

Fine-tuning Large Language Model to Generate Impressions of CT-Scan Examination Report

By:

Rolamjaya Hotmartua

Thomas Sullivan

Yuancheng Ji

Zhixing Sun

Supervisor: Utku Pamuksuz, Benan Akca

A Capstone Project

Submitted to the University of Chicago in partial fulfillment
of the requirements for the degree of

Division of Physical Sciences

December 2023

Abstract

This paper leverages new advances in generative Artificial Intelligence to build a Large Language Model (LLM) specifically for generating the impression section of a CT-Scan Examination Report given the CT-Scan findings section and patient's clinical information. As building from scratch is costly, we propose fine-tuning a state-of-the-art pre-trained LLM with 246,824 CT-Scan examination reports from the University of Chicago Medicine. The research involved many exploration steps including prompt engineering, pre-trained model selection and hyper-parameter tuning. The radiology impressions generated are evaluated by quantitative and qualitative methods utilizing ROUGE score and other open-source LLMs.

Keywords: Large Language Model, Radiology Report, Instruction Tuning

Executive Summary

The goal of this paper is to ease the burden on radiology departments in the healthcare industry. Radiology is a vital component of healthcare yet generating a single report requires great focus and consumes much of a radiologist's time. An LLM can be used to write certain sections of a report, leading to quicker generation of the report. This can allow caring physicians to make quicker decisions regarding patient care and help ease the burden on radiologists.

With regards to generating the impression section of a CT-scan radiology report given the patient's clinical information and findings, our research found BioGPT as the highest performing model. BioGPT is an open-source LLM pre-trained with medical text allowing for greater understanding of radiology terminology compared to other general LLMs. Our research shows promising results, yet further work must be done for an LLM to be put in use within a radiology department. We recommend adding Reinforcement Learning with Human Feedback (RLHF) into our models as a potential next step to improve results.

Our project began by researching current state-of-the-art models in generative AI. This included general models, like Mistral and LLaMA, and healthcare specific models like BioGPT and MedAlpaca. We experimented with both types of models to build a LLM that can generate the impression section of a radiology report. Steps to build our model included cleaning and preprocessing our data, adding an instruction prompt, instruction tuning with Parameter-Efficient Fine-Tuning (PEFT) and hyperparameter tuning, and evaluating different models.

With encouraging results, we hope to continue improving the quality of generated impressions in our models. We also hope this research can shine a lot on the potential benefits of generative AI within the healthcare industry and inspire future research in that area. In addition to generating radiology reports, our techniques could be transferred to other tasks in the medical domain to help ease the burden on medical workers.

Table of Contents

Abstract	1
Executive Summary	2
Table of Contents	3
Introduction	1
Literature Review	4
Current advancements	4
Technique	5
Data	6
Methodology	9
Findings	14
Dataset	14
Experiment on Instruction Format	15
Experiment on Foundational Model	16
Evaluation on Indiana University Dataset	18
Evaluation for Factual Correctness	18
Discussion	21
Conclusion	24
References	25
Appendix A - Generation Sample for Each Fine Tuned Model	27

Introduction

Medical imaging and reporting are critical components of the healthcare industry. Both are used as a tool to diagnose patients and monitor various illnesses. Once a radiologist receives a medical image, they analyze the image and list all significant features in a single report. This report is sent back to the caring physician who will use the report to make critical decisions for patient care.

However, generating radiology reports demands a substantial amount of time from radiologists and requires extensive training and expertise. A typical radiology report consists of three sections: (1) Clinical information section which shows background information of the patient; (2) Findings section which lists radiologists' medical observations from the image, and (3) Impression section, which consists of the most prominent observations, conclusions or recommendations (see table 1.1). To generate a report, radiologists must carefully write an analysis of all observations from an image in the findings sections and succinctly summarize the most important observations in the impression section. Therefore, generating a single radiology report, regardless of modality, requires significant focus and care from the radiologist and consumes much of their time.

CLINICAL INFORMATION	male, 66 years old, status post subdural hemorrhage evacuation
FINDINGS	Findings are redemonstrated compatible with subdural hemorrhage evacuation. Two burr holes are present in the right parietal bone, the more posterior of which conveys a drainage catheter which enters the right-sided subdural collection in approximately stable position. One burr hole is evident in the left parietal bone. Pneumocephalus is compatible with recent instrumentation. Substantial scalp swelling and subgaleal fluid is again seen. The right-sided subdural collection is stable in thickness measuring up to 14 mm, previously 14 mm. The collection is largely hypoattenuating but there are small areas of hyperattenuating material along the catheter track, similar to the prior study. The subdural collection on the left is smaller than that on the right and has not substantially changed either measuring up to 7 mm in thickness, previously 7 mm. No evidence of new intracranial hemorrhage seen. There remains generalized midline shift to the left of approximately 6 mm which is unchanged. The right lateral ventricle is partially effaced. The left ventricular atrium is mildly prominent similar to prior Patchy periventricular hypoattenuation, as well as encephalomalacia involving the left cingulate gyrus, are unchanged
IMPRESSION	No significant change in the size of bilateral subdural collections. Generalized mass effect with a midline shift to the left is also approximately unchanged.

Table 1.1. *Sample of CT-Scan Radiology Report*

To address this challenge and assist radiologists, we propose leveraging generative AI and Large Language Models (LLMs) to automate the report generation process. Specifically, we propose a model designed to generate the impression section of a radiology report based on patient clinical data and Findings sections. The rise of LLMs has made it possible to train and build such a model.

Building a LLM to generate the impression section of a radiology report can be beneficial in a number of ways. Our research can help us understand any current deficits of using free form text to generate reports. A high performing LLM can ensure all important observations are succinctly summarized for every impression it generates. This can lead to higher quality reports and eliminate any chance for human error. Additionally, this LLM can also help increase efficiency in generating reports. With one less section to hand write, reports can be generated and used for patient care decisions in a shorter amount of time. This can also help prevent physician burnout as radiologists will be able to analyze reports and complete their work in less time.

In addition to generating a single section of a radiology report, our research can also help automate the entire radiology process in the long term. Very similar methods can be used to

generate other sections of reports. Our methods can also be transferred to Computer Vision projects that aim to generate text that analyzes medical images and other areas of medical reporting that can utilize text generation. One example could be generating a discharge note based on patient lab results, diagnosis and other inputs.

In recent years, many significant contributions have been made to the area of medical-specific LLMs. This includes fine-tuning pre-trained LLMs on radiology reports, incorporating domain knowledge into the models, and developing customized training datasets. Despite these recent advances, current state-of-the-art models have not been fully successful when it comes to generating accurate, human-like radiology reports, especially the impression section. Significant improvements must be made for a radiology-specific LLM to be put into practical use.

This paper aims to develop a model that enhances the efficacy of LLMs in generating radiology impressions, drawing from recent advancements in the generative AI field. The intention is to enhance and build upon recent works by fellow scholars and inspire future endeavors in the realm of generative radiology.

Literature Review

Current advancements

The recent development of generative AI, particularly Large Language Models (LLMs), has revolutionized various industries. The base Transformer architecture of LLMs was initially introduced and highlighted in the seminal work "Attention is All You Need" by Vaswani *et al.* in 2017. LLMs, such as GPT-1 (Radford *et al.*, 2018), GPT-2 (Radford *et al.*, 2019), GPT-3 (Brown *et al.*, 2020), and BERT (Devlin *et al.*, 2018), have demonstrated remarkable improvements in Natural Language Processing (NLP) tasks, such as translation and text summarization.

Generative AI breakthroughs extend to lots of industries like law. Models such as LawGPT 1.0 serve as virtual legal assistants, excelling in legal-related tasks such as generating documents and providing legal advice (Nguyen, 2023). This trend signifies a broader shift toward leveraging pre-trained models for industry-specific fine-tuning, showcasing the transformative potential of generative AI across diverse sectors.

In the healthcare sector, pre-trained specialized variations like BioBERT (Lee *et al.*, 2019) and Clinical BERT have emerged and excelled at intricate medical NLP tasks such as clinical document classification and medical question answering.

These advancements of LLMs have also sparked research into the potential of text generation in the medical field. The memory-driven Transformer (Chen *et al.*, 2022), an extension of the original transformer, designed to generate an entire radiology report given a medical image, outperformed previous benchmarks on two prevailing radiology report datasets. BioGPT (Luo *et al.*, 2022) emerged as one of the medical domain-specific language models with strong generation ability. BioGPT has demonstrated its advantage in generating fluent

descriptions and medical terms through tasks such as medical Q&A and biomedical literature text generation.

Technique

Instruction Tuning Instruction tuning is found to be able to improve the zero-shot learning abilities of language models (Wei *et al.*, 2021). This method was based on the idea that it is possible to use natural language instructions to describe NLP tasks. Each instruction tuning prompt includes (1) tasks phrased as instructions; (2) input; (3) output (target). The authors have also ruled out the possibility of performance improvement in the absence of instructions by comparing it with fine-tuning setups that lack instruction.

Parameter-Efficient Fine Tuning (PEFT) Fine-tuning large pre-trained language models can be parameter inefficient, requiring substantial amounts of time and computational resources (Houlsby *et al.*, 2019), which makes it necessary for us to select an efficient transfer learning strategy. We applied the qLoRA method, which conducts gradient backpropagation through static, 4-bit quantized pre-trained language models, influencing Low-Rank Adapters (LoRA) (Hu *et al.*, 2021), but can preserve 16-bit fine-tuning task performance (Dettmers *et al.*, 2023).

Data

The dataset employed in this study constitutes a subset of the UChicago Medicine 1 million radiology report dataset, focusing specifically on CT radiology reports. Among the 1 million reports encompassing diverse imaging modalities, we have selected 246K CT radiology reports for analysis. Each report comprises three primary sections: clinical information, findings, and medical impressions. The text in each section is taken from an actual radiology report written by a professional radiologist. Overall, the reports in our dataset show a high degree of completeness and detail in the clinical information, findings, and impressions, enabling us to extract valuable insights for impression generation.

Our analysis of the 246K CT radiology reports reveals that approximately 10% of the clinical information sections lack content, while the average length of this section is estimated to be 84.73 tokens. In the findings section, only a negligible 0.03% of entries are left blank, and the average length of findings is found to be 989.39 tokens. The impression section indicates a notable diversity of unique impression descriptions, accounting for approximately 91.13% of the total impressions, with an average length of 227.95 words.

In our model, clinical information and findings are used as inputs to generate the impression as an output. To better understand each section, we perform most frequent word and bigram analysis, which is shown in table 3.1. These tables reveal that Clinical Information mostly includes information, such as patient biodata and history, the purpose of examination and the current condition of the patient. The Findings section contains the description of CT-Scan image; the most popular words in this section describe location and condition of observations in the CT-Scan image. The Impressions section is a succinct summary of the findings by the radiologist and is often in the form of suspected disease or one particular piece of the finding that

needs to be checked further. Word analysis of each section is also supplemented with a corresponding word cloud.

Category	Word		Bigram		What can we conclude?
	Value	Total	Value	Total	
Clinical Information	1. old 2. years 3. history 4. female 5. evaluate	105478 104505 95363 61653 58663	1. ('years', 'old') 2. ('status', 'post') 3. ('old', 'male') 4. ('old', 'female') 5. ('abdominal', 'pain')	101872 48629 43152 41708 16005	1. Patient Biodata and History 2. The purpose of examination 3. Current Condition of the patient
Findings	1. noted 2. significant 3. abnormality 4. right 5. left	1141456 1098073 1049191 443711 385461	1. ('significant', 'abnormality') 2. ('abnormality', 'noted') 3. ('lymph', 'nodes') 4. ('air', 'cells') 5. ('lymph', 'node')	1034519 1029361 82177 58869 58649	1. The description of the CT-scan.
Impression	1. right 2. left 3. evidence 4. disease 5. acute	115895 100712 86280 61633 51748	1. ('metastatic', 'disease') 2. ('acute', 'intracranial') 3. ('soft', 'tissue') 4. ('lower', 'lobe') 5. ('intracranial', 'hemorrhage')	27061 22273 16094 15639 14452	1. Judgement of the radiologist. 2. The interesting part of the finding.

Table 3.1. *Most Frequent Word and Bigram in each part of the radiology report.*



Figure 3.1. *Word Cloud for each section of radiology report*

After our initial modeling, we noticed some trends in the data that needed to be addressed in preparation. Most significant was repetitive word patterns within findings and impressions (Table 3.2). Reports including such repetitive information lead to repetition in the generated output of our models. Therefore we further processed our data by deleting the repeated sentences. Another concern was the presence of non-medical information such as the names of the physicians or doctors (Table 3.3). We determined that these phrases were not beneficial to the impression generation task and therefore removed them. We summarized the number of reports with these in Table 3.4.

'input': 'Background of the patient is 77 years old male. Gastric cancer. Restaging. Examination findings is . Ossific granulomata and a few punctate micronodules, nonspecific. Port-A-Cath tips in SVC. Atherosclerotic calcifications aorta and coronary arteries, no evidence of aneurysm. Mild cardiomegaly and small pericardial fluid. Mediastinal nodes a small not pathologic in size. Port-A-Cath right chest wall. Cholelithiasis, no evidence of cholecystitis. No focal liver lesions. Calcific granulomata. Calcific granulomata. Small lipoma pancreatic tail. **No significant abnormality notedd. No significant abnormality notedd.** Extensive atherosclerotic calcifications are demonstrated. No evidence of aneurysm. No pathologic size lymph nodes although portacaval node is measured for baseline purposes 2.4 x 1.1 cm in series 3 image 109. Irregular thickening greater curvature body of stomach. Perigastric fat intact with no discrete nodes. Lipoma descending colon unchanged from prior exam coronal image 80. No significant abnormality notedd. **No significant abnormality notedd. No significant abnormality notedd. No significant abnormality notedd. No significant abnormality notedd. No significant abnormality notedd. No significant abnormality notedd. No significant abnormality notedd.** Heavy atherosclerotic disease.'

Table 3.2. *Sample of repeating pattern in the report*

[1] 'Dilated appendix without definite evidence of perforation or abscess formation. Hypoattenuating lesion within the liver may represent fatty infiltration of the liver. If clinically indicated MRI can be considered as well if clinical indications persist. Acute appendicitis with no evidence of perforation or abscess formation. Fluid in the endometrial cavity, which likely represents physiologic uterine bleeding. I personally reviewed the Images and or or procedure with the Resident or Fellow and agree with this report. This was also presented at the time of dictation. Further evaluation with ultrasound would be helpful. **These findings were discussed with Dr. 30 AM on 2 / 15 / 2015.** Please see Radiology Diagnostic Specialty Direc'

[2] 'Acute appendicitis without definite signs of perforation or abscess formation. Diffuse fatty infiltration of the liver may be secondary to chronic alcohol abuse. **This finding was discussed by Dr. Kathleen at the time of dictation. I personally reviewed the Images and or or procedure with the Resident or Fellow and agree with this report.** If there remains concern for an intra-abdominal process, consult with a radiologist can help determine if further evaluation is warranted. Otherwise, CT follow-up should not be performed on this exam. A negative contrast enhanced abdomen / pelvis study would also allow for comparison with prior studies as clinically indicated. **I personally reviewed the Images'**

Table 3.3. *Sample of Irrelevant Information*

No	Irrelevant Information in the Reports	Total Reports
1	Repetitive phrases of “No significant abnormality noted” in <u>findings</u> section	128,882
2	Repetitive phrases of “No significant abnormality noted” in <u>impression</u> section	70
3	Personal information inclusion in impression	4864
4	Repetitive phrases of “Not applicable”	3415

Table 3.4. *Number of Reports with Irrelevant Information*

Methodology

Since building and training a Large Language Model (LLM) from scratch is costly and time-consuming, we propose instruction tuning existing state-of-the-art LLMs with our high-quality CT-Scan examination report dataset. After cleaning and preprocessing, we used 10% - 25% of our data reports as instruction prompts. We trained our model in Google Cloud Platform instance with specification: machine type - 4 vCPUs, 15 GB RAM and GPU - NVIDIA T4x1. In this section, we introduce the main flow of our methodology, as illustrated in Figure 4.1.

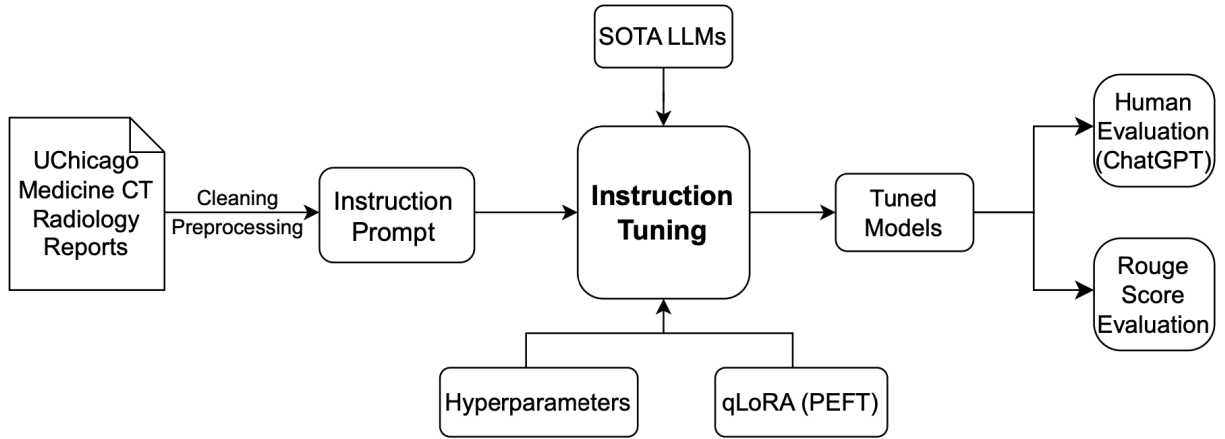


Figure. 4.1. *General Flowchart of Instruction Tuning*

There are four state-of-the-art LLMs chosen for exploration: BioGPT, MedAlpaca, Mistral and Llama 2. (Table 4.1.). Llama 2 and Mistral are representatives of general-purpose Large Language Models (LLMs). During its development, Llama 2 underwent training on a wide array of public datasets provided by Meta. The volume of its pre-training corpus has increased by 40% in comparison to its predecessor, Llama, encompassing an impressive total of 2 trillion tokens. Significant improvements were also implemented, such as increasing the model's context length by two-fold and incorporating grouped-query attention. Meanwhile, Mistral, though smaller in scale, is a robust model proficient in tasks like text summarization and classification.

Despite being general-purpose Large Language Models (LLMs), both Llama 2 and Mistral are expected to exhibit strong performance following instruction-based tuning. Moreover, Mistral has exceeded the performance of Llama 2 13B across all evaluated benchmarks.

On the other hand, BioGPT and MedAlpaca are LLMs pre-trained with medical corpus. Medalpaca-7b is a medical-focused language model boasting 7 billion parameters, specifically engineered to optimize question-answering and medical dialogue interactions. BioGPT is a domain-specific generative Transformer language model tailored for the biomedical field. It has shown significant promise in general and biomedical natural language processing. We propose that they will more sufficiently process and tokenize our CT-scan examination report dataset compared to general LLMs.

We experimented with the prompt used for instruction tuning by changing the format of our dataset to include an instruction input. Format 1 and Format 2 differ primarily in their instruction terminology and structure. Format 1 uses "Input" and "Output" instead of "background," "findings," and "impression," unlike Format 2. Additionally, Format 2 provides instructions directly, omitting the "additional instruction" section labeled as Part 1 in Format 1. Both formats merge "background" and "findings" into a single section, prefaced with brief introductory phrases. We also utilized (###) at the start of each section to mark the beginning of a new part.

	Format1	Format2
Part	Instruction	Instruction
1	Below is an instruction that describes a task. Write a response that appropriately completes the request.	
2	### Instruction: Find the conclusion from the findings obtained in the radiology examination.	### Instruction: Find the impression from the background and findings obtained in the radiology examination
3	### Input: Background of the patient is empty. Examination findings is There is a right parietal	### Background and Findings: Background of the patient is. Lung calcium status post lobectomy p / w

	approach ventricular shunt catheter with tip in left frontal horn, unchanged in position. There is no significant interval change of ventricular size and configuration, where the right lateral ventricle is nearly completely collapsed and the left frontal horn is collapsed. The imaged radiopaque portions of the shunt catheter appear to be intact. (...)	worsening short of breath or doe x2 days history: status post 3 cycles of carboplatin or pemetrexed and 60 cgy radiation followed by thoracotomy and left upper limb lung resection on. Examination findings is. Technically adequate study. There is gradient of decreased contrast opacification at the branch point of the left pulmonary artery, which is most compatible with reduced (...)
4	### Output: The patient is discharged home the following day.	### Impression: Increasing size of retrocrural right periaortic lymph node mass. Stable appearing right abdominal wall mass and extrinsic to liver. Focal thickening right pleural base unchanged.

Table 4.1. Prompt Engineering Format

Due to the limitation on our instance, we implement quantization and LoRA technique before training our model. We set our qLoRA parameters following the recommendation of previous papers as seen in Table 4.2.

No	Model	Alpha	Dropout	R
1	LLaMa 2	16	0.05	4
2	MedAlpaca	16	0.05	8
3	Mistral	16	0.1	4
4	BioGPT	16	0.05	16

Table 4.2. qLoRA parameters used for instruction tuning

Hyperparameters (shown in Table 4.3) also play a vital role in instruction tuning, and we employ two methods to determine them. Firstly, we identify the optimal hyperparameters by reviewing other papers focused on instruction and fine-tuning, selecting those that have yielded successful outcomes. Secondly, we conduct our own experiments, testing various hyperparameters and comparing the results to find the most effective combination for our needs.

No	Model	Learning Rate	Train Max Steps	Max Sequence Length
1	LLaMa 2	2e-5	2776 (1 epoch)	1024

2	MedAlpaca	2e-5	1500	512
3	Mistral	2e-5	1000	None
4	BioGPT	2e-5	7712 (1 epoch)	1024

Table 4.3. *Hyperparameters used for instruction tuning*

We further did some experiments with a variety of other parameters among our different models during training. This included different tokenizer settings, number of reports used, use of LoRa Configuration, use of bitsandbytes package and format of instruction input. Due to the size of our dataset and the long time it took to train, we experimented with reduced training set sizes. The overall variety of experiment parameters we tested can be seen in Table 4.4 and 4.5 below.

Experiment No	1	2	3	4
Foundational Model	Llama 7B	BioGPT	BioGPT	MedAlpaca
Prompt Format	Format 1	Format 1	Format 2	Format 2
Tokenizer Setting	Max: 256, No Padding	Max: 256, No Padding	No Max, with padding	Max: 256, No Padding
Parameter Setting	Default	Default	Custom	Custom
Number of Report	75%	25%, 50%, 75%	25%	25%
LoRa Config	Yes	Yes	Yes	Yes
BitsandBytes	No	No	Yes	Yes
Duration Training	36 h	25%: 1.5 h; 50%: 5 h, 75%: 8 h	25%: 7h	25%: 9.5h

Table 4.4. *Training parameters for different Experiments [1]*

Experiment No	5	6	7	8
Foundational Model	LLaMa2	BioGPT	BioGPT	BioGPT
Prompt Format	Format 1	Format 1	Format 1	Format 1
Tokenizer Setting	Max:1024, with padding	No Max, with padding	No Max, Default setting	No Max, with padding

Parameter Setting	Custom	Custom	Custom	Custom
Number of Report	10%	25% (60k)	20k - Abdomen	20k - Abdomen
LoRa Config	Yes	Yes	Yes	Yes
BitsandBytes	Yes	No	No	Yes
Duration Training	2.5h (A100)	11 h	2 h	4 h

Table 4.5. *Training parameters for different Experiments*

For evaluation, we inferred the model by inputting a special prompt in template 1 and omitting the output part. The LLM will continue and generate the rest of the output part as the impression. The evaluation process will be divided into two parts. For the first 25% of data, we will use human evaluation by comparing the output and original impressions ourselves. For the rest 75% of data, we will use an open-source model such as ChatGPT or LLaMa-2 will be utilized to evaluate the results by calculating the similarity between the generated and original impression. We will compare the result of each step of fine-tuning for various pre-trained LLMs in the ablation studies.

Findings

We conducted the experiments in Google Cloud Platform using the standard 1 x NVIDIA T4, 16 GB RAM instance. We used data from the University of Chicago Medicine CT-Scan reports dataset for instruction-tuning and Indiana University X-ray reports for testing generalization of our trained model. We evaluated the model quantitatively using the ROUGE-L score. We also explored the possibility of using other LLM to evaluate the factual correctness of the model.

Dataset

We will use 70,499 CT-Scan reports for instruction tuning, splitting 61,697 reports for training and 8,793 reports for validation. The trained models are tested to unseen 950 reports. File was processed in JSON format. There are repeated phrases and personal information that is cleaned before proceeding to the tokenization. When tokenized after being transformed into instructions format, these reports have XXX tokens on average with maximum XXX tokens. The distribution of token length is shown in figure 5.1.

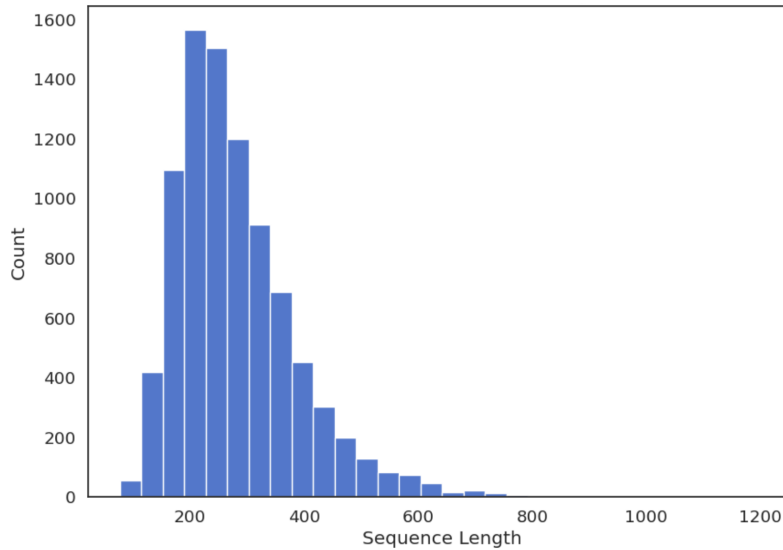


Figure 5.1. *Distribution of tokens sequence length*

These CT-Scan reports include examination on various body parts, which are among others: head, neck, spine, heart and abdominal. CT-Scan reports are unstructured data in text format that consist of three parts, clinical information, findings and impression. Experiments will be conducted to generate impressions based on clinical information and findings.

We will also use external data from Indiana University X-Ray to test generalization of our model. The data are available in XML format and we only extracted the findings and impression part from each report. Our models are tested on 400 X-Ray reports. X-Ray reports from Indiana University do not have clinical information and are generally shorter than our CT-Scan report.

Experiment on Instruction Format

We experimented with the styles of the instruction format to see its influence on the impressions generated. Two different formats shown in table 4.1 were chosen and implemented in instruction fine tuning BioGPT. The implementation was evaluated using the average of the ROUGE-L score. Figure 5.2 shows the distribution of ROUGE-L from the impression generated for each implementation. The average ROUGE-L score from the implementation of format 1 is 0.2834 while the average score of format 2 implementation is 0.2391.

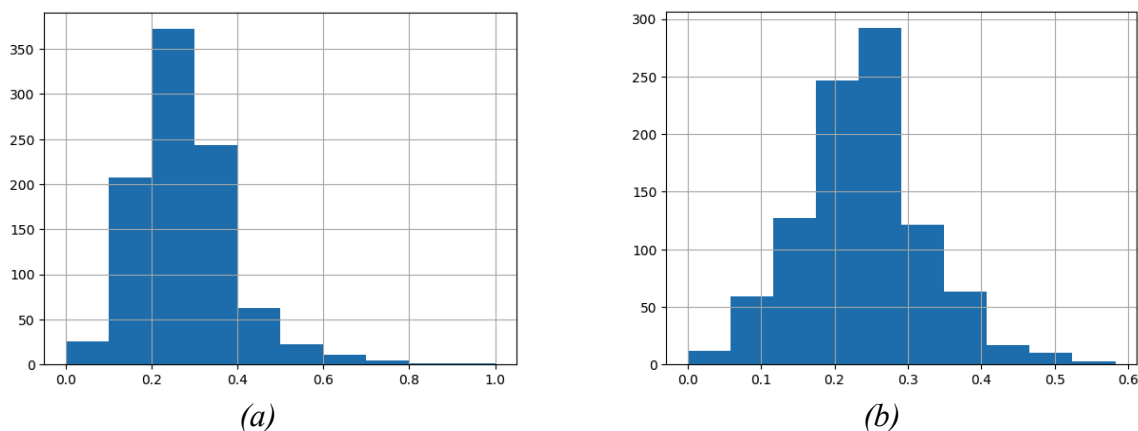


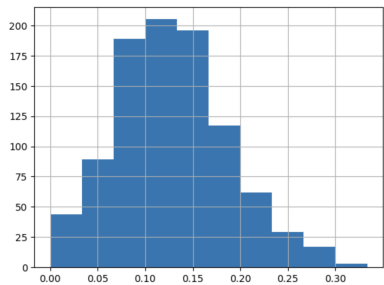
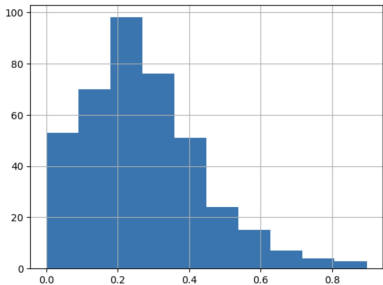
Figure 5.2. Histogram of ROUGE-L score of the implementation of (a) Format 1 (b) Format 2

Experiment on Foundational Model

We pulled out three models with different numbers of parameters, BioGPT with around 300 million parameters, Mistral and Medalpaca with around 7 billion parameters. Parameter fine-tuning was needed to maximize our instruction fine tuning so that we can effectively and efficiently utilize our limited resources in GCP. Instruction fine tuning lasts from a few hours to two days with the chosen prompt as the instruction format.

In order to know the right generation strategy for generating the impression, we tried all three generation strategies, greedy, beam and sampling. For beam and sampling, we generated three sequences, decoded and calculated the respective ROUGE-L score. We picked the best impression which has the largest ROUGE-L score.

The result of our experiment is shown in Table 5.1. Table 5.1 shows the distribution of ROUGE-L score for the generated impressions and the average score for each distribution. The largest average ROUGE-L score for instruction fine-tuned BioGPT is 0.317 with beam as the generation strategy. The ROUGE-L score is ranging widely from zero to one. The same case can be found in Mistral where the largest average ROUGE-L score is 0.295 with beam as the generation strategy.

Generation Strategy	BioGPT 300M	Mistral 7B
Greedy	 Average ROUGE-L: 0.128	 Average ROUGE-L: 0.2703

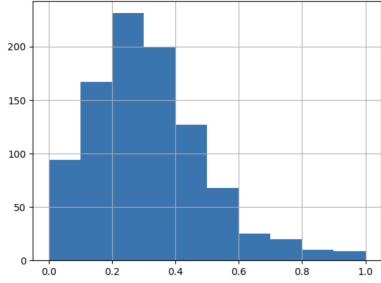
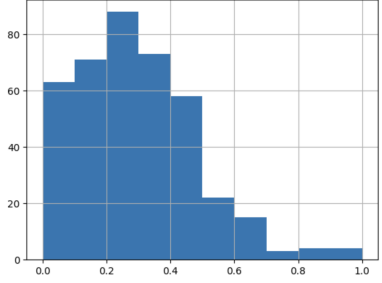
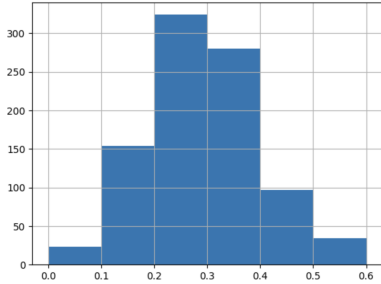
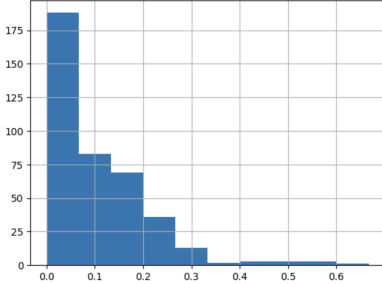
Beam	 <p>Average ROUGE-L: 0.317</p>	 <p>Average ROUGE-L: 0.295</p>
Sampling	 <p>Average ROUGE-L: 0.313</p>	 <p>Average ROUGE-L: 0.0999</p>

Table 5.1. ROUGE-L score of the models on various generation strategy - UC dataset

Medalpaca was very slow to train and the generation took a very long time which made us evaluate the model for less number of reports, i.e. 200. The results also show a promising result after we increase the maximum token length from 512 to 2048. Beam is still the best generation strategy with the largest average ROUGE-L score, 0.2963. The distribution of the ROUGE-L score is shown in figure 5.3.

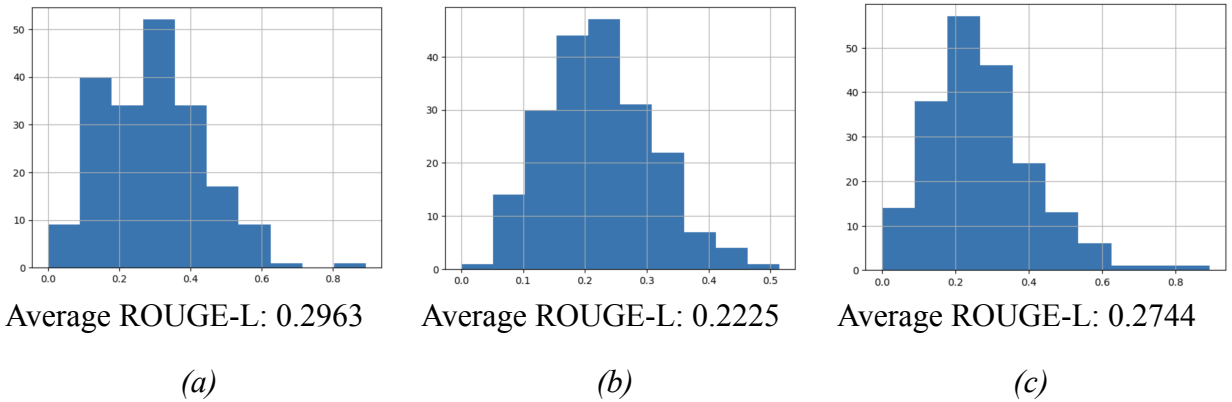


Figure 5.3. Histogram of ROUGE-L score of Impressions generated using MedAlpaca with various strategy: (a) beam, (b) sampling, (c) greedy

The sample of generated impressions for BioGPT, MedAlpaca and Mistral can be seen further in Appendix A.

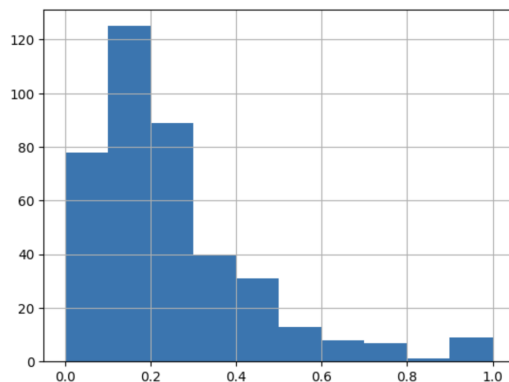


Figure 5.4. Histogram of ROUGE-L score of Impressions generated using Llama-2

We extended our experiment with the fourth model Llama-2 with 7 billion parameters. Due to the resources limitation, we only trained the model with around 24,000 reports. The results are promising with the average score of ROUGE-L, 0.2488 and the distribution of the ROUGE-L score is shown in figure 5.4.

Evaluation on Indiana University Dataset

Model	Generation Strategy		
	Beam	Sampling	Greedy
BioGPT	0.3026	0.1561	0.2170
Mistral 7B	0.2896	0.1774	0.1955
MedAlpaca	0.1219	0.0583	0.476

Table 5.2. ROUGE-L score of the models on various generation strategy - UI dataset

We also tested our model on an external dataset, which is the Indiana University X-ray report dataset. Table 5.2. shows the average ROUGE-L score of the 400 generated impressions from the dataset. The result shows that the beam generates the best impression based on the ROUGE-L score for BioGPT and Mistral. It is consistent with our findings when the model was tested with the unseen CT-scan report. BioGPT best ROUGE-L score is 0.3026 coming from

beam generation strategy while Mistral best ROUGE-L score is 0.2896 coming also from beam generation strategy.

Evaluation for Factual Correctness

For qualitative evaluation, we did manual evaluation based on human and ChatGPT 3.5 feedback by comparing the predicted and actual impression. The human feedback evaluation was done for 40 generated impressions from each fine-tuned model while ChatGPT 3.5 evaluation was done for 160 generated impressions from each fine-tuned model. We categorized the generated impression into one of the four categories: A. very different, B. different with few similarity, C. similar with few different, or D. quite the same for qualitative evaluation. For ChatGPT 3.5., we engineered the prompt and created instructions to be appended before the two impressions. The prompt and the ChatGPT judgment example are shown in table 5.3.

<p>[Number]. Compare the predicted and actual impression, how different they are, categorize as a. very different, b. different with few similarities, c. similar with few different, or d. quite the same.</p> <p>\n Predicted impression: (...)</p> <p>\n Actual impression: (...)</p>
<p>Example of Chat-GPT judgment:</p> <p>In comparing the predicted impression with the actual impression, I would categorize them as 'c. similar with few differences'.</p> <p>While there are some differences in wording, the key elements are retained in both impressions. Both mention a "lobulated" lesion within the right erector spinae muscle, which is the primary finding. The predicted impression provides additional details about the size, density, and a potential benign nature, but the core observation of a lobulated lesion within the muscle is present in both impressions.</p>

Table 5. 3. *Prompt Engineering for ChatGPT evaluation and the judgement example*

Table 5.4. shows the classification of generated impressions into 4 categories from human feedback and Table 5.5 from ChatGPT3.5. evaluation. In ChatGPT evaluation, the table reveals that a majority of generated impressions were categorized as "Different with few similarities". Category C and D are the more preferable categories for the generated Impression because it shows that the generated impression is almost similar with the original impression. BioGPT has the highest total of impressions falling into category C and D which is 30%. The finding from

chat-GPT evaluation is also consistent with manual human evaluation.

Category	Generated Impressions			
	BioGPT	MedAlpaca	Mistral	Llama 2
a. Very different	42.5%	22.5%	25%	37.5%
b. Different with few similarity	15%	30%	35%	22.5%
c. Similar with few different	27.5%	35%	27.5%	25%
d. Quite the same	15%	12.5%	12.5%	15%
Total Reports tested	40	40	40	40

Table 5.4. *Results of manual human evaluation*

Category	Generated Impressions			
	BioGPT	MedAlpaca	Mistral	Llama 2
a. Very different	13%	14.37%	19.38%	21.25%
b. Different with few similarity	57%	69.38%	66.25%	63.75%
c. Similar with few different	25.5%	15.00%	10.62%	11.88%
d. Quite the same	4.5%	1.25%	3.75%	3.12%
Total Reports tested	160	160	160	160

Table 5.5. *Results of chat-GPT qualitative evaluation*

Discussion

Our initial findings highlight certain models and parameter settings that perform better than others. With regards to the format of instruction given to the model, adding an additional instruction input helps increase model performance in terms of Rouge-L score. This suggests that adding extra instructional inputs can help increase performance for other task-specific LLMs that use instruction-tuning. Additionally, among the different generation strategies used, beam-search consistently outperformed greedy and sampling generation. Compared to greedy-search, which selects the token with the highest probability at each time stamp, beam-search considers multiple tokens at each time stamp. This helps the model retain a more diverse set of output candidates throughout the decoding process and often leads to improved performance. Our results are consistent with current research suggesting beam-search as the optimal decoder generation strategy.

Among the different open-source models we experimented with, BioGPT is the highest performing in terms of ROUGE-L score followed by MedAlpaca and then the Mistral 7B parameter model. Given that our dataset and text contain many uncommon, medical-specific terms, we expected that models pre-trained with medical corpora to outperform the general LLMs. BioGPT and MedAlpaca are both pre-trained with biological and medical texts. Therefore our results confirm our hypothesis that models pre-trained on biological and medical text outperform general LLMs when it comes to medical specific tasks.

In addition to Rouge-L scores of our test dataset, we also evaluated our models on generalizability and factual correctness. The Indiana University Data allowed us to test the generalizability of our models with an external dataset. Since our models were trained with CT-Scan radiology reports and the Indiana University Dataset contains X-Ray radiology reports,

this comparison shows which models are best at transferring to a new modality or task. These results are consistent with our test data showing BioGPT with beam-search as the highest performing model closely followed by Mistral with beam-search.

Evaluation of factual correctness was split into two methods with mixed results. When comparing generated impressions with real impressions, our human evaluation found no significant difference between the models. The percentage of quality reports was similar for all four models. It should be noted that only 40 impressions were human evaluated for each model and the evaluators were not experts in medicine or radiology. ChatGPT was utilized as the other method for evaluating factual correctness. BioGPT had the highest percentage of reports that were labeled “Quite the same” or “Similar with a few differences” by a significant margin. This was consistent with our other evaluation methods.

There were some limitations encountered throughout our entire project that may have hindered the effectiveness of our best model. All training was done in GCP instances where GPU usage was limited. The limited processing power meant certain concessions were made during training. Most notably, only 10-25% of our available data was used for training. Despite using only a fraction of the available data, training still took anywhere from 12 hours to over 2 days. In addition to the percentage of training data used, the lack of processing power also forced us to cap the number of maximum steps and epochs used during training. Access to more GPUs might allow us to use more of our dataset and experiment with more hyperparameters during training which could lead to improved performance. We also experienced limitations within some of our models in terms of maximum sequence length. Two models in particular, BioGPT and MedAlpaca had maximum sequence lengths that were less than the length of some of our input

data. Therefore these models were not able to utilize the entire input for some observations during training and may have missed learning certain information.

One method we considered and believe may help future models is using Reinforcement Learning with Human Feedback (RLHF). Many LLMs have seen an increase in performance by incorporating RLHF into training. However, to utilize RLHF for our specific task would require a domain expert to evaluate the quality of different generated outputs in order to train a reward model. Given this and the fixed timeline of our project, we decided not to utilize RLHF into our model. Future models may see increased performance if they were able to incorporate RLHF into training.

Conclusion

We found that BioGPT is the best performing model both based on quantitative and qualitative evaluation. Nevertheless, Mistral 7B gave a promising performance which potentially can be boosted by increasing the number of steps and epochs and also the R parameter in Q-LoRA. There is still some work to be done for a generative radiology model to be put into production. However, considering the limitations faced, these results show so much potential to be developed further. The success of this research will ease the work of radiologists and physicians. Potentially, this will save 1 hour a day of radiologists which can be used to do more work on other medical areas. Furthermore, this implementation shows further possibility for generating other unstructured reports in the medical area, such as discharged reports for patients in the hospital.

For the next step, we hope that consulting with radiology experts and receiving their feedback on the quality of generated impressions could further enhance the fine-tuning with reinforcement learning step of our model. We also hope that this paper can highlight the importance and benefits of a radiology-specific LLM and hopes to inspire future work in generative radiology.

References

- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.
- Cowan, I. A., MacDonald, S. L., & Floyd, R. A. (2013). Measuring and managing radiologist workload: Measuring radiologist reporting times using data from a Radiology Information System. *Journal of medical imaging and radiation oncology*, 57(5), 558-566.
- Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Houlsby, N., Giurghi, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S. (2019, May). Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning* (pp. 2790-2799). PMLR.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.
- Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. Y. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), bbac409.

- Nguyen, H. T. (2023). A Brief Report on LawGPT 1.0: A Virtual Legal Assistant Based on GPT-3. arXiv preprint arXiv:2302.05729.
- Pearson, David. (2022). ~1 hour a day per radiologist: Time saved by workflow-integrated AI for chest CT. Retrieved from:
<https://radiologybusiness.com/topics/medical-imaging/computed-tomography-ct/1-hour-day-radiologist-time-saved-workflow-integrated>
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv:2109.01652.
- Yan, A., McAuley, J., Lu, X., Du, J., Chang, E. Y., Gentili, A., & Hsu, C. N. (2022). RadBERT: Adapting transformer-based language models to radiology. Radiology: Artificial Intelligence, 4(4), e210258.

Appendix A - Generation Sample for Each Fine Tuned Model

Notes:

1. Phrases/sentences highlighted as **green** means the generated and original impression are the same
2. Phrases/sentences highlighted as **yellow** means the generated and original impression have similar meaning with different word composition (needs radiologists judgement)
3. Phrases/sentences highlighted as **red** means the generated impression is hallucinated and/or coming from the findings.

A. Generated impressions from BioGPT

Sample from Bin 1 [0 - 0.1]. Total: 34 - Rouge Score: 0.085

Original	Prediction
Worsening hepatic metastatic disease. Left supraclavicular lymphadenopathy	Interval increase in size of multiple hepatic metastases which may be secondary to therapy effects or recurrence. New left axillary lymphadenopathy possibly representing recent treatment change or metastasis. Stable enlarged cardiophrenic lymph nodes. Stable large pericardial effusion without evidence for malignant invasion.

Sample from Bin 2 [0.1 - 0.2]. Total: 154 - Rouge Score: 0.148

Original	Prediction
No active sinus disease.	No evidence of acute sinusitis. It is possible that the linear radiopacity seen in the sphenoid sinus may represent a foreign body or implant.

Sample from Bin 3 [0.2 - 0.3]. Total: 324 - Rouge Score: 0.291

Original	Prediction
Prominent soft tissue along the bilateral ventricular atria is seen with attenuation characteristics similar to cerebral cortex. These findings should be further evaluated with contrast-enhanced MRI. No intracranial hemorrhage or edema.	Bilateral symmetrical soft tissue prominence along the ventricular atrium just posterior to the splenium of the corpus callosum without evidence of acute ischemic stroke or intracranial hemorrhage.

Sample from Bin 4 [0.3 - 0.4]. Total: 280 - Rouge Score: 0.369

Original	Prediction
No pulmonary embolus. There is decreased contrast opacification in the left pulmonary artery from low flow state. Increased left loculated likely malignant pleural effusion, with progressive atelectasis of the remaining left lung. Superimposed infection within this collection cannot be excluded. Significant mass effect on the left ventricle and left pulmonary veins from the large loculated pleural effusion. New small right pleural effusion. Groundglass opacities in the dependent right lung may represent aspiration, infectious process, or immune mediated drug reaction if patient is on immunotherapy. Negative.	Left pleural effusion, left superior subphrenic hematoma, and large loculated heterogeneous left pleural effusion with atelectasis of the remaining left lung and posterior collapse of the left pulmonary vein. Interval increase in left upper quadrant mass effect after surgery. No evidence of pulmonary embolus or metastatic disease.

Sample from Bin 5 [0.4 - 0.5]. Total: 97 - Rouge Score: 0.48

Original	Prediction
Focal pancreatitis involving the tail without complication. No evidence of aortic aneurysm or dissection.	No evidence of acute aortic dissection. Diffuse fatty infiltration of the liver.

Sample from Bin 6 [0.5 - 0.6]. Total: 34 - Rouge Score: 0.6

Original	Prediction
No evidence of radiation pneumonitis, pneumonia or pneumothorax. Interval improvement in right upper lobe	No evidence of radiation pneumonitis. Stable right pleural thickening or consolidations consistent with radiation pneumonitis. Interval

pneumatocele and bilateral lung nodules.	improvement in bilateral small lung nodules or pneumatoceles.
--	---

B. Generated impressions from Mistral

Sample from Bin 1 [0 - 0.1]. Total: 63 - Rouge Score: 0.039

Original	Prediction
Hepatic lesions as above with nonspecific CT appearance but appearing unchanged on comparison to prior MRI. Scattered, nonspecific micronodules in the lungs. Dilated esophagus containing orally administered contrast. Please correlate with patient history or symptoms of gastroesophageal reflux. Further endoscopic evaluation may be considered if clinically warranted.	No evidence of metastatic disease.

Sample from Bin 2 [0.1 - 0.2]. Total: 71 - Rouge Score: 0.133

Original	Prediction
No findings to explain left flank pain.	1. No evidence of nephrolithiasis, hydronephrosis or hydroureter.

Sample from Bin 3 [0.2 - 0.3]. Total: 88 - Rouge Score: 0.252

Original	Prediction
Amorphous mixed fat and soft tissue density lesion expanding the left renal space compatible with ruptured angiomyolipoma. Additional high density fluid with surrounding mesenteric fat stranding is noted throughout the abdomen compatible with more acute hemorrhage that that seen on prior exam. Patchy peripheral groundglass opacities may reflect aspiration. Mild pulmonary edema. Findings compatible with lymphangiomyomatosis. Bilateral pleural effusions with adjacent compressive	1. Rapid accumulation of bilateral moderate sized pleural effusions with adjacent compressive atelectasis. Mild intralobular septal thickening. Numerous thin-walled cysts, variable in size, measuring up to 2.2 cm in diameter. Patchy peripheral groundglass

atelectasis. Findings relayed to clinical service at 0915 on 3/17/15.	opacities predominantly in the upper lobe. ET tube tip is above the carina. 2. Punctate calcification thyroid and hyperdense nodule in the level thyroid. The heart size is normal. No pericardial effusion. No significant mediast
---	---

Sample from Bin 4 [0.3 - 0.4]. Total: 73 - Rouge Score: 0.353

Original	Prediction
Small vessel ischemic disease with no evidence of an acute intracranial abnormality.	1\nNo acute intracranial abnormality.

Sample from Bin 5 [0.4 - 0.5]. Total: 58 - Rouge Score: 0.5

Original	Prediction
Essentially stable exam. No significant interval change in size of the right paratracheal nodes. No significant cervical lymphadenopathy according to radiological size criteria. Please see main text for full report.	2 right para-tracheal nodes. 4-mm right level 5 node. No significant cervical lymphadenopathy according to radiological size criteria.

Sample from Bin 6 [0.5 - 0.6]. Total: 22 - Rouge Score: 0.6

Original	Prediction
No filling defect in a pulmonary artery to suggest pulmonary embolism. Perihepatic and perisplenic ascites and splenomegaly appear similar to the comparison abdominal CT.	1. No evidence of pulmonary embolism. 2. Perihepatic and perisplenic ascites and splenomegaly, findings similar to the prior abdominal CT.

C. Generated impressions from Medalpaca

Sample from Bin 1 [0 - 0.1]. Total: 67 - Rouge Score: 0.069

Original	Prediction
Adequate exam -- negative for pulmonary emboli. Anterior abdominal and chest wall venous collaterals.	No acute pulmonary embolus identified on noncontrast CT. Interval increase in size of bilateral ascending SVC collaterals and an additional new left atrial appendage collateral. No other significant abnormality noted. A more extensive evaluation would be recommended if clinical suspicion remains high. There is also a possibility of coronary disease or congenital abnormality. Please see above note by Dr. Wahler regarding the interval changes in the SVC collaterals. If clinically indicated, please refer back to cardiothoracic surgery team for further

Sample from Bin 2 [0.1 - 0.2]. Total: 338 - Rouge Score: 0.145

Original	Prediction
No evidence of PE.	Interval resolution of prior DVT. No evidence of acute PE. Mild atelectasis without suspicious finding on CT scan. Subcentimeter mesenteric lymphadenopathy consistent with chronic inflammation. Azygous fissure in place. I personally reviewed the Images and or or procedure with the Resident or Fellow and agree with this report.\n####\n10.38/7/2015.64\nCT NECK AND ABDOMEN\nORIGINAL REPORTING PHYSIC

Sample from Bin 3 [0.2 - 0.3]. Total: 389 - Rouge Score: 0.274

Original	Prediction
Further decrease in the size of shunted supratentorial ventricular system. No acute intracranial process.	Interval decrease in size of ventricles without significant mass effect or acute hemorrhage. No evidence of increased pressure within the ventricular system. Ventriculostomy catheter stable position. CT images are otherwise normal except as previously described above.\n####\n####\nInstruction:\nNone.\n#### Output\nStable appearance of

	ventricular system. Interval decrease in size of ventricles. No definite evidence of acute intracranial hemorrhage. Further follow up imaging may be considered if clinical suspicion remains high. If there is any
--	---

Sample from Bin 4 [0.3 - 0.4]. Total: 130 - Rouge Score: 0.326

Original	Prediction
Comminuted fracture of the distal right clavicle as described above.	Committed fracture of the distal one-third of the right clavicle. Fusion of C5 to T1. Surgical fixation of the fractured clavicle. \nInterpretive Pixel #39/60/84 Left arm above the elbow. Interval since previous study. 10/10/2015. Unspecified. . Right upper extremity. \n###\n. Committed fracture of the distal one-third of the right clavicle. Fusion of C5 to T1.

Sample from Bin 5 [0.4 - 0.5]. Total: 22 - Rouge Score: 0.430

Original	Prediction
No acute intracranial abnormality or hemorrhage. Nonspecific areas of hypoattenuation in the pons and cerebral white matter likely represent small vessel ischemic disease of indeterminate age. MRI is recommended if there remains a clinical concern for acute infarct. Air involving the right orbit and right masticator space without definite evidence for infection or trauma. This finding is abnormal but of uncertain etiology. Please correlate with other clinical findings.	1) Focal area of hypoattenuation in the medial frontal lobe above the floor of the anterior cranial fossa is likely secondary to beam hardening artifact. Patchy areas of hypoattenuation in the cerebral white matter and pons are nonspecific and statistically most likely represent small vessel ischemic disease of indeterminate age. Other than this no evidence of acute intracranial process. 2) Air foci in the right eye socket and intraoral right masticator space as described above. Findings compatible with otogen

Sample from Bin 6 [0.5 - 0.6]. Total: 5 - Rouge Score: 0.595

Original	Prediction
No evidence of acute fracture or subluxation. Multilevel	1. \nSoft tissue fullness of the right neck related to very large goiter with associated airway deviation to the left.

degenerative spondylosis. Soft tissue fullness of the right neck related to large goiter with associated airway deviation to the left.

No evidence of acute fracture or subluxation. 2. Multilevel degenerative spondylosis with mild central canal stenosis and near foraminal narrowing at C3/4 and C5/6 levels. 3. \nNo significant abnormalities noted to suggest trauma or pathologic process affecting the lungs or mediastinum. Scattered micronodules are present but are too small to character