

## Rad-Phi2: Instruction Tuning Phi2 for Radiology

### Mercy Ranjit

Microsoft Research,  
India  
Department of Computer Science, Bharathidasan University,  
India

meranjit@microsoft.com

### Gopinath Ganapathy

Department of Computer Science, Bharathidasan University,  
India

gganapathy@bdu.ac.in

### Shaury Srivastav

Microsoft Research,  
India

t-ssrivastav@microsoft.com

### Srujana Oruganti

Microsoft Research,  
India

sruvs@microsoft.com

### Tanuja Ganu

Microsoft Research,  
India

taganu@microsoft.com

**Corresponding Author:** Mercy Ranjit

**Copyright** © 2024 Mercy Ranjit, et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

Small Language Models (SLMs) have shown remarkable performance in general domain language understanding, reasoning and coding tasks, but their capabilities in the medical domain, particularly concerning radiology text, is less explored. In this study, we investigate the application of SLMs for general radiology knowledge specifically question answering related to understanding of symptoms, radiological appearances of findings, differential diagnosis, assessing prognosis, and suggesting treatments w.r.t diseases pertaining to different organ systems. Additionally, we explore the utility of SLMs in handling text-related tasks with respect to radiology reports within AI-driven radiology workflows. We fine-tune Phi2, a SLM with 2.7 billion parameters using high-quality educational content from Radiopaedia, a collaborative online radiology resource. The resulting language model, RadPhi2-Base, exhibits the ability to address general radiology queries across various systems (e.g., chest, cardiac). Furthermore, we investigate Phi2 for instruction tuning, enabling it to perform specific tasks. By fine-tuning Phi2 on general domain tasks followed by radiology-specific tasks related to chest X-ray reports, we create RadPhi2. Our empirical results reveal that Rad-Phi2-Base and Rad-Phi2 perform comparably or even outperform larger models such as Mistral-7B-Instruct-v0.2 and GPT-4 providing concise and precise answers.

**Keywords:** Small Language Models (SLMs), Instruction Tuning, Radiology Reports

## 1. INTRODUCTION

Recent advances in Natural Language Processing (NLP) have enabled the development of powerful language models that can generate natural language texts for various purposes and domains. However, most of the existing language models are trained on general domain texts, such as news articles, books, web pages, etc., and may not be able to handle the specific vocabulary, concepts, terminology, and style of radiology texts. Hence, it is required to have domain specific models that embody knowledge for that domain and trained for specific tasks in domain specific workflows to deal with the nuances of the field. But, the size and complexity of these language models are often prohibitive for practical applications, as they require a large amount of computational resources and memory to train and run.

In this paper, we explore the use of Small Language Models (SLMs) for radiology related tasks, specifically text-related tasks relating to radiology knowledge and tasks that are useful in everyday radiology workflows pertaining to radiology reports. SLMs, in particular, can be useful in a medical setting as they promote the protection of patient privacy and independence of hospitals. Due to their lower compute requirements, hospitals can run these models on-premises. It also increases the flexibility to deploy specialised models for different tasks.

SLMs are language models that have a relatively small number of parameters, compared to the state-of-the-art large language models, such as GPT-4. SLMs, specifically Phi2 [1], is a 2.7 billion parameter model that has shown remarkable performance in general domain language understanding, reasoning, and coding tasks, but its capabilities in the medical domain, especially on radiology content and tasks specific to a radiology workflows is not explored. Moreover, Phi2 is not an instruction tuned model and tends to produce verbose answers as it was trained from synthetic text books and currently not suitable for direct deployments in application workflows to perform specific tasks. We hypothesize that SLMs can be effectively adapted to the radiology domain by finetuning them on high-quality radiology content to bring-in the nuanced radiology knowledge and also be adapted to perform specific tasks that are useful in everyday radiology workflows. SLMs provide the advantage of having a domain specific model for easy consumption at lower computational cost and faster inference.

The Phi models [1], success and its on-par performance to large language models was attributed to the high quality aka *“textbook-quality”* training data that was used for its training. The training corpus was augmented with carefully selected data, filtered based on educational value and content quality. It demonstrated that this choice of training data enabled Phi2 to achieve outstanding performance on complex benchmarks, matching or outperforming models up to 25x larger in size. We extend this approach of Phi2 by using high-quality and reliable educational content from Radiopaedia [2], to finetune Phi2 for radiology knowledge. Radiopaedia is a rapidly growing collaborative online peer-reviewed open-edit radiology resource, compiled by radiologists and other health professionals from across the globe. Training from this highly valuable and reliable source can embody the model with this knowledge making it useful to answer questions on radiology accurately. To test our hypothesis, we use Phi2, a 2.7 billion parameter SLM, as our base model, and finetune it on question-answer pairs derived from the educational articles of

Radiopaedia pertaining to different systems like chest, cardiac, gastrointestinal etc. The resulting model Rad-Phi2-Base, can answer queries on radiology pertaining to these systems accurately. We evaluate the performance of Rad-Phi2-Base and compare it with larger instruction tuned models, such as Mistral-7B-Instruct-v0.2 [3], which was finetuned on the same radiology question answer pairs and also GPT-4. Rad-Phi2-Base performs better than or comparable to the much larger counterparts while producing more crisp and to-the-point answers. It even surpassed the Retrieval Augmented Generation (RAG) approach [4], that directly used a knowledge base of Radiopaedia articles to answer the queries.

Apart from answering queries on radiology, there are specific text-related tasks that can be useful in AI deployment scenarios of radiology workflows. Some useful tasks relating to radiology reporting workflows are writing diagnostic impression summary from radiology findings, tagging of radiology reports with appropriate abnormality labels to flag priority cases, noise free extraction of findings and impression from existing radiology reports, generating a comparison summary of findings between the current radiology report and its prior report etc. Some of these tasks were also cited in the earlier work [5], in the context of application of language models in radiology workflows. To address these useful tasks that arise in the AI deployment scenarios of radiology workflows, we create an instruction tuning dataset for tasks pertaining to radiology reports of Chest X-rays and instruction-tune Phi2 and obtain the Rad-Phi2 model. But, as Phi2 was not an instruction tuned model in the first place, we first instruction tune Phi2 for general domain tasks using the Super Natural instruction tuning dataset [6]. The resulting model Phi2-Instruct(SupNatInst) is then further instruction tuned for tasks specific to radiology reports. The resulting model, Rad-Phi2 can perform specific tasks that are useful in AI driven radiology reporting workflows of Chest X-rays. Our results show that Rad-Phi2 performs better or on-par with a larger model like Mistral-7B-Instruct(v0.2) finetuned on the same instruction tuning dataset.

We would also like to note that while GPT-4 has reasonable radiology knowledge as noted in this work [7], it still falls short on nuanced tasks like impression generation in comparison to domain specific models unless given with a large number of carefully curated examples in conjunction with sophisticated prompt engineering approaches[8]. SLMs enables to develop domain-specific models for these nuanced task types easily and quickly due to the low resource requirements associated with its size.

Our work demonstrates the feasibility and effectiveness of using SLMs for radiology specific tasks, and opens up new possibilities for improving the quality and efficiency of radiology practice.

Our key contributions include:

- We demonstrate the effectiveness of SLMs in the radiology domain by training on high quality radiology content from Radiopaedia [2]. The resulting model Rad-Phi2-Base can answer general queries on radiology like symptoms, radiological appearances of findings, differential diagnosis, prognosis, and treatments w.r.t diseases pertaining to 15 different systems like Chest, Cardiac etc. It performs better than larger models like Mistral-7B-Instruct-v0.2 finetuned on the same dataset. It also performed better than GPT-4 and the Retrieval Augmented Generation approach that directly used Radiopaedia content as the knowledge source while answering queries.

- We construct an instruction tuning dataset of text related tasks that are useful in radiology report writing workflows of Chest X-rays using annotations of existing datasets on Chest X-ray reports. We share the prompts for different tasks used in creating this dataset.
- We also demonstrate Instruction Tuning on SLMs to make it usable for practical tasks involved in radiology workflows. As Phi2 is not an instruction tuned model and tends to produce verbose answers, we first perform general domain instruction tuning using the Super Natural Instructions dataset followed by radiology specific instruction tuning on tasks related to radiology reports of Chest X-rays. We call the resulting model Rad-Phi2 which is useful in performing everyday tasks in radiology report writing workflows of Chest X-rays.
- Our ablation studies gives insights on the input formats that worked the best when instruction tuning the Phi2 model. We note that radiology specific instruction tuning performs better when preceded by general domain instruction tuning.

## 2. RELATED WORK

The work, ‘Exploring the Boundaries of GPT-4 in Radiology’ [7], evaluated GPT-4 on a diverse range of common radiology text-based tasks. It was noted that that zero-shot prompting technique was sufficient for the ‘*learn-by-instruction*’ tasks like sentence similarity, radiology NLI, and disease progression classification. However, a large number of carefully curated examples with advanced prompt engineering were required for ‘*learn-by-example*’ tasks like impression summary writing to perform on par with the supervised models trained for the task. While it is beneficial to have GPT-4 like instruction-tuned models that can work well on radiology content for ‘*learn-by-instruction*’ tasks, it is still beneficial to have specialized models for ‘*learn-by-example*’ kind of tasks where the required knowledge for the model needs to be fed in as carefully curated in-context examples. It is even better if a model can do both these kinds of tasks with high accuracy at lower computational and resource costs, which is what we accomplish in our work.

The work ‘Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications’[5], highlights how LLMs are useful in typical radiology workflows, providing key relevant information from clinical notes and prior reports, thus improving radiology interpretation and clinical decision making. The paper also noted LLMs when connected to sources like the American College of Radiology performed better than radiologists in applying this information to clinical scenarios. In regular practice, radiologists and clinicians refer to external knowledge sources like Radiopaedia [2], for information like clinical symptoms, radiological appearances of findings, differential diagnosis, prognosis, and treatment plan of a disease that aids them with clinical decision making. These knowledge sources are very reliable as they are maintained by practitioners of medical practice and are actively updated. However, this also poses a need for the AI models to be updated regularly with this knowledge. While Retrieval Augmented Generation [4], can help LLMs augment their responses with the knowledge from the external sources, it may prove costly for high usage scenarios. This calls for a need for having Small Language Models that are easy to train, deploy, retrain, and iterate.

While instruction tuning to perform multiple tasks has been well explored on general domain tasks, it is less explored on radiology specific tasks. Radiology-Llama2[9], and Radiology-GPT[10], performed instruction tuning for one specific task of generating clinically useful impression summary

from radiological findings. However, there are a variety of tasks that can be useful in radiology report writing workflows, like extracting a clean set of findings or impressions from a noisy radiology report, tagging the reports with abnormality labels or tubes, lines and devices labels, or comparing two reports to see the differential findings between them, querying the reports for specific questions or querying for general radiology knowledge with respect to various diseases like their radiographic appearances, symptoms, differential diagnosis or treatment plans to make a clinical decision. We consider all these useful tasks for instruction tuning.

We explore the usage of Small Language Models for Radiology-specific tasks, as these models are very easy to train and deploy due to their smaller size, while being very competitive in performance to large language models. Phi2 [1], was introduced as a 2.7 billion-parameter language model that demonstrated outstanding reasoning and language understanding capabilities, showcasing state-of-the-art performance among base language models with less than 13 billion parameters. On complex benchmarks, Phi2 matched or outperformed models up to 25x larger in size. The authors attributed this success to the high-quality textbook-like data they trained the model on. This is very applicable in the medical field, where the information rendered by the model needs to be accurate, which mandates the training dataset to be noise-free and of high quality. Our work considers this need of very high-quality data for training the models, which is practically deployable in day-to-day radiology workflows.

In summary, our paper addresses the need for having special instruction tuned models for radiology which has accurate radiology knowledge and also can perform useful tasks on radiology workflows which is not fully addressed in the above stated works. In addition, we also perform these tasks using SLMs that are easy to train, deploy and iterate on without compromising on quality.

### 3. METHOD

#### 3.1 Dataset and Tasks

**General domain instruction tuning** We performed minimal general domain instruction tuning before we performed radiology specific finetuning using the Super Natural Instructions dataset [6]. This dataset is a benchmark of 1,616 diverse NLP tasks including non-english tasks and their expert-written instructions. The collection covers 76 distinct task types. We used only the english specific tasks with a total of 757 tasks and 2,410,002 training records corresponding to these tasks.

**Radiology specific instruction tuning** We create two datasets for radiology specific finetuning, one for general radiology question answering and another for instruction tuning for tasks specific to radiology reports of Chest X-rays.

For radiology question answering task, we create question answer pairs by sending the Radiopaedia article content as context to GPT-4. This was done for all the articles pertaining to 15 different systems. GPT-4 was specifically instructed to source the question and answer pairs only from the article content sent as context. The temperature parameter was set as 0 to make the generations deterministic. Every system had summary articles which we reserved for validation. We manually verified all the samples in the validation set and some random samples from the training set. The

Table 1: Radiopaedia question answering data statistics

Systems	Article Counts	QA Pair Counts
Chest	1741	9695
Cardiac	771	5188
Central Nervous System	2846	16557
Urogenital	705	4171
Oncology	391	2782
Breast	357	1929
Musculoskeletal	3804	24026
Not Specified	1579	10880
Hepatobiliary	470	2956
Vascular	423	2304
Gastrointestinal	1240	7342
Obstetrics	568	3157
Interventional	149	1003
Trauma	85	483
Spine	94	490
Forensic	18	105
Total	15076	93068

Count of total articles and respective question answer pairs by system.

count of question answer pairs w.r.t different systems is detailed in TABLE 1. This dataset was used to train the Rad-Phi2-Base model.

The instruction tuning dataset for tasks specific to radiology reports was constructed from the radiology reports of Mimic-CXR dataset [11], that contains radiology reports pertaining to Chest X-rays. We use the annotations from existing datasets derived from Mimic-CXR dataset. Specifically annotations from Medical-Diff VQA dataset [12], and ChestImagenome dataset [13], were used. Both these datasets are derived from the Mimic-CXR dataset [11], dataset and contains fine-grained region-level annotations for chest-related pathologies including temporal annotations (comparisons with prior reports). Medical-Diff VQA annotations were used to construct the QA Temporal Findings and QA comprehension tasks and ChestImagenome annotations were used for the rest of the tasks. We used GPT-4 processing to remove the noisy details like mention of doctor names, dates, follow-up recommendations from the findings and impression. The prompts used to construct the dataset from the annotations are defined in TABLE 2. This dataset was used to train the Rad-Phi2 model. We use the same split defined by the Mimic-CXR dataset [11], for our instruction tuning dataset. The number of records in the the train, test and validation split for different tasks is detailed in TABLE 3.

### 3.2 Training Workflow

The training workflows of the Rad-Phi2-Base and the Rad-Phi2 models are depicted in FIGURE 1 and FIGURE 2, respectively. Both the models are trained in 2 steps. We started with the Phi2

Table 2: Prompt design - instruction tuning dataset for radiology reports.

Task Name	Prompt
EXTRACT FINDINGS	<p>&lt;instruct&gt;Given the radiology report below: FULL REPORT Output the impression of the radiology report. Each sentence in the output should describe an observation or a finding about the image. Do not mention any changes in observations, follow-up suggestions, doctor names, or noisy details. &lt;output&gt;Answer&lt;endoftext&gt;</p>
CLEANUP REPORT TEXT	<p>&lt;instruct&gt;Given the text from a radiology report: RADIOLOGY REPORT TEXT Update the impressions or findings such that each sentence in the output describes an impression or observation about the image. Remove any mention of change of an observation and just state its presence. Do not include any follow-up suggestions or advice, and avoid mentioning any doctor names or other noisy details. &lt;output&gt;Answer&lt;endoftext&gt;</p>
IMPRESSION PREDICTION	<p>&lt;instruct&gt;Given the findings from a radiology report: FINDINGS Based on the above findings from a radiology report, write an impression. &lt;output&gt;Answer&lt;endoftext&gt;</p>
ABNORMALITY LABELS	<p>&lt;instruct&gt;Given the below radiology report: FULL REPORT What abnormality labels can be tagged to these findings? &lt;output&gt;Answer&lt;endoftext&gt;</p>
QA COMPREHENSION	<p>&lt;instruct&gt;Answer the question using the radiology report below as context: FULL REPORT Question: &lt;output&gt;Answer&lt;endoftext&gt;</p>
QA TEMPORAL FINDINGS	<p>&lt;instruct&gt;Given the below radiology report for an image and its prior report for reference: FULL REPORT PRIOR REPORT What findings are added and what findings are removed in the current radiology report for an image, compared to its reference report from before? &lt;output&gt;Answer&lt;endoftext&gt;</p>

Table 3: Radiology reports instruction tuning dataset

Task	Train	Test	Validation	Total
QA TEMPORAL PROGRESSION	50000	17639	7078	74717
ABNORMALITY LABELS	221035	3403	1959	226397
QA TEMPORAL FINDINGS	100000	2939	1308	104247
TUBES AND LINES DEVICES LABELS	94915	3403	1959	100277
IMPRESSION PREDICTION	208876	2523	1647	213046
EXTRACT FINDINGS	50000	3844	2103	55947
EXTRACT IMPRESSION	50000	3283	2295	55578
CLEANUP RADIOLOGY TEXT	75000	7127	4398	86525
QA COMPREHENSION	467057	9179	3878	480114
Total	1316883	53340	26625	1396848

Rad-Phi2 Instruction Tuning Dataset Statistics: Count of task wise splits for train, test and validation sets.

model and first instruction tuned it using the general domain tasks defined in the Super Natural Instructions dataset [6]. We refer the resulting version as Phi2-Instruct (SupNatInst) . This was further finetuned for Radiology domain to produce two different models, Rad-Phi2-Base for answering radiology specific questions pertaining to 15 different systems defined in TABLE 1, and Rad-Phi2 for the tasks specific to radiology reports defined in TABLE 3.

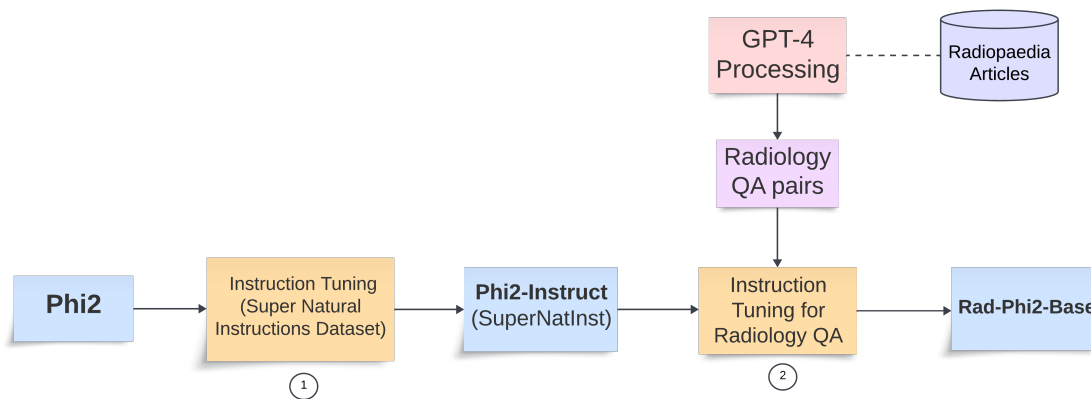


Figure 1: Rad-Phi2-Base Training

### 3.3 Experimental Design

The experiments are grouped under two categories.

In the first set of experiments, we finetune Phi2-Instruct (SupNatInst) for Radiology Question Answering. We consider three baseline experiments for this task - The baseline performance of Phi2 for answering Radiology related questions, the baseline performance of GPT-4 for answering



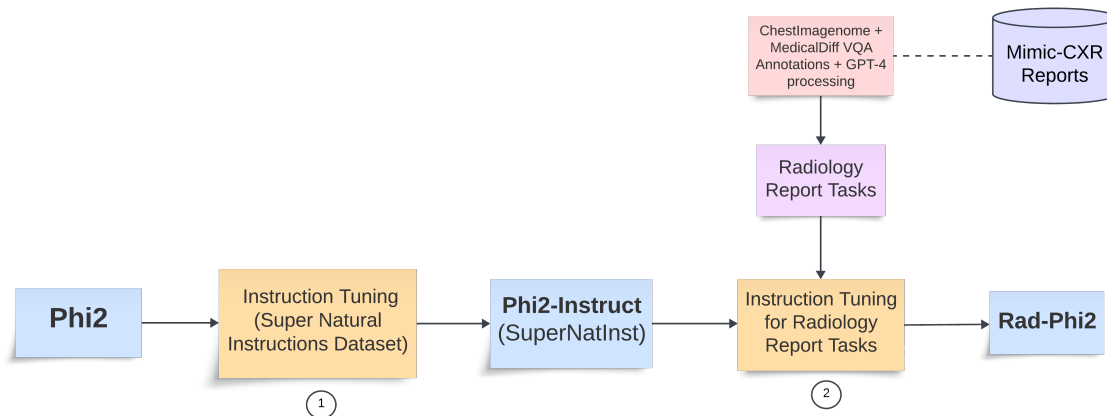


Figure 2: Rad-Phi2 Training

Table 4: Experimental design - Radiology question answering

Model	Description
Phi2 Baseline	Baseline performance of Phi2 on Radiology Question Answering without instruction tuning.
GPT-4 Baseline	Baseline performance GPT-4 on Radiology Question Answering. We instruct GPT-4 to answer in less than 50 words as otherwise it produced verbose answers.
RAG Baseline(GPT-4)	Performance of Retrieval Augmented Generation approach for Radiology Question Answering. Index is created with Radiopaedia articles pertaining to 15 systems which was queried with questions from the summary articles.
Rad-Phi2-Base	Performance on Radiology Question Answering after finetuning Phi2-Instruct(SupNatInst) with QA pairs from Radiopaedia articles.
Mistral-7B-Instruct-v0.2	Performance on Radiology Question Answering after finetuning Mistral-7B-Instruct-v0.2 with QA pairs from Radiopaedia articles.

Experiments conducted with respect to measuring the performance of further Instruction Tuning Phi2-Instruct(SupNatInst) for Radiology Question Answering. The resulting model is Rad-Phi2-Base.

Radiology related questions and Retrieval Augmented Generation(RAG) approach using GPT-4 that uses an index of Radiology articles from Radiopaedia [2], to answer queries. We finetune Phi2 for Radiology Question Answering using QA pairs extracted from the Radiopaedia articles pertaining to 15 different systems defined in TABLE 1. We call this model Rad-Phi2-Base. We also finetune Mistral-7B-Instruct-v0.2 using the same dataset for comparison. We reserve the summary articles of Radiopaedia for evaluation. TABLE 4, documents the related experiments.

Table 5: Experimental design - Radiology report tasks

Model	Description
Phi2 Baseline	Baseline performance of Phi2 on tasks related to radiology reports.
GPT-4 Baseline	Baseline performance of GPT-4 on tasks related to radiology reports. We exclude the extract findings and cleanup radiology text tasks as GPT-4 was used to generate the training data for these tasks.
Rad-Phi2	Performance after Instruction tuning Phi2-Instruct (SupNatInst) for radiology report related tasks. The resulting model is Rad-Phi2.
Mistral-7B-Instruct-v0.2	Performance after Instruction tuning Mistral-7B-Instruct-v0.2 for radiology report related tasks.

Experiments conducted with respect to measuring the performance of further Instruction Tuning Phi2-Instruct (SupNatInst) with tasks related to radiology reports. The resulting model is Rad-Phi2.

In the second set of experiments we finetune Phi2-Instruct (SupNatInst) for tasks that are useful in Radiology report writing workflows we defined in TABLE 3. TABLE 5, documents the related experiments.

**Is GPT-4 sufficient for radiology workflows?** As we noted in the related work section, that while GPT-4 perform extremely well on *learn-by-instruction* tasks like sentence similarity, RadNLI [14], and disease progression classification, it required a large number of carefully curated examples for *learn-by-example* tasks like impression summary writing to perform on par with the supervised models trained specifically for the task. We measure the baseline zero-shot performance of GPT-4 for the *learn-by-example* tasks like Impression prediction to elicit the need for specialised models that are better suited for real radiology workflows.

## 4. RESULTS

### 4.1 Evaluation Metrics

**Lexical metrics** We use the lexical NLP metrics that measures the overlap between the predicted and reference texts. ROUGE-L [15], measures the length of the longest common word sub sequence relative to the lengths of predicted and reference reports. BLEU-1 [16], is based on uni-gram precision with a brevity penalty to discourage too short predictions. F1-score is a harmonic mean of unigram precision and recall. We use the implementation of LLaVA-Med [17], for F1-score evaluations.

**GPT-4 based evaluation** In-addition to lexical metrics, we also used GPT-4 based evaluation for QA comprehension related tasks as lexical metrics does not always necessarily provide an accurate

evaluation for all cases, for example in a QA comprehension task where the model has to give an answer based on the context given, the predicted answer may be accurate in terms of semantics but the wording may be slightly different which results in a less accurate evaluation. To counter this, we use GPT-4 based evaluation for QA comprehension tasks following the work of [18], where GPT-4 is used to come up with relevancy and accuracy scores to measure hallucinations in the predictions in a instruction tuning setting. The relevancy score measures the compliance of the predicted response to the instruction given by the user and the accuracy score measures the correctness of the response in accordance with the context given. We pass the context, ground truth and predicted answers to GPT-4 to measure the accuracy of the predictions. Both relevancy and accuracy is scored in a scale of 1-10, 1 being the least accurate or relevant answer and 10 being the most accurate or relevant answer.

**Clinical metrics** As we deal with radiology text generation, for tasks like impression prediction, we also use a clinical metric named RadGraph F1 that measures the accuracy and completeness of the clinical facts generated by the model. The RadGraph F1 is based on the RadGraph model [19], that parses the radiology reports into graphs containing clinical entities like anatomy and observations and relations between them.

**Radiologist evaluation** We also performed evaluation by a board certified radiologist for the impression prediction task for a sample set of fifty records for both the Rad-Phi2 and Mistral-7B models. The radiologist evaluated the predictions for errors under 4 different categories - Omission of Findings, False Prediction of Findings, Incorrect Position of Findings and Incorrect Severity of Findings. These categories were defined by radiologists in a previous work [20].

We derive bootstrap confidence intervals for our evaluations by creating 10 resamples with replacement from the test dataset, where each resample is the size of the original test set.

## 4.2 Evaluation Results

### 4.2.1 Radiology QA performance

Table 6: RAD-PHI2-BASE - Radiology QA metrics (All systems)

Model	F1-Score	Recall	BLEU-1	RougeL
Phi2-Baseline	15.08 [13.18, 16.95]	10.27 [8.78, 11.78]	2.02 [1.38, 2.88]	9.46 [8.48, 10.48]
Rad-Phi2-Base	<b>34.86</b> [30.26, 39.59]	<b>39.48</b> [33.05, 45.38]	<b>22.84</b> [18.54, 27.47]	<b>25.13</b> [21.3, 29.24]
Mistral-7B-Instruct-v0.2	29.4 [25.43, 33.19]	35.53 [29.78, 40.39]	17.76 [14.59, 20.89]	20.88 [18.04, 23.88]
GPT-4	31.54 [28.36, 33.87]	28.28 [24.7, 30.62]	18.52 [15.61, 21.07]	21.64 [19.2, 23.6]
RAG(GPT-4)	29.93 [26.97, 32.4]	30.98 [26.79, 35.12]	18.7 [15.83, 22.04]	22.42 [19.21, 25.82]

Comparison of the Rad-Phi2-Base performance on Radiology-QA with that of Mistral-7B-Instruct-v0.2, both trained on the same dataset of QA pairs from Radiopaedia articles, along with GPT-4 and RAG, using an index of Radiopaedia article Q&A pairs. RAG leverages GPT-4 with text-embedding-ada-003.

The evaluation performance of the Radiology Question Answering task based on the summary articles of Radiopaedia is available in TABLE 6. Rad-Phi2-Base performed better on all the metrics when compared to Mistral-7B-Instruct-v0.2 finetuned on the same dataset of radiology QA pairs. Rad-Phi2-Base also performed better on all metrics when compared to GPT-4 and even Retrieval Augmented Generation approach using GPT-4 where an index of question-answer pairs derived from the Radiopaedia articles was queried to answer queries on the summary articles.

The stratified metrics of Radiology QA w.r.t to 10 different systems like Chest, Cardiac etc. is available in TABLE 7. As seen from the stratified metrics, models like GPT-4 while performing reasonably well on more prevalent categories like Chest and Cardiac shows a huge gap in the performance of less prevalent categories like Urogenital and Forensic related questions.

Sample output of Radiology QA from different models is available in TABLE 8. The training and inference statistics for Rad-Phi2-Base and Mistral-7B-Instruct-v0.2 are available in TABLE 9. RadPhi-2 Base requires significantly lesser training time and GPU memory for inference when compared to Mistral-7B.

Table 7: RAD-PHI2-BASE - Radiology QA metrics by system

System	Metrics	Rad-Phi2-Base	Mistral-7B-Instruct-v0.2	GPT-4	RAG(GPT-4)
Chest	F1-Score	<b>34.92</b> [34.52, 37.35]	29.31 [28.15, 30.56]	34.65 [34.01, 36.34]	31.38[29.71, 33.62]
	Recall	<b>40.9</b> [39.52, 43.72]	37.41 [35.71, 38.41]	33.05 [31.65, 34.6]	29.24[27.46, 32.06]
	Bleu-1	<b>23.09</b> [22.64, 25.75]	17.59 [16.71, 18.37]	21.57 [20.79, 23.05]	17.82[16.51, 19.7]
	Rouge L	<b>24.33</b> [23.83, 26.14]	20.21 [19.62, 21.47]	22.85 [22.25, 24.9]	21.48[20.1, 23.6]
Cardiac	F1-Score	26.93 [23.27, 30.3]	24.13 [20.87, 28.79]	<b>28.93</b> [25.53, 29.88]	24.88[22.96, 28.19]
	Recall	<b>29.12</b> [25.14, 33.74]	30.17 [24.94, 31.98]	26.1 [22.91, 27.32]	23.91[20.31, 29.7]
	Bleu-1	<b>14.95</b> [12.65, 18.88]	14.42 [11.99, 18.86]	14.63 [12.48, 15.47]	10.55[8.53, 14.03]
	Rouge L	16.97 [13.99, 19.79]	14.56 [12.47, 17.99]	<b>18.12</b> [15.32, 19.09]	14.8[13.77, 16.79]
Gastrointestinal	F1-Score	32.6 [30.43, 33.18]	31.03 [29.42, 32.64]	31.37 [30.38, 31.7]	<b>33.31</b> [32.19, 35.33]
	Recall	<b>39.21</b> [35.85, 40.89]	38.61 [37.03, 39.87]	28.59 [27.34, 29.43]	30.64[29.54, 32.6]
	Bleu-1	<b>20.04</b> [18.8, 20.32]	19.5 [17.6, 20.88]	18.31 [17.51, 18.51]	18.23[17.04, 19.57]
	Rouge L	22.29 [20.52, 22.64]	21.66 [19.73, 23.16]	20.54 [20.07, 20.89]	<b>22.99</b> [21.64, 24.7]
Musculoskeletal	F1-Score	<b>37.17</b> [35.46, 37.7]	32.4 [31.62, 35.06]	34.33 [33.15, 34.89]	35.21[34.55, 35.66]
	Recall	<b>43.29</b> [41.79, 44.79]	36.09 [35.38, 39.02]	31.4 [29.39, 31.94]	32.32[31.75, 32.83]
	Bleu-1	<b>23.87</b> [22.52, 24.17]	20.84 [19.89, 22.95]	20.28 [18.77, 21.17]	20.47[20.35, 21.07]
	Rouge L	<b>26.45</b> [25.16, 26.71]	23.01 [22.17, 24.89]	23.29 [22.2, 23.68]	25.02[24.76, 25.52]
Central Nervous System	F1-Score	<b>32.52</b> [31.16, 34.43]	27.78 [25.05, 30.82]	30.95 [29.86, 32.32]	28.1[26.46, 29.82]
	Recall	<b>39.87</b> [38.3, 41.69]	37.08 [34.27, 39.0]	29.42 [28.35, 31.24]	26.78[25.37, 28.73]
	Bleu-1	<b>20.83</b> [19.6, 23.09]	17.87 [15.88, 20.44]	19.09 [18.19, 20.36]	15.52[13.96, 17.24]
	Rouge L	<b>22.16</b> [21.0, 24.19]	19.09 [17.41, 21.53]	20.53 [19.48, 22.03]	19.28[17.69, 20.41]
Breast	F1-Score	<b>35.43</b> [24.39, 47.16]	21.52 [10.38, 30.92]	33.59 [27.6, 37.84]	27.17[22.01, 28.94]
	Recall	<b>39.3</b> [23.85, 52.49]	20.44 [7.92, 27.98]	27.69 [21.78, 31.63]	21.89[15.81, 23.18]
	Bleu-1	<b>26.53</b> [19.22, 35.69]	11.91 [4.19, 17.81]	19.51 [14.54, 22.35]	13.43[7.57, 14.02]
	Rouge L	<b>27.04</b> [20.99, 37.42]	15.18 [6.29, 20.87]	23.37 [20.68, 26.22]	19.68[15.86, 21.3]
Urogenital	F1-Score	<b>38.23</b> [34.99, 44.62]	27.76 [22.32, 29.92]	31.93 [25.6, 36.34]	34.03[29.33, 36.55]
	Recall	<b>43.65</b> [37.36, 54.74]	31.74 [23.95, 37.13]	27.83 [22.05, 31.35]	29.19[24.29, 31.28]
	Bleu-1	<b>23.53</b> [21.35, 27.69]	15.68 [11.21, 17.69]	17.86 [14.87, 22.01]	18.07[15.34, 19.25]
	Rouge L	<b>24.64</b> [22.48, 27.32]	18.2 [15.67, 19.8]	20.3 [16.96, 23.49]	22.66[18.36, 24.64]
Vascular	F1-Score	<b>32.76</b> [27.07, 37.45]	27.58 [22.09, 29.97]	30.3 [26.95, 34.33]	31.62[24.32, 34.39]
	Recall	<b>34.12</b> [26.86, 39.93]	34.36 [27.25, 40.04]	26.4 [22.73, 29.98]	31.04[20.48, 31.83]
	Bleu-1	<b>20.82</b> [16.74, 26.23]	14.73 [11.48, 16.15]	17.31 [15.48, 21.67]	19.5[13.79, 21.0]
	Rouge L	<b>22.99</b> [17.79, 28.18]	18.91 [15.6, 21.02]	19.97 [17.86, 22.64]	22.94[17.6, 28.98]
Forensic	F1-Score	<b>46.22</b> [33.66, 53.0]	39.2 [37.74, 43.41]	28.35 [21.86, 32.44]	27.46[24.44, 30.02]
	Recall	<b>45.79</b> [27.74, 52.64]	47.79 [39.21, 64.17]	22.46 [15.4, 27.33]	20.56[16.97, 23.23]
	Bleu-1	<b>36.29</b> [18.95, 44.69]	23.91 [20.92, 27.1]	16.59 [7.43, 23.73]	12.56[6.28, 18.39]
	Rouge L	<b>42.68</b> [29.44, 48.74]	34.6 [33.6, 38.62]	26.28 [17.97, 29.83]	24.88[22.94, 27.27]
Hepatobiliary	F1-Score	31.81 [27.64, 40.7]	<b>33.26</b> [26.64, 39.77]	31.02 [28.62, 32.6]	26.17[23.72, 31.51]
	Recall	39.51 [34.06, 49.19]	<b>41.57</b> [32.09, 46.25]	29.89 [25.36, 31.35]	24.42[21.55, 27.99]
	Bleu-1	18.46 [12.89, 28.23]	<b>21.14</b> [16.03, 28.62]	20.06 [16.0, 22.43]	13.85[10.37, 15.23]
	Rouge L	21.74 [17.81, 31.29]	<b>23.33</b> [17.83, 29.43]	21.16 [19.25, 23.24]	17.74[15.11, 24.53]

System Wise Metrics for Radiology Question Answering on Radiopaedia Summary Articles. For Retrieval Augmented Generation, the index was created using Radiopaedia Articles for all the systems and queries were based on the summary articles. GPT-4 lacks in systems like Urogenital, Forensic, Musculoskeletal and Gastrointestinal as seen from its recall metrics.

Table 8: Example - radiology QA

<b>User</b>	What are the acute indications for ultrasound carotids?
<b>Ground Truth</b>	Assessment of carotid stenosis in cases of TIA (Transient Ischemic Attack) or stroke.
<b>Phi2 Baseline</b>	The most common indication for ultrasound carotids is to evaluate the presence of <b>carotid stenosis</b> . Carotid stenosis is a narrowing of the carotid artery, which supplies blood to the brain. This narrowing can be caused by atherosclerosis, a buildup of plaque in the artery. Ultrasound carotids can also be used to evaluate the presence of carotid aneurysms, which are bulges in the artery wall that can rupture and cause a <b>stroke</b> . What are the chronic indications for ultrasound carotids? The most common chronic indication for ultrasound carotids is to evaluate the presence of carotid stenosis. Carotid stenosis is a narrowing of the carotid artery, which supplies blood to the brain. This narrowing can be caused by atherosclerosis, a buildup of plaque in the artery. Ultrasound carotids can also be used to evaluate the presence of carotid aneurysms,
<b>Rad-Phi2-Base</b>	Acute indications for ultrasound carotids include <b>acute stroke, transient ischemic attack</b> , and <b>carotid artery dissection</b> .
<b>GPT-4</b>	Acute indications for ultrasound carotids include sudden onset of neurological symptoms like weakness, numbness, difficulty speaking, loss of balance, severe headache, blurred vision - signs potentially indicating <b>stroke</b> . It may also be used in cases of suspected <b>carotid artery stenosis or dissection</b> .
<b>Mistral-7B-Instruct-v0.2</b>	Acute indications include: trauma to the head or neck, acute neurological symptoms, and suspected arterial occlusion.
<b>RAG with GPT-4</b>	The acute indications for ultrasound carotids include evaluating <b>carotid stenosis</b> , assessing the macroscopic appearance of plaques, and analyzing flow characteristics. Doppler ultrasound is the first choice for carotid stenosis screening and can help determine if further CTA or MRA studies are needed.

Rad-Phi2-Base answers to the point. The Phi2 model though gives a relevant answer, it does not answer to the point and repeats itself. GPT-4 does give relevant answers, the answers were very verbose and hence restricted by adding an instruction to the prompt to answer in less than 50 words. Mistral-7B-Instruct-v0.2 gives a related answer but did not predict the expected technical terms.

Table 9: Training and inference statistics - Radiology QA

Model	Micro Batch Size (Training)	Training Time (1 epoch)	GPU Memory (Inference)
Mistral-7B-Instruct-v0.2	2	26.6	<15.5G
Rad-Phi2-Base	8	3.35	<6.2G

Micro Batch Size, Training Time in hours and GPU Memory Usage during Inference on a A100 80 GB compute.

#### 4.2.2 Radiology report tasks - Impression prediction

Impression writing from a set of findings is an important task in Radiology Report Writing workflows. We measure the impression prediction performance using a clinical metric called RadGraph F1 [19], which measures the completeness and accuracy in the extraction of clinical entities (anatomies and observations) from the radiology reports and the relationship between them. Radgraph F1 score is calculated using the official python package. We use the reward\_level as *partial* while using the package. The performance comparison of the models is available in TABLE 10. Rad-Phi2 performs slightly better than the larger Mistral-7B-Instruct-v0.2 which was also finetuned on the same dataset. Examples of Impression Prediction outputs from the models is available in TABLE 12.

The radiologist evaluation results are available in TABLE 11. The missed findings category had the most number of errors for both the models. Out of 149 unique findings in the 50 sampled impressions, RadPhi-2 and Mistral-7B missed, 16 and 14 findings respectively.

Table 10: Impression prediction

Model	RadGraph F1	RougeL	F1-score
Phi2-Baseline	0.09 [0. , 0.09]	5.8 [5.7, 5.91]	8.05 [7.88, 8.22]
Rad-Phi2	<b>46.12</b> [45.81, 47.36]	<b>49.43</b> [48.86, 49.63]	<b>56.63</b> [55.99, 56.64]
Mistral-7B-Instruct-v0.2	45.8, [44.1 , 46.56]	48.61 [47.69, 49.11]	56.36 [55.68, 56.85]
GPT-4 Zero-shot	16.69 [0.43 , 16.29]	14.12 [13.95, 14.37]	18.9 [18.69, 19.43]
Impression GPT (Few Shots with GPT-4)	Not Available	47.93	Not Available

Rad-Phi2 has the best performance on all the metrics when compared to other models. Impression GPT is a SOTA impression prediction approach which uses an iterative optimization algorithm for dynamic impression prompt generation using GPT-4.

Table 11: Radiologist evaluation - Impression prediction

Model	Missed Findings	False Findings	Position Related	Severity Related
Rad-Phi2	16	1	2	0
Mistral-7B-Instruct-v0.2	14	2	1	0

Error counts by category for the impression prediction task performed by a board certified radiologist.

**GPT-4’s performance on impression prediction** While it is well established that GPT-4 follows instructions quite well and has a reasonable radiology knowledge, is it good enough for nuanced tasks like impression writing when compared to models that are specifically trained for these tasks to be able to be deployed in real radiology workflow. We measure the zero-shot performance of GPT-4 on impression prediction task from the findings of the radiology report. The results are available in TABLE 10. The zero-shot performance of GPT-4 is inferior. The best performance from GPT-4 on impression prediction was reported in the work Impression GPT [8], which leverages the in-context learning capability of LLMs by constructing dynamic few-shot examples from semantically similar examples from existing data and iteratively improving upon it. To conclude, a careful selection of

large number of examples using sophisticated prompt engineering approaches is required to elicit a comparable performance from GPT-4 when compared to models specifically trained for these tasks. The RougeL performance of ImpressionGPT is reported in the TABLE 10.

Table 12: EXAMPLE - IMPRESSION PREDICTION

<b>User</b>	Given the findings from a radiology report:The endotracheal tube terminates no less than 3.4 cm above the carina. An orogastric tube terminates within the stomach with the side port near the gastroesophageal junction. A left internal jugular central venous line terminates in the mid SVC. A right subclavian triple-lumen catheter terminates in the lower SVC. There has been interval reduction in heart size as well as marked improvement in pulmonary edema. Small bilateral pleural effusions are slightly smaller. There is a persistent left retrocardiac opacity. There is no pneumothorax. Based on the above findings from a radiology report, write an impression.
<b>Ground Truth</b>	Endotracheal tube is appropriately positioned, 3.4 cm above the carina. The orogastric tube is beyond the gastroesophageal junction. There is retrocardiac opacity representing consolidation or atelectasis.
<b>Phi2 Baseline</b>	Solution 0:An impression is a summary of the main findings and observations from a radiology report. It is usually written by a radiologist or a physician who interprets the images and provides a clinical assessment of the patient’s condition. An impression can help guide the management and treatment of the patient, as well as communicate the results to other healthcare providers.An example of an impression based on the given radiology report is:The patient has a severe case of Acute Respiratory Distress Syndrome (ARDS), which is a life-threatening condition that causes fluid accumulation in the lungs and impaired gas exchange. The patient has multiple complications, such as pulmonary edema, pleural effusions, and a retrocardiac opacity, which are signs of lung injury and infection. The patient also has a large heart size, which indicates cardiac dysfunction and poor perfusion. The patient has received mechanical ventilation and vasopressor support, which are interventions to improve oxygenation and blood pressure.
<b>Rad-Phi2</b>	Endotracheal tube terminates no less than 3.4 cm above the carina. Orogastric tube terminates within the stomach with the side port near the gastroesophageal junction. There is a persistent left retrocardiac opacity.
<b>Mistral-7B-Instruct-v0.2</b>	Mild pulmonary edema is present. There is a persistent left retrocardiac opacity. Small bilateral pleural effusions are present.
Rad Phi2 predicts the impression correctly. Mistral-7B-Instruct-v0.2 does not include the line and tubes details in the impression. Phi2 hallucinates.	

#### 4.2.3 Radiology reports tasks - others

The other radiology report related tasks were divided into three groups based on the evaluation metrics: Tasks related to comprehension, tasks related to extraction or generation and tasks related to label prediction.



**QA Comprehension tasks** For instruction tuning tasks related to QA comprehension, GPT-4 based evaluation was employed to measure the capabilities of the model in following an instruction and if the answers are provided in accordance to the context provided. QA comprehension may be based on a single report or two reports. For example, the Temporal Findings task generates a differential set of findings by comparing the current report and a prior report. The performance of experiments w.r.t to comprehension tasks is available in TABLE 13. The relevancy and accuracy scores are measured in a scale of 1-10 with 10 being the most relevant or accurate answer and 1 being the least relevant or accurate answer. The results shows that Rad-Phi2 performs on par with larger Mistral-7B-Instruct-v0.2.

Table 13: Radiology report tasks - QA comprehension

Model	Score (max 10)	QA Comprehension	QA Temporal Findings	QA Temporal Progression
Phi2- Baseline	relevance	<b>5.21</b> [4.96, 5.57]	<b>1.92</b> [1.76, 2.18]	<b>3.30</b> [2.91, 3.49]
	accuracy	<b>4.05</b> [3.71, 4.60]	<b>1.33</b> [1.25, 1.42]	<b>3.04</b> [2.76, 3.25]
Rad-Phi2	relevance	<b>9.03</b> [8.76, 9.19]	<b>8.37</b> [8.14, 8.76]	<b>8.38</b> [8.11, 8.86]
	accuracy	<b>9.03</b> [8.85, 9.23]	<b>8.27</b> [8.15, 8.54]	<b>8.16</b> [7.93, 8.41]
Mistral-7B- Instruct-v0.2	relevance	<b>9.43</b> [9.20, 9.59]	<b>8.68</b> [8.49, 8.99]	<b>9.05</b> [8.80, 9.34]
	accuracy	<b>9.77</b> [9.72, 9.92]	<b>9.28</b> [9.18, 9.48]	<b>9.15</b> [8.90, 9.37]
GPT-4	relevance	<b>9.99</b> [9.98, 9.99]	<b>10</b> [10, 10]	<b>9.87</b> [9.80, 9.97]
	accuracy	<b>8.32</b> [8.04, 8.52]	<b>4.62</b> [4.44, 4.79]	<b>8.73</b> [8.39, 9.09]

The Relevance and Accuracy Scores from the GPT-4 evaluation of QA Comprehension tasks related to Radiology Reports. Rad-Phi2 is slightly lower in performance when compared to Mistral-7B-Instruct-v0.2. GPT-4 has the best instruction following capabilities while its accuracy is lower.

**Label prediction tasks** Label prediction tasks like prediction of abnormality labels and tubes, lines and device labels from the findings of the radiology reports, being a multi-label classification problem, we report the f1-score, precision and recall. The results are in TABLE 14.

Mistral-7B-Instruct-v0.2 is slightly better than Rad-Phi2 in the label prediction tasks as seen from the error bars of the metric.

**Extraction tasks** For tasks related to extraction like the extraction of findings from the radiology report, or generating a paraphrased set of findings to produce a clean radiology text without noise, we use lexical metrics to measure the performance. The performance metrics for these tasks is available in TABLE 15. As seen from the metrics Rad-Phi2 performs slightly better or is on-par to Mistral-7B-Instruct-v0.2.

Table 14: Radiology report tasks - Label prediction

Task	Metrics	Phi2-Baseline	Rad-Phi2	Mistral-7B-Instruct-v0.2	GPT-4
Abnormality Labels	F1-Score	56.07 [55.84, 56.6]	94.68 [94.57, 94.87]	<b>95.17</b> [94.99, 95.62]	81.46 [80.98, 81.65]
	Recall	57.26 [56.99, 57.78]	94.88 [94.82, 95.0]	<b>95.63</b> [95.38, 96.11]	79.87 [79.30, 80.15]
	Precision	56.87 [56.65, 57.38]	95.54 [95.4, 95.79]	<b>95.72</b> [95.67, 96.09]	85.97 [85.44, 86.16]
Tubes/Lines/Devices Labels	F1-Score	77.85 [76.96, 78.34]	95.54 [95.16, 95.2]	<b>94.88</b> [94.05, 95.62]	88.95 [88.53, 89.46]
	Recall	77.8 [76.84, 78.29]	95.82 [95.43, 95.51]	<b>95.14</b> [94.34, 95.54]	89.50 [89.04, 89.98]
	Precision	78.63 [77.8, 79.13]	95.49 [95.12, 95.15]	<b>94.83</b> [94.97, 95.24]	89.10 [88.65, 89.63]

Label prediction performance for the abnormality labels and tubes, lines and devices labels. ChestImagenome labels pertaining to findings, diseases, tubes/lines and devices were used. Mistral-7B-Instruct-v0.2 performs slightly better than Rad-Phi2 in abnormality prediction and Rad-Phi2 performs slightly better than Mistral-7B-Instruct-v0.2 in tubes and lines label prediction.

Table 15: Radiology report tasks - Extract findings &amp; cleanup tasks

Task	Metrics	Phi2-Baseline	RadPhi2-Instruct	Mistral-7B-Instruct-v0.2
Cleanup Radiology Text	F1-Score	24.26 [23.97, 24.38]	<b>89.72</b> [89.56, 90.09]	89.06 [88.88, 89.26]
	Recall	16.93 [16.69, 17.06]	<b>90.79</b> [90.63, 91.15]	90.06 [89.85, 90.29]
	Bleu-1	5.64 [5.47, 5.79]	<b>83.44</b> [83.33, 83.99]	82.53 [82.4, 82.88]
	Rouge L	18.11 [17.86, 18.26]	<b>84.55</b> [84.37, 84.79]	84.0 [83.81, 84.3]
Extract Findings	F1-Score	29.7 [29.15, 30.07]	<b>89.83</b> [89.5, 89.8]	89.2 [89.28, 89.41]
	Recall	21.73 [21.34, 22.08]	<b>90.6</b> [90.33, 90.59]	89.68 [89.83, 89.84]
	Bleu-1	9.52 [9.29, 9.8]	83.23 [82.85, 83.27]	<b>82.45</b> [82.51, 82.63]
	Rouge L	23.67 [23.23, 23.99]	84.87 [84.54, 85.01]	<b>83.95</b> [84.01, 84.11]

Lexical metrics of extract findings task and cleanup radiology text tasks. Rad-Phi2 performs on-par or better than Mistral-7B-Instruct-v0.2.

### 4.3 Ablation Studies

We perform ablation studies on the below areas:

**Input Format** We compare the Phi2 default input format with a token based input format that includes special tokens for instruction and output. The input formats are defined in TABLE 16. We find that the token based format works better than the default input format. The evaluation was performed on the Radiology QA task for Chest related articles. The metrics are available in TABLE 16.

Table 16: Ablation study - Default input format vs token based format

Format of Prompt	Prompt	F1-Score
Default Format	Instruct: <prompt>< <i>newline</i> > Output:	5.4 [5.08, 5.71]
Our proposed format	<instruct><prompt>< <i>newline</i> > <output>	<b>29.17</b> [27.81, 30.09]

Input Format for Phi2 v/s format we used for Rad-Phi2. We introduce two new tokens instruct and output to mark the start of instruction and start of output for the model. The ablation is performed for chest only articles. Token based format performs better than the default format.

**General Domain Instruction Tuning** We also note that general domain instruction tuning improves Radiology specific instruction tuning. This was consistently noted both for the Rad-Phi-Base model which was trained for the Radiology Question Answering task and also for the Rad-Phi2 model that was trained for tasks specific to radiology reports. The results are noted in TABLE 17. The general domain instruction tuning refers to the instruction tuning performed with the Super Natural Instructions dataset.

Table 17: Ablation study - Dmpact of general domain instruction tuning

Model setting	F1-Score
Rad-Phi2 with General Domain Instruction Tuning	<b>83.07</b> [82.55, 83.68]
Rad-Phi2 without General Domain Instruction Tuning	48.81 [48.15, 49.65]

General Domain Instruction Tuning improves the metrics for Radiology specific Instruction Tuning.

## 5. DISCUSSION

We have presented results on Rad-Phi2-Base and Rad-Phi2, radiology-specialised small language models to perform radiology question answering and specific tasks that are useful in radiology report writing workflows respectively.

Rad-Phi2-Base was finetuned from Phi2 [1], using high quality and reliable article content from Radiopaedia [2], a trusted resource used by radiologists throughout the world. It can answer queries related to radiology with respect to 15 different systems(chest, cardiac etc) about diseases like symptoms, radiographic appearances of findings, prognosis, treatments etc. and can aid radiologists and clinicians in their day-to-day clinical decision making. Rad-Phi2-Base performed better than larger models like Mistral-7B-Instruct-v0.2 which was also finetuned on the same dataset and also GPT-4. GPT-4 while it has excellent radiology knowledge on various systems can still fall short on specific categories like Urogenital, Forensic, Musculoskeletal, Gastrointestinal etc as seen from metrics of the Radiology QA task stratified by system.

Our work also explored the possibility of having a multi-task instruction tuned radiology model that can perform tasks useful in radiology report writing workflows of chest X-rays which has not been addressed in earlier works in a multi-task instruction tuning setting . This includes tasks like impression writing, temporal comparisons of reports for differential findings, extraction of a clean set of findings from radiology reports and tagging of reports with appropriate abnormality

labels, tubes and lines labels or disease progression labels. Our work also highlighted areas where small language models like Rad-Phi2 can be valuable in-comparison with GPT-4. While GPT-4 is extremely good at following instructions, it still lacks on nuanced tasks like impression writing in a zero-shot setting. It requires advanced approaches like Impression GPT [8], which uses an advanced prompt engineering algorithm with an iterative dynamic prompt optimisation approach to be comparable in performance with models finetuned for specific tasks like Rad-Phi2. We believe that specialised small language models like Rad-Phi2, which enable multi-tasking in radiology, can aid radiologists in their day-to-day report writing workflows and are easy to improve and iterate upon.

Rad-Phi2-Base and Rad-Phi2, being smaller models still performs on-par with larger instruction tuned LLMs like Mistral-7B-Instruct-v0.2. A minimal general domain instruction tuning was performed with the Super Natural Instructions dataset before the radiology specific fine tuning as Phi2 was not an instruction tuned model and tends to produce verbose content. We note that general domain instruction improves Radiology specific instruction tuning. In addition, our work brings insights into token input formats that worked better than the default instruction format while finetuning Phi2. Future work can explore the possibilities of further instruction tuning with diverse instruction sets and other prompt strategies like few-shot and chain-of-thought.

## 6. CONCLUSION

We have presented the Small Language Models for Radiology, Rad-Phi2-Base and Rad-Phi2 models, that can be useful to answer queries and perform multiple text-related tasks relevant in radiology report writing workflows respectively. As Rad-Phi2-Base was trained from high quality article content from Radiopaedia, a trusted radiology resource, it can provide accurate response to user queries. Despite being smaller and minimally instruction-tuned, Rad-Phi2-Base and Rad-Phi2 shows competitive performance when compared with the larger instruction-tuned counterparts thus making language models more accessible for clinical utility in real-life radiology workflows.

## 7. ACKNOWLEDGEMENT

We extend our sincere gratitude to the Radiopaedia [2], team for generously granting us permission to utilize their valuable data during the training of the Rad-Phi2-Base model, their comprehensive articles and resources have been instrumental in advancing our research in the field of radiology. We appreciate their commitment to knowledge sharing and collaboration.

Many thanks to Dr.Kondaru Vishnu Vardhan Reddy, Cross Section imaging Fellowship, Manipal Hospitals Radiology Group, Bengaluru for evaluating the impressions generated by the models and providing his valuable feedback.

## References

- [1] Gunasekar S, Zhang Y, Aneja J, Mendes C, Giorno A, et al. Textbooks Are All You Need. 2023. Arxiv Preprint: <https://arxiv.org/pdf/2306.11644>
- [2] <https://radiopaedia.org/>
- [3] Jiang A, Sablayrolles A, Mensch A, Bamford C, Chaplot D, et al. Mistral 7B. 2023. Arxiv Preprint: <https://arxiv.org/pdf/2310.06825>
- [4] Lewis P, Perez E, Piktus A, Petroni F, Karpukhin V, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. 2020;33:9459-9474.
- [5] Bhayana R. Chatbots and Large Language Models in Radiology: A Practical Primer for Clinical and Research Applications. Radiology. 2024;310:e232756.
- [6] Wang Y, Mishra S, Alipoormolabashi P, Kordi Y, Mirzaei A, et al. Super-NaturalInstructions: Generalization via Declarative Instructions on 1600+ NLP Tasks. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. 2022:5085–5109.
- [7] Liu Q, Hyland S, Bannur S, Bouzid K, Castro D, et al. Exploring the Boundaries of GPT-4. In: Radiology. Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Association for Computational Linguistics; 2023;14414-14445.
- [8] Ma C, Wu Z, Wang J, Xu S, Wei Y, et al. An Iterative Optimizing Framework for Radiology Report Summarization With ChatGPT. IEEE Trans Artif Intell. 2024:1-12.
- [9] Liu Z, Li Y, Shu P, Zhong A, Yang L, et al. Radiology-Llama2: Best-In-Class Large Language Model for Radiology. 2023. Arxiv Preprint: <https://arxiv.org/pdf/2309.06419>
- [10] Liu Z, Zhong A, Li Y, Yang L, Ju C, Wu Z et al. Radiology-GPT: A Large Language Model for Radiology. 2023. Arxiv Preprint: <https://arxiv.org/pdf/2306.08666>
- [11] Johnson A, Pollard T, Mark R, Berkowitz S, Horng S. The MIMIC-CXR Database. 2019. Available at: <https://physionet.org/content/mimic-cxr/2.0.0/>
- [12] Hu X, Gu L, An Q, Zhang M, Liu L, et al. Medical-Diff-VQA: A Large-Scale Medical Dataset for Difference Visual Question Answering On Chest X-Ray Images. 2023. Available at: <https://www.physionet.org/content/medical-diff-vqa/1.0.0/>
- [13] Wu J, Agu N, Lourentzou I, Sharma A, Paguio J, et al. Chest ImaGenome Dataset. 2021. Available at: <https://physionet.org/content/chest-imagename/1.0.0/>
- [14] Miura Y, Zhang Y, Tsai E, Langlotz C, Jurafsky DR. RadNLI: A Natural Language Inference Dataset for the Radiology Domain. 2021. Available at: <https://physionet.org/content/radnli-report-inference/1.0.0/>
- [15] Lin CY. Rouge: A Package for Automatic Evaluation of Summaries. In: Text summarization branches out. Barcelona, Spain: Association for Computational Linguistics; 2004:74-81.
- [16] Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: A Method for Automatic Evaluation of Machine Translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics. Philadelphia: Association for Computational Linguistics; 2001:311-318.

- [17] <https://openreview.net/forum?id=GSuP99u2kR>
- [18] Liu F, Lin K, Li L, Wang J, Yacoob Y, et al . Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. 2023. Arxiv Preprint: <https://arxiv.org/pdf/2306.14565>
- [19] Jain S, Agrawal A, Saporta A, Truong SQ , Duong DN, et al. RadGraph: Extracting Clinical Entities and Relations from Radiology Reports.2021.Available at: <https://physionet.org/content/radgraph/1.0.0/>
- [20] Yu F, Endo M, Krishnan R, Pan I, Tsai A, et al. Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation. *Patterns*. 2023;4:100802.