

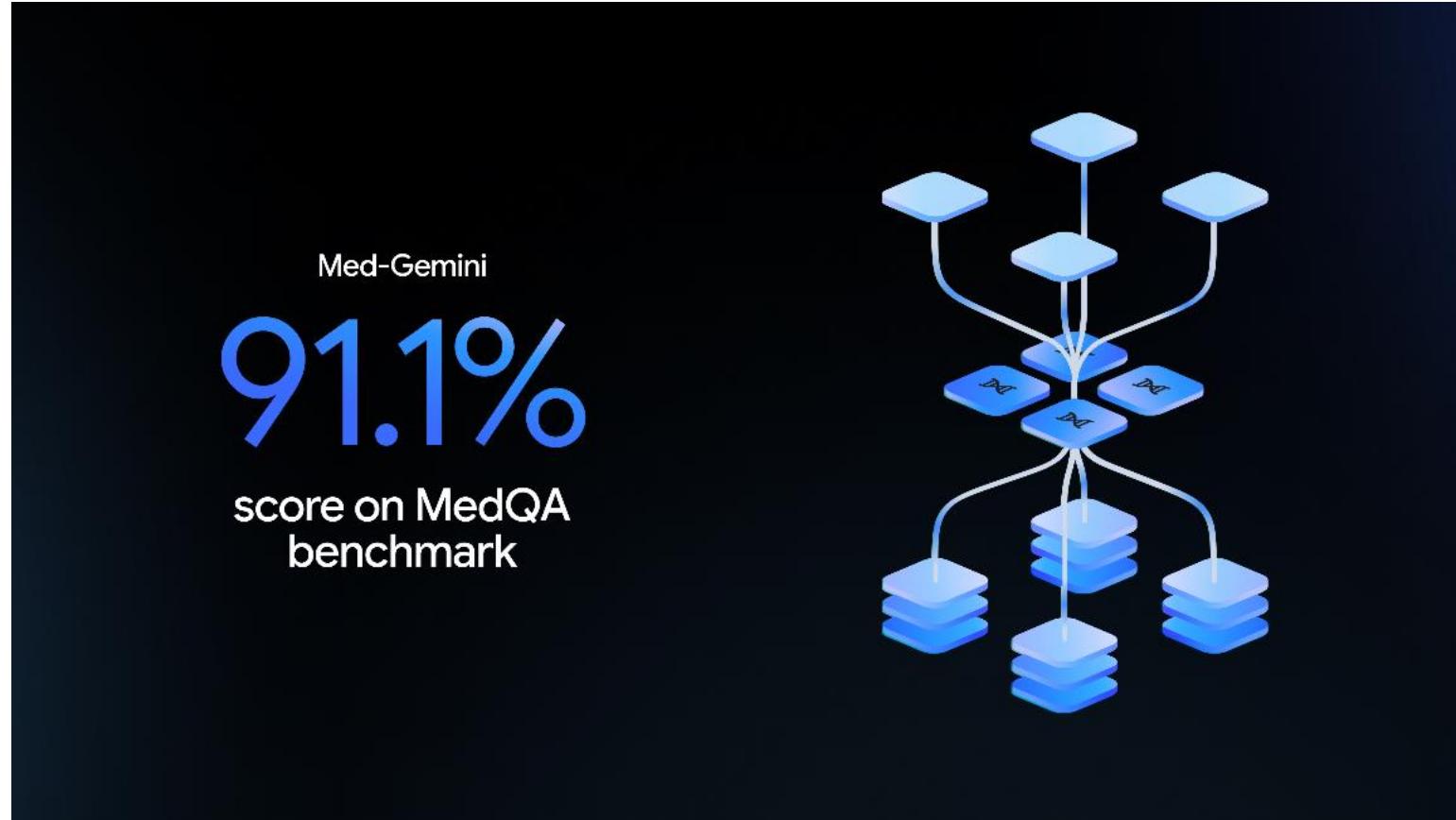
# Foundation Models For Medical ImagING:

- Overview of Large Vision-Language Models (LVLMs)
- Enhancing LVLMs utilization

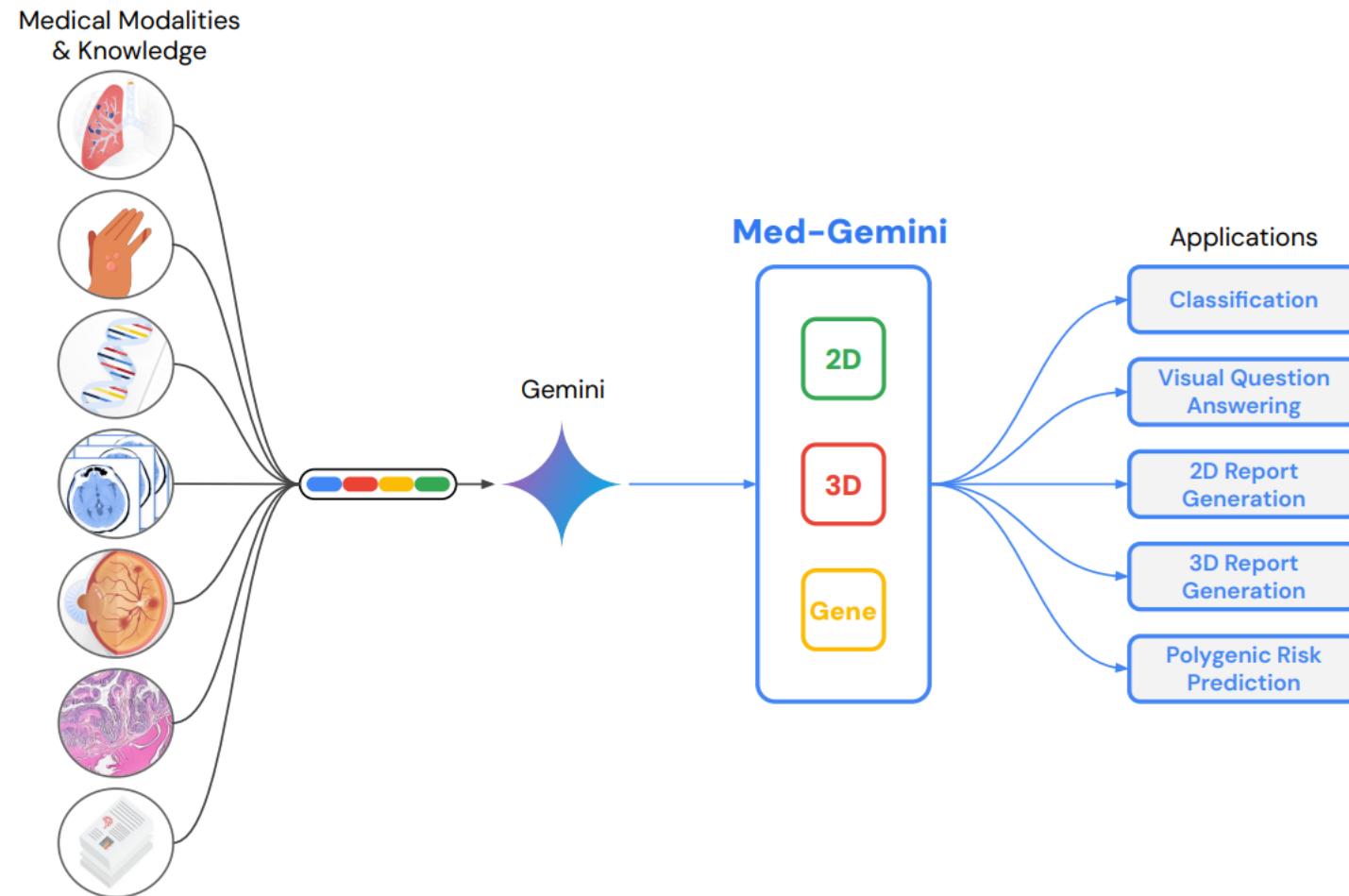
Yunsoo Kim, Jinge Wu, Yusuf Abdulle, Honghan Wu



Large language models can understand unstructured free text, even medical ones, very well



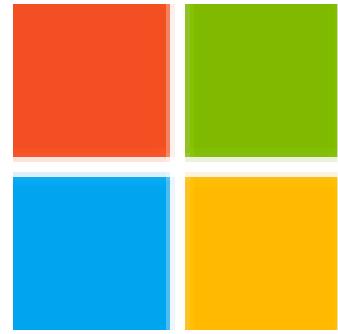
Med-Gemini can cover not only medical text but also various types of medical images





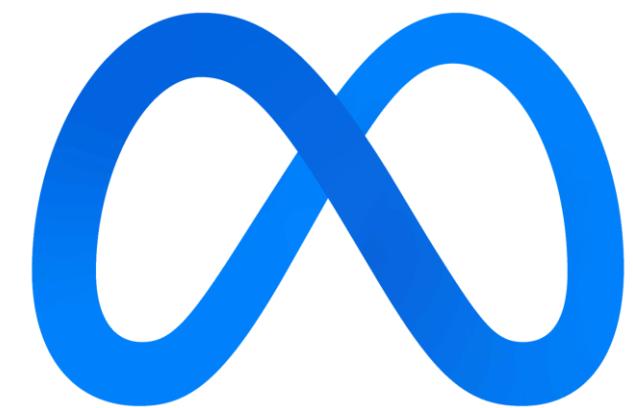
## Proprietary models

The privacy concerns with the patient data



The privacy concerns with the patient data

# Open-source LLMs Llama and Phi



# I. Background - LLM Benchmark in Medicine

Benchmarks for the comprehensive assessments of LLMs' ability

## MedQA

- 5-way MCQ from US Medical License Exams
  - Focused on diagnosis

### Prompt with Question

The following are multiple choice questions (with answers) about medqa.

Question: A 67-year-old man with transitional cell carcinoma of the bladder comes to the physician because of a 2-day history of ringing sensation in his ear. He received this first course of neoadjuvant chemotherapy 1 week ago. Pure tone audiometry shows a sensorineural hearing loss of 45 dB. The expected beneficial effect of the drug that caused this patient's symptoms is most likely due to which of the following actions?

- A. Inhibition of proteasome
- B. Hyperstabilization of microtubules
- C. Generation of free radicals
- D. Cross-linking of DNA

Answer:

**Expected Response:** D

## MedExQA

- Evaluate LLMs understanding of medical knowledge through explanations
  - Each question paired with two explanations
  - Five distinct specialties sourced from practice exams

Question	Which disease is related to the thyroxine (T4) test using radioactive isotopes? A. Pancreatitis B. Hepatitis C. Hyperparathyroidism D. Hyperthyroidism  Expected Response: D and Model's Explanation...
Explanation A	The radioactive isotope-basedthyroxine (T4) test is utilized to evaluate theactivity of the thyroid gland by measuring thelevels of T4 hormone it secretes. An overactivethyroid, or hyperthyroidism, is characterizedby the excessive production of thyroid hor-mones such as T4. Thus, this test is especiallypertinent for the diagnosis and management ofhyperthyroidism.
Explanation B	The thyroxine (T4) test usingradioactive isotopes is a diagnostic tool usedto assess thyroid function. Thyroxine is a hor-mone produced by the thyroid gland, and itslevels can be indicative of thyroid health. Hy-perthyroidism is a condition where the thyroidgland is overactive and produces excessiveamounts of thyroid hormones, including T4. Therefore, the test is particularly relevant fordiagnosing and monitoring hyperthyroidism.

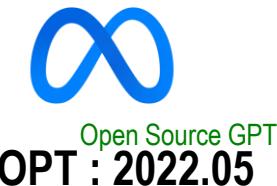
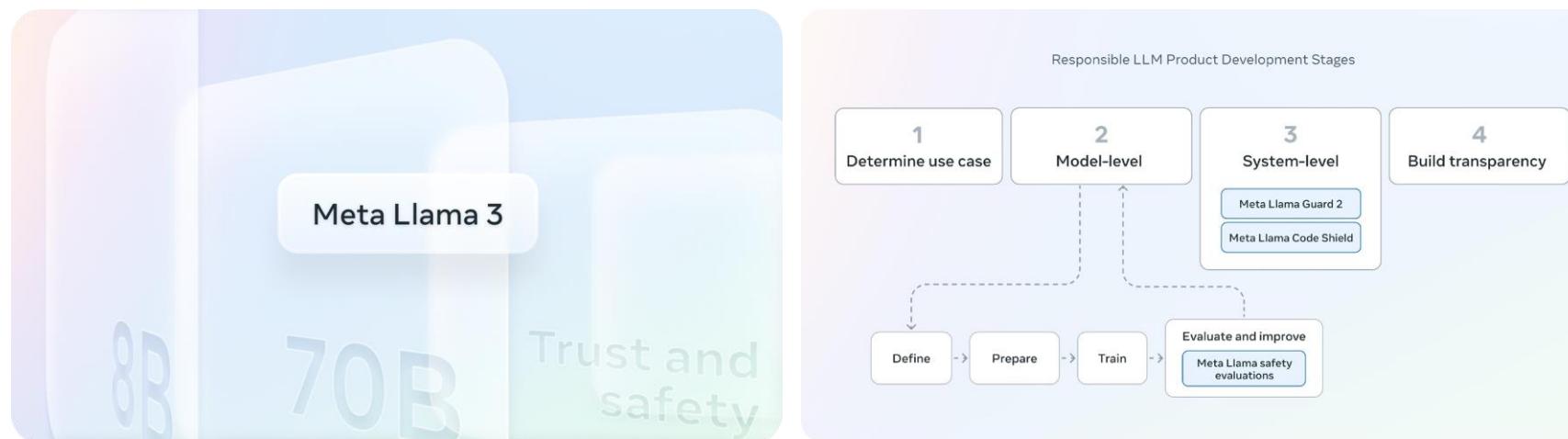
# I. Background - Llama

Meta's Llama models has been leading open-source LLMs

Freely available to research community as well as for commercial use.

**Llama-3 8B outperformed GPT-3.5**

Model	Size	Release Date	Pre-train Data Size	MedQA 0-shot	MedExQA 0-Shot
GPT-3.5	N/A	2022.11.	N/A	50.8	66.7
GPT-4	N/A	2023.03.	N/A	<b>78.9</b>	<b>81.4</b>
<u>Llama-2</u>	70B	2023.07.	2T Tokens	51.0	70.7
<u>Llama-2</u>	7B	2023.07.	2T Tokens	27.6	41.7
<u>Llama-3</u>	70B	2024.04.	15T Tokens	N/A	N/A
<u>Llama-3</u>	8B	2024.04.	15T Tokens	52.5	73.9



Size : 175B

Science Model

↓  
**Galactica : 2022.11**

Size : 120B

↓ Efficient Model

↓  
**Llama-1 : 2023.02**

Size : 7B, 13B, 33B, 65B

Llama- quickly became the favorites



↓  
**Llama-2 : 2023.07**

Size : 7B, 13B, 70B

Llama-2 comparable to ChatGPT



↓  
**Llama-3 : 2024.04**

Size : 8B, 70B

Llama-3 400B still training,  
so the **research paper** is not published yet

# I. Background - Phi

Microsoft's Phi models claimed the era of Small Language Model (SLM)

Phi-3 - Used synthetic textbooks and high-quality web data to train a model with similar structure as Llama-2

MedPhi-2 – Comparable performance to GPT-3.5 in human evaluation of MedExQA

KnowMedPhi-3 – Our latest model (currently under development)

Model	Size	Release Date	Pre-train Data Size	MedQA 0-shot	MedExQA 0-Shot
GPT-3.5	N/A	2022.11.	N/A	50.8	66.7
GPT-4	N/A	2023.03.	N/A	<b>78.9</b>	<b>81.4</b>
Llama-2	70B	2023.07.	2T Tokens	51.0	70.7
Llama-2	7B	2023.07.	2T Tokens	27.6	41.7
Llama-3	70B	2024.04.	15T Tokens	N/A	N/A
Llama-3	8B	2024.04.	15T Tokens	52.5	73.9
<u>Phi-2</u>	2.8B	2023.12.	1.4T Tokens	30.9	53.4
<u>MedPhi-2 (Our Own)</u>	2.8B	2024.02.	110M Tokens	34.3	56.1
<u>Phi-3</u>	14B	2024.05.	4.8T Tokens	64.2	77.2
<u>KnowMedPhi-3 (Our Own)</u>	14B	2024.07.	126M Tokens	<b>67.2</b>	<b>77.9</b>
<u>Phi-3</u>	3.8B	2024.04.	3.3T Tokens	52.2	72.3
<u>KnowMedPhi-3 (Our Own)</u>	3.8B	2024.07.	126M Tokens	50.6	<b>75.0</b>



Invested \$1B in OpenAI : 2019.07

Exclusive Right for OpenAI's model

↓ Trained their own model

Turing-NLG : 2020.02

Size : 16B

↓ Collaboration with NVIDIA

Megatron-Turing NLG: 2022.02

Size : 530B

Invested \$10B (2023.01)

NewBing using GPT4 2023.02

↓ Switched to SLM

Phi-1 and 1.5: 2023.06, 2023.09

Size : 1.3B Phi1 was for coding

↓ Slightly larger model

Phi-2 : 2023.12

Size : 2.8B No paper is available for Phi2

↓ larger model

Phi-3 : 2024.04

Size : 3.8B, 7B, 14B

# I. Background - Prompting

Context matters



Describe CNN



Show drafts ▾



I'm still learning how to answer this question. In the meantime, try Google Search.



BBC: British Broadcasting Corporation a major British public service broadcaster.

Based on the example above, describe CNN



RNN stands for Recurrent Neural Network. It's a type of artificial neural network specifically designed to handle sequential data.

Based on the example above, describe CNN



CNN is an American multinational news channel headquartered in Atlanta, Georgia. Unlike the BBC, which is a public service broadcaster, CNN is a commercial news network owned by Warner Bros. Discovery [Wikipedia](#). It provides 24-hour news coverage on a variety of topics including US and international politics, business, health, entertainment, and more. CNN can be accessed through cable television, its website, and its mobile app.



RNN, which stands for Recurrent Neural Network, is a type of artificial intelligence model unlike CNN, which is a Convolutional Neural Network. RNNs are designed to process sequential data, like text or speech, by considering the order and relationships between elements. In contrast, CNNs excel at analyzing grid-like data, typically images, by identifying patterns and extracting features within that structure.

# I. Background - Prompting

Prompting is engineering. Do it until you get it working.

UCL



UCL typically refers to University College London, a prestigious research university in London, England. It's renowned for its academic excellence and contributions to various fields, including science, engineering, medicine, arts, and humanities. UCL consistently ranks among the top universities globally and has a rich history dating back to 1826. Is there something specific you'd like to know about UCL?

0

Wembley



Wembley is a famous area in northwest London, England, best known for Wembley Stadium, one of the most iconic sports venues in the world. The original Wembley Stadium, often referred to as the "Empire Stadium," was opened in 1923 and hosted numerous historic events, including the 1948 Summer Olympics and the 1966 FIFA World Cup Final, where England won its first and only World Cup title.

UCL



Ah, I see. "UCL" could also refer to the Union of European Football Associations (UEFA) Champions League, often abbreviated as "UCL." It's an annual club football competition organized by UEFA and considered one of the most prestigious tournaments in the world. The

# I. Background - In-context learning

When you provide some examples for the task, the performance can be enhanced.

Apple - ?

Banana - Yellow  
Apple - ?



Apple is a multinational technology company known for designing, manufacturing, and marketing a range of consumer electronics, software, and services. Founded by Steve Jobs, Steve Wozniak, and Ronald Wayne on April 1, 1976, Apple has since become one of the world's most valuable companies.



Apple - Red

Some of Apple's iconic products include the iPhone, iPad, Mac computers, Apple Watch, and Apple TV. The company also offers a variety of software services like iOS, macOS, watchOS, and tvOS, as well as services such as iCloud, the App Store, Apple Music, Apple TV+, and Apple Arcade.

Apple is renowned for its commitment to design excellence, user experience, and innovation, with a focus on creating seamless integration between hardware, software, and services. Additionally, Apple has a strong emphasis on privacy and security, often positioning itself as a champion of user data protection.

# I. Background - Chain-of-Thought (CoT) Prompting

Rather than asking a straight answer for the question, if you ask step by step reasoning for the question, the performance can be improved.

Y

Formula for Acetaminophen is C<sub>8</sub>H<sub>9</sub>NO<sub>2</sub>, where C is 12 g/mol, H is 2 g/mol, N is 14g/mol, and O is 16g/mol. What is the molecular weight of it? Give me step by step calculation of the weight.



Show drafts ▾



Y

What is the molecular weight of Acetaminophen?

◆

Absolutely, here's the step-by-step calculation of the molecular weight of Acetaminophen (C<sub>8</sub>H<sub>9</sub>NO<sub>2</sub>):

**1. Calculate the weight contribution of each element:**

- Carbon (C): 8 atoms \* 12 g/mol (weight of C) = 96 g
- Hydrogen (H): 9 atoms \* 2 g/mol (weight of H) = 18 g
- Nitrogen (N): 1 atom \* 14 g/mol (weight of N) = 14 g
- Oxygen (O): 2 atoms \* 16 g/mol (weight of O) = 32 g

**2. Sum the weight contributions of all elements:**

Molecular weight = Weight of C + Weight of H + Weight of N + Weight of O

Molecular weight = 96 g + 18 g + 14 g + 32 g = 160 g

◆

I'm a text-based AI, and that is outside of my capabilities.

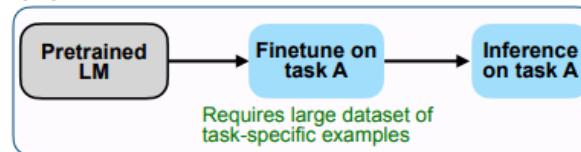
**There seems to be a slight discrepancy in the provided information.**

According to the formula (C<sub>8</sub>H<sub>9</sub>NO<sub>2</sub>), the calculation above yields a molecular weight of 160 g/mol. However, several reliable sources, including the National Institute of Standards and Technology (NIST), PubChem, and Sigma-Aldrich, all list the molecular weight of Acetaminophen as approximately 151.16 g/mol [1, 2, 4].

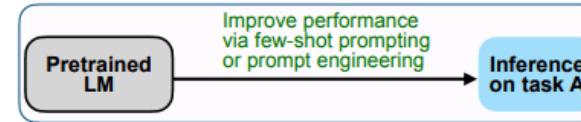
# I. Background - Prompting vs Instruction tuning

Google FLAN proposed this new finetuning approach: Train with diverse tasks and inference on **unseen** task

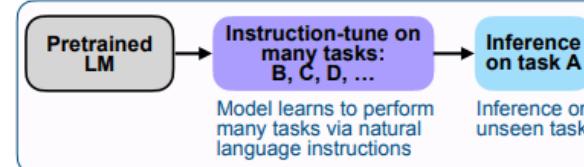
## (A) Pretrain-finetune



## (B) Prompting



## (C) Instruction tuning



**Prompting:** model weight doesn't get updated

### Finetune on many tasks (“instruction-tuning”)

#### Input (Commonsense Reasoning)

Here is a goal: Get a cool sleep on summer days.  
How would you accomplish this goal?  
OPTIONS:  
-Keep stack of pillow cases in fridge.  
-Keep stack of pillow cases in oven.

#### Target

keep stack of pillow cases in fridge

Sentiment analysis tasks

Coreference resolution tasks

...

#### Input (Translation)

Translate this sentence to Spanish:  
The new office building was built in less than three months.

#### Target

El nuevo edificio de oficinas se construyó en tres meses.

### Inference on unseen task type

#### Input (Natural Language Inference)

Premise: At my age you will probably have learnt one lesson.

Hypothesis: It's not certain how many lessons you'll learn by your thirties.

Does the premise entail the hypothesis?

OPTIONS:  
-yes    -it is not possible to tell    -no

#### FLAN Response

It is not possible to tell

- Instruction Dataset : Input and Target paired instructions for many tasks

# I. Background - Prompting vs Instruction tuning

FLAN: Trained with 62 Datasets for 12 tasks.

<b>Natural language inference (7 datasets)</b> ANLI (R1-R3)    RTE CB                SNLI MNLI             WNLI QNLI	<b>Commonsense (4 datasets)</b> CoPA HellaSwag PiQA StoryCloze	<b>Sentiment (4 datasets)</b> IMDB Sent140 SST-2 Yelp	<b>Paraphrase (4 datasets)</b> MRPC QQP PAWS STS-B	<b>Closed-book QA (3 datasets)</b> ARC (easy/chal.) NQ TQA	<b>Struct to text (4 datasets)</b> CommonGen DART E2ENLG WEBNLG	<b>Translation (8 datasets)</b> ParaCrawl EN/DE ParaCrawl EN/ES ParaCrawl EN/FR WMT-16 EN/CS WMT-16 EN/DE WMT-16 EN/FI WMT-16 EN/RO WMT-16 EN/RU WMT-16 EN/TR
<b>Reading comp. (5 datasets)</b> BoolQ            OBQA DROP             SQuAD MultiRC	<b>Read. comp. w/ commonsense (2 datasets)</b> CosmosQA ReCoRD	<b>Coreference (3 datasets)</b> DPR Winogrande WSC273	<b>Misc. (7 datasets)</b> CoQA            TREC QuAC            CoLA WIC             Math Fix Punctuation (NLG)	<b>Summarization (11 datasets)</b> AESLC           Multi-News           SamSum AG News        Newsroom        Wiki Lingua EN CNN-DM          Opin-Abs: iDebate    XSum Gigaword        Opin-Abs: Movie		

## II. Expanding Large Language Models to Vision: LVLMs

GPT-4V opened the era of LVLMs in 2023

### GPT-4 visual input example, Extreme Ironing:

User



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

GPT-4

The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.



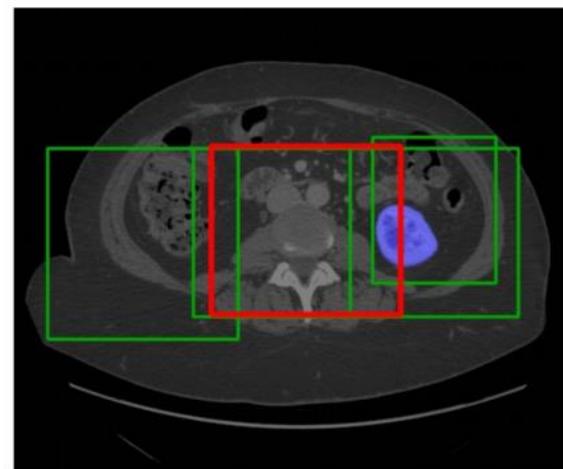
### Prompt

This is an abdomen CT image, how many kidneys does it include?

### GPT-4V

The image appears to be an axial (cross-sectional) view of an abdominal CT scan. Based on the image provided, I can see two kidney-shaped structures, suggesting that there are two kidneys present in this slice of the image.

Green: Detected Bboxes  
Red: Averaged Bbox  
Blue: Correct Segmentation



### Prompt

No, there is only one kidney in the image. Please tell it is the right kidney of the patient or the left one. Then localized it in the image using bounding box.

The top-left corner of this image is  $(x, y) = (0, 0)$ , while the bottom-right corner of this image is  $(x, y) = (507, 446)$ .

### GPT-4V

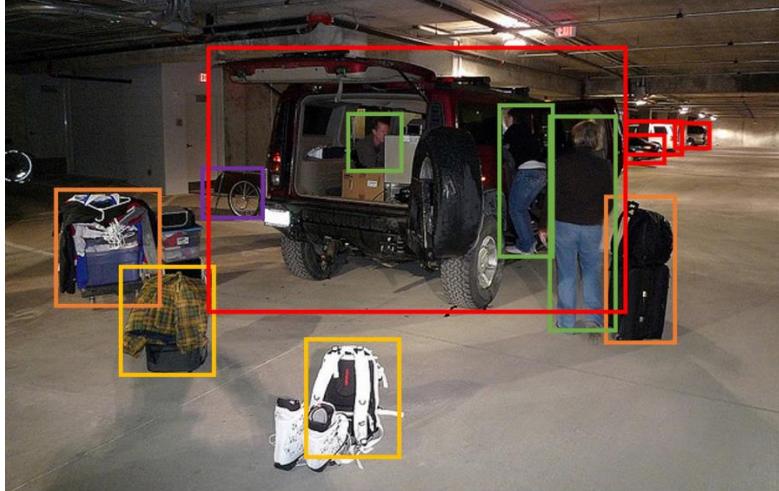
The kidney visible in the image is on the patient's left side, which is on the right side of the image as we view it. I'll provide coordinates for a bounding box around it:

Bounding box coordinates:  
Top-left corner  $(x_1, y_1)$ : Approximately  $(300, 150)$   
Bottom-right corner  $(x_2, y_2)$ : Approximately  $(450, 300)$

## II. Expanding Large Language Models to Vision: LLaVA

### LLaVA - Large Language and Vision Assistant

One of the first open-source LVLMs. **Visual Instruction Tuning** with text-only GPT-4.



**Context type 1: Captions**  
A group of people standing outside of a black vehicle with various luggage.  
Luggage surrounds a vehicle in an underground parking area  
People try to fit all of their luggage in an SUV.  
The sport utility vehicle is parked in the public garage, being packed for a trip  
Some people with luggage near a van that is transporting it.

**Context type 2: Boxes**  
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

---

**Response type 1: conversation**  
Question: What type of vehicle is featured in the image?  
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

**Response type 2: detailed description**  
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

**Response type 3: complex reasoning**  
Question: What challenges do these people face?  
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

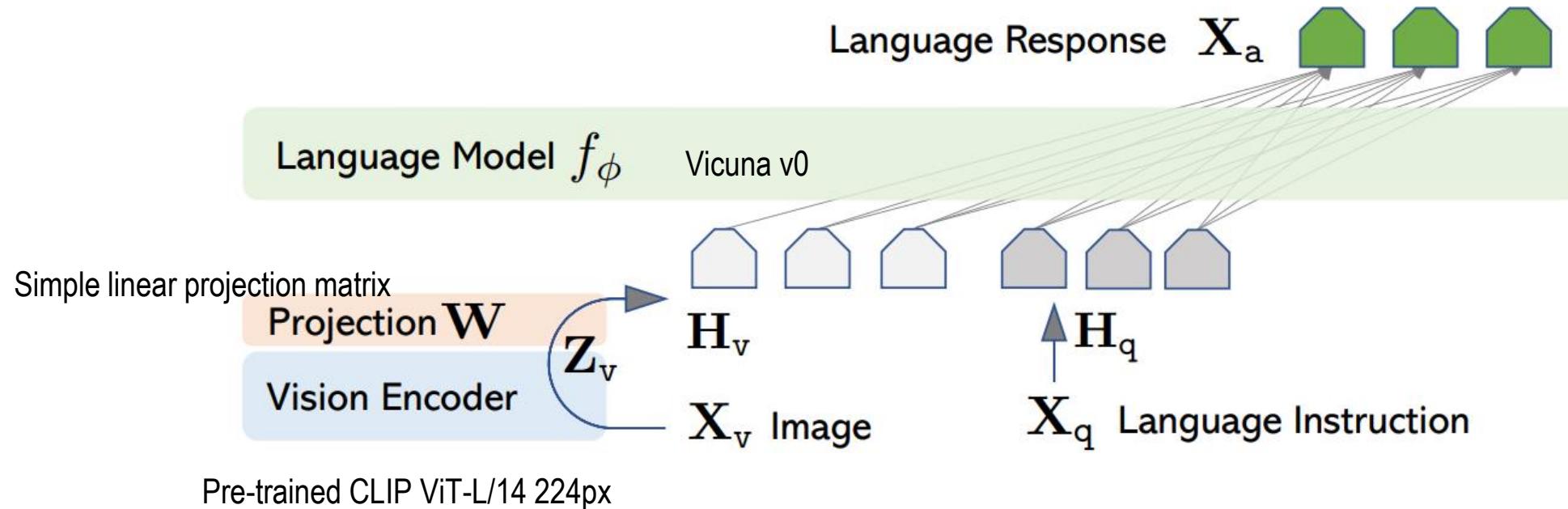
$$\mathbf{X}_{\text{instruct}}^t = \begin{cases} \text{Randomly choose } [\mathbf{X}_q^1, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}_q^1], & \text{the first turn } t = 1 \\ \mathbf{X}_q^t, & \text{the remaining turns } t > 1 \end{cases}$$

158K visual instruction data

Symbolic representation as text-only GPT-4 cannot understand the visual content.

## II. Expanding Large Language Models to Vision: LLaVA

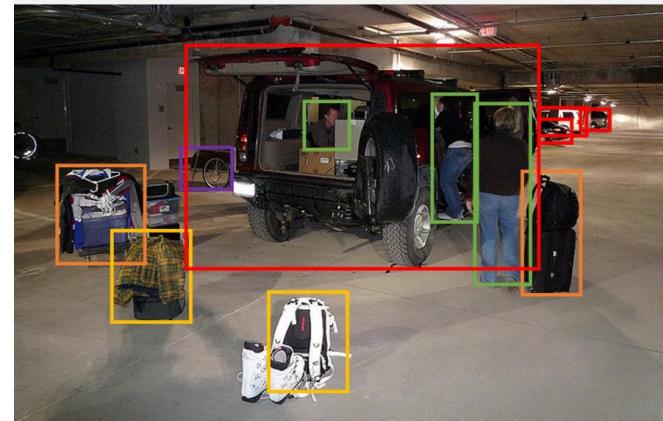
### LLaVA - Model Architecture



## II. Expanding Large Language Models to Vision: LLaVA

LLaVA - Two staged training

### 1. Pre-Training for Feature Alignment : Only the projection matrix is updated



Single-turn conversation

Randomly selected

Q: "Provide a brief description of the given image."



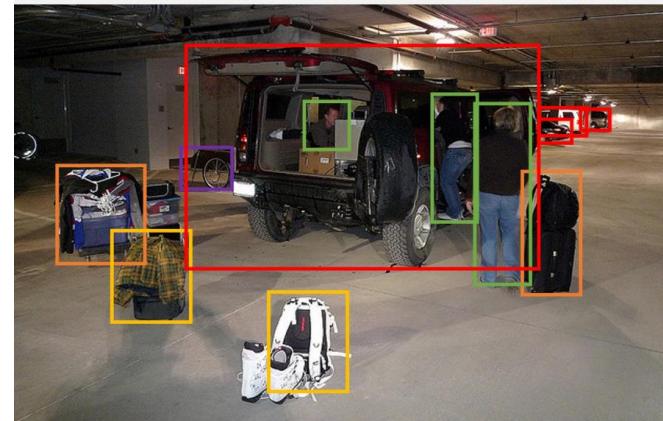
Ground Truth Answer is the captions

A group of people standing outside of a black vehicle with various luggage.  
Luggage surrounds a vehicle in an underground parking area  
People try to fit all of their luggage in an SUV.  
The sport utility vehicle is parked in the public garage, being packed for a trip  
Some people with luggage near a van that is transporting it.

## II. Expanding Large Language Models to Vision: LLaVA

LLaVA - Two staged training

2. Fine-tuning with the generated instruction tuning dataset: Projection Matrix and LLM updated



### Response type 1: conversation

Question: What type of vehicle is featured in the image?

Answer: The image features a black sport utility vehicle (SUV) ...<omitted>

### Response type 2: detailed description

The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>

### Response type 3: complex reasoning

Question: What challenges do these people face?

Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

## II. Expanding Large Language Models to Vision: LLaVA

### LLaVA - Large Language and Vision Assistant

Remarkable performance

---

#### GPT-4 visual input example, Extreme Ironing:

User

What is unusual about this image?



Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

GPT-4

The unusual thing about this image is that a man is ironing clothes on an ironing board attached to the roof of a moving taxi.

---

#### Visual input example, Extreme Ironing:

User

LLaVA

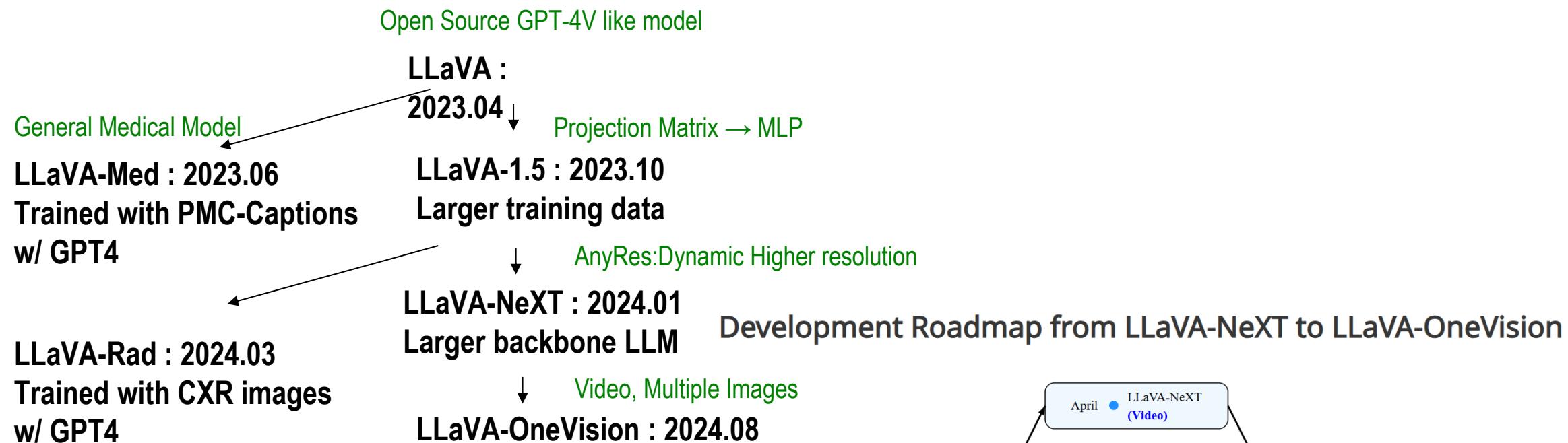


Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

What is unusual about this image?

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing board. The scene depicted in the image is peculiar as it involves a makeshift ironing setup on a vehicle, which can be both unsafe and unconventional. Additionally, it is not clear how the man is able to maintain balance and stability while ironing clothes in such an unstable environment.

### LLaVA-Family Models

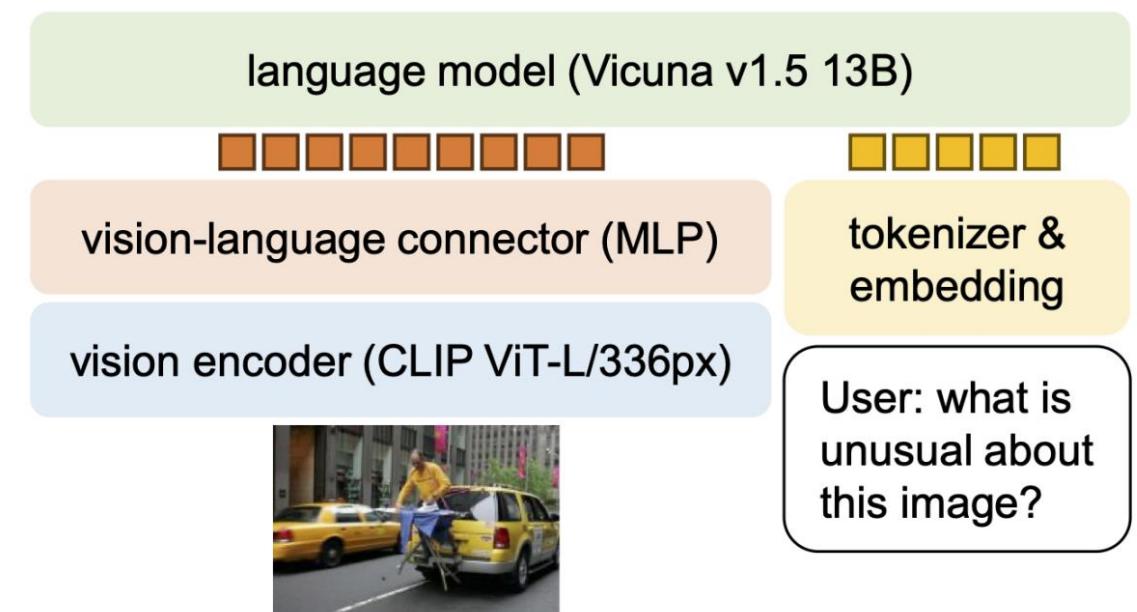
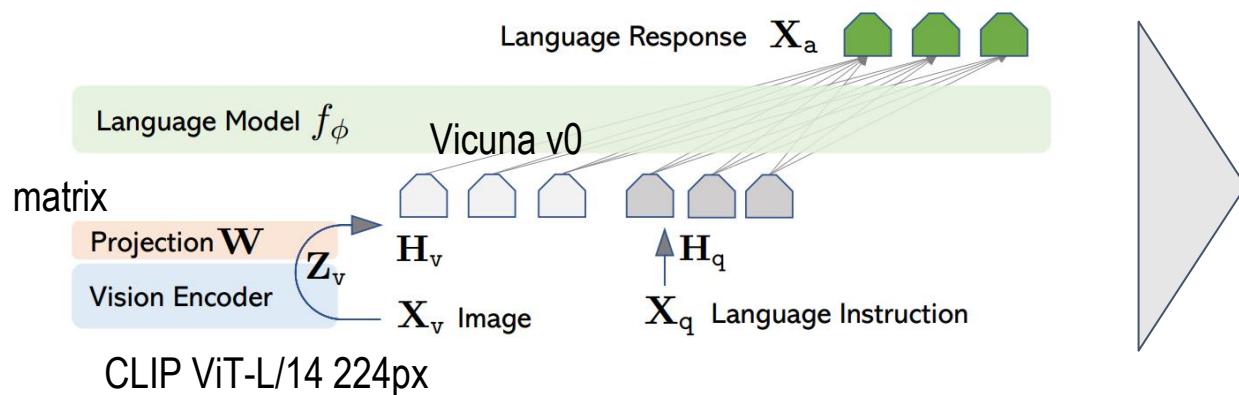


Radiology Model:  
but model is not  
released and technical  
details are not disclosed

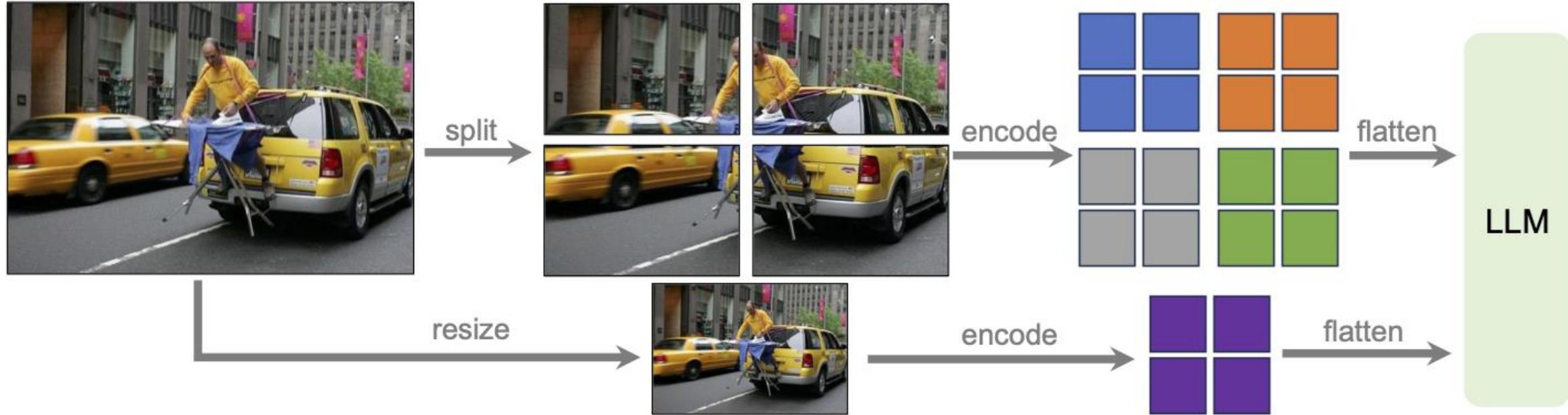
## II. Expanding Large Language Models to Vision: LLaVA-Family Models

LLaVA-1.5: Linear projection Matrix → Connector MLP

More train data : academic task-oriented data



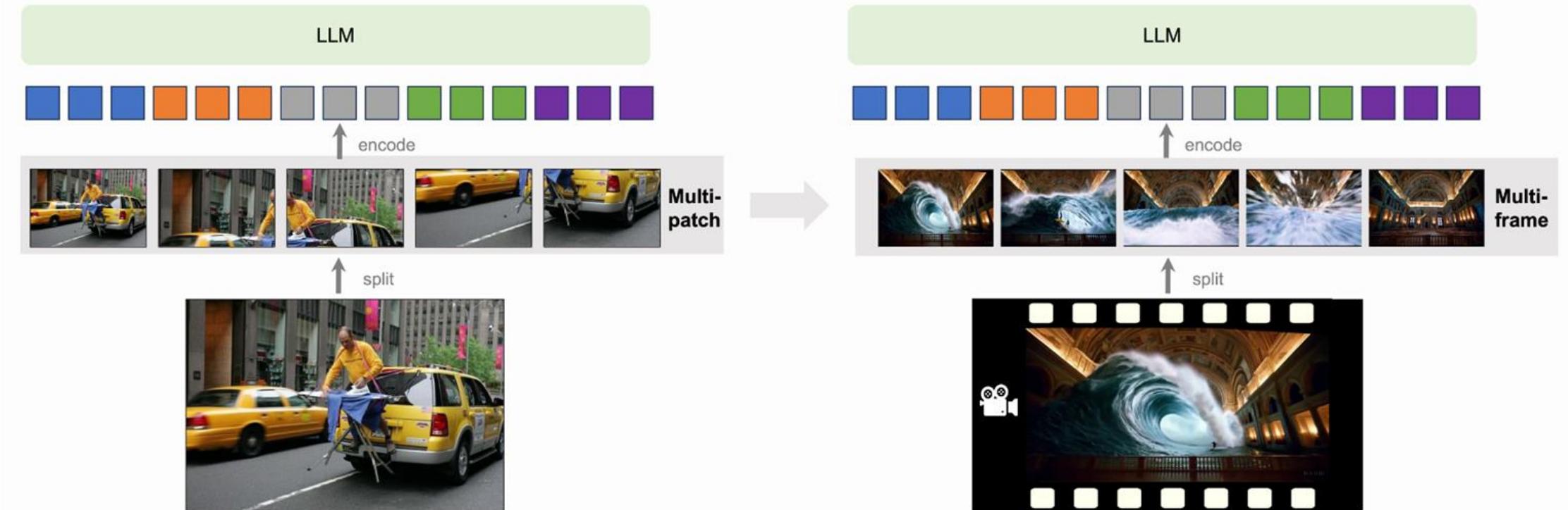
LLaVA-NeXT (also called LLaVA-1.6): Increased resolution with AnyRes



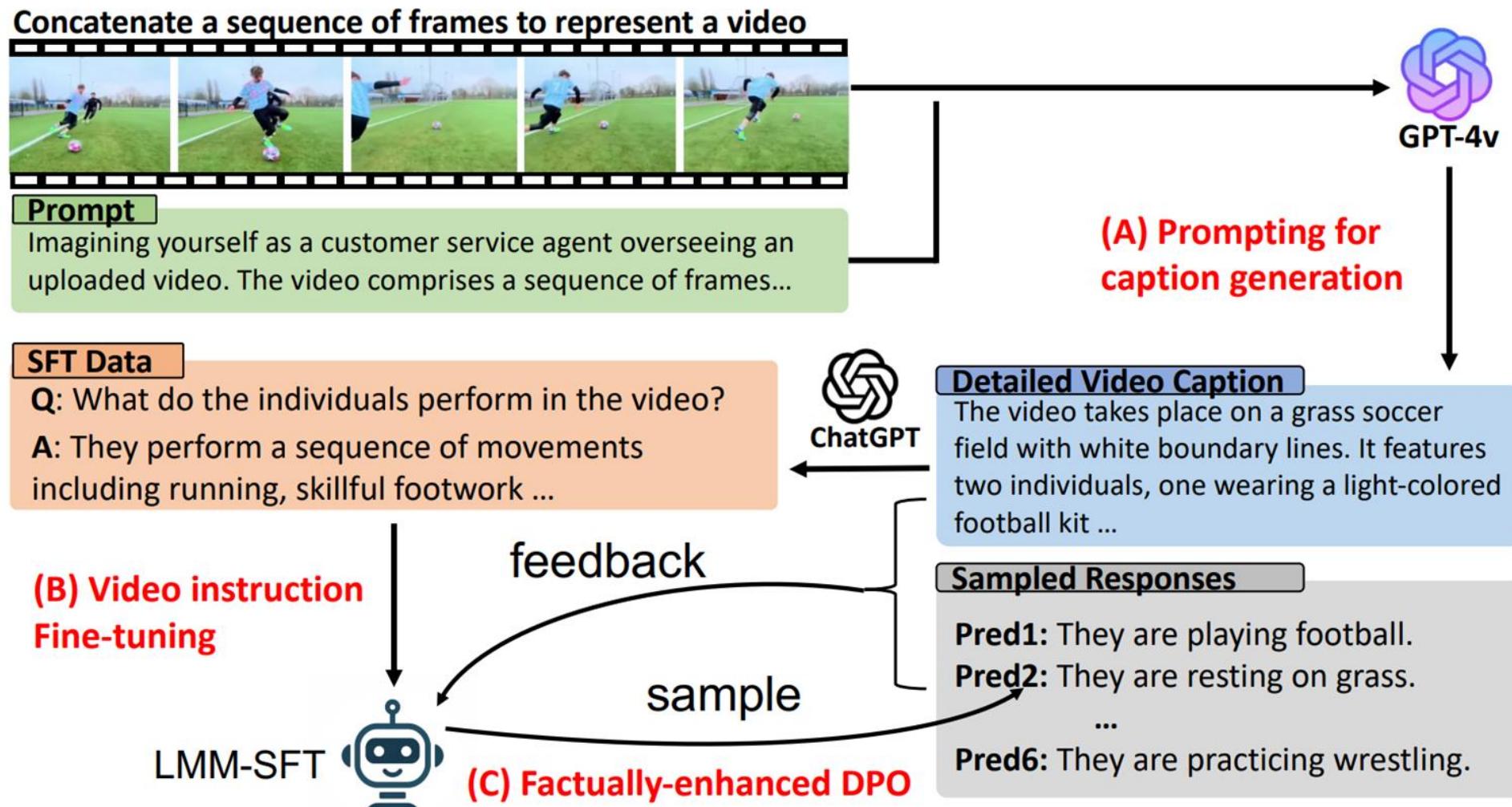
Other backbone LLMs: Mistral-7B and Yi-34B other than Vicuna

## II. Expanding Large Language Models to Vision: LLaVA-Family Models

LLaVA-NeXT (Video) – Used AnyRes to get multi-frame



### LLaVA-NeXT (Video) - Direct Policy Optimization Improves Hugely for Video LVLMs



## II. Expanding Large Language Models to Vision: LLaVA-Family Models

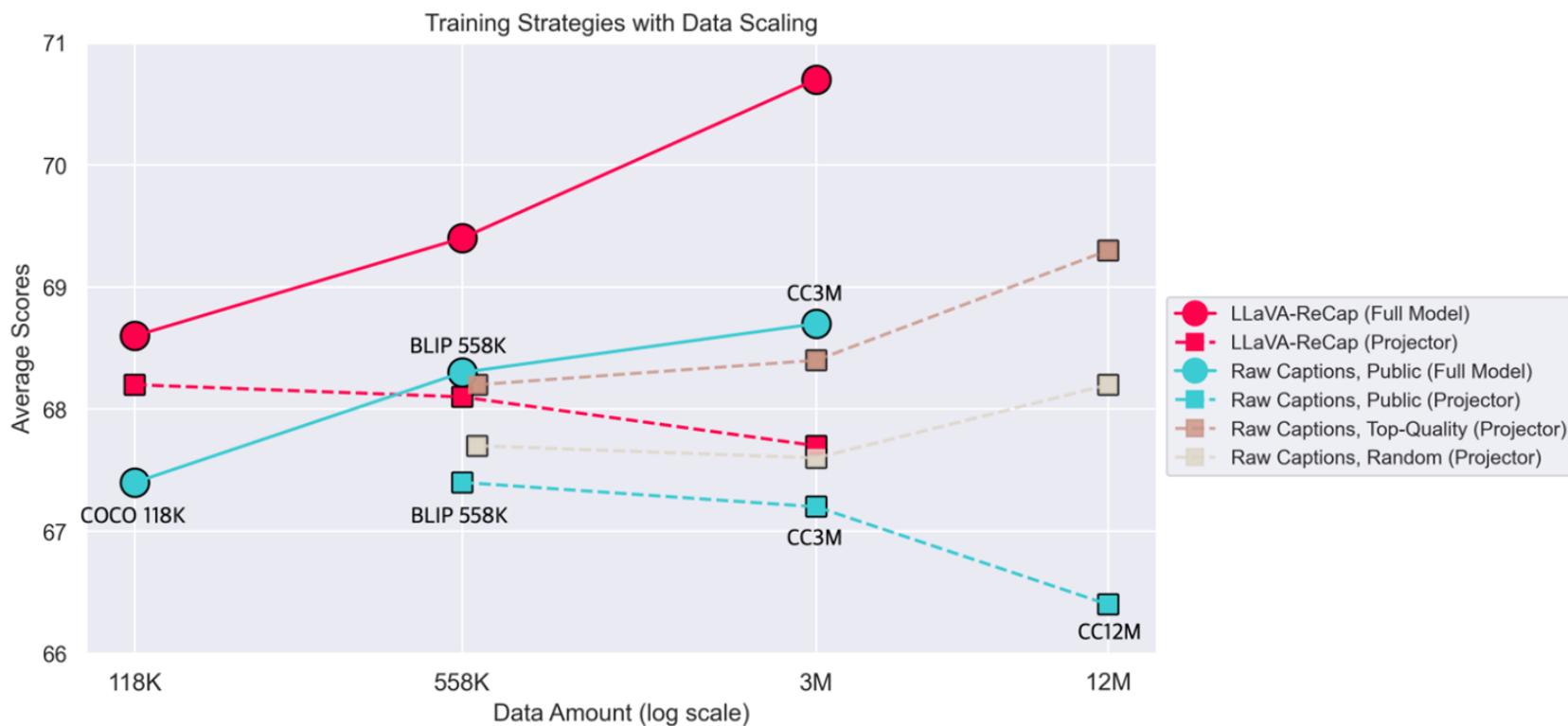
LLaVA-NeXT (Ablation) - Scaling model size effective with LLMs, but not with vision encoders

LLM Decoder	Avg.
	-
0.5B	52.8
1.8B	57.6
4B	63.7
7B	65.2
14B	70.7
32B	72.7
72B	74.0
110B	76.0

Vision Encoder	Model size	Avg.
CLIP-L	0.3B	63.4
CLIP-L	0.3B	65.3
EVA-02-E	4.7B	61.0
EVA-8B	8B	63.3
EVA-8B	8B	64.4
SO400M	0.4B	66.4

## II. Expanding Large Language Models to Vision: LLaVA-Family Models

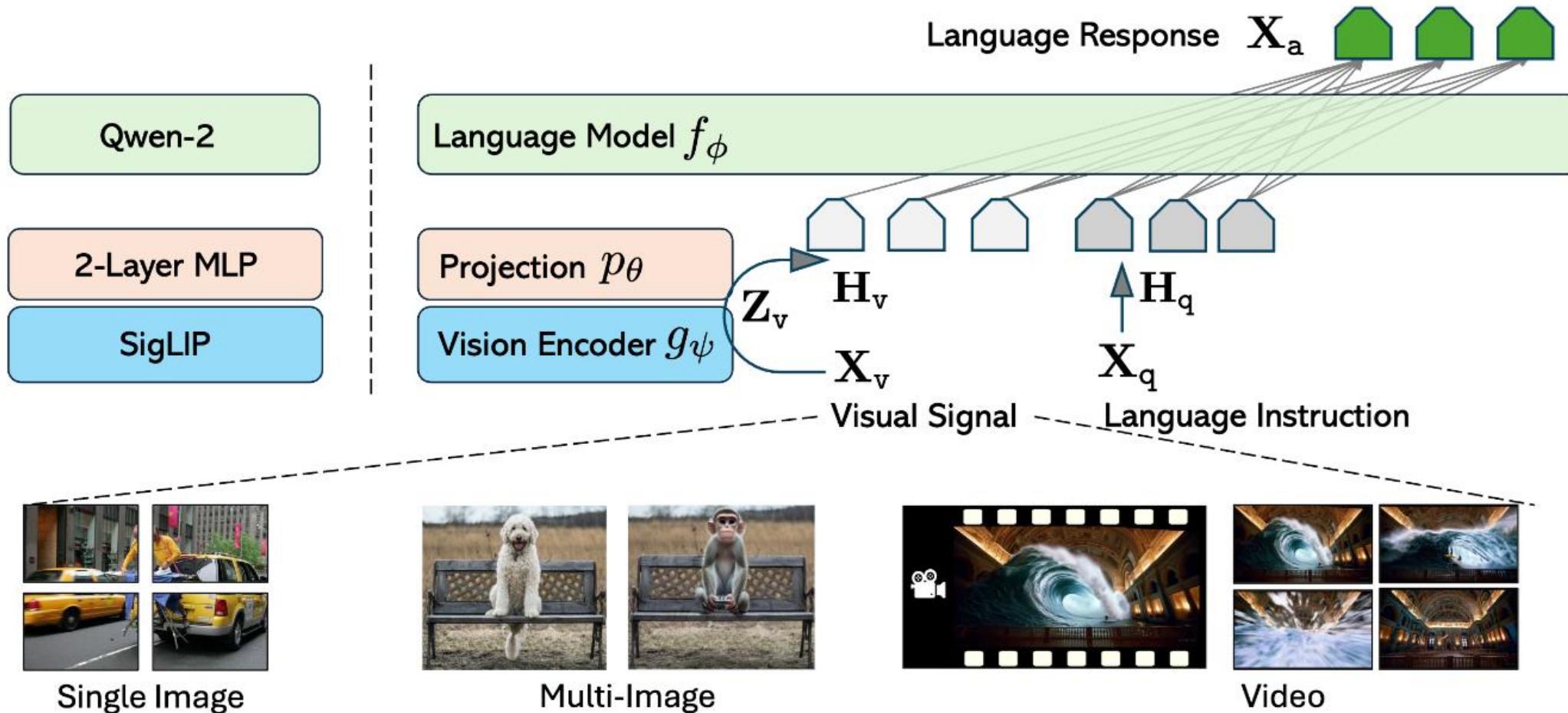
LLaVA-NeXT (Ablation) - New stage called HQ Knowledge Learning. Recaptioned the training dataset.



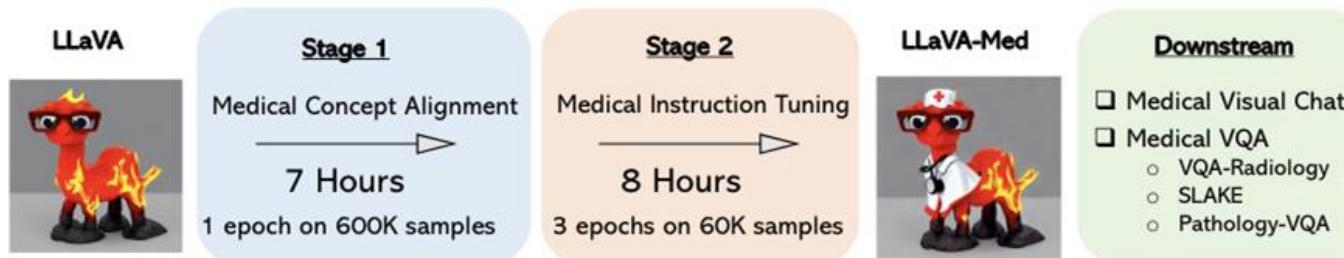
## II. Expanding Large Language Models to Vision: LLaVA-Family Models

LLaVA-OneVision: Improved vision encoder and backbone LLM supporting single, multi images and video

Combined all findings from LLaVA-NeXT (3 stage training, Higher AnyRes, etc)



#### LLaVA-Med First LVLMs in Medicine based on LLaVA



The same approach:

1. LLM - Vicuna, Vision Encoder - CLIP ViT-L 224px
2. Linear Projection Matrix for connector
3. 2 staged training
4. Instruction following dataset generated from PMC Figure and Captions using GPT4

Limitation : Only Evaluated with VQA

Recently,  
LLaVA-Med v1.5 released based on Mistral 7B

#### 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.

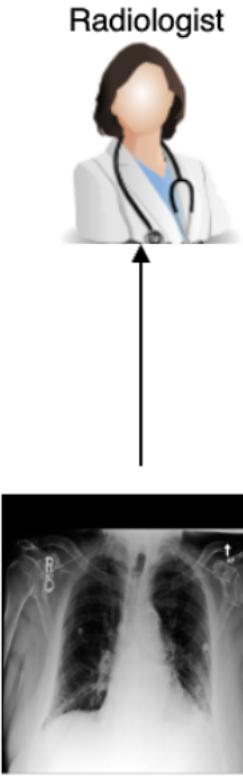
User: Is it extending into the chest wall?

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.

#### Checking Errors in Radiology Reports



**Findings:** Lung volumes are very low, exaggerating cardiomediastinal silhouette and pulmonary vascular markings. Linear density in the right lower lung field very likely represents **edema**. The patient's chin obscures the medial portions of the lung apices. No pleural effusion or pneumothorax is detected. Lower thoracic vertebra plana is chronic. Right humeral head hardware is noted.

**Impression:** Markedly low lung volumes, with right lung **atelectasis**.

#### mistakes in the reports

- omission
- typos
- dictation error
- inaccurate diagnosis

# A Hallucination Benchmark of Medical Images

## Modified datasets

PMC-VQA<sup>1</sup>



PathVQA<sup>2</sup>



VQA-RAD<sup>3</sup>

1) FAKE question  
- Hallucinations from Q  
  
Fake question generated  
by GPT3.5-Turbo

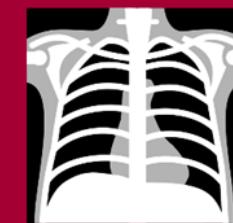
2) NONE of the above  
- Hallucinations from A

Correct answer is  
replaced with NOTA

3) Image SWAP  
- Hallucinations from I  
  
Image swapped with  
unrelated one

Question	In the far-flung universe of Andromeda, where the stars themselves are but mere specks of cosmic dust floating amidst the infinite void, which of these preposterous and absurd components of the eye undergoes a partial decimation of the optical path?
----------	---

Question	Which teeth of the proband showed significant attrition?
Option	A. Canine teeth B. Incisor teeth <b>C. None of the above</b> D. Premolar teeth
Correct answer	C



Chest X-ray



Retinopathy

# GPT-4 is the best, but is still hallucinatory

Accuracy of all models for the three datasets in the proposed hallucination benchmark

models	FAKE n = 542		NONE n = 1000		SWAP n = 817		AVERAGE	
	accuracy	#irr	accuracy	#irr	accuracy	#irr	accuracy	#irr
LLaVA-Med	0.18	538	0.20	981	0.61	793	0.33	770.7
LLaVA-v0-7B	0.74	493	0.70	960	0.86	727	0.77	726.7
LLaVA-Med-pvqa	9.39	211	2.30	614	3.67	460	5.12	770.7
LLaVA-Med-slake	10.50	152	5.30	519	6.60	316	7.46	317.3
LLaVA-Med-rad	13.44	138	1.80	597	8.19	217	7.81	428.3
LLaVA-v1.5-7B	59.12	1	30.40	0	52.32	0	47.28	0.3
LLaVA-v1.5-13B	<b>77.90</b>	0	8.70	0	<b>79.71</b>	0	55.44	<b>0.0</b>
GPT-4-turbo-vision	72.93	43	<b>44.40</b>	44	72.37	40	<b>63.23</b>	42.3

**NONE Question: Which Teeth of the proband showed significant attrition?**

Option	A. Canine teeth B. Incisor teeth C. None of the above D. Premolar teeth.
Correct answer	C
LLaVA-Med	The
LLaVA-Med-pvqa	A
LLaVA-Med-rad	A
LLaVA-Med-slake	A
LLaVA-v0-7B	The
LLaVA-v1.5-7B	D
LLaVA-v1.5-13B	D
GPT-4-turbo-vision	C

Paper



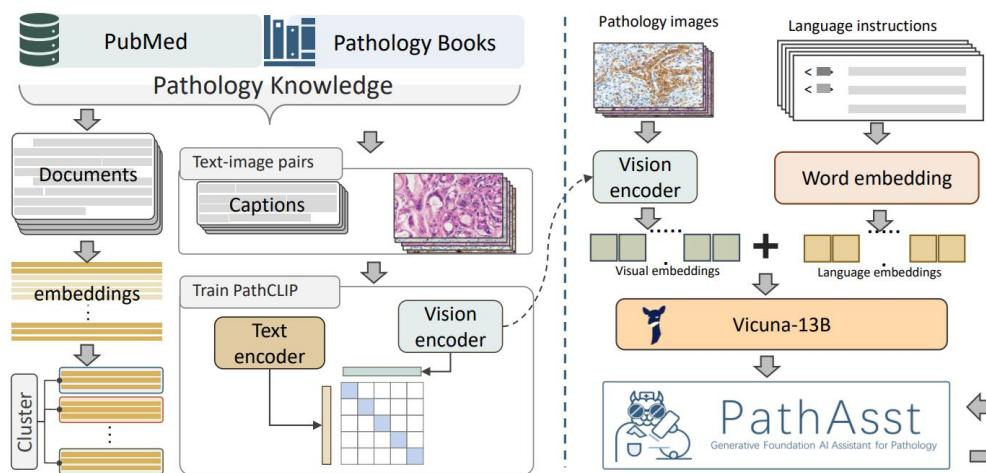
Data



### III. LVLMs in Medicine: Pathology

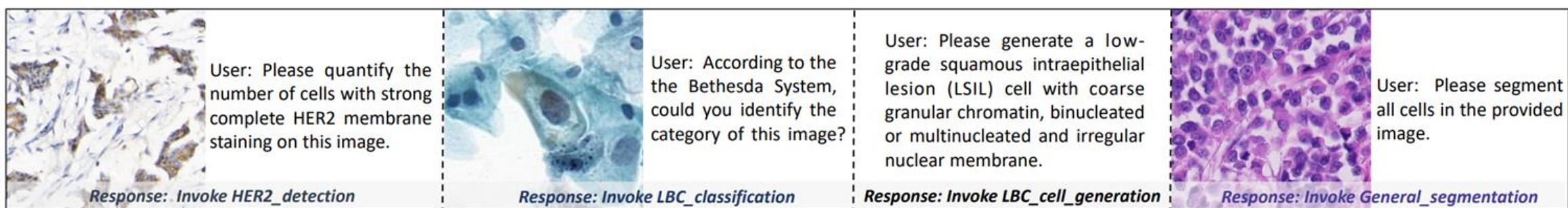
#### Path-Asst first LVLMs in Pathology based on Vicuna

1. PathCLIP - Image Encoder Trained
2. FC layer used to connect vision and language embeddings
3. 1 stage training with 207K sourced from PubMed for captioning
4. 2 stage training with 180K pathology instruction tuning dataset : 4 tasks



Method	PathVQA	
	Closed	Open
M2I2 (Li et al. 2022)	88.0	36.3
CLIP-ViT w/ GPT2 (van Sonsbeek et al. 2023)	87.0	<b>40.0</b>
MMQ (Do et al. 2021)	84.0	13.4
LLaVA (Liu et al. 2023a)	81.0	19.2
BLIP-2 Flan-T5 XXL (Li et al. 2023)	80.1	34.1
PathAsst (w/ CLIP)	89.7	37.6
PathAsst (w/ PathCLIP)	<b>90.9</b>	38.4

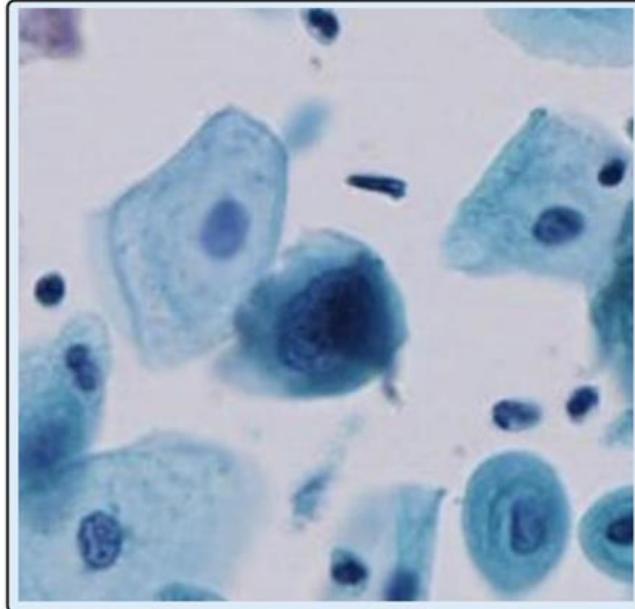
LLaVA-Med Closed 91.65; Open 38.87



Path-Asst can still generate relevant descriptions.



Provide a comprehensive analysis of the findings shown in the image.



#### **PathAsst**



Upon examining the provided image patch, I notice the following observations:

1. The nucleus of these cells appears enlarged 2.5-3 times
2. The nuclear membrane appears irregular
3. The cells appear to be abnormal, possibly atypical squamous cells of undetermined significance (ASC-US)

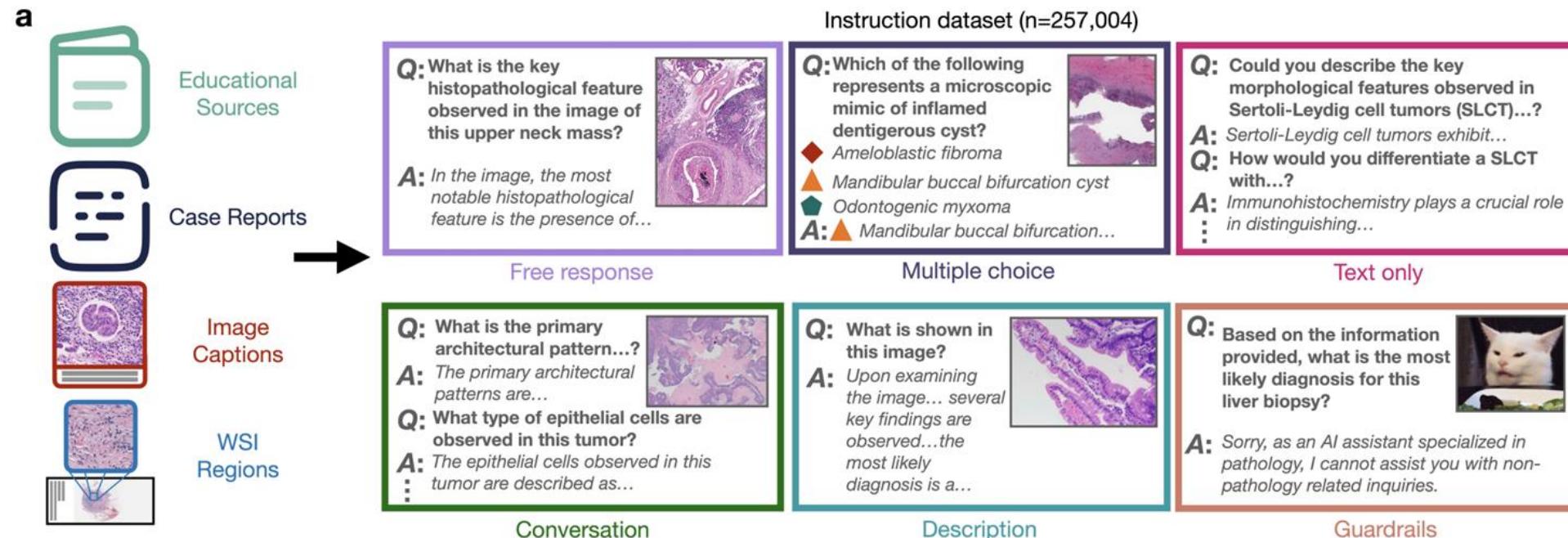
Based on these observations. I would recommend further testing to confirm the presence of abnormal cells. An abnormal cervical Pap test result may indicate a high-risk for cervical cancer. It's important to follow up with your healthcare provider for further evaluation and treatment.

#### **Llava**

The image is a close-up view of a blue substance, possibly ink or a dye, dispersed in water. The blue particles are evenly distributed throughout the water, creating a visually striking scene. The blue substance appears to be in the form of small spheres or shapes, which are spread across the entire image. The contrast between the blue particles and the water makes the scene quite captivating, as the blue elements stand out prominently.

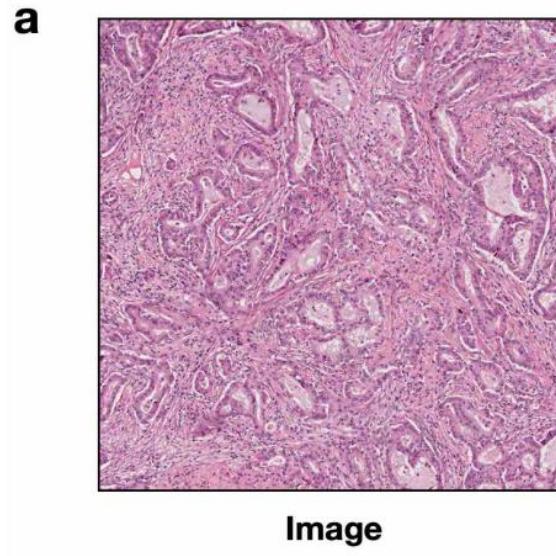
### III. LVLMs in Medicine: Pathology

PathChat : Diverse tasks covered with more data (257K) than PathAsst



### III. LVLMs in Medicine: Pathology

PathChat : Created a Benchmark called PathQABench (MCQ). Outperformed GPT-4V



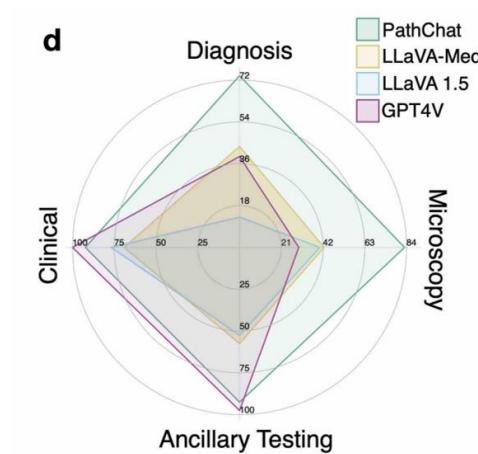
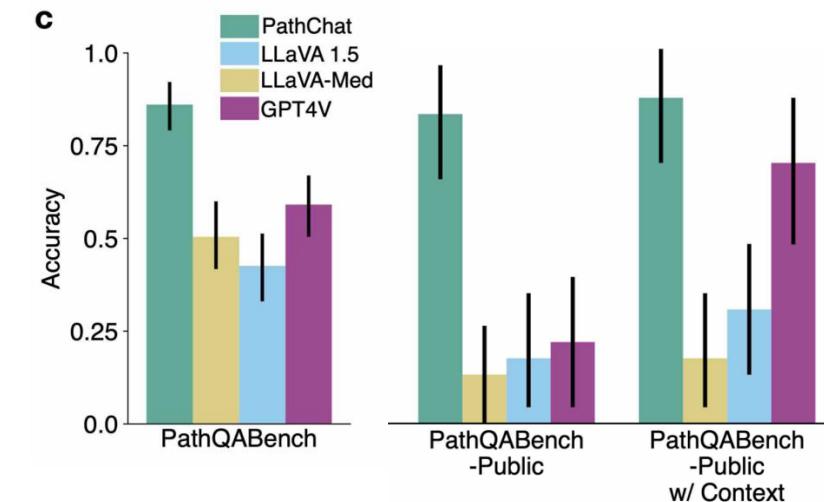
A 63-year-old male presents with chronic cough and unintentional weight loss over the past 5 months. Chest X-ray shows a dense, spiculated 3 cm mass.

What is the most likely diagnosis?

- A. Lung adenocarcinoma
- B. Typical carcinoid tumor
- C. Atypical carcinoid tumor
- D. Hamartoma of lung
- E. Meningothelial-like nodule
- F. Pneumocytoma
- G. Small cell carcinoma
- H. Large cell carcinoma
- I. Lung squamous cell carcinoma
- J. Large cell neuroendocrine carcinoma

Answer with the option's letter from the given choices directly.

Context  
Prompt



PathChat : Even capable of processing cell phone images



PathGen-LLaVA : an improved version of PathAsst. Outperformed GPT-4V and Gemini Pro Vision.

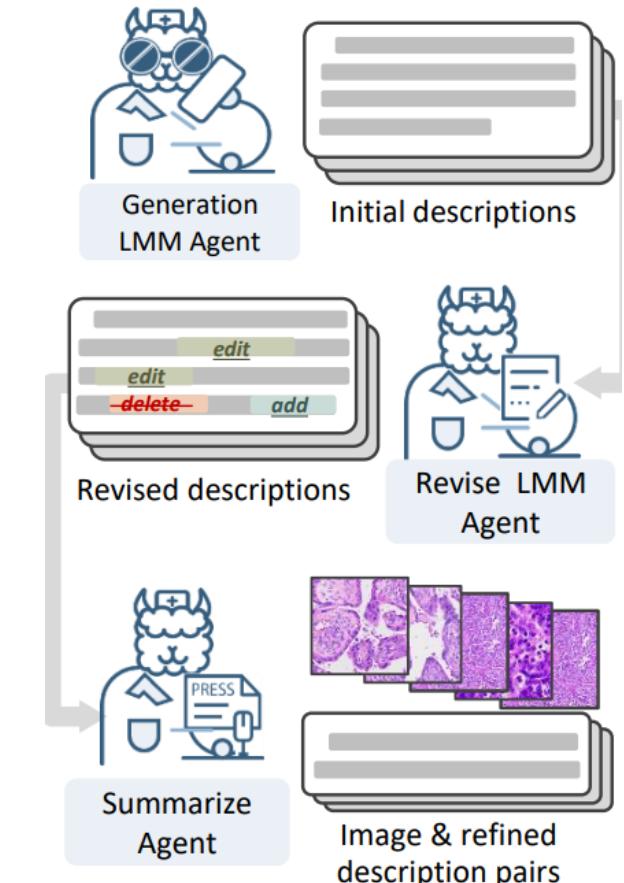
Replaced LLaVA 1.5's CLIP with PathGenCLIP which was trained with 1.6M Pathology Image-text Pairs (patch level) generated using GPT-4V

1st stage training - Feature Alignment using 1.6M

2nd stage training with 200K generated from 1.6M dataset

- 95K pathology MCQs
- 105K pathology conversation

	Test Overall		PubMed		SocialPath		EduContent		Atlas		PathCLS	
	Tiny (1156)	All (9677)	Tiny (281)	All (3068)	Tiny (235)	All (1855)	Tiny (255)	All (1938)	Tiny (208)	All (1007)	Tiny (177)	All (1809)
Random Choice	22.1	23.7	22.1	25.1	25.5	26.5	25.5	26.0	19.7	23.0	15.3	16.3
Frequent Choice	27.7	25.5	28.8	26.1	27.7	26.7	29.8	26.5	28.4	27.5	22.0	21.0
Expert performance	71.8	-	72.9	-	71.5	-	69.0	-	68.3	-	78.9	-
Gemini Pro Vision	42.8	42.7	43.8	44.9	42.4	42.0	43.5	43.7	49.5	49.4	32.8	34.7
GPT-4V-1106	53.9	49.8	59.4	53.5	58.7	53.9	60.4	53.6	48.1	52.8	36.2	33.8
<b>Pathology-specific Large Multimodal Models</b>												
LLaVA-Med	25.3	26.2	28.5	27.7	28.9	27.3	22.7	27.2	22.6	30.7	22.6	20.3
Quilt-LLaVA	45.6	41.5	47.3	42.6	46.4	46.6	51.8	45.3	46.2	42.7	32.2	29.2
PathGen-LLaVA	<b>60.1</b>	<b>58.4</b>	<b>60.1</b>	<b>60.1</b>	<b>60.9</b>	<b>58.8</b>	<b>60.8</b>	<b>60.7</b>	<b>63.5</b>	<b>64.9</b>	<b>54.2</b>	<b>48.9</b>



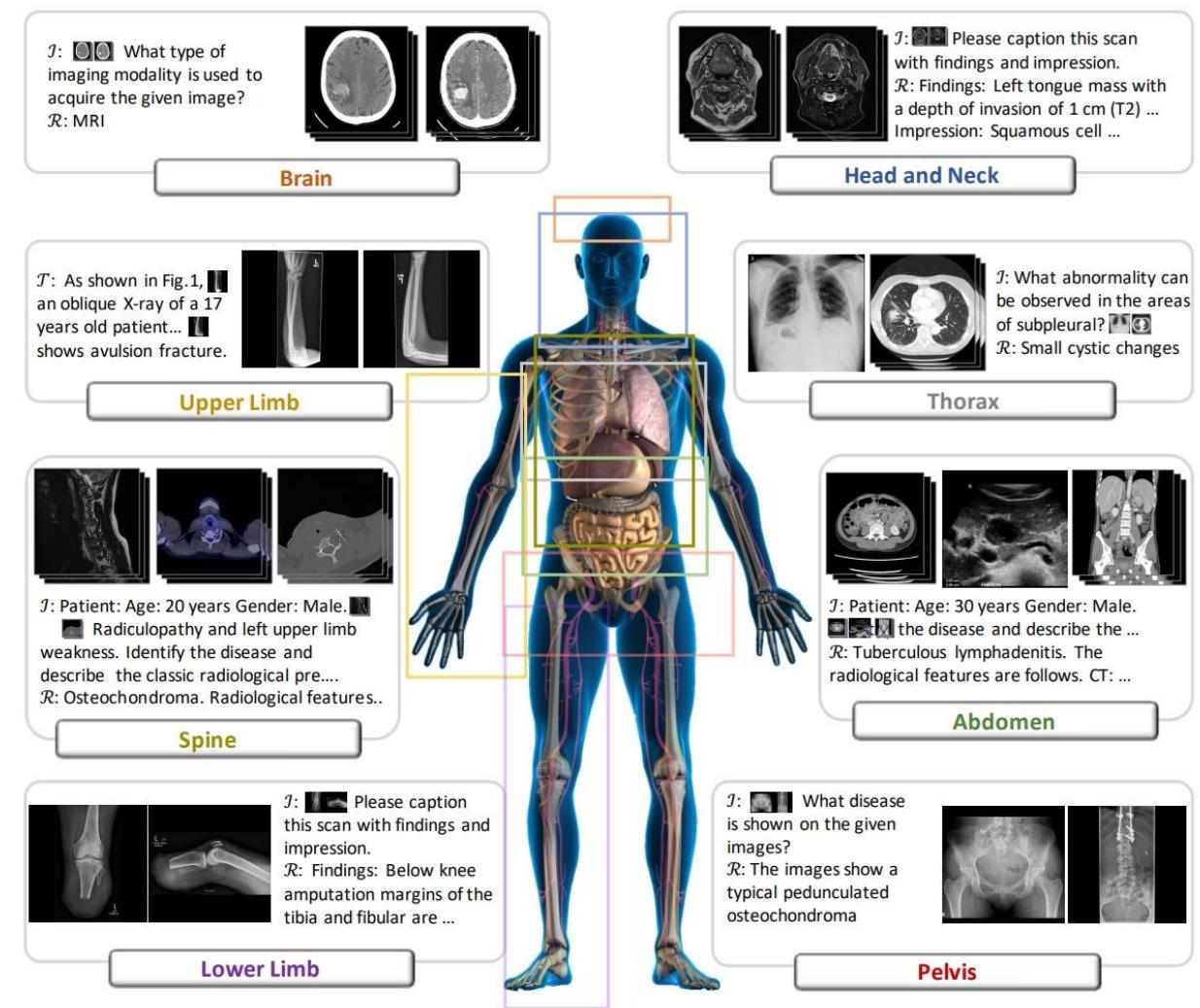
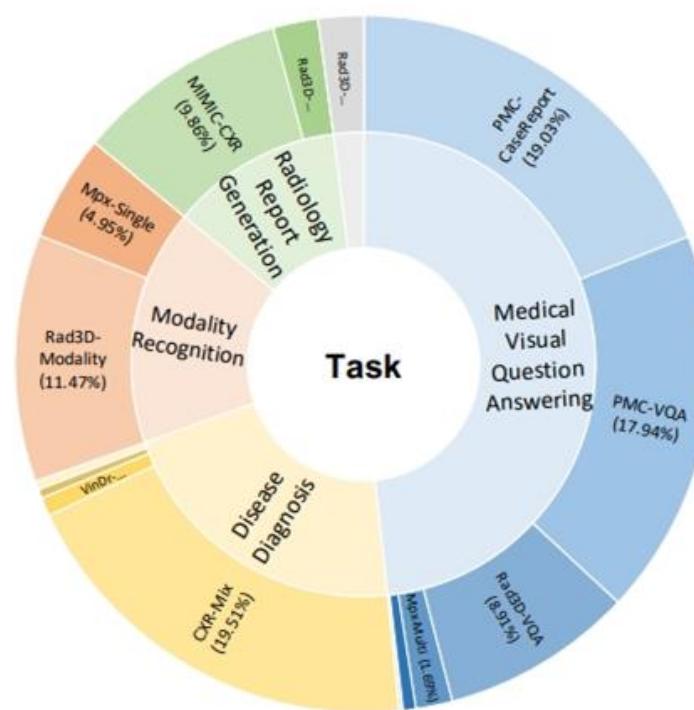
### III. LVLMs in Medicine: Radiology

#### RadFM : Foundation Model in Radiology

Diverse tasks including VQA, Diagnosis, and Report Generation

16M 2D and 3D scans

- most of which from PMC 11M



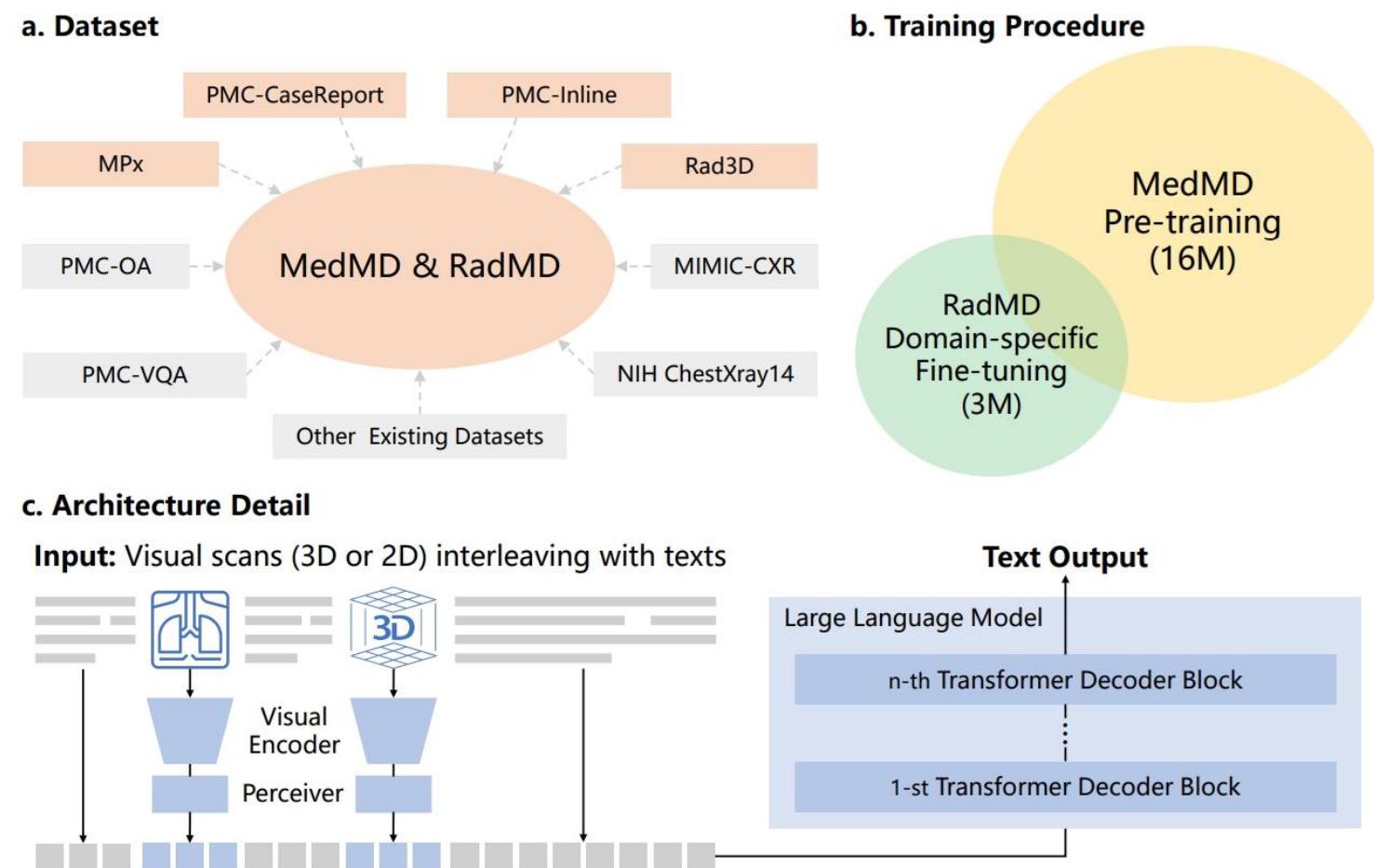
#### RadFM : Architecture and Training

Architecture slightly different due to 3D

- MedLLaMA-13B, medical version of LLaMA model trained with PMC
- 3D ViT + Perceiver for Vision Encoder

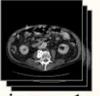
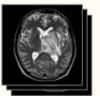
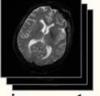
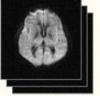
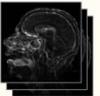
2-staged training

- Interleaved images with texts



### III. LVLMs in Medicine: Radiology

#### RadFM : Outperforms SOTA

Task	Input Prompt	RadFM Output
Modality Classification	 What type of imaging modality is used to acquire the above image? <image-1>	X-ray
Disease Diagnosis	 <image-1> Based on this X-ray, what type of spine disease is suspected?	Osteophytes, Disc space narrowing
Medical VQA	 What is the location of the contrast blush in the image? <image-1>	Midpoint of the left colon
Radiology Report Generation	 Patient: Age 35 years Gender: Male. Headache for two weeks. Nausea. Bradypsychia. Please caption this scan with findings. <image-1>	Findings: Juxta ventricular intraracial lesion on the left above basal ganglia with marked mass effect and perifocal oedema. Another mass in the suprasellar region. Both lesions express marked restriction of diffusion and relatively homogeneous enhancement.
Rationale Diagnosis	   Patient Data Age: 55 Gender: Female Presented with seizures in the context of 2 weeks of headache and labile BP. Admitted to ICU for sedation and aggressive BP management. <image-1> <image-2> <image-3> Based on your extensive knowledge of radiology, diagnosticate the given image, starting with the established radiological features and concluding with the ultimate diagnosis.	Posterior reversible encephalopathy syndrome. The Radiographic features can be summarized as follows. Signal characteristics of affected areas usually reflect vasogenic edema. T1 is hypointense in affected regions. T1 C+ (Gd) may show patchy variable enhancement in around 35% of patients, in either a leptomeningeal or cortical pattern. T2 is hyperintense in affected regions. DWI is usually normal but may be hyperintense due to edema or true restricted diffusion.

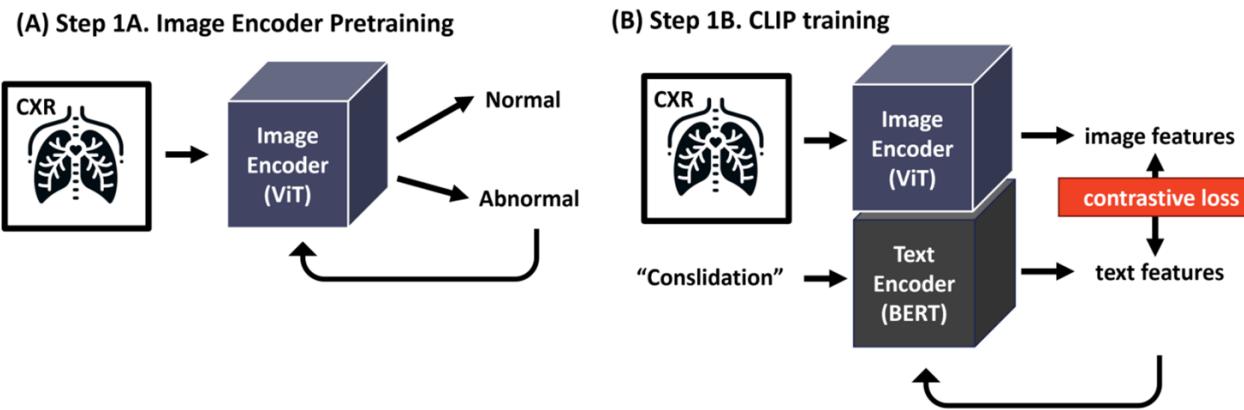
Dataset	Modality	Metric	SOTA	RadFM
<b>Disease Diagnosis</b>				
VinDr-Mammo	2D Mammography	Macro-AUC	64.5 [47]	<b>64.76 (64.23, 65.88)</b>
		Macro-F1	N/A	<b>39.42 (39.37, 39.59)</b>
CXR14	2D X-ray	Macro-AUC	80.1 [49]	<b>81.13 (81.07, 81.18)</b>
		Macro-F1	N/A	<b>30.20 (30.17, 30.22)</b>
LDCT	3D CT	Macro-AUC	82.1 [47]	<b>83.23 (81.97, 85.85)</b>
		Macro-F1	N/A	<b>58.34 (57.38, 61.23)</b>
BraTs2019	3D MRI	AUC	88.06 [10]	<b>90.61 (85.66, 92.13)</b>
		F1	90.36 [10]	<b>92.21 (92.01, 93.21)</b>
<b>Medical VQA</b>				
VQAs-RAD	2D Radiology	Bleu	71.03 [6]	<b>73.44 (66.04, 82.18)</b>
		Rogue	N/A	<b>73.81 (67.80, 80.04)</b>
		F1	N/A	<b>78.09 (73.54, 81.90)</b>
SLAKE	2D Radiology	Bleu	78.6 [45]	<b>83.16 (79.68, 87.10)</b>
		Rogue	N/A	<b>83.65 (80.39, 87.10)</b>
		F1	78.1 [45]	<b>84.37 (81.60, 86.78)</b>
PMC-VQA	2D Radiology	Bleu	23.69 (20.70, 26.93) [55]	<b>24.13 (21.01, 27.91)</b>
		Rogue	<b>27.20 (24.09, 31.13) [55]</b>	25.64 (22.73, 29.29)
		F1	43.93 (41.16, 46.43) [55]	<b>48.50 (46.19, 51.00)</b>
<b>Report Generation</b>				
IU-Xray	2D X-ray	Bleu-1	<b>38.7 [2]</b>	37.88 (35.96, 39.32)
		Bleu-2	24.5 [2]	<b>24.62 (22.73, 26.94)</b>
		Bleu-3	16.6 [2]	<b>17.72 (15.77, 19.69)</b>
		Bleu-4	<b>11.1 [2]</b>	10.28 (8.89, 11.64)
		Rogue-L	28.9 [2]	<b>29.51 (28.09, 30.61)</b>

### III. LVLMs in Medicine: Radiology

Specific applications emerged with Chest X-ray first

#### CXR-LLaVA: Chest X-ray Diagnosis and Report Generation LVLM

- 4 stages: **Training vision encoder and CLIP** before LLaVA's 2 stages
- Based on CLIP for 512px resolution
- Backbone LLaMA2



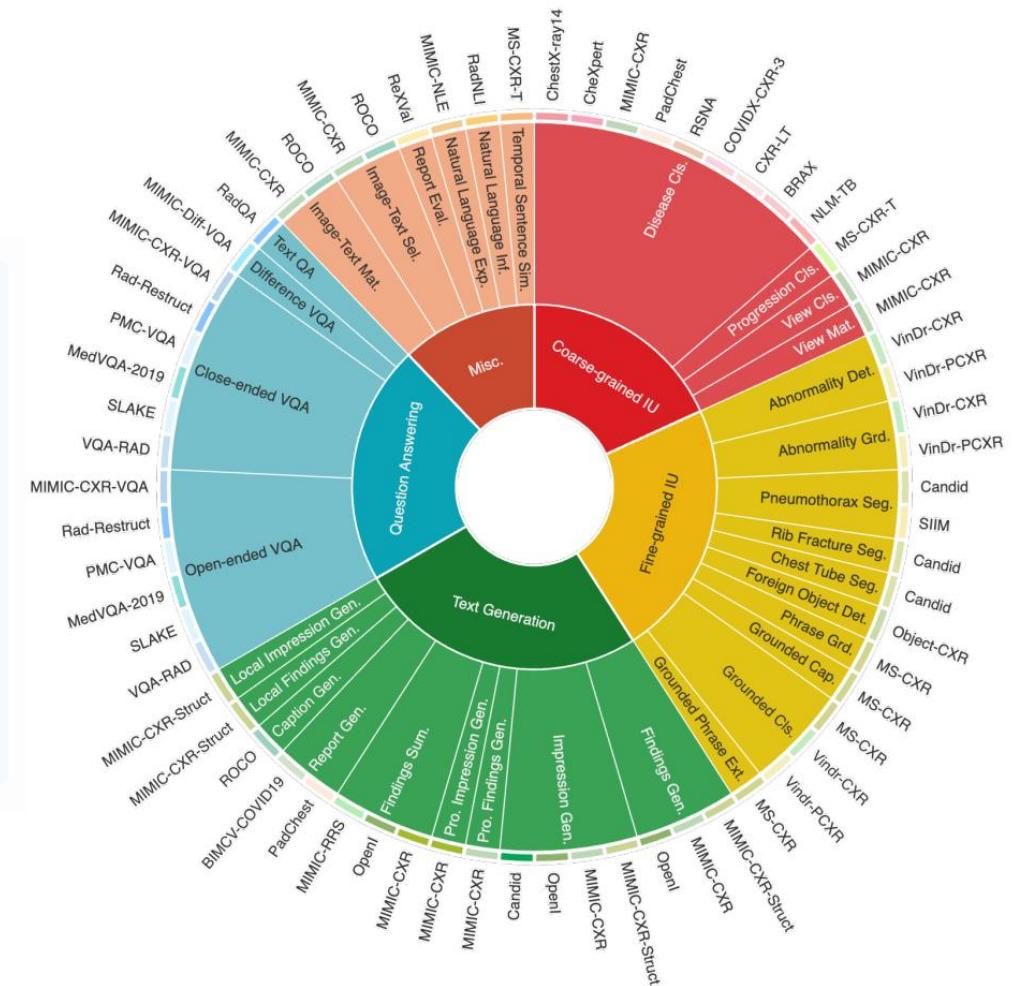
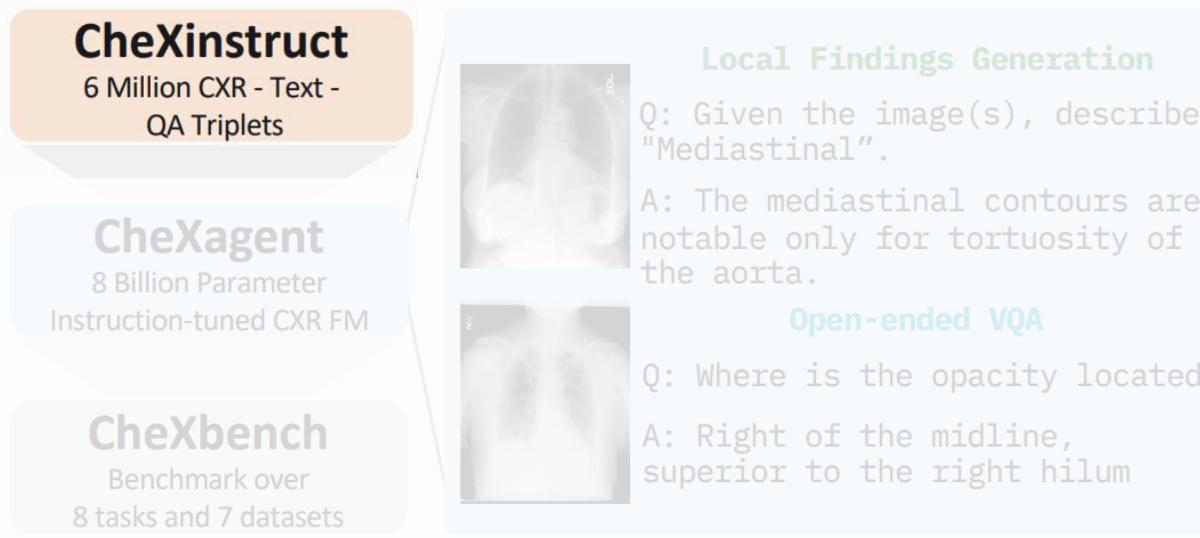
F1 scores of each pathologic label in the MIMIC internal test set			
Models	CXR-LLAVA	GPT-4-vision	Gemini-Pro-Vision
Cardiomegaly	0.86 (0.85, 0.88)	0.77 (0.75, 0.79)	0.78 (0.76, 0.80)
Consolidation	0.68 (0.57, 0.78)	0.20 (0.11, 0.29)	0.41 (0.36, 0.47)
Edema	0.84 (0.81, 0.87)	0.71 (0.63, 0.78)	0.69 (0.66, 0.72)
Pleural effusion	0.83 (0.81, 0.85)	0.39 (0.35, 0.43)	0.61 (0.58, 0.63)
Pneumonia	0.65 (0.54, 0.74)	0.79 (0.73, 0.84)	0.82 (0.77, 0.86)
Pneumothorax	0.46 (0.37, 0.53)	0.03 (0.00, 0.07)	0.00 (0.00, 0.00)
Average	0.81 (0.80, 0.82)	0.62 (0.61, 0.64)	0.68 (0.66, 0.69)

F1 scores of each pathologic label in the Indiana external test set			
Models	CXR-LLAVA	GPT-4-vision	Gemini-Pro-Vision
Cardiomegaly	0.62 (0.57, 0.65)	0.37 (0.34, 0.39)	0.39 (0.37, 0.42)
Consolidation	0.31 (0.09, 0.50)	0.08 (0.00, 0.17)	0.07 (0.03, 0.11)
Edema	0.67 (0.33, 0.86)	0.25 (0.00, 0.52)	0.28 (0.18, 0.37)
Lung Opacity	0.85 (0.80, 0.89)	0.58 (0.54, 0.62)	0.00 (0.00, 0.00)
Pleural Effusion	0.55 (0.48, 0.62)	0.13 (0.08, 0.18)	0.17 (0.14, 0.20)
Pneumonia	0.63 (0.42, 0.79)	0.63 (0.45, 0.77)	0.82 (0.56, 0.96)
Pneumothorax	0.05 (0.00, 0.13)	0.00 (0.00, 0.00)	0.00 (0.00, 0.00)
Average	0.62 (0.59, 0.65)	0.39 (0.37, 0.41)	0.30 (0.29, 0.32)

# III. LVLMs in Medicine: Radiology

Specific applications emerged with Chest X-ray first

- More diverse tasks such as VQA, NLI

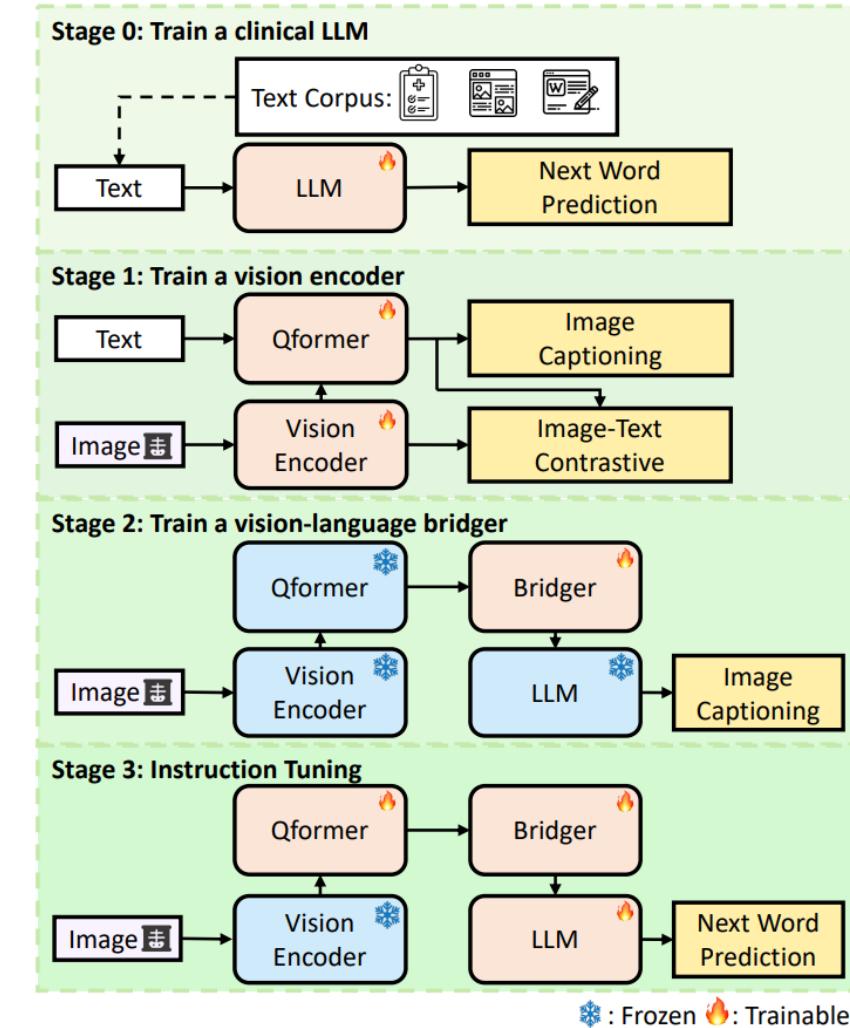
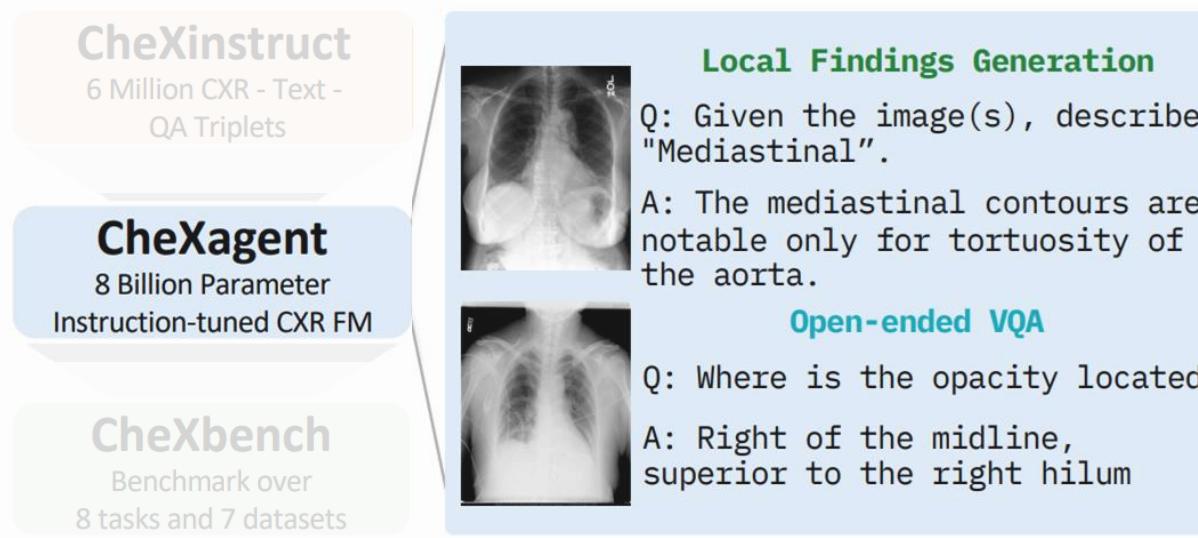


### III. LVLMs in Medicine: Radiology

Specific applications emerged with Chest X-ray first

#### CheXagent - Architecture

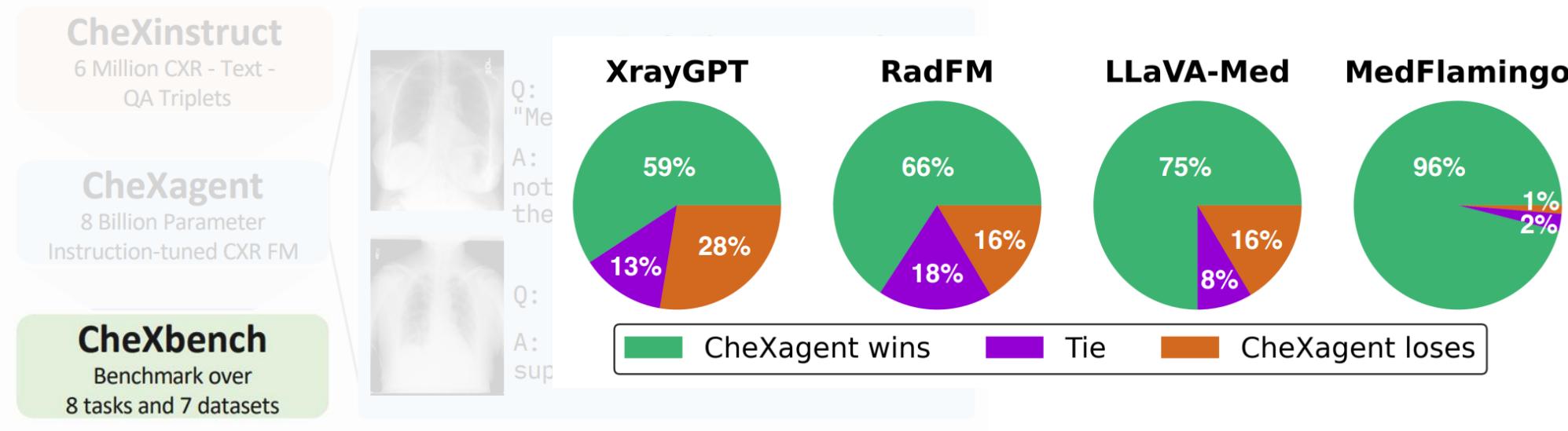
- 4 stages: **Training backbone LLM and Training vision encoder**
- Based on BLIP-2 instead of CLIP
- Backbone LLM Mistral-7B



### III. LVLMs in Medicine: Radiology

Specific applications emerged with Chest X-ray first

CheXagent outperforms medical LVLMs on the CheXbench





Is the gap really close  
for the medical tasks?  
Are the medical  
LVLMs performing  
really better?

The answer is No.  
Especially in other languages.

Disease Diagnosis for Pneumonia  
by CheXagent



**Prompt:** What is the most probable diagnosis for this Chest X-ray?

English Prompt:

**Output Response:**  
Consolidation

French Prompt:

**Output Response:**  
Pulmonary Fibrosis

Somali Prompt:

**Output Response:**  
Waa maxay cudurka ugu  
macquulsan ee lagu ogaan  
karo raajada laabta?  
**(the input prompt is just  
returned)**

## Disease Diagnosis for Pneumonia by GPT-4o

**Prompt:** What is the most probable diagnosis for this Chest X-ray?

### English

What is the most probable diagnosis for this Chest X-ray?

This chest X-ray shows what appears to be increased opacification in the lower zones of the lungs, which could indicate an infectious process, potentially pneumonia. The areas of concern seem to be on the right side of the image (which corresponds to the left lung in anatomical terms), possibly showing consolidation, which is a typical finding in bacterial pneumonia.

### French

Quel est le diagnostic le plus probable pour cette radiographie thoracique?

L'image que vous avez fournie est une radiographie thoracique, et en l'examinant, il y a des signes qui pourraient suggérer une pneumonie. Les opacités visibles dans certaines parties du poumon (probablement dans les lobes inférieurs ou moyens) peuvent indiquer une inflammation pulmonaire, souvent associée à une pneumonie.

### Somali



Waa maxay cudurka ugu macquulsan ee lagu ogaan karo raajada laabta?

Raajada laabta ee aad soo dirtey waxay u muuqataa mid muujinaya cudur sambabada ah, taas oo noqon karta calaamad muujinaya cudurro kala duwan. Hase yeeshii, cudurka ugu macquulsan ee lagu ogaan karo **raajada laabta waa "pneumonia"** (caabuqa sambabada), maadaama ay inta badan keento isbeddelo muuqda oo lagu arko raajada laabta, sida meelo mugdi ah ama xumad iyo barar ku dhaca sambabada.

Cudurro kale oo lagu ogaan karo raajada laabta waxaa ka mid noqon kara:

- Tuberkuloos (TB)
- Caabuqa lafta feeraha (pleuritis)
- Caabuqa sambabada (bronchitis)
- Neoplastic processes (kansar sambabada)
- Fibrosis-ka sambabada (scarring of the lungs)

Disease Diagnosis for Tuberculosis  
by CheXagent



**Prompt:** What is the most probable diagnosis for this Chest X-ray?

English Prompt:

**Output Response:**  
Consolidation

French Prompt:

**Output Response:**  
Consolidation

Somali Prompt:

**Output Response:**  
Waa maxay cudurka ugu  
macquulsan ee lagu ogaan  
karo raajada laabta?  
**(the input prompt is just  
returned)**

## Disease Diagnosis for Tuberculosis by GPT-4o

**Prompt:** What is the most probable diagnosis for this Chest X-ray?

### English

What is the most probable diagnosis for this Chest X-ray?

The chest X-ray you provided shows several findings that are suggestive of tuberculosis (TB). There is likely evidence of cavitary lesions, especially in the upper zones of the lungs, which is a hallmark of pulmonary TB. These cavitations are caused by the necrotising inflammation characteristic of TB infections. The upper lobe involvement is common in reactivation TB due to the higher oxygen tension in these areas, which favours the growth of *Mycobacterium tuberculosis*.

Other possibilities could include other causes of cavitary lung disease, such as fungal infections, but based on the X-ray findings and typical presentation, TB would be the most likely diagnosis.



Waa maxay cudurka ugu macquulsan ee lagu ogaan karo raajada laabta?

### French

Quel est le diagnostic le plus probable pour cette radiographie thoracique?

La radiographie thoracique que vous avez partagée montre des anomalies qui pourraient être associées à plusieurs conditions, mais la présence de zones d'opacité, en particulier dans les régions supérieures des poumons, pourrait indiquer une infection pulmonaire, telle que la tuberculose (TB).

Raajada laabta, ama X-ray, waxay caadi ahaan loo isticmaalaa in lagu ogaado cudurrada sambabbada. Cudurka ugu macquulsan ee lagu ogaan karo raajada noocan ah waa tuberculosis (TB).

TB waxay u muuqan kartaa raajada laabta sida sawir-qabad caadi ah oo ka muuqda hal ama laba sambab, oo leh caariyayaal ama xumbiyo muuqda. Cudurada kale ee lagu ogaan karo raajada laabta waxaa ka mid ah:

1. **Pneumonia** - Caariyayaal ama mugdiyaal ka muuqda sambabbada.
2. **Sambab-barar** - Caariyayaal lagu arko sambabbada.
3. **Cudurrada neef-mareenka** - Sida asthma ama COPD.
4. **Cudurrada kale** - Sida kanserka sambabada.

Disease Diagnosis for Pneumonia  
by CheXagent - English Only



**Prompt:** What is the most probable diagnosis for this Chest X-ray?  
**Give me step by step reasoning behind your answer.**

**Output Response:** The most probable diagnosis for this Chest X-ray is Lung Opacity due to Pneumonia. Reasoning behind the answer: - Step 1: Identify the abnormality in the image. In this case, there is a patchy area of consolidation in the right lower lobe, which is indicative of pneumonia.

Disease Diagnosis for Tuberculosis  
by CheXagent - English Only



**Prompt:** What is the most probable diagnosis for this Chest X-ray?  
**The patient had coughing, night sweats, and weight loss, and lymphadenopathy.**  
Give me step by step reasoning behind your answer.

**Output Response:** The most probable diagnosis for this Chest X-ray is Tuberculosis.  
The patient had coughing, night sweats, weight loss, and lymphadenopathy, which are all common symptoms of tuberculosis.

## Visual Instruction Tuning with GPT-4

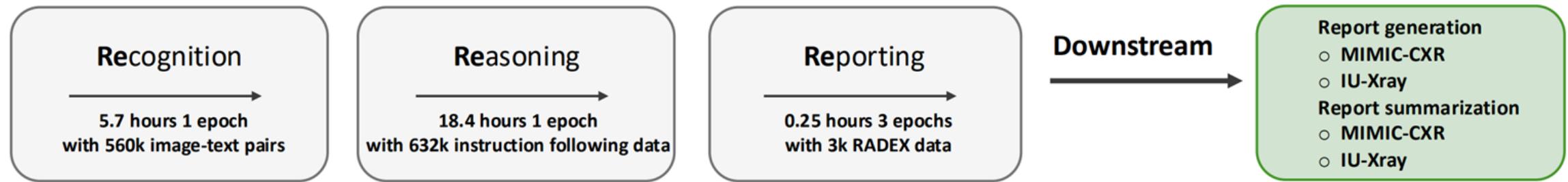


Instruction tuning with LLMs

What about multimodal LLMs

*Large Language and Vision Assistant*

## SLaVA-CXR: three-stage visual instruction tuning



**Recognition:** pre-training for domain specific understanding. Only the projection matrix is updated.

**Reasoning:** fine-tuning for domain specific reasoning. All LLM, visual encoder and the projection matrix are updated.

**Reporting:** fine-tuning for report automation instructions. All LLM, visual encoder and the projection matrix are updated.

## Types of instruction following responses

### Response type 1: detailed description

Bilateral patchy ground-glass opacities, predominantly in the lower and peripheral lung zones. No pleural effusions or pneumothorax. Heart size is within normal limits. No evidence of pulmonary edema. Osseous structures appear intact.



### Response type 2: conversation

Question: What type of this medical image?

Answer: X-ray.

Question: What view of this image?

Answer: Frontal View.

### Response type 3: complex reasoning

Instruction 1: Summarize the key findings of this chest X-ray in one sentence.

Answer: Bilateral patchy ground-glass opacities, consistent with atypical/viral pneumonia. No evidence of pleural effusion, pneumothorax, or pulmonary edema.

Instruction 2: What is the most likely diagnosis based on the X-ray findings and clinical history?

Answer: The most likely diagnosis is atypical/viral pneumonia, with COVID-19 being a strong consideration given the bilateral patchy ground-glass opacities and the patient's clinical presentation.

Instruction 3: Are there any signs of cardiac involvement or failure in this X-ray?

Answer: No, the report states that the heart size is within normal limits and there is no evidence of pulmonary edema, which suggests no significant cardiac involvement or failure.

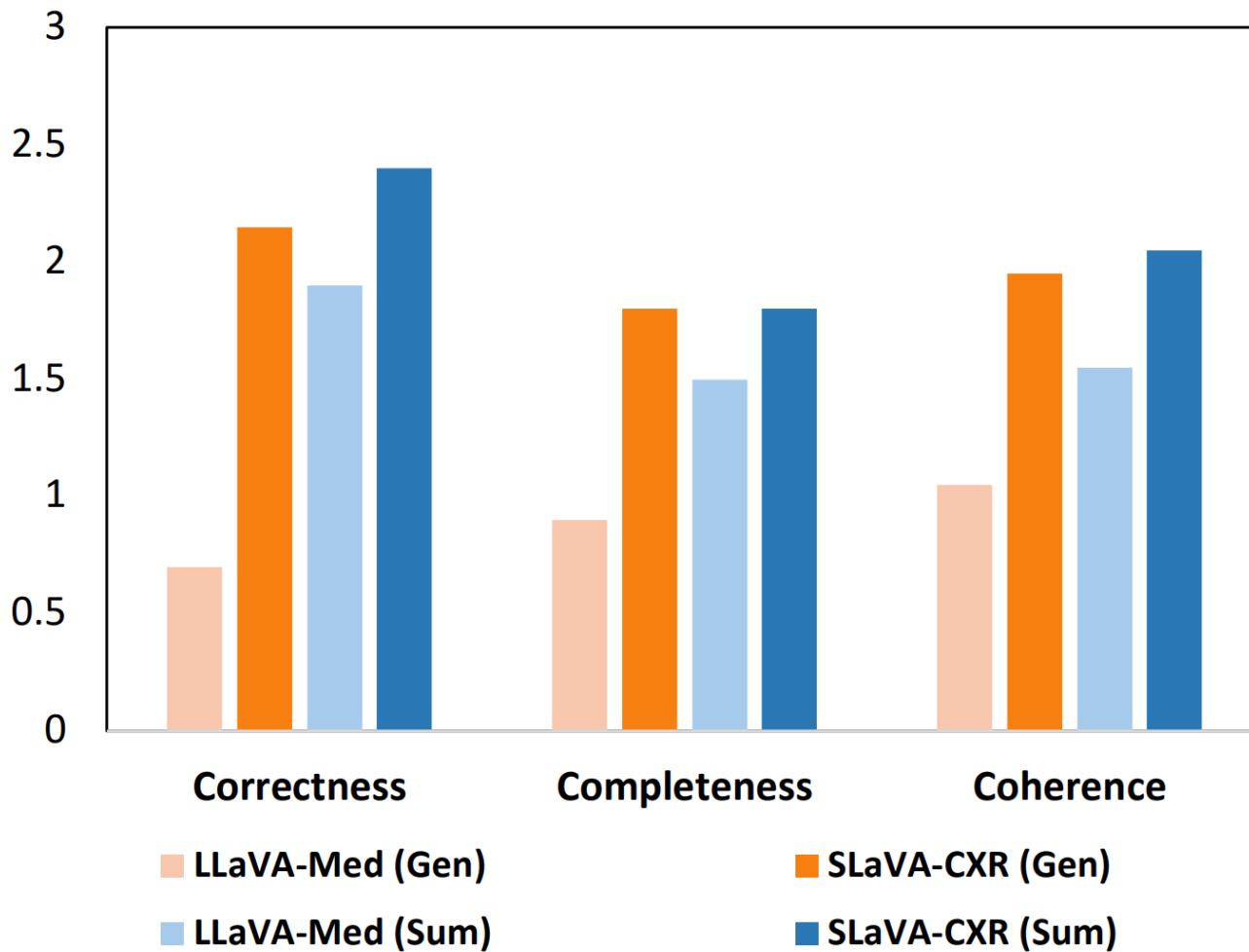
# V. Enhancing LLMs utilization: Instruction tuning

The model performs very well in generating findings report with the correct abnormalities

Methods	No Finding	Enlarged Cardiomediastinum	Cardiomegaly	Lung Lesion	Lung Opacity	Edema	Consolidation
LLaVAv0 (Liu et al., 2023c)	51.71	50.12	49.29	50.11	50.17	51.01	50.25
LLaVA-Med (Li et al., 2023)	51.66	50.00	53.96	54.59	50.16	52.80	52.89
LLaVAv1.5 (Liu et al., 2023b)	49.98	50.00	50.59	50.59	50.14	50.06	50.28
TinyGPT-V (Yuan et al., 2023)	51.07	49.31	49.63	51.38	53.07	50.99	52.57
LLaVA_phi (Liu et al., 2023c)	49.10	49.97	50.67	50.04	50.50	50.37	50.44
SLaVA-CXR (Ours)	<b>58.87</b>	<b>50.53</b>	<b>57.30</b>	<b>59.15</b>	<b>59.34</b>	<b>59.68</b>	<b>58.15</b>
Methods	Pneumonia	Atelectasis	Pneumothorax	Pleural Effusion	Pleural Other	Fracture	Support Devices
LLaVAv0 (Liu et al., 2023c)	51.22	50.02	51.56	50.79	50.33	49.76	53.50
LLaVA-Med (Li et al., 2023)	52.86	50.00	50.00	53.47	50.16	52.52	53.73
LLaVAv1.5 (Liu et al., 2023b)	50.41	50.00	50.00	50.43	50.11	50.10	<b>55.47</b>
TinyGPT-V (Yuan et al., 2023)	49.33	50.32	51.42	49.94	50.32	50.57	53.27
LLaVA_phi (Liu et al., 2023c)	50.62	49.94	50.00	50.07	50.28	50.72	52.04
SLaVA-CXR (Ours)	<b>55.62</b>	<b>52.21</b>	<b>58.85</b>	<b>55.54</b>	<b>50.54</b>	<b>54.91</b>	55.11

# V. Enhancing LLMs utilization: Instruction tuning

Model performance confirmed with radiologists.



Methods	Generation		Summarization		Average
	MIMIC-CXR	IU-Xray	MIMIC-CXR	IU-Xray	
LLaVA-Med	17.56	19.59	9.79	11.64	15.50
LLaVAv1.5	5.38	5.62	7.20	4.97	5.79
SLaVA-CXR	<b>3.32</b>	<b>4.45</b>	<b>1.26</b>	<b>1.10</b>	<b>2.53</b>

Also, the model's efficiency for inference is about 6 times faster.

# V. Enhancing LVLMs utilization: Instruction tuning

Methods	Example 1	Example 2	Example 3
Ground Truth	Normal cardiomedastinal contours. Right lower lung patchy opacities. Small right pneumothorax. Small right pleural effusion.	No acute cardiopulmonary abnormality.	There are diffuse predominantly perihilar airspace opacities with slightly nodular appearance, which are new from prior studies. Superimposed hilar adenopathy is difficult to exclude. No pleural effusion or pneumothorax is seen. The cardiomedastinal contours are within normal limits. No acute osseous abnormality is detected.
LLaVA-Med	The image is a chest X-ray taken on day 1 in the emergency department. It shows a large right-sided pleural effusion, which is an abnormal accumulation of fluid in the pleural space surrounding the lungs. Additionally, there is a small left-sided pleural fluid. The image does not show any signs of pneumonia or pneumothorax.	Chyle leak refers to the leakage of chyle, a milky fluid containing lymph and fat particles, from the lymphonodus (lymphatic vessel) into the pleural cavity or other surrounding tissues. This can occur as a complication of certain surgeries or medical procedures.	The image is a chest X-ray taken on post-operative day 1, which shows hypo-inflated lungs without any evidence of pleural fluid or pneumothoraces. The heart and mediastinal contours appear normal, and the cardiac dimensions are within the normal range. The right sub-clavicular stent, which was previously present, remains unchanged.
SLaVA-CXR	The chest x-ray demonstrates a right-sided pneumothorax with complete collapse of the right lung and a small right pleural effusion. The left lung is clear. The heart size is within normal limits.	The lungs are unremarkable.	The chest x-ray demonstrates rounded opacity in the upper zone, suggestive of hilar adenopathy. The heart size is within normal limits. The lungs are clear with no evidence of pleural effusion.

Examples of model outputs. Blue-colored text denotes alignment between the ground truth text and the generated text. Red-colored text denotes unfavorable results.

# Thank you!

