Appendix

A. Datasets Construction Details

A.1 Visual Factual Hallucination

Close-ended. In the close-ended evaluation, we curate two datasets for comparison using GPT-4, referred to as **Mixed-H** and **Single-H**.

The Mixed-H dataset is from the test sets from SLAKE and VQA-RAD. As illustrated in Figure 4, for VQA-RAD, we use GPT-4 to generate new question-answer pairs within the scope of the four targeted types, based on the existing question-answer pairs and metadata, such as the imaging technique and organs associated with a medical image. The bounding box annotations in SLAKE are also provided as additional context to GPT-4 for VQA generation, with the corresponding prompts for SLAKE depicted in Figure 5.

The Single-H dataset is specifically constructed using images and radiology reports from IU-Xray and test sets of MIMIC-CXR, focusing exclusively on chest X-rays. As depicted in Figure 4, we prompt GPT-4 to generate a variety of question-answer pairs using the corresponding chest X-ray reports as context.

Open-ended. This open-ended scenario assesses the model's capability to directly and accurately generate detailed descriptions of anatomical structures, measurements, and symptoms observed in chest X-rays. We utilize image-report pairs from the MIMIC-CXR test set. Based on the medical reports associated with each image, we prompt GPT-4 to generate structured lists of anatomical structures, measurements, and symptoms, which serve as the ground truth. The prompt is shown in Figure 6.

A.2 Knowledge Hallucination beyond Images

Close-ended. The close-ended evaluation data are derived from the MIMIC-CXR test set, utilizing the imaging report as the basis for interpreting the images. This report is then used to prompt GPT-4 to generate diagnostic questions, as illustrated in Figure 7.

Open-ended. The dataset is constructed from the test set of the MIMIC-CXR dataset. In addition to imaging reports, we use a Retrieval Augmentation Generation (RAG) database, which was created using the United States Medical Licensing Examination (USMLE) dataset and radiology textbooks[10], to retrieve relevant knowledge for specific questions.

Specifically, we firstly retrieve relevant knowledge of the image report, and then, as shown in Figure 9, we filter the noisy and useless knowledge from the original retrieved knowledge and then prompt the GPT-4 with filter knowledge and the image reports to generate high-quality QA pairs. Specifically, in Figure 9, we use two separate inference processes using GPT-4 API, implemented as two chains in LangChain[11]. The frist chain and API request is used to filter the noisy knowledge retrieved, the second chain is implemented for dataset generation.

[10]https://www.who.int/publications/i/item/9241545550

[11]https://www.langchain.com/langchain

A.3 Contextual Hallucination

In the dataset design process, we link the MIMIC-CXR data with a de-identified EHR dataset, MIMIC-IV, using the subject ID to provide comprehensive medical notes for each individual's chest X-rays. Clinical notes are specifically utilized as additional context for GPT-4 to generate questions.

The evaluation follows a close-ended design for consistent and reliable assessment. Given the clinical context and the medical report of an image, we prompt GPT-4 to generate questions that could induce contextual hallucinations using some crafted examples. The prompt is illustrated in Figure 10, where the crafted examples are also included.

A.4 Implementation Details

The construction process is implemented using GPT-4 and LangChain. Specifically, the API version of GPT-4 is "2024-05-01-preview". Our open-ended evaluation process is implemented via the Azure OpenAI ChatGPT API.

A.5 Datasets

The dataset we constructed is available in the "data" folder within the supplementary materials. Please note that the images are not included. To obtain the images for SLAKE[12] and VQA-RAD[13], you can download them using the provided links. For MIMIC-CXR[14], you will need to apply for access. For IU-Xray, the images can be downloaded via the link provided in the R2GenGPT repository[15]. Specifically, for MIMIC-CXR, we used the test data split from the MIMIC-CXR-JPG dataset[16].

B. Experiment Setting Details

B.1 Implementation Details

We follow the original settings of all baseline models for inference during the evaluation, as provided in their official implementations. Specifically, the version of the GPT-4o API used is "2024-05-01-preview." Inference for all baseline models, except for GPT-4o, was conducted on A6000 GPUs.

C. Evaluation Details of Open-ended Datasets

C.1 Open-ended Evaluation of Visual Factual Hallucination

The model responses are assessed on their ability to accurately identify hallucinated components across three visual aspects: anatomical structures, measurements, and symptoms. These evaluations are then used to calculate CHAIR. The identification process is carried out using the GPT-4 API, with the prompt design illustrated in Figure 11.

C.2 Evaluation of Knowledge Hallucination beyond Images

The hallucinated responses are identified through a binary assessment by GPT-4. The prompt used for GPT-4 is shown in Figure 12.

D. Case Study

We include several cases in the open-ended evaluation to illustrate hallucinations in (Med)-LVLMs. In Figure 13 and Figure 14, the hallucinated text is highlighted in red.

[12]https://www.med-vqa.com/slake/

[13]https://osf.io/89kps/

[14]https://physionet.org/content/mimic-cxr/2.1.0/

[15]https://github.com/wang-zhanyu/R2GenGPT

[16]https://physionet.org/content/mimic-cxr-jpg/2.1.0/

Close-ended Datasets Generation Prompt for Visual Factual Hallucination -- VQA-RAD

System Message:
You are provided with the metadata and a set of existing QA pairs about one medical image. You task is to synthesize a set of new close-ended QA pairs according to the requirements. Unfortunately, you don't have access to the actual image. Below are requirements for generating the questions and answers in the conversation.

- Do not use phrases like "mentioned", "QA", "context" in the conversation. Instead, refer to the information as being "in the image."
- Answer responsibly within the given information and knowledge context, avoiding information not included in the given context.
- Make sure your generation contains a key "type" for each QA pair, indicating its question type (as detailed in the following part).
- Make the questions diverse.
- The ground truth type can be "yes or no" or one choice from multi-choice (you need to synthesize several choices in this case), condition on the question and available materials.

Here are a set of question types for your generation, you must assign each new QA pair a question type.
- type_1: Anatomical Hallucination. Example questions: "Which part of the body does this image belong to?" "Does the picture contain liver?"
- type_2: Measurement Hallucination like location, size. Example questions: "Where is the liver?"
- type_3: Symptom-Based Hallucination. Example questions: "Is the lung healthy?" "Is there evidence of a pneumothorax?" "Is there a fracture?"
- type_4: Technique Hallucination. Example questions: "What modality is used to take this image? "

Instructions:
- Add one key to each QA pair, key= "ground_truth_type", value= "binary" if the type is "yes or no" else "multi-choice"
- When you see diagnosis information of a disease (e.g. lung cancer) in QA pairs, you should generate new QA pair by asking the symptoms of the disease
- For "multi-choice" type QA, you must include one key "choices" of string type.
- Avoid the question that you can not generate a ground truth, for example avoid the answer "The image does not provide information".

User Message:
Given the metadata: {metadata} and existing QA pairs: {QA_pairs}, generate high-quality questions for which the correct answers can be inferred solely from the provided information. Ensure the questions align with the specified question types.

Figure 4: **Prompt for constructing the close-ended datasets for visual factual hallucination using VQA-RAD**.

Close-ended Datasets Generation Prompt for Visual Factual Hallucination -- SLAKE

System Message:
You are provided with the metadata, object bounding boxes and a set of existing QA pairs about one medical image. You task is to synthesize a set of new close-ended QA pairs according to the requirements. Unfortunately, you don't have access to the actual image. Below are requirements for generating the questions and answers in the conversation.
- Do not use phrases like "mentioned", "bounding boxes", "context" in the conversation. Instead, refer to the information as being "in the image."
- Answer responsibly within the given information and knowledge context, avoiding information not included in the given context.
- Make sure your generation contains a key "type" for each QA pair, indicating its question type (as detailed in the following part).
- Make the questions more diverse.
- The ground truth type can be "yes or no" or one choice from multi-choice (you need to synthesize several choices in this case), condition on the question and available materials.

Here are a set of question types for your generation, you must assign each new QA pair a question type.
- type_1: Anatomical Hallucination. Example questions: "Which part of the body does this image belong to?" "Does the picture contain liver?"
- type_2: Measurement Hallucination like location, size. Example questions: "Where is the liver?"
- type_3: Symptom-Based Hallucination. Example questions: "Is the lung healthy?" "Is there evidence of a pneumothorax?" "Is there a fracture?"
- type_4: Technique Hallucination. Example questions: "What modality is used to take this image? "

Instructions:
- Add one key to each QA pair, key= "ground_truth_type", value= "binary" if the type is "yes or no" else "multi-choice"
- When you see diagnosis information of a disease (e.g. lung cancer) in QA pairs, you should generate new QA pair by asking the symptoms of the disease
- For "multi-choice" type QA, you must include one key "choices" of string type.
- Avoid the question that you can not generate a ground truth, for example avoid the answer "The image does not provide information".

User Message:
The metadata: {metadata} , object bounding boxes:{bounding_boxes}, existing QA pairs: {QA_pairs}. Generate high-quality questions for which the correct answers can be inferred solely from the provided information. Ensure the questions align with the specified question types.

Figure 5: **Prompt for constructing the close-ended datasets for visual factual hallucination using SLAKE**.

Close-ended Datasets Generation Prompt for Visual Factual Hallucination -- IU-Xray + MIMIC-CXR

System Message:

You are provided with the clinical report about a medical image. You task is to synthesize a set of new close-ended QA pairs according to the requirements. Unfortunately, you don't have access to the actual image. Below are requirements for generating the questions and answers in the conversation.

- Do not use phrases like "mentioned", "report", "context" in the conversation. Instead, refer to the information as being "in the image."
- Answer responsibly within the given information and knowledge context, avoiding information not included in the given context.
- Make sure your generation contains a key "type" for each QA pair, indicating its question type (as detailed in the following part).
- Make the questions more diverse.
- The ground truth type can be "yes or no" or one choice from multi-choice (you need to synthesize several choices in this case), condition on the question and available materials.

Here are a set of question types for your generation, you must assign each new QA pair a question type.

- type_1: Anatomical Hallucination. Example questions: "Which part of the body does this image belong to?" "Does the picture contain liver?"
- type_2: Measurement Hallucination like location, size. Example questions: "Where is the liver?"
- type_3: Symptom-Based Hallucination. Example questions: "Is the lung healthy?" "Is there evidence of a pneumothorax?" "Is there a fracture?"
- type_4: Technique Hallucination. Example questions: "What modality is used to take this image? "

Instructions:

- Add one key to each QA pair, key= "ground_truth_type", value= "binary" if the type is "yes or no" else "multi-choice"
- When you see diagnosis information of a disease (e.g. lung cancer) in QA pairs, you should generate new QA pair by asking the symptoms of the disease
- For "multi-choice" type QA, you must include one key "choices" of string type.
- Avoid the question that you can not generate a ground truth, for example avoid the answer "The image does not provide information".

User Message:

The medical report of the image: {report}. generate high-quality questions for which the correct answers can be inferred solely from the provided report.

Figure 6: **Prompt for constructing the close-ended datasets for visual factual hallucination using IU-Xray and MIMIC-CXR**.

Open-ended Datasets Generation Prompt for Visual Factual Hallucination

System Message:

You are provided with the clinical report about a medical image. You task is to synthesize a set of open-ended QA pairs according to the requirements. Unfortunately, you don't have access to the actual image. Below are requirements for generating the questions and answers in the conversation.

- Do not use phrases like "mentioned", "report", "context" in the conversation. Instead, refer to the information as being "in the image."
- Answer responsibly within the given report, avoiding information not included in the given context.

You need to generate the question to query three types of components: (1)anatomical structure; (2)anatomical measurement(organ location, size); (3)symptoms(such as normal or abnormal symptoms, not direct diagnosis).
Here is an example question: "List your findings of anatomical structures and measurements in detail, as well as the possible symptoms, abnormal findings on these structures."

Instructions:

- Additionally return a structured answer with clear classification of three components, which means simply classify the answer into this form:
  {"Sturctured_Answer": {"anatomy": List[str], "measurement": List[str], "symptom": List[str]} }
- Make sure the classification is precise and accurate. If you are not sure about the category, do not include it in the structured result.
- The structured output of "measurement" should be the measurements of organs or important structures, it could be an empty list if there is no important measurements.

User Message:

The medical report of the image: {report}. Generate a question-answer pair of high quality according to the instructions:

Figure 7: **Prompt for constructing the open-ended datasets for visual factual hallucination using MIMIC-CXR**.

Close-ended Datasets Generation Prompt for Knowledge Hallucination beyond Images

System Message:
You are provided with the clinical report about a medical image. You task is to synthesize a set of close-ended QA pairs (diagnosis) according to the requirements. Unfortunately, you don't have access to the actual image. Below are requirements for generating the questions and answers in the conversation.

- Do not use phrases like "mentioned", "report", "context" in the conversation. Instead, refer to the information as being "in the image".
- Answer responsibly within the given report, avoiding information not included in the given context.

Instructions:
- Ensure balanced labels in your generated questions. For example, "yes-or-no" questions should have an equal number of "yes" and "no" answers. To achieve this balance, you may use negative sampling to generate questions with the answer "no".

User Message:

The medical report of the image: {report}. Generate high-quality close-ended question-answer pairs focused on the diagnosis.

Figure 8: **Prompt for constructing the close-ended datasets for knowledge hallucination beyond images using MIMIC-CXR**.

Open-ended Datasets Generation Prompt for Knowledge Hallucination beyond Images

Chain 1: Knowledge Filtering

System Message:
You should filter out useless and noisy retrieval knowledge, keep the important and useful knowledge about the given report, especially the knowledge about the medical terminologies.

User Message:

The medical report: {report}, retrieved knowledge: {knowledge_retrieved}, return the filtered knowledge.

Chain 2: QA Generation

System Message:
You are provided with the clinical report about a medical image, and relevant retrieved knowledge. You task is to synthesize a set of open-ended QA pairs (asking medical terminologies, such as disease, clinical symptoms) according to the requirements. Unfortunately, you don't have access to the actual image. Below are requirements for generating the questions and answers in the conversation.

- Do not use phrases like "mentioned", "report" in the conversation. Instead, refer to the information as being "In the image."
- Answer responsibly within the given information.
- You could rely on the knowledge if it is useful.
- Only focus on the most crucial several terminologies in the reports.

Here is one example question: "What does mediastinal lipomatosis indicate when seen in an image?"

User Message:

The medical report: {report}, retrieved knowledge: {filtered_knowledge}. Generate high-quality open-ended question-answer pairs.

Figure 9: **Prompt for constructing the open-ended datasets for knowledge hallucination beyond images using MIMIC-CXR**.

Close-ended Datasets Generation Prompt for Contextual Hallucination

System Message:
You are provided with the report about a medical image, and the additional clinical notes of this patient. You task is to synthesize a set of close-ended QA pairs according to the requirements. Unfortunately, you don't have access to the actual image. Below are requirements for generating the questions and answers in the conversation.

Answer responsibly within the given information. Given the clinical notes and x-ray reports provided, you should design questions to test for contextual visual hallucination. The goal is to ensure that the model interprets the x-ray images accurately within the specific context provided by the clinical notes, without generating clinically inappropriate or inconsistent responses.

Example yes-or-no questions:
(1) Diagnostic Relevance: "Given the patient's history of severe back pain and the x-ray findings, is it likely that the back pain is due to a cardiopulmonary issue?" (Expected Answer: No)
(2) Complication Risk: "Is it necessary to monitor the patient for potential complications related to the spine, given the findings of thoracic kyphosis and vertebral wedging?" (Expected Answer: Yes)
(3) Family History: "Given the patient's family history of colorectal cancer, as mentioned in the clinical notes, are there any signs of colorectal abnormalities or precancerous lesions visible in the abdominal CT scan?" (Expected Answer: Yes)
(4) Symptom Analysis: "Does the absence of focal consolidation in the chest x-ray suggest that the patient's cough is unrelated to a pulmonary infection?" (Expected Answer: Yes)

Instructions:
- Generate "yes-or-no" questions, ensuring a balanced distribution of labels (yes and no).

User Message:
The medical report of the image: {report}, clinical notes: {notes}. Generate high-quality close-ended question-answer pairs.

Figure 10: **Prompt for constructing the close-ended datasets for contextual hallucination using MIMIC-CXR**.

Open-ended Evaluation Prompt for Visual Factual Hallucination

System Message:
You are provided with a JSON object of model answers from eleven models and a structured ground truth. The structured ground truth contains lists from three aspects: anatomy, symptom, and measurement. Your task is to evaluate each model's answer based on these aspects. Specifically, for each aspect, you need to count: The number of correctly recalled components (recall_number). The number of incorrect components that do not exist in the chest X-ray image and ground truth (wrong_number).

The input format, A JSON object in the following structure:
{ "structured_ground_truth": { "anatomy": List[str], "symptom": List[str], "measurement": List[str] }, "models": { "model_1": "model_1_answer", ... }}
Your output format:
{ "model_1": { "anatomy": {"recall_number": int, "wrong_number": int}, "symptom": {"recall_number": int, "wrong_number": int}, "measurement": {"recall_number": int, "wrong_number": int} }, ...}

Instructions:
- The recall should consider similar mentions, not just exact matches. Ensure you capture all possible correct components.
- Do not count common anatomical structures in chest X-rays that are absent from the ground truth as incorrect components..

User Message:
The structured ground_truth and models' model_response: {input_pairs}. Ensure that you read the model answers carefully and find the similar mentioned components as the recalled components. Your evaluation output in a JSON object format, without extra explanation:

Figure 11: **Prompt for open-ended evaluation of visual factual hallucination**.

Open-ended Evaluation Prompt for Knowledge Hallucination beyond Images

System Message:

You are provided with a JSON object that includes a question-ground_truth pair and the corresponding answers from eleven different models. The question pertains to medical knowledge derived from an image. Your task is to evaluate the correctness of each model's answer, returning a binary output: 1 for correct and 0 for incorrect.

The input format, A JSON object in the following structure:
{ "question": str, "ground_truth": str, "models": { "model_1": "model_1_answer", ... }}
Your output format:
{ "model_1": 1 or 0, "model_2": 1 or 0, ...}

Instructions:
- Consider an answer correct if it expresses the correct medical knowledge, even if the wording differs from the ground truth.
- Aim to capture all possible correct expressions, not just those that match the ground truth verbatim.

User Message:

The question, ground_truth and models' model_response: {input_pairs}. Ensure that you read the model answers carefully and evaluate its correctness. Your evaluation output in a JSON object format, without extra explanation:

Figure 12: **Prompt for open-ended evaluation of knowledge hallucination beyond images**.

List your findings of anatomical structures and measurements in detail, as well as the possible symptoms, abnormal findings on these structures.

Ground Truth: The image reveals a spiculated and cavitary nodule in the left mid lung at the level of the third left anterior rib, measuring 2.5 cm in diameter, which is larger than previous findings and suggests a primary lung cancer or less likely a granulomatous infection. The lungs are clear without new focal areas of consolidation, indicating no acute pneumonia, but show signs of linear scar or atelectasis in the mid lung regions. There is also sclerosis of the medial left clavicle, likely from prior trauma.

GPT-4o

1. Anatomical Structures: Lungs, heart, ribs, clavicles, trachea, and diaphragm. 2. Measurements: Not specified in the image. 3. Possible Symptoms: Shortness of breath, chest pain, cough. 4. Abnormal Findings: Presence of medical devices (e.g., central venous catheter, ECG leads), possible opacity in the lower lung fields indicating fluid or infection, tracheostomy tube in place."

LLaVA-1.6-13B

The image you've provided appears to be an X-ray of a human chest. Here are the findings based on the visible anatomical structures and measurements: 1. Ribs: There are 12 ribs visible on each side, which is the typical number for an adult human. 2. Sternum: The central bone of the chest, the sternum, is also visible. 3. Ribcage: The ribcage is intact with no fractures or dislocations visible. 4. Heart: The heart is located in the center of the chest, and its outline is visible. The heart appears to be of normal size and shape, with no significant abnormalities. 5. Lungs: The lungs are visible as dark areas within the chest cavity. They appear to be of normal size and there are no significant abnormalities such as masses or consolidation. 6. Portable Oxygen Concentrator: There is a portable oxygen concentrator attached to the right side of the chest, which suggests that the individual may require supplemental oxygen. 7. Wires: There are wires running from the chest to the portable oxygen concentrator, which are likely part of the device's electrical connections. 8. Labels: There are labels on the wires, which are likely to indicate the type of wire or its function. 9. Arteries: The arteries are visible as dark lines running alongside the ribcage, which is typical. 10. Airway: The airway is visible as a dark line running vertically down the center of the chest, which is typical. 11. Heart Shadow: The heart shadow is visible as a dark area on the right side of the chest, which is typical. 12. Diaphragm: The diaphragm is not visible in this X-ray, but it is typically located at the bottom of the chest cavity. 13. Spine: The spine is not visible in this X-ray, but it is typically located at the back of the chest cavity. 14. Bone Density: The bone density appears to be normal, with no signs of osteoporosis or other bone diseases. 15. Lymph Nodes: There are no enlarged lymph nodes visible, which is typical. 16. Pleura: The pleura, the membrane that lines the chest cavity, appears to be intact with no signs of inflammation or thickening. 17. Mediastinum: The mediastinum, the central part of the chest cavity, is visible and appears to be normal. 18. Heart Size: The heart appears to be of normal size, with no signs of enlargement or other abnormalities. 19. Lung Expansion: The lungs appear to be fully expanded, which is typical. 20. Airway Diameter: The airway diameter appears to be normal, with no signs of narrowing or obstruction. As for symptoms and abnormal findings, without additional context, it is difficult to provide a definitive diagnosis. However, based on the visible anatomical structures, there are no obvious signs of trauma, infection, or other pathological conditions. The presence of the portable oxygen concentrator suggests that the individual may have a condition that requires supplemental oxygen, such as chronic obstructive pulmonary disease (COPD), but this is not definitively indicated by the X-ray alone. It's important to note that medical diagnoses should be made by a qualified healthcare professional who has access to the patient's full medical history and clinical presentation.

LLaVA-1.6-7B	The image you've provided appears to be a chest X-ray, which is a medical imaging technique used to visualize the internal structures of the chest. Here are the findings and measurements I can identify: 1. **Lung Fields**: The lungs are the large, dark areas on the X-ray, which are typically filled with air. There is a normal distribution of lung tissue, with no significant mass or abnormal density. 2. **Heart and Great Vessels**: The heart is located in the upper left quadrant of the image, and the large vessels (aorta, pulmonary artery, and superior and inferior vena cava) are visible. The heart appears to be in a normal position, and there are no signs of abnormal enlargement or congestion. 3. **Trachea and Bronchus**: The trachea is the tube-like structure in the center of the image, and the bronchus are the branches that extend from the trachea into the lungs. There is no evidence of obstruction or narrowing of these airways. 4. **Esophagus**: The esophagus is the tube-like structure in the upper part of the image, leading from the throat to the stomach. It appears to be in a normal position. 5. **Ribs and Skeletal Structure**: The ribs are the bony structures that form the chest wall. They are straight and aligned, which is typical. 6. **Portable X-ray Device**: There is a label on the image indicating that the X-ray was taken using a portable device, which is a common method for quick and convenient imaging. 7. **Annotations**: There are two annotations on the image: \"L 357\" and \"Portable.\" The \"L 357\" likely refers to a patient identification number or a location marker, and \"Portable\" indicates that the X-ray was taken using a portable device. 8. **Cables and Wires**: There are several cables and wires visible in the image, which are likely part of the portable X-ray device. As for symptoms, the absence of any significant abnormalities or pathology on the X-ray would suggest that there are no immediate signs of illness or injury. However, it is important to note that a chest X-ray is just one part of a comprehensive medical assessment, and a full evaluation by a healthcare professional is necessary to determine the overall health of the individual. If there are any specific symptoms or concerns, it would be necessary to consult a healthcare professional for a full evaluation and interpretation of the X-ray findings.
MiniGPT4-7B	1. The image shows a chest X-ray of a patient with a right upper lobe pneumonia. 2. The patient's right upper lobe is partially filled with air, indicating pneumonia. 3. The left lung appears normal. 4. The patient's heart is visible in the left side of the image, and there is no evidence of cardiac abnormalities. 5. The patient's ribs are visible on the right side of the image, and there is no evidence of rib fractures or other injuries.
LLaVA-Med	The image is a chest X-ray that shows the anatomical structures and measurements of the patient's chest, as well as any potential abnormalities. It can help identify issues related to the heart, lungs, blood vessels, and other structures within the chest cavity. Some common abnormalities that may be visible on a chest X-ray include lung infections, fluid accumulation, tumors, or structural changes. However, without specific details about the findings, it's difficult to provide a precise interpretation. If you have any questions about particular structures or measurements, feel free to ask, and I'll do my best to help you understand the image.
LLaVA-Med-1.5	In the chest X-ray, there are several findings: 1. Cardiomegaly: An enlarged heart, which can be a sign of heart disease or other conditions. 2. Pulmonary congestion: Fluid accumulation in the lungs, which can be caused by various factors such as heart failure, lung infections, or other lung diseases. 3. Pleural effusion: The presence of excess fluid in the pleural space, which is the area between the layers of tissue that line the lungs and the chest cavity. 4. Rib fractures: Broken or cracked ribs, which can be caused by trauma, injury, or other factors. 5. Pneumothorax: A collapsed lung, which can be due to various reasons such as lung disease, injury, or spontaneous rupture of air-filled sacs in the lung tissue. These findings may be associated with various symptoms and conditions, depending on the underlying cause. It is important to consult a healthcare professional for a thorough evaluation and proper diagnosis of the underlying cause of these findings.
XrayGPT	The lung volumes appear decreased on the present exam, with the left lower lobe being the most severely affected, and the right side appearing slightly decreased. The left basilar opacity could be due to atelectasis or aspiration, but could also represent pneumonia in the appropriate clinical context. There is an associated left apical opacity with associated pleural thickening, which is likely related to a prior aspiration/pneumonia process. The cardiac size appears to be stable. The overall impression is that there is decreased lung volume, with possible atelectasis/pneumonia and associated left

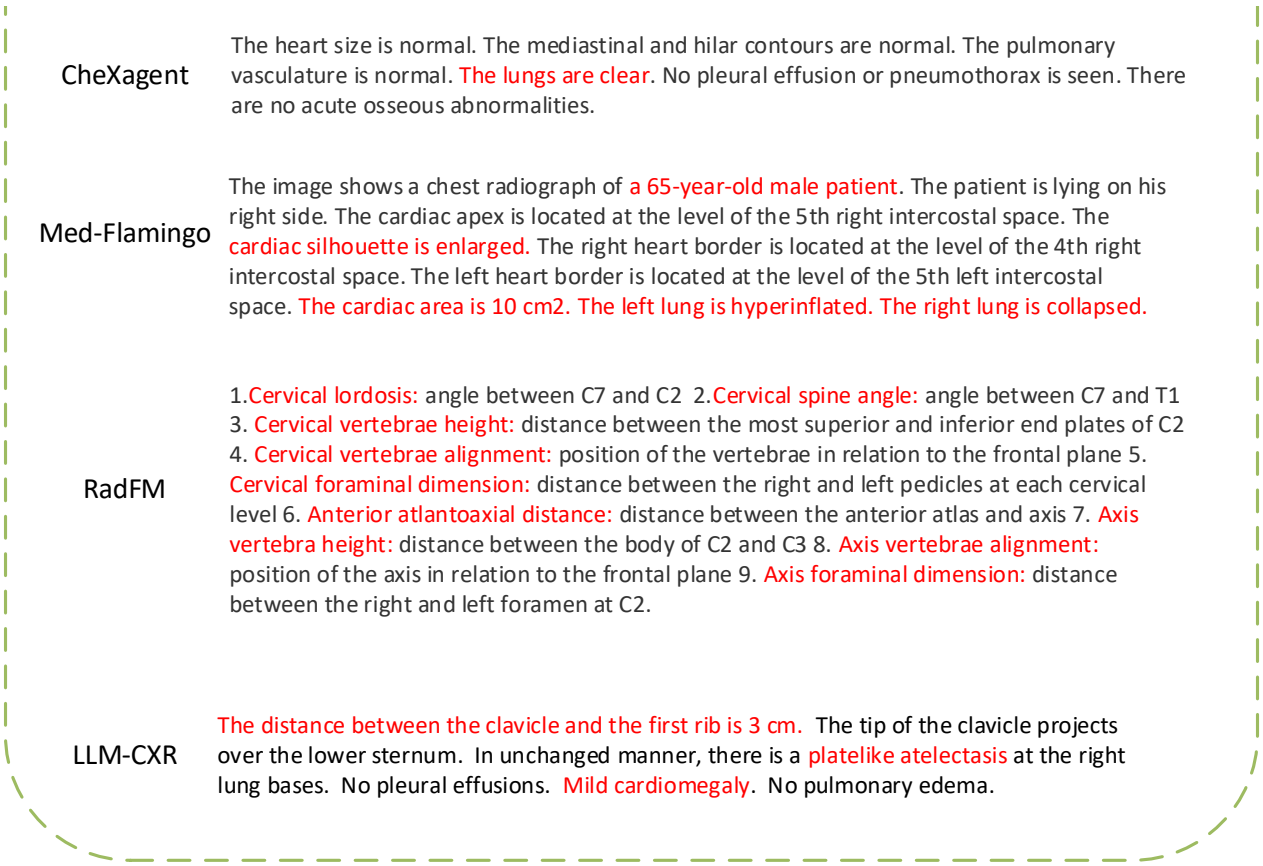

Figure 13: **Cases for open-ended question in visual factual hallucination**.

What could be the implications of the heterogeneous opacities observed in the right lower lung?

Ground Truth: The heterogeneous opacities in the right lower lung suggest the presence of a chronic pneumonia. This could be indicative of exogenous lipoid pneumonia or tuberculosis, both of which are conditions that affect the lungs and can lead to serious health complications if not treated appropriately.

GPT-4o	The heterogeneous opacities could indicate infection, inflammation, or atelectasis, and further evaluation is needed.
LLaVA-1.6-13B	The presence of heterogeneous opacities in the right lower lung, as seen in the X-ray image, could indicate several potential medical conditions. These may include infections such as pneumonia, bronchitis, or tuberculosis; lung cancer; or chronic obstructive pulmonary disease (COPD). The exact cause would require further medical evaluation, including a physical examination, medical history, and possibly additional diagnostic tests, such as a CT scan or a biopsy. The opacities could also be due to atelectasis, which is the collapse of lung tissue, or other less common conditions. It is important for the patient to consult with a healthcare professional for a proper diagnosis and treatment plan.
LLaVA-1.6-7B	The presence of heterogeneous opacities in the right lower lung on a chest X-ray could indicate a variety of conditions, such as infections, inflammation, tumors, or lung diseases like chronic obstructive pulmonary disease (COPD) or bronchitis. The specific implications would depend on the underlying cause of the opacities, which would require further diagnostic testing and medical evaluation to determine the appropriate course of treatment.
MiniGPT4-7B	The heterogeneous opacities observed in the right lower lung could be indicative of a number of different conditions, including pneumonia, lung cancer, or a lung infection. It is important to consult with a medical professional to determine the cause of these opacities and to receive appropriate treatment.
LLaVA-Med	Heterogeneous opacities in the right lower lung on a chest X-ray can have various implications. These opacities may indicate the presence of an infection, inflammation, or other lung abnormalities. The underlying cause could be due to pneumonia, atelectasis (collapse or incomplete expansion of the lung), or other lung diseases. It is important to consider the patient's clinical history, symptoms, and additional diagnostic tests to determine the exact cause of the heterogeneous opacities and provide appropriate treatment. Consulting a healthcare professional is essential for a thorough evaluation and proper diagnosis.
LLaVA-Med-1.5	The heterogeneous opacities observed in the right lower lung could be indicative of various conditions, such as infection, inflammation, or other lung abnormalities. These opacities appear as irregular, cloudy areas on the chest X-ray. It is important to consider the patient's clinical history, symptoms, and additional diagnostic tests to determine the underlying cause of these findings and provide appropriate treatment.

XrayGPT	The patient has a right-sided chest tube and an ET tube, with the endotracheal tube tip extending to the carina. Lung volumes are low, and there is crowding of bronchovascular markings without focal consolidation. Heart size and pulmonary vascularity are unremarkable. There is no pneumothorax, and bony structures are intact. The impression is that there is a right lower lobe pneumonia with no left lobe involvement, and the right lower lobe collapse is due to the pleural effusion.
CheXagent	Pneumonia.
Med-Flamingo	The heterogeneous opacities observed in the right lower lung could be due to atelectasis, pneumonia, pulmonary edema, or atelectasis with pneumonia. The heterogeneous opacities observed in the right lower lung could also be due to atelectasis with pneumonia.
RadFM	The most important implication of the observed heterogeneous opacities in the right lower lung is to rule out aspiration pneumonia and other potential causes of infection. The presence of such opacities may indicate the need for further investigation and treatment.
LLM-CXR	The heterogeneous opacities may indicate aspiration and/or pneumonia, as well as atelectasis.

Figure 14: **Cases for open-ended question in knowledge hallucination beyond images**.