MELBA Journal Machine Learning for Biomedical Imaging

Prompting Medical Large Vision-Language Models to Diagnose Pathologies by Visual Question Answering

Danfeng Guo [1], Demetri Terzopoulos [1,2]

1 Computer Science Department, University of California, Los Angeles, CA, USA
2 VoxelCloud, Inc., Los Angeles, CA, USA

Abstract

Large Vision-Language Models (LVLMs) have achieved significant success in recent years, and they have been extended to the medical domain. Although demonstrating satisfactory performance on medical Visual Question Answering (VQA) tasks, Medical LVLMs (MLVLMs) suffer from the hallucination problem, which makes them fail to diagnose complex pathologies. Moreover, they readily fail to learn minority pathologies due to imbalanced training data. We propose two prompting strategies for MLVLMs that reduce hallucination and improve VQA performance. In the first strategy, we provide a detailed explanation of the queried pathology. In the second strategy, we fine-tune a cheap, weak learner to achieve high performance on a specific metric, and textually provide its judgment to the MLVLM. Tested on the MIMIC-CXR-JPG and Chexpert datasets, our methods significantly improve the diagnostic F1 score, with the highest increase being 0.27. We also demonstrate that our prompting strategies can be extended to general LVLM domains. Based on POPE metrics, it effectively suppresses the false negative predictions of existing LVLMs and improves Recall by approximately 0.07.

Keywords

Medical Visual Question Answering, Large Vision-Language Models, Prompt Engineering

Article informations

https://doi.org/10.59275/j.melba.2025-1a8b

Received: 2024-08-02, Published 2025-03-14

Corresponding author: lyleguo@g.ucla.edu

1. Introduction

Research on Large Language Models (LLMs) has yielded astonishing results in recent years. LLMs with billions of parameters have achieved outstanding abilities in a wide range of application scenarios (OpenAI, 2022; OpenAI et al., 2023; Chiang et al., 2023). The success of LLMs has quickly extended into the Vision-Language (VL) domain. Large Vision-Language Models (LVLMs) are built upon LLMs by training adapters that project visual features into tokens that can be interpreted by LLMs (Li et al., 2023b; Zhang et al., 2023a; Liu et al., 2023b). Visual Question Answering (VQA) is an essential skill of LVLMs, and VQA accuracy serves as a test metric for most of these models (Li et al., 2023b; Zhang et al., 2023a; Zhu et al., 2023; Liu et al., 2023b). LVLMs have been pretrained on medical datasets (Li et al., 2023a; Liu et al., 2023c; Singhal et al., 2023) and they have been tested on medical VQA tasks (Lau et al., 2018; He et al., 2020). These Medical LVLMs (MLVLMs) have been able to answer questions regarding the imaging modalities, organs, and abnormalities depicted by the input medical scans.

However, "hallucination" has been a major problem for LVLMs. This refers to the generation of content that is contradictory to the input images. Hallucination can be measured via VQA. One may ask the model questions regarding the existence of objects in the input image(s) and the hallucination level is assessed as the percentage of correctly answered questions. VQA can also potentially serve for medical image diagnosis. Users pose questions regarding a pathology and the MLVLM responds based on its analysis of the medical scans. However, most of the available datasets involve simple questions such as "what is the modality of this image" and "what is the organ/tissue in this image". MLVLMs have yet to be thoroughly evaluated on VQA accuracy across a broad spectrum of pathologies. Additionally, general VQA models are usually tested by the commonly known accuracy: the percentage of correctly answered questions, which is an unsuitable measure for medical VQA. Medical image classification metrics such as the Precision, Recall, and F1 are more suitable for the evaluation of medical VQA models. Several strategies have been explored to enhance the question answering of LLMs/LVLMs, including chain-of-thought prompting (Zheng et al., 2023), self-consistency (Wang et al., 2023), and retrieval-based augmentation (Caffagni et al., 2024). All these methods involve fine-tuning the models, which is expensive. Training-free methods to improve the VQA accuracy of MLVLMs are desirable.

For MLVLMs, hallucination is exacerbated by imbalanced training data. Many pathologies are minority categories in medical datasets. Models trained on large-scale medical data may easily fail to learn the features of less common pathologies. Addressing data bias typically involves strategies such as including more data of better quality, but given the scarcity of medical data, significantly enlarging the dataset may not be feasible. Common remediations involve re-sampling the data such that the positive and negative cases are better balanced, but this poses challenges when the data involves multiple categories of pathology. Additionally, re-sampling may undermine the training needs of LVLMs, which generally require large quantities of data. These problems highlight the need of a cost-effective approach to navigate the problem of minority categories in datasets.

Our study focuses on the VQA abilities of MLVLMs. In particular, we test an existing MLVLM, LLaVA-Med (Li et al., 2023a) for chest X-ray VQA across 5 categories of pathologies. The results show that the model has low accuracy, especially on minority pathologies. To enhance its VQA accuracy, we propose two prompting strategies. The first involves enriching prompts with detailed explanations of the queried pathology. The explanations include how the queried pathology is defined and how it appears in images. Our second strategy involves introducing an auxiliary weak-learner model as another agent. We train a small image classifier and fine-tune it to identify negative images accurately. Then, the negative predictions of this classifier are appended to the prompt as a reference for the MLVLM.

We run our experiments on the MIMIC-CXR-JPG (Goldberger et al., 2000) and Chexpert (Irvin et al., 2019) datasets. The results show that our prompt strategies improve the F1 score significantly in most pathology categories (highest +0.27). We also show that our weak-learner-prompting strategy is applicable to the general domain. It reduces the false negative predictions of general domain LVLMs and improves the Recall by around 0.07 according to POPE metrics (Li et al., 2023c).

To summarize, our contributions include the following:

1. We improve the VQA accuracy of MLVLMs by prompting with detailed explanations of pathologies.
2. We introduce a low-cost weak learner model as a reference for LLaVA-Med, and this effectively reduces the false positive answers.
3. We show that our second prompting strategy can be extended to general domains to help models adapt to specialized accuracy needs.

Section 2 reviews related work, Section 3 describes our methodology, Section 4 presents our empirical study and its results, and Section 5 draws conclusions from our research.

2. Related Work

LVLMs and VQA LVLMs are built upon LLMs. A pre-trained visual encoder extracts the visual features and an adapter module projects the extracted features to ones that can be understood by the LLM. Models of this type include those by Liu et al. (2023b), Zhu et al. (2023), and Zhang et al. (2023a). During training, the visual encoder and the LLM are usually fixed. VQA is an essential skill of LVLMs. Given an input image, the models should be able to answer questions correctly regarding that image.

Hallucination in LVLM VQA The hallucination problem usually refers to the LVLM generating a response that is not consistent with the input image. For VQA, in their generated answers the models may make mistakes on object presence, location, attributes, or the mutual relationship between objects. Li et al. (2023c) find that frequently occurring objects are easily hallucinated by LVLMs, in that they tend to mention such objects even if it they are absent in the image. Qian et al. (2024) and Liu et al. (2023a) show that LVLMs sometimes presume the assumptions in questions are true and easily give wrong answers when asked about some objects not in the given image.

Causes of LVLM VQA Hallucination Hallucination can result from bias in the training data, missing fine-grained visual features, and LLM decoding strategies (Liu et al., 2024). For data bias, the imbalanced distribution of data is an important aspect. When most of the answers to a question in the training data are “Yes”, the model tends to answer “Yes” to that question. Missing fine-grained visual features usually result from the pretraining of the visual encoder. Most LVLMs use the visual encoder of CLIP trained through contrastive learning. The encoder mainly focuses on salient features and ignores fine-grained features (Jain et al., 2023). LVLM decoding strategies mostly choose the next word as the one having maximum conditional probability given previous text and the input image. This can lead to hallucination when the model overly relies on the knowledge learned in its training texts. Other causes include model simplicity and insufficient attention (Liu et al., 2024).

Mitigation of LVLM VQA Hallucination Strategies to mitigate hallucination in LVLMs mainly fall into two categories: prompt engineering and model improvement. Prompt engineering is a well developed technique in the natural language domain that provides LLMs with instructions and/or additional information to perform tasks. Vu et al. (2024) and Peng et al. (2023) improve the LLM performance by providing external information. Regarding the former, Liu et al. (2023a) leverage visual instructions constructed from the bounding box information in the input image to prompt LLMs. Zheng et al. (2023) use a chain of thought scheme to prompt the models to perform step-by-step visual-language reasoning like humans, which eventually leads to the correct answers. Wang et al. (2023) generate multiple chains of thought and use the one with the majority vote as the answer. Caffagni et al. (2024) prompt the model with explanations of the terms in questions. Wang et al. (2024) retrieve external knowledge to assist the model in VQA tasks. With regard to the model improvement strategy for reducing hallucination, Sun et al. (2023) improve the visual and text feature alignment through reinforcement learning. Leng et al. (2023) propose a contrastive decoding strategy to reduce reliance on pretrained knowledge. Favero et al. (2024) and Zhao et al. (2024) also focus on the inference stage and propose specialized decoding strategies to mitigate hallucination. Other strategies for reducing hallucination have been proposed. For example, Zhou et al. (2024) design a post-processing model to detect hallucinated objects and rephrase the generated answers, and Sun et al. (2023) adapt a reinforcement learning strategy that uses human evaluation of the hallucination level to improve the model.

Assessment of LVLM Hallucination There are two approaches to assessing hallucination in LVLMs. The first is VQA. The ground truth information of the input images is leveraged to construct questions regarding the existence of objects in the images (e.g., “Is there a black cat in the image?”), as well as questions about objects which do not exist in the images. The models are evaluated in terms of the percentage of correctly answered questions. Metrics of this type include POPE (Li et al., 2023c), CIEM (Hu et al., 2023), and NOPE (Lovenia et al., 2023). The second approach is to use pre-designed prompts from which the models produce various generations that are then evaluated. Examples include CHAIR (Rohrbach et al., 2018), which counts the hallucinated objects in generated image captions, and MMHAL-BENCH (Sun et al., 2023), which uses GPT-4 (OpenAI et al., 2023) to compare the generations with human answers and determine the propensity toward hallucination.

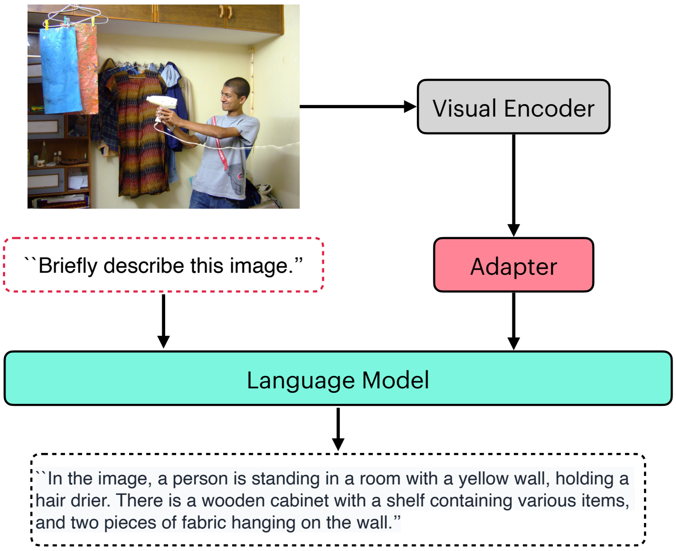
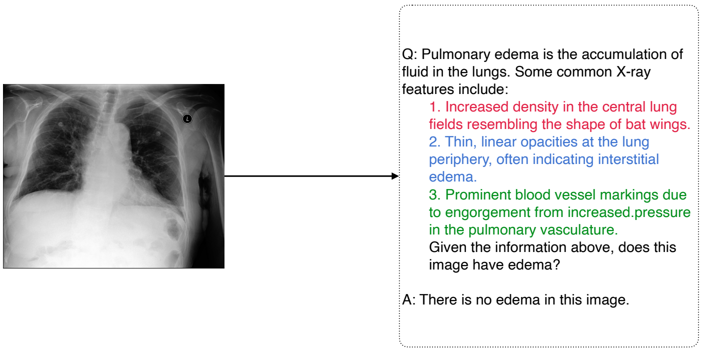

Visual Encoder
``Briefly describe this image.”
Adapter
Language Model
``In the image, a person is standing in a room with a yellow wall, holding a hair drier. There is a wooden cabinet with a shelf containing various items, and two pieces of fabric hanging on the wall.”

Figure 1: The structure of common LVLMs.

VQA in MLVLMs For MLVLMs, given a medical scan, models such as LLaVA-Med (Li et al., 2023a) and Med-PALM (Liu et al., 2023c) are able to answer questions regarding the types of modalities, the scanned organs, and medical indicators such as opacity. They have demonstrated good performance on medical VQA datasets such as VQA-RAD (Lau et al., 2018), SLAKE (Liu et al., 2021), and Path-VQA (He et al., 2020). XrayGPT (Thawakar et al., 2023) improved model performance by performing an additional round of training on selected high-quality data. CheXagent (Chen et al., 2024) further developed the training process such that the visual encoder, projection layer, and the whole model are trained separately in three steps. However, most medical questions in existing datasets are simple (e.g., view classification). MLVLMs have not yet been tested on a broader range of complex pathologies. Recently, RaDialog (Pellegrini et al., 2023b) trains a separate image classifier and fine-tunes the MLVLM with the classification results on their designed datasets including various medical tasks. The image classifier helps the model generate medical reports with high medical correctness.

Medical Image Classification via VLMs The most notable vision-language model for medical image classification is ConVIRT (Zhang et al., 2022), which employed a contrastive learning approach to pretrain the model for various tasks. It inspired CLIP (Radford et al., 2021), which is able to utilize the pretrained model for zero-shot classification by searching for the best match between the image features and text features of disease categories. Several CLIP variants, such as BioMedCLIP (Zhang et al., 2023b), ChexZero (Tiu et al., 2022), MedCLIP (Wang et al., 2022), Xplainer (Pellegrini et al., 2023a), Seibold et al. (2022), and Jang et al. (2022), perform well on medical image classification



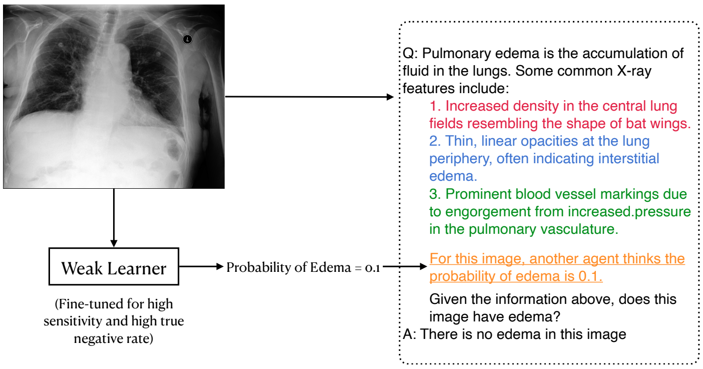
Q: Pulmonary edema is the accumulation of fluid in the lungs. Some common X-ray features include:

1. Increased density in the central lung fields resembling the shape of bat wings.
2. Thin, linear opacities at the lung periphery, often indicating interstitial edema.
3. Prominent blood vessel markings due to engorgement from increased.pressure in the pulmonary vasculature.

Given the information above, does this image have edema?

A: There is no edema in this image.

Figure 2: An example of including pathology explanations when prompting an MLVLM for medical VQA.



Weak Learner
(Fine-tuned for high sensitivity and high true negative rate)
Probability of Edema = 0.1

Q: Pulmonary edema is the accumulation of fluid in the lungs. Some common X-ray features include:

1. Increased density in the central lung fields resembling the shape of bat wings.
2. Thin, linear opacities at the lung periphery, often indicating interstitial edema.
3. Prominent blood vessel markings due to engorgement from increased.pressure in the pulmonary vasculature.

For this image, another agent thinks the probability of edema is 0.1.

Given the information above, does this image have edema?

A: There is no edema in this image

Figure 3: An example of prompting an MLVLM for medical VQA using both pathology explanations and reference predictions from a weak learner.

tasks. They can also be fine-tuned for impressive performance on other tasks, such as segmentation and report generation. Compared with LVLMs, they are single-purpose models rather than generative AIs; however, these models may be used as the backbones of encoders in MLVLMs.

3. Methodology

Figure 1 illustrates the structure of common LVLMs. They are based on a pretrained unimodal LLM such as Llama (Touvron et al., 2023) and Vicuna (Chiang et al., 2023). A pretrained visual encoder, such as ViT (Dosovitskiy et al., 2021) or conventional CNNs, is applied to extract image features that are projected to the text feature space by an adapter. The projected visual features are concatenated with the text prompt embeddings and fed to the LLM. The adapter usually consists of several linear layers with nonlinear activations. The visual encoder and the LLM are usually frozen during training.

In our work, we choose the pretrained LLaVA-Med (Li et al., 2023a) as our model, which is a MLVLM built upon LLaVA (Liu et al., 2023b). The model structure resembles Figure 1. It uses pretrained Vicuna (Chiang et al., 2023) as the LLM and the pretrained ViT encoder from CLIP (Radford et al., 2021) as the visual encoder. The adapter is simply a trainable projection matrix. Both the visual encoder and LLM weights are frozen during training. LLaVA-Med fine-tunes LLaVA in two steps. First, it fine-tunes LLaVA to generate medical reports from input medical images. Second, it uses GPT-4 to generate various questions from the ground truth reports and fine-tunes the model to perform question answering.

Most MLVLMs are currently trained by medical VQA such that medical diagnosis can be performed by asking questions related to various pathologies; e.g., "Does this image have lung lesion?". To reduce model hallucination and improve VQA accuracy, we propose two prompting strategies at the inference stage: (1) providing the model with detailed explanations about the queried pathologies and (2) asking the model to consider the inferences of a weak learner.

3.1 Prompting With Detailed Explanations

Given imbalanced training data, MLVLMs might not adequately be able to learn the features of the minority pathologies. To compensate for insufficient training, we provide a detailed explanation of the queried pathology as a prompt at the inference stage. The explanation briefly defines the pathology and lists several key findings in medical images that may indicate its existence. An example is shown in Figure 2. The model is informed that Pulmonary Edema is defined as the accumulation of fluid in the lungs. Then several chest X-ray findings that may suggest its existence are provided. The model can determine if the given image has Pulmonary Edema by linking the given findings with the image features.

Prompt templates for a number of pathologies are listed in Section A.

3.2 Prompting With Detailed Explanations and Weak Learners

Data re-sampling is a commonly-used strategy to deal with imbalanced datasets that are responsible for the tendency of traditional image classification models to return negative predictions for minority pathologies. Models trained on re-sampled datasets often exhibit improvements in Precision and Recall scores; however, this strategy may not be suitable to MLVLMs for two reasons. First, it is difficult to balance a dataset containing many categories of pathologies. Second, MLVLMs usually demand much larger datasets and fine-tuning is also expensive.

One can nevertheless enable MLVLMs to benefit by leveraging small models trained on re-sampled datasets. Our method resembles multiagent LLM systems, such as

Pathology	+Cases
Atelectasis	64
Cardiomegaly	31
Consolidation	335
Edema	1,276
Pleural Effusion	260

Table 1: Positive (+) case counts for the 5 test pathologies in the LLaVA-Med training set.

Du et al. (2023), where multiple LLMs debate each other and hallucination can be corrected by referring to the generated outputs of other models. Given that traditional image classifiers are smaller, it is feasible to train multiple small classifiers each of which is trained on re-sampled datasets of a particular pathology. Those models can be further fine-tuned to optimize a single aspect, such as fewer False Positives (FPs) or fewer False Negatives (FNs). The classifiers are applied to the medical images and return preliminary predictions. These predictions are selectively included in the prompts as references for the MLVLM. Hence, MLVLMs can benefit indirectly from the nuanced understanding that these specialized models can provide. This method is meaningful because clinicians usually must balance the trade-off between overtreatment and undertreatment when making healthcare decisions. For instance, they may prefer models having a low FP rate if the cost of overtreatment is higher than that of undertreatment.

An example is shown in Figure 3, which queries about the presence of Edema. We first provide the model with the detailed explanation of Edema. Then, we use the weak learner to suppress the FPs. The image is input to an Edema classifier that has been fine-tuned on a balanced dataset for high sensitivity and high true negative (TN) rate. If its prediction is negative, we append after the pathology explanation the prompt "For this image, another agent thinks the probability of Edema is 0.1". Instead of using the actual predicted probability, the probability value is manually chosen because the decision threshold has been fine-tuned and is no longer 0.5. We do not use a zero probability value because we do not want the model overly to trust the weak learner. Although in this example our goal is only to reduce FPs, our strategy can also be applied to reduce FNs, simply by fine-tuning the classifier for a high True Positive (TP) rate and applying the prompt in the case of positive predictions.

4. Empirical Study

4.1 Datasets

LLaVA-Med is pretrained on the PMC-15M dataset (Zhang et al., 2024), which contains image-text pairs of multiple modalities; e.g., CT, MRI, X-ray, etc. In the first stage, 467,710 image-report pairs were selected for training. In the second stage, 56,708 question-answer pairs were created from the data of the first stage to fine-tune the model. Table 1 shows the count of reports in the LLaVA-Med training data (second stage) that mention one of the five test pathologies as positive. Relative to the total amount of data, all five categories are minorities.

To assess the zero-shot performance of the MLVLM, we used the MIMIC-CXR-JPG (Goldberger et al., 2000) and Chexpert (Irvin et al., 2019) chest X-ray test sets. They include 5,159 and 668 images, respectively. Neither dataset overlaps with PMC-15M.

MIMIC-CXR-JPG includes images and medical reports covering 13 categories of findings: Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomediastinum, Fracture, Lung Lesion, Lung Opacity, Pleural Effusion, Pneumonia, Pneumothorax, Pleural Other, and Support Devices. The raw reports are parsed and rough image-level tags are automatically generated by a rule-based approach (Irvin et al., 2019). Each label contains four values: 1 (positive), 0 (negative), -1 (uncertain), and missing. For simplicity, we treat both uncertain and missing as negative. We also use the MIMIC-CXR-JPG training set, which contains 227,827 chest X-rays with reports, to train the weak learner models.

Chexpert covers the same 13 categories as MIMIC-CXR-JPG. However, it does not include medical reports and has only image-level labels. There is no overlap between MIMIC-CXR-JPG and Chexpert.

Table 2 shows the split of pathology categories (excluding normal) in the MIMIC-CXR-JPG and Chexpert test sets. Clearly, almost all pathology categories are minor classes with much fewer positive than negative occurrences.

For our main testing regimen, we selected the five pathologies in the Chexpert Competition (Irvin et al., 2019): Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion.

4.2 Implementation Details

As was mentioned in Section 3, we use the pretrained LLaVA-Med MLVLM without any further fine-tuning. We convert the classification task into a VQA task by using the prompt template shown in Row 1 of Table 3, which we name Prompt Template 1 (PT1). We first run the pretrained LLaVA-Med with PT1. Next, we incorporate pathology explanations (Row 2 of Table 3), yielding Prompt Template 2 (PT2). Finally, we integrate the predictions of weak learners into the prompts (Row 3 of Table 3), resulting in Prompt Template 3 (PT3).

As will be justified by our experiments, our weak learner is designed to suppress FP predictions. To this end, we use

	MIMIC-CXR-JPG (5,159)		Chexpert (668)	
Category	+Cases	−Cases	+Cases	−Cases
Atelectasis	1,034	4,125	178	490
Cardiomegaly	1,258	3,901	175	493
Consolidation	326	4,833	35	633
Edema	959	4,200	85	583
Enlarged Cardiomediastinum	200	4,959	298	370
Fracture	167	4,992	6	662
Lung Lesion	202	4,957	14	654
Lung Opacity	1,561	3,598	310	358
Pleural Effusion	1,542	3,617	120	548
Pleural Other	119	5,040	8	660
Pneumonia	539	4,620	14	654
Pneumothorax	144	5,015	10	658
Support Devices	1,457	3,702	315	353

Table 2: Splits of positive (+) and negative (−) cases ('uncertain' is regarded as negative) for the 13 finding categories in the MIMIC-CXR-JPG and Chexpert test sets.

Prompt	Template
PT1	"Does this image have {target}?"
PT2	"{explanation} Given the information above, does this image have {target}?"
PT3	"{explanation} For this image, another agent thinks the probability that it has {target} is {n} percent. Given the information above, does this image have {target}?"

Table 3: The Prompt Templates (PTs). {target} is the pathology cited in the questions. {explanation} contains a pathology explanation among those listed in Section A. {n} is the probability associated with the weak learner.

the pretrained ResNet50 (He et al., 2016). Given the low cost of the weak learner, we train a model separately for each pathology with each training dataset sampled such that the ratio of positive and negative cases is $2:1$. The model was trained for 10 epochs with a $1e-4$ learning rate. The training process was monitored using the AUC score and the one with the highest validation AUC was kept. Then, the decision threshold d was fine-tuned to optimize a weighted sum of Specificity and Negative Predictive Value (NPV); i.e.,

$$d = w_1 \frac{\text{TN}}{\text{TN}+\text{FP}} + w_2 \frac{\text{TN}}{\text{TN}+\text{FN}}, \qquad (1)$$

where weights w_1 and w_2 are preset to 0.2 and 0.8, respectively. The medical images were input to the weak learners to obtain preliminary predictions for each pathology and only the negative predictions were selected to craft the PT3 prompts.

The responses returned by LLaVA-Med can take various forms, such as "This image has Edema", "Edema is found", "The fluid in the lung indicates Edema", etc. An off-the-shelf Llama-7B (Touvron et al., 2023) serves to summarize long responses into Yes/No answers such that accuracies could easily be computed.

4.3 Results

To demonstrate the efficacy of our prompting strategies, starting from the PT1 baseline, the pathology explanations were provided first (strategy PT2) and then, based on the results, weak learners were introduced to improve performance on specific aspects, resulting in strategy PT3.

PT2: Adding Pathology Explanations Table 4 reports Precision, Recall, and F1 scores of the PT1 and PT2 strategies on the MIMIC-CXR-JPG and Chexpert test sets.[1] On MIMIC-CXR-JPG, after adding pathology explanations, the F1 scores increased for detecting Atelectasis, Cardiomegaly,

[1] The AUC and ROC scores commonly reported in the literature to assess the performances of most medical image classification models on the MIMIC-CXR-JPG and Chexpert datasets are unsuitable in our context because MLVLMs output text rather than probabilities.

		MIMIC-CXR-JPG		Chexpert	
Pathology	Metric	PT1	PT2	PT1	PT2
Atelectasis	Precision	19.5	20.0	30.5	26.5
	Recall	41.5	92.9	44.4	91.6
	F1	26.5	33.0	36.5	41.0
Cardiomegaly	Precision	25.8	24.6	27.1	26.0
	Recall	22.5	89.4	20.0	86.3
	F1	24.0	38.6	23.0	40.0
Consolidation	Precision	6.8	6.3	6.0	5.2
	Recall	42.3	98.5	40.0	97.1
	F1	11.7	11.9	10.4	9.8
Edema	Precision	19.6	18.5	11.7	13.7
	Recall	36.0	72.7	29.4	76.5
	F1	25.4	29.5	16.8	23.2
Pleural Effusion	Precision	30.4	30.0	22.3	17.9
	Recall	42.8	92.7	49.2	90.0
	F1	35.6	45.3	30.7	29.9

Table 4: LLaVA-Med VQA performance evaluated by Precision, Recall, and F1 (%) scores of five pathologies on the MIMIC-CXR-JPG and Chexpert test datasets.

Pathology	TP	FP	FN
Atelectasis	163	453	15
Cardiomegaly	151	430	24
Consolidation	28	557	7
Edema	65	410	20
Pleural Effusion	108	495	12

Table 5: True positive (TP), false positive (FP), and false negative (FN) counts of LLaVA-Med with the PT2 strategy on the Chexpert test set.

Pathology	AUC	Precision	Recall	F1
Atelectasis	82.2	62.0	56.7	59.2
Cardiomegaly	85.0	74.4	38.3	50.6
Consolidation	81.7	100	2.9	5.6
Edema	87.3	46.5	61.2	54.2
Pleural Effusion	91.2	36.6	94.2	52.7

Table 6: Performance of the weak learner on the Chexpert test sets for the 5 pathologies.

Edema, and Pleural Effusion, albeit only minimally for Consolidation. On Chexpert, after adding pathology explanations, the F1 scores for detecting Atelectasis, Cardiomegaly, and Edema increased, whereas they did not for Consolidation and Pleural Effusion. The Precision and Recall scores reveal that adding explanations generally leads to a large increase in Recall, but only minimally influences Precision. For minority pathologies such as Consolidation whose F1 score is dominated by low Precision, improving the Recall would not have much effect. Thus, PT2's performance bottleneck is Precision.

PT3: Referring to Weak Learners Going beyond our PT2 strategy, we applied our PT3 strategy to further improve diagnostic accuracy. Table 5 provides the TP, FP, and FN prediction counts of LLaVA-Med on the Chexpert test set using the PT2 strategy. Note the large number of FP cases. Hence, we designed our weak learners to suppress FP predictions. As mentioned in Section 4.2, we trained the model for the highest AUC and then fine-tuned the decision threshold. Table 6 reports the AUC, Precision, Recall and F1 (after fine-tuning) of the weak learner. It is important to note that we fine-tuned the decision threshold to achieve high specificity and negative predictive value; thus, the reported Precision, Recall, and F1 scores are based on this fine-tuned threshold and may not be directly comparable to those of other models. Table 7 compares the performance on Chexpert before and after referring to the weak learner. It shows that the F1 prediction accuracy can be substantially increased by introducing weak learner predictions into the prompts. The F1 scores of Cardiomegaly, Edema, and Pleural Effusion increase by 0.115, 0.194 and 0.089, respectively. To further demonstrate the efficacy of our PT3 strategy, Table 8 compares the FP predictions of the PT2 and PT3 strategies. The reduction of FP cases is

Pathology	Metric	PT2	PT3
Atelectasis	Precision	26.5	28.8
	Recall	91.6	83.1
	F1	41.0	42.8
Cardiomegaly	Precision	26.0	38.1
	Recall	86.3	79.4
	F1	40.0	51.5
Consolidation	Precision	5.2	7.5
	Recall	97.1	34.3
	F1	9.8	12.2
Edema	Precision	13.7	36.8
	Recall	76.5	50.6
	F1	23.2	42.6
Pleural Effusion	Precision	17.9	25.0
	Recall	90.0	85.0
	F1	29.9	38.8

Table 7: Diagnostic accuracies of LLaVA-Med with the PT2 and PT3 strategies on the Chexpert test set.

noteworthy, especially on Edema, for which the FP count is reduced by 78.5% (322).

Additional VQA Experiments Table 9 shows the results of applying the PT1, PT2, and PT3 strategies with LLaVA-Med on the MIMIC-CXR-JPG and Chexpert datasets across another five medical findings: Enlarged Cardiomediastinum, Lung Lesion, Lung Opacity, Pneumonia, and Pneumothorax. Providing pathology explanations (PT2) generally yields better results over the PT1 baseline, albeit inconsistently. Introducing weak learner references (PT3) yields only limited increases in Precision, but large decreases in Recall. Generally, it offers insignificant improvement. Enlarged Cardiomediastinum, Lung Lesion, Pneumonia, and Pneumothorax are minor categories and all our experimental settings, including for the weak learner, fail to learn them. Prompting is apparently unhelpful in such situations.

SOTA Benchmark Tiu et al. (2022) report F1 scores for detecting Atelectasis, Cardiomegaly, Consolidation, Edema, and Pleural Effusion on the Chexpert dataset using their deep learning model, as well as for the performance of radiologists. Their work offers a state-of-the-art chest X-ray diagnosis benchmark. Table 10 compares the F1 scores of radiologists, the model of Tiu et al. (2022), and LLaVA-Med. It shows that LLaVA-Med's VQA performance of with the baseline PT1 strategy is unsatisfactory, rendering the model far from being deployable in clinical practice. However, while still underperforming radiologists, our PT3 strategy yields a significant improvement, especially on Atelectasis, Cardiomegaly, and Edema for which the F1 score increases by approximately 17% to 21%.

Pathology	PT2	PT3
Atelectasis	453	365
Cardiomegaly	430	226
Consolidation	557	149
Edema	410	88
Pleural Effusion	495	304

Table 8: False positive counts of LLaVA-Med with the PT2 and PT3 strategies on the Chexpert test set.

Application to General Domain LVLMs Our prompt strategies can also be applied to general domain LVLMs. We studied the performance of LLaVA (Liu et al., 2023b) and MiniGPT-v2 (Zhu et al., 2023) using POPE metrics (Li et al., 2023c). POPE evaluates the hallucination of LVLMs by asking questions about the existing/non-existing objects on given images. The input images are from MSCOCO dataset and there are three questions categories: Random (ramdom sample objects for questions), Popular (frequent objects), and Adversarial (frequent but non-existent objects). The performance is evaluated by the Precision, Recall and F1 of correctly answered questions. The POPE scores of LLaVA and MiniGPT-v2 have high Precision and low Recall. Hence, our weak learner strategy can be used to reduce the FN predictions. We selected an off-the-shelf Fast-RCNN (Girshick, 2015) as the weak learner, fine-tuned the detection threshold of bounding box scores to achieve high Recall, and introduced the positive predictions of the weak learner into the prompts. The results in Table 11 show that the Recall scores across three POPE categories increased by around 7% (Precision scores decrease slightly), thus improving the F1 scores.

5. Conclusions and Discussion

We have tested the visual question answering abilities of the LLaVA-Med medical large vision-language model when applied to the diagnosis of pathologies. Our results show that the model has unsatisfactory performance when asked questions regarding the presence of complex pathologies. We proposed two prompt engineering strategies to improve the visual question answering accuracy of the model: providing explanations of pathologies and referring to the predictions of weak learners. The first strategy helps the model understand minority pathologies that it does not learn well in the training stage. The second strategy can help improve diagnostic accuracy in specific ways; e.g., by suppressing false positives. This strategy can also be applied to LVLMs in other, non-medical domains.

However, our two strategies are not effective on pathologies with extremely scarce data. For example, providing

		MIMIC-CXR-JPG			Chexpert		
Pathology	Metric	PT1	PT2	PT3	PT1	PT2	PT3
Enlarged Cardiomediastinum	Precision	4.3	3.9	5.0	49.3	44.1	49.4
	Recall	15.0	89.0	53.5	12.4	85.2	59.1
	F1	6.7	7.4	9.1	19.8	58.1	53.8
Lung Lesion	Precision	3.9	3.9	4.2	2.0	2.1	2.9
	Recall	77.2	100.0	66.3	71.4	100.0	92.9
	F1	7.4	7.5	8.0	3.9	4.1	5.7
Lung Opacity	Precision	31.4	30.4	31.9	50.0	47.2	51.4
	Recall	67.6	88.8	84.1	70.3	90.7	84.8
	F1	42.8	45.3	46.2	58.5	62.1	63.3
Pneumonia	Precision	11.4	10.5	12.3	2.4	1.7	6.3
	Recall	20.0	74.6	18.0	21.4	57.1	28.6
	F1	14.6	18.4	14.6	4.4	3.4	10.4
Pneumothorax	Precision	3.0	2.6	3.6	0.0	1.7	2.0
	Recall	16.7	78.5	50.7	0.0	90.0	50.0
	F1	5.1	5.1	6.8	0.0	3.3	3.8

Table 9: LLaVA-Med VQA performance on the MIMIC-CXR-JPG and Chexpert test sets for another 5 pathologies.

Pathology	Radiologist	(Tiu et al., 2022)	PT1	PT3
Atelectasis	69.2	64.6	26.5	41.3
Cardiomegaly	67.8	74.3	24.0	51.5
Consolidation	38.5	33.3	11.7	12.2
Edema	58.3	60.2	25.4	42.6
Pleural Effusion	73.7	70.4	35.5	46.8

Table 10: F1 scores (%) on 5 pathologies in the Chexpert test set, including for the radiologist diagnoses, the state-of-the-art benchmark (Tiu et al., 2022), as well as LLaVA-Med VQA with the PT1 scenario and the PT3 scenario, which is the best result achieved by applying both our prompting strategies.

	POPE Adversarial			POPE Popular			POPE Random		
Model	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
LLaVA	91.0	78.8	84.5	95.2	78.8	86.2	97.4	78.8	87.1
with referral	88.4	85.7	87.0	92.8	85.7	89.0	97.3	85.7	91.1
MiniGPT-v2	88.2	77.2	82.3	92.7	77.2	84.2	97.2	77.2	86.1
with referral	86.8	84.2	85.5	91.9	84.2	87.9	97.3	84.2	90.3

Table 11: Comparison of POPE scores for LVLM models with and without referring to the predictions of weak learners.

text explanations for Consolidation, Fracture, Lung Lesion, Pneumonia, and Pneumothorax may not suffice since the visual encoder does not adequately learn meaningful visual features. Moreover, the data may not suffice to adequately train weak learners. A promising direction for future research would be to devise a strategy for handling these rare categories. Retrieval Augmented Generation (RAG) could be a potential solution. For instance, in addition to textual explanations of pathologies, typical example images can be provided to help the model make diagnostic decisions.

Acknowledgments

DG was supported in part by an unrestricted donation to UCLA from VoxelCloud, Inc.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We have no conflicts of interest.

Data Availability

Only publicly available datasets were used.

Appendix A. Pathology Explanations

The explanations of the five pathologies are as follows:

Atelectasis: Atelectasis refers to the partial or complete collapse of a lung or a section of lung. The features of atelectasis on an X-ray can vary depending on the cause and extent of the collapse. Some common X-ray features include: 1. The affected area may appear denser or whiter than normal lung tissue due to the collapse, leading to increased opacity on the X-ray. 2. The affected portion of the lung may appear smaller or compressed compared to the surrounding healthy lung tissue. 3. Atelectasis can cause a shift or displacement of nearby structures, such as the trachea or heart, toward the affected area. 4. In obstructive atelectasis (caused by a blockage in the airways), there might be signs of hyperinflation in the unaffected areas of the lung and a visible blockage or narrowing in the affected bronchus. 5. Linear or band-like opacities may be visible, often referred to as plate or band atelectasis, which can occur due to the collapse of small airways. Given the information above, does this image have Atelectasis?

Cardiomegaly: Cardiomegaly is enlargement of the heart. The definition is when the transverse diameter of the cardiac silhouette is greater than or equal to 50% of the transverse diameter of the chest (increased cardiothoracic ratio) on a posterior-anterior projection of a chest radiograph or a computed tomography. Given the information above, does this image have Cardiomegaly?

Consolidation: Consolidation on an X-ray refers to the filling of the lung's air spaces with fluid inflammatory exudate, or cellular material. Typical X-ray findings suggesting consolidation include: 1. Areas of increased density in the lung tissue, appearing as an opaque or hazy patch on the X-ray. Given the information above, does this image have Consolidation?

Edema: Pulmonary edema is the accumulation of fluid in the lungs. Some common X-ray features include: 1. Increased density in the central lung fields resembling the shape of bat wings. 2. Thin, linear opacities at the lung periphery, often indicating interstitial edema. 3. Prominent blood vessel markings due to engorgement from increased pressure in the pulmonary vasculature. Given the information above, does this image have Edema?

Pleural Effusion: Pleural effusion is the accumulation of fluid in between the parietal and visceral pleura. Some common X-ray features include: 1. blunting of the costophrenic / cardiophrenic angle. 2. fluid within the horizontal or oblique fissures. 3. meniscus is seen. 4. mediastinal shift occurs away from the effusion.

References

D. Caffagni, F. Cocchi, N. Moratelli, S. Sarto, M. Cornia, L. Baraldi, and R. Cucchiara. Wiki-LLaVA: Hierarchical retrieval-augmented generation for multimodal LLMs. *arXiv Preprint arXiv:2404.15406*, 2024.

Z. Chen, M. Varma, J.-B. Delbrouck, M. Paschali, L. Blankemeier, D. V. Veen, J. M. J. Valanarasu, A. Youssef, J. P. Cohen, E. P. Reis, E. B. Tsai, A. Johnston, C. Olsen, T. M. Abraham, S. Gatidis, A. S. Chaudhari, and C. Langlotz. CheXagent: Towards a foundation model for chest X-ray interpretation. *arXiv preprint arXiv:2401.12208*, 2024.

W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90% ChatGPT quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Y. Du, S. Li, A. Torralba, J. B. Tenenbaum, and I. Mordatch. Improving factuality and reasoning in language models through multiagent debate. *arXiv Preprint arXiv:2305.14325*, 2023.

A. Favero, L. Zancato, M. Trager, S. Choudhary, P. Perera, A. Achille, A. Swaminathan, and S. Soatto. Multi-modal hallucination control by visual information grounding. *arXiv Preprint arXiv:2403.14003*, 2024.

R. Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

A. L. Goldberger, L. A. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23): e215–e220, 2000.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie. PathVQA: 30000+ questions for medical visual question answering. *arXiv Preprint arXiv:2003.10286*, 2020.

H. Hu, J. Zhang, M. Zhao, and Z. Sun. CIEM: Contrastive instruction evaluation method for better instruction tuning. *arXiv Preprint arXiv:2309.02301*, 2023.

J. A. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. L. Ball, K. S. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. Langlotz, B. N. Patel, M. P. Lungren, and A. Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI Conference on Artificial Intelligence*, 2019.

J. Jain, J. Yang, and H. Shi. VCoder: Versatile vision encoders for multimodal large language models. *arXiv Preprint arXiv:2312.14233*, 2023.

J. Jang, D. Kyung, S. Kim, H. Lee, K. Bae, and E. Choi. Significantly improving zero-shot X-ray pathology classification via fine-tuning pre-trained image-text encoders. *ArXiv*, abs/2212.07050, 2022.

J. J. Lau, S. Gayen, A. Ben Abacha, and D. Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5(1):1–10, 2018.

S. Leng, H. Zhang, G. Chen, X. Li, S. Lu, C. Miao, and L. Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv Preprint arXiv:2311.16922*, 2023.

C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *arXiv Preprint arXiv:2306.00890*, 2023a.

J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023b.

Y. Li, Y. Du, K. Zhou, J. Wang, X. Zhao, and J.-R. Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing*, pages 292–305, Dec. 2023c.

B. Liu, L.-M. Zhan, L. Xu, L. Ma, Y. Yang, and X.-M. Wu. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1650–1654, 2021.

F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, pages 1–45, 2023a.

H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In *Neural Information Processing Systems*, 2023b.

H. Liu, W. Xue, Y. Chen, D. Chen, X. Zhao, K. Wang, L. Hou, R. Li, and W. Peng. A survey on hallucination in large vision-language models. *arXiv Preprint arXiv:2402.00253*, 2024.

Y. Liu, M. Wang, X. Liu, W. Chang, M. Sun, and P. Li. Large language models encode clinical knowledge. *Nature*, 603:589–593, 2023c.

H. Lovenia, W. Dai, S. Cahyawijaya, Z. Ji, and P. Fung. Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models. *arXiv Preprint arXiv:2310.05338*, 2023.

OpenAI. ChatGPT blog, 2022. `https://openai.com/blog/chatgpt`.

OpenAI et al. GPT-4 technical report. *arXiv Preprint arXiv:2303.08774*, 2023.

C. Pellegrini, M. Keicher, E. Özsoy, P. Jiraskova, R. Braren, and N. Navab. Xplainer: From X-ray observations to explainable zero-shot diagnosis. In *Medical Image Computing and Computer Assisted Intervention — MICCAI 2023*, pages 420–429, Cham, 2023a. Springer Nature Switzerland.

C. Pellegrini, E. Özsoy, B. Busam, N. Navab, and M. Keicher. RaDialog: A large vision-language model for radiology report generation and conversational assistance. *arXiv preprint arXiv:2311.18681*, 2023b.

B. Peng, M. Galley, P. He, H. Cheng, Y. Xie, Y. Hu, Q. Huang, L. Liden, Z. Yu, W. Chen, and J. Gao. Check your facts and try again: Improving large language models with external knowledge and automated feedback. *arXiv preprint arXiv:2302.12813*, 2023.

Y. Qian, H. Zhang, Y. Yang, and Z. Gan. How easy is it to fool your multimodal LLMs? An empirical analysis on deceptive prompts. *arXiv Preprint arXiv:2402.13220*, 2024.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763, 2021.

A. Rohrbach, L. A. Hendricks, K. Burns, T. Darrell, and K. Saenko. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium, 2018.

C. Seibold, S. Reiß, M. S. Sarfraz, R. Stiefelhagen, and J. Kleesiek. Breaking with fixed set pathology recognition through report-guided contrastive training. In *Medical Image Computing and Computer Assisted Intervention — MICCAI 2022*, pages 690–700, Cham, 2022. Springer Nature Switzerland.

K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, M. Schaekermann, A. Wang, M. Amin, S. Lachgar, P. Mansfield, S. Prakash, B. Green, E. Dominowska, B. A. y Arcas, N. Tomasev, Y. Liu, R. Wong, C. Semturs, S. S. Mahdavi, J. Barral, D. Webster, G. S. Corrado, Y. Matias, S. Azizi, A. Karthikesalingam, and V. Natarajan. Towards expert-level medical question answering with large language models. *arXiv Preprint arXiv:2305.09617*, 2023.

Z. Sun, S. Shen, S. Cao, H. Liu, C. Li, Y. Shen, C. Gan, L.-Y. Gui, Y.-X. Wang, Y. Yang, K. Keutzer, and T. Darrell. Aligning large multimodal models with factually augmented RLHF. *arXiv Preprint arXiv:2309.14525*, 2023.

O. Thawakar, A. M. Shaker, S. S. Mullappilly, H. Cholakkal, R. M. Anwer, S. S. Khan, J. Laaksonen, and F. S. Khan. XrayGPT: Chest radiographs summarization using large medical vision-language models. In *Workshop on Biomedical Natural Language Processing*, 2023.

E. Tiu, E. Talius, P. Patel, C. P. Langlotz, A. Y. Ng, and P. Rajpurkar. Expert-level detection of pathologies from unannotated chest X-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.

H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. LLaMA: Open and efficient foundation language models. *arXiv Preprint arXiv:2302.13971*, 2023.

T. Vu, M. Iyyer, X. Wang, N. Constant, J. Wei, J. Wei, C. Tar, Y.-H. Sung, D. Zhou, Q. Le, and T. Luong. FreshLLMs: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics (ACL)*, pages 13697–13720, 2024.

B. Wang, S. Min, X. Deng, J. Shen, Y. Wu, L. Zettlemoyer, and H. Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, 2023.

Q. Wang, R. Ji, T. Peng, W. Wu, Z. Li, and J. Liu. Soft knowledge prompt: Help external knowledge become a better teacher to instruct LLM in knowledge-based VQA. In *Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6132–6143, 2024.

Z. Wang et al. MedCLIP: Contrastive learning from unpaired medical images and text. In *Conference on Empirical Methods in Natural Language Processing*, pages 3876–3887, 2022.

R. Zhang, J. Han, C. Liu, P. Gao, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, and Y. Qiao. LLaMA-Adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv Preprint arXiv:2303.16199*, 2023a.

S. Zhang, Y. Xu, N. Usuyama, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, M. P. Lungren, T. Naumann, and H. Poon. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv Preprint arXiv:2303.00915*, 2023b.

S. Zhang, Y. Xu, N. Usuyama, H. Xu, J. Bagga, R. Tinn, S. Preston, R. Rao, M. Wei, N. Valluri, C. Wong, A. Tupini, Y. Wang, M. Mazzola, S. Shukla, L. Liden, J. Gao, M. P. Lungren, T. Naumann, S. Wang, and

H. Poon. BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. *arXiv Preprint arXiv:2303.00915*, 2024.

Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, volume 182 of *Proceedings of Machine Learning Research*, pages 2–25, 2022.

L. Zhao, Y. Deng, W. Zhang, and Q. Gu. Mitigating object hallucination in large vision-language models via classifier-free guidance. *arXiv Preprint arXiv:2402.08680*, 2024.

G. Zheng, B. Yang, J. Tang, H.-Y. Zhou, and S. Yang. DDCoT: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 5168–5191, 2023.

Y. Zhou, C. Cui, J. Yoon, L. Zhang, Z. Deng, C. Finn, M. Bansal, and H. Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, 2024.

D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv Preprint arXiv:2304.10592*, 2023.