# How Well Do Multi-modal LLMs Interpret CT Scans? An Auto-Evaluation Framework for Analyses

**Anonymous ACL submission**

## Abstract

Automatically interpreting CT scans can ease the workload of radiologists. However, this is challenging mainly due to the scarcity of adequate datasets and reference standards for evaluation. This study aims to bridge this gap by introducing a novel evaluation framework, named "GPTRadScore". This framework assesses the capabilities of multi-modal LLMs, such as GPT-4 with Vision (GPT-4V), Gemini Pro Vision, LLaVA-Med, and RadFM, in generating descriptions for prospectively-identified findings. By employing a decomposition technique based on GPT-4, GPTRadScore compares these generated descriptions with gold-standard report sentences, analyzing their accuracy in terms of body part, location, and type of finding. Evaluations demonstrated a high correlation with clinician assessments and highlighted its potential over traditional metrics, such as BLEU, METEOR, and ROUGE. Furthermore, to contribute to future studies, we plan to release a benchmark dataset annotated by clinicians. Using GPTRadScore, we found that while GPT-4V and Gemini Pro Vision fare better, their performance revealed significant areas for improvement, primarily due to limitations in the dataset used for training these models. To demonstrate this potential, RadFM was fine-tuned and it resulted in significant accuracy improvements: location accuracy rose from 3.41% to 12.8%, body part accuracy from 29.12% to 53%, and type accuracy from 9.24% to 30%, thereby validating our hypothesis.

## 1 Introduction

In current clinical practice, a radiologist communicates the results of an imaging exam for a patient to their referring doctor through a signed report. While reading the patient exam, the radiologist routinely use Speech Recognition Software (SRS) that converts dictated speech into text. SRS has significantly reduced the report turn-around time. However, any errors resulting from the dictation have to be corrected by the radiologists themselves, and persistent errors can negatively impact the interpretation of patient diagnoses and can have medico-legal ramifications (Smith and Berlin, 2001). These errors are most common for cross-sectional imaging (Ringler et al., 2017), such as CT and MR, and the volume of these exams has steadily increased each year (Mahesh et al., 2023). This has led to a 54-72% radiologist burn-out rate (Fawzy et al., 2023) where they are under increased pressure to deal with a substantially higher number of patients while maintaining a high level of accuracy.

To ameliorate the radiologist workload, various transformer-based approaches have been proposed to generate radiology reports in one shot (Chen et al., 2020, 2021). However, these efforts focus mainly on chest radiographs (CXR), with limited attention to CT (Ichinose et al., 2023). Developing CT-based reporting methods is challenging due to the 3D nature of CT data, computational complexity, and the factual accuracy of reporting needed. Recent advances with Large Language Models (LLMs) like GPT-4 (Achiam et al., 2023), GPT-4 Vision (GPT-4V), Gemini Pro Vision (Team Gemini et al., 2023), LLaVA-Med (Li et al., 2024), and Radiology Foundation Model (RadFM) (Wu et al., 2023) show potential for various tasks, such as taking medical exams, note-taking, and disease diagnosis (Tian et al., 2024; Nori et al., 2023; Jin et al., 2023). These multi-modal models could pre-fill the "findings" section of radiology reports for quick review by radiologists (Zhu et al., 2023).

Despite these advances, crucial factors determining their clinical use involve: (1) radiologist trust, and (2) easy interpretation and evaluation of the generated content. Current evaluation metrics, including Natural Language Generation (NLG) and Clinical Efficacy (CE) metrics, are notoriously limited (Irvin et al., 2019; Zhu et al., 2023, 2024; Jin et al., 2024) when it comes to capturing the semantic richness and clinical relevance necessary for radiology reports. Additionally, they lack the explanatory power that is required for clinical use.

In this paper, we present a novel evaluation framework to assess the capability of multi-modal LLMs to generate diagnostically accurate descriptions of CT-based findings for radiology reports. CT slices with an abnormal finding were fed to a multi-modal LLM (e.g., GPT-4V) that generated a description of the abnormality. A language-centric GPT-4 model decomposed the summary into its characteristics (body part, location, type), evaluated them against gold-standard references, and scored the description based on its clinical relevance and accuracy. Our contribution can be summarized as follows:

- We introduced a new framework named "GPTRadScore", designed to evaluate the accuracy of multi-modal LLMs in describing CT scan findings, specifically focusing on the precision of identifying body parts, locations, and types of findings.
- To validate this approach, we conducted human evaluations on 500 cases in collaboration with clinicians. Furthermore, we intend to publicly release these expert annotations (with CC-BY-NC-SA 4.0 licence) to establish a new benchmark for accuracy in future assessments.
- Four recent multi-modal LLMs were evaluated for their ability to describe CT findings.
- RadFM was fine-tuned with domain-specific data to improve its generation accuracy.

## 2 Related Works

Early efforts in extracting pathologies utilized NLG rules, which were crafted to isolate specific disease features. Notable examples include the cheXpert-labeler and NegBio (Wang et al., 2017; Peng et al., 2018; Irvin et al., 2019), both of which were employed to derive disease labels in chest X-rays. With the advent of transformer models, notably the BERT model, a more advanced solution, the cheXbert-labeler, was introduced. The cheXbert-labeler is a model specifically trained on the CheXpert dataset to perform this task.

As LLMs gain popularity, their integration into radiology becomes increasingly inevitable. These models, including multimodal LLMs, are set to assist in clinical decisions, extract information from clinical notes, and generate radiological reports, showcasing their broad utility in the field (Zhou et al., 2023b,a; Bhayana, 2024; Tian et al., 2024).

LLMs possess the advanced capability for complex reasoning, making them highly suitable for analyzing AI-generated radiological reports in comparison to ground truth. Leveraging LLMs to evaluate radiological reports harnesses their analytical power and provides a scalable solution for managing large datasets, potentially containing thousands of reports. Relying on clinicians to validate these reports is an inefficient use of their time, given their essential roles in direct patient care and decision-making. By using LLMs for initial evaluations, healthcare systems can reserve clinicians' expertise for tasks where human judgment is crucial, optimizing resources and potentially speeding up the diagnostic process.

Wang et al. (2024) recently introduced LLM-RadJudge, a method that compares the performance of various LLMs and demonstrates that using GPT-4, their proposed metric achieves evaluation consistency close to that of radiologists. Furthermore, they constructed a dataset based on LLM evaluation results and used knowledge distillation to train a smaller model, which achieves evaluation capabilities comparable to GPT-4. Similarly, Liu et al. (2024) proposed MRScore, a framework akin to LLM-RadJudge. Zhu et al. (2024) proposed a method that combines the expertise of professional radiologists with LLMs such as GPT-3.5 and GPT-4. Using In-Context Instruction Learning (ICIL) and Chain of Thought (CoT) reasoning to align LLM evaluations with radiologist standards, experimental results demonstrated greater alignment with expert evaluations, surpassing traditional NLG metrics such as BLEU, ROUGE, and METEOR.

Despite these advancements, there is still no automated system for validating the clinical accuracy of CT reports, largely because of the scarcity of high-quality datasets and the complexity of CT imaging, which involves a broad range of body parts and requires extensive anatomical knowledge. This paper introduces GPTRadScore, a novel evaluation framework that assesses the capabilities of multimodal LLMs. It uses a decomposition method based on GPT-4, which mimics clinicians' evaluation processes, comparing AI-generated descriptions with the actual ground truth across factors such as body part, location, and type.

## 3 Methods

This study introduces a novel "GPTRadScore" framework for evaluating the accuracy of multi-modal LLMs in generating clinical descriptions of CT-based findings. Figure 1 illustrates the experimental design. We break down the experimental setting into two integral steps: 1. Generating Descriptions of CT Findings: (1)Visual Context Integration: CT slices with abnormalities are marked with bounding boxes to provide clear visual context to the multi-modal LLMs. (2) Text-Based Chain-of-Thought (CoT): The multi-modal LLMs generate free-text descriptions of the abnormalities, focusing on body part, specific location, and type of finding. (3) Fine-Tuning RadFM: RadFM was fine-tuned using domain-specific data from the DeepLesion dataset to improve its accuracy in generating clinically relevant descriptions of findings. 2. Evaluation Process (GPTRadScore): GPT-4 was used to compare the generated descriptions against gold-standard report sentences. Scores were assigned based on clinical relevance and accuracy, mimicking clinician assessment.

### 3.1 Dataset

To the best of our knowledge, no publicly available dataset pairs CT exams with corresponding radiology reports for lesions. For this retrospective study, the DeepLesion dataset (Yan et al., 2018) was utilized. The dataset comprises 23,436 CT slices and 8,340 studies with reports from 3,832 patients (mean age: 51, s.d.: 17; 2,085 males). Report sentences containing prospective RECIST-based measurements, made by radiologists and referred to as "bookmarks", were extracted using regular expressions. An enclosed bounding box was also created from the prospective measurement to highlight the finding in the CT slice.

The main portion of this experiment utilizes a subset of the DeepLesion dataset (Yan et al., 2019), which comprises 496 CT volumes (496 studies) from 486 patients (mean age: 52.2, s.d.: 17.7; 294 males). The subset contained 500 lesions of various kinds (e.g., liver, kidney, bone, etc.) that were prospectively marked in 500 CT slices. The subset also provided specific characteristics of lesions that were extracted from the sentences in the radiology reports using an automated method. These included the body part where the lesion is located, the fine-grain location within that region (e.g., upper pole of left kidney), and the type of lesion attributes. As certain lesion characteristics were missed by the automated extraction, two board-certified radiologists, each with 10+ years of experience, manually reviewed and comprehensively annotated any missing lesion characteristics.

### 3.2 Visual Context Integration

In addition to the CT slice that shows the abnormal finding, a visual prompt was also provided. This prompt, in the form of a bounding box, delineated the abnormality before being input into the multi-modal LLM. The clear visual context was hypothesized to enhance the accuracy of the generated descriptions of findings.

### 3.3 Text-Based CoT

Through Text-Based CoT, the free-text abnormality description generated by a multi-modal LLM should contain the following aspects: Body Part, Location (specific), and Lesion Type. The prompt used for this task was designed to allow the model to concentrate on each aspect individually, thereby optimizing the use of its natural language generation capabilities to produce clinically relevant and informative descriptions. This approach contrasted with the one-shot methods (Wei et al., 2022) that attempted to generate entire reports in a single step without explicit intermediate reasoning.

"Body Part" is the larger anatomical region or organ of the body (e.g. liver) where the lesion or abnormality is situated. "Location" refers to the specific site within a body part (e.g., Couinaud segment 2 of liver) where the abnormality is located. "Type" includes classifications, such as nodule, mass, or enlarged lymph node. The description should be concise and clinically relevant, such that the characteristics of the findings can be pre-filled in the findings section of a radiology report.

Through experiments, it was observed that LLaVA-Med and RadFM were unable to leverage text-based CoT as shown in Figure 2. GPT-4V and Gemini Pro Vision effectively used CoT to provide detailed and relevant descriptions, demonstrating stronger comprehension skills. In contrast, LLaVA-Med, despite being tasked with using CoT, did not produce an analysis related to CoT, focusing instead on the visual elements of the scan, such as the bounding box. RadFM also showed limited capability and offered minimal output, which aligned with findings in literature (Kim et al., 2023). This

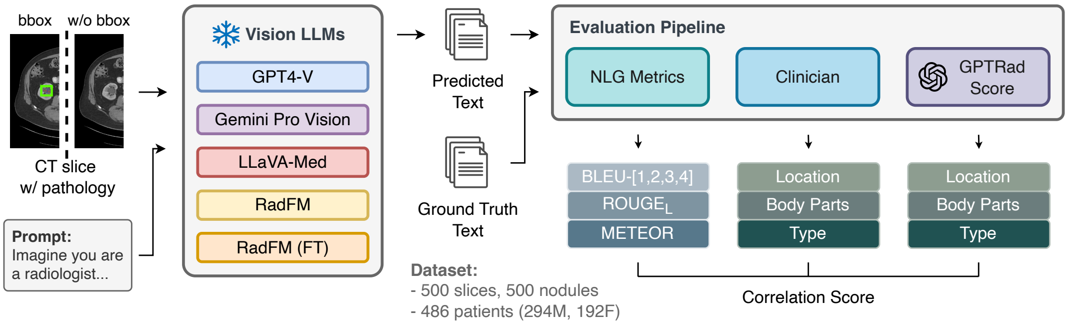
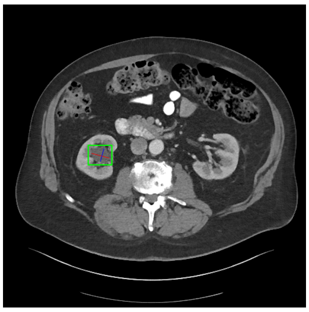


Figure 1: "GPTRadScore" framework for the auto-evaluation of LLM descriptions of CT-based findings. CT slices with outlined lesions were fed to vision-based LLMs that generated a description of the finding. They were then evaluated against the gold-standard sentences by a clinician, with NLG metrics, and auto-evaluation with GPT-4.

Case Study



**Ground Truth**: right renal parapelvic cyst.

**GPT-4V**: Location: Right abdomen, near the kidney; Body Part: Abdomen, kidney; Type: Mass; Impression: Well-circumscribed homogenous mass in the right abdomen, adjacent to the right kidney.

**Gemini Pro Vision**: Location: Right kidney; Body Part: Kidney; Type: Mass; Impression: Right renal mass, likely benign.

**LLaVA-Med**: The image is an axial computed tomography (CT) scan that has been annotated with a bounding box by a radiologist. The bounding box is a rectangular box that encloses the lesion, which is indicated in green. This helps to highlight the area of interest and provides a clear view of the lesion's location, size, and shape.

**RadFM**: { Lesion 1 }

Figure 2: Comparison of the responses from multi-modal LLMs (using CoT reasoning) for a renal cyst in the right kidney. Red, blue and purple fonts denote incorrect, correct, and uncertain descriptions respectively.

discrepancy highlighted the architectural or design limitations that hinder certain models from effectively processing input information in a sequential manner. Additional comparative analyses, including those with and without CoT, are detailed in the supplementary material.

### 3.4 Fine-Tuning RadFM

These models have not been specifically fine-tuned for lesion detection on Chest CT scans. Instead, they are often applied in a zero-shot setting, where they are expected to generalize without prior training on the specific task (Li et al., 2024). To address this, RadFM was fine-tuned using domain-specific non-overlap data from the DeepLesion subset (Yan et al., 2019), to enhance its ability to produce clinically accurate descriptions of CT findings. Fine-Tuning the model required 1 x 80GB A100 GPU, and took approx. 4 days. We believe a major reason for these issues is the lack of public datasets with paired CT studies and detailed descriptions of findings, which are essential for training effective medical imaging models, so we prepossess this dataset fist. The initial dataset, created by radiologists, was not formatted suitably for direct fine-tuning. Therefore, we utilized the GPT-4 API to systematically organize the findings; relevant descriptions of findings and measurements were extracted (sample examples are in the supplementary material). Cases lacking informative descriptions were excluded, resulting in a refined dataset comprising 17,907 descriptions linked with CT images for fine-tuning. These descriptions served as the ground truth for the fine-tuning process. The enhanced model, designated as RadFM (FT), utilized these datasets. The effectiveness of RadFM (FT) was subsequently assessed using the "GPTRadScore" evaluation framework, confirming the enhancements in its performance.

### 3.5 GPTRadScore: Evaluation using GPT-4

"GPTRadScore" is the cornerstone of our framework, leveraging GPT-4 to replicate the evaluation processes traditionally conducted by radiologists. This system assesses the ability of other multimodal LLMs to generate accurate descriptions for prospectively-identified radiological findings. Specifically, GPT-4 evaluates the accuracy of summaries provided by multimodal LLMs against gold-standard sentences derived from the DeepLesion dataset. Prioritizing criteria most significant to radiologists, the evaluation is segmented into three key aspects: body part, location, and type. For each category, GPT-4 assigned one of the following categorical scores: “Correct”, “Partially Correct”, “Incorrect”, and “Not Applicable”. A “Correct” score indicated a completely accurate interpretation. “Partially Correct” suggested the interpretation captured some aspects accurately, but lacked complete precision or detail. “Incorrect” implied the interpretation did not align with the gold-standard at any level. “Not Applicable” was used when relevant information was omitted from the description and thus evaluated was not possible. Detailed instructions provided to GPT-4, as outlined in the supplementary materials, guided the model to analyze these predictions in a manner akin to clinical judgment. Meanwhile, GPT-4 provided relevant explanations when scoring (see supplementary material) and this capability underscored the advantages of leveraging GPT-4 for complex medical evaluation tasks, where it is crucial to understand detailed anatomical context.

# 4 Experiment Setup

## 4.1 Human Evaluation Process

To assess the effectiveness of GPT-4 in automatically evaluating findings from multimodal LLMs, we need to establish a human evaluation baseline. This involves comparing AI-generated findings with the ground truth as evaluated by a human expert. We undertook a human evaluation through a structured, collaborative, and iterative process. For this analysis, we randomly selected 100 lesions from a total pool of 500 for each of the five models evaluated: GPT-4V, Gemini Pro Vision, LLaVA-Med, RadFM, and RadFM (FT). Initially, the 500 cases (100 from each model) were analyzed by three graders (PhDs) with bioinformatics backgrounds. The grading guidelines provided mirror the prompt issued to GPT-4. These assessments were subsequently reviewed and enriched with clinical insights during discussions with a clinician. Any ambiguous findings were collaboratively refined and confirmed, ensuring that the final evaluations were both scientifically robust and clinically relevant. The outcomes of this process is then compared with GPT-4’s evaluation of the same report.

## 4.2 Implementation Details

The Advanced 1.5 Pro setting of Gemini Pro Vision was used. LLaVA-Med and RadFM were run using the default configurations. For the model evaluation, we employed the Azure API for GPT-4, configured with a “temperature” of 0, “top_p” of 0.95, “max_tokens” of 4000, and the “model_version” set to “2024-02-15-preview”.

## 4.3 Metrics

The quality of the generated descriptions were initially assessed using traditional NLG metrics from Huggingface evalaute package, including BLEU, METEOR, and ROUGE. Then, following the approach suggested (Zhu et al., 2024), we conducted an auto-evaluation using GPTRadScore, where the model’s predictions were compared against gold-standard annotations. Additionally, these evaluations were compared with assessments conducted by a clinician. The Pearson’s Correlation Coefficient (Pearson, 1895) between the GPTRadScore and clinician evaluations served as an indicator of GPT-4’s reliability for auto-evaluation tasks.

# 5 Results and Discussion

## 5.1 Traditional NLG Metrics Analysis

**Results:** Table 1 evaluates multi-modal LLMs using traditional NLG metrics, and differentiates their performance in scenarios with and without bounding box constraints. RadFM (FT) bbox model exhibited outstanding performance across all NLG metrics, and particularly excelled at structural alignment and linguistic matching. GPT-4V bbox with CoT and Gemini Pro Version bbox with CoT also performed well; GPT-4V bbox with CoT achieved the highest METEOR score of 0.165. Conversely, the non-fine-tuned versions of RadFM, LLAVA-Med, and GPT-4V without CoT exhibited substantially lower performance. This decline is likely due to significant domain shift of the test dataset (radiology reports) in contrast to the model training dataset. Despite this discrepancy, the experiment was setup to expose the limitations of traditional NLG metrics.

**Limitations of traditional metrics:** While traditional NLG metrics are valuable for assessing linguistic quality, they do not fully capture the clinical

| Model | | BLEU_1 | BLEU_2 | BLEU_3 | BLEU_4 | ROUGE | METEOR |
|---|---|---|---|---|---|---|---|
| **GPT-4V** | bbox CoT | 0.164 | 0.048 | 0.015 | **0.003** | 0.171 | **0.165** |
| | w/o bbox CoT | 0.099 | 0.016 | 0.004 | 0.000 | 0.103 | 0.107 |
| | bbox w/o CoT | 0.057 | 0.007 | 0.002 | 0.000 | 0.057 | 0.146 |
| | w/o bbox w/o CoT | 0.022 | 0.002 | 0.000 | 0.000 | 0.029 | 0.071 |
| **Gemini Pro Vision** | bbox CoT | 0.160 | 0.037 | 0.007 | 0.000 | 0.180 | 0.137 |
| | w/o bbox CoT | 0.116 | 0.025 | 0.006 | 0.000 | 0.137 | 0.108 |
| | bbox w/o CoT | 0.061 | 0.014 | 0.004 | 0.001 | 0.085 | 0.140 |
| | w/o bbox w/o CoT | 0.025 | 0.005 | 0.001 | 0.000 | 0.045 | 0.075 |
| **LLAVA-Med** | bbox | 0.024 | 0.002 | 0.000 | 0.000 | 0.025 | 0.062 |
| | w/o bbox | 0.028 | 0.002 | 0.000 | 0.000 | 0.031 | 0.069 |
| **RadFM** | bbox | 0.096 | 0.011 | 0.000 | 0.000 | 0.075 | 0.095 |
| | w/o bbox | 0.094 | 0.010 | 0.001 | 0.000 | 0.075 | 0.095 |
| **RadFM (FT)** | bbox | **0.203** | **0.058** | **0.016** | 0.002 | **0.205** | 0.159 |
| | w/o bbox | 0.187 | 0.052 | 0.015 | 0.001 | 0.195 | 0.151 |

Table 1: Comparative performance of various natural language generation models using BLEU, ROUGE, and METEOR metrics, including fine-tuned RadFM (FT). "bbox" meant with bounding box, "w/o bbox" meant without bounding box. These metrics, which measured word overlap, showed low scores across the board. This suggested limitations in handling tasks that require deep contextual understanding. This highlighted the need for more sophisticated evaluation methods to gauge true performance.

relevance of the generated descriptions. In clinical settings, the priority lies in the factual accuracy and clinical relevance of generated summary over mere linguistic fidelity. This highlights the need for more robust evaluation methods that better mirror the utility in medical contexts.

### 5.2 Correlation between Clinician, GPTRadScore, Traditional Metrics

Figure 3 shows the correlation between the various metrics computed based on the ground truth evaluation, GPTRadScore, and traditional NLG metrics. Traditional NLG metrics like BLEU and METEOR demonstrate strong correlations amongst themselves (purple box), particularly at lower levels of precision like BLEU-1 and BLEU-2. This indicated a consistency in evaluating the linguistic quality of generated texts at these levels. However, at higher precision levels (BLEU-3 and BLEU-4), these correlations significantly weaken, particularly for LLaVA-Med, where scores frequently register at zero and indicate no correlation. This pattern again reflects the limitations of traditional metrics in evaluating complex sentence structures typical in radiology reports.

Furthermore, the decomposition of the description into specific aspects (location, body part, and lesion type) revealed insightful patterns (peach box). These aspects showed a lack of strong correlation with one another, and other pairings also displayed no significant correlation. These observations affirmed the efficacy of the approach in dissecting the findings into their granular elements, such that the distinct parts of report quality can be assessed independently.

Comparing traditional metrics with the ground truth evaluation showed a weak correlation, suggesting that these metrics may not serve as reliable indicators of clinical accuracy for radiology reports (blue box). This highlighted a potential gap in utilizing NLG metrics for assessing the clinical relevance of generated reports, pointing to the necessity for domain-specific evaluation methods.

Lastly, the comparison between GPTRadScore and the ground truth evaluation showed the strength of our framework (pink box), and summarized in Table 2. The results showed a strong correlation with ground truth, suggesting that GPTRadScore closely aligned with the clinical assessment paradigms utilized by radiologists. This observation underscored the potential of LLMs like GPT-4 in accurately mirroring radiologists' evaluations, offering promise for automating assessment with a high degree of fidelity to clinical standards.

| | Location | Body Part | Type | Avg. | p-value |
|---|---|---|---|---|---|
| GPT-4V | 0.86 | 0.90 | 0.84 | 0.87 ± 0.02 | <0.001 |
| Gemini Pro Vision | 0.87 | 0.91 | 0.96 | 0.91 ± 0.03 | <0.001 |
| LLaVA-Med | 0.59 | 0.83 | 0.76 | 0.75 ± 0.10 | <0.001 |
| RadFM | 0.99 | 0.92 | 0.82 | 0.90 ± 0.07 | <0.001 |
| RadFM (FT) | 0.96 | 0.83 | 0.89 | 0.89 ± 0.05 | <0.001 |

Table 2: Correlation scores between the clinician and GPT-4 grading of reports.

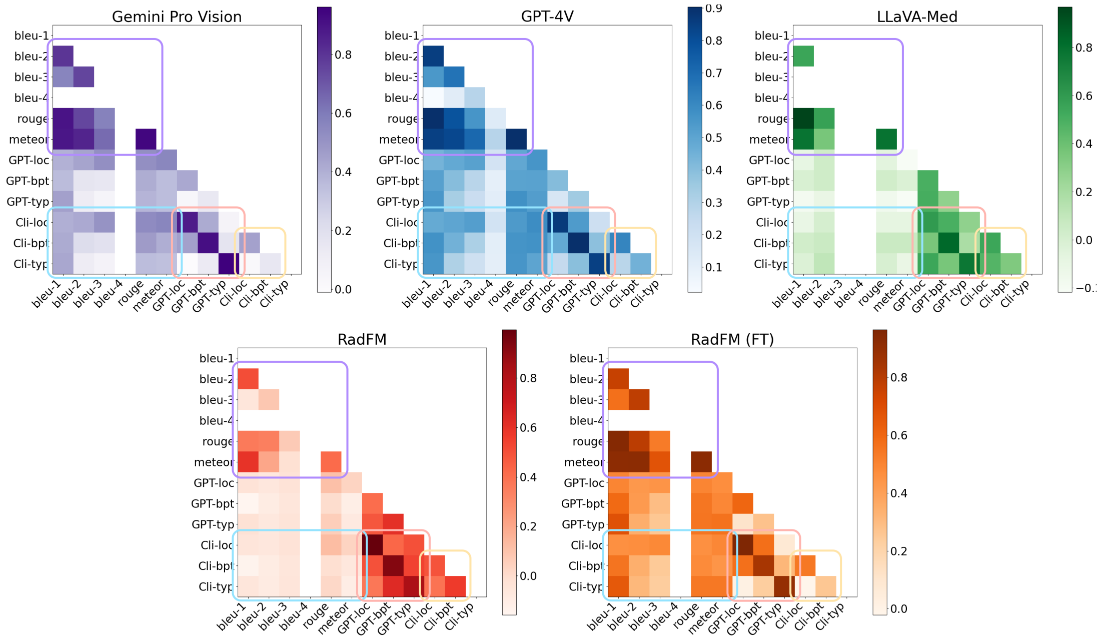


Figure 3: Heatmap of pairwise Pearson's Correlation Coefficient among various grading scores; traditional metrics, Clinician evaluations and GPTRadScore for Gemini Pro Vision, GPT-4V, LLaVA-Med, RadFM, and RadFM (FT). Color intensity indicates the strength of correlation, with darker shades representing higher correlation.

### 5.3 GPTRadScore Evaluation

Due to the strong correlation between the GPTRadScore and clinician evaluations, GPTRadScore was employed to assess the predicted findings against the ground truth for all 500 lesions for three categories: location, body part, and type of abnormality. Figure 4 displays the grading scores across four configurations: with and without lesion bounding boxes in the CT slice, and with and without text-based CoT. Notably, LLaVA-Med and RadFM do not support text-based CoT processes; thus, this figure exclusively presents the use of CoT in the GPT-4V and Gemini Pro Vision models. *Bounding boxes consistently enhanced identification of body part and location across all models, thereby indicating a dependency on strong visual cues for accurate recognition*. For example, GPT-4V and Gemini Pro Vision performed better with bounding boxes, particularly for accurate body part identification.

To better illustrate the impact of CoT reasoning, besides this figure, we have included correct classification percentages for GPT-4V and Gemini Pro Version with bounding boxes in Table 3. *CoT significantly boosts type classification accuracy*, with improvements of 28.1% for GPT-4V (from 16.5% to 44.6%) and 14.39% for Gemini Pro Vision (from 28.78% to 43.17%). However, its influence on location and body part classifications is less pronounced and may even hinder performance in some cases. This suggests that the structured reasoning provided by CoT particularly benefits complex decision-making tasks requiring nuanced interpretation and detailed contextual understanding. Conversely, for simpler tasks like identifying locations or body parts, a direct approach without CoT tends to be more effective. Overall, the enhanced performance of CoT indicates that its architecture is well-suited to sequential reasoning, akin to a radiologist's thought process, thus leading to more accurate descriptions of CT findings.

| **Category** | **GPT-4V** | | **Gemini Pro Vision** | |
|---|---|---|---|---|
| | **CoT** | **w/o CoT** | **CoT** | **w/o CoT** |
| Location | 17.1% | 13.0% | 14.02% | 17.27% |
| Body Part | 46.4% | 53.5% | 53.51% | 58.64% |
| Type | 44.6% | 16.5% | 43.17% | 28.78% |

Table 3: Correct Classification Percentages for GPT-4V and Gemini Pro Version with Bounding Boxes.

*Among the non-fine-tuned models, GPT-4V and Gemini Pro Vision excel in medical imaging tasks, which is likely attributed to their extensive pretraining on diverse datasets.* Specifically, in tasks requiring identification within bounding boxs, GPT-4V scores 46.4% in type recognition and 53.5% in body part identification. Gemini Pro Vision follows closely with scores of 44.6% and 43.17%, respectively. Despite outperforming other models, both

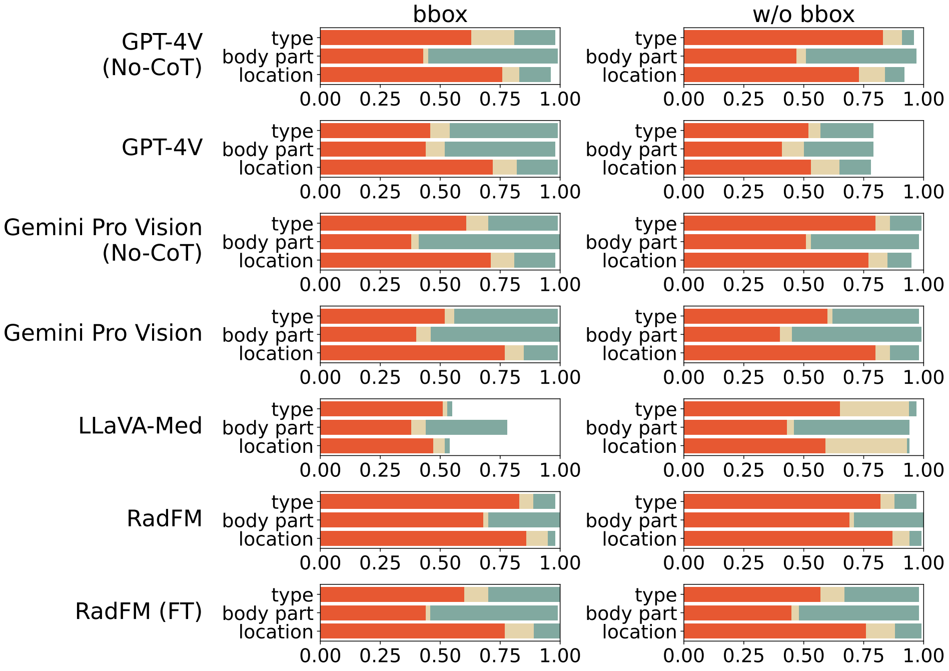


Figure 4: Comparison of results of abnormality characterization by GPT-4V, Gemini Pro Vision, LLaVA-Med, RadFM, and RadFM (FT) with bounding boxes (bbox) vs. without bounding boxes (w/o bbox). Color mapping = {orange: 'Incorrect', beige: 'Partially Correct', teal: 'Correct', white: 'Not Applicable'}. $x$-axis denotes scores $\{\, x \in \mathbb{R} \mid 0 < x < 1 \,\}$, $N$ = 500 samples.

GPT-4V and Gemini Pro Vision have room for improvement in lesion location accuracy, with scores of 17.1% and 14% respectively. In contrast, models like LLaVA-Med and RadFM demonstrate significantly weaker performance, particularly in tasks without spatial cues. For example, The LLaVA-Med bbox model achieves accuracies of 1.81% for identifying lesion locations and types, and 33.73% for recognizing body parts. These models struggle to generalize from their training data, highlighting significant challenges in adapting AI to real-world medical tasks. The suboptimal performance of models stems from the fact that these models have not been specifically fine-tuned for lesion detection on Chest CT scans (Li et al., 2024).

To prove this, we fine-tuned the RadFM model by using domain-specific, non-overlapping data. RadFM (FT) exhibited improvement across all three categories compared to the standard RadFM. Especially when bounding boxes were employed, the location accuracy rose from 3.41% to 12.8%, body part accuracy increased from 29.12% to 53%, and type accuracy improved from 9.24% to 30%. *This indicated that fine-tuning and targeted optimizations with data effectively address specific weaknesses in model performance, suggesting a pathway for further enhancing the reliability of multi-modal LLMs for medical imaging.*

## 6 Conclusion

In summary, we proposed the novel "GPTRadScore" framework for automatically evaluating AI-generated descriptions of findings prospectively identified in CT exams. These descriptions were intended to be pre-filled into the radiology reports' findings section. Four multi-modal LLMs were tested for the ability to generate a description of a CT-based finding when fed with an input CT slice. GPT-4V and Gemini Pro Vision notably outperformed other recent multi modal LLMs in accurately predicting lesion characteristics. Bounding boxes outlining the lesions in the CT slices provided strong visual cues and consistently helped these multi-modal LLMs to identify the body part and location correctly. GPTRadScore auto-evaluation results demonstrated a strong correlation with clinician assessments as measured by Pearson's correlation coefficient. Our evaluation highlighted specific weaknesses in various multi-modal LLMs, primarily due to the dataset limitations that these models were trained on. By fine-tuning RadFM on domain-specific data, significant enhancements substantially improve the utility of multi-modal LLMs in radiology.

## 7 Limitions

One limitation of our study is the lack of investigation into prompt engineering. We utilized the prompts recommended by the model developers, assuming these would optimize performance. However, more meticulously crafted prompts could potentially yield better outcomes. This reliance on predefined prompts mirrors the early days of image-based pattern recognition, suggesting that just as image recognition evolved to require less manual intervention, prompt engineering may also become more automated and effective in the future. To address this, future research could explore automated prompt generation techniques or machine learning algorithms that optimize prompt selection based on task specifics and data context.

Another limitation involves the cost and practicality of implementing such advanced AI models in clinical settings. As LLMs continue to evolve, the associated deployment costs are expected to decrease, making the technology more accessible and feasible for wider implementation. Our model shows a key direction that can be used in the future to further this progress. To mitigate high costs and enhance practicality, solutions such as developing computationally efficient models, utilizing cloud-based deployments, forming partnerships with technology providers, and initiating pilot projects could be pursued. These strategies can demonstrate the benefits of AI technologies and support broader adoption, aligning with the ongoing advancements and cost reductions in the field of LLMs.

## 8 Ethical Statement

The data used in this study are fully anonymized, and comply with the Health Insurance Portability and Accountability Act (HIPAA). Use of the data was approved by the Institutional Review Board (IRB), and the requirement for informed consent was waived. For the generation task, publicly available images from DeepLesion were used. For the evaluation task, we accessed GPT-4, an online large language model, via Microsoft Azure services to guarantee secure and privacy-compliant data handling.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rajesh Bhayana. 2024. Chatbots and large language models in radiology: A practical primer for clinical and research applications. *Radiology*, 310(1):e232756.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. Cross-modal memory networks for radiology report generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5904–5914, Online. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.

N. A. Fawzy, M. J. Tahir, A. Saeed, M. J. Ghosheh, T. Alsheikh, A. Ahmed, and Z. Lee, K. Y. Yousaf. 2023. Incidence and factors associated with burnout in radiologists: A systematic review. *European journal of radiology open*, 11:100530.

Akimichi Ichinose, Taro Hatsutani, Keigo Nakamura, Yoshiro Kitamura, Satoshi Iizuka, Edgar Simo-Serra, Shoji Kido, and Noriyuki Tomiyama. 2023. Visual grounding of whole radiology reports for 3d ct images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 611–621, Cham. Springer Nature Switzerland.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Thirty-Third AAAI Conference on Artificial Intelligence*.

Qiao Jin, Fangyuan Chen, Yiliang Zhou, Ziyang Xu, Justin M Cheung, Robert Chen, Ronald M Summers, Justin F Rousseau, Peiyun Ni, Marc J Landsman, et al. 2024. Hidden flaws behind expert-level accuracy of gpt-4 vision in medicine. *arXiv preprint arXiv:2401.08396*.

Qiao Jin, Robert Leaman, and Zhiyong Lu. 2023. Retrieve, summarize, and verify: How will chatgpt impact information seeking from the medical literature? *Journal of the American Society of Nephrology*, pages 10–1681.

Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. 2023. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

Yunyi Liu, Zhanyu Wang, Yingshu Li, Xinyu Liang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2024. Mrscore: Evaluating radiology report generation with llm-based reward system. *arXiv preprint arXiv:2404.17778*.

M. Mahesh, A. J. Ansari, and Jr Mettler, F. A. 2023. Patient exposure from radiologic and nuclear medicine procedures in the united states and worldwide: 2009-2018. *Radiology*, 301.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Karl Pearson. 1895. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. 2018. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188.

Michael D Ringler, Brian C Goss, and Brian J Bartholmai. 2017. Syntactic and semantic errors in radiology reports associated with speech recognition software. *Health Informatics Journal*, 23(1):3–13. PMID: 26635322.

John J. Smith and Leonard Berlin. 2001. Signing a colleague's radiology report. *American Journal of Roentgenology*, 176(1):27–30. PMID: 11133532.

Team Gemini, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C Comeau, et al. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 2017. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106.

Zilong Wang, Xufang Luo, Xinyang Jiang, Dongsheng Li, and Lili Qiu. 2024. Llm-radjudge: Achieving radiologist-level evaluation for x-ray report generation. *arXiv preprint arXiv:2404.00998*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Ke Yan, Yifan Peng, Veit Sandfort, Mohammadhadi Bagheri, Zhiyong Lu, and Ronald M Summers. 2019. Holistic and comprehensive annotation of clinically significant findings on diverse ct images: learning from radiology reports and label ontology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8523–8532.

Ke Yan, Xiaosong Wang, Le Lu, and Ronald M Summers. 2018. Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. *Journal of medical imaging (Bellingham, Wash.)*, 5(3):036501.

Juexiao Zhou, Xiuying Chen, and Xin Gao. 2023a. Path to medical agi: Unify domain-specific medical llms with the lowest cost. *medRxiv*, pages 2023–06.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, and Xin Gao. 2023b. Skingpt-4: an interactive dermatology diagnostic system with visual large language model.

Qingqing Zhu, Xiuying Chen, Qiao Jin, Benjamin Hou, Tejas Sudharshan Mathai, Pritam Mukherjee, Xin Gao, Ronald M Summers, and Zhiyong Lu. 2024. Leveraging professional radiologists' expertise to enhance llms' evaluation for radiology reports. *Preprint*, arXiv:2401.16578.

Qingqing Zhu, Tejas Sudharshan Mathai, Pritam Mukherjee, Yifan Peng, Ronald M. Summers, and Zhiyong Lu. 2023. Utilizing longitudinal chest x-rays and reports to pre-fill radiology reports. In *MICCAI (5)*, volume 14224 of *Lecture Notes in Computer Science*, pages 189–198. Springer.

# 9 Appendix

| Scenario | Prompt |
|---|---|
| w/o CoT, bbox | This image is with a bounding box created by a radiologist. Imagine you are a radiologist. Generate a short radiological impression based on this image. |
| w/o CoT, w/o bbox | Imagine you are a radiologist. Generate a short radiological impression based on this image. |
| CoT, bbox | Please describe this image in detail, which is with a bounding box created by a radiologist. When describing this image, please point this: 1. Location: Refers to the specific area where the lesion is found. For example: the outer edge of the lower left lung; 2. Body Part: Indicates the larger region of the body where the lesion is located. For example: lung; 3. Types, which include general terms (e.g., nodule, mass) and more specific ones (e.g., liver mass); 4. Impression: Summarize the most significant findings. |
| CoT, w/o bbox | Please describe this image in detail. When describing this image, if this image contains a lesion, please point this: 1. Location: Refers to the specific area where the lesion is found. For example: the outer edge of the lower left lung; 2. Body Part: Indicates the larger region of the body where the lesion is located. For example: lung; 3. Types, which include general terms (e.g., nodule, mass) and more specific ones (e.g., liver mass); 4. Impression: Summarize the most significant findings. |
| Evaluation | This task involved evaluating the accuracy of a predicted diagnostic interpretation (pred_result) against a ground truth description (gt_text) and its expanded terms in the context of medical imaging findings. What’s Checked: Location: If the model finds the problem in the right spot. Body Part: If the model correctly names which part of the lesion located. Type: If the model accurately describes what kind of issue it sees (like a nodule or mass). Grading Terms: Correct: The model’s guess is just right. Partially Correct: The model’s guess is somewhat right but lacks full accuracy or completeness. Incorrect: The model’s guess doesn’t match the correct answer at all. Not Applicable: The model’s guess omits relevant information and thus cannot be evaluated |

Table 4: Prompts for CT Reporting Generation and Evaluation

| Original Sentence | Current Description | Current Size |
|---|---|---|
| There is no mediastinal adenopathy however there is a nodule in the prevascular space measuring BOOKMARK ( 1.8 cm x 1.0 cm ) ( series 3 , image 88 ) . | There is no mediastinal adenopathy; however, there is a nodule in the prevascular space. | 1.8 cm x 1.0 cm |
| Smaller retroperitoneal nodules and masses for example iliac artery OTHER_BMK ( 1.6 cm x 1.4 cm ) , prior exam was OTHER_BMK ( 3.4 cm x 1.8 cm ) and left internal iliac BOOKMARK ( 2.0 cm x 1.2 cm ) , prior exam OTHER_BMK was ( 5.1 cm x 4.4 cm ) | Retroperitoneal nodules and masses, for example left internal iliac. | 2.0 cm x 1.2 cm |

Table 5: This table illustrates the methodical extraction and analysis of radiological data from medical imaging reports, specifically focusing on measurements based on the RECIST criteria, referred to as ’bookmarks.’ It details how the Current Description and Current Size are extracted from the original text using the GPT-4 API.

**gt_text**: Right posterior hemithorax mass

**pred_result**: The image is a computed tomography (CT) scan of the chest, taken during the arterial phase. It shows a mass in the right upper lobe of the lung, which appears to have lobulated contours and blurry margins. The mass has a central caliber and a diameter of around 5 cm. Additionally, there is a solid lesion measuring approximately 3 x 1.5 cm at the posterior wall of the trachea.

| Criterion | Evaluation | Explanation |
|---|---|---|
| Location | Partially Correct | While the prediction does identify a ’right’ sided issue and a finding in the ’posterior’ aspect, it describes the mass being in the ’right upper lobe’, whereas the ground truth indicates a ’right posterior hemithorax’ location. |
| Body Part | Correct | The prediction correctly identifies the ’lung’ and ’chest’ as the body parts involved, which is consistent with the ground truth. |
| Type | Correct | The prediction correctly identifies a ’mass’, which is consistent with the ground truth. |
| Attributes | Partially Correct | The prediction describes some attributes of the mass (e.g., ’lobulated contours’, ’blurry margins’, ’central caliber’, ’diameter of around 5 cm’) but these are not specified in the ground truth. The prediction is however incorrect in describing the lesion at the ’posterior wall of the trachea’ because this is not mentioned in the ground truth. |

Table 6: Evaluation from GPT-4 for LLaVA-Med Generated Findings. The *gt_text* represents the ground truth, while *pred_result* contains the sentences generated by LLaVA-Med.

**gt_text**: destructive left iliac bone ilium mass with extraosseous mass invading the adjacent iliopsoas and gluteus minimus muscles

**pred_result**: The image is a pre-operative computed tomography (CT) scan of the pelvis. It shows multiple osteolytic sacral and pelvic lesions, as well as a large anterior sacral wall defect and soft tissue invasion. The sacral region appears to have irregular bone destruction, and there are several smaller lesions scattered throughout the pelvis. The largest lesion is located in the left sacral bone, with a diameter of 27 mm.

| **Criterion** | **Evaluation** | **Explanation** |
|---|---|---|
| Location | Partially Correct | The GT mentions ’left iliac bone ilium’ while the prediction mentions the ’sacral and pelvic’ region. The iliac bone is a part of the pelvic bone, so this is partially correct. |
| Body Part | Correct | The ground truth mentions ’iliac bone ilium’ and ’iliopsoas and gluteus minimus muscles.’ The prediction mentioned ’sacral and pelvic’ bone which includes the body parts in GT, so it is correct. |
| Type | Incorrect | The prediction indicates that there are multiple osteolytic lesions while the GT mentions a destructive mass. There is a mismatch here. An osteolytic lesion is different from a destructive mass. |
| Attributes | Incorrect | While the prediction provides a measure of the ’largest lesion’ and mentions ’several smaller lesions,’ the ground truth does not mention size or number of masses but describes an ’extraosseous mass invading adjacent muscles,’ which the prediction does not detail. |

Table 7: Evaluation from GPT-4 for RadFM Generated Findings. The *gt_text* represents the ground truth, while *pred_result* contains the sentences generated by RadFM.

**gt_text**: destructive left iliac bone ilium mass with extraosseous mass invading the adjacent iliopsoas and gluteus minimus muscles

**pred_result**: The image is a pre-operative computed tomography (CT) scan of the pelvis. It shows multiple osteolytic sacral and pelvic lesions, as well as a large anterior sacral wall defect and soft tissue invasion. The sacral region appears to have irregular bone destruction, and there are several smaller lesions scattered throughout the pelvis. The largest lesion is located in the left sacral bone, with a diameter of 27 mm.

| **Criterion** | **Evaluation** | **Explanation** |
|---|---|---|
| Location | Partially Correct | The GT mentions ’left iliac bone ilium’ while the prediction mentions the ’sacral and pelvic’ region. The iliac bone is a part of the pelvic bone, so this is partially correct. |
| Body Part | Correct | The ground truth mentions ’iliac bone ilium’ and ’iliopsoas and gluteus minimus muscles.’ The prediction mentioned ’sacral and pelvic’ bone which includes the body parts in GT, so it is correct. |
| Type | Incorrect | The prediction indicates that there are multiple osteolytic lesions while the GT mentions a destructive mass. There is a mismatch here. An osteolytic lesion is different from a destructive mass. |
| Attributes | Incorrect | While the prediction provides a measure of the ’largest lesion’ and mentions ’several smaller lesions,’ the ground truth does not mention size or number of masses but describes an ’extraosseous mass invading adjacent muscles,’ which the prediction does not detail. |

Table 8: Evaluation from GPT-4 for RadFM Generated Findings. The *gt_text* represents the ground truth, while *pred_result* contains the sentences generated by RadFM.

| Model | gt_text | CoT | w/o CoT |
|---|---|---|---|
| Gemini | right renal hypodensity; cyst | Location: Right kidney; Body Part: Kidney; Type: Cyst; Impression: Right renal cyst. | There is a 5 mm non-obstructing calculus in the mid ureter on the right. |
| GPT-4V |  | Location: Right lower quadrant of the abdomen, Body Part: Appendix, Type: Appendicolith, Impression: 'Hyperdense focus within the appendix suggestive of appendicolith | There is a well-circumscribed, low-density lesion in the left kidney, which is highlighted by the bounding box. This finding is suggestive of a simple renal cyst. No features suggestive of malignancy are noted in the visualized portion of the abdomen. Further evaluation with ultrasound can be done for confirmation and to exclude any complex features. |
| Gemini | large necrotic right hepatic mass | Location: Right lower quadrant of the abdomen; Body Part: Abdomen; Type: Mass; Impression: Right lower quadrant abdominal mass, likely representing an abscess. | There is a 3.8 x 3.3 x 4.1 cm enhancing mass in the mid right kidney. The mass demonstrates central low attenuation. There is no evidence of perinephric extension or lymphadenopathy." |
| GPT-4V |  | Location: Right upper quadrant of the abdomen, Body Part: Liver, Type: Lesion, Impression: Well-circumscribed hypodense lesion in the right lobe of the liver. | The axial CT image of the abdomen with contrast demonstrates a well-circumscribed lesion in the left kidney, as demarcated by the bounding box. The lesion shows homogeneous density and does not appear to involve the renal hilum or adjacent structures. There are no signs of local invasion or distant metastasis. Further evaluation with dedicated renal protocol imaging and potential urology consultation is recommended for characterization and management of the renal lesion. No other significant abnormalities are noted in the visualized portions of the abdomen. |

Table 9: Generated descriptions with and without the CoT by Gemini and GPt-4V models.