

Lecture 10: Vision-Language Generative Models in Biomedicine

Announcements

- Project proposal is due Wed Oct 23
- A2 will be released Oct 30, due Nov 13
- Discussion presentations start Mon Nov 4
- Before starting discussions, we will summarize and synthesize what we have seen in class so far at the start of the Wed Oct 30 lecture

Finishing up from last lecture: Vision-Language Generative Models

Model Choice

Fully Open-Sourced Large VLMs (Data+Training+Inference):

- ❖ Cambrian-1

Open-Sourced Large VLMs (Inference Only):

- ❖ Qwen2-VL

Proprietary Large VLMs (API Only):

- ❖ Image: GPT-4o
- ❖ Video: Gemini-1.5 Pro

2024/10 ; My Personal Opinion + Benchmark Results

How to Link Vision to LLMs?

Internal Linkage → Vision-Language Models

- ❖ Architecture
- ❖ Evaluation
- ❖ Training Recipe
- ❖ Improvement
- ❖ Takeaway

External Linkage → Vision-Language Agents

How to Link Vision to LLMs?

Internal Linkage → Vision-Language Models

- ❖ Architecture
- ❖ Evaluation
- ❖ Training Recipe
- ❖ Improvement
- ❖ Takeaway

External Linkage → Vision-Language Agents

- ❖ Language as Interface

Open-Source Computer Vision Models

Computer Vision

- Depth Estimation
- Image Classification
- Object Detection
- Image Segmentation
- Text-to-Image
- Image-to-Text
- Image-to-Image
- Image-to-Video
- Unconditional Image Generation
- Video Classification
- Text-to-Video
- Zero-Shot Image Classification
- Mask Generation
- Zero-Shot Object Detection
- Text-to-3D
- Image-to-3D
- Image Feature Extraction
- Keypoint Detection



Hugging Face

```
from transformers import DetrImageProcessor, DetrForObjectDetection
import torch
from PIL import Image
import requests

url = "http://images.cocodataset.org/val2017/000000039769.jpg"
image = Image.open(requests.get(url, stream=True).raw)

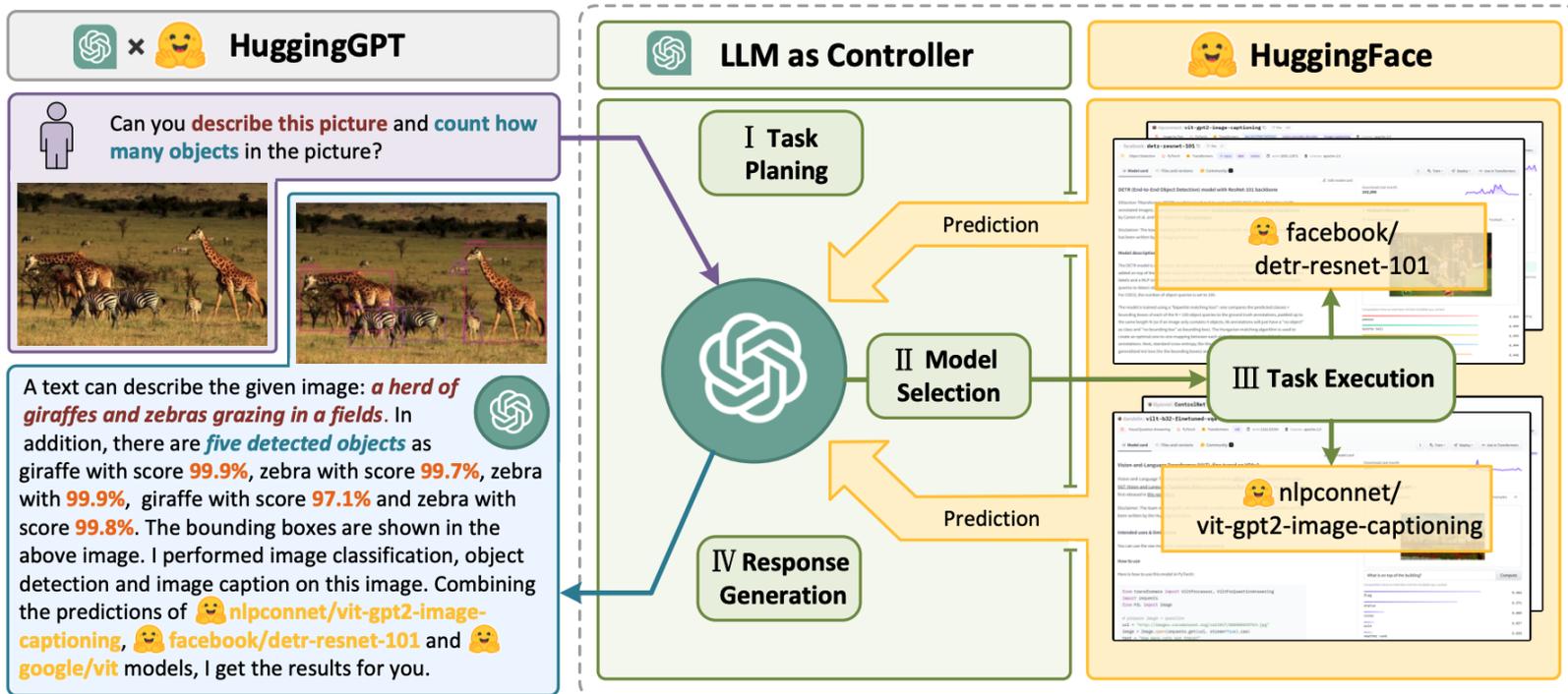
# you can specify the revision tag if you don't want the timm depende
processor = DetrImageProcessor.from_pretrained("facebook/detr-resnet-
model = DetrForObjectDetection.from_pretrained("facebook/detr-resnet-

inputs = processor(images=image, return_tensors="pt")
outputs = model(**inputs)

# convert outputs (bounding boxes and class logits) to COCO API
# let's only keep detections with score > 0.9
target_sizes = torch.tensor([image.size[:-1]])
results = processor.post_process_object_detection(outputs, target_size

for score, label, box in zip(results["scores"], results["labels"], re
    box = [round(i, 2) for i in box.tolist()]
    print(
        f"Detected {model.config.id2label[label.item()]} with con
        f"{round(score.item(), 3)} at location {box}"
    )
```

Language to Connect Vision Models



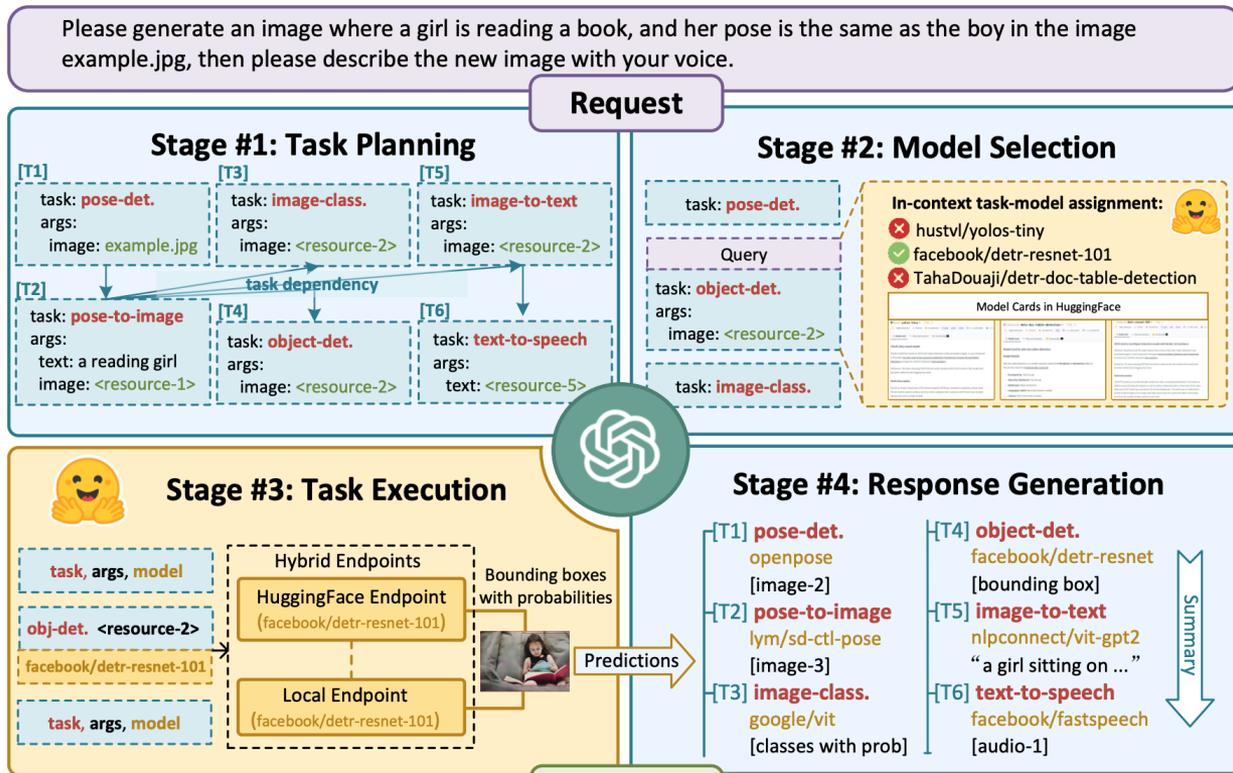
HuggingGPT (Shen et al. 2023)

Serena Yeung-Levy
Xiaohan Wang

BIODS 276: Adv. Topics in CV and Biomedicine

Lecture 10 - 8

Language to Connect Vision Models



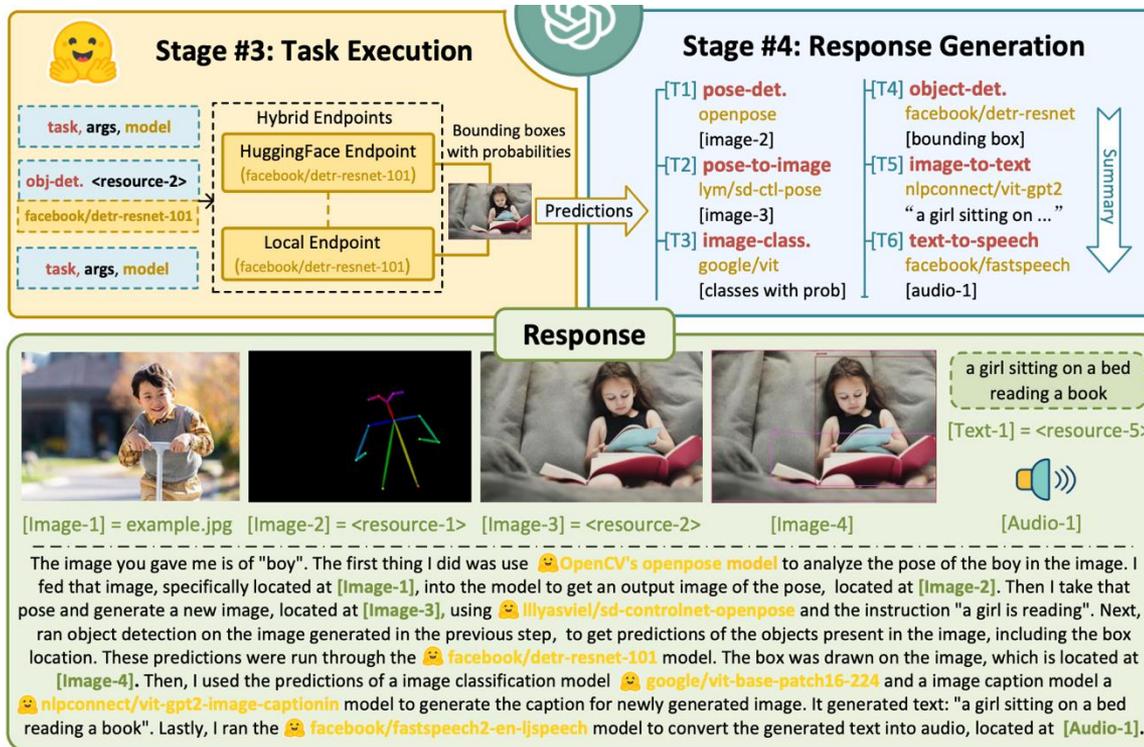
HuggingGPT (Shen et al. 2023)

Serena Yeung-Levy
Xiaohan Wang

BIODS 276: Adv. Topics in CV and Biomedicine

Lecture 10 - 9

Language to Connect Vision Models



HuggingGPT (Shen et al. 2023)

Serena Yeung-Levy
Xiaohan Wang

BIODS 276: Adv. Topics in CV and Biomedicine

Lecture 10 - 10

How to Link Vision to LLMs?

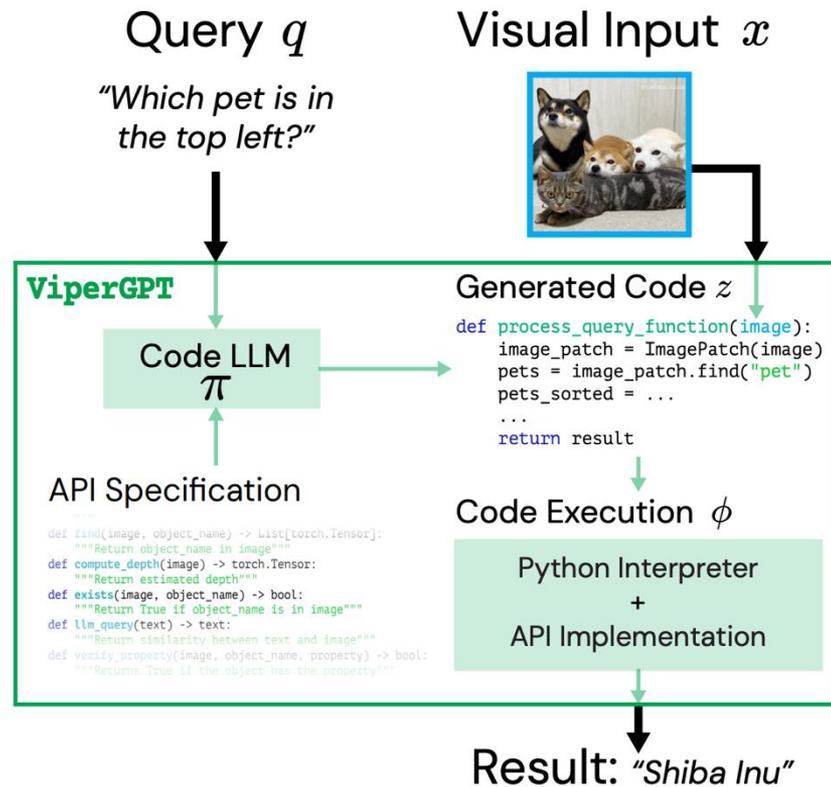
Internal Linkage → Vision-Language Models

- ❖ Architecture
- ❖ Evaluation
- ❖ Training Recipe
- ❖ Improvement
- ❖ Takeaway

External Linkage → Vision-Language Agents

- ❖ Language as Interface
- ❖ Code as Interface

Code to Connect Vision Models



ViperGPT (Surís et al. 2023)

Serena Yeung-Levy
Xiaohan Wang

BIODS 276: Adv. Topics in CV and Biomedicine

Lecture 10 - 12

Code to Connect Vision Models

find (GLIP):

- Inputs: image, noun phrase (e.g., "car"); Output: list of image patches containing the noun phrase

exists (GLIP):

- Inputs: image, noun phrase; Output: boolean indicating if the noun phrase is present

verify_property (XVLM):

- Inputs: image, object (noun phrase), property (attribute); Output: boolean indicating if the property is present

best_image_match (X-VLM):

- Inputs: list of image patches, noun phrase; Output: best matching image patch for the noun phrase

best_text_match (X-VLM):

- Inputs: list of noun phrases, image; Output: best matching noun phrase for the image

compute_depth (MiDaS):

- Input: image patch; Output: median depth of the image patch

distance:

- Computes pixel-distance between two patches using Python built-in tools

Code to Connect Vision Models

Query: How many muffins can each kid have for it to be fair?



Generated Code

```
def execute_command(image):  
    image_patch = ImagePatch(image)  
    muffin_patches = image_patch.find("muffin")  
    kid_patches = image_patch.find("kid")  
    return str(len(muffin_patches) // len(kid_patches))
```

Execution

```
muffin_patches =  
image_patch.find("muffin")
```



```
kid_patches =  
image_patch.find("kid")
```



```
► len(muffin_patches)=8  
► len(kid_patches)=2
```

```
► 8//2 = 4
```

Result: 4

Query: Drink with zero alcohol



```
def execute_command(image):  
    image_patch = ImagePatch(image)  
    drink_patches = image_patch.find("drink")  
    for drink_patch in drink_patches:  
        drink_name = drink_patch.simple_query("What is this?")  
        alcoholic = llm_query(f"Does the {drink_name} have alcohol?")  
        if alcoholic == "no":  
            return drink_patch  
    return None
```

► drink_patches=



```
► drink_name = 'tullamore dew'  
► alcoholic = 'yes'
```

```
► drink_name = 'bacardi'  
► alcoholic = 'yes'
```

```
► drink_name = 'gin'  
► alcoholic = 'yes'
```

```
► drink_name = 'dr pepper'  
► alcoholic = 'no'
```

Result:



Code to Connect Vision Models

Query: What did the boy do after he dropped the sparkles on the floor?

Generated code

```
def execute_command(video, question, possible_answers):
    video_segment = VideoSegment(video)
    drop_detected = False
    for i, frame in enumerate(video_segment.frame_iterator()):
        if frame.exists("boy") and frame.exists("sparkles") and \
            frame.simple_query("is the boy dropping the sparkles?") == "yes":
            drop_detected = True
            break
    if drop_detected:
        index_frame = i + 1
    else:
        index_frame = video_segment.num_frames // 2
    frame_of_interest = ImagePatch(video_segment, index_frame)
    boys = frame_of_interest.find("boy")
    if len(boys) == 0:
        boys = [frame_of_interest]
    boy = boys[0]
    caption = boy.simple_query("What is this?")
    info = {
        "Caption of frame after dropping the sparkles": caption,
    }
    answer = select_answer(info, question, possible_answers)
    return answer
```

Execution

In:



```
frame.exists("boy") and \
frame.exists("sparkles") and \
frame.simple_query("is the boy
dropping the sparkles?") == "yes":
▶ frame = {ImagePatch}
```



```
▶ i = {int} 25
```

```
index_frame = i + 1
```

```
▶ index_frame = {int} 26
▶ frame_of_interest = {ImagePatch}
```



```
boys = frame_of_interest.find("boy")
```

```
▶ boy = {ImagePatch}
```



```
▶ caption = {str} "a child running
with fire in his hands"
```

```
▶ answer = {str} "pick it up"
```

Result: "Pick it up"

Code to Connect Vision Models

Table 3. **OK-VQA Results.**

		Accuracy (%) ↑
Sup.	TRiG [14]	50.5
	KAT [17]	54.4
	RA-VQA [33]	54.5
	REVIVE [34]	58.0
	PromptCap [22]	58.8
ZS	PNP-VQA [54]	35.9
	PICa [63]	43.3
	BLIP-2 [31]	45.9
	Flamingo [1]	50.6
	ViperGPT (ours)	51.9

Table 4. **NExT-QA Results.** Our method gets overall state-of-the-art results (including *supervised* models) on the hard split. “T” and “C” stand for “temporal” and “causal” questions, respectively.

		Accuracy (%) ↑		
		Hard Split - T	Hard Split - C	Full Set
Sup.	ATP [7]	45.3	43.3	54.3
	VGT [61]	-	-	56.9
	HiTeA [64]	48.6	47.8	63.1
ZS	ViperGPT (ours)	49.8	56.4	60.0

How to Link Vision to LLMs?

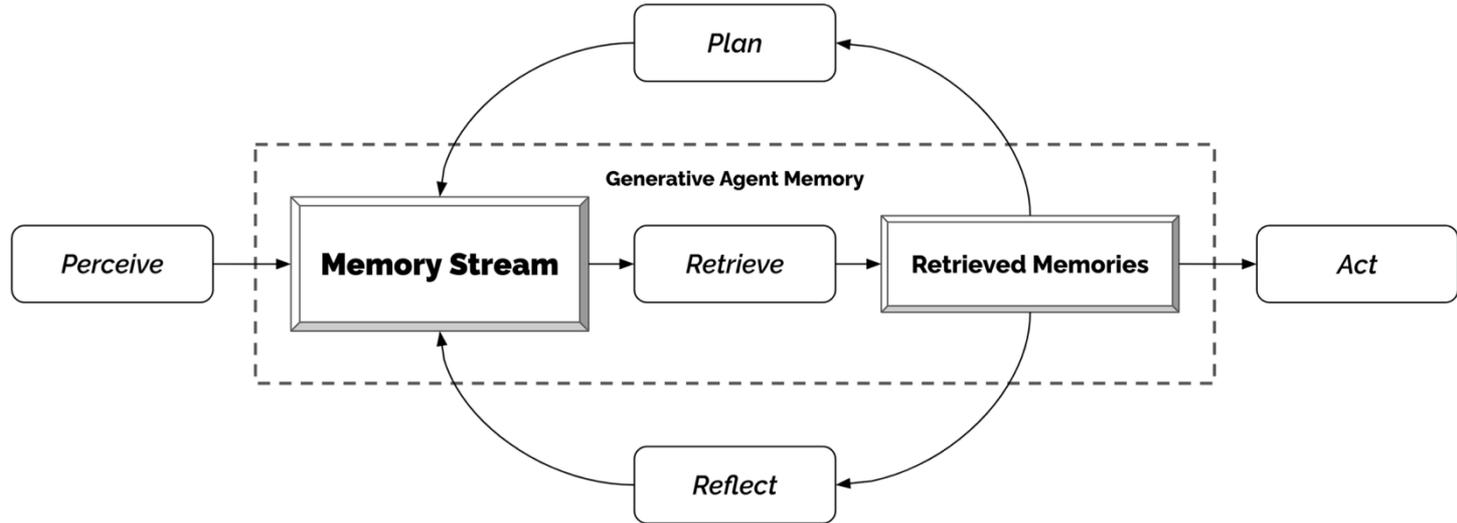
Internal Linkage → Vision-Language Models

- ❖ Architecture
- ❖ Evaluation
- ❖ Training Recipe
- ❖ Improvement
- ❖ Takeaway

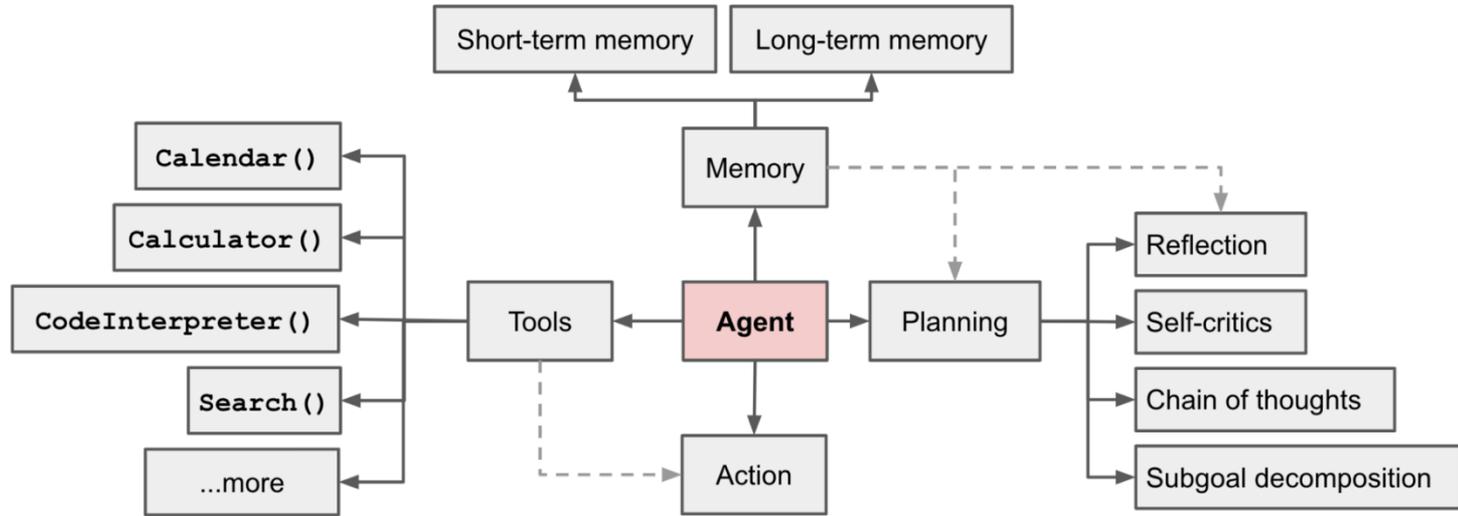
External Linkage → Vision-Language Agents

- ❖ Language as Interface
- ❖ Code as Interface
- ❖ Takeaway

The Generative Agent Architecture



Overview of Agent System



What's next?

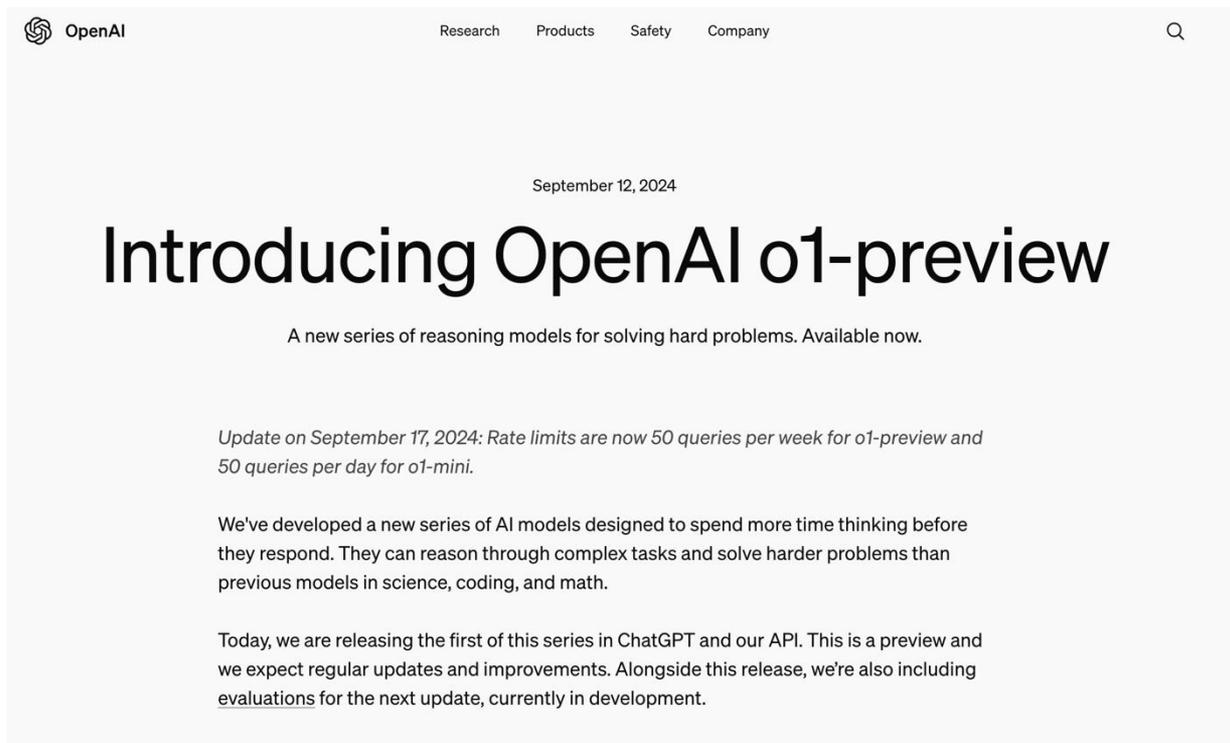
Internal Linkage → Vision-Language Models

- ❖ Architecture
- ❖ Evaluation
- ❖ Training Recipe
- ❖ Improvement
- ❖ Takeaway

External Linkage → Vision-Language Agents

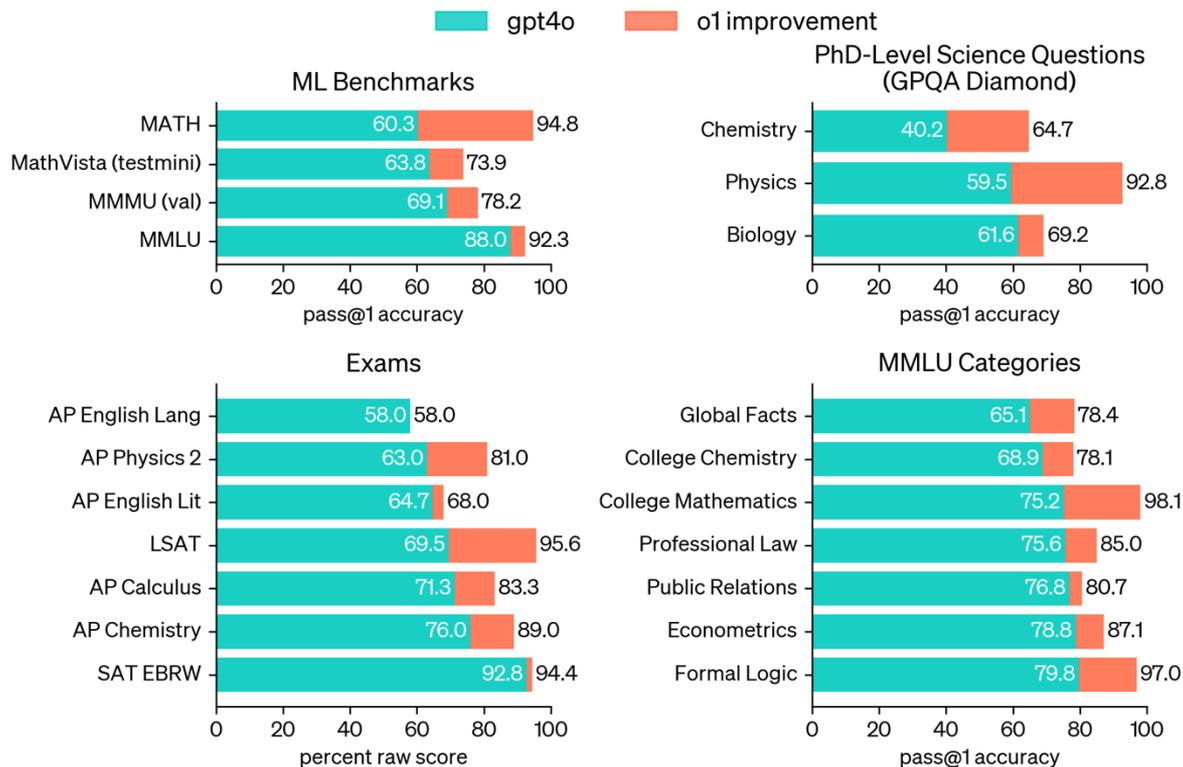
- ❖ Language as Interface
- ❖ Code as Interface
- ❖ Takeaway

Learning to Reason



The image is a screenshot of the OpenAI website. At the top left is the OpenAI logo. To its right are navigation links for 'Research', 'Products', 'Safety', and 'Company'. Further right is a search icon. The main content area features a date 'September 12, 2024' centered above a large heading 'Introducing OpenAI o1-preview'. Below the heading is a sub-headline: 'A new series of reasoning models for solving hard problems. Available now.' A paragraph of italicized text follows: 'Update on September 17, 2024: Rate limits are now 50 queries per week for o1-preview and 50 queries per day for o1-mini.' Another paragraph states: 'We've developed a new series of AI models designed to spend more time thinking before they respond. They can reason through complex tasks and solve harder problems than previous models in science, coding, and math.' The final paragraph reads: 'Today, we are releasing the first of this series in ChatGPT and our API. This is a preview and we expect regular updates and improvements. Alongside this release, we're also including evaluations for the next update, currently in development.'

Learning to Reason



Agent for GUI Control

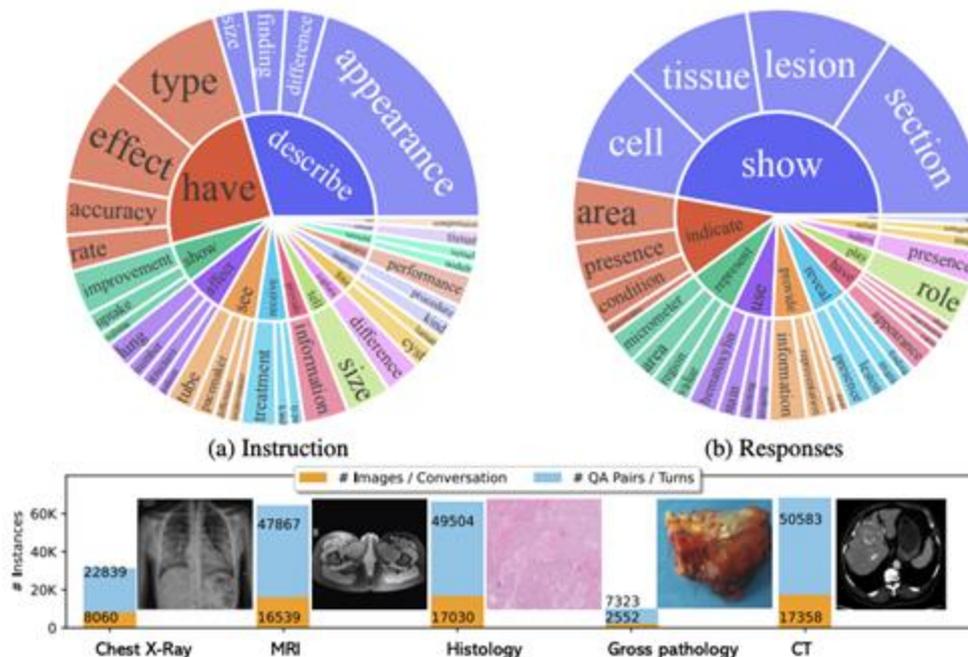


Next:
Vision-Language Generative Models in
Biomedicine

Generalist (broad domain) biomedical VLMs

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

- Extends LLaVA to better answer biomedical questions, using a new biomedical instruction tuning dataset
- Instruction tuning dataset leverages PMC-15M (PubMed Central figure-caption dataset) and covers diverse domains
- Efficiently trained in < 15 hours using eight A100s

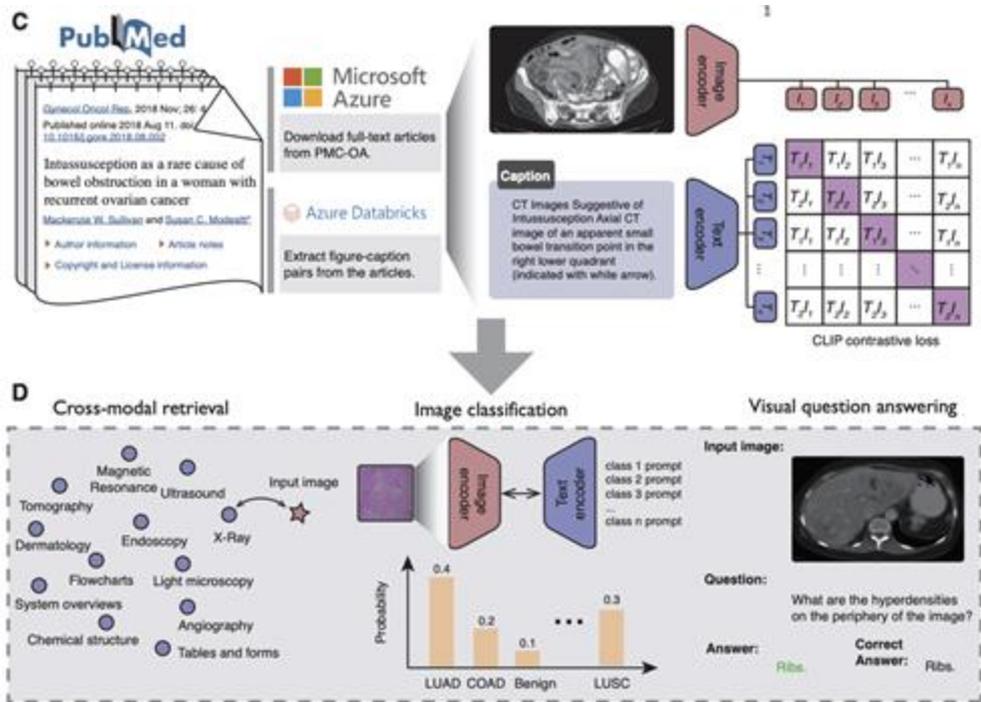


Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

Remember: PMC-15M from BiomedCLIP (Generalist foundation model trained on PubMed)

PMC-15M: 15 million image-caption pairs from 4.4 million publicly available full-text articles in PubMed Central

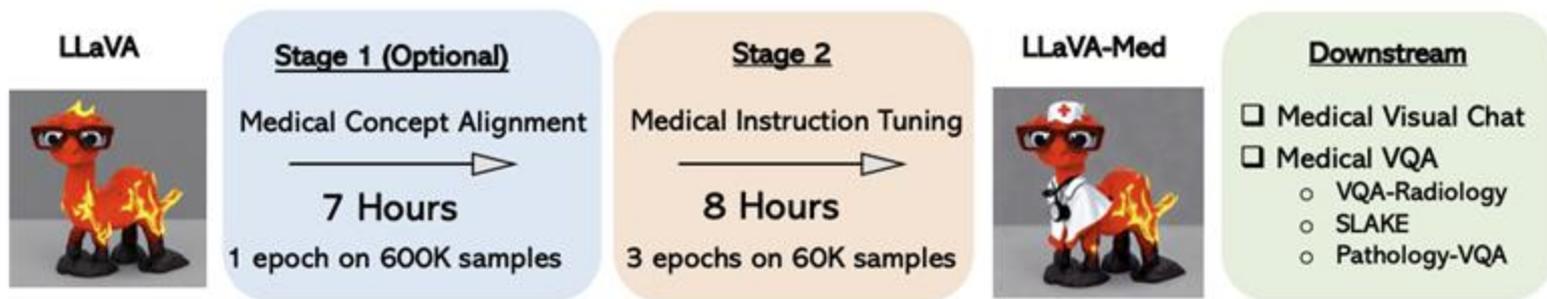
Complete article packages are downloaded, and figure files and matching captions are extracted



Zhang et al. BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv, 2024.

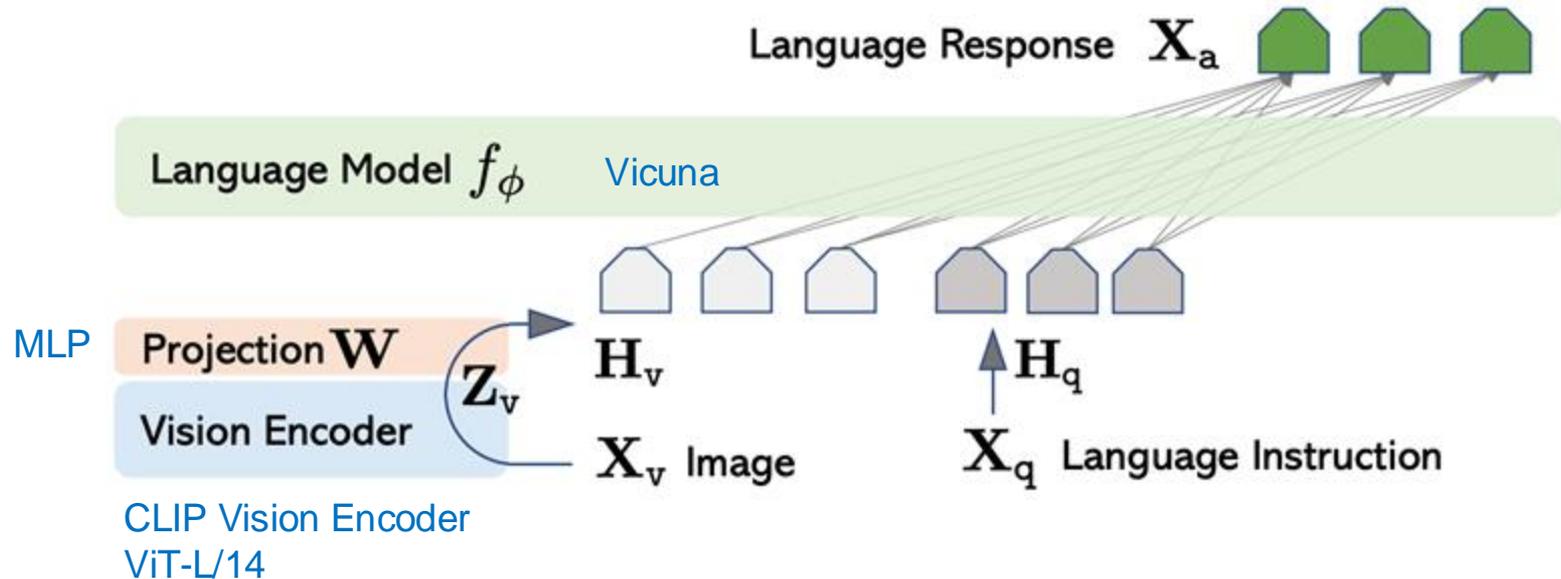
LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

- Two stage medical fine-tuning approach starting from LLaVA



Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

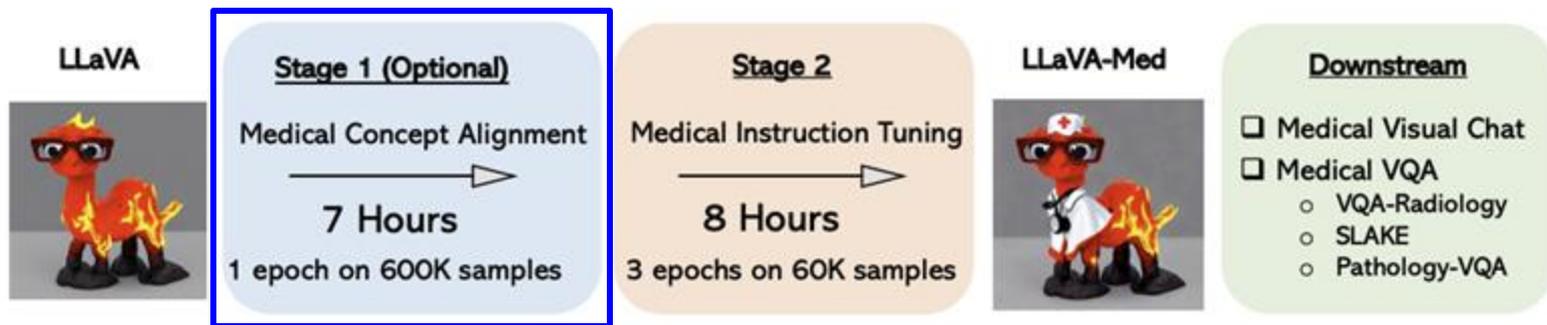
Remember: LLaVA



Stage 1: Pre-training for Feature Alignment
Stage 2: Fine-tuning End-to-End (W and ϕ)

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

- Two stage medical fine-tuning approach starting from LLaVA
- Stage 1: **Medical concept alignment** using the single-task instruction of image captioning. Trained on 600K image-text caption pairs sampled from PMC-15M.



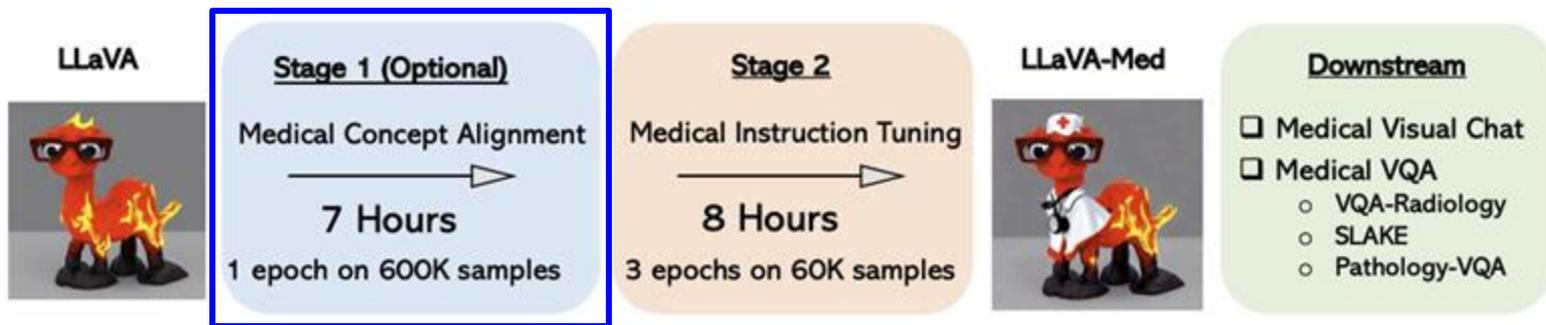
Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

- Two stage medical fine-tuning approach starting from LLaVA
- Stage 1: **Medical concept alignment** using the single-task instruction of image captioning. Trained on 600K image-text caption pairs sampled from PMC-15M.

Biomedical Concept Alignment Data. For a biomedical image X_v and its associated caption X_c , we sample a question X_q , which asks to describe the biomedical image. With (X_v, X_c, X_q) , we create a single-round instruction-following example:

Human : $X_q X_v$ <STOP>\n Assistant : X_c <STOP>\n



Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

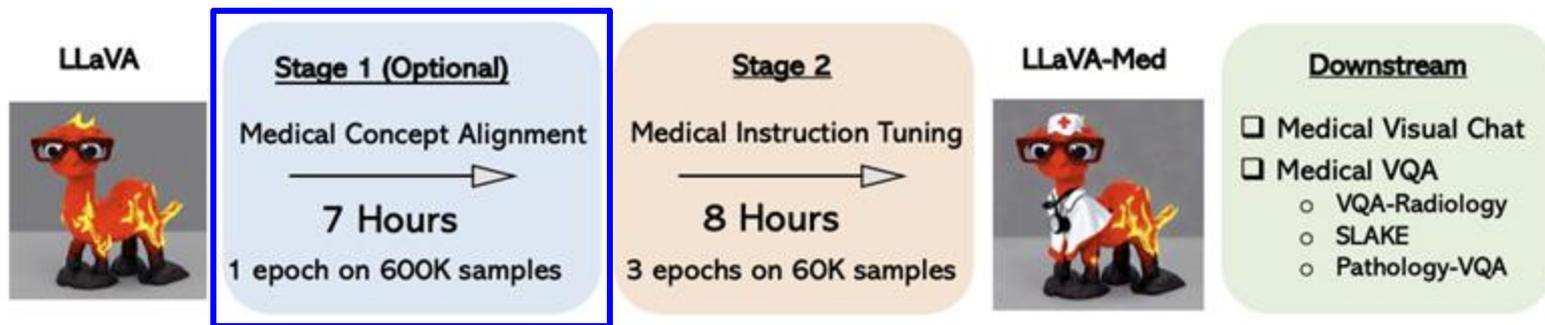
LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

In stage 1, update only projection matrix, freeze visual encoder and LM weights

- Two stage medical fine-tuning approach starting from LLaVA
- Stage 1: **Medical concept alignment** using the single-task instruction of image captioning. Trained on 600K image-text caption pairs sampled from PMC-15M.

Biomedical Concept Alignment Data. For a biomedical image X_v and its associated caption X_c , we sample a question X_q , which asks to describe the biomedical image. With (X_v, X_c, X_q) , we create a single-round instruction-following example:

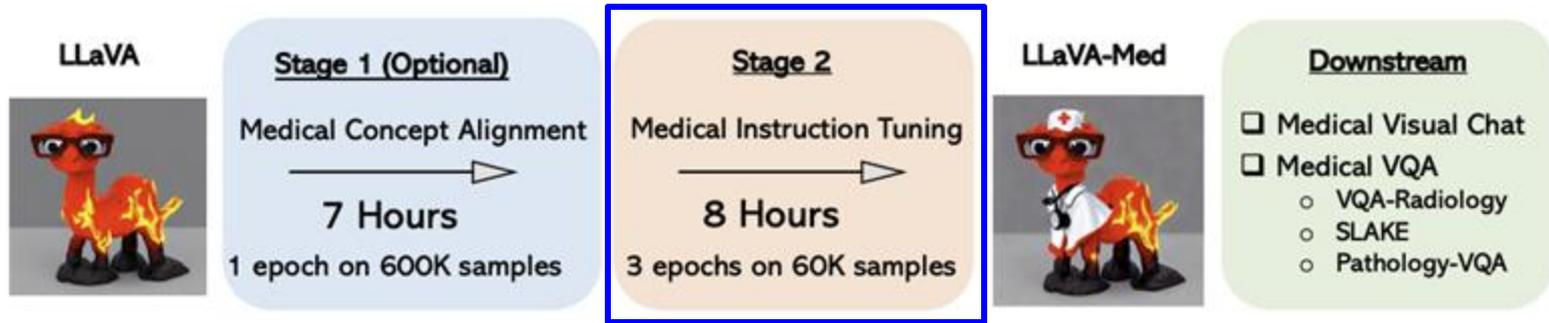
Human : $X_q X_v$ <STOP>\n Assistant : X_c <STOP>\n



Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

- Two stage medical fine-tuning approach starting from LLaVA
- Stage 2: **Medical instruction tuning** by prompting language-only GPT-4 to generate multi-round questions and answers given an image caption, in a tone as if it could see the image (even though it only has access to the text).

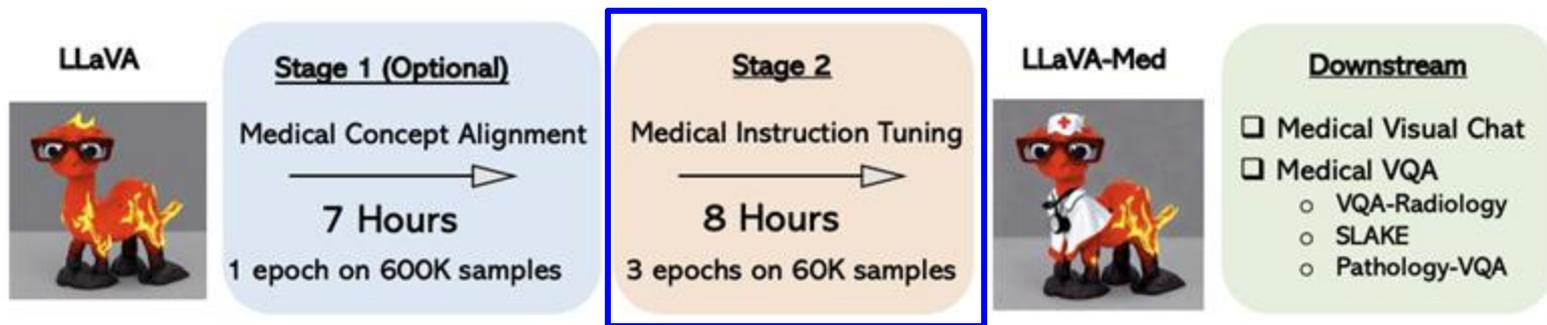


Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

In stage 2, update projection matrix and LM weights, but still keep visual encoder frozen

- Two stage medical fine-tuning approach starting from LLaVA
- Stage 2: **Medical instruction tuning** by prompting language-only GPT-4 to generate multi-round questions and answers given an image caption, in a tone as if it could see the image (even though it only has access to the text).



Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

Prompting GPT-4 to generate instruction tuning data (Stage 2)

Prompting GPT-4 to generate medical visual instruction-following data

```
messages = [{"role": "system", "content": ""You are an AI assistant specialized in biomedical topics.
```

You are provided with a text description (Figure Caption) of a figure image from a biomedical research paper. In some cases, you may have additional text (Figure Context) that mentions the image. Unfortunately, you don't have access to the actual image.

Below are requirements for generating the questions and answers in the conversation:

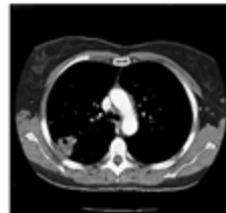
- Avoid quoting or referring to specific facts, terms, abbreviations, dates, numbers, or names, as these may reveal the conversation is based on the text information, rather than the image itself. Focus on the visual aspects of the image that can be inferred without the text information.
- Do not use phrases like "mentioned", "caption", "context" in the conversation. Instead, refer to the information as being "in the image."
- Ensure that questions are diverse and cover a range of visual aspects of the image.
- The conversation should include at least 2-3 turns of questions and answers about the visual aspects of the image.
- Answer responsibly, avoiding overconfidence, and do not provide medical advice or diagnostic information. Encourage the user to consult a healthcare professional for advice."

```
]  
for sample in fewshot_samples:  
    messages.append({"role": "user", "content": sample['context']})  
    messages.append({"role": "assistant", "content": sample['response']})  
messages.append({"role": "user", "content": query})
```

Biomedical Visual Instruction-Following Example

Caption:

Figure 4: Contrast-enhanced CT scan of the chest for patient #1. A large cavitating lesion was identified in the right upper lobe with pleural involvement and likely extension into the chest wall. In addition, right hilar lymphadenopathy was suspected. This was the primary malignant lesion in this patient.



Sentences with in-line figure mentions:

- Computerized tomography (CT) scans of the chest (Figure 4) were obtained.
- (Figure 4) and demonstrated a large cavitating lesion in the posterior aspect of the right upper lobe with probable pleural involvement likely representing the primary lesion.

Conversation:

User: Where is the lesion located?

Assistant: The lesion is located in the posterior aspect of the right upper lobe of the lung.

User: Is it extending into the chest wall?

Assistant: Yes, the CT scan suggests that the large cavitating lesion in the right upper lobe is likely extending into the chest wall.

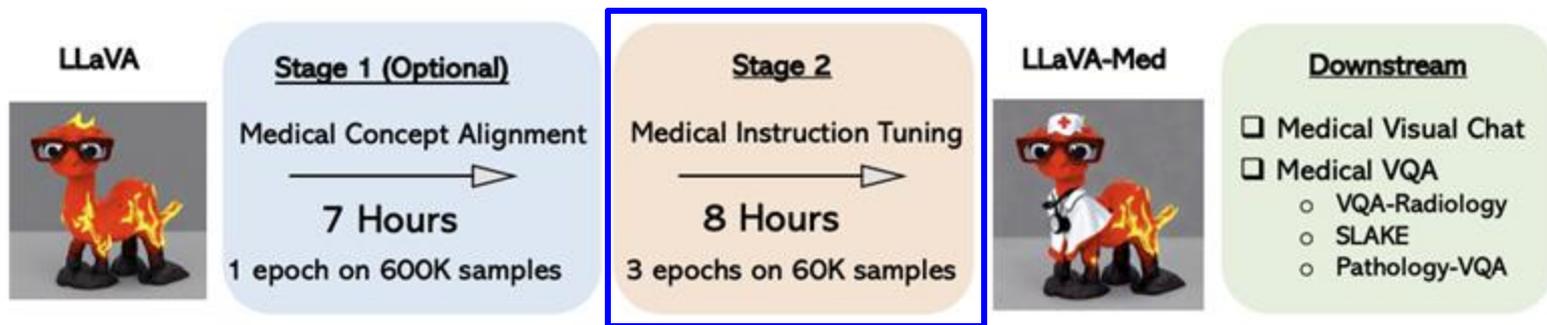
User: Can you explain more about the cavitating lesion?

Assistant: Certainly. A cavitating lesion refers to an abnormal mass or growth within the lung tissue that has a cavity or hollow space inside it. This cavity can be formed due to the necrosis (death) of the tissue inside the lesion, leading to the breakdown and disintegration of the affected tissue. Cavitating lesions can be caused by various factors, such as infections (like tuberculosis or fungal infections), autoimmune diseases, or malignancies (cancer). In the context of the CT scan provided, the large cavitating lesion in the right upper lobe of the lung is suggestive of a malignant (cancerous) growth. The cavity within the lesion may be a result of the rapid and uncontrolled growth of cancer cells, which can lead to the destruction of surrounding healthy tissue. It's important to note that further diagnostic tests, such as a biopsy, would be required to confirm the malignancy and determine the specific type of cancer involved.

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

In stage 2, update projection matrix and LM weights, but still keep visual encoder frozen

- Two stage medical fine-tuning approach starting from LLaVA
- Stage 2: **Medical instruction tuning** by prompting language-only GPT-4 to generate multi-round questions and answers given an image caption, in a tone as if it could see the image (even though it only has access to the text).



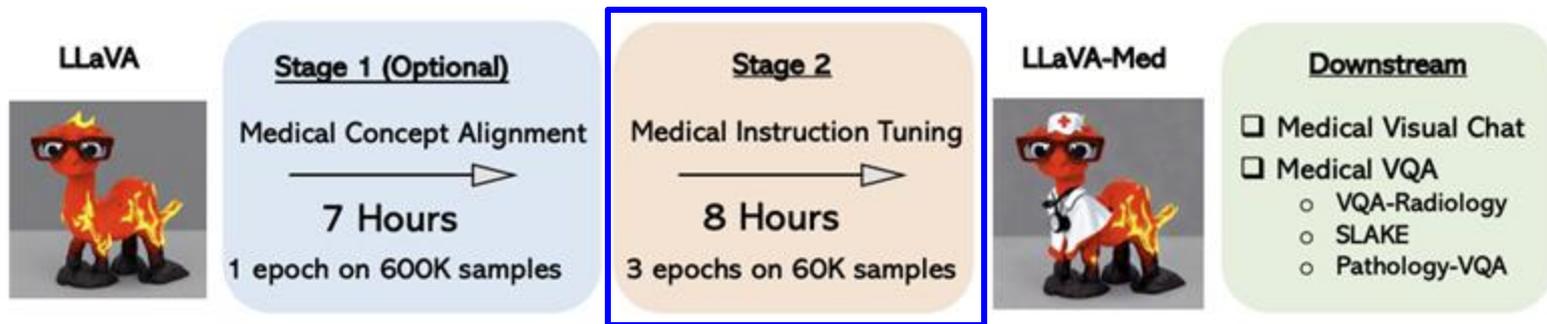
Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

In stage 2, update projection matrix and LM weights, but still keep visual encoder frozen

- Two stage medical fine-tuning approach starting from LLaVA
- Stage 2: **Medical instruction tuning** by prompting language-only GPT-4 to generate multi-round questions and answers given an image caption, in a tone as if it could see the image (even though it only has access to the text).

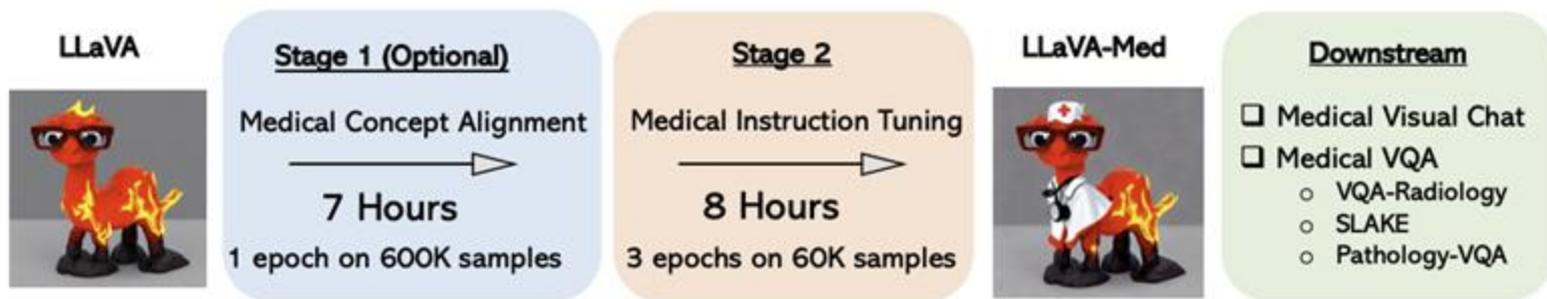
After the two-stage training, additionally fine-tune on three biomedical VQA datasets to better target to some specific biomedical scenarios. (Both Stage 1 and Stage 2 used the broad PMC-15M data.)



Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day

- Two stage medical fine-tuning approach starting from LLaVA
- Stage 2: **Medical instruction tuning** by prompting language-only GPT-4 to generate multi-round questions and answers given an image caption, in a tone as if it could see the image (even though it only has access to the text).



Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

Evaluation on downstream medical VQA datasets

Method	VQA-RAD			SLAKE			PathVQA		
	Ref	Open	Closed	Ref	Open	Closed	Ref	Open	Closed
<i>Supervised finet-tuning results with our own experiment runs</i>									
LLaVA		50.00	65.07		78.18	63.22		7.74	63.20
LLaVA-Med (From LLaVA)		61.52	84.19		83.08	85.34		37.95	91.21
LLaVA-Med (From Vicuna)		64.39	81.98		84.71	83.17		38.87	91.65
LLaVA-Med (BioMed CLIP)		64.75	83.09		87.11	86.78		39.60	91.09
<i>Representative & SoTA methods with numbers reported in the literature</i>									
VL Encoder-Decoder [5]	71.49		82.47				71.49		85.61
Q2ATransformer [30]	79.19		81.20				54.85		88.85
Prefix T. Medical LM [44]				84.30		82.01	40.00		87.00
PubMedCLIP [9]	60.10		80.00	78.40		82.50			
BiomedCLIP [51]	67.60		79.80	82.05		89.70			
M2I2 [24]	66.50		83.50	74.70		91.10	36.30		88.00

Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

Qualitative example of LLaVA-Med used for medical visual chat

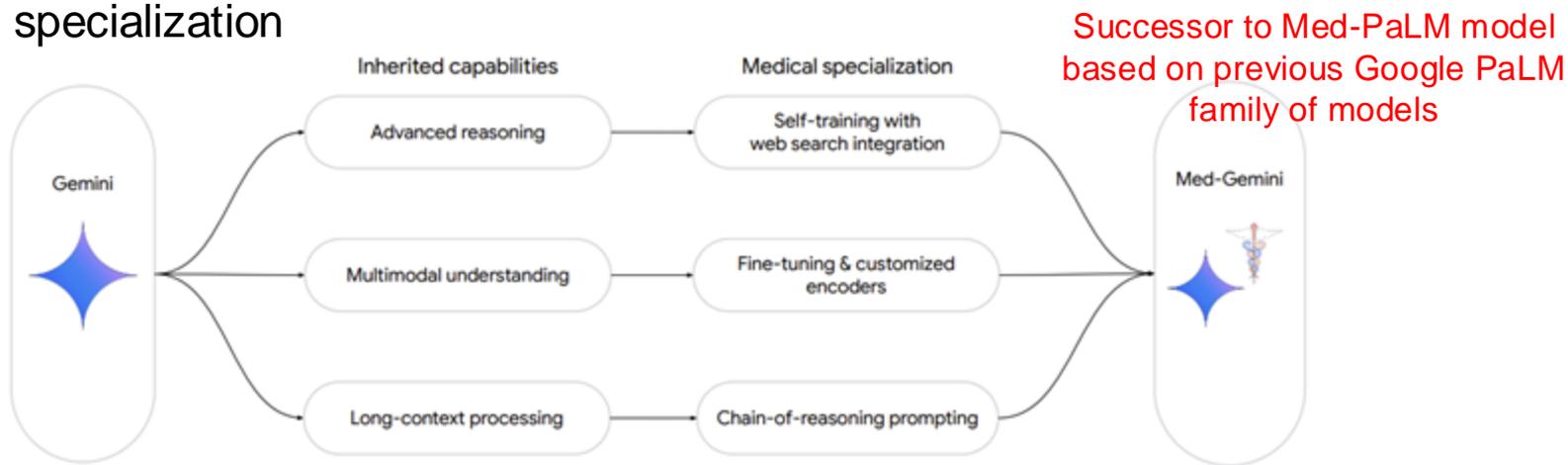
Visual input example, Biomedical image from [6]

		<p>Caption: Chest x-ray (CXR) on Day 2 of admission post-intubation (yellow line showing the level of the endotracheal tube). Rapidly worsening ground-glass opacities are seen throughout the lungs with relative subpleural sparing (red arrows)</p>
User	<p>Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7386722/</p> <p>What is shown in this image?</p>	
LLaVA-Med	<p>The image is a chest X-ray (CXR) that shows bilateral patchy infiltrates, which are areas of increased opacity in the lungs. These infiltrates can be indicative of various lung conditions, such as infections, inflammation, or other lung diseases.</p>	
User	<p>Are there any devices implanted in this patient?</p>	
LLaVA-Med	<p>Yes, it appears that the patient has several devices implanted. These devices may include endotracheal tubes, central venous catheters, and/or pulmonary artery catheters. These devices are typically used for monitoring and managing various medical conditions, as well as for providing support during surgical procedures.</p>	
User	<p>What is shown in this image?</p>	
LLaVA	<p>The image shows a ribcage with wires coming out of it, indicating that it is likely a medical illustration or diagram.</p>	
User	<p>Are there any devices implanted in this patient?</p>	
LLaVA	<p>Yes, there are several devices protruding from the patient's chest, which are likely medical devices used for treatment or monitoring.</p>	
User	<p>What is shown in this image?</p>	
Language GPT-4	<p>The image shows a chest x-ray taken on Day 2 of a patient's admission after intubation. You can see the endotracheal tube marked by a yellow line, and rapidly worsening ground-glass opacities throughout the lungs. There is also relative subpleural sparing, which is indicated by red arrows.</p>	
User	<p>Are there any devices implanted in this patient?</p>	
Language GPT-4	<p>Yes, there is an endotracheal tube implanted in the patient, as indicated by the yellow line in the image.</p>	

Li et al. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS Datasets and Benchmarks 2023.

Med-Gemini: state-of-the-art generalist biomedical VLM

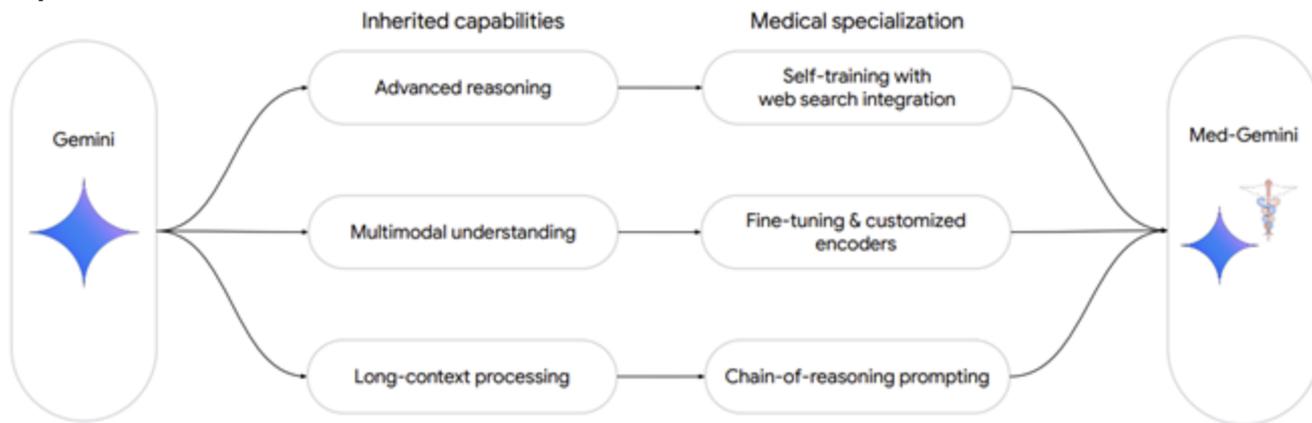
- Also a generalist (broad domains) biomedical VLM like LLaVA-Med, but extends from the much more powerful Gemini models (and built by internal Google DeepMind team)
- SoTA (state-of-the-art) due to Gemini foundation and additional techniques for medical specialization



Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Med-Gemini: state-of-the-art generalist biomedical VLM

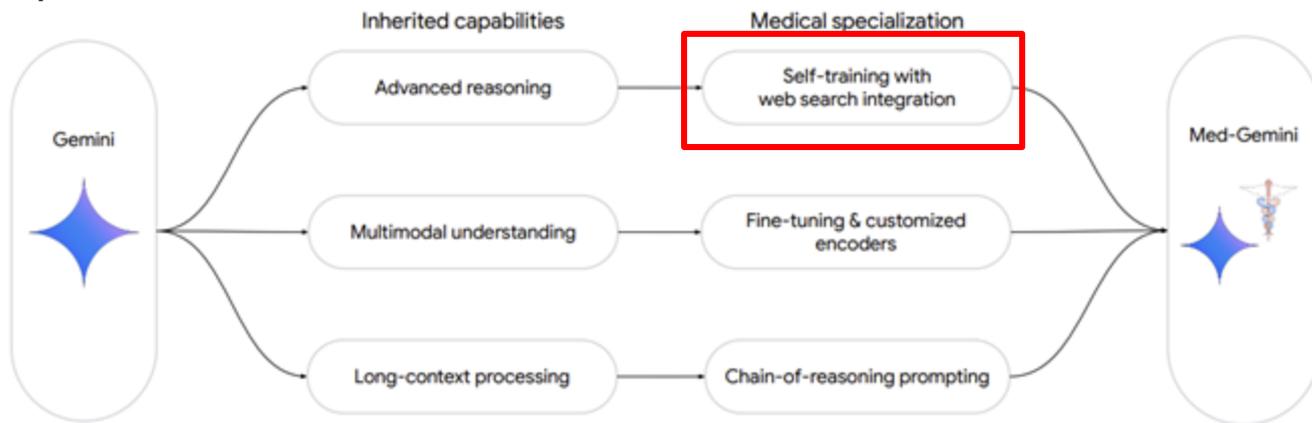
- Also a generalist (broad domains) biomedical VLM like LLaVA-Med, but extends from the much more powerful Gemini models (and built by internal Google DeepMind team)
- SoTA (state-of-the-art) due to Gemini foundation and additional techniques for medical specialization



Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Med-Gemini: state-of-the-art generalist biomedical VLM

- Also a generalist (broad domains) biomedical VLM like LLaVA-Med, but extends from the much more powerful Gemini models (and built by internal Google DeepMind team)
- SoTA (state-of-the-art) due to Gemini foundation and additional techniques for medical specialization



Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

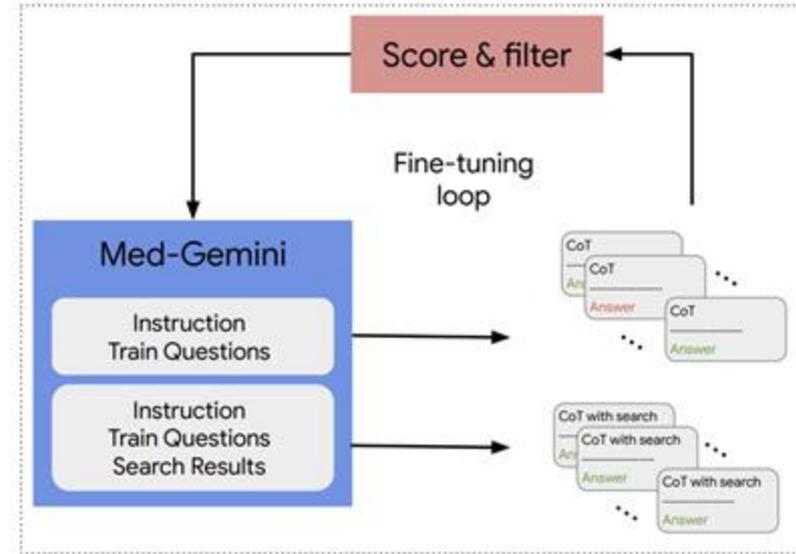
Self-training with web search integration

Med-Gemini is instruction-tuned in an iterative “fine-tuning loop” on MedQA medical exam questions (USMLE) using instructions that are generated leveraging intermediate Med-Gemini models

Fine-tuning loop:

- (1) Prompt Med-Gemini to generate chain-of-thought (CoT) i.e. step-by-step reasoning explanations to answer questions in fine-tuning dataset
- (2) Additionally prompt Med-Gemini to generate relevant web search queries, whose results can be included in the input context for repeating (1)
- (3) Score for correctness, and use filtered subset of CoTs to fine-tune Med-Gemini. The improved model can then be used to generate new, better CoTs.

Self-training with search



Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Med-Gemini: SoTA generalist biomedical VLM

Input
Instruction
You are a medical expert answering a multiple choice question about medical knowledge. To help you answer the question, you are given access to search results.

Question
A 20-year-old woman presents with menorrhagia for the past several years. She says that her menses “have always been heavy”, and she has experienced easy bruising for as long as she can remember. Family history is significant for her mother, who had similar problems with bruising easily. The patient’s vital signs include: heart rate 98/min, respiratory rate 14/min, temperature 36.1°C (96.9°F), and blood pressure 110/87 mm Hg. Physical examination is unremarkable. Laboratory tests show the following: platelet count 200,000/mm³, PT 12 seconds, and PTT 43 seconds. Which of the following is the most likely cause of this patient’s symptoms? (A) Hemophilia A (B) Lupus anticoagulant (C) Protein C deficiency (D) Von Willebrand disease

Search Results
Partial thromboplastin time (PTT) and activated partial thromboplastin time (aPTT) are used to test for the same functions; however, in aPTT, an activator is added that speeds up the clotting time and results in a narrower reference range. The aPTT is considered a more sensitive version of the PTT and is used to monitor the patient’s response to heparin therapy. The reference range of the aPTT is 30-40 seconds. The reference range of the PTT is 60-70 seconds. In patients receiving anticoagulant therapy, the reference range is 1.5-2.5 times the control value in seconds. Critical values that should prompt a clinical alert are as follows: aPTT: More than 70 seconds (signifies spontaneous bleeding). PTT: More than 100 seconds (signifies spontaneous bleeding). eMedicine Logo. Next: Interpretation. A prolonged aPTT result may indicate the following : Congenital deficiencies of intrinsic system clotting factors such as factors VIII, IX, XI, and XII, including hemophilia A and hemophilia B (Christmas disease), two inherited bleeding disorders resulting from a deficiency in factors VIII and IX, respectively. Congenital deficiency of Fitzgerald factor (prekallikrein). Von Willebrand disease, which is the most common inherited bleeding disorder, affecting platelet function owing to decreased von Willebrand factor activity. Hypofibrinogenemia · Liver cirrhosis (the liver makes most of the clotting factors, including those that are vitamin K-dependent ones); diseases of the liver may result in an inadequate quantity of clotting factors, prolonging the aPTT. Vitamin K deficiency: The synthesis of some clotting factors requires vitamin K, so vitamin K deficiency results in an inadequate quantity of intrinsic system and common pathways clotting factors, as a result the aPTT is prolonged. Disseminated intravascular coagulation (DIC): The clotting factors involved in the intrinsic pathway are consumed, prolonging the aPTT.
Source: <https://emedicine.medscape.com/article/2085837-overview>

Example of web search results relevant to a question



Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Med-Gemini: SoTA generalist biomedical VLM

Input

Instruction

You are a medical expert answering a multiple choice question about medical knowledge. To help you answer the question, you are given access to search results.

Question

A 20-year-old woman presents with menorrhagia for the past several years. She says that her menses “have always been heavy”, and she has experienced easy bruising for as long as she can remember. Family history is significant for her mother, who had similar problems with bruising easily. The patient’s vital signs include: heart rate 98/min, respiratory rate 14/min, temperature 36.1°C (96.9°F), and blood pressure 110/87 mm Hg. Physical examination is unremarkable. Laboratory tests show the following: platelet count 200,000/mm³, PT 12 seconds. Which of the following is the most likely cause of this patient’s symptoms? (A) Hemophilia A (B) Hemophilia B (C) Protein C deficiency (D) Von Willebrand disease

Search Results

Partial thromboplastin time (PTT) and activated partial thromboplastin time (aPTT) are used to evaluate the intrinsic and common pathways of the coagulation cascade. PTT is used to monitor heparin therapy, while aPTT is used to monitor heparin therapy in patients with renal impairment. In patients receiving anticoagulant therapy, the reference range is 1.5-2.5 times the normal range. Critical values that should prompt a clinical alert are as follows: aPTT: More than 70 seconds (spontaneous bleeding). PTT: More than 100 seconds (signifies spontaneous bleeding). eMedicine Interpretation. A prolonged aPTT result may indicate the following : Congenital deficiencies of clotting factors such as factors VIII, IX, XI, and XII, including hemophilia A and hemophilia B (Christmas disease), two inherited bleeding disorders resulting from a deficiency in factors VIII and IX, respectively. Congenital deficiency of Fitzgerald factor (prekallikrein). Von Willebrand disease, which is the most common inherited bleeding disorder, affecting platelet function owing to decreased von Willebrand factor activity. Hypofibrinogenemia · Liver cirrhosis (the liver makes most of the clotting factors, including those that are vitamin K-dependent ones); diseases of the liver may result in an inadequate quantity of clotting factors, prolonging the aPTT. Vitamin K deficiency: The synthesis of some clotting factors requires vitamin K, so vitamin K deficiency results in an inadequate quantity of intrinsic system and common pathways clotting factors, as a result the aPTT is prolonged. Disseminated intravascular coagulation (DIC): The clotting factors involved in the intrinsic pathway are consumed, prolonging the aPTT.

Source: <https://emedicine.medscape.com/article/2085837-overview>

Example of chain-of-thought leading to an answer

Target

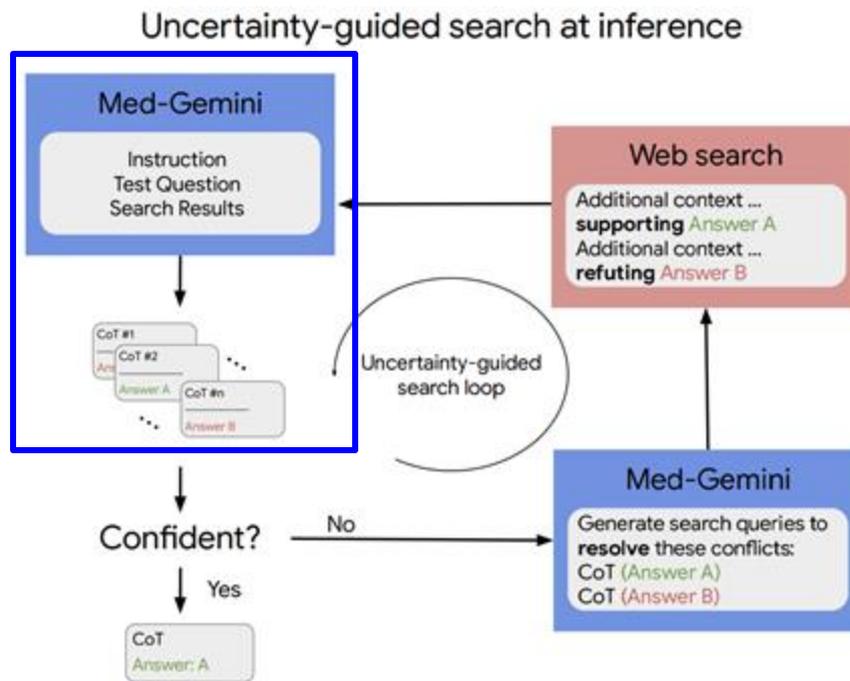
The patient’s symptoms of menorrhagia and easy bruising, along with a prolonged PTT, suggest a bleeding disorder. Hemophilia A (option A) is an X-linked recessive disorder that typically presents with a prolonged PTT and normal PT. However, it is unlikely in this case as the patient is female and has a family history of her mother having similar symptoms. Lupus anticoagulant (option B) can cause a prolonged PTT, but it is typically associated with thrombosis rather than bleeding. Protein C deficiency (option C) can cause thrombosis, but it does not affect PTT. According to the search results, “Von Willebrand disease (vWD) is the most common inherited bleeding disorder, affecting platelet function owing to decreased von Willebrand factor activity.” It can present with menorrhagia, easy bruising, and a prolonged PTT. The patient’s family history of her mother having similar symptoms also supports this diagnosis. Therefore, the most likely cause of this patient’s symptoms is (D) Von Willebrand disease. Answer: (D) Von Willebrand disease

Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Uncertainty-guided search at inference

Multi-step inference process loop:

1. Generate multiple CoT reasoning pathways to answer a question

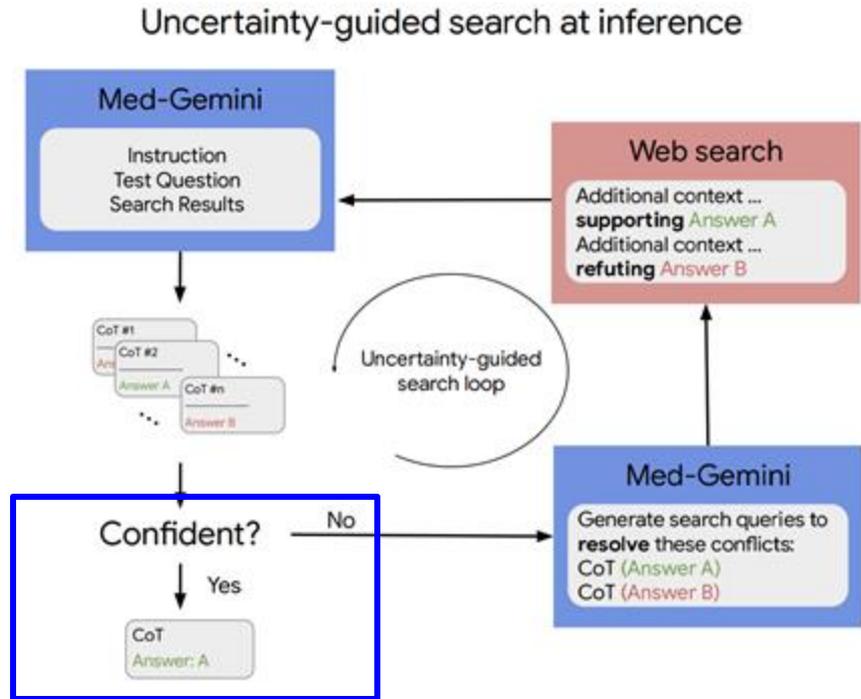


Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Uncertainty-guided search at inference

Multi-step inference process loop:

1. Generate multiple CoT reasoning pathways to answer a question
2. If these generally agree (using an entropy measure), the majority vote answer is final. If not, proceed to Steps 3 and 4.

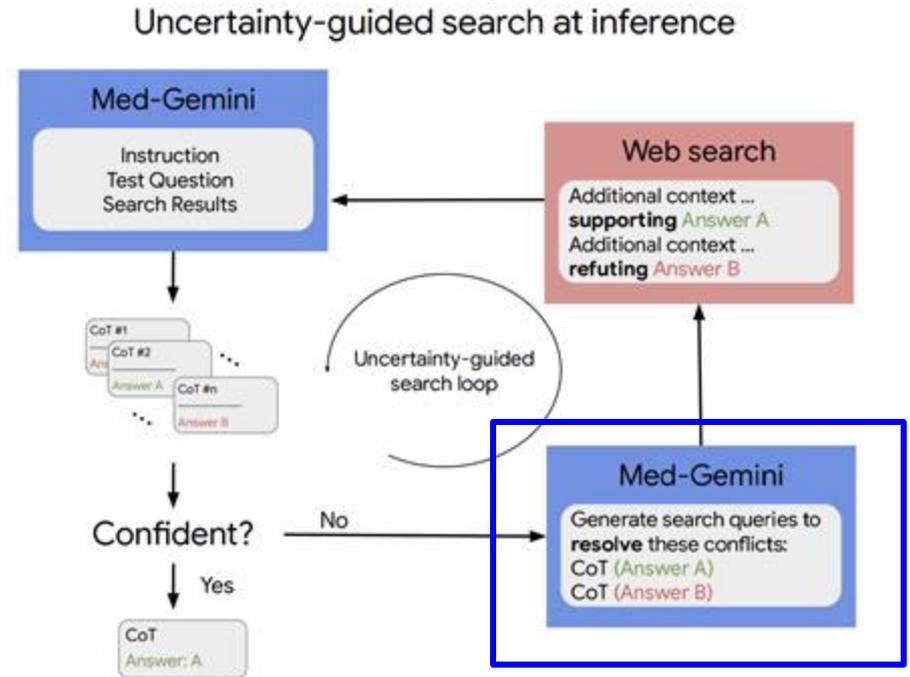


Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Uncertainty-guided search at inference

Multi-step inference process loop:

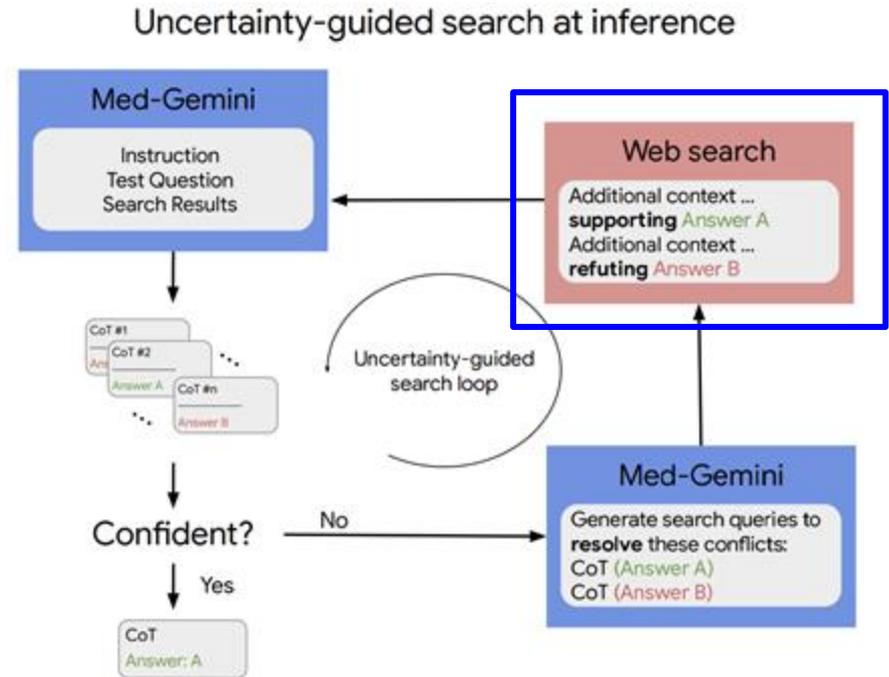
1. Generate multiple CoT reasoning pathways to answer a question
2. If these generally agree (using an entropy measure), the majority vote answer is final. If not, proceed to Steps 3 and 4.
3. Prompt Med-Gemini to generate 3 search queries whose results would help resolve the conflict produced by Step 1.



Uncertainty-guided search at inference

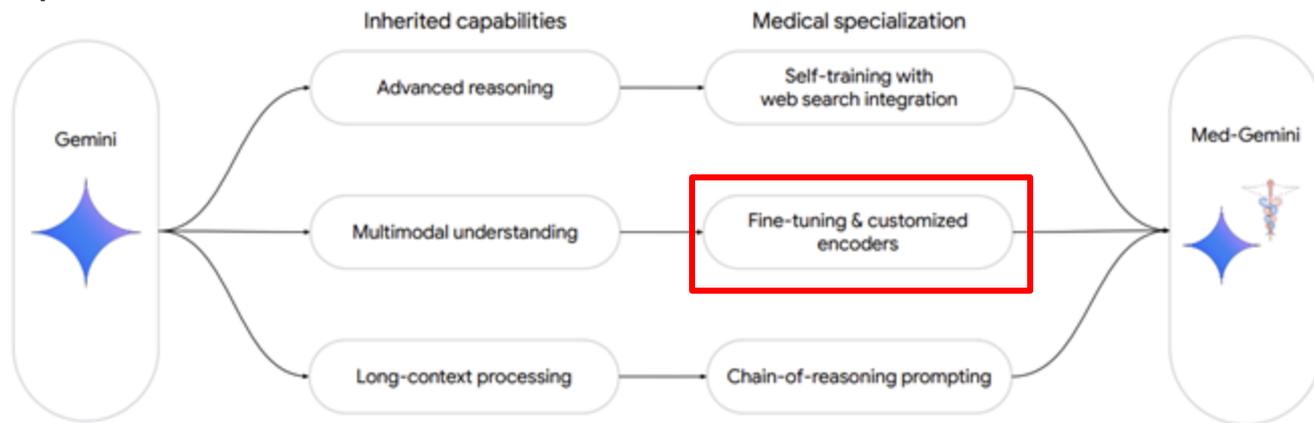
Multi-step inference process loop:

1. Generate multiple CoT reasoning pathways to answer a question
2. If these generally agree (using an entropy measure), the majority vote answer is final. If not, proceed to Steps 3 and 4.
3. Prompt Med-Gemini to generate 3 search queries whose results would help resolve the conflict produced by Step 1.
4. Web results are retrieved and incorporated into the next iteration of Step 1.



Med-Gemini: state-of-the-art generalist biomedical VLM

- Also a generalist (broad domains) biomedical VLM like LLaVA-Med, but extends from the much more powerful Gemini models (and built by internal Google DeepMind team)
- SoTA (state-of-the-art) due to Gemini foundation and additional techniques for medical specialization

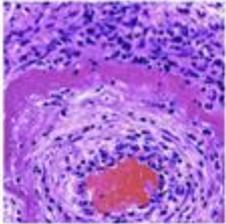


Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Further fine-tuning on a variety of benchmark datasets with custom instructions for each dataset (8 tasks across 6 datasets)

Open-ended Visual QA (Path-VQA)

Visual input



Instruction
You are a helpful medical assistant. The following are questions about medical knowledge. Solve them in a step-by-step fashion, referring to authoritative sources as needed.
Question: What does the wall of the artery show with protein deposition and inflammation?

Response
a circumferential bright pink area of necrosis

Image Classification (PAD-UFES-20 6-condition classification)

Visual input



Instruction
You are a helpful dermatology assistant. The following are questions about skin lesions. Categorize the skin lesions into the most likely class given the patient history. Output a single option letter from the provided options as the final answer.
Patient History: Age: 51, Gender: female, Smoke: false, Drink: false, Family skin cancer history: true, Family any cancer history: false, Lesion region: back, Lesion itch: false, Lesion grew: false, Lesion bled: false, Lesion elevation: false, Fitzpatrick scale: 1.0, Diameters (mm): [12.0, 8.0].
Question: Which of the following is the most likely diagnosis of the patient's skin lesion? (A) Nevus (B) Basal Cell Carcinoma (C) Squamous Cell Carcinoma (D) Actinic Keratosis (E) Seborrheic Keratosis (F) Melanoma.

Response
(A)

Open-ended Visual QA in Chinese (Slake-VQA)

Visual input

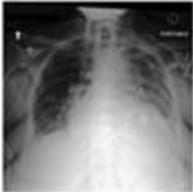


Instruction
You are a helpful medical assistant. The following are questions about medical knowledge. Solve them in a step-by-step fashion, referring to authoritative sources as needed.
Question: 图像里包含的区域属于身体哪个部分?

Response
腹部

Image Classification (MIMIC-CXR 13-condition classification)

Visual input



Instruction
You are a helpful radiology assistant. The following are questions about findings in chest X-ray in the frontal view. Identify if a specific type of abnormality is shown in the X-ray.
Given the <VIEW> X-ray image,
Question: Which of the following abnormalities are indicated by the image? (A) Atelectasis (B) Cardiomegaly (C) Consolidation (D) Edema (E) Enlarged Cardiomediastinum (F) Fracture (G) Lung Lesion (H) Lung Opacity (I) Pleural Effusion (J) Pleural Other (K) Pneumonia (L) Pneumothorax (M) Support Devices

Response
(A)

Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Further fine-tuning on a variety of benchmark datasets with custom instructions for each dataset (8 tasks across 6 datasets)

Close-ended Visual QA (NEJM Image Challenge, USMLE-MM)

Visual input



Instruction
You are a medical expert. Only output the final (diagnosis, answer). Do not output the reasoning or explanation. Output the final diagnosis in the format "Final (Diagnosis, Answer): X" where X is the most (possible medical diagnosis, correct letter choice).
Question: Infection with which one of the following organisms is the most likely cause of this rash? (A) Coxsackie virus type A16 (B) Echovirus type 16 (C) Group A streptococcus (D) Herpes simplex virus type 1 (E) Norwalk virus

Response
Final Answer: (A)

Image Classification (MIMIC-CXR normal vs abnormal classification)

Visual input



Instruction
You are a helpful radiology assistant. The following are questions about findings in chest X-ray in the frontal view. Identify if a specific type of abnormality is shown in the X-ray.
Given the <VIEW> X-ray image.
Question: are there any abnormalities indicated by the image? (A) Yes (B) No.

Response
(A)

Waveform Signal Visual QA (ECG-QA)

Raw sensor input*



Instruction
Given this ECG sequence, please answer the following question. From the provided options, select all that apply. List your selections alphabetically, separated by commas.
Question: What signs of a rhythm-related disorder can be found in this ECG recording?
Options: atrial fibrillation, atrial flutter, bigeminal pattern, normal functioning artificial pacemaker, sinus arrhythmia, sinus bradycardia, sinus rhythm, sinus tachycardia, supraventricular tachycardia

Response
atrial fibrillation, atrial flutter

Text Report Generation (MIMIC-CXR)

Visual input



Instruction
You are a helpful radiology assistant.
Question: Describe the findings and impression in the chest X-ray following the instructions.

Response
FINDINGS: Low lung volumes are present. This accentuates the size of the cardiac silhouette which is likely mildly enlarged. Mediastinal and hilar contours are likely within normal limits. A right brachiocephalic venous stent is re-demonstrated. There is crowding of the bronchovascular structures with probable mild pulmonary vascular congestion. No pleural effusion or pneumothorax is identified.
IMPRESSION: Low lung volumes with mild pulmonary vascular congestion.

Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Further fine-tuning on a variety of benchmark datasets with custom instructions for each dataset (8 tasks across 6 datasets)

Close-ended Visual QA (NEJM Image Challenge, USMLE-MM)

Visual input



Instruction
You are a medical expert. Only output the final (diagnosis, answer). Do not output the reasoning or explanation. Output the final diagnosis in the format "Final (Diagnosis, Answer): X" where X is the most (possible medical diagnosis, correct letter choice).
Question: Infection with which one of the following organisms is the most likely cause of this rash? (A) Coxsackie virus type A16 (B) Echovirus type 16 (C) Group A streptococcus (D) Herpes simplex virus type 1 (E) Norwalk virus

Response
Final Answer: (A)

Image Classification (MIMIC-CXR normal vs abnormal classification)

Visual input



Instruction
You are a helpful radiology assistant. The following are questions about findings in chest X-ray in the frontal view. Identify if a specific type of abnormality is shown in the X-ray.
Given the <VIEW> X-ray image.
Question: are there any abnormalities indicated by the image? (A) Yes (B) No.

Response
(A)

New custom encoder for waveforms



Waveform Signal Visual QA (ECG-QA)

Raw sensor input*



Instruction
Given this ECG sequence, please answer the following question. From the provided options, select all that apply. List your selections alphabetically, separated by commas.
Question: What signs of a rhythm-related disorder can be found in this ECG recording?
Options: atrial fibrillation, atrial flutter, bigeminal pattern, normal functioning artificial pacemaker, sinus arrhythmia, sinus bradycardia, sinus rhythm, sinus tachycardia, supraventricular tachycardia

Response
atrial fibrillation, atrial flutter

Text Report Generation (MIMIC-CXR)

Visual input



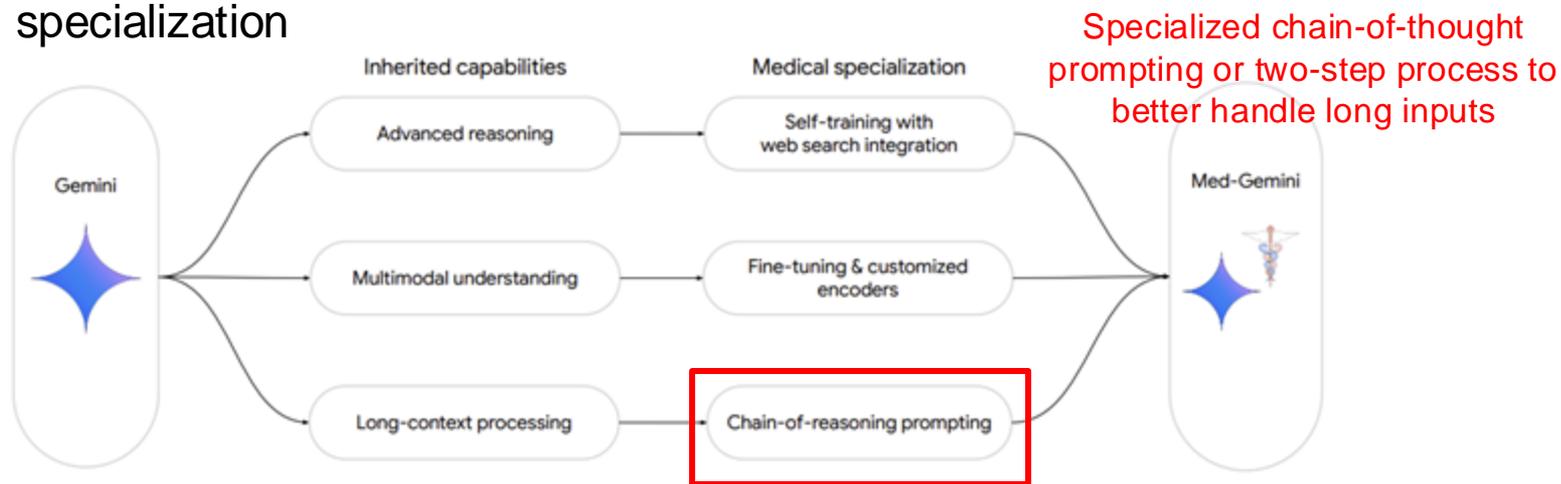
Instruction
You are a helpful radiology assistant.
Question: Describe the findings and impression in the chest X-ray following the instructions.

Response
FINDINGS: Low lung volumes are present. This accentuates the size of the cardiac silhouette which is likely mildly enlarged. Mediastinal and hilar contours are likely within normal limits. A right brachiocephalic venous stent is re-demonstrated. There is crowding of the bronchovascular structures with probable mild pulmonary vascular congestion. No pleural effusion or pneumothorax is identified.
IMPRESSION: Low lung volumes with mild pulmonary vascular congestion.

Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Med-Gemini: state-of-the-art generalist biomedical VLM

- Also a generalist (broad domains) biomedical VLM like LLaVA-Med, but extends from the much more powerful Gemini models (and built by internal Google DeepMind team)
- SoTA (state-of-the-art) due to Gemini foundation and additional techniques for medical specialization



Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Example of recognizing events in a long surgical video

Prompt

HEAD AND NECK SURGERY DEPARTMENT

- Male, 36
- Right enlarged level III tender neck mass
- 3 years of slow growth, painless
- No fever or weight loss
- Ultrasound suggests enlarged abnormal lymph node, 3.5 x 2.5cm
- FNA negative for malignancy

...

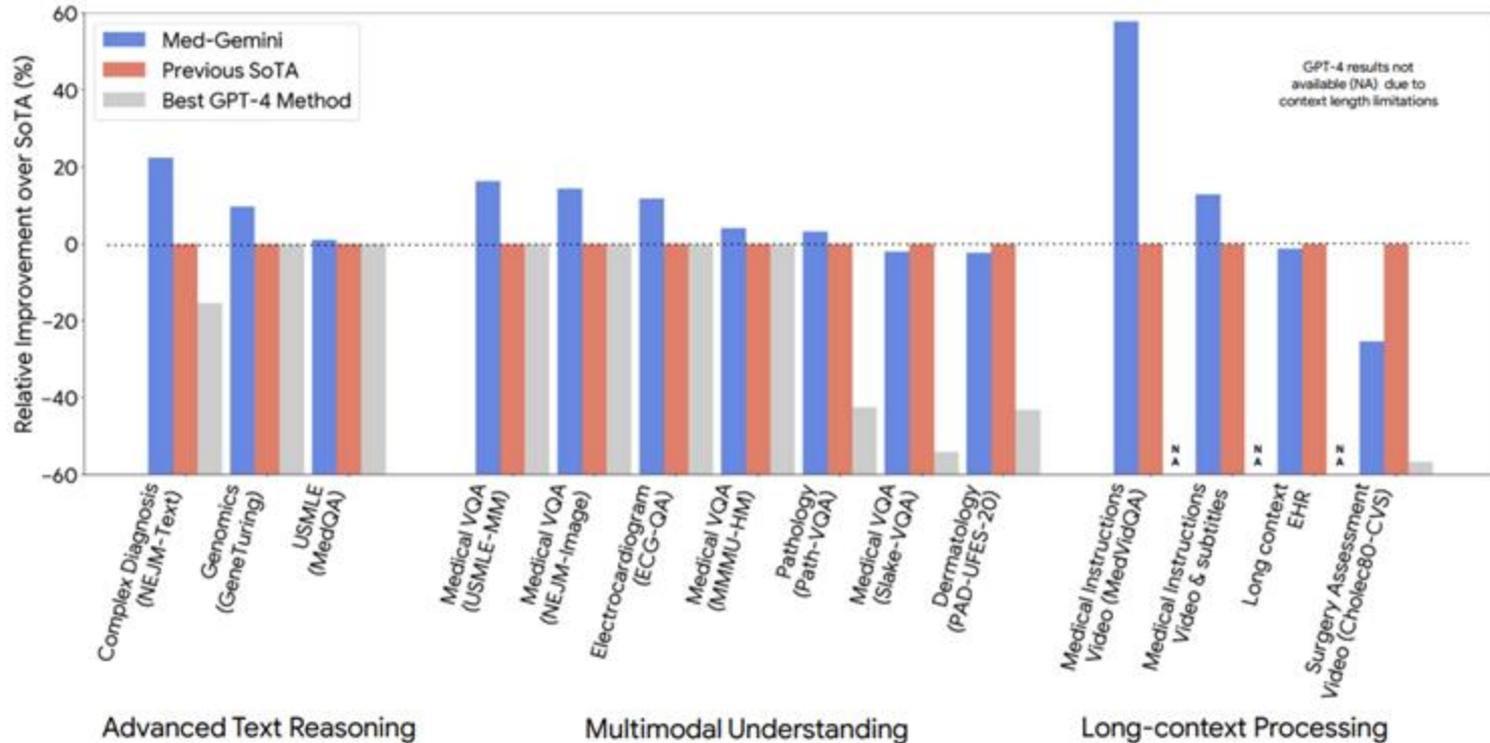


...



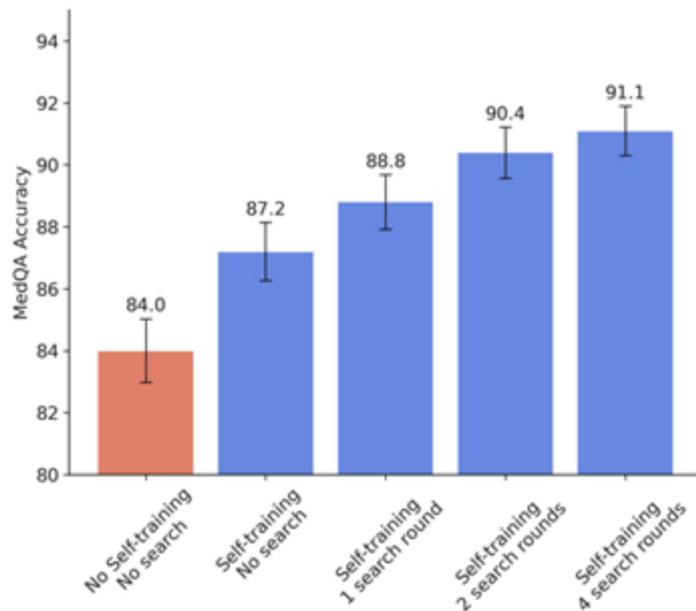
Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Med-Gemini benchmarking



Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

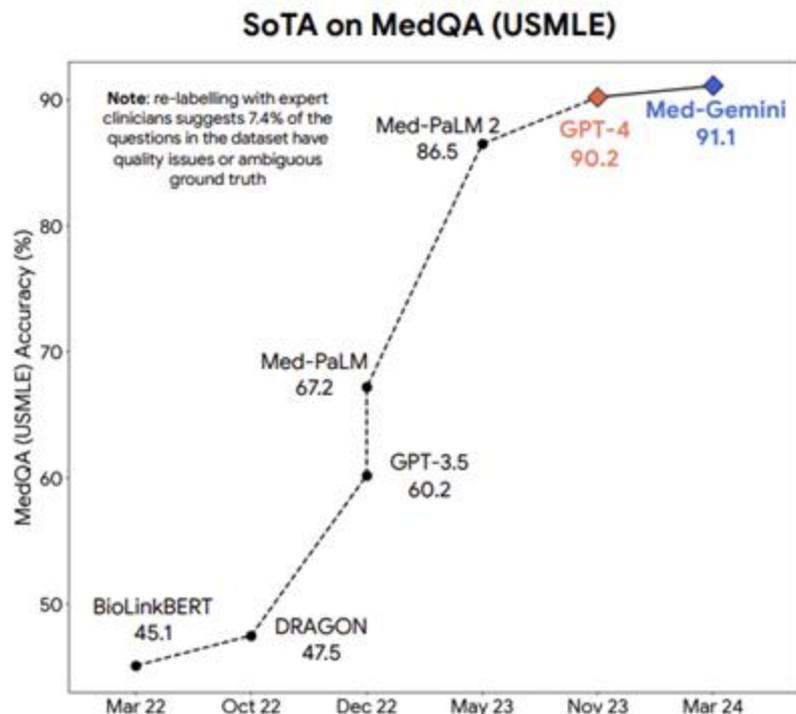
Effect of web search in self-training loop



(a) MedQA Self-training & search ablation

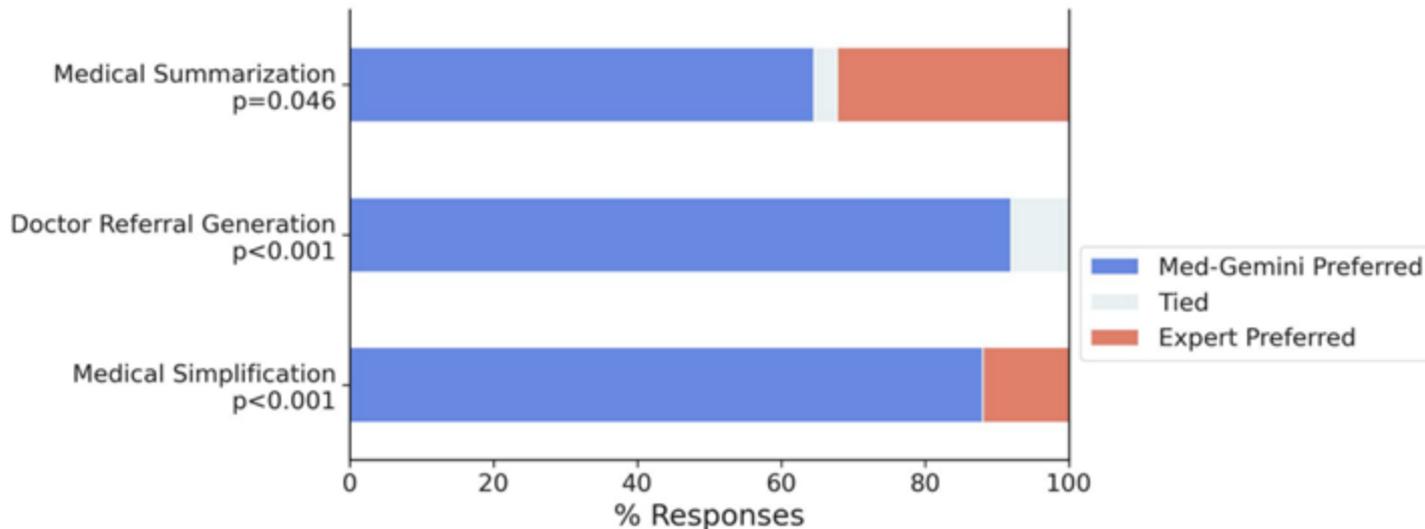
Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Some interesting evaluation results (even though these are text-only :))



Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Some interesting evaluation results (even though these are text-only :))



Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Some interesting evaluation results (even though these are text-only :))

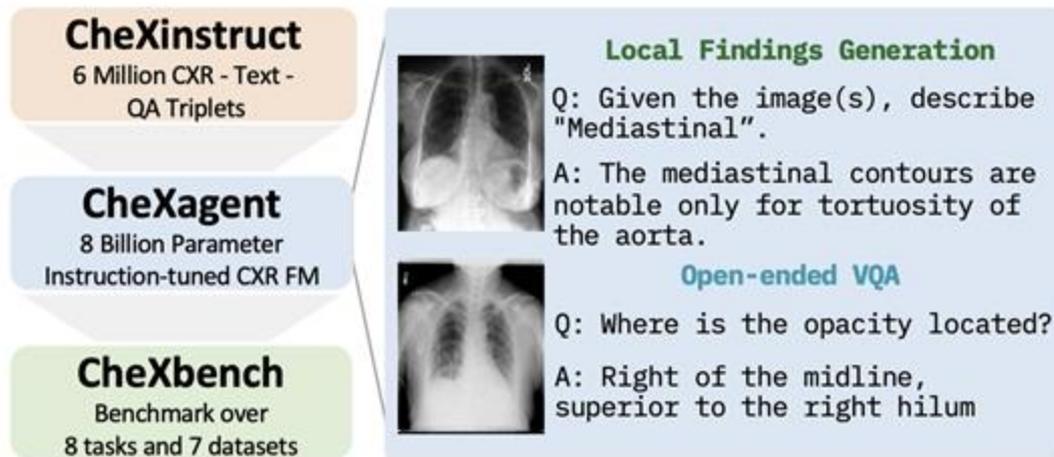
Task	Num. Examples	Evaluation-Axis	Expert Preferred	Tied	Med-Gemini	p-value
After-Visit Summary	31	Accuracy	19%	55%	26%	p < 0.001
		Coverage	48%	16%	35%	p = 0.861
		Succinctness	29%	10%	61%	p = 0.017
		Coherence	29%	13%	58%	p = 0.017
		Overall	32%	3%	65%	p = 0.046
Referral Letter	25	Overall	0%	8%	92%	p < 0.001
Cochrane Simplification	25	Accuracy	52%	8%	40%	p = 0.846
		Coverage	12%	12%	76%	p < 0.001
		Succinctness	4%	4%	92%	p < 0.001
		Reading Level	0%	0%	100%	p < 0.001
		Overall	12%	0%	88%	p < 0.001

Saab et al. Capabilities of Gemini Models in Medicine. arXiv 2024.

Specialist (narrower domain) biomedical VLMs

CheXagent: VLM for CXR interpretation

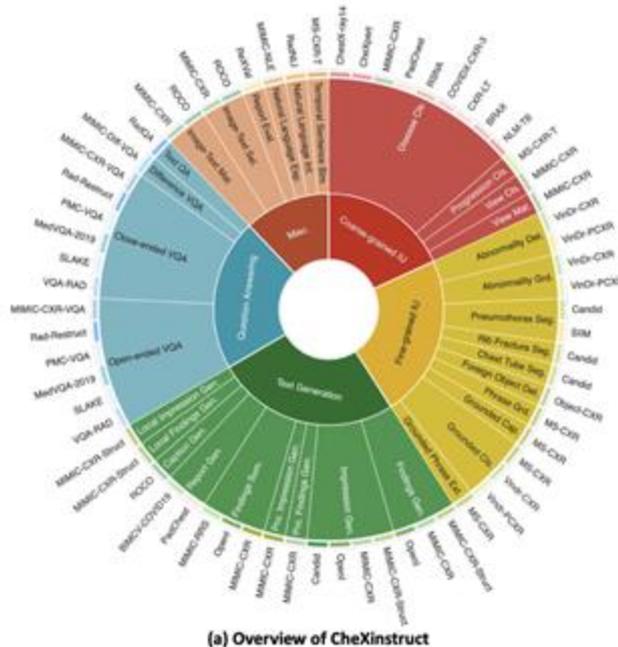
Contributed an instruction tuning dataset, instruction-tuned model, and evaluation benchmark for CXR interpretation tasks spanning coarse-grained image understanding, fine-grained image understanding, question answering, text (report) generation, and other miscellaneous categories.



Chen et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024.

CheXagent: VLM for CXR interpretation

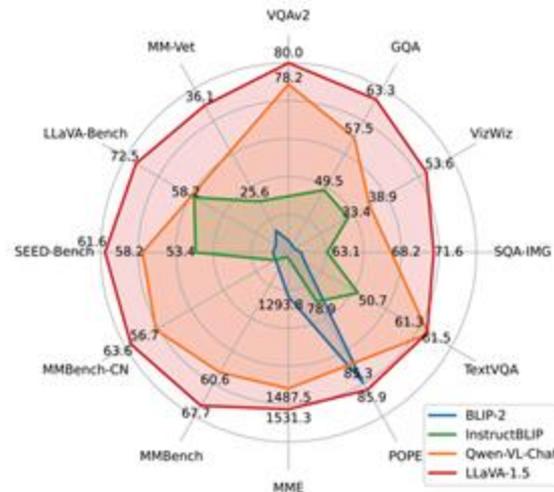
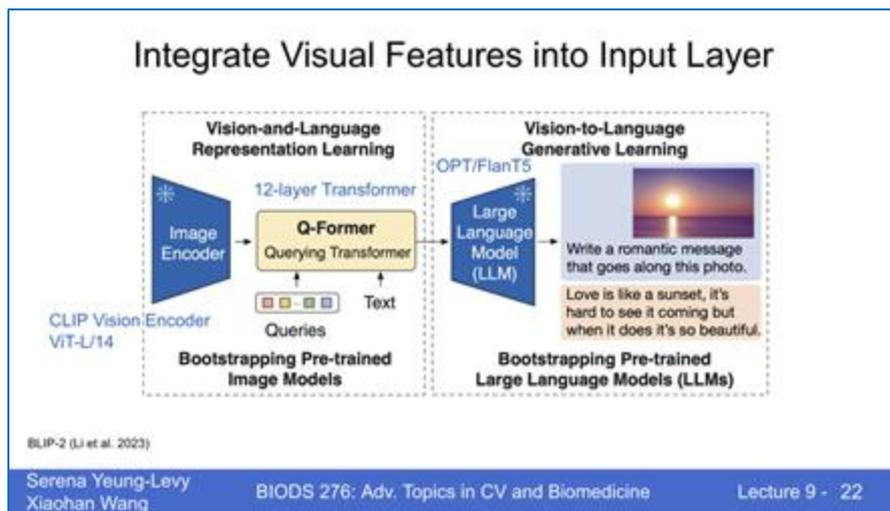
- CheXinstruct instruction-tuning dataset contains 6M instruction-image-answer triplets curated from across 34 tasks and 65 existing datasets



Chen et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024.

CheXagent: VLM for CXR interpretation

- CheXagent model is based on the BLIP-2 architecture (one of the VLM models mentioned last lecture, under the “integrate visual features into input layer” category together with LLaVA (but LLaVA-1.5 generally outperforms BLIP-2))



Chen et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024.

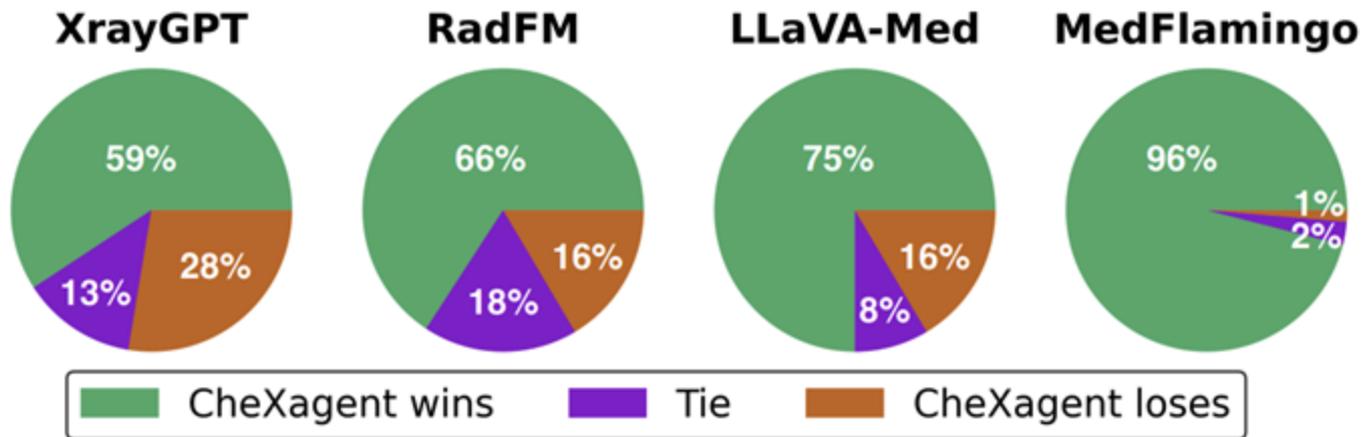
Evaluation on image perception tasks

Task	Dataset	General-domain FMs			Medical-domain FMs			CheXagent (Ours)
		BLIP-2	InstructBLIP	XrayGPT	MedFlamingo	RadFM	LLaVA-Med	
View Classification	MIMIC-CXR	28.8	25.3	24.0	25.0	28.5	23.8	97.5
	CheXpert	38.0	34.0	33.0	39.0	37.0	30.0	96.7
Binary Disease Classification	SIIM	53.0	54.0	50.0	50.0	50.0	49.0	64.0
	RSNA	50.0	60.0	50.0	50.0	50.0	44.0	81.0
	CheXpert	51.5	53.2	51.5	48.5	55.8	47.6	76.0
Single Disease Identification	OpenI	40.2	40.2	45.4	39.0	42.2	43.8	47.0
	MIMIC-CXR	25.6	22.6	24.1	25.6	27.2	26.7	30.3
	CheXpert	21.3	19.5	23.7	26.0	26.6	26.0	29.6
Multi Disease Identification	OpenI	48.5	54.4	57.7	46.1	52.8	53.9	55.6
	MIMIC-CXR	30.0	25.3	39.0	14.7	22.3	28.7	55.3
	CheXpert	4.3	6.1	3.9	7.1	23.6	2.1	52.1
Visual Question Answering	Rad-Reconstruct	41.2	42.4	38.6	45.5	48.5	34.9	57.1
	SLAKE	74.3	86.4	52.4	64.8	85.0	55.5	78.1
Image-Text Reasoning	OpenI	47.9	52.6	52.4	54.7	54.0	45.8	59.0

Chen et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024.

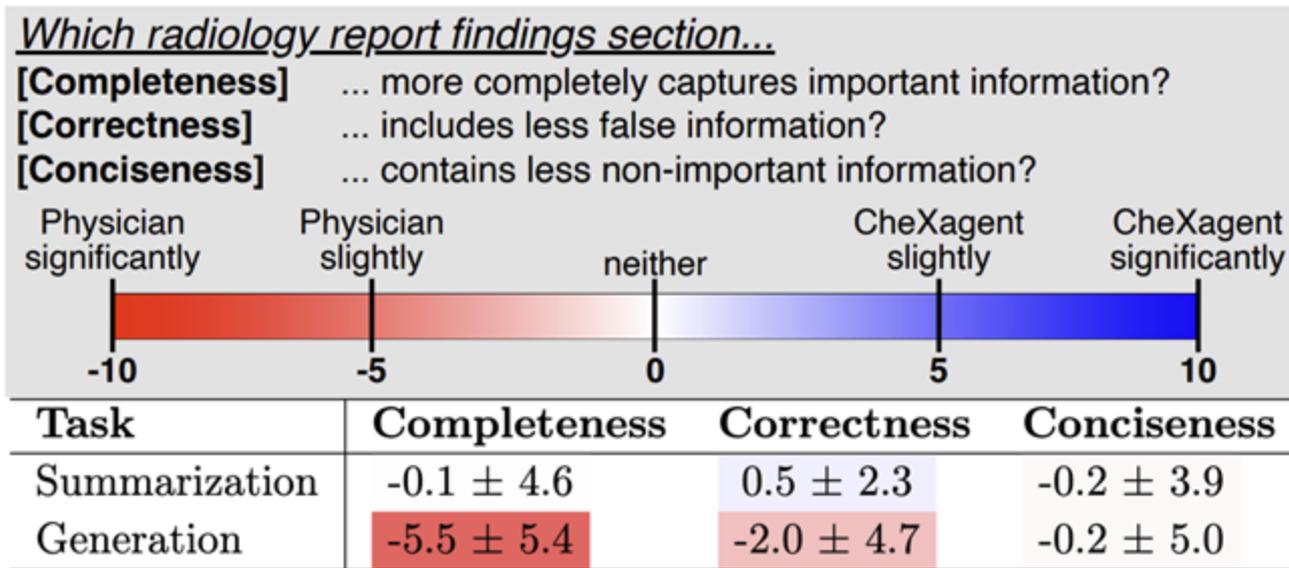
Evaluation on findings generation in reports

- Automated GPT-4 based evaluation, where GPT-4 is provided with a reference report, findings generated by CheXAgent, and findings generated by comparison models, and prompted to select the report with the highest accuracy



Chen et al. CheXAgent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024.

Evaluation on findings generation in reports: reader study (assessment by five radiologists)



Chen et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024.

Evaluation on findings generation in reports: qualitative comparison of CheXagent with human physician report



CheXagent: the right-sided chest tube has been removed. there is no evidence of pneumothorax. there is a small right pleural effusion. bibasilar atelectasis is present. there is no pulmonary edema. the heart size is normal. the mediastinal contours are normal. the hilar contours are normal. there is no pneumothorax.

Physician: right-sided chest tube remains in place, with slight increase in size of a small right pleural effusion, but no visible pneumothorax. bibasilar linear atelectasis has slightly worsened, and there is a persistent small left pleural effusion.

Color key: Correct Error Refers to prior study

Evaluation on findings generation in reports: qualitative comparison of CheXAgent with human physician report



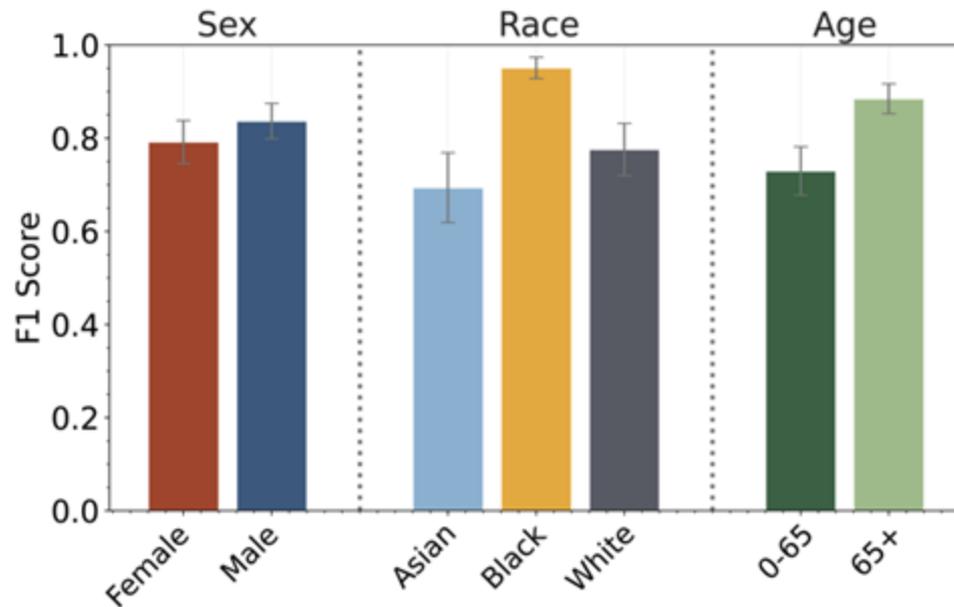
CheXAgent: the right-sided chest tube has been removed. there is no evidence of pneumothorax. there is a small right pleural effusion. bibasilar atelectasis is present. there is no pulmonary edema. the heart size is normal. the mediastinal contours are normal. the hilar contours are normal. there is no pneumothorax.

Physician: right-sided chest tube remains in place, with slight increase in size of a small right pleural effusion, but no visible pneumothorax. bibasilar linear atelectasis has slightly worsened, and there is a persistent small left pleural effusion.

Color key: Correct Error Refers to prior study

Physician reports more frequently refer to past studies

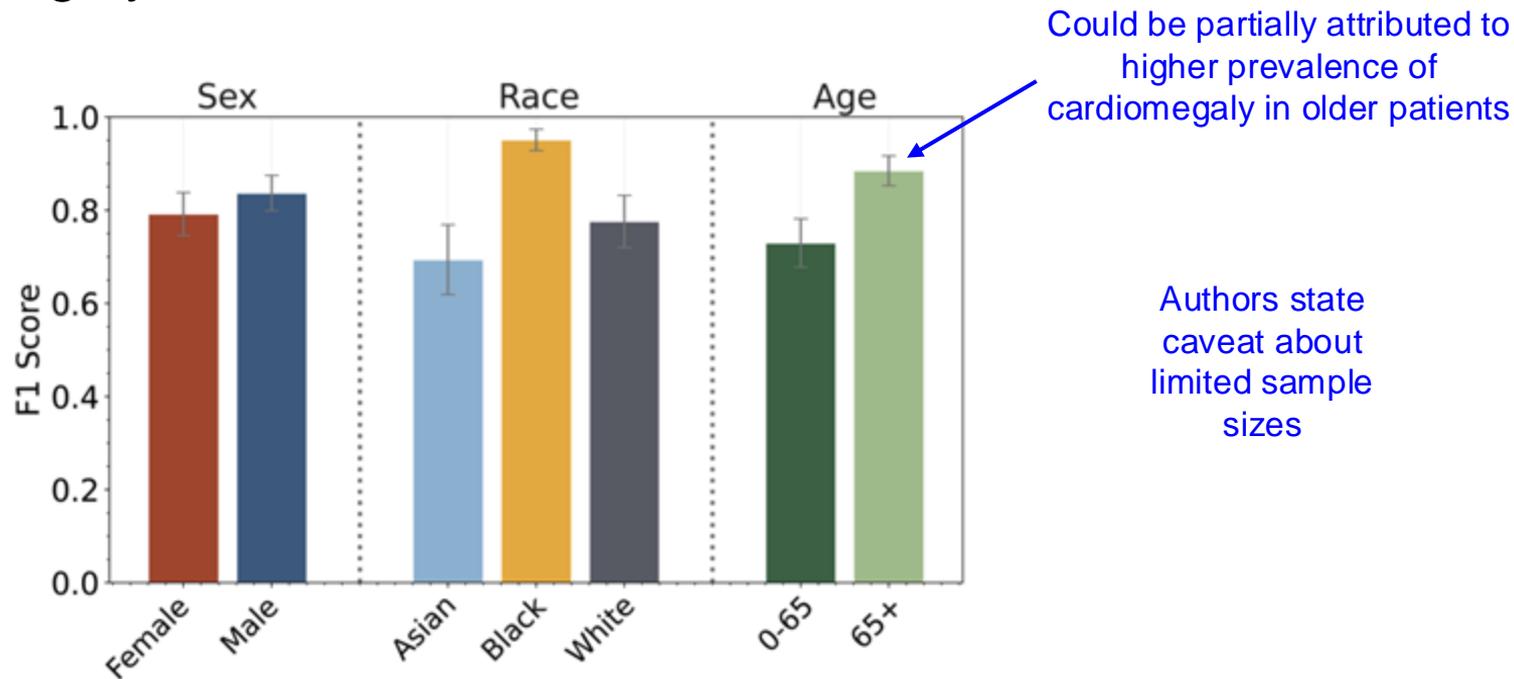
Example of fairness evaluation: subgroup performance comparison on cardiomegaly classification



Authors state
caveat about
limited sample
sizes

Chen et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024.

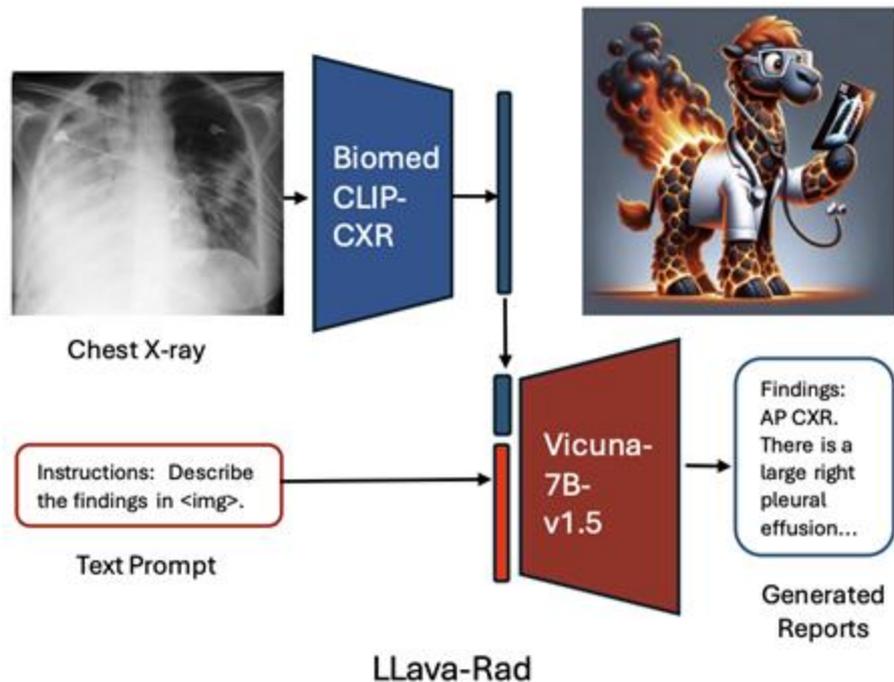
Example of fairness evaluation: subgroup performance comparison on cardiomegaly classification



Chen et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024.

LLaVA-Rad: specialized CXR VLM that improves over CheXpert

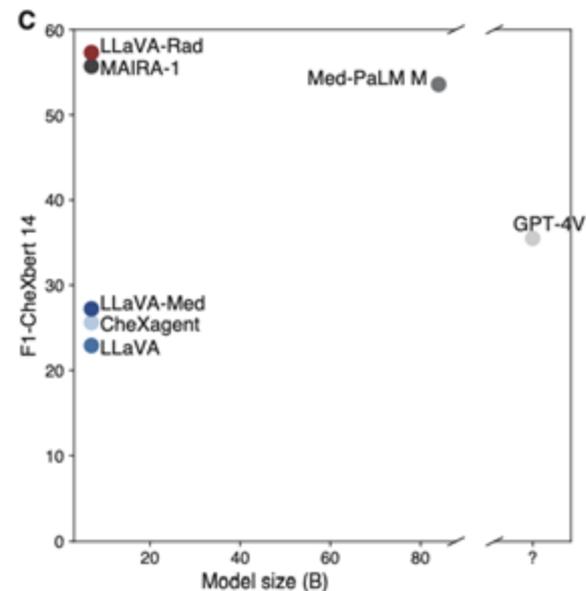
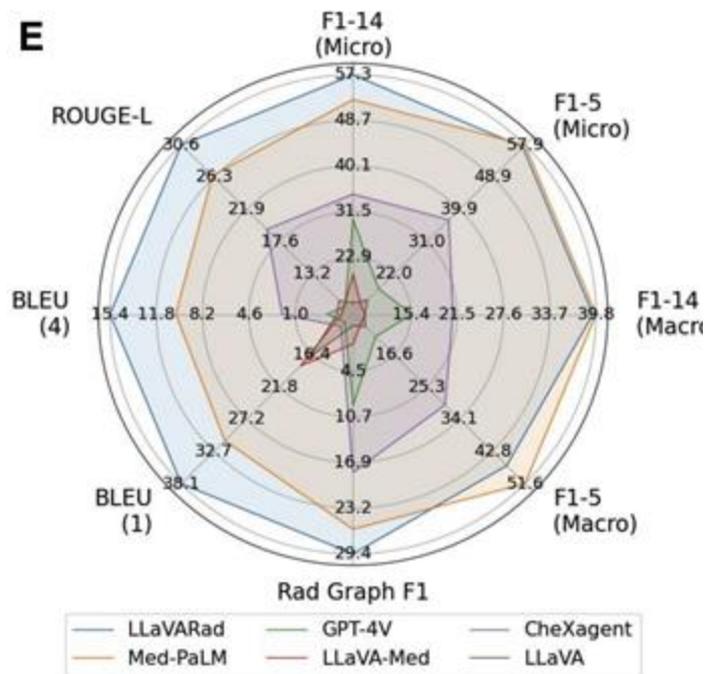
- Based on LLaVA (similar to LLaVA-Med), but trains from scratch (instead of fine-tuning LLaVA) since it uses a biomedical encoder (Biomed CLIP-CXR) instead of LLaVA's CLIP encoder.
- Trains using 697K image-report pairs from a collection of datasets



Chaves et al. Towards a clinically accessible radiology multimodal model: open-access and lightweight, with automatic evaluation. arXiv 2024.

LLaVA-Rad: specialized CXR VLM that improves over CheXpert

Outperforms LLaVA-Med, CheXagent, and GPT-4V, among other models (but probably not Med-Gemini)

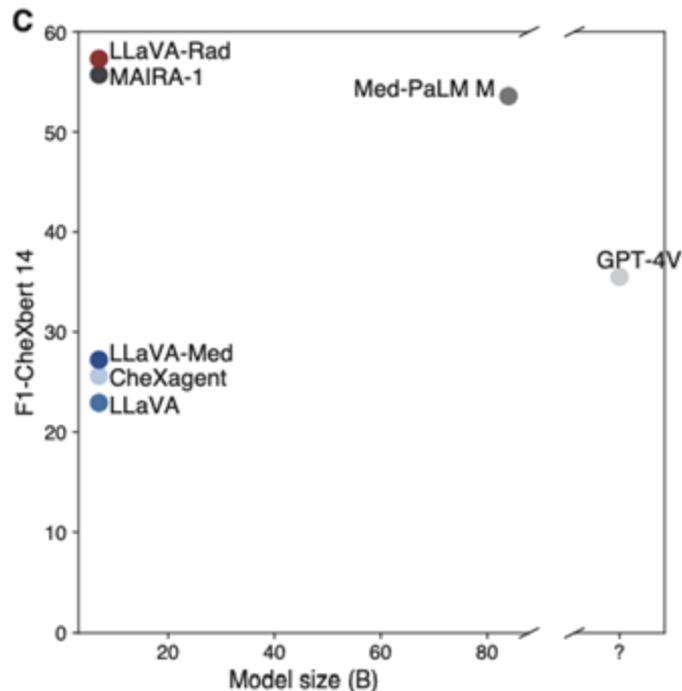


Results

Chaves et al. Towards a clinically accessible radiology multimodal model: open-access and lightweight, with automatic evaluation. arXiv 2024.

LLaVA-Rad: specialized CXR VLM that improves over CheXpert

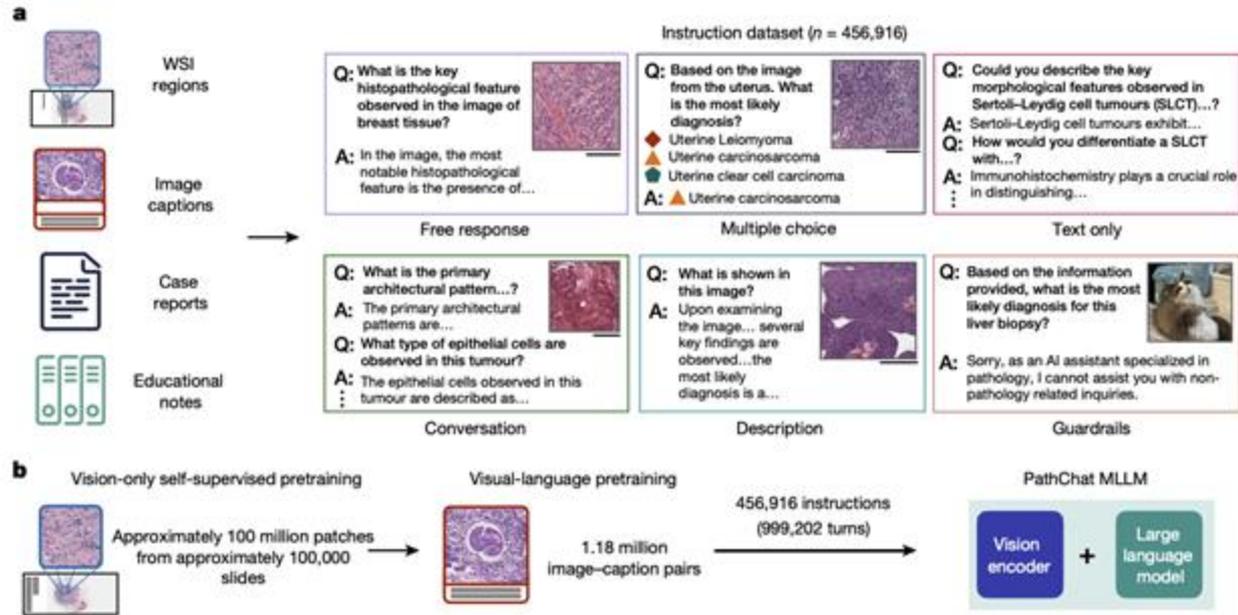
- Training and inference is reasonably fast:
 - Training takes one day on an 8xA100 server
 - Inference can be run on single (older generation) GPUs



Chaves et al. Towards a clinically accessible radiology multimodal model: open-access and lightweight, with automatic evaluation. arXiv 2024.

PathChat: specialized VLM for pathology

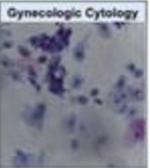
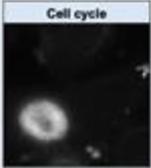
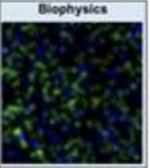
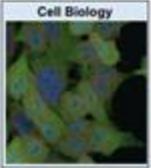
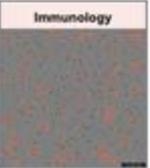
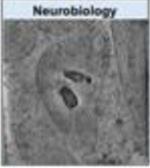
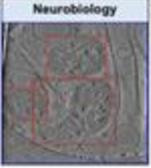
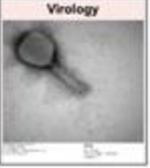
Trained on over 450K instructions from pathology. Will cover further in the discussion presentations!



Lu et al. A multimodal generative AI copilot for human pathology. Nature 2024.

μ -Bench: benchmarking VLMs on microscopy image interpretation tasks across scales

Evaluated proprietary and open-source VLMs on 22 biomedical tasks across various scientific disciplines (biology, pathology), microscopy modalities (electron, fluorescence, light), and scales (subcellular, cellular, tissue)

	Coarse-Grained Perception (Modality, Sub-modality, Domain, Sub-domain, Stain)	Fine-Grained Perception (Pap smear grading, Molecule colocalization...)	Fine-Grained Perception (Cell, Nucleus, Mitochondria localization)	COGNITION (Reasoning) (Pathology, Biology...)
Tissue	<p>What is the most likely technique used to stain this micrograph?</p> <p>A) IHC(DAB)</p> 	<p>Liquid-based cytology pap smear of human cancer lesions. Based on the cytogram, what is the most likely finding?</p> <p>E) Low-grade (LSIL) lesions</p> 	<p>Detect glands in colorectal tissue</p> <p>{ "x1": 0.0, "y1": 136.0, "x2": 382.0, "y2": 49.0 }, ...</p> 	<p>What is the significance of the pale area with the red outline?</p> <p>The pale area is a region of hypomyelination in the brain white matter. Loss of myelin is associated with disease processes, including demyelinating diseases or cerebrovascular injury</p> 
Cellular	<p>What is the most likely microscopy modality used to acquire this image?</p> <p>C) Fluorescence microscopy</p> 	<p>What is the most accurate description for the colocalization patterns displayed in the micrograph?</p> <p>B) Very weak</p> 	<p>Detect Nucleus</p> <p>{ "x1": 0.0, "x2": 599.4140625, "y1": 0.0, "y2": 0.0 }, ...</p> 	<p>What action is depicted in the image?</p> <p>HT55 cancer cells expressing a red cytoplasmic protein</p> 
Subcellular	<p>What is the most likely microscopy sub-modality used to acquire this image?</p> <p>D) Serial blockface scanning electron microscopy</p> 	<p>Based on the mitochondria cryo-et, what is the most likely mitochondrial morphology?</p> <p>A) Abnormal mitochondrial morphology</p> 	<p>Detect mitochondria</p> <p>{ "x1": 70.0, "y1": 136.0, "x2": 382.0, "y2": 391.0 }, ...</p> 	<p>What is shown in the image?</p> <p>A transmission electron microscope image of T4 bacteriophage, which infects Escherichia coli bacteria</p> 

Lozano et al. μ -Bench: A Vision-Language Benchmark for Microscopy Understanding. NeurIPS Datasets and Benchmarks 2024.

μ -Bench: benchmarking VLMs on microscopy image interpretation tasks across scales

GPT-4o still outperforms all other compared models, and non-specialist biomedical models often outperform specialist counterparts, suggesting room for improvement in how to leverage specialist domain data

μ -Bench					
Perception (Coarse-Grained)		Perception (Fine-Grained)		Cognition (Reasoning)	
Model	Accuracy (\pm CI)	Model	Accuracy (\pm CI)	Model	Accuracy (\pm CI)
GPT-4o	62.68 (\pm 0.35)	GPT-4o	51.73 (\pm 0.82)	GPT-4o	62.00 (\pm 9.00)
CogVLM	52.05 (\pm 0.35)	BiomedCLIP	34.65 (\pm 0.75)	QwenVLM	41.00 (\pm 10.00)
QwenVLM	49.85 (\pm 0.35)	CONCH	33.64 (\pm 0.72)	CogVLM	41.00 (\pm 10.00)
BiomedCLIP	47.57 (\pm 0.34)	ALIGN	31.9 (\pm 0.72)	OpenCLIP	38.33 (\pm 8.33)
ALIGN	40.7 (\pm 0.34)	CLIP	30.09 (\pm 0.71)	ALIGN	31.00 (\pm 9.00)
OpenCLIP	36.34 (\pm 0.33)	OpenCLIP	29.36 (\pm 0.69)	CLIP	28.00 (\pm 9.00)
PaliGemma	36.29 (\pm 0.33)	CogVLM	28.18 (\pm 0.70)	PaliGemma	25.00 (\pm 8.00)
CLIP	35.41 (\pm 0.34)	QuiltNet	27.85 (\pm 0.69)	BiomedCLIP	25.00 (\pm 8.00)
PLIP	31.11 (\pm 0.32)	QwenVLM	27.81 (\pm 0.70)	CONCH	18.00 (\pm 7.00)
CONCH	27.84 (\pm 0.31)	PLIP	25.49 (\pm 0.68)	Random	17.00 (\pm 7.00)
QuiltNet	26.58 (\pm 0.31)	PaliGemma	21.29 (\pm 0.64)	PLIP	17.00 (\pm 7.00)
Random	18.34 (\pm 0.27)	Random	19.13 (\pm 0.60)	QuiltNet	13.00 (\pm 6.00)

+ ■ General autoregressive VLMs ■ General contrastive VLMs ■ Pathology contrastive VLMS ■ Biomedical contrastive VLMS.

Lozano et al. μ -Bench: A Vision-Language Benchmark for Microscopy Understanding. NeurIPS Datasets and Benchmarks 2024.

Today we covered

Examples of vision-language generative models in biomedicine:

- Generalist (broad domain) biomedical VLMs
 - LLaVA-Med (open-source) and Med-Gemini (SoTA)
- Specialist biomedical VLMs:
 - CXR: CheXagent (earlier, though still 2024), LLaVA-Rad (improved)
 - Pathology: PathChat
 - Benchmarking VLMs across cell and tissue imaging domains: μ -Bench

Next time

- A Deeper Dive into GPUs and Compute