GEMeX: A Large-Scale, Groundable, and Explainable Medical VQA Benchmark for Chest X-ray Diagnosis

Supplementary Material

	Chest ImaGenome	Ours
Reserve	right lung, right mid lung zone, right hilar structures, right hemidiaphragm, left lung, left mid lung zone, left hilar structures, left hemidiaphragm, trachea, spine, abdomen, svc	right lung, right mid lung zone, right hilar structures, right hemidiaphragm, left lung, left mid lung zone, left hilar structures, left hemidiaphragm, trachea, spine, abdomen, svc
Incorporate	left upper lung zone left apical zone	left upper lung zone
	right upper lung zone right apical zone	right upper lung zone
	mediastinum upper mediastinum	mediastinum
	right lower lung zone right costophrenic angle	right lower lung zone
	left lower lung zone left costophrenic angle	left lower lung zone
	cardiac silhouette cavoatrial junction right atrium	cardiac silhouette
Delete	carina right clavicle left clavicle aortic arch	-
Merge	left lung + right lung	bilateral lung
	left upper + right upper	bilateral upper lung zone
	left mid + right mid	bilateral mid lung zone
	left lower + right lower	bilateral lower lung zone
	left hilar + right hilar	bilateral hilar structures
	left hemidiaphragm + right hemidiaphragm	bilateral hemidiaphragm
	left mid + left lower	left mid-to-lower lung zone
	right mid + right lower	right mid-to-lower lung zone
	left mid + left upper	left mid-to-upper lung zone
	right mid + right upper	right mid-to-upper lung zone
	left mid-to-lower + right mid-to-lower	bilateral mid-to-lower lung zone
	left mid-to-upper + right mid-to-upper	bilateral mid-to-upper lung zone
Sum	29	30

Table 8. Anatomical regions transformation from the Chest ImaGenome to our GEMeX version. The left column indicates the detailed operation.

7. GEMeX Construction Details

7.1. Transformation and Distribution of Anatomical Regions

As we said in Section 3.1, we provide detailed operations to transform anatomical regions from Chest ImaGenome to our GEMeX. The process is summarized in Table 8. The resulting anatomical region distribution corresponding to each sentence is shown in Figure 5. Overall, there are 30 regions, and the merged area occupies a large proportion, such as "bilateral lung" and "bilateral hilar structures".

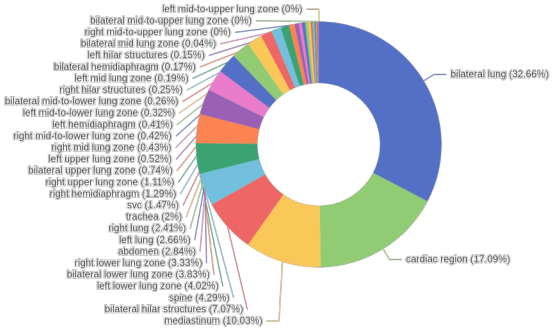


Figure 5. Distribution of anatomical regions corresponding to each sentence after transformation from the Chest ImaGenome dataset.

7.2. Prompt for Re-grounding Report

Here, we provide detailed instructions to re-ground reports with medical LLM, as we elaborated in Section 3.1. The prompt is shown in Table 9, where we add clinical guidance (like (2) (4) (5) (6)) and split and re-written requirements (*e.g.*, (3)) to ensure correct sentence-region correspondence. Moreover, we provide some manually labeled pairs as demonstrations (*i.e.*, "here are some cases") for in-context learning, aiming to improve overall performance.

7.3. Prompt for VQA Generation

In Table 10, we provide a detailed prompt to guide GPT-4o in VQA generation. Specifically, for each CXR, we generate 3 open-ended questions, 2 closed-ended questions, 3 single-choice questions, and 3 multi-choice questions, containing diverse content, like "abnormality", "disease", etc. GPT-4o is required to provide detailed reasoning and specify relevant visual regions (rule (2)). Additionally,

messages = [“role”:“**system**”, “content”:
f“‘You are a helpful chest X-ray radiologist. Given an input sentence, your task is to map it to an anatomical region on X-ray for better observation from a predefined list [right lung, cardiac silhouette,..., bilateral lower lung zone].
Here are some rules:
(1) If there is no corresponding region for this sentence, leave it out.
(2) If the sentence describes the overall anatomical characteristics without specifying a particular region, you can choose “bilateral lung” as its region. For example, “No focal consolidation, pleural effusion or pneumothorax is present”:“bilateral lung”.
(3) One sentence can only correspond to one region. If a sentence’s main symptom involves several anatomical regions, rephrase it into multiple sentences with corresponding regions. Note that all derived sentences must be syntactically complete, not phrases (i.e., containing a subject and a predicate at least). For example: “The cardiomediastinal silhouette is normal.” can be segmented into “The cardiac silhouette is normal.”:“cardiac region”, “The mediastinal silhouette is normal.”:“mediastinum”, where “cardiomediastinal” corresponds to the “cardiac region” and “mediastinum”.
(4) Small (tiny) pleural effusion (fluid) usually appears in the lower lung zone, a moderate pleural effusion appears in the mid-to-lower lung zone, and a large (substantial) pleural effusion can even occupy the entire lung. If the severity (like small, moderate and large) is not indicated, output the left lung or right lung directly.
(5) The main anatomical region for observing pulmonary venous is the bilateral hilar structures on the X-ray.
(6) The region where the atrium and ventricle can be observed is the cardiac region.
Here are some cases: (1)... (2)... (3)... (4)...
Organize your output in a json formatted as DictStr(sentence):Str(region), without other words.”’]

---

messages += [“role”:“**user**”, “content”: “Input: “Bibasilar atelectasis is seen without discrete focal consolidation.”

Table 9. Our proposed prompt guided by radiologist feedback for refining sentence-region pairs.

messages = [“role”:“**system**”, “content”:
f“‘You are a chest X-ray AI assistant, and you are seeing a frontal view chest X-ray image, described by several phrases with visual regions. Generate **3** open-ended questions, **2** closed-ended questions, **3** single-choice questions, and **3** multi-choice questions about this chest X-ray. Format your output in JSON format.
Here are some rules:
(1) Include questions asking about the visual content of the image, containing abnormality, disease, location, severity, cause of disease, size, and implication. For a CXR, the types of questions generated need to be diverse. Do not ask any questions that cannot be answered confidently.
(2) For each question, generate its type (abnormality, location, ...), provide the answer, explain the reason for obtaining such answer, and output the corresponding visual regions as a visual clue.
(3) For open-ended questions, the answers must be concise. You should generate detailed reasons based on the provided CXR phrases and your medical knowledge. Do not refer to the text description in your questions or answers.
(4) Avoid questions that cannot be answered by looking at the given CXR image itself, such as asking about changes/comparisons from previous scans, asking about staff notifications, or asking about view types or other scans.
Here is one example:
Chest X-ray: {...}, One open-ended question can be: {...}, One closed-ended question can be: {...}, One single-choice question can be: {...}, One multi-choice question can be: {...}”’

---

messages += [“role”:“**user**”, “content”: “Chest X-ray: There is also fullness of the right hilum which is new. [visual location: right hilar structures] ...”

Table 10. Our designed prompt for generating groundable and explainable medical VQA, using a grounded report as input.

comparison-based questions are excluded, as only a single CXR is provided (rule (4)). To better align with our objectives, we incorporate manually crafted questions as demonstrations (the inputs after “Here is one example”).

7.4. Prompt for Calculating AR-score

We provide the detailed prompt for calculating the AR-score with GPT-4o (proposed in Section 5.2) in Table 11. When both the ground truth and the LLM’s output are provided, the prompt directs GPT-4o to evaluate the two answers without indicating which is correct. Consequently, it assigns scores purely based on its own judgment, resulting in a more objective and reliable outcome.

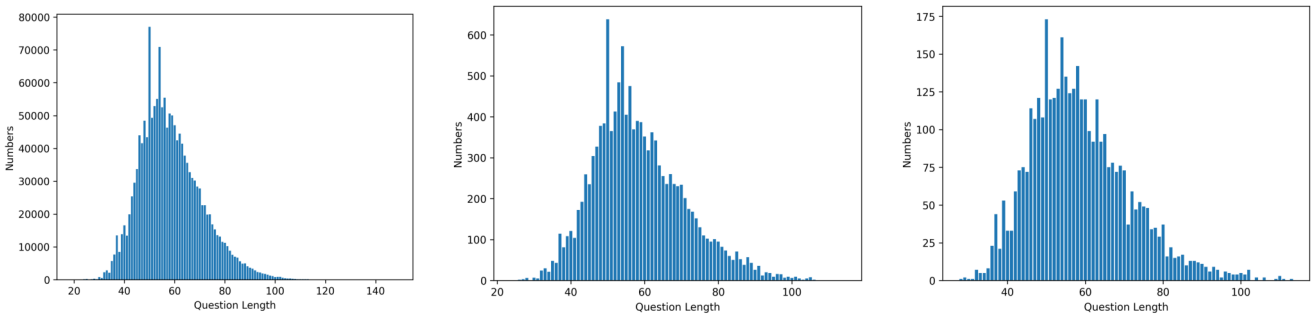
8. Benchmark Details

8.1. Distribution Analysis of Question and Reason Lengths Across Data Splits

We provide a detailed distribution of question and reason lengths across data splits, as presented in Figure 6. It can be seen that under different data splits, the distributions of question (reason) lengths are generally similar. Further-

messages = [“role”:“**system**”, “content”: “You are a helpful and precise assistant for checking the quality of the answer.”
messages += [“role”:“**user**”, “content”: f“‘
[Context]
Figure Caption:
{YOUR_REPORT_CAPTION}
[Question]:
{YOUR_QUESTION}
[Assistant 1]
[Answer]: {GT_ANSWER} [Reason]: {GT_REASON}
[End of Assistant 1]
[Assistant 2]
[Answer]: {LLM_ANSWER} [Reason]: {LLM_REASON}
[End of Assistant 2]
[System]
We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above. The user asks the question on observing an image. For your reference, the visual content in the image is represented with caption describing the same image. Please rate the accuracy (most important), relevance, helpfulness of their responses, considering both answer and reason (if any). Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance. Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.”’

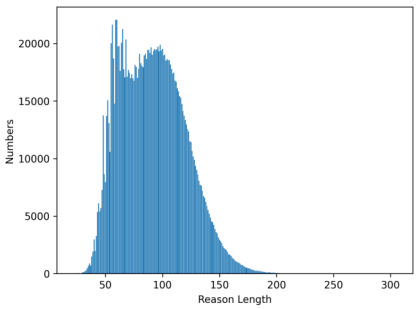
Table 11. Our proposed prompt for calculating the AR-score.



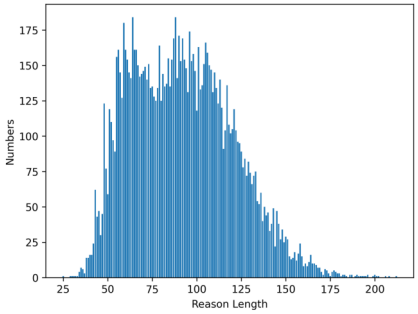
(a) Distribution of question lengths in training set.

(b) Distribution of question lengths in valid set.

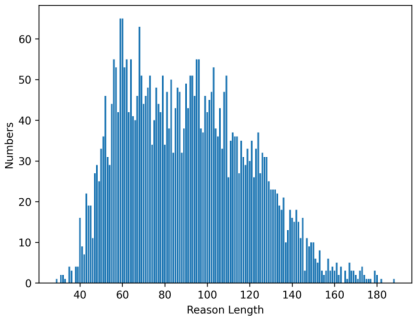
(c) Distribution of question lengths in test set.



(d) Distribution of reason lengths in training set.



(e) Distribution of reason lengths in valid set.



(f) Distribution of reason lengths in test set.

Figure 6. Detailed distribution of question lengths and reason lengths across data splits.

more, from the perspective of reason, the reason lengths mostly fall between 60 and 150, demonstrating the level of detail in the reasons as textual explanations.

8.2. Word Frequency Analysis of Questions and Reasons

Besides providing length distribution, we also explore the frequency of words from both questions and reasons, as

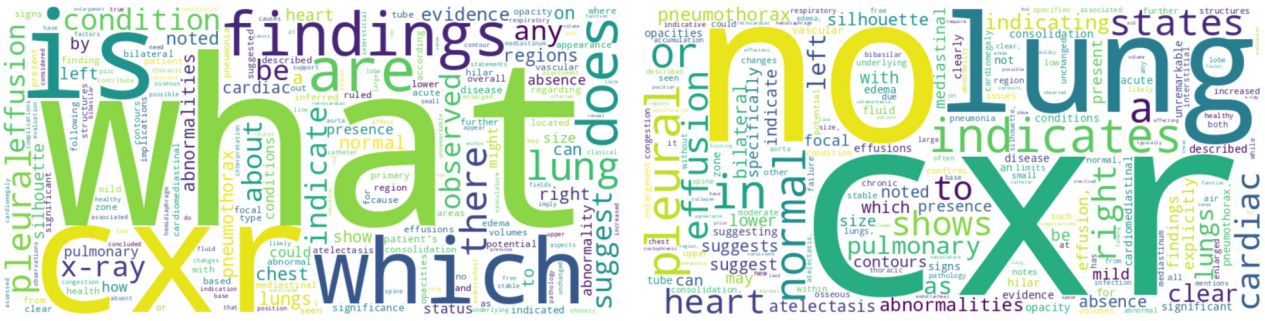


Figure 7. Visualization of word frequencies in questions (left) and reasons (right).

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions. Input an open-ended question, and the assistant will output its answer with a detailed reason and corresponding visual location.\n\n### Human:<image>\n What abnormalities are observed in the left lower lung zone?\n### Assistant: <answer>Linear atelectasis. <reason>The CXR indicates that the lungs are clear except for linear atelectasis located specifically at the left base. <location>[[126, 110, 203, 167]]\n###

Table 12. An input sample of open-ended questions

shown in Figure 7. From the left part (regarding questions), we can observe that the majority of words are question-related terms, such as "what", "which", "is", and "are". Additionally, some disease-related terms, such as "abnormality", "findings", and "pleural effusion", are also quite common. Lastly, content words related to the questions, such as "regions", "evidence", and "size", are frequently mentioned. These demonstrate the diversity of questions; On the right, we show the word cloud of reasons. It can be seen that the vocabulary mainly falls into two categories: one is related to diseases or anatomical regions, such as "normal", "heart", and "pleural", and the other consists of words used to convey explanations, such as "indicates" and "states".

8.3. Input Samples for the Proposed Fine-tuning

We here provide some input samples for a better understanding of how to fine-tune LLaVA-Med [26] on our GEMeX. An input sample of open-ended questions can be seen in Table 12; An input sample of closed-ended questions can be seen in Table 13; An input sample of single-choice questions can be seen in Table 14; An input sample of multi-choice questions can be seen in Table 15.

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions. Input a closed-ended question, and the assistant will output its answer (yes or no) with a detailed reason and corresponding visual location.\n\n### Human:<image>\n Are there any lung abnormalities present in this CXR?\n### Assistant: <answer>No. <reason>The CXR clearly shows that the lungs are clear, indicating no lung abnormalities. <location>[[30, 34, 185, 178]]\n###

Table 13. An input sample of closed-ended questions

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions. Input a single-choice question, and the assistant will output its answer (an option) with a detailed reason and corresponding visual location.\n\n### Human:<image>\n Which of the following is absent in this CXR? <choices>: [A: Pulmonary edema, B: Pleural effusion, C: Pneumothorax, D: All of the above]\n### Assistant: <answer>D <reason>The CXR shows that there is no pulmonary edema, effusion, or pneumothorax present. <location>[[48, 48, 175, 180]]\n###

Table 14. An input sample of single-choice questions

8.4. Fine-tuning Details

We fine-tune both the visual projection layers and the LLM components of LLaVA-Med-v1 (after stage II) by calculating the auto-regressive loss to predict the assistant's responses and the dialogue termination token <STOP>. Particularly, the model is trained for 3 epochs on four NVIDIA H100 GPUs with a batch size of 64, taking around 54 hours. The network is warmed up in the first 0.03 epochs with a linear learning rate from 3e-7 to 2e-5, which further decays by cosine schedule. The optimizer is AdamW. To accelerate training, we employ the Fully Sharded Data Parallel (FSDP) mechanism, the bf16 (Brain Floating Point) data

A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions. Input a multi-choice question, and the assistant will output its answer (some options) with a detailed reason and corresponding visual location.\n\n### Human:<image>\n What abnormalities are mentioned regarding the lung fields? <choices>: [A: Clear lung fields, B: Atelectasis, C: Effusions, D: Congestion]\n### Assistant: <answer>[B, C] <reason>The CXR shows bibasilar atelectasis and small pleural effusions. There is no impressive congestion shown. <location>[[26, 119, 217, 183]]\n###

Table 15. An input sample of multi-choice questions

format, and gradient checkpointing.

8.5. LVLMs Introduction

Besides fine-tuning a task-oriented model, we perform a zero-shot evaluation on our GEMeX dataset across the other 12 LVLMs, with 7 in the general domain and the other 5 in the medical domain:

- **In the General Domain:** `LLaVA-v1` [34] and `Mini-GPT4-v1` [55] are two pioneering works, achieving remarkable results in multimodal tasks; `mPLUG-Owl` [47] is a multimodal model from the mPLUG series focused on visual-language tasks; `LLaVA-v1.5` [33] is an improved version of LLaVA-v1 with enhancements in multimodal alignment, accuracy, and efficiency; `Deepseek-VL` [35] uses the SigLIP-L and SAM-B as the hybrid vision encoder and the DeepSeek-LLM as the base LLM for solving multi-modal tasks; `Qwen-VL-Chat` [5] originates from the Qwen family, possessing capabilities such as multi-lingual dialogue and multi-image interleaved dialogue; `GPT-4o-mini` is a smaller, optimized version of GPT-4 [1] intended for lighter computational environments. Note that we did not test GPT-4o because its safety protection policy prohibits it from analyzing medical images.
- **In the Medical Domain:** `LLaVA-Med-v1` [26] is designed specifically for medical applications based on LLaVA-v1 and `LLaVA-Med-v1.5` [26] is an advanced version; `MiniGPT-Med` [3] is an medical version of Mini-GPT4 [55]; `XrayGPT` [39] is a specialized GPT model for interpreting chest X-rays; `RadFM` [42] is a radiology foundation model.

To make a fair comparison, the evaluated models (except GPT-4o-mini and RadFM (with MedLLaMA-13B [43])) are based on 7B-LLMs in this section. Specifically, LLaVA-v1, LLaVA-Med-v1, and Mini-GPT4-v1 are based on Vicuna-v0-7B [12] while LLaVA-v1.5 and XrayGPT are based on Vicuna-v1-7B; LLaVA-Med-v1.5 is built upon Mistral-7B-Instruct-v0.2 [20]; mPLUG-Owl is using LLaMA-7B [40]; Deepseek-VL [35] is based on DeepSeek-LLM-7B; and Qwen-VL-Chat [5] is based on Qwen-7B. All models' configurations are set according to their open-source codes.

9. More Case Studies

Here, we present more questions with answers from `GPT-4o-mini`, `LLaVA-Med`, and our fine-tuned `LLaVA-Med-GEMeX`, for a better understanding of our dataset GEMeX and the corresponding performance of LVLMs.

- We first present some cases of **open-ended questions**, as in Table 16. It can be seen that our fine-tuned model can generally provide correct (or partially correct) answers and identify relatively accurate visual locations. However, other models fail to deliver both precise textual answers and accurate visual positions simultaneously.
- Furthermore, we provide some cases from **closed-ended questions** in Table 17. Although LLaVA-Med can correctly answer the first two questions, it fails to provide visual grounding. For the third question, GPT-4o-mini provides a correct answer, but there is a discrepancy between its grounded visual location (mediastinum) and the ground truth (cardiac region). In contrast, our fine-tuned model can provide both correct answers and accurate visual grounding.
- Next, we show three cases of **single-choice questions**, presented in Table 18. Overall, GPT-4o-mini and LLaVA-Med demonstrate insufficient image understanding capabilities. For instance, in the third example (CASE III), both models incorrectly identify pleural effusion on both sides, whereas it is actually present only in the left lower lung. In comparison, the fine-tuned model shows significant improvement in visual understanding, as evidenced by the grounding results.
- Finally, some cases from **multi-choice questions** are illustrated in Table 19. It can be observed that multi-choice questions are generally more challenging. GPT-4o-mini can only partially identify the correct options; LLaVA-Med tends to answer questions directly based on the question. For example, in all CASEs, it outputs all textually corresponding answers and provides seemingly reasonable explanations, but some of these answers are incorrect when judged based on visual content; The fine-tuned model may also make analytical errors. In CASE III, it incorrectly determines the presence of pleural effusion.

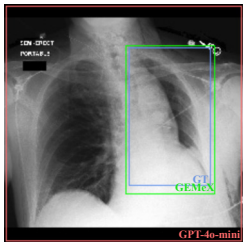
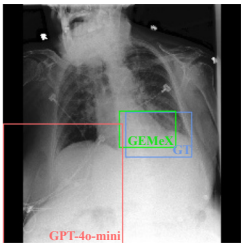
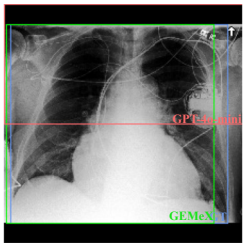
Open-ended questions from GEMeX:			
	CASE I  (Bbox predicted by each model)	**CASE II**  (Bbox predicted by each model)	**CASE III**  (Bbox predicted by each model)
Question *Ground Truth*	What might be the cause of the overexpansion of the right lung? <Answer>: Compensatory hyperinflation due to left lung removal. <Reason>: After a pneumonectomy, the remaining lung may overexpand to compensate for the loss of the left lung.	What disease is indicated at the left lung base? <Answer>: Atelectasis and/or pulmonary edema. <Reason>: The increased opacity at the left lung base suggests either atelectasis or a combination of atelectasis and pulmonary edema.	What is the condition of the lungs in this CXR? <Answer>: Clear. <Reason>: The CXR indicates that the lungs appear clear without definite signs of pneumonia or congestive heart failure (CHF).
`GPT-4o-mini`	<Answer>: The overexpansion of the right lung may be caused by conditions such as a right-sided pneumothorax or chronic obstructive pulmonary disease (COPD) (✗). <Reason>: Overexpansion can occur when air enters the pleural space (as in pneumothorax) or when there is an obstruction that prevents air from leaving the lung, causing it to become hyperinflated. In COPD, the lung tissue loses elasticity, leading to overinflation. <GPTScore>: 44.44	<Answer>: Possible pneumonia or pleural effusion (✗). <Reason>: The left lung base shows increased opacity, which may indicate fluid accumulation (pleural effusion) or consolidation (pneumonia). These conditions often present as areas of increased density on chest X-rays. <GPTScore>: 55.56	<Answer>: The lungs appear to have a significant opacity, suggesting possible fluid accumulation or consolidation (✗). <Reason>: The central area of the chest X-ray shows increased density, which may indicate the presence of pleural effusion or other pathological processes affecting lung visibility. The clear outlines of the lungs are obscured, which is not typical in a healthy CXR. <GPTScore>: 33.33
`LLaVA-Med`	<Answer>: The overexpansion of the right lung with a mediastinal shift to the left could be caused by various factors, such as a buildup of air, fluid, or pressure within the lung cavity. This can be due to conditions like pneumothorax (collapsed lung), pleural effusion (fluid buildup around the lungs), or a mass causing pressure on the lung (✗). It's important to consider the patient's clinical history, symptoms, and additional imaging findings to determine the exact cause of the overexpansion and mediastinal shift. <GPTScore>: 55.56	<Answer>: The image shows left lung base fibrocystic disease (✗), which is a lung condition characterized by the presence of fibrous and cystic tissue in the lung base. <GPTScore>: 22.22	<Answer>: In this post-operative chest X-ray (CXR), the lungs appear to be clear and without any signs of infection (✓). <GPTScore>: 77.78
`LLaVA-Med-GEMeX`	<Answer>: Possible chronic lung disease or compensatory mechanism (✓). <Reason>: Overexpansion of the lung can be due to chronic lung diseases or compensatory mechanisms, especially after a left lobectomy (✓). <GPTScore>: 66.67	<Answer>: Atelectasis (✓). <Reason>: The CXR shows moderate atelectasis at the left lung base (✓), which is a collapse or incomplete expansion of the lung. <GPTScore>: 77.78	<Answer>: Clear. <Reason>: The CXR indicates that the lungs are clear with normal pulmonary vasculature (✓). <GPTScore>: 88.89

Table 16. Open-ended questions from GEMeX answered by GPT-4o-mini, LLaVA-Med, and our fine-tuned LLaVA-Med-GEMeX.

Closed-ended questions from GEMeX:

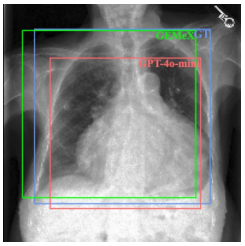
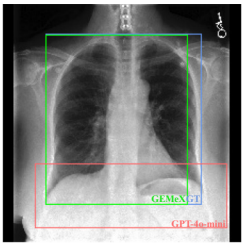
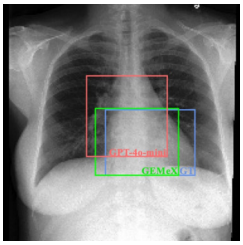
	CASE I	**CASE II**	**CASE III**
	 (Bbox predicted by each model)	 (Bbox predicted by each model)	 (Bbox predicted by each model)
Question *Ground Truth*	Are there signs of pneumonia in this CXR? <Answer>: No. <Reason>: The CXR explicitly states that there is no evidence of pneumonia, indicating clear lung fields.	Is there any pleural effusion present in the CXR? <Answer>: No. <Reason>: The CXR explicitly states that there is no pleural effusion.	Is the heart size abnormal in this CXR? <Answer>: No. <Reason>: The CXR indicates that the heart size is within normal limits.
`GPT-4o-mini`	<Answer>: Yes. <Reason>: The chest X-ray shows areas of increased opacity, which may indicate the presence of fluid or consolidation (✗) typically associated with pneumonia. <GPTScore>: 37.50	<Answer>: Yes. <Reason>: The presence of pleural effusion is indicated by the blunting of the costophrenic angles and the increased opacity in the lower lung fields (✗), suggesting fluid accumulation in the pleural space. <GPTScore>: 20.00	<Answer>: No. <Reason>: The heart size appears to be within normal limits (✓) based on the overall silhouette and the ratio of the heart to the thoracic cavity. <GPTScore>: 100.00
`LLaVA-Med`	<Answer>: No, there are no signs of pneumonia in this chest X-ray (✓). <GPTScore>: 87.50	<Answer>: No, there is no pleural effusion present in the chest X-ray (CXR) (✓). <GPTScore>: 100.00	<Answer>: Yes, the heart size in this chest X-ray appears to be abnormally enlarged (✗). This is referred to as cardiomegaly. <GPTScore>: 20.00
`LLaVA-Med-GEMeX`	<Answer>: No. <Reason>: The CXR indicates no focal consolidation, which rules out pneumonia (✓). <GPTScore>: 100	<Answer>: No. <Reason>: The CXR clearly states that there is no pleural effusion (✓). <GPTScore>: 100.00	<Answer>: No. <Reason>: The heart is described as normal in size (✓). <GPTScore>: 100.00

Table 17. Closed-ended questions from GEMeX answered by GPT-4o-mini, LLaVA-Med, and our fine-tuned LLaVA-Med-GEMeX.

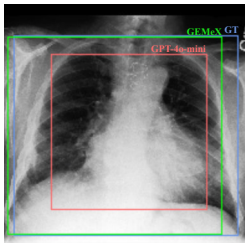
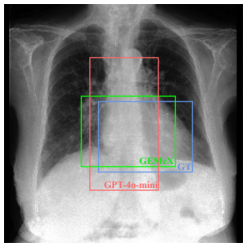
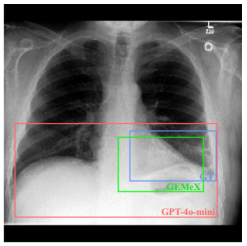
Single-choice questions from GEMeX:			
	CASE I  (Bbox predicted by each model)	**CASE II**  (Bbox predicted by each model)	**CASE III**  (Bbox predicted by each model)
Question	What is the overall condition of the lungs based on the CXR? Choices: [A: Clear, B: Consolidated, C: Effused, D: Thickened]	What does the CXR indicate about the heart size? Choices: [A: Normal, B: Mild cardiomegaly, C: Moderate cardiomegaly, D: Severe cardiomegaly]	Which regions are assessed for pleural effusion in this CXR? Choices: [A: Right lower lung zone, B: Left lower lung zone, C: Cardiac region, D: Bilateral lung]
Ground Truth	<Answer>: A. <Reason>: The CXR indicates that the lungs are clear of consolidation or effusion, although right pleural-based thickening at the base.	<Answer>: C. <Reason>: The CXR indicates unchanged moderate cardiomegaly.	<Answer>: B. <Reason>: The pleural effusion is specifically noted in the left lower lung zone.
`GPT-4o-mini`	<Answer>: B. <Reason>: The presence of increased opacity in the lung fields suggests consolidation, indicating that the lungs are not clear and may have fluid or solid material in the alveoli (✗). <GPTScore>: 33.33	<Answer>: B. <Reason>: The CXR shows an enlarged heart silhouette, indicating mild cardiomegaly. The heart appears larger than normal but not excessively so (✗). <GPTScore>: 55.56	<Answer>: D. <Reason>: Pleural effusion can accumulate in the lower lung zones, and both sides of the chest The bilateral lung regions are crucial for evaluating pleural effusion (✗). <GPTScore>: 66.67
`LLaVA-Med`	<Answer>: The chest X-ray (CXR) shows bilateral diffuse interstitial thickening, which means that both lungs have a thickened appearance in their interstitial areas (✗). This finding can be associated with various lung conditions, such as infections, inflammation, or fibrosis. It is important to consider the patient's clinical history and symptoms, as well as consult a healthcare professional for a thorough evaluation and proper diagnosis of the underlying cause of these findings. <GPTScore>: 37.50	<Answer>: The CXR indicates that there is mild cardiomegaly, which means that the heart appears slightly enlarged (✗). <GPTScore>: 44.44	<Answer>: In this chest X-ray (CXR), the regions assessed for pleural effusion include the right lower lung zone, the left lower lung zone, the cardiac region, and the bilateral lung (✗). <GPTScore>: 50.00
`LLaVA-Med-GEMeX`	<Answer>: A. <Reason>: The CXR indicates that the lungs are otherwise clear except for the noted pleural thickening (✓). <GPTScore>: 88.89	<Answer>: C. <Reason>: The CXR indicates moderate cardiomegaly (✓). <GPTScore>: 100.00	<Answer>: B. <Reason>: The CXR specifically indicates that trace pleural effusion is difficult to exclude in the left lower lung zone (✓). <GPTScore>: 87.50

Table 18. Single-choice questions from GEMeX answered by GPT-4o-mini, LLaVA-Med, and our fine-tuned LLaVA-Med-GEMeX.

Multi-choice questions from GEMeX:

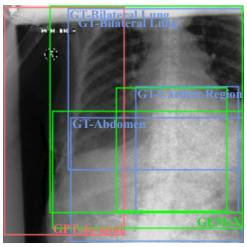
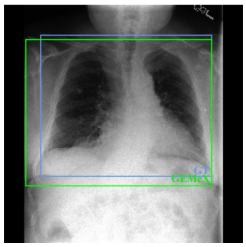
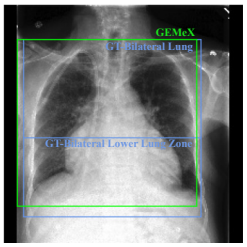
	CASE I	**CASE II**	**CASE III**
	 (Bbox predicted by each model)	 (Bbox predicted by each model)	 (Bbox predicted by each model)
Question	What regions on the X-ray indicate abnormal findings? Choices: [A: Bilateral lung, B: Cardiac region, C: Abdomen, D: Spine]	Which findings indicate abnormal lung conditions? Choices: [A: Mild interstitial abnormality, B: Hyperinflation, C: Focal consolidation, D: No pleural effusion]	Which observations rule out serious lung conditions in this CXR? [A: No pneumonia, B: No lung nodules, C: No pleural effusions, D: No lymphadenopathy]
Ground Truth	<Answer>: [A, B, C]. <Reason>: The bilateral lung shows pleural effusion, the cardiac region shows enlargement, and the abdomen shows the nasogastric tube. The spine is normal.	<Answer>: [A, B]. <Reason>: Mild interstitial abnormality and hyperinflation indicate abnormal lung conditions. There is no focal consolidation.	<Answer>: [A, B, D] <Reason>: The CXR shows no pneumonia, lung nodules, or lymphadenopathy, which rules out serious lung conditions. However, there are mild bilateral pleural effusions.
`GPT-4o-mini`	<Answer>: [A: Bilateral lung]. <Reason>: The bilateral lung region shows potential abnormalities such as opacities or other irregularities that may indicate pathology (✗). <GPTScore>: 37.50	<Answer>: [A, B, C]. <Reason>: Mild interstitial abnormality, hyperinflation (✓), and focal consolidation (✗) are all findings that indicate abnormal lung conditions. <GPTScore>: 66.67	<Answer>: [A, B, C, D]. <Reason>: The absence of pneumonia, lung nodules, pleural effusions (✗), and lymphadenopathy (✓) in the chest X-ray suggests that there are no serious lung conditions present. Each of these findings is associated with significant lung pathology, and their absence helps rule out serious issues. <GPTScore>: 62.50
`LLaVA-Med`	<Answer>: The chest X-ray shows abnormal findings in multiple regions. The left side of the image indicates bilateral lung abnormalities, which could be related to lung issues such as infection, inflammation, or other lung conditions (✗). The right side of the image shows abnormal findings in the cardiac region, which may suggest issues with the heart or surrounding structures (✓). The lower part of the image reveals abnormal findings in the abdomen, which could be related to gastrointestinal or other abdominal issues (✓). Lastly, the spine region also shows abnormal findings, which could be related to spinal issues or other back-related problems. <GPTScore>: 50.00	<Answer>: The chest x-ray shows findings A, B, and C, which indicate abnormal lung conditions. These findings are likely a combination of interstitial thickening, hyperinflation (✓), and focal consolidation (✗). These abnormalities can be associated with various lung conditions, such as chronic obstructive pulmonary disease (COPD), asthma, or other respiratory disorders. <GPTScore>: 55.56	<Answer>: The image shows a chest X-ray that demonstrates the absence of certain conditions, which are marked as B on the image. These conditions include pneumonia, lung nodules, lymphadenopathy (✓), and pleural effusions (✗). This means that the chest X-ray does not show any signs of these lung conditions, ruling them out based on the image. <GPTScore>: 62.50
`LLaVA-Med-GEMeX`	<Answer>: [A, B, C]. <Reason>: The bilateral lung shows interstitial opacities (✗), the cardiac region shows enlargement (✓) , and the abdomen shows the feeding tube location (✓). <GPTScore>: 87.50	<Answer>: [A, B]. <Reason>: The CXR shows mild interstitial abnormality and hyperinflation (✓) as abnormal lung conditions. <GPTScore>: 88.89	<Answer>: [A, B, C, D]. <Reason>: The CXR indicates the absence of pneumonia, lung nodules, pleural effusions (✗), and lymphadenopathy (✓). <GPTScore>: 62.50

Table 19. Multi-choice questions from GEMeX answered by GPT-4o-mini, LLaVA-Med, and our fine-tuned LLaVA-Med-GEMeX.