CSCE 689 - Special Topics in NLP for Science

Lecture 8: Scientific Vison-Language Models (Bioimaging)

Yu Zhang

yuzhang@tamu.edu

February 11, 2025

Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html

Submit Pre-Lecture Questions via Google Form

- https://docs.google.com/forms/d/e/1FAIpQLSdKAGdPP41dsKXyIloWJCCFXWaNqobX-u4DL7b5IIw2Yy2OBw/viewform?usp=dialog
- The next three consecutive lectures will all be given by students or guests.

	2/13	Scientific VLMs: Geometry	* UniMath: A Foundational and Multimodal Mathematical Reasoner [EMNLP 2023] * G-LLaVA: Solving Geometric Problem with Multi-Modal Large Language Model [arXiv 2023] * Math-LLaVA: Bootstrapping Mathematical Reasoning for Multimodal Large Language Models [EMNLP 2024]		Shuo
W6	2/18	[Guest Lecture] Hanwen Xu (University of Washington): Towards Patient Level Representations for Better Clinical Outcome * Suggested Reading: A Whole-Slide Foundation Model for Digital Pathology from Real-World Data [Nature 2024]			Guest Lecturer
	2/20	Scientific VLMs: Miscellaneous	* UrbanCLIP: Learning Text-Enhanced Urban Region Profiling with Contrastive Language-Image Pretraining from the Web [WWW 2024] * BioCLIP: A Vision Foundation Model for the Tree of Life [CVPR 2024] * MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI [CVPR 2024]		Hasnat

Guest Lecture (Next Tuesday)

- The guest speaker will give the lecture via Zoom.
- You may attend the guest lecture either in person or online.

- If you choose to attend in person:
  - I will be in the classroom and project Zoom onto the screen.

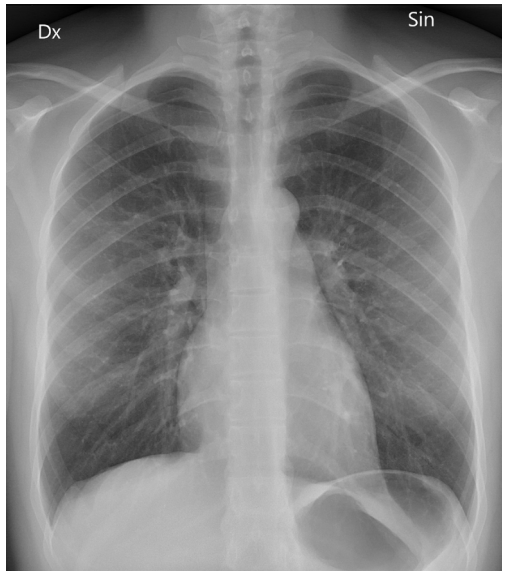
- If you choose to attend online:
  - Remember the Zoom link: https://tamu.zoom.us/j/6411788612 (you can find it on the course website)
  - I will take a screenshot at a random time during the lecture to take attendance.
  - For other (student/instructor) lectures, you still need to attend in person.

Project Proposal

- Due 2/23
- A team of 2 or 3 people
- The proposal should be 2-3 pages (ACL 2024 template, excluding references).
  - https://www.overleaf.com/latex/templates/acl-2023-proceedings-template/qjdgcrdwcnwp
  - List the team members, motivation, task definition, potential datasets, baselines, and expected outcomes of the project

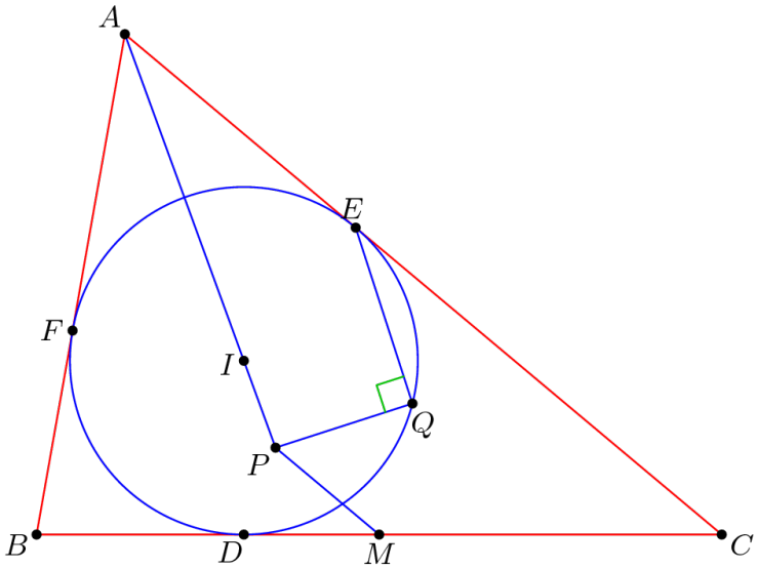
- If you do not know what to work on, come to my office hour this or next Thursday for a discussion.
- If you do not know whom to work with, drop me an email by next Tuesday. I will group all the students who email me.

Scientific Images Are Usually Accompanied by Text!



Chest X-Ray

A normal posteroanterior (PA) chest radiograph of someone without any signs of injury …



Geometry

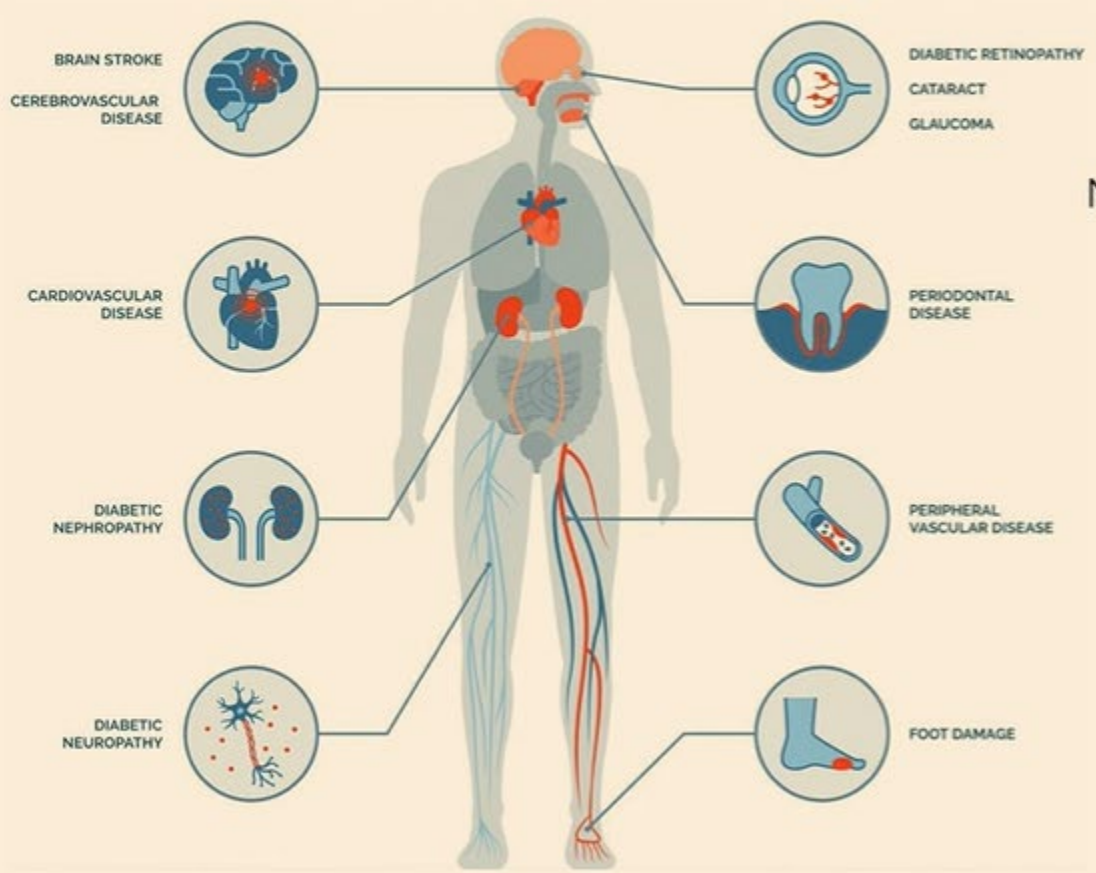
Let ABC be a triangle with incenter I whose incircle is tangent to BC, CA, AB at …

Aerial View

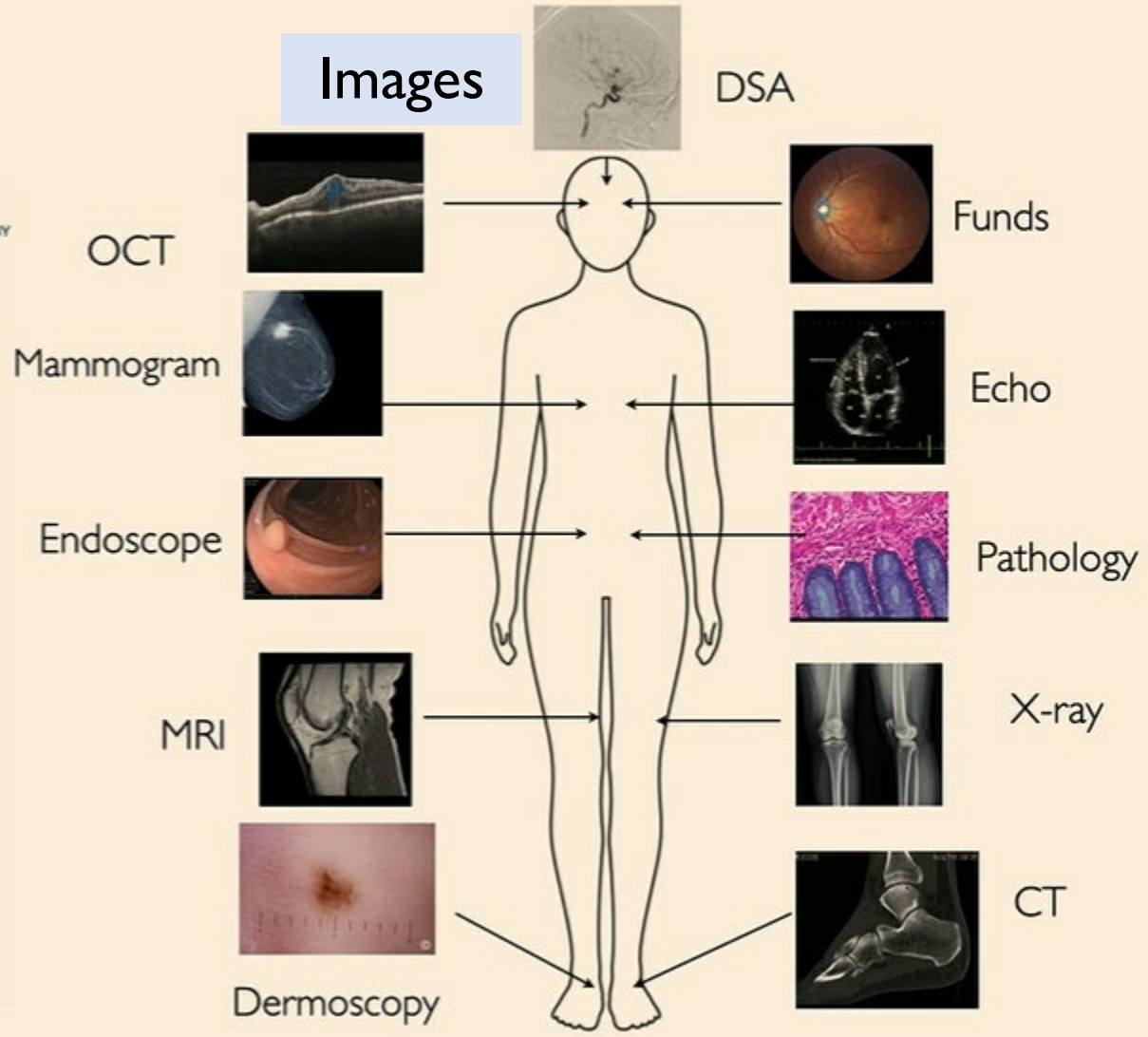
A satellite view of a large city with a mix of residential and commercial …

Biomedical visual and textual signals can mutually benefit each other.

Papers, Clinical Reports, …



Images



Agenda

- Encoder-Only
  - MedCLIP: Contrastive Learning from Unpaired Medical Images and Text
  - PLIP: Harvesting Image-Text Pairs from Twitter
- Decoder-Only
  - LLaVA-Med: Visual Instruction Tuning
- Encoder-Decoder
  - BiomedGPT: Diverse Modalities and Tasks

Agenda

- Encoder-Only
  - MedCLIP: Contrastive Learning from Unpaired Medical Images and Text
  - PLIP: Harvesting Image-Text Pairs from Twitter
- Decoder-Only
  - LLaVA-Med: Visual Instruction Tuning
- Encoder-Decoder
  - BiomedGPT: Diverse Modalities and Tasks

Assume we have a collection of image-text pairs, …

- MIMIC-CXR
  - Image: Radiographs
  - Text: Radiology reports

```
EXAMINATION:  CHEST (PA AND LAT)

INDICATION:  ___ year old woman with ?pleural effusion // ?pleural effusion

TECHNIQUE:  Chest PA and lateral

COMPARISON:  ___

FINDINGS:

Cardiac size cannot be evaluated. Large left pleural effusion is new.  Small
right effusion is new.  The upper lungs are clear.  Right lower lobe opacities
are better seen in prior CT.  There is no pneumothorax.  There are mild
degenerative changes in the thoracic spine

IMPRESSION:

Large left pleural effusion
```

https://physionet.org/content/mimic-cxr/2.0.0

Database | Credentialed Access

MIMIC-CXR Database

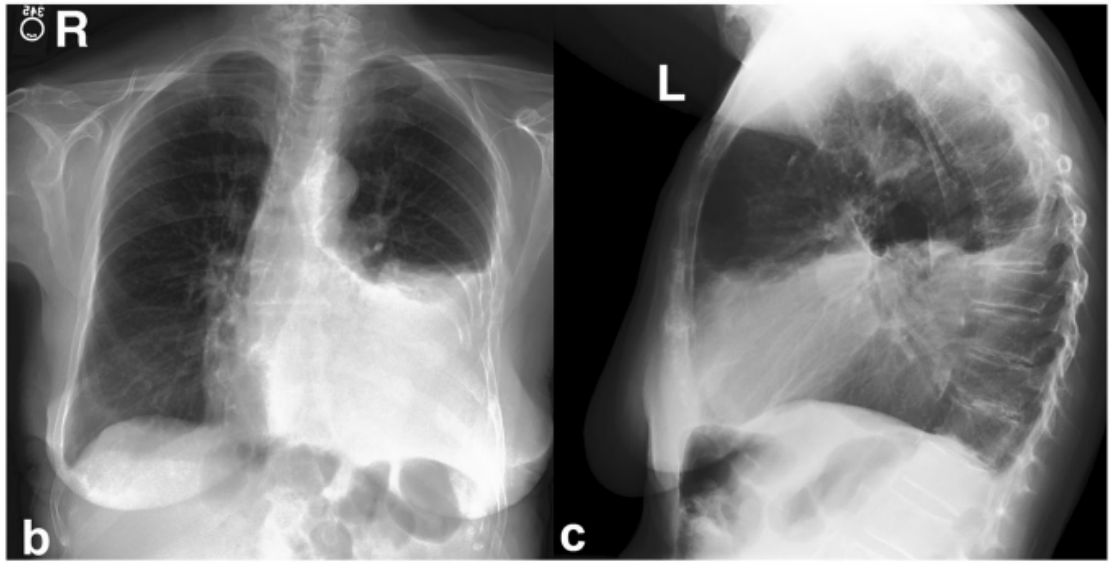
Alistair Johnson, Tom Pollard, Roger Mark, Seth Berkowitz, Steven Horng

Published: Sept. 19, 2019. Version: 2.0.0 <View latest version>

This is **not** the latest version. Click here for the latest version. ×

MIMIC-CXR paper published! *(Feb. 10, 2020, 4:06 p.m.)*

A journal article describing the MIMIC-CXR database was recently published in *Scientific Data*. The article provides detail regarding the collection, curation, and processing done in order to create the database. The article is open access and available online [1].



MIMIC-CXR, a De-Identified Publicly Available Database of Chest Radiographs with Free-Text Reports. Scientific Data 2019.

Assume we have a collection of image-text pairs, …

- ROCO [1] and MedICaT [2]
  - Images: Figures in full-text biomedical papers (PMC/S2ORC)
  - Text: Figure captions

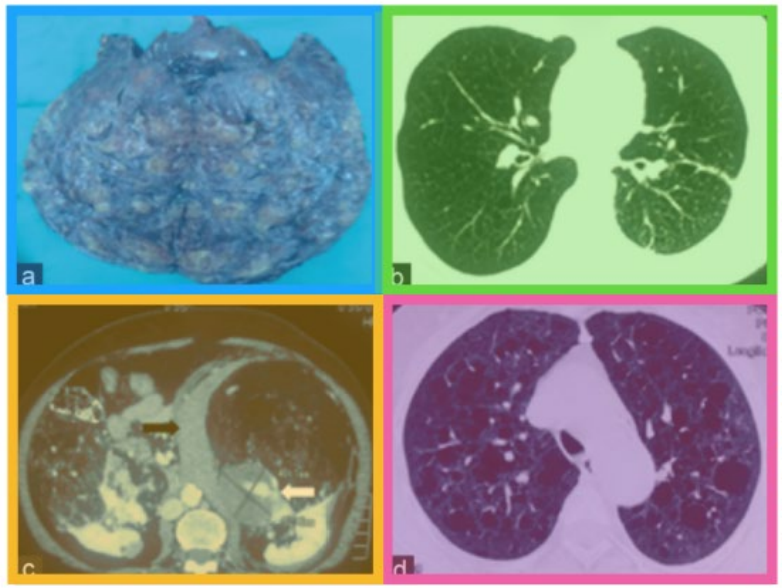


Figure 1: (a) Right renal angiomyolipoma (gross specimen postexcision). **(b)** High-resolution computed tomography chest images of Case 1 showing multiple variable sized cysts uniformly scattered in both lungs. **(c)** Computed tomography abdomen showing bilateral renal angiomyolipomas with fat densities, tortuous vessels, and pseudoaneurysm (white arrow). There is also the presence of perinephric hematoma (black arrow). **(d)** High-resolution computed tomography image of Case 2 showing bilateral lung cysts.

https://github.com/razorx89/roco-dataset

README

Radiology Objects in COntext (ROCO): A Multimodal Image Dataset

This repository contains the *Radiology Objects in COntext (ROCO)* dataset, a large-scale medical and multimodal imaging dataset. The listed images are from

https://github.com/allenai/medicat

README Apache-2.0 license

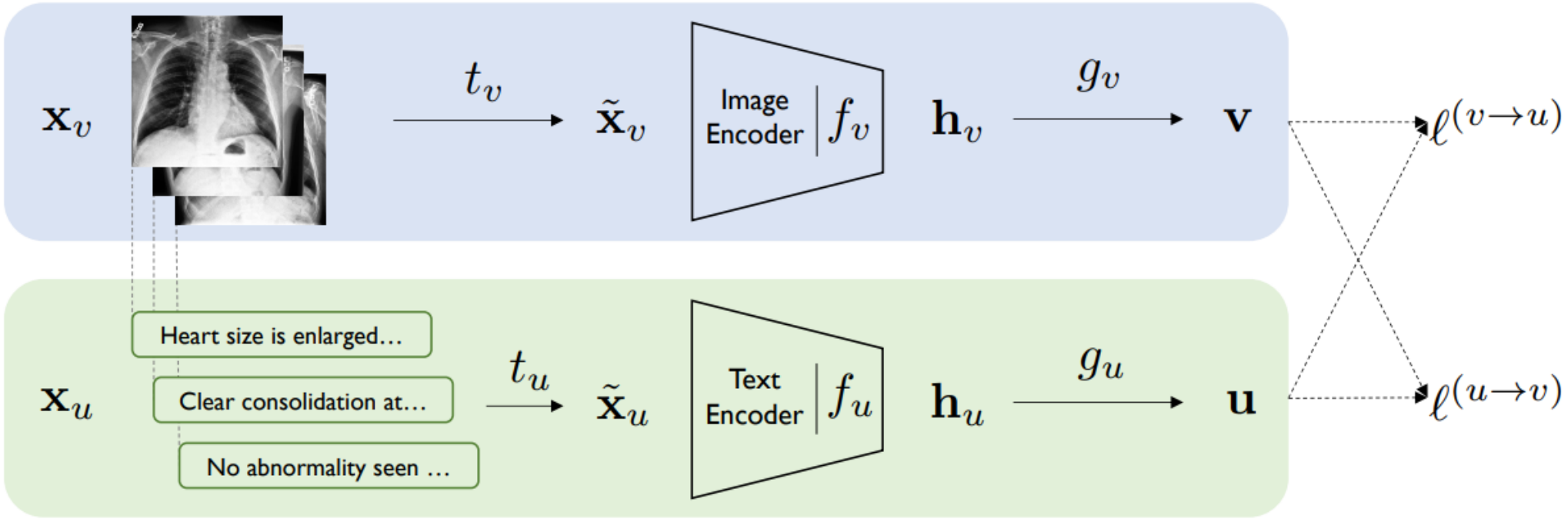
MedICaT

MedICaT is a dataset of medical images, captions, subfigure-subcaption annotations, and inline textual references. Instructions for access are provided here.

[1] *Radiology Objects in COntext (ROCO): A Multimodal Image Dataset.* MICCAI 2018 Workshop.
[2] *MedICaT: A Dataset of Medical Images, Captions, and Textual References.* EMNLP 2020 Findings.

Assume we have a collection of image-text pairs, …

- … we can simply perform contrastive learning.

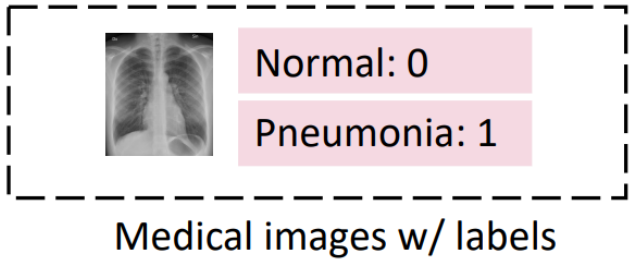


$$\ell_i^{(v \to u)} = -\log \frac{\exp(\langle \mathbf{v}_i, \mathbf{u}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{v}_i, \mathbf{u}_k \rangle / \tau)}, \qquad \ell_i^{(u \to v)} = -\log \frac{\exp(\langle \mathbf{u}_i, \mathbf{v}_i \rangle / \tau)}{\sum_{k=1}^{N} \exp(\langle \mathbf{u}_i, \mathbf{v}_k \rangle / \tau)}.$$

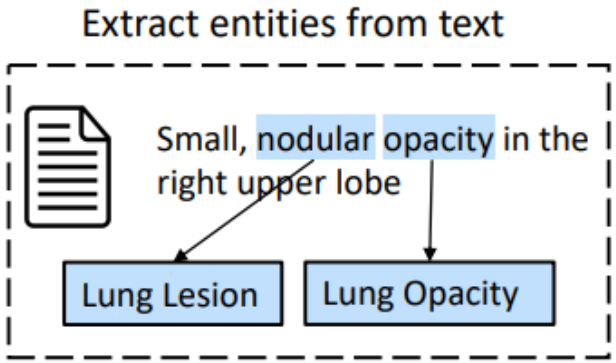
Contrastive Learning of Medical Visual Representations from Paired Images and Text. MLHC 2022.

Contrastive Learning from Unpaired Medical Images and Text

- What if we do not have a collection of image-text pairs?
  - For example, what if we want to train a vision-language model for mammograms?
- We may have some images with categorical labels.

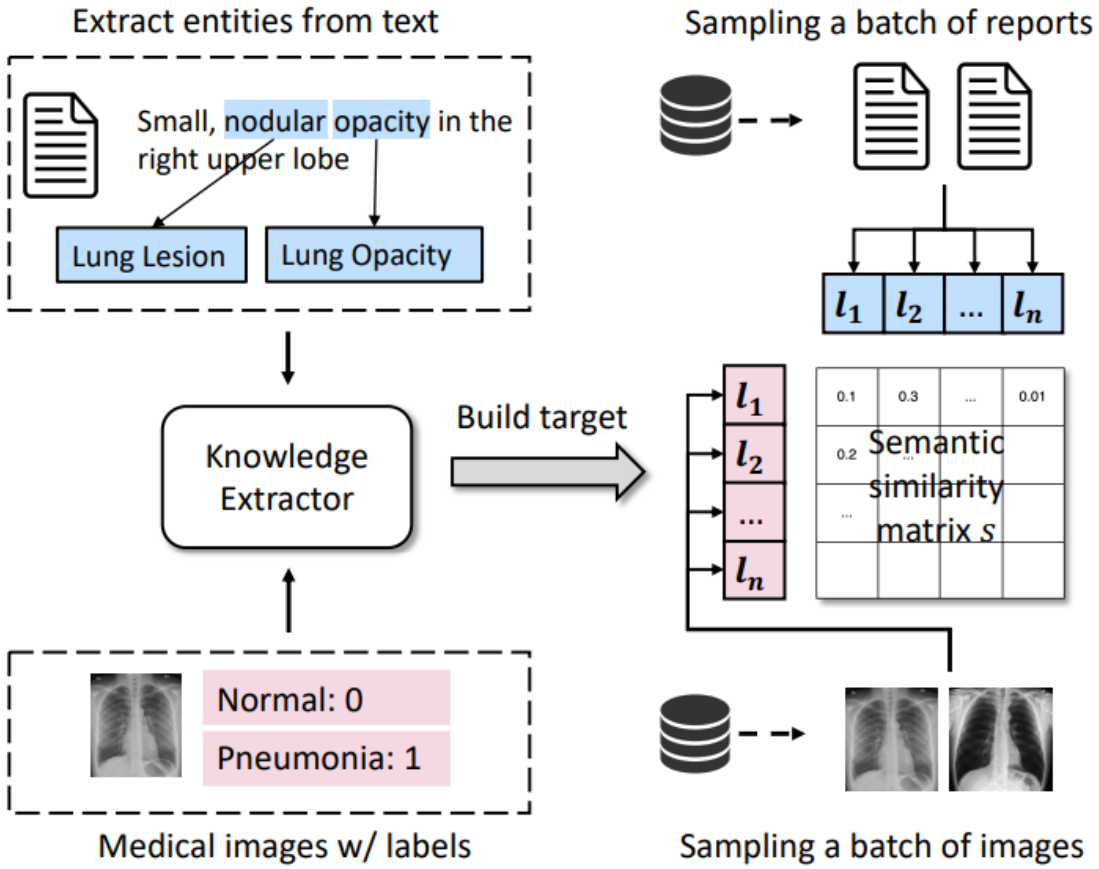


- We may have some medical sentences, where we can extract entities from.



Contrastive Learning from Unpaired Medical Images and Text

- Similarity(Image, Text) = Similarity(Labels of Image, Entities in Text)



$$s = \frac{\mathbf{l}_{img}^{\top} \cdot \mathbf{l}_{txt}}{\|\mathbf{l}_{img}\| \cdot \|\mathbf{l}_{txt}\|}.$$

$$y_{ij}^{v \to t} = \frac{\exp s_{ij}}{\sum_{j=1}^{N_{batch}} \exp s_{ij}}.$$

$$y_{ji}^{t \to v} = \frac{\exp s_{ji}}{\sum_{i=1}^{N_{batch}} \exp s_{ji}}.$$

MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. EMNLP 2022.

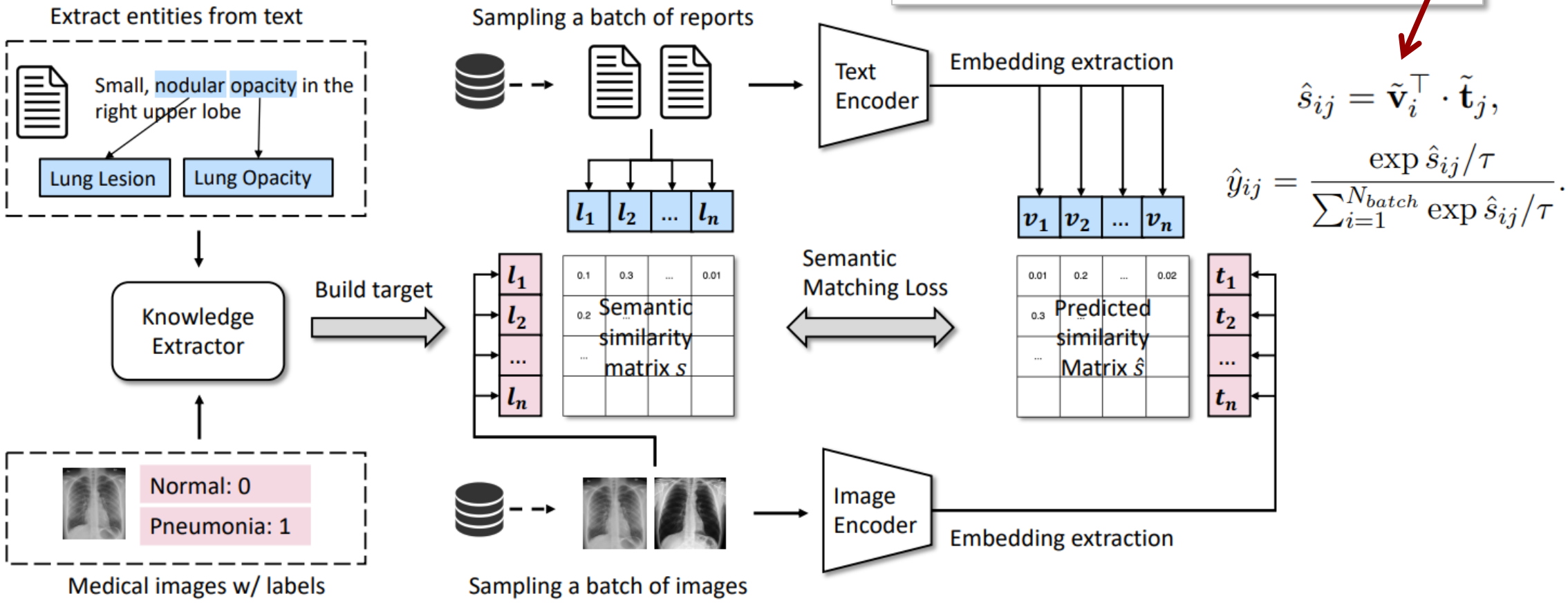
Contrastive Learning from Unpaired Medical Images and Text

- Train a CLIP model to learn the derived similarity.

$$\mathcal{L}^{v \to l} = -\frac{1}{N_{batch}} \sum_{i=1}^{N_{batch}} \sum_{j=1}^{N_{batch}} y_{ij} \log \hat{y}_{ij}.$$

$$\hat{s}_{ij} = \tilde{\mathbf{v}}_i^{\top} \cdot \tilde{\mathbf{t}}_j,$$

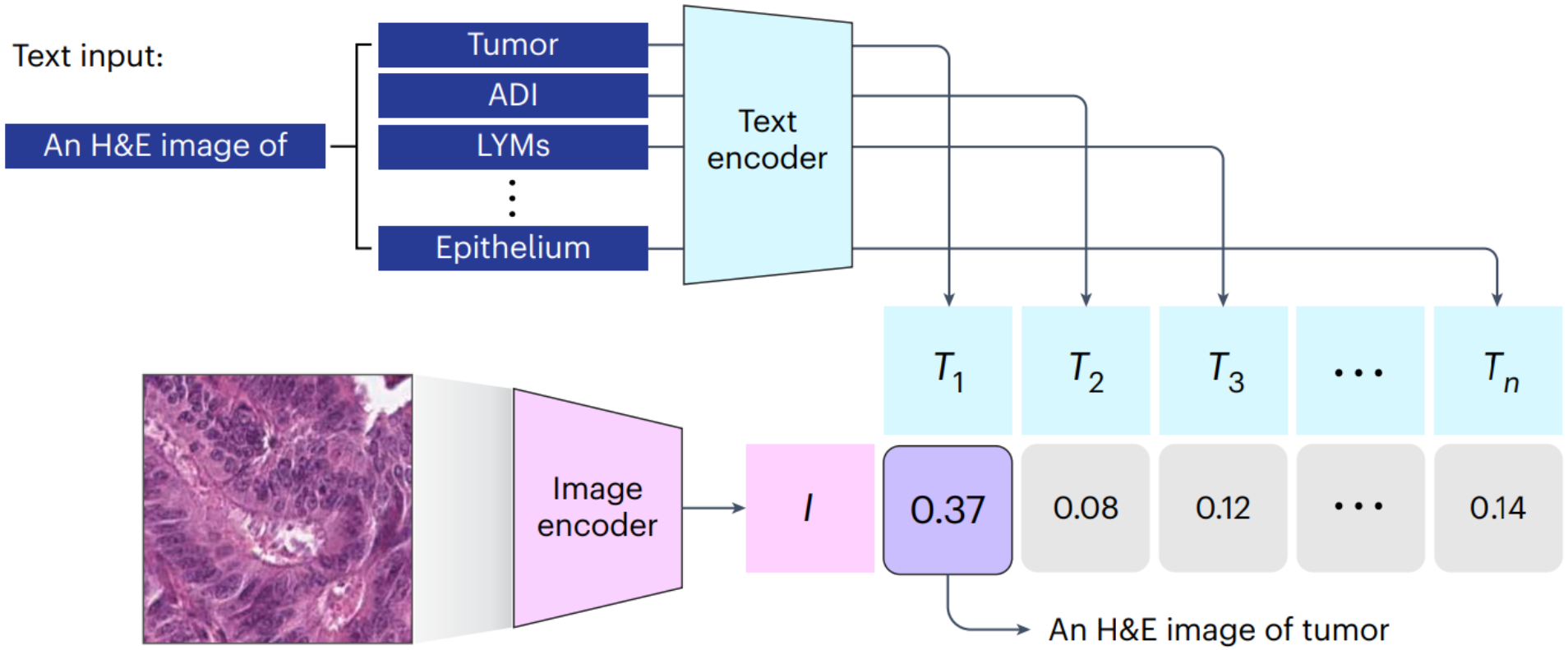
$$\hat{y}_{ij} = \frac{\exp \hat{s}_{ij}/\tau}{\sum_{i=1}^{N_{batch}} \exp \hat{s}_{ij}/\tau}.$$



MedCLIP: Contrastive Learning from Unpaired Medical Images and Text. EMNLP 2022.

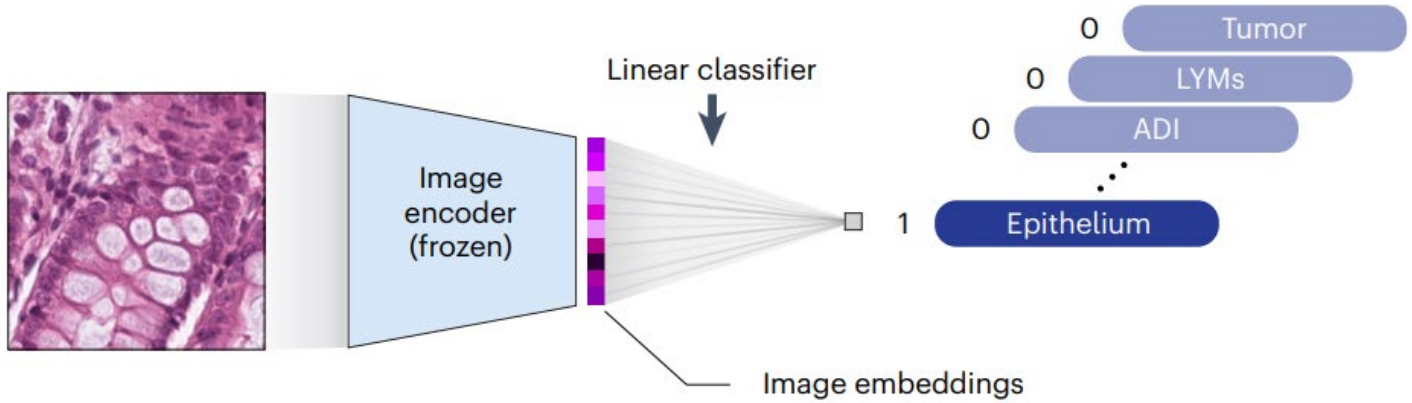
Tasks for Evaluating MedCLIP

- Zero-shot Image Classification
  - Enumerate all classes with a text template and calculate its similarity with the image



Tasks for Evaluating MedCLIP

- Supervised Image Classification
  - Add a classification layer upon the pre-trained image encoder



  - Fine-tune this architecture using labeled data
    - "*Fine-Tuning*": fine-tune all parameters in this architecture
    - "*Linear Probing*": fine-tune the classification layer only (more commonly used when evaluating CLIP-based models)

- Image-to-Text Retrieval
  - Given an image as a query, retrieve text relevant to this query,

Performance of MedCLIP: Zero-Shot Image Classification

- The text encoder of MedCLIP is initialized from BioClinicalBERT.

ACC(STD)	CheXpert-5x200	MIMIC-5x200	COVID	RSNA
CLIP	0.2016(0.01)	0.1918(0.01)	0.5069(0.03)	0.4989(0.01)
CLIP_{ENS}	0.2036(0.01)	0.2254(0.01)	0.5090(<0.01)	0.5055(0.01)
ConVIRT	0.4188(0.01)	0.4018(0.01)	0.5184(0.01)	0.4731(0.05)
$\text{ConVIRT}_{\text{ENS}}$	0.4224(0.02)	0.4010(0.02)	0.6647(0.05)	0.4647(0.08)
GLoRIA	0.4328(0.01)	0.3306(0.01)	0.7090(0.04)	0.5808(0.08)
$\text{GLoRIA}_{\text{ENS}}$	0.4210(0.03)	0.3382(0.01)	0.5702(0.06)	0.4752(0.06)
MedCLIP-ResNet	0.5476(0.01)	0.5022(0.02)	**0.8472(<0.01)**	0.7418(<0.01)
$\text{MedCLIP-ResNet}_{\text{ENS}}$	0.5712(<0.01)	**0.5430(<0.01)**	0.8369(<0.01)	0.7584(<0.01)
MedCLIP-ViT	0.5942(<0.01)	0.5006(<0.01)	0.8013(<0.01)	0.7447(0.01)
$\text{MedCLIP-ViT}_{\text{ENS}}$	**0.5942(<0.01)**	0.5024(<0.01)	0.7943(<0.01)	**0.7682(<0.01)**

Performance of MedCLIP: Supervised Image Classification and Image-to-Text Retrieval

ACC	CheXpert -5x200	MIMIC -5x200	COVID	RSNA
Random	0.2500	0.2220	0.5056	0.6421
ImageNet	0.3200	0.2830	0.6020	0.7560
CLIP	0.3020	0.2780	0.5866	0.7303
ConVIRT	0.4770	0.4040	0.6983	0.7846
GLoRIA	0.5370	0.3590	0.7623	0.7981
MedCLIP	**0.5960**	**0.5650**	**0.7890**	**0.8075**

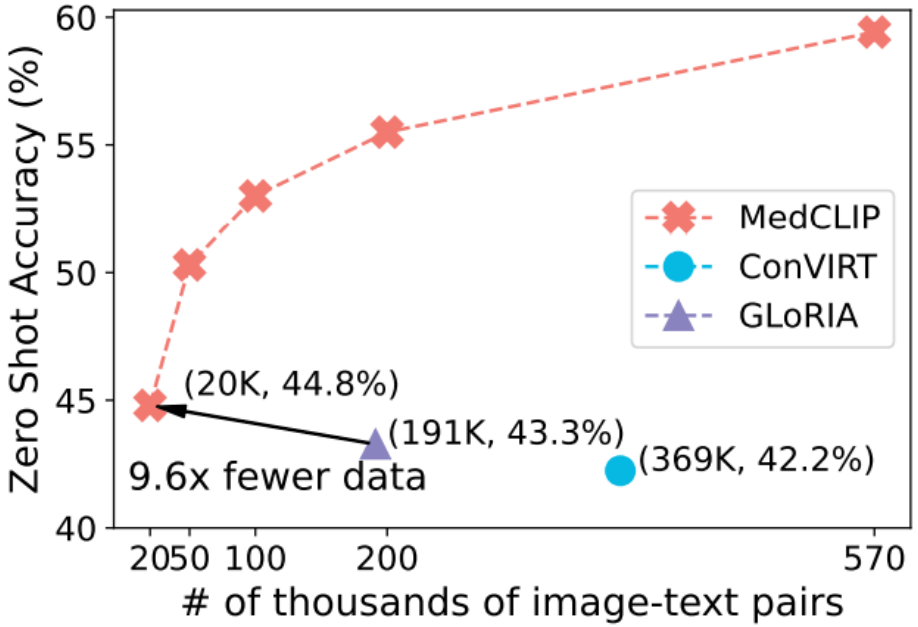
Linear Probing

Model	P@1	P@2	P@5	P@10
CLIP	0.21	0.20	0.20	0.19
ConVIRT	0.20	0.20	0.20	0.21
GLoRIA	**0.47**	0.47	0.46	0.46
MedCLIP	0.45	**0.49**	**0.48**	**0.50**

Image-to-Text Retrieval

Take-Away Messages

- Contrastive Image-Text Pre-training enables zero-shot image classification and image-to-text/text-to-image retrieval.
- Unpaired images and text also benefit the contrastive learning process.
  - The model using the combination of unpaired and paired data can beat the model using much more paired data only.

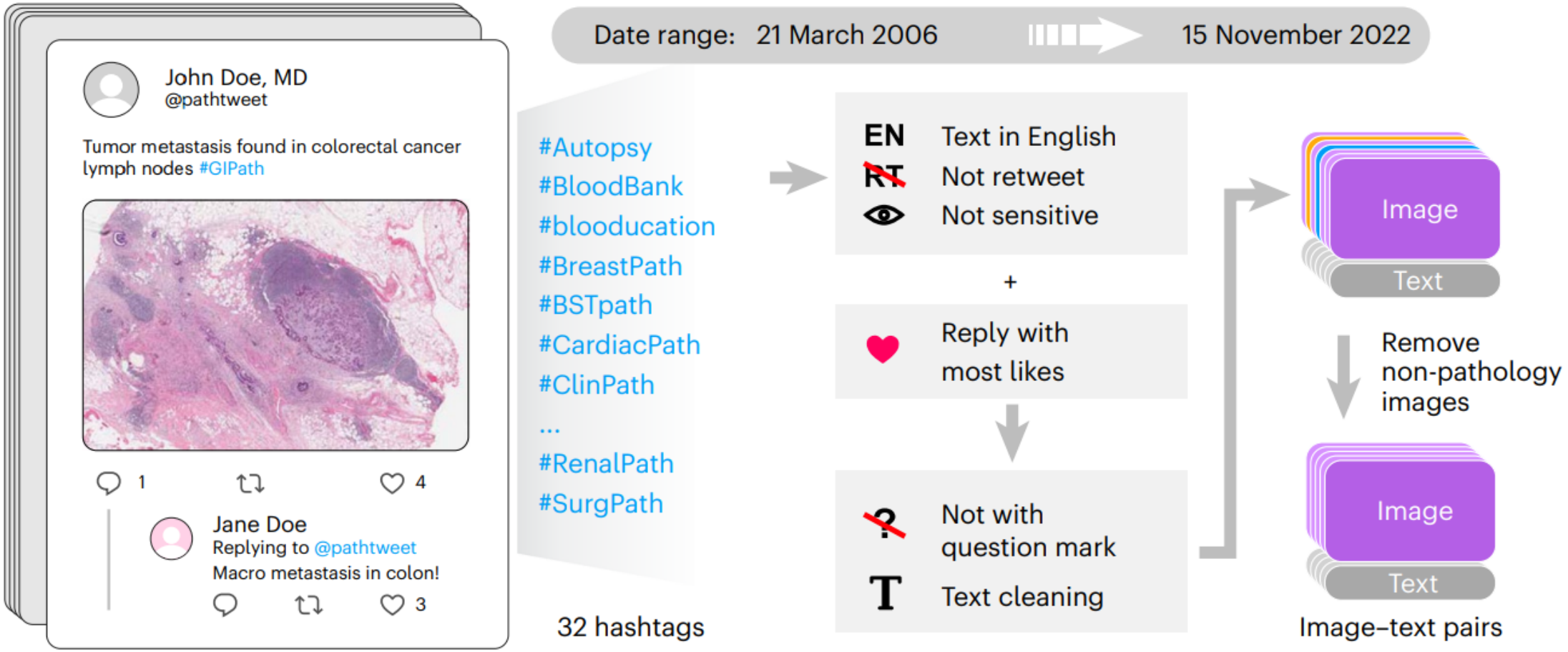


- Limitation:
  - Experiments on chest X-rays only (but one motivation for considering unpaired images and text is we need to deal with a new type of images).
  - The authors remove the links between paired images and text for pre-training, which is still different from unpaired images and text in practice.

Agenda

- Encoder-Only
  - MedCLIP: Contrastive Learning from Unpaired Medical Images and Text
  - PLIP: Harvesting Image-Text Pairs from Twitter
- Decoder-Only
  - LLaVA-Med: Visual Instruction Tuning
- Encoder-Decoder
  - BiomedGPT: Diverse Modalities and Tasks

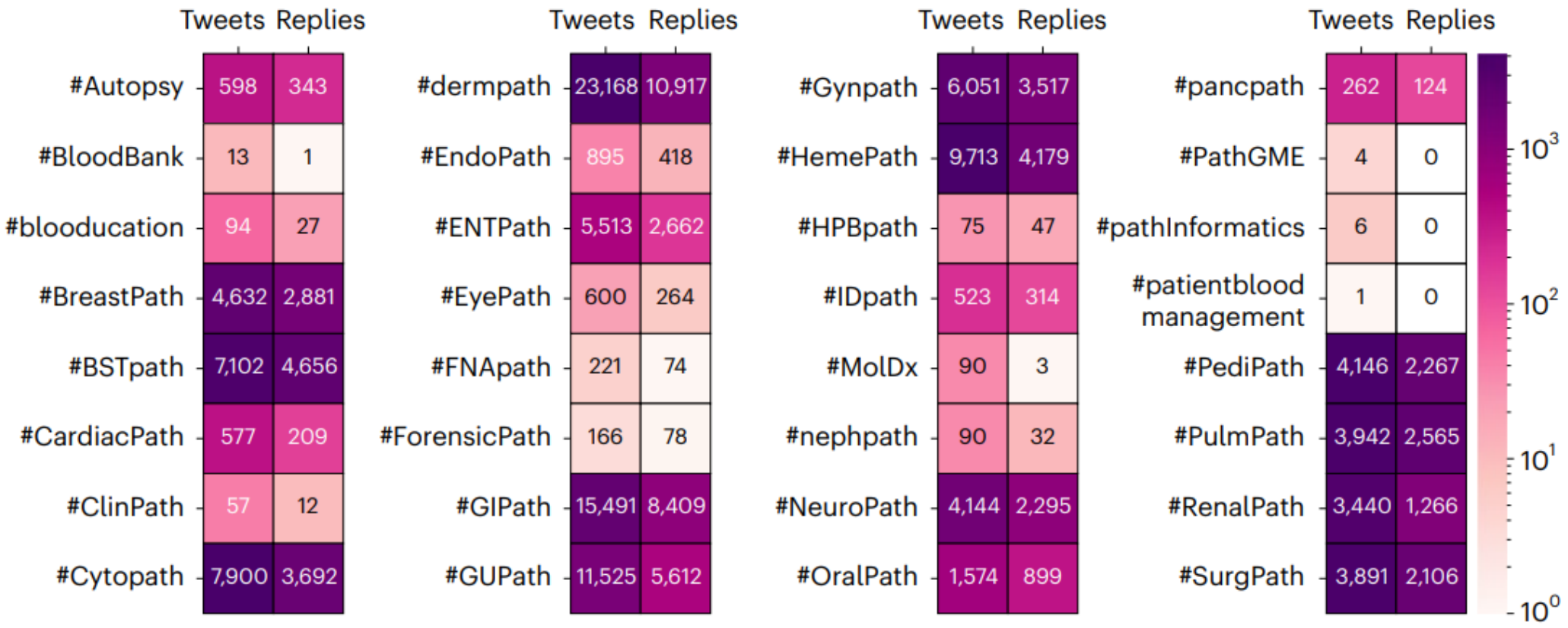
Pathology Images on Twitter



A Visual–Language Foundation Model for Pathology Image Analysis using Medical Twitter. Nature Medicine 2023

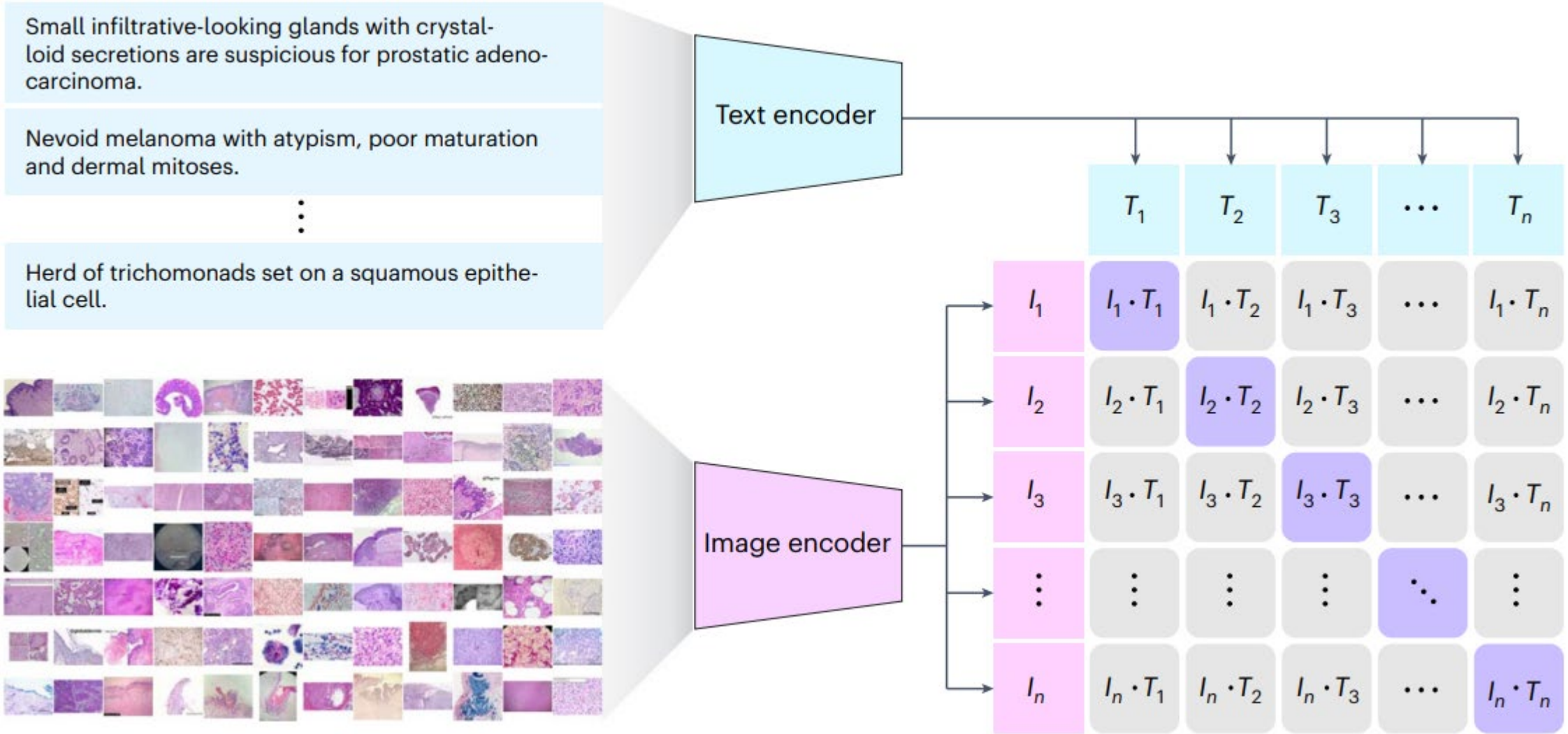
Dataset Statistics

116,504 image-text pairs from tweets and 59,869 image-text pairs from the top replies.



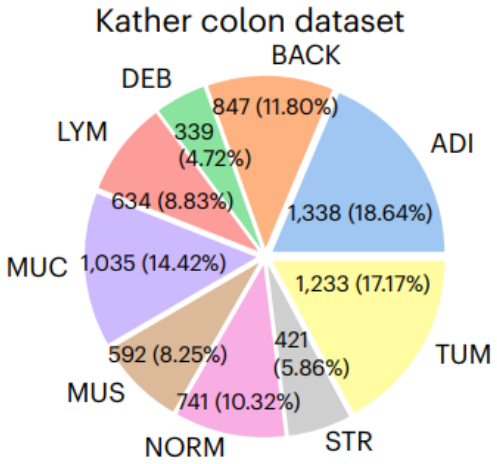
A Visual–Language Foundation Model for Pathology Image Analysis using Medical Twitter. Nature Medicine 2023

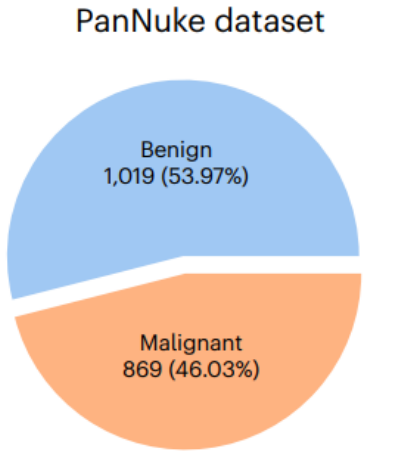
PLIP: The CLIP Architecture

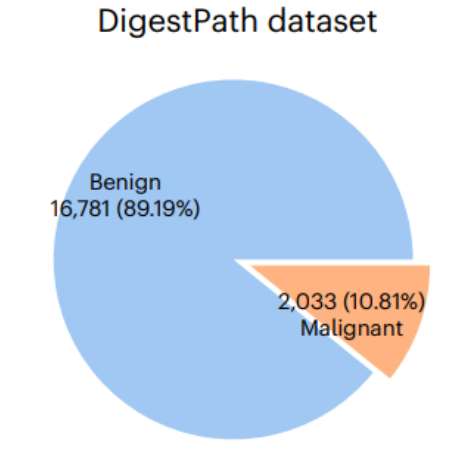


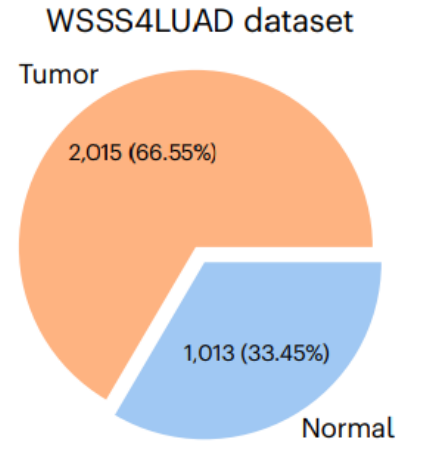
A Visual–Language Foundation Model for Pathology Image Analysis using Medical Twitter. Nature Medicine 2023

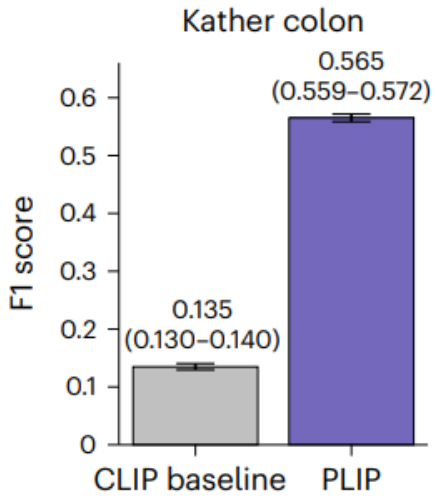
Performance of PLIP: Zero-Shot Classification

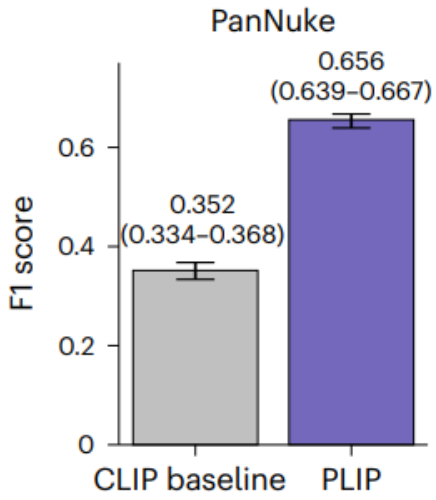


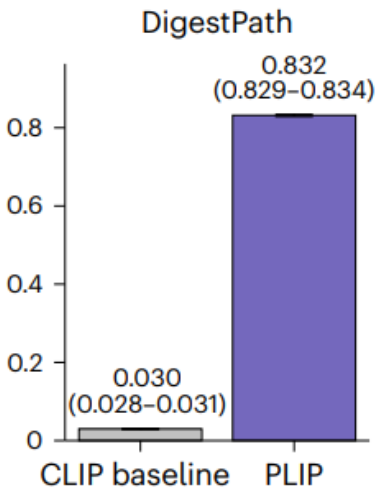


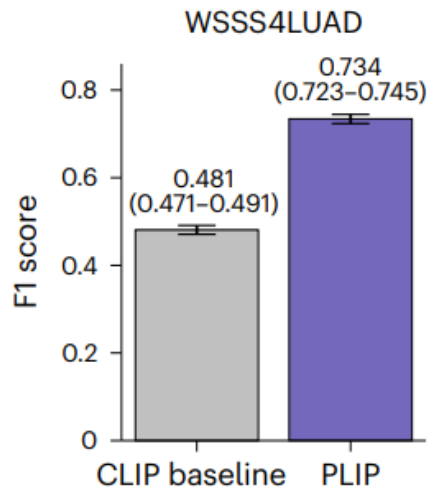












Performance of PLIP: Linear Probing and Text-to-Image Retrieval

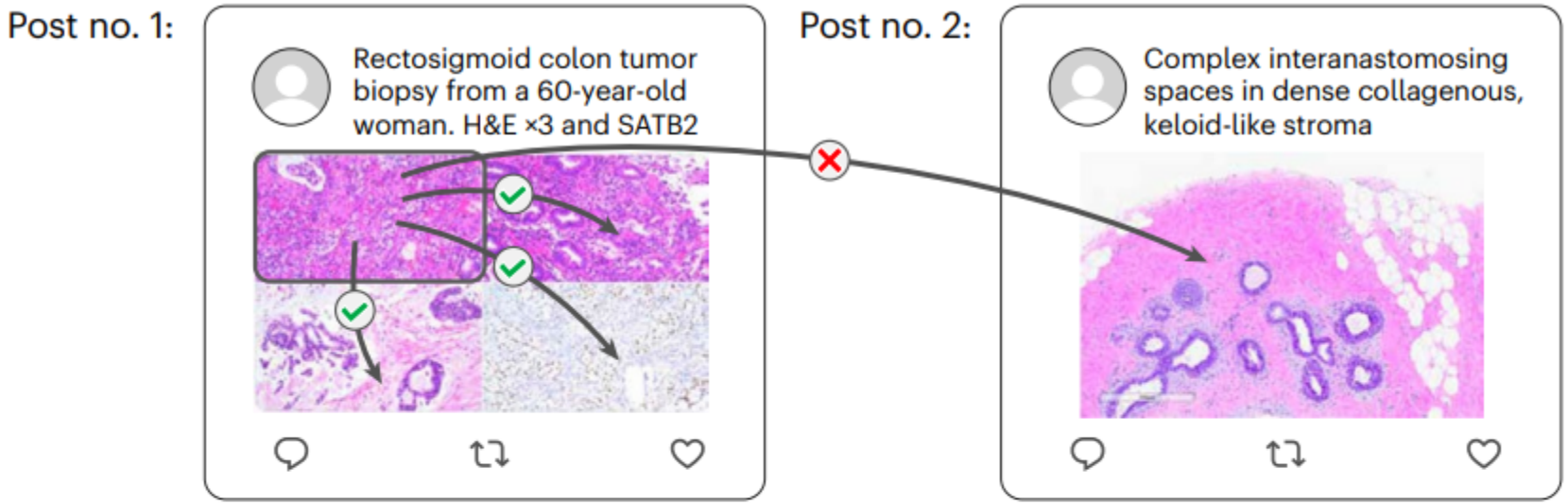
Linear Probing

	Kather colon	PanNuke	DigestPath	WSSS4LUAD	Average
CLIP	0.797 (±0.006)	0.852 (±0.002)	0.753 (±0.009)	0.850 (±0.022)	0.813 (±0.043)
MuDiPath	0.825 (±0.001)	0.896 (±0.001)	0.827 (±0.007)	0.917 (±0.003)	0.866 (±0.041)
PLIP	0.877 (±0.001)	0.902 (±0.010)	0.856 (±0.008)	0.927 (±0.007)	0.891 (±0.028)
PLIP versus CLIP	2.9×10^{-9}	9.4×10^{-6}	1.5×10^{-7}	1.5×10^{-4}	—
PLIP versus MuDiPath	9.4×10^{-12}	0.249	6.2×10^{-4}	3.0×10^{-2}	—

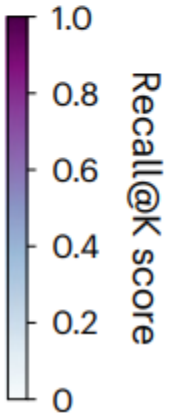
Image-to-Text Retrieval

Dataset	Number of candidates	Metric	PLIP	CLIP
Twitter	2,023	Recall@10	0.271	0.061
		Recall@50	0.527	0.128
PathPedia	210	Recall@10	0.409	0.167
		Recall@50	0.752	0.476
PubMed	1,419	Recall@10	0.069	0.015
		Recall@50	0.206	0.082
Books	558	Recall@10	0.265	0.045
		Recall@50	0.659	0.165

Performance of PLIP: Image-to-Image Retrieval



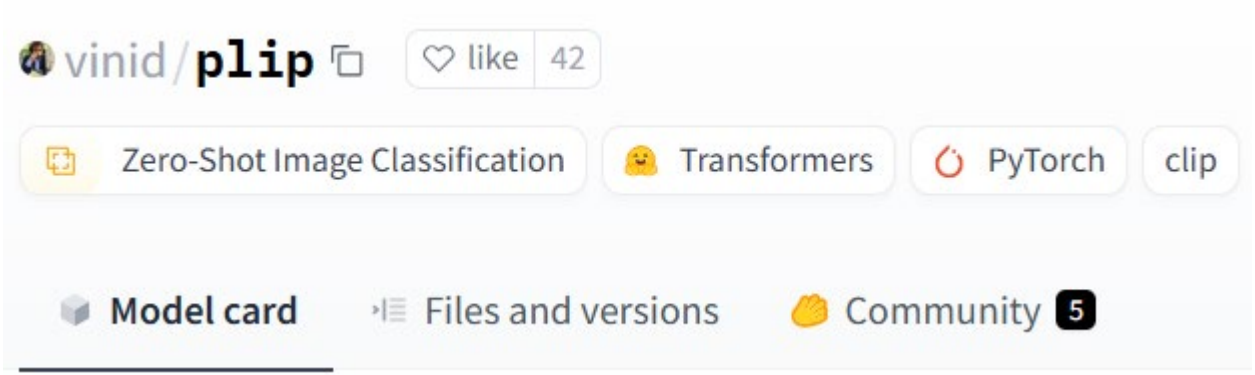
	PLIP	CLIP	MuDiPath (Euclidean)	SISH
Recall@10	0.646 (116.1x)	0.353 (63.5×)	0.336 (60.3×)	0.356 (64.0×)
Recall@50	0.814 (34.5x)	0.513 (21.7×)	0.485 (20.6×)	0.474 (20.1×)



Take-Away Messages

- Publicly shared medical information, such as Twitter, is a tremendous source of pathology image-text pairs.
  - What if you do know that many relevant hashtags?

https://huggingface.co/vinid/plip

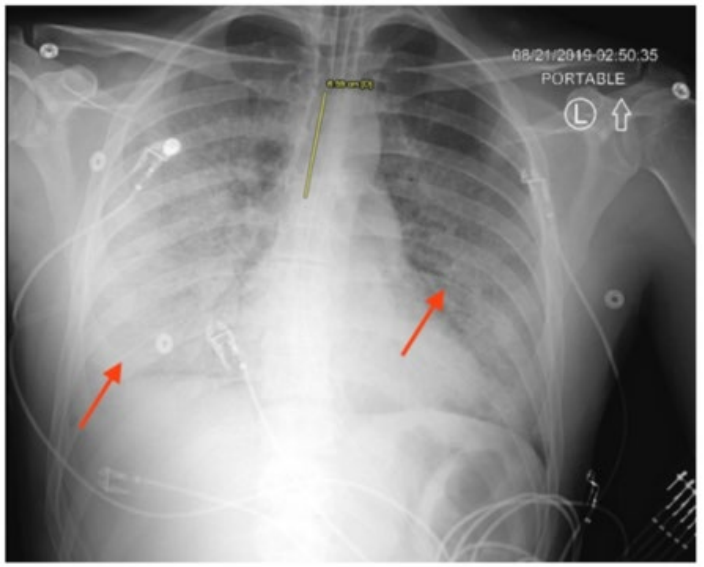


- Limitation:
  - The paper only introduces a way to get more high-quality data. Technically, it still exactly follows the CLIP architecture.

Agenda

- Encoder-Only
  - MedCLIP: Contrastive Learning from Unpaired Medical Images and Text
  - PLIP: Harvesting Image-Text Pairs from Twitter
- Decoder-Only
  - LLaVA-Med: Visual Instruction Tuning
- Encoder-Decoder
  - BiomedGPT: Diverse Modalities and Tasks

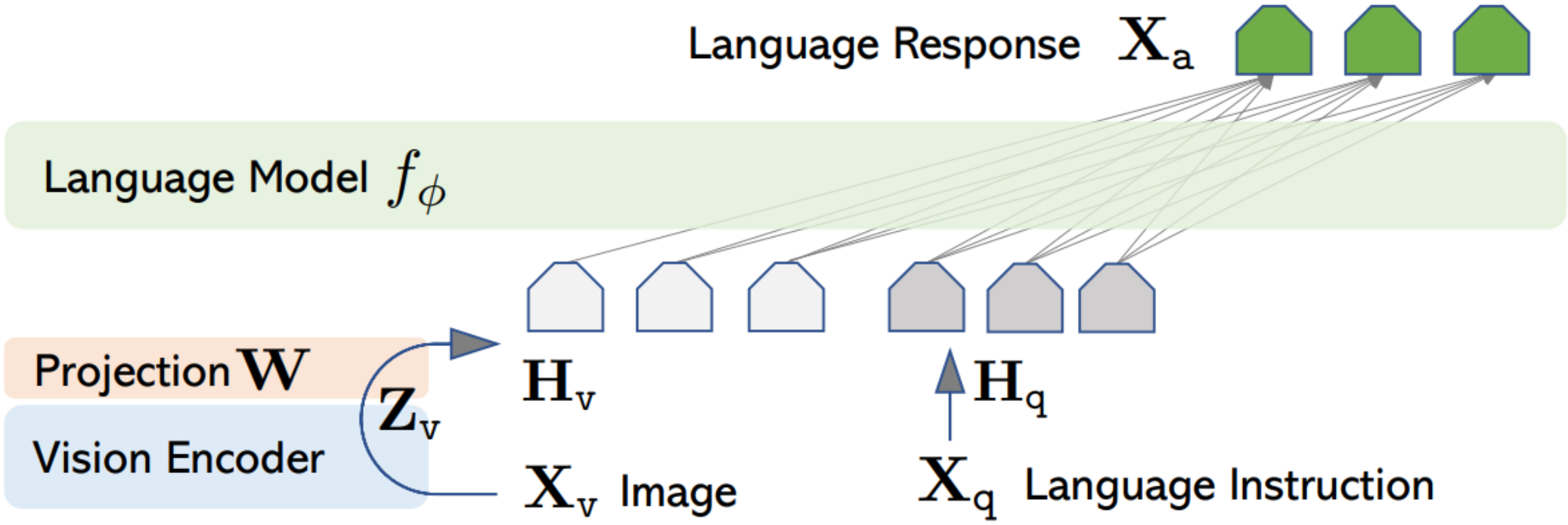
Biomedical Visual Question Answering (VQA)



- Question: What is shown in this image?
- Answer: 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.

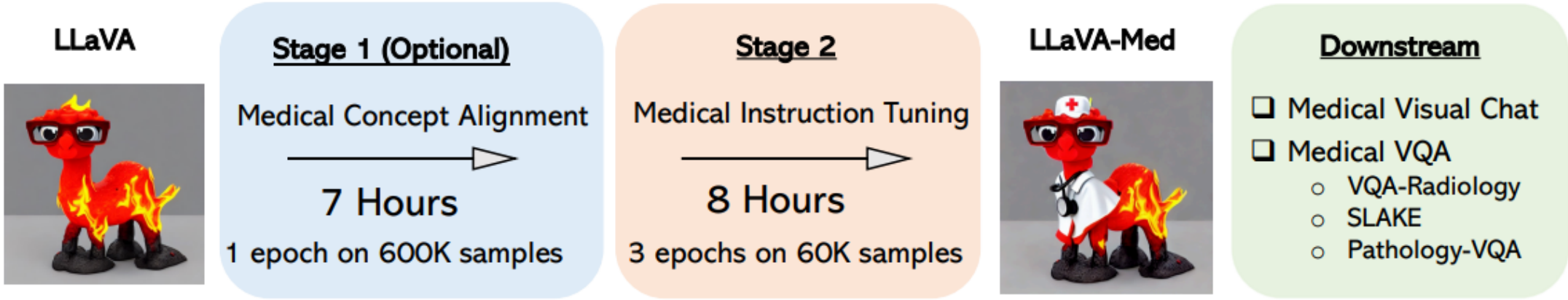
Adding Images into a Decoder-Based Architecture – LLaVA

- Project images onto several vision tokens.
- Prepend vision tokens to text tokens for next token prediction.



Visual Instruction Tuning. NeurIPS 2023.

LLaVA-Med: Adapting LLaVA to the Biomedical Domain

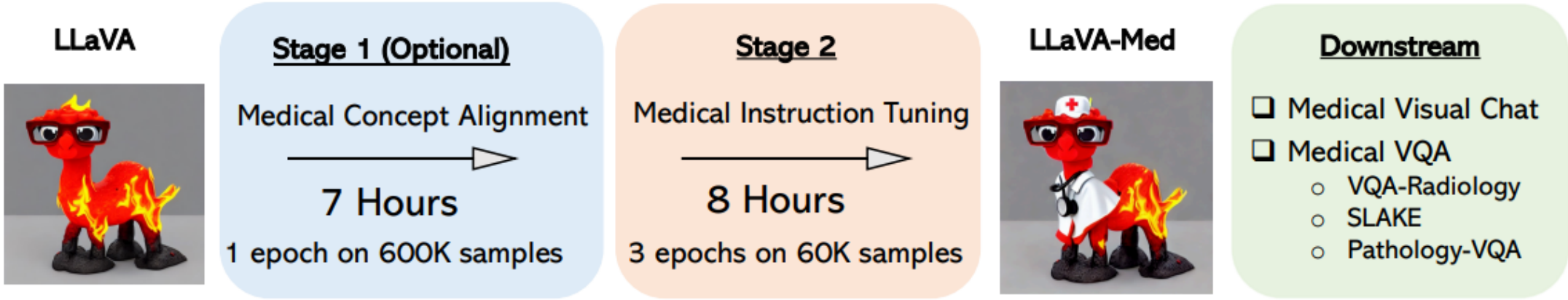


- Step 1: Medical Concept Alignment
  - Crawl figure-caption pairs from PMC full-text papers
  - Given a figure, ask the model "*What is shown in this image?*" and train it to generate the corresponding caption.

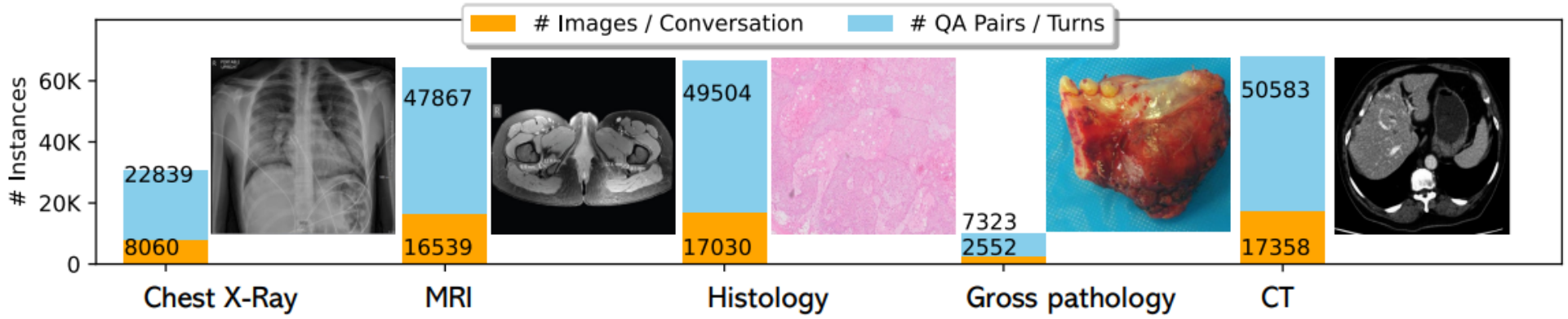
$$\texttt{Human : } \mathbf{X}_q\ \mathbf{X}_v\texttt{<STOP>\n Assistant : } \mathbf{X}_c\texttt{<STOP>\n}$$

LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS 2023.

LLaVA-Med: Adapting LLaVA to the Biomedical Domain



- Step 2: Medical Instruction Tuning
  - Use GPT-4 to construct instruction-following data



LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day. NeurIPS 2023.

Using GPT-4 to Generate Instruction-Following Data

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.
```

Adding Few-shot Examples to the Prompt of GPT-4

One of the few-shot examples used in the prompt

Example input:
Figure Caption:
Figure 2: Chest X-ray. Cardiomegaly with diffuse bilateral interstitial infiltrates and a right-sided pleural effusion.

Figure Context:
- His respiratory status continued to worsen and on hospital day number 14 , he required intubation and mechanical ventilation for hypoxemic respiratory failure (Figure 2) .

Example output:

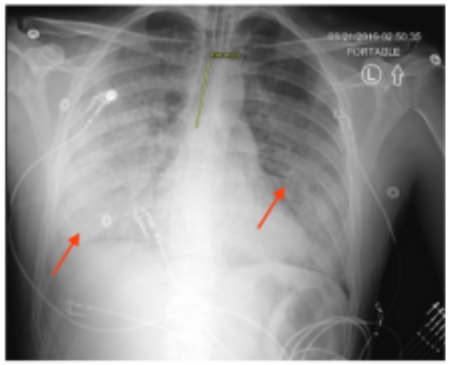
```
User: Which side of the lung has a pleural effusion?

Assistant: The right side of the lung has a pleural effusion.

User: According to the CT image, does the patient has cardiomegaly?

Assistant: The image appeares to be a chest x-ray. However, it
indicates that the patient has cardiomegaly, which means an enlarged
heart. Additionally, there are diffuse bilateral interstitial
infiltrates and a right-sided pleural effusion present in the image.

User: Additional Info: The patient's respiratory status continued
to worsen and on hospital day number 14, requiring intubation and
mechanical ventilation for hypoxemic respiratory failure.

Assistant: Thank you for providing additional information about
the patient's condition. Based on the information you provided, it
appears that the patient's respiratory status continued to worsen and
on hospital day number 14, they required intubation and mechanical
ventilation for hypoxemic respiratory failure. This suggests that
the patient's lung function has significantly declined, and they
may be experiencing respiratory distress. It is important for the
medical team to closely monitor the patient's condition and provide
appropriate treatment to manage their symptoms and improve their lung
function.
```

Performance of LLaVA-Med: Biomedical Visual Chatbot

- Task: Given images of PMC full-text papers, perform conversation with humans.
- Evaluation: Use GPT-4 to quantify the correctness of the model output.

	Question Types		**Domains**					**Overall**
	Conversation	Description	CXR	MRI	Histology	Gross	CT	
(Question Count)	(143)	(50)	(37)	(38)	(44)	(34)	(40)	(193)
LLaVA	39.4	26.2	41.6	33.4	38.4	32.9	33.4	36.1
LLaVA-Med								
Stage 1	22.6	25.2	25.8	19.0	24.8	24.7	22.2	23.3
10K	42.4	32.5	46.1	36.7	43.5	34.7	37.5	39.9
60K	53.7	36.9	57.3	39.8	49.8	47.4	52.4	49.4
60K-IM	55.1	36.4	56.2	40.4	52.7	51.8	50.1	50.2

Performance of LLaVA-Med: Biomedical Visual Chatbot



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)

Source: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7398722/

User	What is shown in this image?
LLaVA-Med	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.
User	Are there any devices implanted in this patient?
LLaVA-Med	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.
User	What is shown in this image?
LLaVA	The image shows a ribcage with wires coming out of it, indicating that it is likely a medical illustration or diagram.
User	Are there any devices implanted in this patient?
LLaVA	Yes, there are several devices protruding from the patient's chest, which are likely medical devices used for treatment or monitoring.

Performance of LLaVA-Med: Biomedical VQA

Method	VQA-RAD			SLAKE			PathVQA		
	Ref	Open	Closed	Ref	Open	Closed	Ref	Open	Closed
Supervised finet-tuning results with our own experiment runs									
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**
Representative & SoTA methods with numbers reported in the literature									
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

Take-Away Messages

- Adapting general-domain decoder-only VLMs to a specific scientific domain (e.g., biomedicine) bears similarity with adapting general-domain decoder-only LLMs to a specific scientific domain.
  - We need domain-specific data (i.e., image-text pairs) to perform next token prediction.
  - We need instruction-following data to perform instruction tuning.
    - If it is hard to manually construct such data, we can use GPT-4.
  - Key technical novelty: Representing images as visual tokens

- Drawbacks
  - Information loss when projecting images onto visual tokens
  - No strategies to handle multiple images as input
  - Cannot generate images (i.e., no strategies to visual tokens back to images)

Agenda

- Encoder-Only
  - MedCLIP: Contrastive Learning from Unpaired Medical Images and Text
  - PLIP: Harvesting Image-Text Pairs from Twitter
- Decoder-Only
  - LLaVA-Med: Visual Instruction Tuning
- Encoder-Decoder
  - BiomedGPT: Diverse Modalities and Tasks

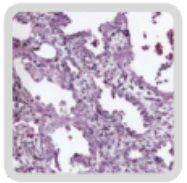
A Generalist Vision-Language Foundation Model

Diverse Modalities

Diverse Tasks

Diverse Modalities and Tasks

Pathology and radiology VQA



Q: What is seen at this stage, associated with regeneration and repair?
A: Numerous reactive type II pneumocytes.
Q: Are bite cells like this one in the smear associated with regeneration and repair at this stage?
A: No.

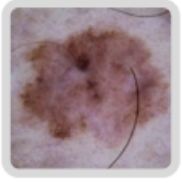
Report generation

Q: What are the findings based on the image?
A: The nasogastric tube is in adequate position, and there is a resolution of the gastric distention. There is still mild bibasilar atelectasis. There are no pneumothorax no pleural effusion.

Disease diagnosis

Q: What disease does this image depict?
A: Breast cancer.

Lesion detection



Q: What skin lesion does this image depict?
A: Melanoma.

Diverse Modalities and Tasks

Mortality prediction

Chief complaint: Dyspnea, abdominal distention
Present illness: 78-year-old female with multiple medical problems, pertinently including CAD status post CABG, hypertension and type 2 diabetes
Medical history: (1) CAD status post CABG [Reg#] (2) Hypertension (3) Type 2 diabetes (4) Pulmonary fibrosis
Allergies: Patient recorded as having no known allergies to drugs
...
Physical exam: BP 107/68, HR 70s, RR 28, 90% on NRB

Q: What is the predicted outcome for the patient before discharge?
A: Deceased.

Treatment suggestion

ID	Age	Sex	Race	Tumor size	Nodes examined	...	ER Status	PR Status
001	54	Female	White	25	14		Negative	Negative
002	44	Female	White	23	34		Positive	Positive
003	47	Male	White	22	3	...	Positive	Positive
004	89	Female	White	17	1		Negative	Positive
005	59	Female	White	36	19		Positive	Negative

The patient is a 44-year-old white female. She has one malignant tumor and five regional lymph nodes that tested positive. The tumor measures 23 mm. Estrogen and progesterone receptor tests are positive. A total of 34 regional nodes were removed.

Q: Please provide treatment suggestion given the patient's information.
A: Recommend using beam radiation, suggesting that the sequence for radiation should be post-surgery. Furthermore, chemotherapy should indeed be considered.

Conversation summarization

What symptoms are you experiencing?

I have been suffering from headache and general weakness and have been diagnosed with typhoid fever.

What treatments have you undergone in the past?

I was put on siprofloxacin for ten days. One week after I completed the course, I started having the same problems.

It seems you are having relapse of typhoid fever. Up to 10% of patients develop mild relapse. You need more tests to check for other causes, including liver tests, because your urine is dark, and hepatitis can sometimes happen with typhoid fever.

Q: What is the summary of conversation?
A: Suggest remedy for recurred symptoms of typhoid after treatment.

Report summarization

There is no evidence of hemorrhage, masses, mass effect or shift of normally midline structures. The ventricles and sulci are mildly prominent, compatible with age-appropriate involutional changes. There is hypoattenuation along the right caudate head and periventricular frontal white matter, unchanged, compatible with small vessel ischemic disease.

Q: What is summary based on the given report?
A: 1. No acute intracranial process.
2. Small vessel ischemic disease.

Clinical-trial matching

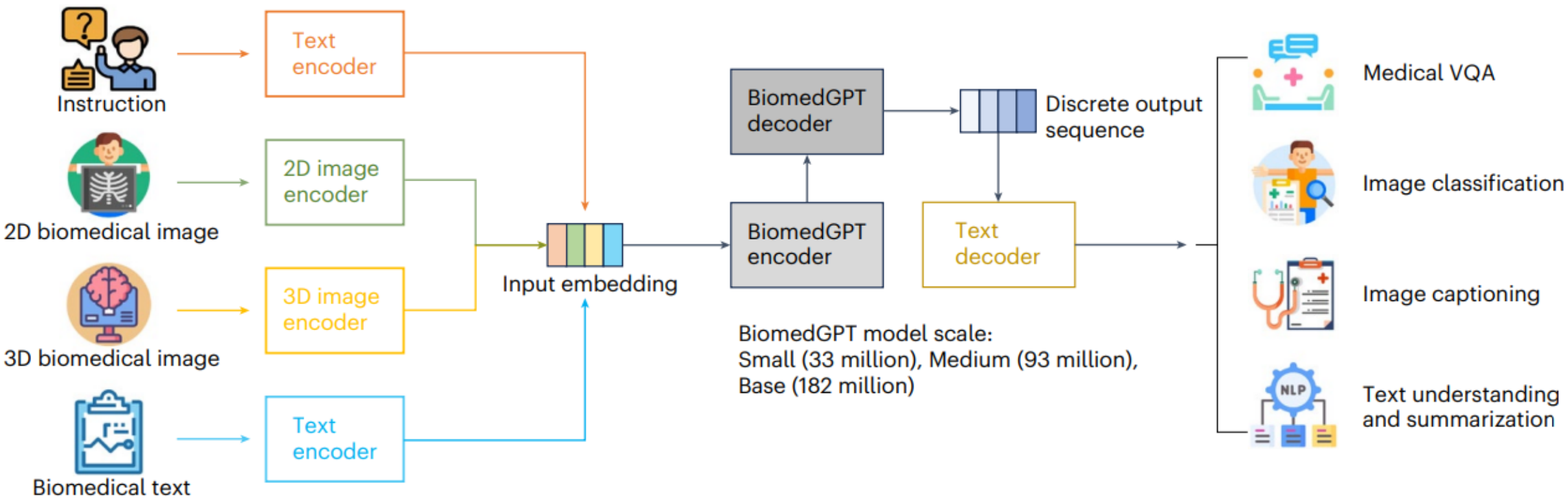
Patient information
A **19**-year-old **male** came to the clinic with some sexual concerns. On physical examination, there are some poorly developed secondary sexual characteristics. Ultrasound reveals a **testes volume of 1-2 ml**. The hormonal evaluation showed a **serum testosterone** level of **65 ng dl^{-1}** with low levels of GnRH.

Clinical trial information
Description: Evaluate the safety and efficacy of Androxal.

Inclusion criteria: Total serum testosterone concentrations < 300 ng dl^{-1}.
Male patients over the age of 18.
Exclusion criteria: Elevated PSA > 3.5 ng ml^{-1}.

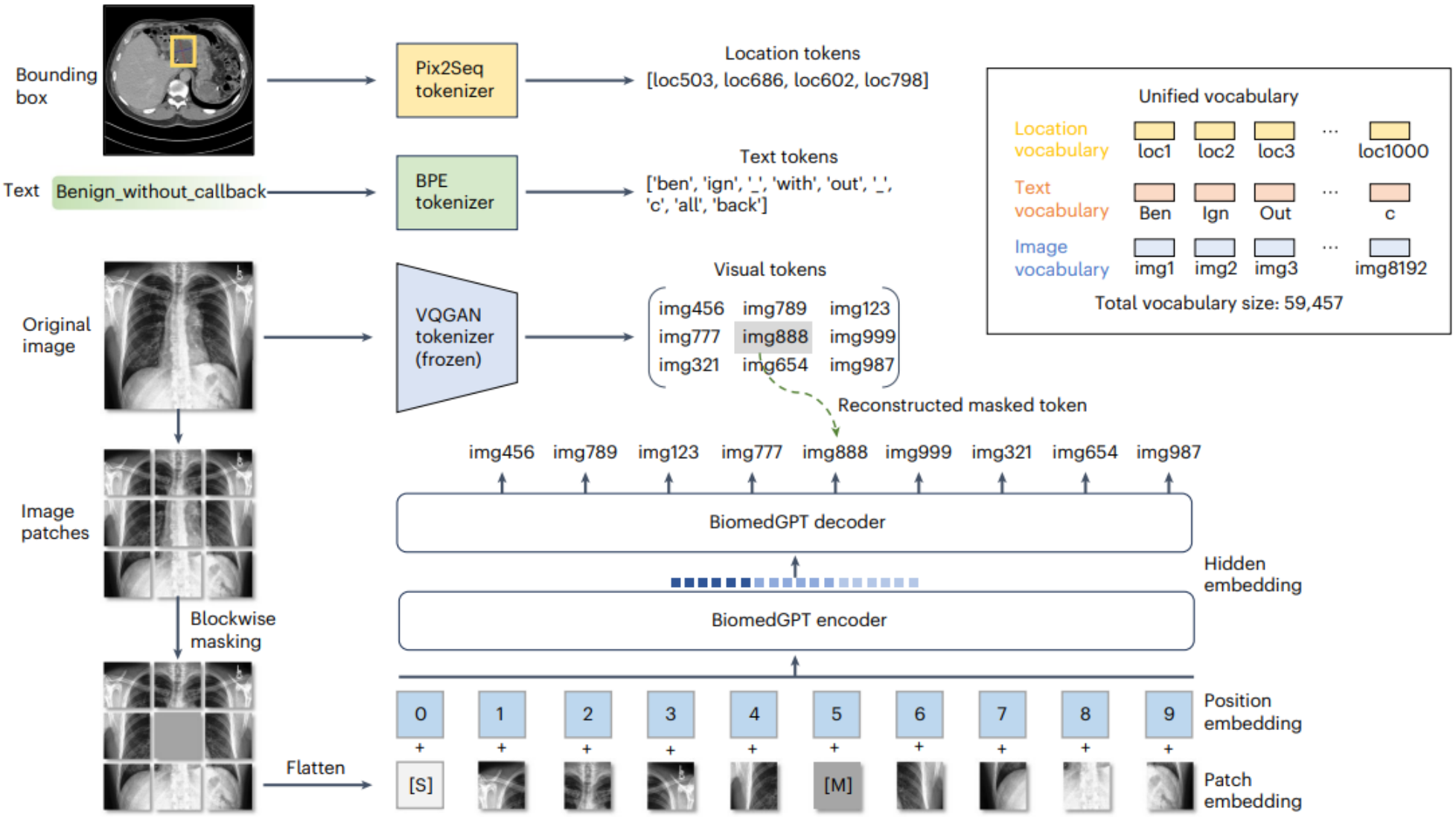
Q: Please determine the patient's eligibility by comparing the given patient note and trial details.
A: The patient is eligible for the clinical trial.

BiomedGPT: Handling Multi-modal Inputs



A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks. Nature Medicine 2024.

BiomedGPT: Handling Multi-modal Inputs



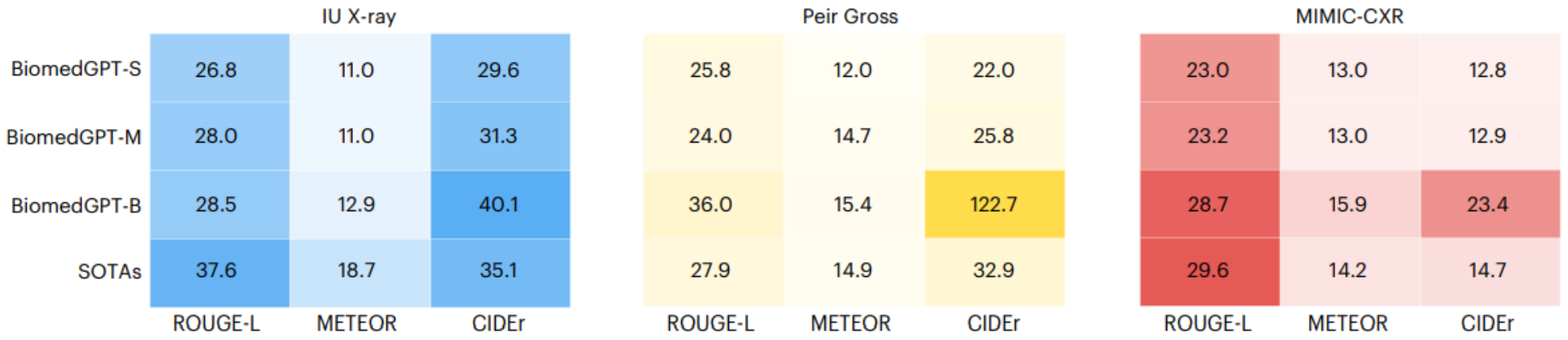
A Generalist Vision-Language Foundation Model for Diverse Biomedical Tasks. Nature Medicine 2024.

Evaluation of BiomedGPT: VQA and Image Captioning

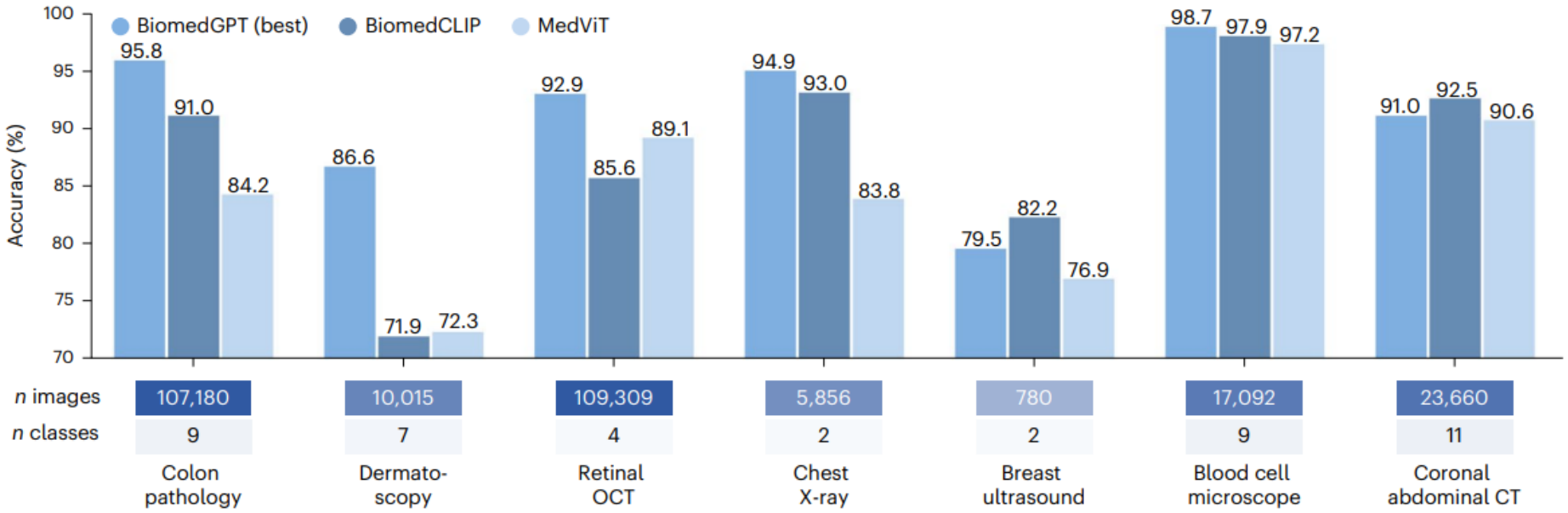
VQA

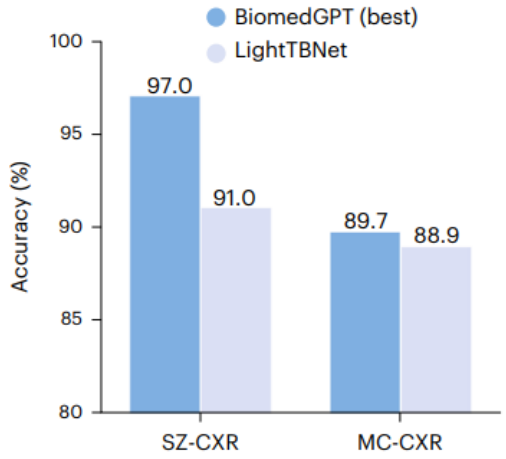
Model	Parameters	VQA-RAD accuracy		SLAKE accuracy		PathVQA accuracy	
		Closed-ended	Open-ended	Closed-ended	Open-ended	Closed-ended	Open-ended
BiomedGPT-S (ours)	33M (0.2×)	57.8 (23.5↓)	13.4 (47.5↓)	73.3 (16.6↓)	66.5 (17.8↓)	84.2 (3.8↓)	10.7 (17.3↓)
BiomedGPT-M (ours)	93M (0.5×)	79.8 (1.5↓)	53.6 (7.3↓)	86.8 (3.1↓)	78.3 (6.0↓)	85.7 (2.3↓)	12.5 (15.5↓)
M2I2	252M (1.4×)	81.6 (0.3↑)	61.8 (0.9↑)	91.1 (0.2↑)	74.7 (9.6↓)	88.0	36.3 (8.3↑)
BiomedCLIP	422M (2.3×)	79.8 (1.5↓)	67.6 (6.7↑)	89.7 (0.2↓)	82.5 (1.8↓)	-	-
CLIP-ViT with GPT2-XL	1.6B (8.8×)	-	-	82.1 (7.8↓)	84.3	87.0 (1.0↓)	40.0 (12.0↑)
MedVInT-TD	7.0B (38.5×)	86.8 (5.5↑)	73.7 (12.8↑)	86.3 (3.6↓)	84.5 (0.2↑)	-	-
BiomedGPT-B (ours)	182M	81.3	60.9	89.9	84.3	88.0	28.0

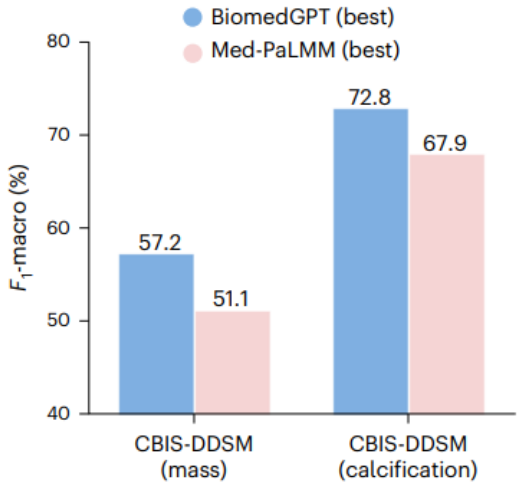
Image Captioning



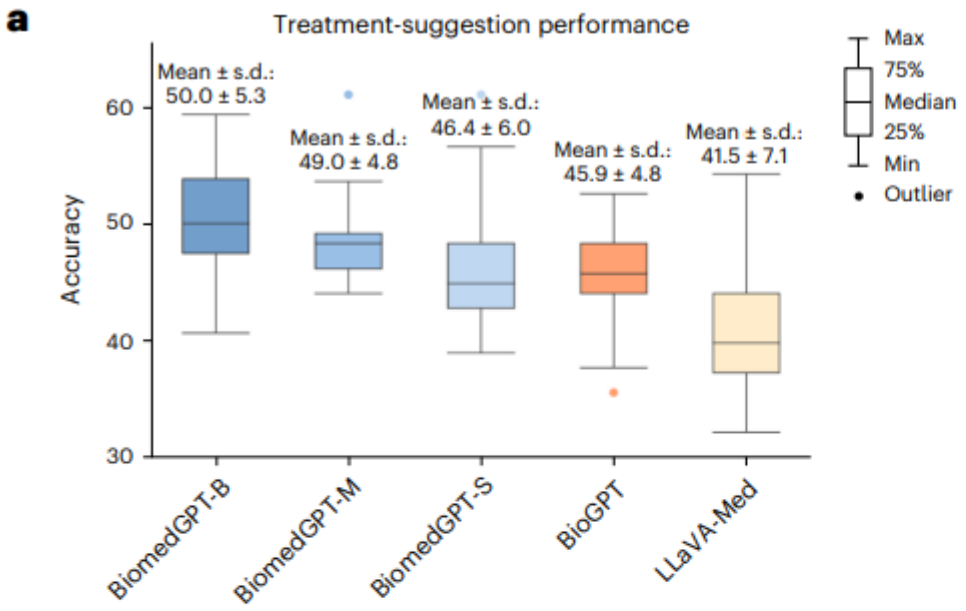
Evaluation of BiomedGPT: Image Classification

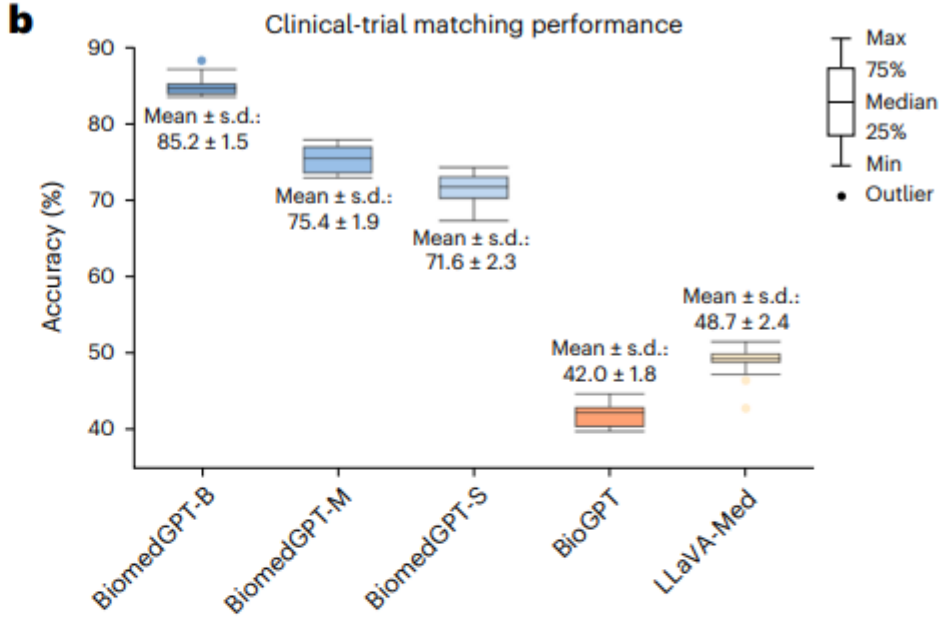






Evaluation of BiomedGPT: Text-Only Tasks

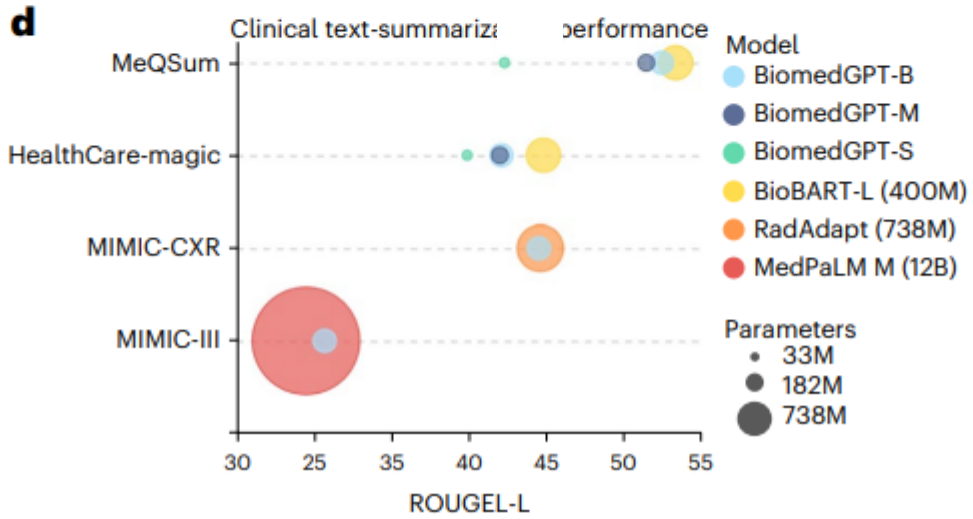


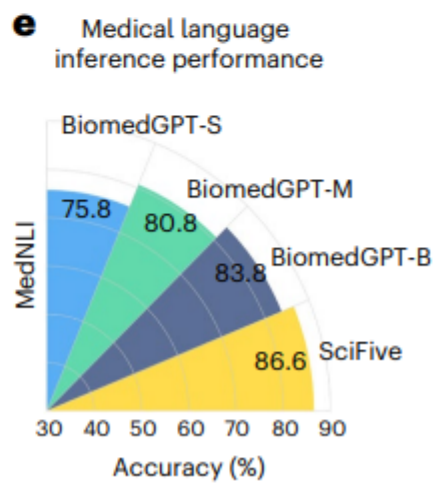


c

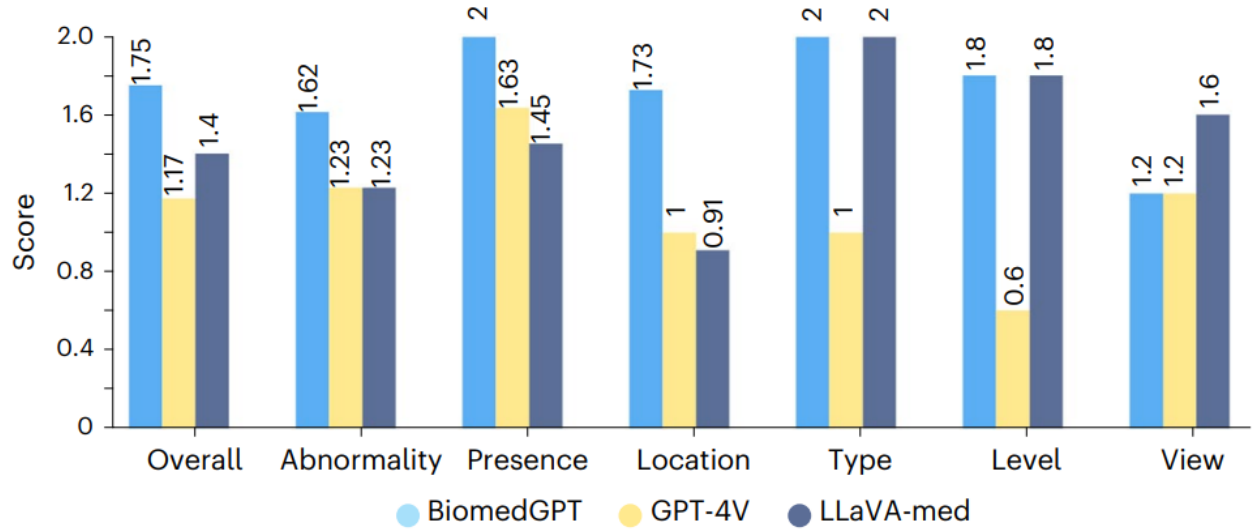
In-hospital mortality prediction performance

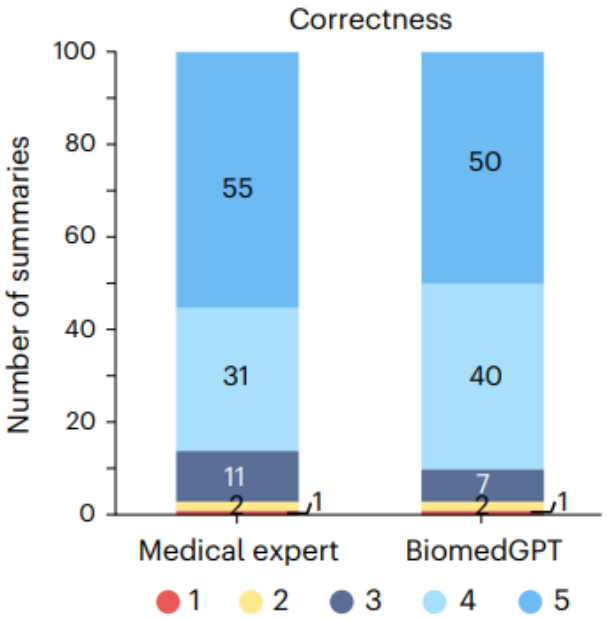
	Accuracy
BiomedGPT-B	89.5
BiomedGPT-M	89.2
BiomedGPT-S	77.8
BioGPT	74.2
LLaVA-Med	72.8

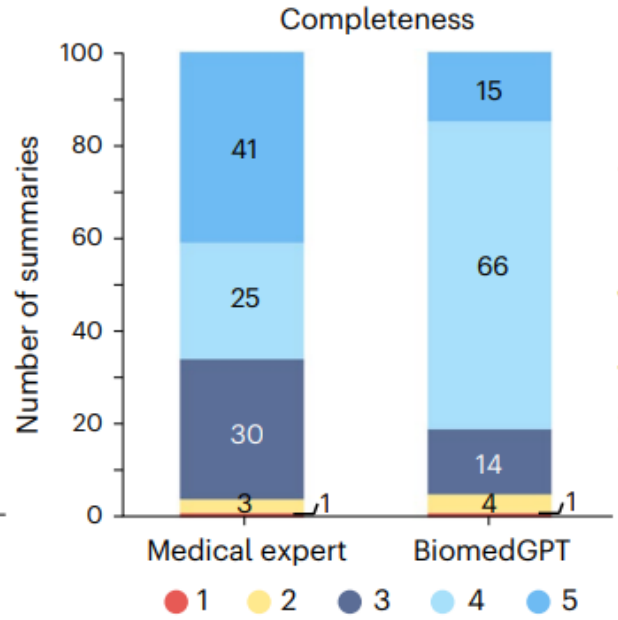


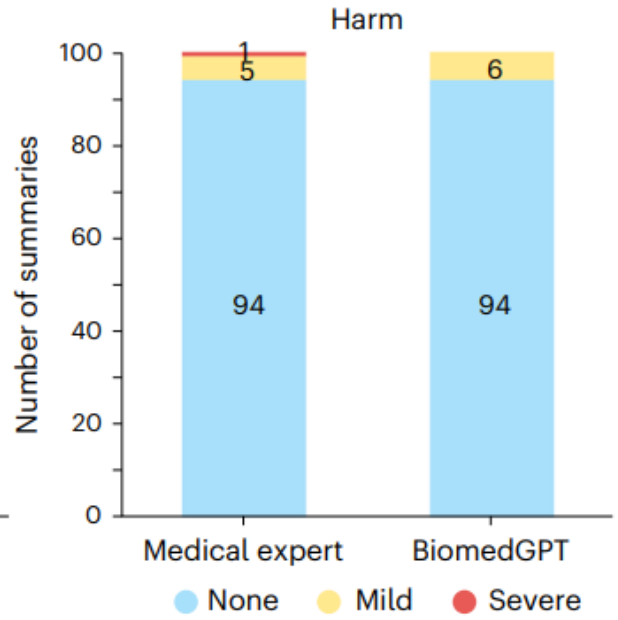


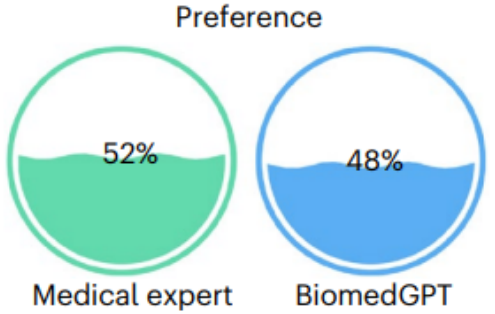
Human Evaluation of VQA and Text Summarization





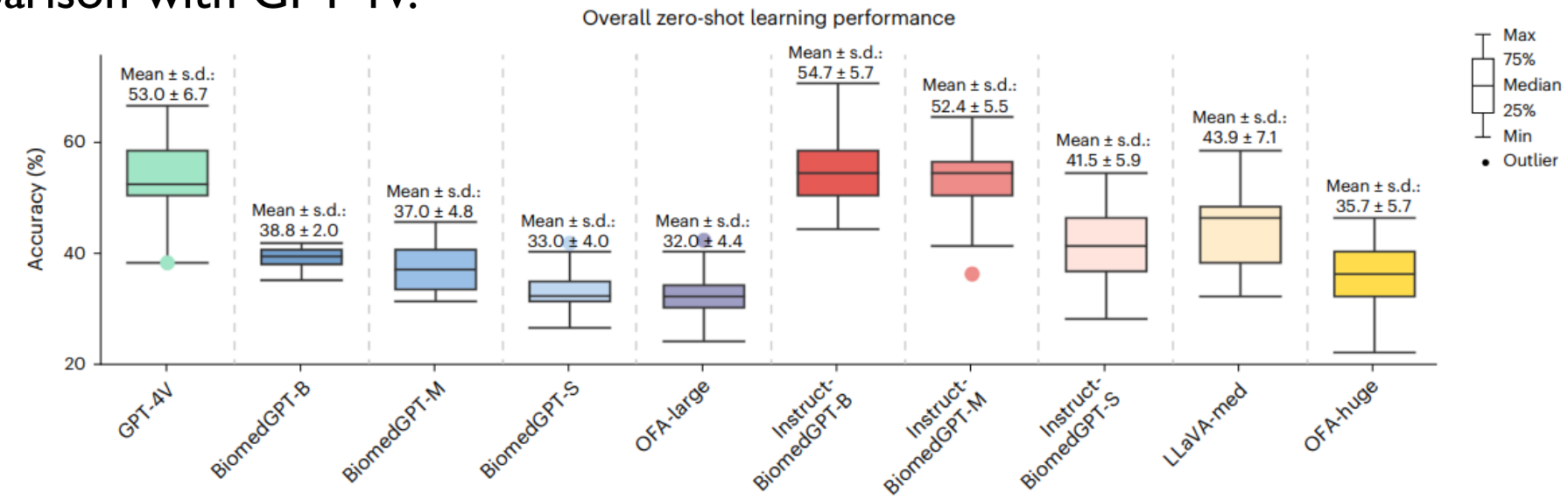






Take-Away Messages

- Biomedical images are beyond chest X-rays or 2-dimensional images only. Collecting instruction-tuning data from a wide spectrum of vision-language and language-only tasks enables the model to generalize to diverse image modalities and tasks.
- Leveraging different tokenizers to map all types of data to tokens paves the way for a generalist vision-language foundation model.
- Drawback: The zero-shot ability of BiomedGPT has not been fully established in comparison with GPT-4V.



Thank You!

Course Website: https://yuzhang-teaching.github.io/CSCE689-S25.html