XrayGPT: Chest Radiographs Summarization using Large Medical Vision-Language Models

Anonymous ACL submission

Abstract

The latest breakthroughs in large language models (LLMs) and vision-language models (VLMs) have showcased promising capabilities toward performing a wide range of tasks. Such models are typically trained on massive datasets comprising billions of image-text pairs with diverse tasks. However, their performance on task-specific domains, such as radiology, is still under-explored. While few works have recently explored LLMs-based conversational medical models, they mainly focus on text-based analysis. In this paper, we introduce XrayGPT, a conversational medical vision-language (VLMs) model that can analyze and answer open-ended questions about chest radiographs. Specifically, we align both medical visual encoder with a fine-tuned LLM to possess visual conversation abilities, grounded in an understanding of radiographs and medical knowledge. For improved alignment of chest radiograph data, we generate 217k interactive and high-quality summaries from free-text radiology reports. Extensive experiments are conducted to validate the merits of XrayGPT. To conduct an expert evaluation, certified medical doctors evaluated the output of our XrayGPT on a test sub-set and the results reveal that more than 70% of the responses are scientifically accurate, with an average score of 4/5. We hope our simple and effective method establishes a solid baseline, facilitating future research toward automated analysis and summarization of chest radiographs. Code, models, and instruction sets will be publicly released.

1 Introduction

The Large-scale Vision-Language models have emerged as a transformative area of research at the intersection of computer vision and natural language processing, enabling machines to understand and generate information from both visual and textual modalities. These models represent a significant advancement in the field, bridging the gap between visual perception and language comprehension, and have demonstrated remarkable capabilities across various tasks, including but not limited to image captioning (Hossain et al., 2019), visual question answering (Lu et al., 2023), and visual commonsense reasoning (Zellers et al., 2019). Training these models requires vast amounts of image and text data, enabling them to learn rich representations that capture the intricacies of both modalities. Additionally, fine-tuning can be employed using task-specific data to better align the models with specific end tasks and user preferences. Recently, Bard and GPT-4 have demonstrated impressive capabilities in various tasks, raising excitement within the research community and industry. However, it is important to note that the models of Bard and GPT-4 are not currently available as open-source, limiting access to their underlying architecture and implementation details.

Recently, Mini-GPT (Zhu et al., 2023) demonstrates a range of impressive capabilities by aligning both vision and language models. It excels at generating contextual descriptions based on the given image. However, it is not as effective in medical scenarios due to the significant differences between medical image-text pairs and general web content. Adopting vision-text pre-training in the medical domain is a challenging task because of two factors: (1) Lack of data, Mini-GPT has trained the projection layer on a dataset of 5M image-text pairs, while the total number of publicly available medical images and reports is orders of magnitude below. (2) Different modalities and domains, while Mini-GPT may involve distinguishing between broad categories like "Person" and "Car" the distinctions within medical domains are much more subtle and fine-grained. For instance, differentiating between terms like "Pneumonia" and "Pleural Effusion" requires more precision by capturing and aligning relevant medical domain knowledge.

Chest radiographs are vital for clinical decisionmaking as they offer essential diagnostic and prognostic insights about the patients. Text summarization tasks can partially address this challenge by providing meaningful information and summaries based on the given radiology reports. In our approach, we go beyond traditional summarization techniques by providing concise summaries that highlight the key findings and the impression based on the X-ray. Additionally, our model allows for interactive engagement, enabling users to ask follow-up questions based on the provided answers. We argue that based on the visual and large language models, the majority of knowledge acquired during the pertaining stage of these models requires a domain-specific high-quality instruction set derived from task-specific data to achieve promising results. The main contributions of our work are:-

- We generate interactive and clean summaries (217k) from free-text radiology reports of two datasets: MIMIC-CXR (Johnson et al., 2019) and OpenI (Demner-Fushman et al., 2015). These summaries serve to enhance the performance of our XrayGPT by fine-tuning the linear transformation layer on high-quality data.
- To obtain a medical LLM, we fine-tuned a standard LLM (Vicuna) on medical data (100k real conversations) and 20k radiology conversations to acquire domain-specific features.
- In our XrayGPT, the frozen specialized medical visual encoder is aligned with a fine-tuned medical LLM, using a simple linear transformation to understand medical meanings and acquire visual conversation capabilities.
- We conduct experiments including an evaluation study through certified medical doctors to validate the merits of our XrayGPT. To promote future research, our codebase, fine-tuned models, and high-quality instruction set along with the recipe for data generation and model training will be publicly released.

2 Related Work

Medical Chatbots: Recent medical chatbots have emerged as valuable tools in healthcare, providing personalized support and assistance to patients and professionals. The recently introduced Chatdoctor (Li et al., 2023), a next-gen AI doctor based on LLaMA (Touvron et al., 2023), aims to be an intelligent healthcare companion, answering patient queries and offering personalized medical advice. After success of ChatGPT (OpenAI, 2022), GPT-4 (OpenAI, 2023) and other open source LLM's (Touvron et al., 2023; Chiang et al., 2023; Taori et al., 2023), many medical chatbots were introduced recently such as Med-Alpaca (Han et al., 2023), PMC-LLaMA (Wu et al., 2023), and DoctorGLM (Xiong et al., 2023). These models utilize open-source LLMs and are finetuned on medical instructions, demonstrating the potential of chatbots to enhance patient engagement and health outcomes with personalized interactions.

Large Vision-Language Models: A significant area of research in natural language processing (NLP) and computer vision is the exploration of Large Vision-Language Model (VLM) learning techniques. This VLM aims to bridge the gap between visual and textual information, enabling machines to understand and generate content that combines both modalities. Recent studies have demonstrated the potential of VLM models in various tasks, such as image captioning (Zhu et al., 2023), visual question answering (Bazi et al., 2023; Liu et al., 2023; Muhammad Maaz and Khan, 2023), and image generation (Zhang and Agrawala, 2023).

3 Method

XrayGPT is an innovative conversational medical vision-language model designed for analyzing chest radiographs. Our approach draws inspiration from the design of vision-language models in general, but with a specific focus on the medical domain. Due to the limited availability of medical image-summary pairs, we adopt a similar methodology by building upon a pre-trained medical vision encoder (VLM) and medical large language model (LLM), as our foundation. The fine-tuning process involves aligning both modalities using high-quality image-summary pairs through a simple transformation layer. This alignment enables XrayGPT to possess the capability of generating insightful summaries about chest radiographs.

3.1 Model Architecture

We show in Fig. 1 an overview of our XrayGPT. Given the X-ray, we align both visual features and textual information from a pre-trained medical vision encoder (VLM), and medical large language model (LLM). Specifically, we utilize MedClip (Wang et al., 2022) as a visual encoder and our large language model (LLM) is built upon the recent Vicuna (Chiang et al., 2023).

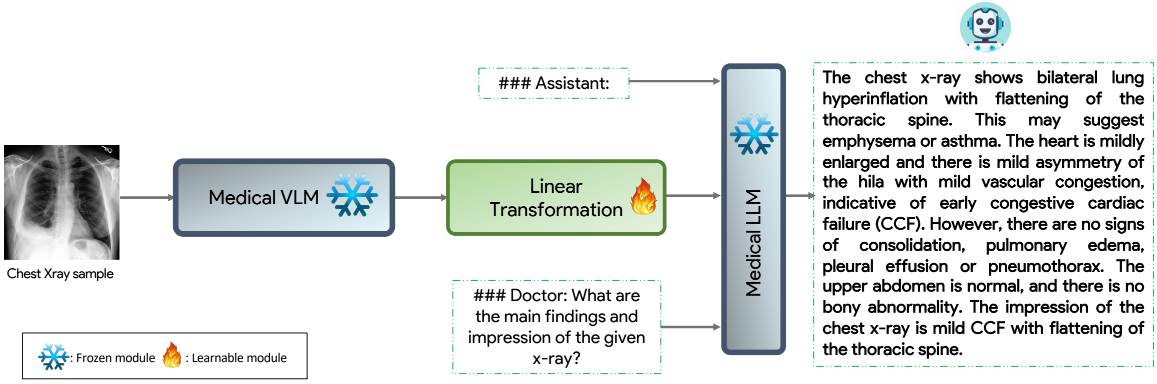

Figure 1: Overview of our XrayGPT framework. The input X-ray is passed sequentially to three components. (i) Frozen medical visual encoder to extract relevant features pertaining to the chest diagnosis. (ii) Learnable linear transformation layer to align the medical visual features with the Medical LLM to learn medical visual-text alignment. (iii) Frozen Medical LLM to generate X-ray summary based on encoded features and the given prompt.

Given X-ray $\mathbf{x} \in R^{H \times W \times C}$, the visual encoder is used to encode the image into embeddings using a vision encoder E_{img}. Then, the raw embeddings are mapped to an output dimension of 512 using a linear projection head f_v.

$$\mathbf{V}_p = f_v(E_{img}(\mathbf{x})) \quad (1)$$

To bridge the gap between image-level features and the language decoder's embedding space, we employ a trainable linear transformation layer, denoted as t. This layer projects the image-level features, represented by $\mathbf{V}_p$, into corresponding language embedding tokens, denoted as $\mathbf{L}_v$:

$$\mathbf{L}_v = t(v_p), \quad (2)$$

We utilize pre-determined prompts in the given format ###Doctor: <Img><ImageFeature></Img> <Instruction> ###Assistant:

###Doctor, corresponds to the prompt. ###Assistant, serves the purpose of determining the system role, which in our case is defined as "*You are a helpful healthcare virtual assistant*." <Instruction> refers to a randomly selected instruction from our pre-defined set. Both text queries undergo tokenization, resulting in dimensions represented by $\mathbf{L}_t$. Finally, $\mathbf{L}_v$ is concatenated with $\mathbf{L}_t$ and fed into the medical LLM, fine-tuned Vicuna, which generates the summary of the chest x-ray.

Our XrayGPT follows a two-stage training approach. In the first stage, we pre-train the model using high-quality curated interactive summaries of the training set of MIMIC-CXR (Johnson et al., 2019) reports. While in the second stage, we use the curated interactive summaries of OpenI (Demner-Fushman et al., 2015) reports.

3.2 Image-text alignment

To align the generated high-quality summaries with the given x-ray, we use similar conversational format of the Vicuna (Chiang et al., 2023) language model as follows:

###Doctor: $X_R X_Q$ *###Assistant:* X_S

where X_R is the visual representation produced by the linear transformation layer for image X, X_Q is a sampled question (e.g. What are the main findings and impression of the given X-ray?), and X_S is the associated summary for image X.

4 Curating high-quality data

Datasets: The MIMIC-CXR consists of a collection of chest radiographs associated with free-text radiology reports. It consists of 377,110 images and 227,827 associated reports, which are used for both training and testing purposes. The dataset is de-identified by removing the health information to satisfy health insurance and privacy requirements. The OpenI dataset is a collection of chest X-ray images from the Indiana University hospital network, composing 6,459 images and 3,955 reports.

High Quality and Interactive Summaries: To generate concise and coherent medical summaries from the unstructured reports, we perform the following pre-processing steps for both datasets: (1) Removal of incomplete reports lacking finding or impression sections. (2) Elimination of reports that have finding sections containing less than 10 words. (3) Exclusion of reports with impression sections containing less than 2 words.

In addition, utilizing the power of gpt-3.5-turbo model, we further implement the following pre-processing techniques to ensure high-quality summaries per image: (1) Elimination of sentences containing comparisons to the patient's prior medical history. (2) Removal of de-defined symbols "__", while preserving the original meaning. (3) As our training relies on image-text pairs, we excluded the provided view from the summary. (4) We combine the clean findings and impressions to generate an interactive and high-quality summary.

Following these steps, we obtained a set of filtered training reports consisting of 114,690 reports associated with 241k training images based on Mimic-CXR dataset. Also, we obtain 3,403 high-quality summaries that used for training based on the OpenI dataset. We show an example before and after our pre-processing in Appendix A.1.

5 Experiments

5.1 Implementation Details

Stage-1 Training: The model is designed in this stage to gain understanding of how Xray image features and corresponding reports are interconnected by analyzing a large set of image-text pairs. We employ our high-quality interactive report summaries as described in sec. 4 of MIMIC-CXR (Johnson et al., 2019) train set with 213,514 image-text pairs. The model is trained for 320k steps with a total batch size of 128 using 4 AMD MI250X GPUs. **Stage-2 Training:** The pretrained model of stage-1 is fine-tuned on 3k highly curated image-text pairs from OpenI dataset (Demner-Fushman et al., 2015). The total training steps are 5k, with a total batch size of 32 using single AMD MI250X GPU.

5.2 Evaluation Metrics

We use Rogue Score as an evaluation metric to assess the contributions of our components over the baseline Mini-GPT (Zhu et al., 2023). The Rogue Score serves as a valuable quantitative measure to assess the performance of different text generation models with the ground truth. Then, we use GPT-based evaluation schema to assess the quality and coherence of the text generated by our approaches, compared to the baseline. Furthermore, we provide certified medical doctors evaluation for 50 samples derived from the MIMIC-CXR testing set.

5.3 Results

In this section, we highlight a key contribution of our XrayGPT compared to our baseline (Zhu et al., 2023). We conduct quantitative evaluation using advanced metrics such as Rogue score and GPT-based evaluation as described in sec. 5.2. Tab. 1 shows comparison of our key components when progressively integrated into our baseline (Zhu et al., 2023) frame. From Tab. 1 our XrayGPT (row 4) has a significant improvement of 19% over the state-of-the-art baseline (Zhu et al., 2023) on the MIMIC-CXR test set. Also, we did LLM's based evaluation by asking ChatGPT model to choose "which response is closer to reference between baseline vs XrayGPT" where our model scored 82% compared to baseline 6% showing the superiority of our XrayGPT for radiology-specific summary.

Method	R-1	R-2	R-L
Baseline	0.1313	0.0221	0.0879
+ MedCLIP	0.1517	0.0308	0.0973
+ MedVicuna	0.2099	0.0551	0.1284
+ RadVicuna	**0.3213**	**0.0912**	**0.1997**

Table 1: Comparison of our XrayGPT components with Baseline (Zhu et al., 2023) using Rogue scores on MIMIC-CXR (Johnson et al., 2019) Test set. Our approach outperforms the recent Minigpt-4 with an absolute gain of 19% in terms of R-1 score.

To assess the responses of XrayGPT scientifically, we asked certified medical doctors to evaluate the responses of our model alongside the baseline, compared to their real findings and impression. This evaluation shows that our XrayGPT has accurate output in 72% of the cases, with an average score of 4/5, outperforming the baseline which achieves only 20%, with an average score of 2/5. Both models provided inaccurate responses in 8% of the cases. Despite occasional inaccuracies, XrayGPT significantly improves upon the baseline, highlighting its potential in assisting radiologists with chest radiograph analysis. Additional details of our evaluations are in Appendix (A.2,A.3). We also show qualitative examples in Appendix A.4.

6 Conclusion

To conclude, we introduce XrayGPT, an innovative medical vision-language model that combines both vision-language modalities to summarize and answer inquiries regarding chest radiographs. By aligning both modalities and leveraging our proposed interactive summaries derived from free-text radiology reports, XrayGPT demonstrates exceptional visual conversation abilities grounded in a deep understanding of chest radiographs.

7 Limitations

While our XrayGPT shows promise toward constructing conversational VLMs for chest radiograph summarization, we acknowledge some limitations here. We observe some potential limitations in generation of responses when presented with side views, as the majority of the trained images primarily consist of frontal views. To address this limitation, our potential future research direction is to focus on enhancing the quality and reliability of the model when handling multiple views of X-rays. Our current model is specifically trained and designed to answer questions pertaining to chest radiographs. Expanding the model's support to encompass multiple modalities is a potential research direction. By broadening its capabilities, the resulting potential model is expected to possess certain capabilities across various medical imaging domains beyond chest radiographs.

References

Yakoub Bazi, Mohamad Mahmoud Al Rahhal, Laila Bashmal, and Mansour Zuair. 2023. Vision–language model for visual question answering in medical imagery. *Bioengineering*, 10(3):380.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. 2015. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexander Löser, Daniel Truhn, and Keno K Bressem. 2023. Medalpaca–an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*.

MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CsUR)*, 51(6):1–36.

Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*.

Yunxiang Li, Zihan Li, Kai Zhang, Ruilong Dan, and You Zhang. 2023. Chatdoctor: A medical chat model fine-tuned on llama model using medical domain knowledge.

Yunyi Liu, Zhanyu Wang, Dong Xu, and Luping Zhou. 2023. Q2atransformer: Improving medical vqa via an answer querying decoder. *arXiv preprint arXiv:2304.01611*.

Siyu Lu, Yueming Ding, Mingzhe Liu, Zhengtong Yin, Lirong Yin, and Wenfeng Zheng. 2023. Multiscale feature extraction and fusion of image and text in vqa. *International Journal of Computational Intelligence Systems*, 16(1):54.

Salman Khan Muhammad Maaz, Hanoona Rasheed and Fahad Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *ArXiv 2306.05424*.

OpenAI. 2022. Chatgpt.

OpenAI. 2023. Gpt-4 technical report.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. 2022. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Pmc-llama: Further finetuning llama on medical papers. *arXiv preprint arXiv:2304.14454*.

Honglin Xiong, Sheng Wang, Yitao Zhu, Zihao Zhao, Yuxiao Liu, Qian Wang, and Dinggang Shen. 2023. Doctorglm: Fine-tuning your chinese doctor is not a herculean task. *arXiv preprint arXiv:2304.01097*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731.

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Appendix

A.1 Example for generating interactive summary

Input findings: *PA and lateral views of the chest were provided demonstrating no focal consolidation, effusion, or pneumothorax. Cardiomediastinal silhouette appears normal and stable. There is a compression deformity involving a mid thoracic vertebral body, which appears new from the prior chest radiograph of ___. No free air below the right hemidiaphragm. There are tiny surgical clips in the left base of neck, likely indicating prior thyroid surgery.* **Input Impression**: *No acute intrathoracic process. Interval development of a mid thoracic spine compression fracture.* **High-quality and interactive summary**: *The chest x-ray findings reveal no evidence of focal consolidation, effusion, or pneumothorax. The cardiomediastinal silhouette appears stable and normal. There is a newly developed mid thoracic spine compression fracture but no free air below the right hemidiaphragm. The presence of surgical clips in the left base of the neck suggests prior thyroid surgery. The impression suggests that there is no acute intrathoracic condition detected in the x-ray aside from the new development of mid thoracic spine compression fracture.*

A.2 LLM based Evaluation

LLM-based evaluation represents a comprehensive and meticulous approach to assessing and analyzing language models such as GPT-3.5 (OpenAI, 2022, 2023). This evaluation methodology aims to evaluate the language model's performance, capabilities, and constraints when generating text that is both coherent and contextually appropriate. The evaluation process employs a diverse range of techniques, including automated metrics like perplexity and fluency scores, as well as human evaluations that rely on expert judgments and comparisons with reference texts. By incorporating quantitative and qualitative measures, LLM-based evaluation offers valuable insights into the language model's grasp of language comprehension, coherence, factual accuracy, and its ability to produce responses that align with the given context. Through this robust evaluation framework, we can continuously enhance and refine language models by addressing potential biases, improving response quality, and maximizing their practicality across various language tasks and domains. We used GPT-3.5 (OpenAI, 2022) Turbo for evaluation of our baseline (Zhu et al., 2023) vs XrayGPT generated responses with the following meta details.

System Role: You are a chest radiologist that evaluates the response of two models: Model_1 and Model_2 and say which one is closer to the ground truth. You should print which model is closer to the ground truth.

User Role: Perform the following task: **[1]** Which model (Model_1 or Model_2) is closer to the Ground_Truth based on the medical finding and impression? Ground_Truth: <response> Model_1: <response> Model_2: <response>. **[2]** Output should be in valid JSON format without explanation where key=answer and value should be Model_1 or Model_2.

A.3 Evaluation from Certified Medical Doctors

In order to comprehensively evaluate the quality and performance of XrayGPT-generated samples, we conducted an extensive assessment in collaboration with certified medical doctors, utilizing samples from MIMIC-CXR test split. During the evaluation process, we carefully curated summaries generated by two different methods: our baseline (Zhu et al., 2023) method and the XrayGPT model. These summaries were then presented to the certified medical doctors, who were tasked with determining which response provided a relatively superior summary and diagnosis for the given X-ray image. To further gauge the medical doctor's evaluation, we requested them to assign a score ranging from 1 to 5 to each response, indicating the perceived quality of the summaries. This detailed evaluation process allowed us to gather valuable insights into the comparative performance and effectiveness of the baseline (Zhu et al., 2023) method and the XrayGPT model in generating accurate and informative summaries for X-ray images.

Our scoring criteria is defined as follows,

- 1 : poor description
- 2 : fair description
- 3 : moderate description
- 4 : good description
- 5 : excellent description

During the evaluation process, medical doctors also identify and address limitations and challenges associated with our XrayGPT-generated X-ray image summaries. They consider scenarios where our XrayGPT method may fail to detect subtle abnormalities, mischaracterize findings, or provide inaccurate recommendations. The medical doctors explore the causes of such limitations, such as dataset biases, rare or unusual cases not adequately represented during training, or technical factors that may impact the XrayGPT system's performance. Feedback from medical doctors regarding these limitations was utilized to refine and improve our XrayGPT.

A.4 Qualitative examples

We present a series of illustrative examples that highlight the remarkable capabilities of our XrayGPT. In Fig. 2, our XrayGPT effectively summarizes the contents of a provided chest x-ray image, extracting essential findings and generating an impression. While in Fig. 3, the model demonstrates its proficiency in follow-up procedures by recommending appropriate treatments based on the given case. In Fig. 4, the model accurately identifies key findings in the x-ray and provides correct responses to user queries. Notably, this particular case reveals the presence of lung hyperinflation. Fig. 5 serves as a testament to the versatility of XrayGPT, showcasing its ability to address multiple subsequent questions and suggest treatment options for pleural effusion. Lastly, in Fig. 6, the model showcases its competence in identifying various abnormalities within a chest x-ray while also offering medical treatment recommendations for pulmonary edema. Through these diverse examples, we demonstrate how XrayGPT consistently delivers accurate and comprehensive analyses, positioning it as a valuable tool in the field of medical imaging.

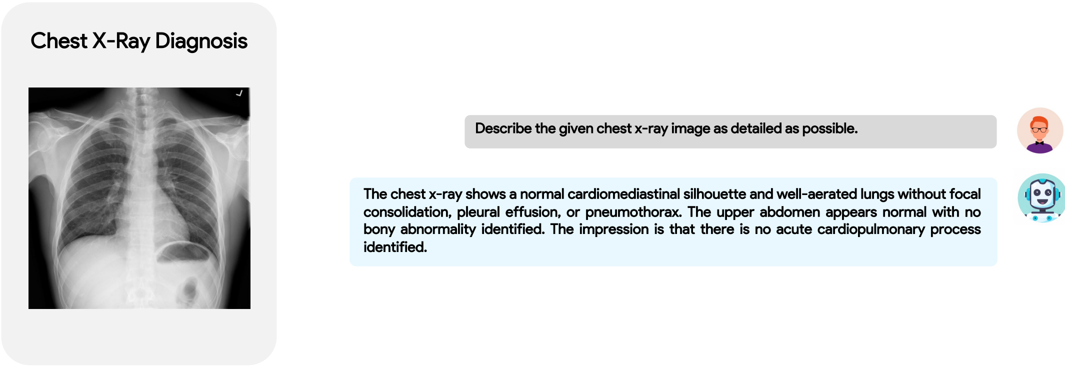

Figure 2: **XrayGPT: Delivering Comprehensive and Succinct Findings and Impressions for Every X-ray Image.** Our XrayGPT provides detailed radiology specific summary as well as concise findings and impression for given xray image.

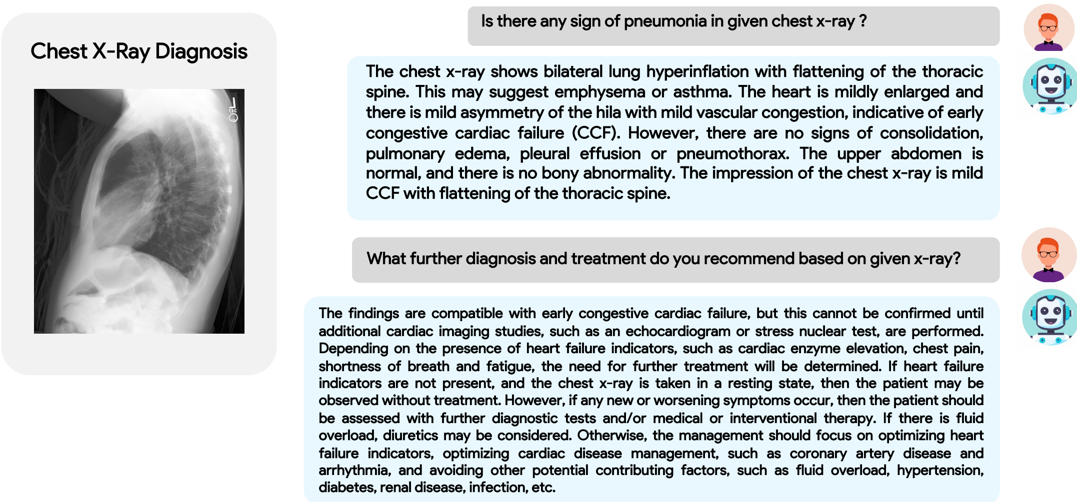

Figure 3: **XrayGPT: Empowering Users with Concise Insights - Explore the Comprehensive Analysis Derived from the xray Image.** Our XrayGPT provides a brief explanation for users quaries based on the overall findings from the given xray image.

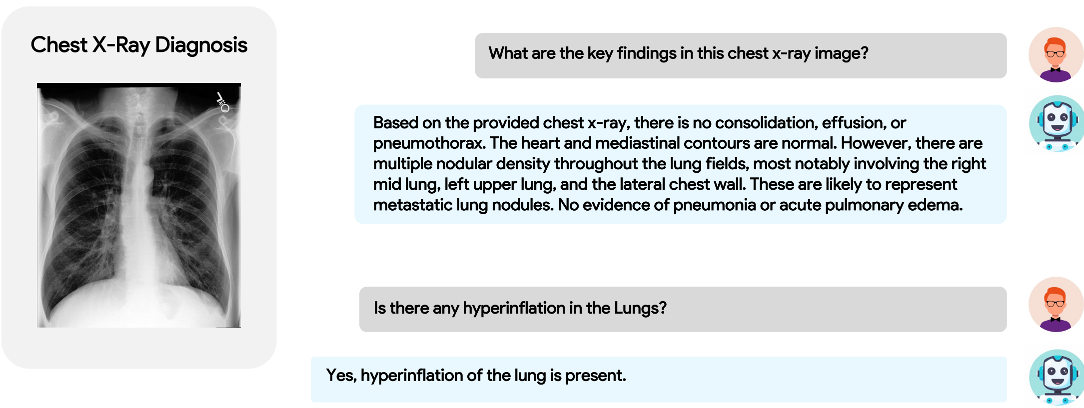

Figure 4: **XrayGPT: Unlocking the Power of Precision, Providing Specific Key Details in Response to User Queries.** Our XrayGPT is able to answer specific key detail in response to user queries.

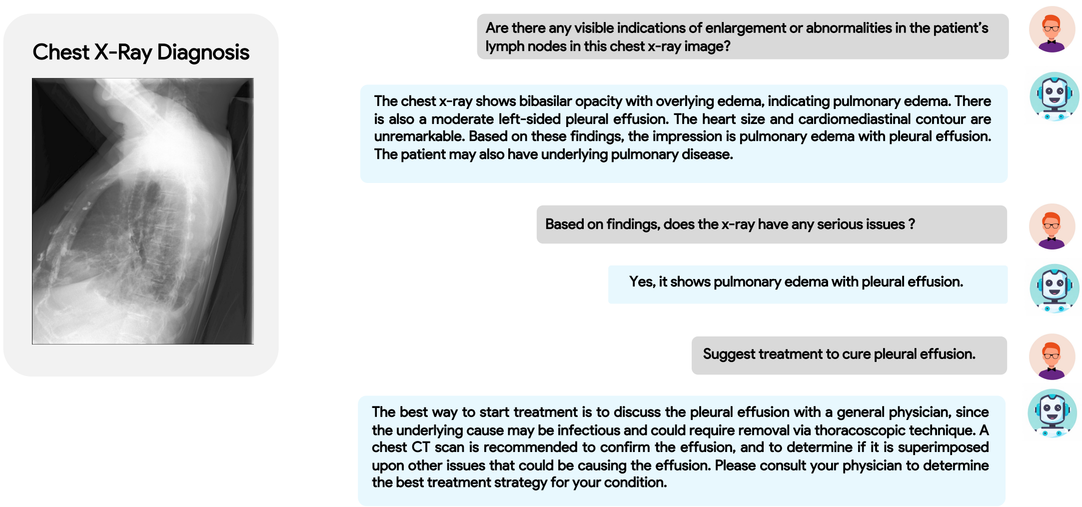

Figure 5: **XrayGPT: The Conversational AI Revolutionizing Radiological Interactions.** Our XrayGPT has radiological conversational capabilities.

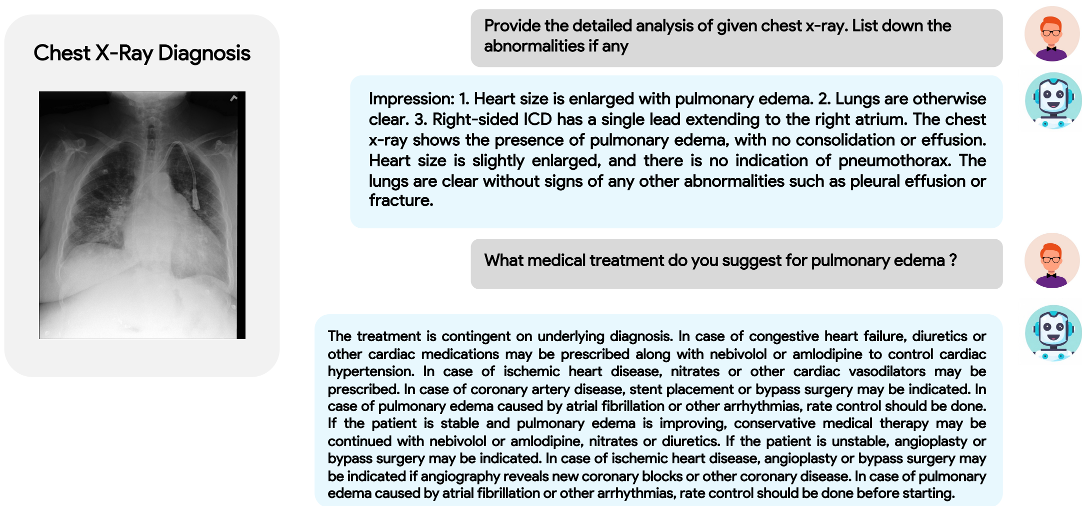

Figure 6: **XrayGPT: Medical Treatment Recommendation.** Our XrayGPT has the capability to suggest treatment based on the diagnosis.