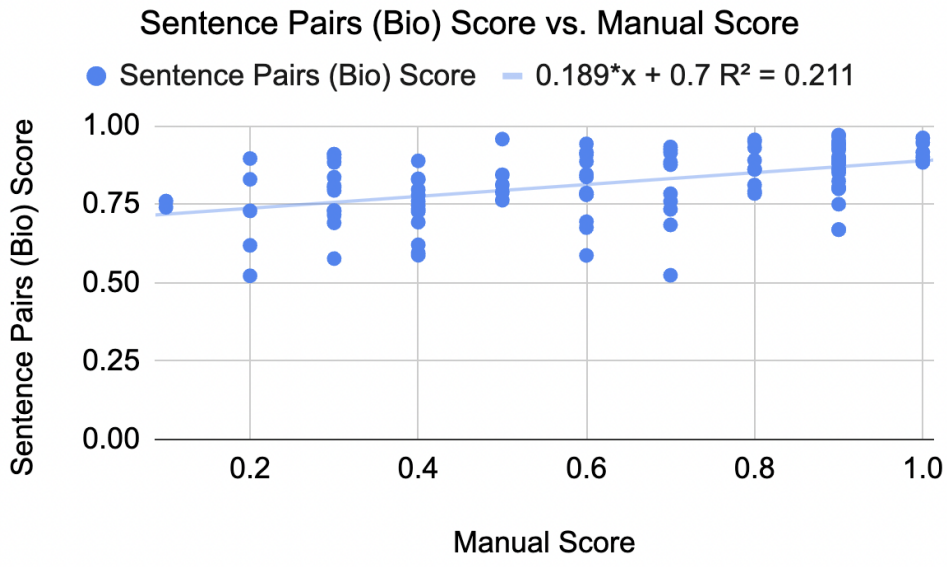

Figure 4.19: Sentence Pairs (Bio) Score vs. Manual Score Scatter Plot

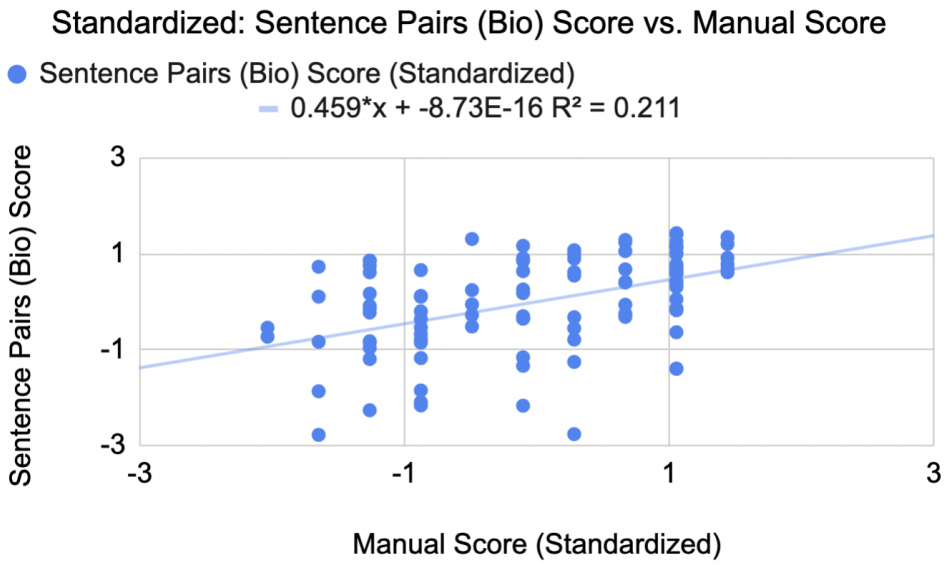

Figure 4.20: Standardized: Sentence Pairs (Bio) Score vs. Manual Score Scatter Plot

square of the difference between the manual score and the metric score. Next, we averaged these values across all 100 rows of data, then took the square root of the average value. We repeated this process for all 10 metrics.

In order to make sure that comparisons were fair across each of the metrics, we also standardized the RMSE values. In order to do this, we standardized each set of 100 measurements per metric using the equation standardized metric = (original metric value - mean across all 100 original metric values)/(standard deviation across all 100 original metric values).

Table 4.6, shown below, shows these values for all 10 of the metrics. As shown by the table, the standardized RMSE value is the lowest for the Sentence Pairs method, which shows that the Sentence Pairs metric is closest to the manual score. We can also see that the RMSE values for the Sentence Average metric is quite high, at around 1.2, which shows that the Sentence Average metric does not accurately measure how similar the generated and reference medical reports are. All 4 of the metrics that we introduce in this research is relatively close to 1, while the BLEU-2 score is much higher, at 1.5. Overall, we can see that there is variance between the quality of the metrics that we introduce, but the Sentence Pairs metric is the most accurate within all of the new metrics, and outperforms all of the prior metrics. This is also shown visually in Figure 4.21.

4.2.3 R-squared Comparison

Another way to measure the quality of these metrics is to use the R-squared value, which stays the same both with and without standardization.

As shown in Table 4.7, the metric with the highest R-squared value is the Sentence Pairs metric, which has an R-squared value of 0.283. This further reinforces our conclusion from the RMSE comparison, and shows that the Sentence Pairs metric more effectively measures the

Standardized RMSE Values for all 10 metrics	
Model	**Standardized RMSE Value**
BLEU-1	1.169
BLEU-2	1.641
ROUGE-1	1.008
ROUGE-2	1.068
ROUGE-L	0.994
RaTE	1.107
Word Pairs	1.122
Sentence Average	1.233
Sentence Pairs	0.963
Sentence Pairs (Bio)	1.035

Table 4.6: Comparison of Standardized RMSE Values for Each Metric

similarity between generated and reference medical reports than any of the past metrics, because the Sentence Pairs metric has the strongest correlation. On the other hand, the Sentence Average metric has a very low R-squared value, at 0.054, which shows that the correlation between the Sentence Average score and the manual score is very weak. This information is also shown in Figure 4.22.

Standardized R-squared Values for all 10 Metrics	
Model	**Standardized R-squared Value**
BLEU-1	0.096
BLEU-2	0.104
ROUGE-1	0.237
ROUGE-2	0.179
ROUGE-L	0.251
RaTE	0.145
Word Pairs	0.133
Sentence Average	0.054
Sentence Pairs	0.283
Sentence Pairs (Bio)	0.211

Table 4.7: Comparison of Standardized R-squared Values for Each Metric

4.2.4 Manual Score Table

Similar to the plot comparison and the RMSE comparison, we created a table of all 100 reports that we manually scored, along with the reasoning for each of the scores given to each of the reports. This table is shown in Table 8.1, which is in the appendix.

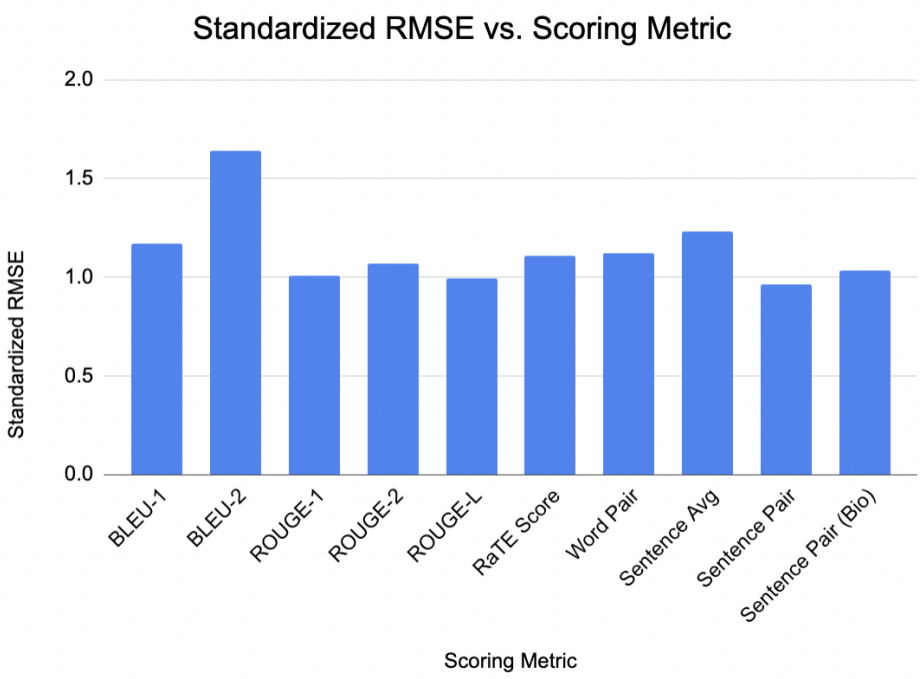

Figure 4.21: Standardized RMSE vs. Scoring Metric

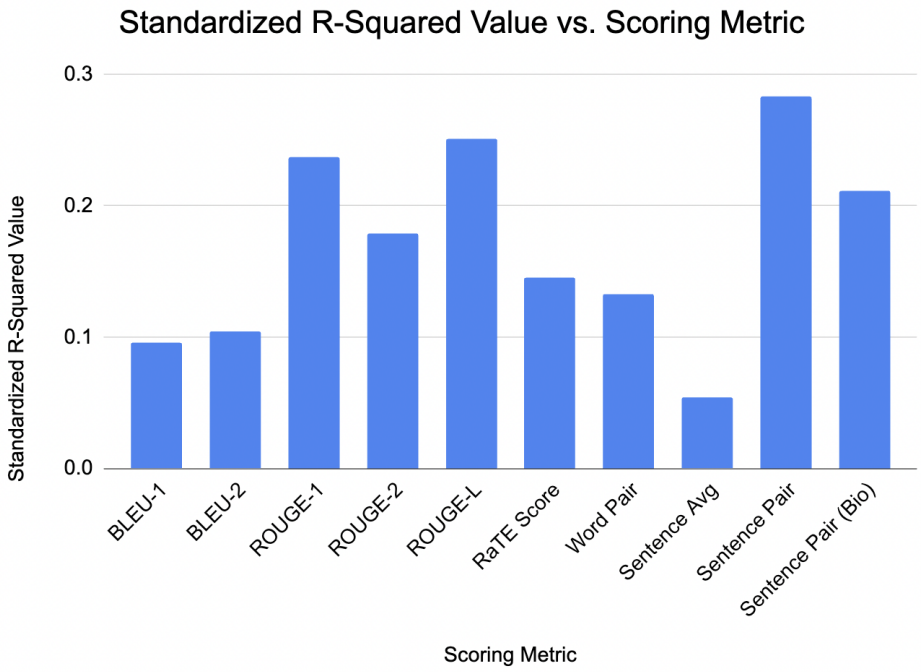

Figure 4.22: Standardized R-Squared Values vs. Scoring Metric

4.2.5 Qualitative Comparison

Another effective way to compare the 4 new metrics that we introduce in this research, as compared to the 6 prior metrics, is by looking at examples where the new metric was much higher than previous metrics. In order to do this, we picked a sample of 10 generated reports. For each of these reports, we show the value of the Sentence Pairs score, the manual score, and the other scores.

In Table 4.8, WP means Word Pairs SA means Sentence Average, SP means Sentence Pair, and SPB means Sentence Pairs (Bio).

Abbreviation for Each Metric	
Metric	**Abbreviation**
Manual Score	M
Word Pairs	WP
Sentence Average	SA
Sentence Pairs	SP
Sentence Pairs (Bio)	SPB

Table 4.8: Abbreviation for Each Metric

In order to make sure that the representative shows all possible accuracies, we picked one example from each of the different manual scores, from 1 to 10. Each generated report, the reference report, the manual score, and the scores from the 4 new metrics are shown in the table below.

As shown in Table 4.9, for the generated report that was manually rated as a 1 out of 10, the Word Pairs and Sentence Pairs values were the lowest, at 0.435, while the Sentence Average and Sentence Pairs (Bio) values were much higher. This makes sense, because, as mentioned before, the Sentence Average and Sentence Pairs (Bio) metrics have extremely high values.

For the generated report that was manually rated as a 10 out of 10, we can see that the two reports have the exact same meaning, but use slightly different words. For example, the generated report says that the osseous structures are unremarkable, while the reference report says that they are without acute abnormality. All of the four new metrics have values greater than 0.65, and we can see that the Sentence Pairs metric has a value of 0.76, which is relatively close to 1. Similarly, although the Sentence Average and Sentence Pairs (Bio) metrics are generally greater than 0.7, in this example, the Sentence Average score is 0.873 and the Sentence Pairs (Bio) score is 0.947. This shows that these scores also show a significant increase when the generated and reference reports have the same meaning.

Qualitative Example Table						
Generated	Reference	M	WP	SA	SP	SPB

The lungs are adequately inflated. No focal airspace opacity pleural effusion or pneumothorax. Unchanged small airways. Normal cardiomediastinal silhouette. Normal imaged portion of the upper abdomen. Degenerative changes are present at the spine.	There are prominent epicardial fat pads unchanged from prior. The cardiomediastinal silhouette and pulmonary vasculature are within normal limits. There is no pneumothorax or pleural effusion. There are no focal areas of consolidation. There is atherosclerosis of the aortic XXXX. Unchanged streaky opacities in the bilateral costophrenic sulci XXXX represent chronic scarring or atelectasis.	1	0.486	0.7	0.435	0.741
There is a right central line with the tip in the right atrium. There is a left central line with the tip in the superior vena cava. Heart size is within normal limits. There is bilateral hilar lymphadenopathy right greater than left consistent with history of sarcoidosis. There is asymmetric right lower lobe airspace disease. There is no pneumothorax or pleural effusion.	Right dual-lumen internal jugular central venous catheter seen with tip overlying the cavoatrial junction. Heart size at the upper limits of normal. Low lung volumes with bronchovascular crowding. Patchy bibasilar air airspace opacities right greater than left. No visualized pneumothorax. Prominence of the mediastinum consistent with history of sarcoid.	2	0.656	0.744	0.587	0.619

The lungs are clear. There is no pneumothorax or pleural effusion. There is no consolidation. There is mild cardiomegaly. Median sternotomy wires are present. There is a component of atherosclerosis of the aortic arch. There are degenerative changes of the thoracic spine.	There has been interval sternotomy with intact midline sternotomy XXXX. The heart is near top normal in size with unfolding of the aorta. The lungs are grossly clear with no focal airspace opacity pleural effusion or pneumothorax. The osseous structures are grossly normal.	3	0.404	0.789	0.575	0.899
The heart is normal in size. The right middle lobe airspace disease is improved. The lungs are clear. No pleural effusion or pneumothorax. The diaphragm mediastinum and hilar regions are unremarkable.	The cardiomediastinal silhouette is normal size and configuration. Pulmonary vasculature within normal limits. There is right middle lobe airspace disease may reflect atelectasis or pneumonia. No pleural effusion. No pneumothorax. Elevated right hemidiaphragm.	4	0.677	0.89	0.707	0.735
The lungs are hypoinflated. No focal airspace opacity pleural effusion or pneumothorax. Minimal left basilar scarring is again demonstrated. The cardiac silhouette is at the upper limit of normal for size. Unchanged hilar contours. Surgical clips project over the upper abdomen. Degenerative changes are present at the spine.	There is some minimal patchy opacity in left base which may represent atelectasis or scarring. The lungs are otherwise clear. The heart and mediastinum are normal for age. There is some arthritic changes of the skeletal structures and there has been previous rotator XXXX repair on the right.	5	0.443	0.711	0.553	0.789

The lungs are adequately inflated. No focal airspace opacity pleural effusion or pneumothorax. Normal cardiomediastinal silhouette. Normal imaged portion of the upper abdomen. Degenerative changes are present at the spine.	Cardiac and mediastinal contours are within normal limits. Atherosclerotic aorta. Mild blunting left costophrenic recess possibly mild atelectasis or scarring. No confluent lobar consolidation or large volume pleural effusion. Thoracic spondylosis.	6	0.438	0.778	0.503	0.91
The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	Normal heart size. Clear hyperaerated lungs. No pneumothorax. No pleural effusion. XXXX substernal density may be related to a pectus deformity.	7	0.434	0.839	0.647	0.885
Heart size is within normal limits. Lungs are without focal airspace consolidation. No evidence of pleural effusion or pneumothorax. Soft tissues and osseous structures are intact.	Heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen.	8	0.738	0.822	0.83	0.951
Lungs are clear without mass consolidation pleural effusion or pneumothorax. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Osseous structures are unremarkable.	Normal heart size and mediastinal contours. The lungs are clear. There is no pneumothorax or pleural effusion. No acute bony abnormalities.	9	0.573	0.728	0.591	0.881

The lungs are clear. No pneumothorax. No pleural effusion. No pulmonary edema. The cardiomediastinal silhouette is normal. The osseous structures are unremarkable.	The lungs are clear bilaterally. Specifically no evidence of focal consolidation pneumothorax or pleural effusion. Cardiomediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are without acute abnormality.	10	0.661	0.873	0.76	0.947

Table 4.9: Examples of Generated Reports and New Metric Values

Chapter 5

Discussion

5.1 Model Comparison

One key observation that we see after comparing the unimodal and multimodal performance is that the multimodal model far outperforms the unimodal model across all metrics. This makes sense, because the unimodal medical LLM doesn't have access to the chest X-ray images, which are the primary inputs needed to generate an accurate chest X-ray medical report.

Another interesting observation from the results is that the multimodal model with symptom data only performs slightly better than the multimodal model without symptom data. We found this surprising, because we originally hypothesized that adding symptom information would result in a significant benefit. According to our results, it seems like the only metric that showed a large benefit was the Word Pairs metric. This benefit in the Word Pairs metric is most likely because including symptom data as an input results in the model generating a medical report that has similar keywords to the symptoms, and the symptoms are likely included in the reference report.

One potential explanation for why the multimodal model with symptom data only performs slightly better than the multimodal model without symptom data is that the average quality of the symptom data might be inaccurate. For example, if the chest X-ray indicates that there are problems, yet the patient for the corresponding chest X-ray mentions that they don't have any symptoms because they are unaware of their problems, the multimodal model with symptom information could be less likely to generate a medical report that focuses on the symptom information. Similarly, if the symptoms mention several problems that aren't found in the chest X-ray, the multimodal model with symptom information could be more likely to mention those problems as keywords in the medical report, even if the given patient doesn't actually have those problems.

5.2 Evaluation Metric Comparison

One thing that we found very surprising was that the radiology-based text-encoder Sentence Pairs method we created, titled "Sentence Pairs (Bio)", had such inflated scores. The core motivation behind creating the "Sentence Pairs (Bio)" metric in the first place was to design a system

that was specifically designed for the radiology embedding case, but the results show that the Sentence Pairs (Bio) metric has worse performance than the regular Sentence Pairs metric.

We think that one potential reason for this is that two sentences in the radiology embedding space are more likely to be similar to each other, which results in the radiology embedding score inflating the final similarity between any two given sentences. In the future, we can try other radiology sentence embedding models to see if these new sentence embedding approaches can make the Sentence Pairs method more accurate.

Based on the evaluation metrics, we can see that the Sentence Pairs metric performs the best, with the lowest standardized RMSE value, at 0.963. This makes sense, because the Sentence Pairs metric combines the best aspects of the Word Pairs metric and the Sentence Average metric. Similar to the Word Pairs metric, the Sentence Pairs metric compares individual sentences to find the pair of sentences that are as similar as possible, instead of taking an average, which can inflate similarity scores. Similar to the Sentence Average metric, the Sentence Pairs metric uses sentence embeddings, which means that comparisons also include the relationship between words in a given sentence. This is especially helpful for cases where there is ”not” followed by a keyword, since using sentence embeddings will be able to effectively differentiate between two sentences, where one has the word ”not”, and the other one doesn't.

Another important result is that the standardized RMSE values of the other metrics aside from Sentence Pairs, do not outperform some of the prior metrics. In particular, the Sentence Average metric has the highest standardized RMSE amongst the new metrics that we introduce, with a standardized RMSE of 1.233. It makes sense that the Sentence Average metric does not effectively measure how similar the generated and reference medical reports are to each other, because two reports in the radiology report domain are always going to be very similar to each other. In other words, averaging sentences might be useful when comparing topic similarity between two different reports, but since all reports that we compare are chest X-ray medical reports, the Sentence Average metric does not serve as an effective method for measuring the similarity between the generated and reference medical reports.

5.3 LLM as a Judge

One interesting approach for comparing generated and reference text is the ”LLM as a Judge” approach, as shown in Figure 5.1. In this method, an LLM is given both the generated text and the ground truth, then is asked to measure how similar the generated text and ground truth text are. This could also involve giving the LLM some structure, like asking the LLM to follow a certain structure similar to a rubric that humans would use to measure how similar the generated and reference medical reports are. Our approaches are quite different from this, because we only use the existing data from the generated and reference text, instead of an external model that determines how similar the text is. However, this is an interesting area of future work.

Some researchers, like Zheng et al, looked at evaluating chat bot assistants using other LLMs, wher these Judge LLMs are able to evaluate the model on more open-ended questions [26]. This same process can be applied in this context, where an LLM is able to use text-based reasoning to identify how similar the generated and reference medical reports are.

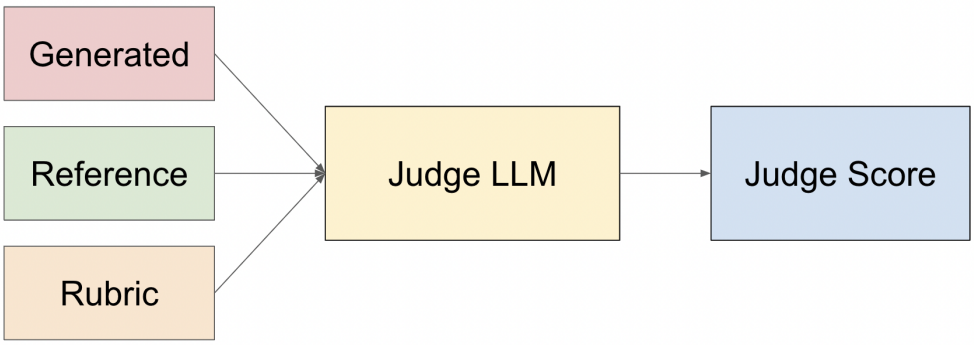

Figure 5.1: Using an LLM to Compare Generated and Reference Medical Reports

5.4 Impact

As mentioned earlier, there are two key research questions that we address with this research.

First, we compare several different types of medical report generation techniques. As mentioned earlier, we split the types of medical report generation techniques into 3 main types, including the unimodal text-based fine-tuned medical LLM, the multimodal MAIRA-2 model without the indication, and the multimodal MAIRA-2 model with the indication.

After comparing these 3 models using 10 different metrics on 500 samples, we found that both multimodal models perform better than the unimodal fine-tuned medical LLM, and the multimodal model with indication information performs slightly better than the multimodal model without indication information.

Second, we introduced 4 new metrics for evaluating how similar generated and reference medical reports are, including Word Pairs, Sentence Average, Sentence Pairs, and Sentence Pairs (Bio). In order to measure how effective these metrics are, we took 100 medical reports from the 500 samples, then measured the R-squared and RMSE between each of the metrics and the manual score, with the end goal of measuring how similar these metrics were to manual scores. Based off of this analysis, we found that the Sentence Pairs metric performs better than every metric in the prior work, across both the standardized R-squared and standardized RMSE values.

There are several key applications that this research has. To begin with, our answer to the first research question shows that multimodal models perform better than unimodal models, but those with symptom information don't perform significantly better than those without symptom information. In order to address this, future researchers can try improving the quality of the symptom information to see if there is further improvement in the model's accuracy with the symptom information.

The new metrics that we introduce also have several key impacts. For example, these metrics can be used in reinforcement learning-based methods, where having an accurate reward function is extremely important. These metrics can also be helpful as a method for determining how accurate future models are for medical report generation.

Chapter 6

Limitations

6.1 Models

One key limitation of this research is the number of models that we considered, along with types of different models. Specifically, we could consider more models than just a medical LLM and multimodal model with and without symptom information. As mentioned earlier in the paper, there are several models aside from just these two, including retrieval-based approaches, graph neural networks, and reinforcement-learning based approaches. In the future, we could compare the performance of all of these other approaches along with the medical LLM and MAIRA-2 model, to better identify which type of multimodal model has the highest performance for medical report generation.

6.2 Dataset

One major limitation in this thesis is the quality of the IU Chest X-Ray dataset that we used. Since the IU Chest X-Ray dataset is publicly available, the dataset creators chose to remove certain personal information from the dataset, including exact ages and other relevant information. Thus, some of the symptom and indication information sections in the dataset are not very useful, and potentially serve as extra noise in all of the models that we tested. In addition, it's possible that the reference medical reports from the dataset are not as long as full medical reports that doctors could write.

Another limitation of this research is the number of data points that we used, at 500 samples. In order to address this, future work could consider using a larger dataset, like MIMIC-IV and MIMIC-CXR. These datasets have around 200,000 images, as opposed to the IU-XRay dataset. In the context of the IU-XRay dataset, we only focused on a subset of the IU-XRay dataset, but we could have used the entire dataset if we wanted to further increase the dataset size.

6.3 Evaluation Metrics

There are several limitations for each of the evaluation metrics that we introduce in this research paper. To begin with, all of our evaluation metrics include some dependency on another metric. For example, the Word Pairs metric is based on word vector embeddings from Word2Vec, the Sentence Pairs metric is based on sentence embeddings from a sentence transformer, and the Sentence Pairs (Bio) metric is based on embeddings from CXR-BERT. These dependencies mean that the metrics can potentially be limited by the performance of the word vector embeddings and the sentence embeddings. In other words, if the word embeddings or sentence embeddings are inaccurate, it is likely that our methods are also inaccurate, since they are based on these existing approaches.

6.3.1 Word Pairs

The main limitation of the Word Pairs metric is that it just checks for keyword overlap, which means that the metric is not robust to cases where the metric is comparing two medical reports, where one has the phrase ”no” + keyword, and the other just has the keyword. In other words, since the Word Pairs metric just looks for keyword overlap and pre-processes out other words, the metric can measure two medical reports with exact opposite meanings as being the same.

6.3.2 Sentence Average

The main limitation of the Sentence Average metric is that it has a very inflated score. This makes sense, because the Sentence Average score takes the average of all sentence embeddings, which is likely to be similar to the average of all sentence embeddings for another medical report since the two reports are in the medical domain. However, this limits how useful this metric is. One area of future work to address this problem would be to scale the metric value down, then see how that impacts how effective the metric is.

6.3.3 Sentence Pairs

Although Sentence Pairs is a major improvement on the Word Pairs and Sentence Average methods, the method still considers the generated and reference reports to both be a bag of sentences. In other words, the model doesn't take into account the relationship between sentences, and instead measures how similarity pairs of sentences are. In addition, it's possible that one word in the reference report is similar to multiple words in the predicted report, in which case the Sentence Pairs metric value will get inflated. In order to address this, future work can focus on adding a penalty so that the generated report and the reference report don't keep using the same reference sentence in finding the best sentence from the predicted report.

6.3.4 Sentence Pairs (Bio)

Beyond the limitations mentioned for the Sentence Pairs metric, the Sentence Pairs (Bio) metric is also limited by the quality of CXR-BERT encoding model. If the two sentences in the two

medical reports are always encoded to be extremely similar to each other, then it is very likely that the CXR-BERT encoding model is not differentiating accurately between two generated medical reports. In order to address this future work can explore different sentence transformers from the medical domain that do a better job of differentiating two sentences in the medical domain.

6.4 GPU Resources

One of the most major limitations across this research thesis was lack of more powerful GPU resources. Due to resource limitations, we chose to run all code for this project on Google Colab with one A100 GPU instance. The A100 GPU instance has a limit of 40 GB of GPU RAM on Colab, which was just barely enough to run the MAIRA-2 model on 500 samples. In the future, we could consider using multiple GPUs or increasing the GPU RAM for the current GPU, with the end goal of being able to train the model on more data.

Chapter 7

Conclusion

7.1 Model Comparison

One important conclusion is that we can see that multimodal models like MAIRA-2 are significantly more accurate than uni-modal models. Furthermore, we can see that giving information on symptoms helps give a small increase in the accuracy of the model. By systematically comparing the accuracy of these models on the IU-XRay dataset, we can see that these results hold across 500 samples.

In order to further reinforce these conclusions, we can run the dataset on more examples from the IU-XRay dataset or from larger datasets, like MIMIC-CXR, which contains over 370,000 chest X-ray images [8]. This is much larger than the IU Chest X-Ray dataset that we used in this research, which only contains around 7,400 chest X-ray images [13].

7.2 Evaluation Metric Comparison

Second, we can see that all of our evaluation metrics are more effective than past evaluation metrics. In this paper, we randomly sampled 100 samples from the total amount of 500 samples, then graded each one of these 100 samples on a scale of 0 to 10, converted the metrics to a score from 0 to 1, and compared these human-graded scores to the generated metrics across all 10 metrics. Based on both the standardized R-squared score and the standardized RMSE, the Sentence Pairs method performs better than the past metrics. The R-squared score measures the association between the generated score values and the manual score, while the RMSE measures how far the generated score is from the manual score across all 100 samples. From both of these metrics, we can see that the Sentence Pairs method performs the best, which shows us that this is the best evaluation metric.

One way to further validate that our metrics are more effective than past metrics is to run a user study with doctors, instead of using the manual score. Since doctors actively write medical reports, they are more likely to be able to accurately measure how similar a generated and reference medical report are. Due to time limitations, we manually graded 100 reports, but future work could include asking doctors to grade a series of reports, in order to get a more accurate ground truth metric for how similar a generated medical report is to a reference medical report

that a doctor would write.

7.3 Evaluation Metric Applications

The new evaluation metric that we discuss also has a wide variety of applications. Future researchers can use it as a method for measuring how similar generated and reference medical reports are for their own medical report generation approaches. Reinforcement-learning based approaches can also use this metric has an effective way to reward models for generating higher quality medical reports, especially for cases where the generated report uses different medical terms or describes the given patient's condition using different words.

Chapter 8

Future Work

8.1 Adversarial Inputs

One interesting avenue of future work is testing adversarial inputs to multimodal models. For example, let's consider the MAIRA-2 model with indication information, where the two different input modalities are a given image of a chest x-ray and some text-based symptom information. We could try changing the image or the indication slightly, with the end goal of identifying whether the model is robust to changes in the symptom information. This could look like adding noise to one of the input images, or changing the symptom information to mention problems when there aren't any problems, then seeing the extent to which the generated medical report changes as a result.

This area is extremely relevant and important, because it's important to measure how robust the models we develop are to attacks that alter data. It's also interesting to see how confident the model is about data, even when it's irregular. These insights can help us build more robust multimodal medical report generation models.

8.2 Medical Context

Another interesting area of future work is comparing the accuracy of models trained without a medical context and those trained with a medical context. For example, in the unimodal example of the LLM fine-tuned on medical data, we could compare the fine-tuned medical LLM's performance on medical report generation with a base LLM's performance on medical report generation.

The impact of this research would be to show the extent to which domain-specific knowledge helps both unimodal and multimodal models make accurate predictions and generate accurate medical reports.

8.3 Trusting Inputs

Another interesting question to consider would be the extent to which the model trusts the image compared to the symptom. For example, if a normal chest X-ray is also given an indication that says that there are negative symptoms, or if a negative chest X-ray is given an indication that the patient is normal, we could look at the predicted medical report to determine how much the model weights the image compared to the indication.

8.4 LLM as a Judge

As mentioned in the discussion section, one interesting approach for future work in developing better evaluation metrics would be the LLM-as-a-judge approach, where we ask an LLM to measure how similar a generated and reference report are. In the future, we could use a similar rubric to the one that we followed manually, but instead of manually judging the similarity between the reports, we could supply it as input to an LLM, which can then judge the reports.

8.5 Ground Truth Labels

In this thesis paper, one assumption that we made is that the best ground truth label is simply the reference report, but it's also possible that predicting keywords or predicting key problems is a better method for comparing predictions made by a multimodal model. We could compare what happens when we assign each type of medical report to a category, and identify the extent to which comparing categories is more or less effective than comparing generated medical reports.

8.6 Datasets

Lastly, with more compute power, we could both expand the dataset and expand the number of models that we compare. First, we could expand the dataset to include MIMIC-IV and MIMIC-CXR, which consist of much more samples [8, 9]. MIMIC-IV consists of data for 364,000 individuals, while MIMIC-CXR consists of 377,100 images of relevant Chest X-rays. When compared to the IU Chest X-Ray dataset, both of these datasets are much larger, which means that they could potentially result in higher quality training data for a larger model.

8.7 Models

In terms of models, we could try additional multimodal models. Although MAIRA-2 is a high-performing model, there are several other multimodal models for medical report generation. For example, we could look at retrieval-based methods, like the one that Endo et al. explored [6]. As an alternative, we could look into

Bibliography

[1] Bio-medical: A high-performance biomedical language model. https://huggingface.co/ContactDoctor/ Bio-Medical-Llama-3-8B, 2024. 3.1.2.1

[2] Zaheer Babar, Twan van Laarhoven, and Elena Marchiori. Encoder-decoder models for chest x-ray report generation perform no better than unconditioned baselines. *Plos one*, 16 (11):e0259639, 2021. 2.1.1.2

[3] Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*, 2024. 2.1.2, 3.1.2.2

[4] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer, 2022. 3.2.1.4

[5] Kenneth Ward Church. Word2vec. *Natural Language Engineering*, 23(1):155–162, 2017. 3.2.1.1

[6] Mark Endo, Rayan Krishnan, Viswesh Krishna, Andrew Y Ng, and Pranav Rajpurkar. Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Machine Learning for Health*, pages 209–219. PMLR, 2021. 2.1.3, 8.7

[7] Daibing Hou, Zijian Zhao, Yuying Liu, Faliang Chang, and Sanyuan Hu. Automatic report generation for chest x-ray images via adversarial reinforcement learning. *IEEE Access*, 9: 21236–21250, 2021. 2.1.3

[8] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1): 317, 2019. 7.1, 8.6

[9] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023. 8.6

[10] HyoJe Jung, Yunha Kim, Heejung Choi, Hyeram Seo, Minkyoung Kim, JiYe Han, Gaeun Kee, Seohyun Park, Soyoung Ko, Byeolhee Kim, et al. Enhancing clinical efficiency through llm: Discharge note generation for cardiac patients. *arXiv preprint*

arXiv:2404.05144, 2024. 2.1.1.1

[11] Mingjie Li, Rui Liu, Fuyu Wang, Xiaojun Chang, and Xiaodan Liang. Auxiliary signal-guided knowledge encoder-decoder for medical report generation. *World Wide Web*, 26(1): 253–270, 2023. 2.1.1.2

[12] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004. 2.2.3

[13] Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. Clinically accurate chest x-ray report generation. In *Machine Learning for Healthcare Conference*, pages 249–269. PMLR, 2019. 3.1.1, 7.1

[14] Harpal Nandhra, Graham Murray, Nigel Hymas, and Neil Hunt. Medical records: doctors' and patients' experiences of copying letters to patients. *Psychiatric bulletin*, 28(2):40–42, 2004. 1.1

[15] Mohammed Yasser Ouis and Moulay Akhloufi. Deep learning for report generation on chest x-ray images. *Computerized Medical Imaging and Graphics*, page 102320, 2023. 2.2.1

[16] Matt Post. A call for clarity in reporting bleu scores. *arXiv preprint arXiv:1804.08771*, 2018. 2.2.2

[17] Mehreen Sirshar, Muhammad Faheem Khalil Paracha, Muhammad Usman Akram, Norah Saleh Alghamdi, Syeda Zainab Yousuf Zaidi, and Tatheer Fatima. Attention based automated radiology report generation using cnn and lstm. *Plos one*, 17(1):e0262209, 2022. 2.1.1.2

[18] Omkar Thawkar, Abdelrahman Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Shahbaz Khan. Xraygpt: Chest radiographs summarization using medical vision-language models. *arXiv preprint arXiv:2306.07971*, 2023. 2.1.2

[19] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2.1.1.1

[20] Xing Wu, Jingwen Li, Jianjia Wang, and Quan Qian. Multimodal contrastive learning for radiology report generation. *Journal of Ambient Intelligence and Humanized Computing*, 14(8):11185–11194, 2023. 2.1.2

[21] Yuan Xue, Tao Xu, L Rodney Long, Zhiyun Xue, Sameer Antani, George R Thoma, and Xiaolei Huang. Multimodal recurrent model with attention for automated radiology report generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I*, pages 457–466. Springer, 2018. 1.1

[22] Dong Yuan, Eti Rastogi, Gautam Naik, Sree Prasanna Rajagopal, Sagar Goyal, Fen Zhao, Bharath Chintagunta, and Jeff Ward. A continued pretrained llm approach for automatic

medical note generation. *arXiv preprint arXiv:2403.09057*, 2024. 2.1.1.1

[23] Biao Zhang, Zhongtao Liu, Colin Cherry, and Orhan Firat. When scaling meets llm finetuning: The effect of data, model and finetuning method. *arXiv preprint arXiv:2402.17193*, 2024. 2.1.1.1

[24] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023. 2.1.1.1

[25] Weike Zhao, Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Ratescore: A metric for radiology report generation. *arXiv preprint arXiv:2406.16845*, 2024. 2.2.4

[26] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023. 5.3

Appendix

8.8 Manual Score Table

Manual Score Table			
Generated Report	Reference Report	Manual Score (0-10)	Justification
Heart size is within normal limits. There is mild tortuosity of the thoracic aorta. The lungs are without focal airspace consolidation. There is no evidence of pleural effusion or pneumothorax. Soft tissues and osseous structures are intact.	The cardiac silhouette and mediastinum size are within normal limits. There is no pulmonary edema. There is no focal consolidation. There are no XXXX of a pleural effusion. There is no evidence of pneumothorax.	8	mentions everything except the mild tortuosity of the thoracic aorta
No acute pulmonary findings. Sternal wires and surgical clips are present. The lungs are clear. No pleural effusion or pneumothorax is identified. The heart size is normal. The mediastinal and hilar contours are normal.	Borderline cardiomegaly. Midline sternotomy XXXX. Enlarged pulmonary arteries. Clear lungs. Inferior XXXX XXXX XXXX.	4	does not mention the cardiomegaly or enlarged pulmonary arteries
The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	No focal areas of consolidation. No suspicious pulmonary opacities. Heart size within normal limits. No pleural effusions. No evidence of pneumothorax. Osseous structures intact.	8	does not mention the osseous structures being intact

PA and lateral views of the chest submitted. Mediastinum: Cardiac silhouette size is normal. No mediastinal or hilar mass or lymphadenopathy. Right lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Left lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Bones and soft tissues: No acute abnormality demonstrated.	Heart size is normal. There is tortuosity of the thoracic aorta stable compared with prior. No focal airspace disease or effusion. No pleural effusions or pneumothoraces. Degenerative changes in the thoracic spine.	6	does not mention the tortuosity of the thoracic aorta or the thoracic spine
The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	The cardiac contours are normal. The lungs are clear. Thoracic spondylosis.	6	does not mention thoracic spondylosis
The lungs are clear. No pleural effusion or pneumothorax is identified. The heart and mediastinal silhouette are normal. The osseous structures are unremarkable.	The heart is normal in size and contour. The lungs are clear without evidence of infiltrate. There is no pneumothorax or effusion.	9	mentions the exact same things
The cardiac silhouette is normal in size. The left lung appears clear of any focal area of consolidation. There is a faint 8 millimeter nodular opacity at the left lung base. Calcified granuloma is identified at the left lung base. No evidence of pleural effusion. No evidence of pneumothorax. There is a calcified granuloma at the right upper lung. The osseous structures appear intact.	The cardiac and mediastinal contours are within normal limits. The lungs are well-inflated and clear. There is an 8mm nodule in the left lower lobe XXXX calcified granuloma. There is no pneumothorax or effusion. Bony structures of the thorax are intact with minimal early degenerative change.	7	mentions calcified granuloma but not the thorax

The lungs are clear. The heart is normal in size. The aorta is mildly ectatic. There is no edema or effusion or congestion. No significant bony abnormality is demonstrated. There is evidence of a prior cervical spine surgery.	The heart is normal in size. The mediastinum is unremarkable. The lungs are clear. Mild scoliosis and degenerative changes of the thoracic spine noted.	5	does not mention scoliosis or degenerative changes of the thoracic spine
The cardiac silhouette is normal in size. The lungs are clear of infiltrates edema or effusions. No lung masses or nodules are seen. The bony structures are unremarkable.	Chest. Both lungs are clear and expanded with no pleural air collections or parenchymal consolidations. Heart and mediastinum remain normal. Lumbosacral spine. XXXX disc spaces and alignment are normal. Sacrum and sacroiliac joints are normal.	7	”mention that everything is fine
Lungs are clear without mass consolidation pleural effusion or pneumothorax. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Osseous structures are unremarkable.	The cardiomediastinal silhouette and vasculature are within normal limits for size and contour. The lungs are normally inflated and clear. Osseous structures are within normal limits for patient age.	9	mean the exact same thing
The lungs are clear. There is no pneumothorax or pleural effusion. The cardiac silhouette is unremarkable. The mediastinum is unremarkable. There are no acute osseous abnormalities.	Lungs are clear without focal consolidation effusion or pneumothorax. Normal heart size. Negative for pneumoperitoneum. Mild degenerative changes of the thoracic spine.	6	do not mention the degenerative changes in the thoracic spine
The cardiomediastinal silhouette is normal. No focal consolidations pleural effusions or pneumothorax. Osseous structures demonstrate no acute abnormality. Bilateral hyperexpansion and interstitial prominence.	There is a single calcified granuloma in the right lung base. The lungs are otherwise grossly clear bilaterally. There is no pneumothorax or pleural effusion. Cardiac and mediastinal silhouettes are normal. There are cholecystectomy clips in the right upper quadrant of the abdomen. Small T-spine osteophytes are noted.	6	does not mention the t-spine osteophytes

The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	Normal heart size. Clear hyperaerated lungs. No pneumothorax. No pleural effusion. XXXX substernal density may be related to a pectus deformity.	7	mention everything esxcept the substernal density
The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	The lungs are clear bilaterally. Specifically no evidence of focal consolidation pneumothorax or pleural effusion.. Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are without acute abnormality.	9	mean the same thing
The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	Heart size and vascularity normal. Lungs are clear. No effusions. No pneumothorax. Visualized osseous structures unremarkable.	10	mean the same thing
The lungs are clear. No pneumothorax or pleural effusion. Unremarkable cardiomediastinal silhouette.	There is a calcified granuloma in the right midlung zone. Lungs are otherwise clear. There is no pleural effusion or pneumothorax. The heart and mediastinum are normal. The skeletal structures are normal. Surgical clips are present in the right upper quadrant.	3	does not mention the calcified granuloma or surgical clips

PA and lateral views of the chest submitted. Mediastinum: Cardiac silhouette size is normal. No mediastinal or hilar mass or lymphadenopathy. Right lung: Nodular opacity projects over the right upper lobe and this is an artifact as it does not project over the lateral projection. No airspace consolidative process. No pleural effusion or pneumothorax. Left lung: Nodular opacity projects over the left upper lobe and this is an artifact as it does not project over the lateral projection. No airspace consolidative process. No pleural effusion or pneumothorax. Bones and soft tissues: No acute abnormality demonstrated.	There are no airspace opacities to suggest pneumonia. There is a vague nodular like opacity in the right midlung measuring 1.2 cm projecting through the posterior 7th and 8th ribs. This may be artifact. Chest fluoroscopy would confirm this. Heart and pulmonary XXXX appear normal. There are calcified subcarinal and right hilar lymph XXXX. The pleural spaces are clear.	4	mentions there being something int he left and right lungs
Lungs are clear without mass consolidation pleural effusion or pneumothorax. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Osseous structures are unremarkable.	The Cardiopulmonary silhouette is normal. The Heart size is normal. The lungs are clear with no pulmonary effusions or pneumothorax.	9	mean the same thing
Cardiomediastinal silhouette is normal. Pulmonary vascularity is normal. Lungs are clear without evidence for infiltrate. No pleural effusions are seen. No evidence of a pneumothorax.	Low lung volumes. Heart size and mediastinal contour within normal limits. No focal air space consolidation pneumothorax or pleural effusion. Mild thoracic spine degenerative change.	3	does not mention the low lung volumes or the thoracic spine degenerative change
Heart size is within normal limits. Lungs are without focal airspace consolidation. No evidence of pleural effusion or pneumothorax. Soft tissues and osseous structures are intact.	Heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen.	8	”mean the same thing

Lungs are clear without mass consolidation pleural effusion or pneumothorax. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Osseous structures are unremarkable.	Normal heart size and mediastinal contours. The lungs are clear. There is no pneumothorax or pleural effusion. No acute bony abnormalities.	9	mean the same thing
Lungs are clear without mass consolidation pleural effusion or pneumothorax. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Osseous structures are unremarkable.	The heart size is normal. The mediastinal contour is within normal limits. The lungs are free of any focal infiltrates. There are no nodules or masses. No visible pneumothorax. No visible pleural fluid. The XXXX are grossly normal. There is no visible free intraperitoneal air under the diaphragm.	9	mean the same thing
The lungs are hyperinflated. No focal airspace opacity pleural effusion or pneumothorax. No pulmonary nodules are identified. Normal cardiomediastinal silhouette. Normal imaged portion of the upper abdomen. Degenerative changes are present at the spine.	Chest: Stable cardiomediastinal silhouette. Pulmonary vascularity is within normal limits. Hyperlucent apices. Negative for focal airspace disease or consolidation. Negative for pneumothorax or pleural effusion. Healed remote left 9th rib fracture. Right shoulder: Negative for fracture or dislocation.	7	”mention the same thing
The lungs are clear. There is no pleural effusion or pneumothorax. There is stable cardiomegaly and aortic calcifications. There is no pulmonary edema. Degenerative changes are seen in the thoracic spine.	Heart size is unchanged. Aortic calcification is noted. No pneumothorax. No large pleural effusions. There are unchanged XXXX opacities throughout the lungs which XXXX represent scarring. Lungs are hyperexpanded.	4	”doesn’t mention that the lungs are hyperexpanded

The lungs are adequately inflated. No focal airspace opacity pleural effusion or pneumothorax. Normal cardiomediastinal silhouette. Normal imaged portion of the upper abdomen. Degenerative changes are present at the spine.	Cardiac and mediastinal contours are within normal limits. Atherosclerotic aorta. Mild blunting left costophrenic recess possibly mild atelectasis or scarring. No confluent lobar consolidation or large volume pleural effusion. Thoracic spondylosis.	6	does not mention the atherosclerotic aorta
PA and lateral views of the chest. No infiltrate effusion or pneumothorax identified. Cardiac and mediastinal contours are within normal limits. The soft tissues are intact.	The heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen.	9	mean the same thing
The lungs are adequately inflated. No focal airspace opacity pleural effusion or pneumothorax. Linear atelectasis or scar is seen near the right lung base. The cardiac silhouette is at the upper limit of normal for size. A prosthetic valve projects over the heart. Surgical clips project over the mediastinum. Normal imaged portion of the upper abdomen. Early degenerative changes are present at the spine.	Atrial septal occluder artifact. Rotated frontal position overall heart size within normal limits no typical findings of pulmonary edema. XXXX densities in the left base small focal XXXX opacity in the right base with focal posterior right hemidiaphragm elevation and obscured right costophrenic XXXX. Biapical pleuroparenchymal irregularities most compatible with scarring chronic appearing right 5th rib contour deformity. No pneumothorax seen.	3	”does not mention the biapical pleuroparenchymal irregularities
The heart is normal in size. No focal infiltrate is seen. There is no marked central vascular congestion. No pleural effusion or pneumothorax is seen. The bones are unremarkable for age.	Overall hyperexpanded lungs with flattening of the diaphragms consistent with obstructive lung disease. Lungs are clear without focal consolidation. No pleural effusions or pneumothoraces. Heart and mediastinum of normal size and contour. Degenerative changes in the spine.	3	does not mention obstructive lung disease

Right upper lobe consolidation consistent with pneumonia. No pleural effusion or pneumothorax. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Osseous structures are unremarkable.	There is a right upper lobe opacity. Cardiomediastinal silhouette is normal. Pulmonary vasculature and XXXX are normal. Osseous structures and soft tissues are normal.	8	mentions the right upper lobe opacity
The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	No pneumothorax pleural effusion or focal airspace disease. Heart size normal. Stable cardiomediastinal silhouette. Nodular opacities consistent with chronic granulomatous disease. Bony structures intact.	4	does not mention chronic granulomatous disease
The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	The heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen.	9	mean the same thing
The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	Heart size is normal. The lungs are grossly clear. No pleural effusions or pneumothoraces. The hilar and mediastinal contours are stable. Normal pulmonary vascularity. No overt edema.	7	”means the same thing
No indwelling catheters are seen. A central venous line is seen with the tip at the caval atrial region. The lungs are clear. The costophrenic angles are sharp. No pneumothorax. The cardiac silhouette is normal in size. The osseous structures are unremarkable.	Heart size within normal limits stable mediastinal and hilar contours right chest XXXX tip in the low SVC. Monitoring device artifacts. No focal alveolar consolidation no definite pleural effusion seen. No typical findings of pulmonary edema.	6	mentions the caval atrial region
No pleural effusion or pneumothorax. No focal consolidation. Cardiomediastinal silhouette and pulmonary vasculature are unremarkable. Osseous structures are unremarkable.	The lungs are clear. There is no pleural effusion or pneumothorax. The heart and mediastinum are normal. The skeletal structures are normal.	9	mean the same thing

PA and lateral views of the chest submitted. Mediastinum: Cardiac silhouette size is normal. No mediastinal or hilar mass or lymphadenopathy. Right lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Left lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Other: Calcified granuloma projects over left upper lobe. Bones and soft tissues: No acute abnormality demonstrated.	Heart size within normal limits. Mild hyperinflation of the lungs. Mild pectus excavatum deformity. Stable left mid lung calcified granuloma. No focal airspace disease. No pneumothorax or effusions.	5	mentions everything except the pectus excavatum deformity and the mild hyperinflation
The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	Normal cardiomediastinal contours. Clear lungs bilaterally. No pneumothorax or large effusion.	10	exact same meaning
The cardiomediastinal silhouette is normal. No focal consolidations pleural effusions or pneumothorax. Osseous structures demonstrate no acute abnormality.	The cardiomediastinal silhouette and pulmonary vasculature are within normal limits. There is no pneumothorax or pleural effusion. There are no focal areas of consolidation.	9	mentions everything except doesn't use the term pulmonary vasculature
The lungs are clear. No pneumothorax or pleural effusion. Unremarkable cardiomediastinal silhouette.	The heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen.	10	same meaning
Lungs are clear without mass consolidation pleural effusion or pneumothorax. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Osseous structures are unremarkable.	The heart size and mediastinal silhouette are within normal limits for contour. The lungs are clear. No pneumothorax or pleural effusions. The XXXX are intact.	8	"same meaning

The cardiac silhouette is mildly enlarged. No mediastinal or hilar mass or lymphadenopathy. No lung nodule in the right lung. No airspace consolidative process in the right lung. No pleural effusion or pneumothorax in the right lung. No lung nodule in the left lung. No airspace consolidative process in the left lung. No pleural effusion or pneumothorax in the left lung. No acute abnormality in bones and soft tissues.	The lungs and pleural spaces show no acute abnormality. Stable left upper lobe calcified granuloma. Heart size is mildly enlarged pulmonary vascularity within normal limits. Mild tortuosity of the descending thoracic aorta.	2	does not mention the calcified granuloma or the mild tortuosity of the thoracic aorta
The lungs are adequately inflated. No focal airspace opacity pleural effusion or pneumothorax. The cardiac silhouette is at the upper limit of normal for size. Atherosclerotic calcifications are present at the aorta. Normal imaged portion of the upper abdomen. Degenerative changes are present at the spine.	Heart size and mediastinal contour are normal. Pulmonary vascularity is normal. is not diffuse interstitial prominence which has chronic appearance. Cannot exclude early pulmonary edema. Two airspace consolidation or effusion. XXXX are osteopenic. No visible pneumothorax.	3	”does not mention pulmonary edema
The cardiac silhouette is normal in size. The lungs are clear of infiltrates edema or effusions. No lung masses or nodules are seen. The bony structures are unremarkable.	Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Lungs are clear. No pneumothorax or pleural effusion. No acute osseous findings. XXXX degenerative changes of the thoracic spine.	4	does not mention degenerative changes in the thoracic spine
The lungs are adequately inflated. No focal airspace opacity pleural effusion or pneumothorax. Normal cardiomediastinal silhouette. Normal imaged portion of the upper abdomen. No acute osseous findings.	The cardiomediastinal silhouette is within normal limits for appearance. The trachea is midline. No focal pulmonary consolidation. No pneumothorax. No pleural effusion. Minimal degenerative changes of the thoracic spine.	5	does not mention the thoracic spine

The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	Normal cardiomediastinal contours. No focal consolidation or pleural effusions. No pneumothorax.	9	mentions everything except the focal consolidation term
The lungs are adequately inflated. No focal airspace opacity pleural effusion or pneumothorax. No pulmonary nodules are identified. Normal cardiomediastinal silhouette. No evidence of lymphadenopathy. Normal imaged portion of the upper abdomen. Degenerative changes are present at the spine.	The lungs are clear. There is hyperinflation. Calcification is seen over the anterior mediastinum XXXX a calcified lymph node at is not identified on the PA projection. The heart is normal. Arthritic changes the spine are seen.	3	”does not mention calcification
The lungs are clear. The costophrenic angles are sharp. No evidence of pleural effusion. No pneumothorax. The cardiac silhouette is within normal limits. There is atherosclerotic calcification of the aortic arch. A percutaneous biliary catheter is seen in the right upper quadrant of the abdomen. Degenerative changes are present in the spine.	The heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen. No non-calcified nodules are identified.	2	”says that there is calcification and degenerative changes in the spine
The lungs are clear. No pneumothorax or pleural effusion. Unremarkable cardiomediastinal silhouette.	Heart size and pulmonary vascularity appear within normal limits. The lungs are free of focal airspace disease. No pleural effusion or pneumothorax is seen. Degenerative changes are present in the spine.	4	does not mention the degenerative changes present in the spine

The right-sided dialysis catheter tip projects over the right atrium. The left-sided dialysis catheter is unchanged in position with the distal tip projecting over the right atrium. Mild cardiomegaly is similar to prior study. There is mild pulmonary vascular congestion. Bibasilar airspace opacities and bilateral pleural effusions are similar to prior study. No pneumothorax. No acute bony abnormalities.	There has been interval placement of a dual-lumen dialysis catheter with the distal tip projected over the right atrium. Moderate cardiomegaly is identified. There is mild calcification of the transverse XXXX. XXXX airspace opacities are identified with bilateral pleural effusions.	9	mentions the cardiomegaly
The lungs are clear. There is no pneumothorax or pleural effusion. The cardiomediastinal silhouette is unremarkable. There are mild degenerative changes of the thoracic spine.	Normal heart size. No focal air space consolidation pneumothorax pleural effusion or pulmonary edema. Anterior osteophytes of the thoracic spine.	7	mentions problems with the thoracic spine
The lung fields are clear. The costophrenic angles are sharp. No pneumothorax. The cardiac silhouette is mildly enlarged. Mild biomechanical degenerative changes are seen of the thoracic spine.	Cardiomegaly is unchanged. Stable superior mediastinal contour with tortuous calcified aorta. Normal pulmonary vascularity. No focal air space consolidation pleural effusion or pneumothorax. No acute bony abnormality. Changes of prior right mastectomy.	4	mentions degenerative changes in the thoracic spine instead of the tortuous calcified aorta
The lungs are adequately inflated. No focal airspace opacity pleural effusion or pneumothorax. Normal cardiomediastinal silhouette. Atherosclerotic calcifications are present at the aortic arch. Normal imaged portion of the upper abdomen. No acute osseous findings.	Heart size mediastinal contour and pulmonary vascularity are within normal limits. No focal consolidation pleural effusion or pneumothorax is identified. No acute osseous abnormality identified.	6	mentions atherosclerotic calcifications when they are not in the reference report
Cardiac silhouette and mediastinal contours are normal. Lungs are clear. No pleural effusion. No osseous abnormality.	Lungs are clear. No pleural effusions or pneumothoraces. Heart and mediastinum of normal size and contour.	9	means the same thing

PA and lateral views of the chest submitted. Mediastinum: The cardiac silhouette is enlarged but unchanged. No mediastinal or hilar mass or lymphadenopathy. Right lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Left lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Other: Reticular opacity within the periphery of the lungs bilaterally is unchanged and may be due to interstitial lung disease. Bones and soft tissues: No acute abnormality demonstrated.	Redemonstration of interstitial opacities consistent with patient's history of pulmonary fibrosis. Unchanged calcified granulomas at the left greater than right hilum and in the pretracheal region. No pneumothorax pleural effusion or focal airspace consolidation. Cardiomediastinal size is the upper limits of normal. Pulmonary vasculature is normal . XXXX XXXX intact.	3	mentions interstitial lung disease instead of calcified granulomas and the cardiomediastinal size being on the larger side
The cardiomediastinal silhouette is normal in size and configuration. No infiltrates pleural effusions or vascular congestion. No acute bony abnormalities.	The heart and mediastinum are unremarkable. The lungs are clear without infiltrate. There is no effusion or pneumothorax. There is a mild levoscoliosis of the thoracic spine. There is mild widening of the right acromioclavicular joint which may be postsurgical or posttraumatic in XXXX.	3	"does not mention the levoscoliosis of the thoracic spine
The lungs are hypoinflated. No focal airspace opacity pleural effusion or pneumothorax. Minimal left basilar scarring is again demonstrated. The cardiac silhouette is at the upper limit of normal for size. Unchanged hilar contours. Surgical clips project over the upper abdomen. Degenerative changes are present at the spine.	There is some minimal patchy opacity in left base which may represent atelectasis or scarring. The lungs are otherwise clear. The heart and mediastinum are normal for age. There is some arthritic changes of the skeletal structures and there has been previous rotator XXXX repair on the right.	5	mentions scarring in the left base

The heart is normal in size. No focal infiltrate is seen. There is no marked central vascular congestion. No pleural effusion or pneumothorax is seen. The bones are unremarkable for age. Bilateral breast implants are seen.	The heart is normal in size. The mediastinum is unremarkable. The lungs are grossly clear. Bilateral breast prostheses are noted.	7	mentions the same things
The lungs are clear. No pleural effusion or pneumothorax is identified. The heart and mediastinal silhouette are normal. The osseous structures are unremarkable.	Cardiac and mediastinal contours are within normal limits. The lungs are clear. Bony structures are intact.	9	”mean the exact same thing
The feeding tube courses beneath the diaphragm with tip out of the field-of-view. The left subclavian line is unchanged in position. Bilateral pleural effusions are again noted right greater than left. Bibasilar airspace disease is again noted. Lucency is again noted in the left upper quadrant of the abdomen.	There is a left subphrenic crescentic lucency this is concerning for pneumoperitoneum. There are low lung volumes and bilateral moderate to large pleural effusions with bibasilar atelectasis/airspace disease that are larger in size in comparison to the prior exam. No pneumothorax. Heart size upper limits of normal. The left central venous catheter tip overlies the lower SVC. The feeding tube has been placed in the interval and extends below the diaphragm and below the XXXX-of-view.	7	mentions the left subphrenic crescentic lucency
Normal cardiac mediastinal contour. Lungs are clear. No consolidation or fluid. No bone lesion.	The cardiomediastinal silhouette is within normal limits for size and contour. The lungs are normally inflated without evidence of focal airspace disease pleural effusion or pneumothorax. Osseous structures are within normal limits for patient age..	9	”means the same thing

PA and lateral views of the chest submitted. Mediastinum: Cardiac silhouette size is normal. No mediastinal or hilar mass or lymphadenopathy. Right lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Left lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Bones and soft tissues: No acute abnormality demonstrated.	On the right there is marked narrowing of the hip joint space uniformly throughout. Osteophyte formation is present with some sclerosis and subchondral cyst formation vertically along the superior acetabulum and femoral head. I do not see evidence for fracture or destructive process. AP view of the femur shows no femoral XXXX destructive process or other significant abnormality. For of the Left hip shows near-complete obliteration of the joint space with severe subchondral sclerosis and cystic formation in both the superior acetabulum and superior aspect of the femoral head. No fracture or destructive process is identified. Surgical markers were XXXX in the images and left hip for the purpose of surgical planning. PA and lateral chest show the lungs to be clear. There may be some hyperinflation. No pleural effusion is identified. The heart is normal in size. There are calcified mediastinal lymph XXXX. The skeletal structures appear normal.	3	does not mention the calcified mediastinal lymph or hyperinflation
The lungs are clear. There is no pleural effusion. There is no pneumothorax. There is mild cardiomegaly. There is a tortuous aorta. There are degenerative changes of the thoracic spine.	Mild cardiomegaly. Mild unfolding of the thoracic aorta. No focal air space opacity. No pleural effusion or pneumothorax. Visualized osseous structures are unremarkable in appearance.	6	mentions the mild cardiomegaly

The cardiac silhouette is normal in size. There is a 4.8 cm opacity in the left lower lobe. This is favored to be pleural based. There is also a small opacity at the right cardiophrenic angle. This is favored to be a fat pad. There is no pleural effusion. There is no pneumothorax. The osseous structures are unremarkable.	There is a approximately 4 cm opacity with one XXXX margin and the other ill-defined in the lateral lower left lung is seen on the PA view. This is not definitely seen on the lateral view. There is no pneumothorax or pleural effusion. The cardiac silhouette is within normal limits. There are T-spine osteophytes. There is no pneumothorax or pleural effusion. There are calcified hilar lymph XXXX there	4	”does not mention the T-spine osteophytes
Lungs are clear without mass consolidation pleural effusion or pneumothorax. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Osseous structures are unremarkable.	The cardiac and mediastinal contours are within normal limits. The lungs are well-inflated and clear. There is no focal consolidation pneumothorax or effusion. The bony structures of the thorax are unremarkable.	9	”mean the same thing
Lungs are clear without mass consolidation pleural effusion or pneumothorax. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Osseous structures are unremarkable.	Cardiomediastinal silhouettes are within normal limits. Lungs are clear without focal consolidation pneumothorax or pleural effusion. Bony thorax is unremarkable.	9	”mean the same thing
The lungs are adequately inflated. No focal airspace opacity pleural effusion or pneumothorax. Normal cardiomediastinal silhouette. Atherosclerotic calcifications are present at the aortic arch. Surgical clips project over the left neck. Degenerative changes are present at the spine.	Stable appearing bilateral calcified lymph XXXX. The cardiac silhouette and mediastinal contours are within normal limits. No focal opacity. No large pleural effusion. There is no pneumothorax.	4	mentions degenerative changes at the spine

Cardiac silhouette is normal. Mild thoracic aortic tortuosity. No airspace consolidations or pleural effusions. No bony abnormality. Two-view left knee demonstrates mild degenerative disease medial compartment. No effusions.	Chest. Lungs are clear and expanded. Heart normal. Left knee. No change marked narrowing large osteophyte formation multiple synovial osteochondromas.	3	mentions thoracic aortic tortuosity
The cardiac silhouette is normal in size. The lungs are clear of infiltrates edema or effusions. No lung masses or nodules are seen. The bony structures are unremarkable.	Heart size normal. Lungs are clear. XXXX are normal. No pneumonia effusions edema pneumothorax adenopathy nodules or masses.	9	mentions no lung masses instead of specifically saying pneumonia
PA and lateral views of the chest submitted. Mediastinum: Cardiac silhouette size is normal. No mediastinal or hilar mass or lymphadenopathy. Right lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Left lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Bones and soft tissues: No acute abnormality demonstrated.	There are no focal areas of consolidation. No suspicious pulmonary opacities. Heart size within normal limits. No pleural effusions. There is no evidence of pneumothorax.	9	”mention the same thing

PA and lateral views of the chest submitted. Mediastinum: Cardiac silhouette size is normal. No mediastinal or hilar mass or lymphadenopathy. Right lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Left lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Bones and soft tissues: No acute abnormality demonstrated.	Chest. The trachea is midline. Negative for pneumothorax pleural effusion or focal airspace consolidation. The heart size is normal. Abdomen. No pneumoperitoneum. There is a normal bowel XXXX pattern. Air and stool visible throughout the entire large colon including the rectum. No abnormally dilated small bowel loops. No evidence for intussusception or small bowel obstruction. No pathologic calcifications XXXX over the abdomen or pelvis. XXXX XXXX are without fracture or destructive lesion though there are mild degenerative changes throughout the lumbar spine. Small hiatal hernia is not as well demonstrated on this exam.	4	does not mention the mild degenerative changes throughout the lumbar spine
The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	Lungs are clear bilaterally.There is no focal consolidation pleural effusion or pneumothoraces. Cardiomediastinal silhouette is within normal limits. XXXX are unremarkable.	9	”mean the same thing
No acute cardiopulmonary abnormality. The lungs are clear. No pneumothorax or pleural effusion. Normal-sized cardiac silhouette. There are degenerative changes of the thoracic spine.	No pneumothorax pleural effusion or airspace consolidation. Cardiomediastinal size is within normal limits. XXXX XXXX intact.	3	mentions degenerative changes in the thoracic spine when it is not in the reference report

The lungs are adequately inflated. No focal airspace opacity pleural effusion or pneumothorax. Normal cardiomediastinal silhouette. Normal imaged portion of the upper abdomen. Degenerative changes are present at the spine. The ribs are poorly penetrated on chest x-ray. No definite fracture of the imaged portion of the right ribs.	No focal consolidation no definite pleural effusion seen. Exaggerated kyphosis with increased AP dimension of the thorax curvilinear density projected over the right anterior 3rd and 4th ribs beyond which lung markings are seen XXXX skin fold artifact. Mild aortic ectasia/tortuosity no typical mediastinal widening to suggest vascular injury. Contour irregularity of the lateral right 9th rib of indeterminate age.	6	”mention the fracture at the right ribs
The lungs are clear. No pneumothorax. No pleural effusion. No pulmonary edema. The cardiomediastinal silhouette is normal. The osseous structures are unremarkable.	The lungs are clear bilaterally. Specifically no evidence of focal consolidation pneumothorax or pleural effusion. Cardiomediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are without acute abnormality.	10	”means the exact same thing
Normal cardiac mediastinal contour. Lungs are clear. No consolidation or fluid. No bone lesion.	No acute osseous abnormality. The soft tissues are within normal limits. Normal cardiomediastinal silhouette and hilar contours. No focal area of consolidation pleural effusion or pneumothorax.	9	”means the same thing
The cardiomediastinal silhouette is normal. No focal consolidations pleural effusions or pneumothorax. Calcified granulomas again present in the right lung. Bilateral hyperexpansion and interstitial prominence.	There is scattered calcified granulomas. The lungs are otherwise grossly clear. Cardiac and mediastinal silhouettes are normal. Pulmonary vasculature is normal. No pneumothorax or pleural effusion. No acute bony abnormality.	5	mentions the calcified granulomas

The lungs are clear. No pleural effusion or pneumothorax is identified. The heart and mediastinal silhouette are normal. The osseous structures are unremarkable.	2 images. Heart size and pulmonary vascular engorgement appear within limits of normal. Mediastinal contour is unremarkable. No focal consolidation pleural effusion or pneumothorax identified. No convincing acute bony findings.	9	”means the same thing
The cardiac silhouette is mildly enlarged. There is a hiatal hernia. No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. No acute abnormality in bones and soft tissues.	Lung volumes are XXXX. XXXX opacities are present in both lung bases. A hiatal hernia is present. Heart and pulmonary XXXX are normal.	4	mentions the hiatal hernia
The cardiomediastinal silhouette is normal. No focal consolidations pleural effusions or pneumothorax. Osseous structures demonstrate no acute abnormality.	Cardiomediastinal silhouette is within normal limits in overall size and appearance. Central vascular markings are symmetric and within normal limits. The lungs are normally inflated with no focal airspace disease pleural effusion or pneumothorax. No acute bony abnormality. Stable scarring in the right lung apex.	7	mentions everything except stable scarring in the right lung apex
PA and lateral views of the chest submitted. Mediastinum: Cardiac silhouette size is normal. No mediastinal or hilar mass or lymphadenopathy. Right lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Left lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Reticular opacity left upper lobe is unchanged and likely due to scarring. Bones and soft tissues: No acute abnormality demonstrated.	Stable appearance of the left upper lung lobe with scarring volume loss and pleural thickening. Cardiomediastinal silhouette is within normal limits normal appearance similar to prior. Volume loss in the left lung stable. Right lung is clear. There is no XXXX focal airspace disease pleural effusion or pneumothorax. Mild scarring at the right apex. No acute bony abnormality.	6	mentions that the reticular opacity in the left upper lobe is unchanged

The cardiac silhouette is normal in size. There is a nodular opacity in the periphery of the left upper lobe which may represent a parenchymal nodule or pleural-based lesion. There is an area of airspace disease in the right upper lobe which is not significantly changed from the reference exam. No pleural effusion or pneumothorax is seen. The left lung is otherwise clear. No acute osseous abnormality is seen.	The cardiomediastinal silhouette and vasculature are within normal limits for size and contour. There is right upper lobe airspace disease.. There is a rounded nodular opacity in the left upper lung measuring approximately 7 mm which may represent further sequela of infectious process versus other pathology. Osseous structures are within normal limits for patient age.	6	mentions the nodular opacity in the left upper lobe
The cardiomediastinal silhouette is normal in appearance. The central pulmonary hila and peripheral vascular markings are normal in appearance. The lungs are adequately expanded without evidence of focal opacification mass or nodule. The pleura chest wall and diaphragms are normal in appearance. There is no pleural effusion or pneumothorax. The osseous structures and soft tissues are unremarkable.	The lungs are clear bilaterally. Specifically no evidence of focal consolidation pneumothorax or pleural effusion.. Cardio mediastinal silhouette is unremarkable. Visualized osseous structures of the thorax are without acute abnormality.	8	”mean the same thing

PA and lateral views of the chest submitted. Mediastinum: Cardiac silhouette size is normal. No mediastinal or hilar mass or lymphadenopathy. Right lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Left lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Other: The lungs are hyperinflated compatible with chronic obstructive pulmonary disease. Bones and soft tissues: No acute abnormality demonstrated.	Heart size normal. Tortuous aorta. Calcified hilar lymph XXXX XXXX sequela of prior granulomatous disease. Hyperinflated lungs. The otherwise lungs are clear. The bilateral apices are partially excluded from the XXXX-of-view. There is the interval fixation of the right humeral fracture XXXX appears grossly intact. Osteopenia. Exaggerated kyphosis of the thoracic spine.	2	mentions that the lungs are hyperinflated
Clear lungs. No pleural effusion or pneumothorax. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Osseous structures are unremarkable.	Lungs are clear. No pleural effusions or pneumothoraces. Heart and mediastinum of normal size and contour.	10	”mean the exact same thing
The right pleural effusion and adjacent atelectasis are stable. There is also a small left pleural effusion. The lungs are otherwise clear. The heart and mediastinum are within normal limits. There are degenerative changes in the spine.	Exam limited by patient rotation. Mild rightward deviation of the trachea. Stable cardiomegaly. Unfolding of the thoracic aorta. Persistent right pleural effusion with adjacent atelectasis. Low lung volumes. No focal airspace consolidation. There is severe degenerative changes of the right shoulder.	2	mentions the right pleural effusion
The lungs are clear. No pleural effusion or pneumothorax is identified. The heart and mediastinal silhouette are normal. The osseous structures are unremarkable.	Frontal and lateral views of the chest show an unchanged cardiomediastinal silhouette. No XXXX focal airspace consolidation or pleural effusion.	8	”means the same thing

Lungs are clear without mass consolidation pleural effusion or pneumothorax. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Osseous structures are unremarkable.	There are no focal areas of consolidation. No suspicious pulmonary opacities. Heart size within normal limits. No pleural effusions. No evidence of pneumothorax. Osseous structures intact.	9	”mean the same thing
The lungs are adequately inflated. No focal airspace opacity pleural effusion or pneumothorax. Unchanged small airways. Normal cardiomediastinal silhouette. Normal imaged portion of the upper abdomen. Degenerative changes are present at the spine.	There are prominent epicardial fat pads unchanged from prior. The cardiomediastinal silhouette and pulmonary vasculature are within normal limits. There is no pneumothorax or pleural effusion. There are no focal areas of consolidation. There is atherosclerosis of the aortic XXXX. Unchanged streaky opacities in the bilateral costophrenic sulci XXXX represent chronic scarring or atelectasis.	1	mentions degenerative changes in the spine instead of atherosclerosis
The heart is normal in size. The right middle lobe airspace disease is improved. The lungs are clear. No pleural effusion or pneumothorax. The diaphragm mediastinum and hilar regions are unremarkable.	The cardiomediastinal silhouette is normal size and configuration. Pulmonary vasculature within normal limits. There is right middle lobe airspace disease may reflect atelectasis or pneumonia. No pleural effusion. No pneumothorax. Elevated right hemidiaphragm.	4	mentions the right lobe airspace disease

PA and lateral views of the chest submitted. Mediastinum: Cardiac silhouette size is normal. No mediastinal or hilar mass or lymphadenopathy. Right lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Left lung: No lung nodule. No airspace consolidative process. No pleural effusion or pneumothorax. Bones and soft tissues: No acute abnormality demonstrated.	The cardiomediastinal silhouette is within normal limits. The lungs are clear without areas of focal consolidation. There is a calcified granuloma within the left lung base. There is suggestion of a deep sulcus sign on the right. No definite pleural line of pneumothorax visualized. There is age-indeterminate wedging of several midthoracic vertebral bodies.	1	”does not mention the calcified granuloma
No acute cardiopulmonary disease. The lungs are clear. The costophrenic angles are sharp. No pneumothorax. The cardiac silhouette is normal. The osseous structures are unremarkable.	The heart pulmonary XXXX and mediastinum are within normal limits. There is no pleural effusion or pneumothorax. There is no focal air space opacity to suggest a pneumonia.	9	”means the same thing
There is a right central line with the tip in the right atrium. There is a left central line with the tip in the superior vena cava. Heart size is within normal limits. There is bilateral hilar lymphadenopathy right greater than left consistent with history of sarcoidosis. There is asymmetric right lower lobe airspace disease. There is no pneumothorax or pleural effusion.	Right dual-lumen internal jugular central venous catheter seen with tip overlying the cavoatrial junction. Heart size at the upper limits of normal. Low lung volumes with bronchovascular crowding. Patchy bibasilar air airspace opacities right greater than left. No visualized pneumothorax. Prominence of the mediastinum consistent with history of sarcoid.	2	mentions the bilateral hilar lymphadenopathy due to the history of sarcoidosis like the reference report
The cardiac silhouette is normal in size. The lungs are clear of infiltrates edema or effusions. No lung masses or nodules are seen. The bony structures are unremarkable.	The heart is normal in size. The mediastinum is unremarkable. Small nodular opacity left upper lobe may represent early infiltrate. The lungs are otherwise clear. There is no pleural effusion.	6	mentions everything except the small nodular opacity in the left upper lobe

The cardiac silhouette is normal in size. No focal infiltrate is seen. There is no marked central vascular congestion. No pleural effusion or pneumothorax is seen. The bones are unremarkable for age.	Lungs are clear. No pleural effusions or pneumothoraces. heart and mediastinum are stable with normal sized heart. Degenerative changes in the spine.	4	mentions everything except the degenerative changes in the spine
The lungs are clear. There is no pneumothorax or pleural effusion. There is no consolidation. There is mild cardiomegaly. Median sternotomy wires are present. There is a component of atherosclerosis of the aortic arch. There are degenerative changes of the thoracic spine.	There has been interval sternotomy with intact midline sternotomy XXXX. The heart is near top normal in size with unfolding of the aorta. The lungs are grossly clear with no focal airspace opacity pleural effusion or pneumothorax. The osseous structures are grossly normal.	3	mentions the sternotomy wires
The cardiac silhouette is normal in size. The lungs are clear of infiltrates edema or effusions. No lung masses or nodules are seen. The bony structures are unremarkable.	Both lungs are clear and expanded. Heart and mediastinum normal.	10	”means the same thing
The lungs are clear. No pneumothorax or effusion. Unremarkable cardiomediastinal silhouette.	The heart size is on the upper limits of normal. There is no mediastinal widening. The lungs are clear bilaterally. No large pleural effusion or pneumothorax. The XXXX are intact.	8	mentions everything except that the heart size is on the upper limits of normal

The cardiac silhouette is enlarged. There is aneurysmal dilatation of the aortic arch and descending thoracic aorta. The lungs are free of confluent infiltrates. No pleural effusions are present. The cardiac silhouette is enlarged. There is a large amount of calcification in the aortic arch.	Frontal and lateral views of the chest with overlying external cardiac monitor leads show an unchanged cardiomediastinal silhouette. Cardiac silhouette at the upper limits of normal in size. Tortuous ectatic aorta. The aortic XXXX is near 5 cm in diameter. There is a retrocardiac left paraspinal bulge concerning for a descending thoracic aortic aneurysm. There is biapical scarring. No XXXX focal airspace consolidation or pleural effusion. XXXX spine spondylitic changes.	3	mentions that the cardiomediastinal silhouette is enlarged
No acute findings in the lungs. No consolidation pleural effusion or pneumothorax. No acute findings in the heart. No acute findings in the mediastinum. No acute findings in the bones.	Heart size within normal limits stable mediastinal and hilar contours. No focal alveolar consolidation no definite pleural effusion seen. No typical findings of pulmonary edema. No pneumothorax. Mild dextrocurvature of the spine again noted.	7	mentions everything except the mild dextrocurvature of the spine
Lungs are clear without mass consolidation pleural effusion or pneumothorax. Cardiomediastinal silhouette and pulmonary vasculature are within normal limits. Osseous structures are unremarkable.	Lungs are clear. Heart size normal. The XXXX are unremarkable.	8	mentions the cardiomediastinal silhouette instead of mentioning that the heart size is normal

The cardiomediastinal silhouette and pulmonary vascularity are normal. Lung volumes are normal. No acute pulmonary infiltrate pulmonary edema pleural effusion or pneumothorax. Bony thorax appears intact. No radiopaque foreign body or focal air trapping. No free intraperitoneal air.	Normal heart size and mediastinal contours. The lungs are hyperinflated but clear. No pneumothorax or pleural effusion. No acute bony abnormalities.	9	”means the same thing

Table 8.1: Manually Scored Generated and Reference Medical Reports, with Justification