

Multimodal Large Language Models for Healthcare

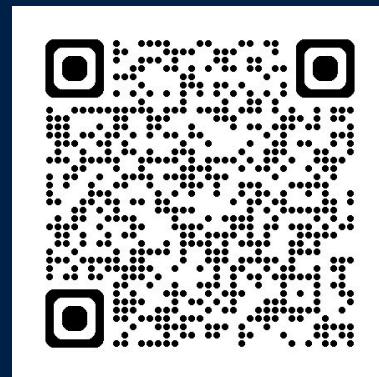
https://knowlab.github.io/glasgow_tutorial/

Yunsoo Kim, Jinge Wu, Yusuf Abdulle, Honghan Wu

School of Health and Wellbeing, University of Glasgow

Institute of Health Informatics, University College London

24 November 2025



Era of Large Language Models

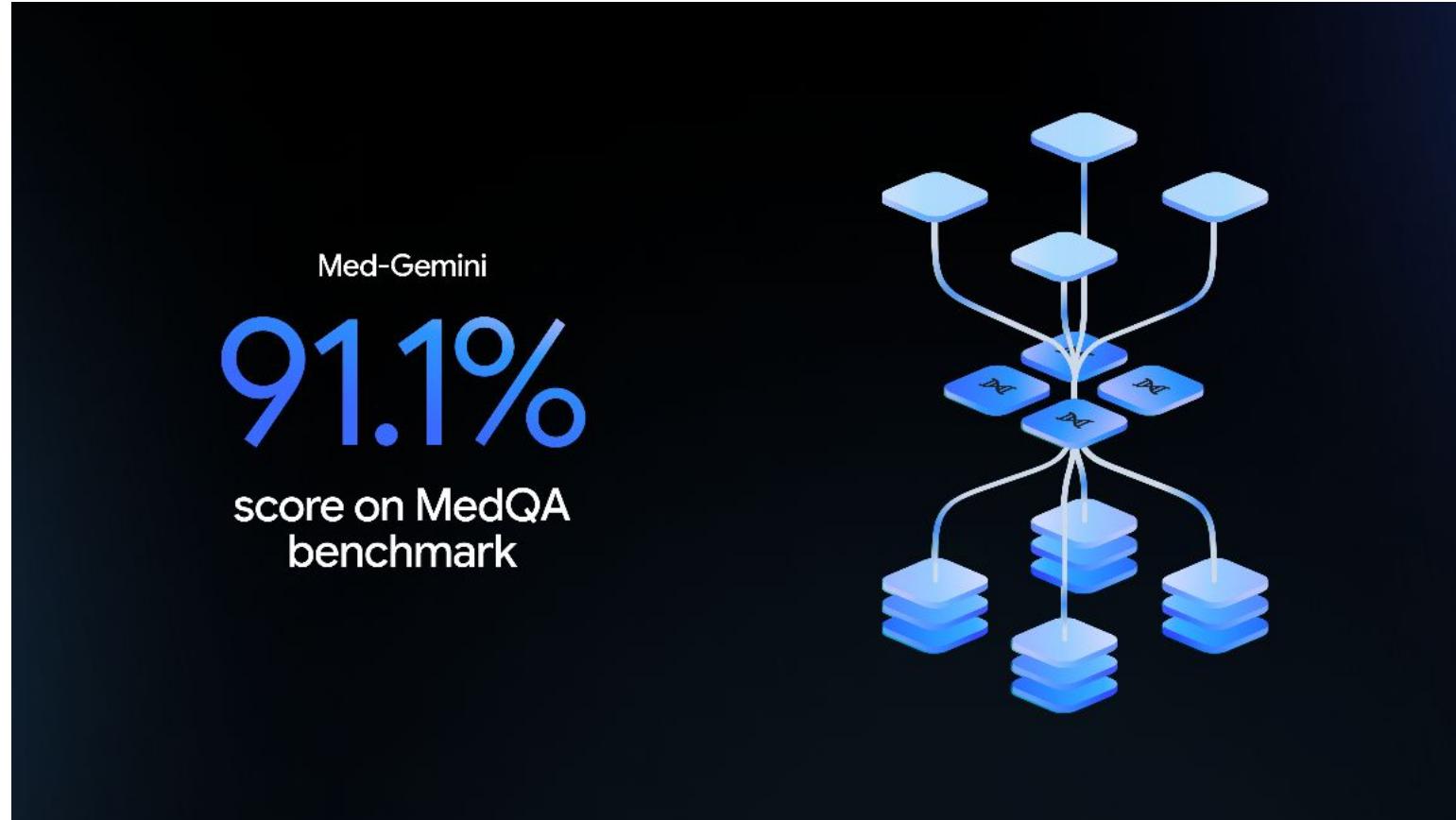
Honghan Wu

School of Health and Wellbeing, University of Glasgow

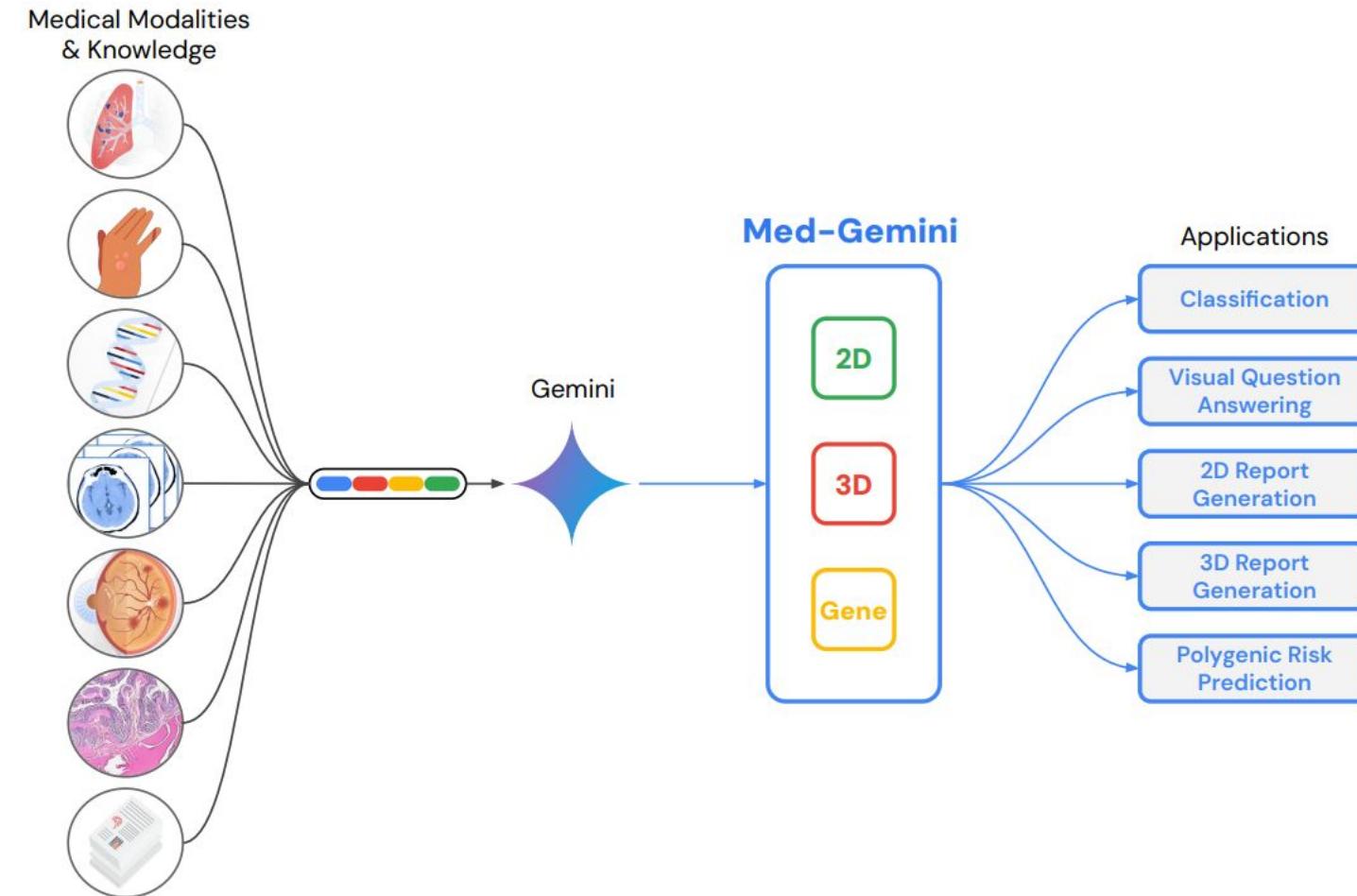
Institute of Health Informatics, University College London

24 November 2025

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

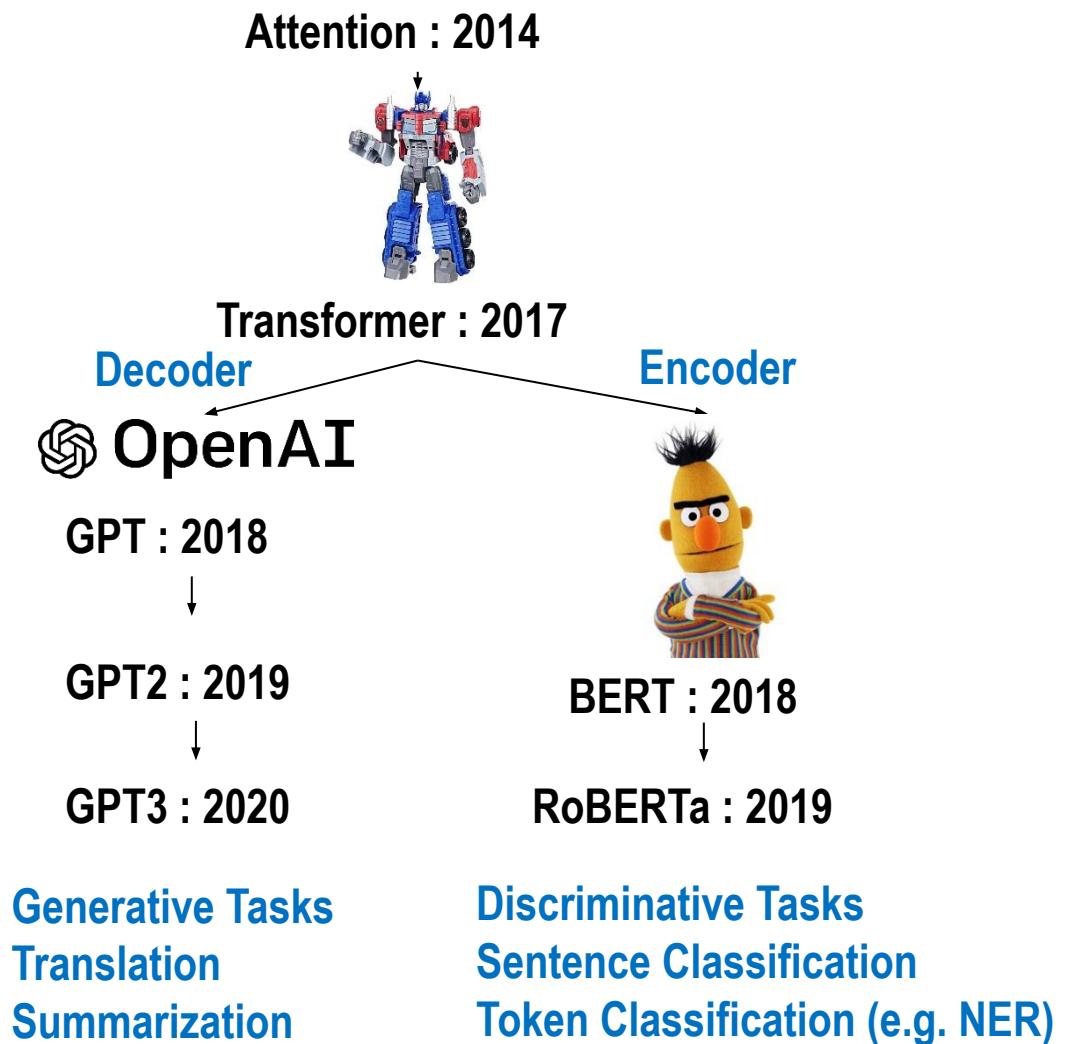


Transformer revolutionized Natural Language Processing (NLP) with pre-training

□ A brief summary of the Transformer models

- Transformer: In 2017, Google revolutionized NLP with the transformer architecture using self-attention mechanisms, enabling parallel processing and **improved performance**.
- GPT (Generative Pre-trained Transformer): OpenAI released the first GPT model in 2018, employing unsupervised learning on vast amounts of text data. It showed the potential of large-scale pre-training.
- BERT (Bidirectional Encoder Representations from Transformers): Google introduced BERT in 2018, utilizing bidirectional training to capture deeper contextual relationships. The 1st model to outperform humans.
- Open AI started to focus on training a larger model and referred their **GPT2 as a large language model**, and it generated text samples of unprecedented quality

Model	Model Size	Training Corpus Size
BERT	Base : 110M Large : 330 M	16GB Books and Wikipedia
RoBERTa	Base : 123M Large : 353 M	160GB Web Crawl
GPT	117M	4.6GB BookCorpus
GPT2	1,500M	40GB Filtered Crawl
GPT3	175,000M	753GB Crawl, Books, Wiki



Transformer architecture uses self-attention to capture contextual relationships between words.

Attention Is All You Need

Ashish Vaswani*
Google Brain
avaswani@google.com

Noam Shazeer*
Google Brain
noam@google.com

Niki Parmar*
Google Research
nikip@google.com

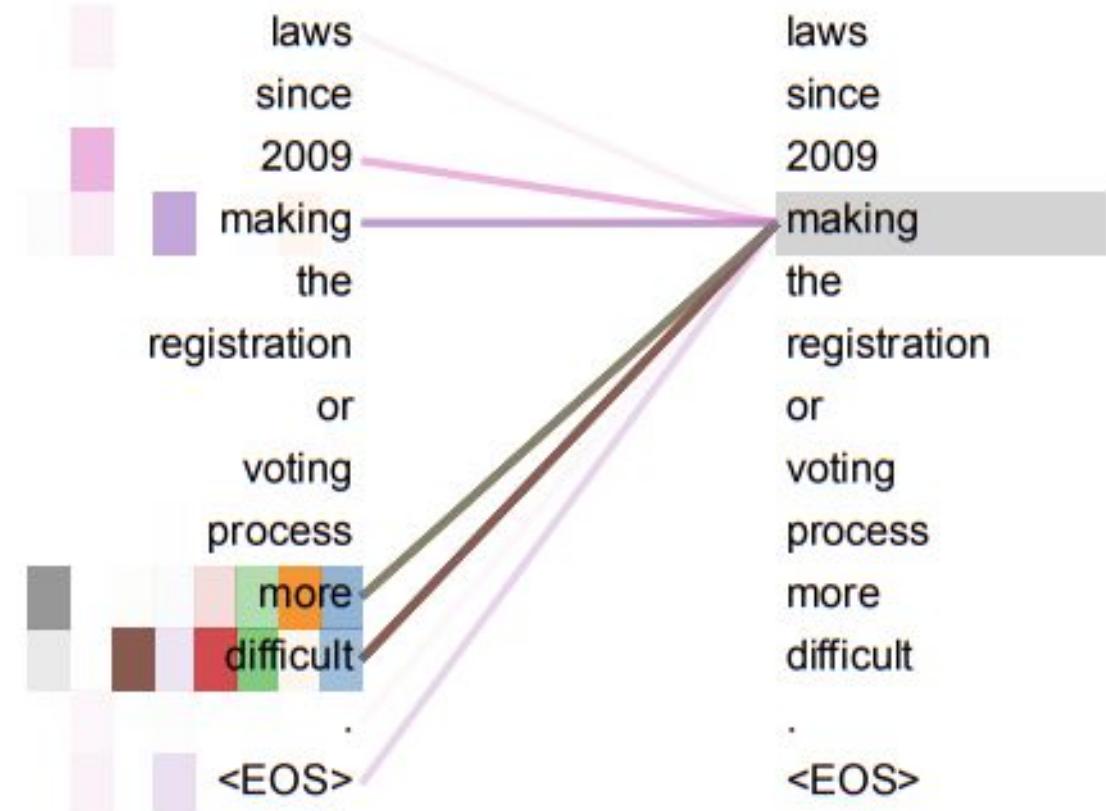
Jakob Uszkoreit*
Google Research
usz@google.com

Llion Jones*
Google Research
llion@google.com

Aidan N. Gomez* †
University of Toronto
aidan@cs.toronto.edu

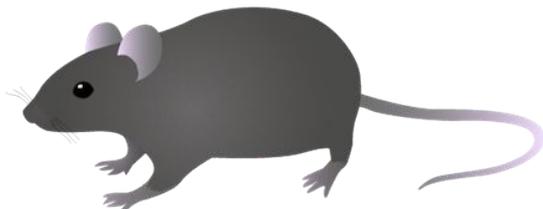
Łukasz Kaiser*
Google Brain
lukaszkaiser@google.com

Illia Polosukhin* ‡
illia.polosukhin@gmail.com

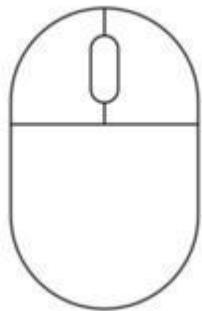


Attention is similar to “Search” function or data retrieval for the dictionary data type.

- Return **Value** if **Query** = **Key**



Query: Mouse



Key	Value	Similarity
Human	<i>Homo sapiens</i>	0
Mouse	<i>Mus musculus</i>	1
...
Round worm	<i>Caenorhabditis elegans</i>	0
Fruit Fly	<i>Drosophila melanogaster</i>	0

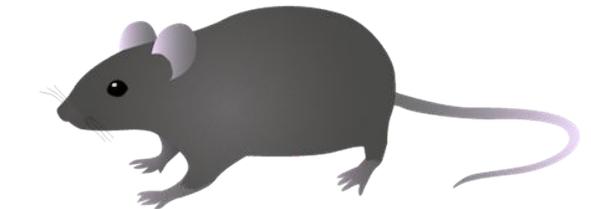
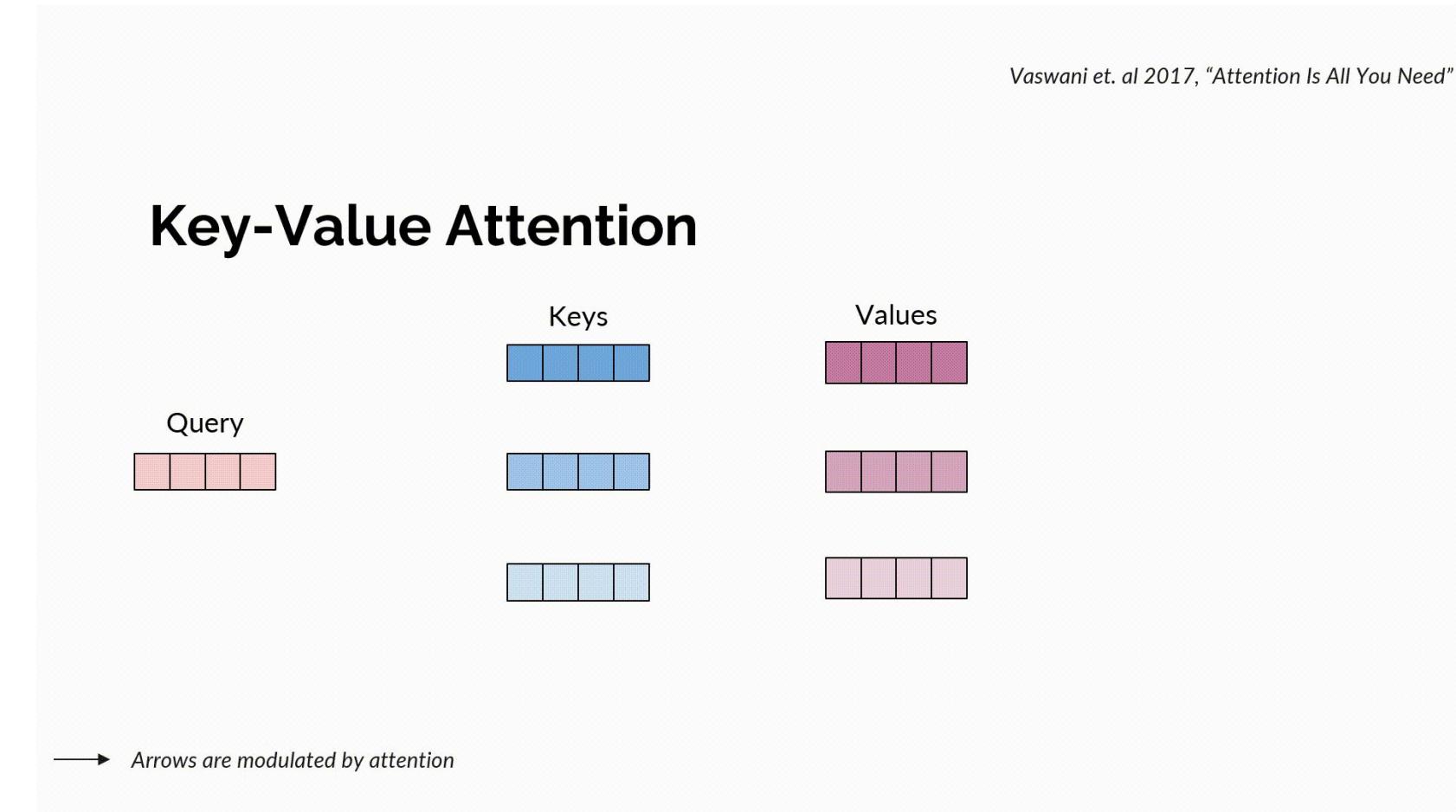
$$result = \sum_i similarity(key, query) * value$$



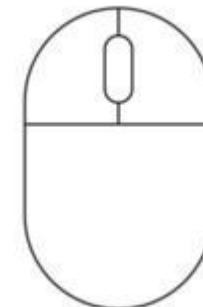
$$A(q, K, V) = \sum_i softmax(f(K, q)) V$$

For self-attention, attention tells us where to look through within the text based on the query word.

▀ Better contextual learning

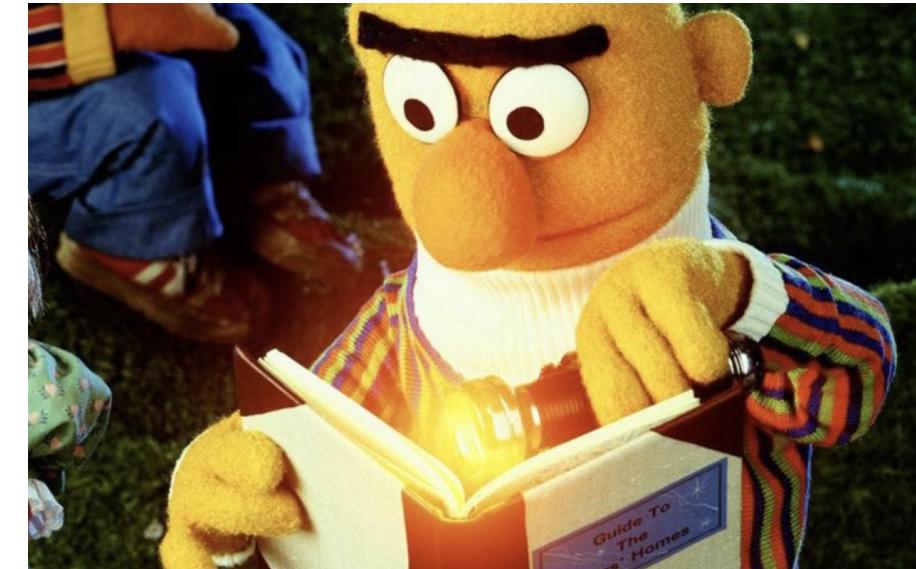
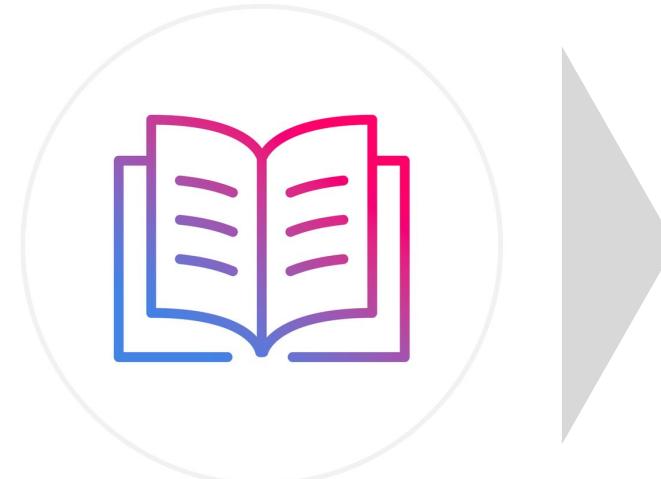


Will be able to
Differentiate



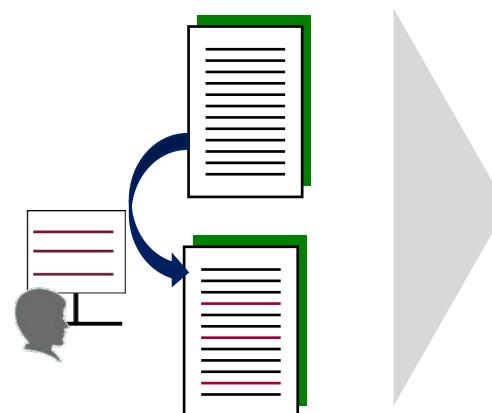
Pre-training of transformer architecture quickly reached the state of the art (SOTA) performance.

Pre-training
Unsupervised learning
Unlabeled large (> million words) corpus of text data (books, crawled data)



Supervised Finetuning

A small number (~10,000 examples) of labeled data for specific task



Cell type Species Gene/Protein DNA Drug/Chemical Disease

Autophagy maintains **tumour** growth through circulating **arginine**. Autophagy captures intracellular components and delivers them to lysosomes, where they are degraded and recycled to sustain metabolism and to enable survival during starvation¹⁻⁵. Acute, whole-body deletion of the essential autophagy gene **Atg7** in adult **mice** causes a systemic metabolic defect that manifests as **starvation intolerance** and gradual loss of white adipose tissue, liver glycogen and muscle mass¹. **Cancer cells** also benefit from autophagy.

Biomedical entity recognition

Large scale pre-training enabled the GPT model to show enhanced few-shot and zero-shot learning

Language Models are Unsupervised Multitask Learners										
	LAMBADA (PPL)	LAMBADA (ACC)	CBT-CN (ACC)	CBT-NE (ACC)	WikiText2 (PPL)	PTB (PPL)	enwik8 (BPB)	text8 (BPC)	WikiText103 (PPL)	1BW (PPL)
SOTA	99.8	59.23	85.7	82.3	39.14	46.54	0.99	1.08	18.3	21.8
117M	35.13	45.99	87.65	83.4	29.41	65.85	1.16	1.17	37.50	75.20
345M	15.60	55.48	92.35	87.1	22.76	47.33	1.01	1.06	26.37	55.72
762M	10.87	60.12	93.45	88.0	19.93	40.31	0.97	1.02	22.05	44.575
1542M	8.63	63.24	93.30	89.05	18.34	35.76	0.93	0.98	17.48	42.16

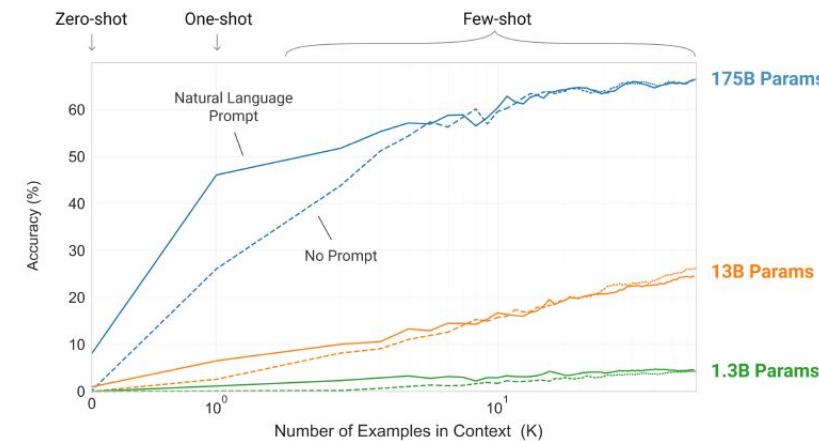
Table 3. Zero-shot results on many datasets. No training or fine-tuning was performed for any of these results. PTB and WikiText-2 results are from (Gong et al., 2018). CBT results are from (Bajgar et al., 2016). LAMBADA accuracy result is from (Hoang et al., 2018) and LAMBADA perplexity result is from (Grave et al., 2016). Other results are from (Dai et al., 2019).

GPT2

\$ 0.5 M

OpenAI did not release GPT2 model, only smaller version was released

GPT3 model was available to the public as an API



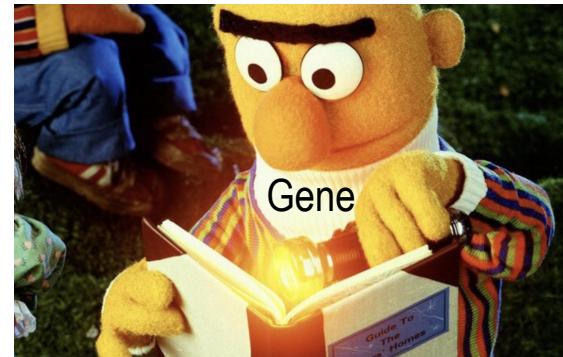
GPT3

\$ 5 M

Few-shot learning enabled better usability of the model in real practice

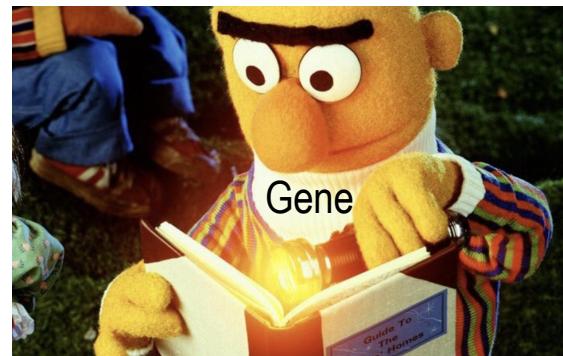
Few-shot learning

1 ~ 10 examples shown



Zero-shot learning

0 examples are shown



Acetaminophen – Drug
Homo sapiens – Species
Escherichia coli – ?

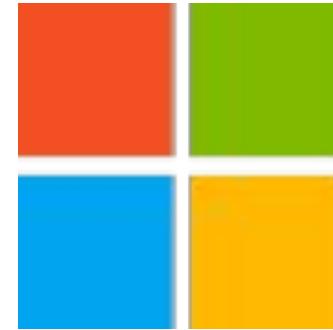


Escherichia coli - ?



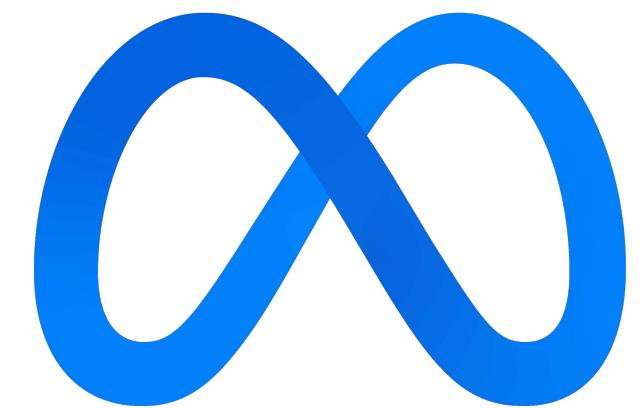
Proprietary models

The privacy concerns with the patient data



The privacy concerns with the patient data

Open-source LLMs
Llama and Phi



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 C₈H₉NO₂, 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 (C₈H₉NO₂):

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.

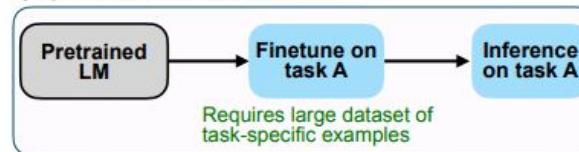
There seems to be a slight discrepancy in the provided information.

According to the formula (C₈H₉NO₂), 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].

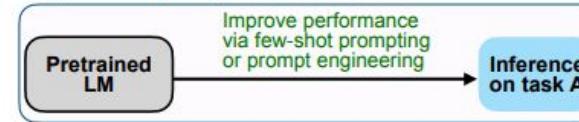
I. Background - Prompting vs Instruction tuning

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

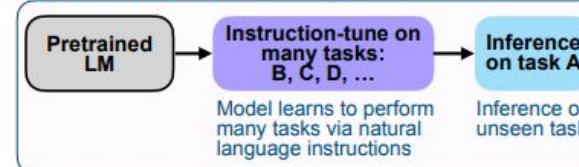
(A) Pretrain-finetune



(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

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.

Sentiment analysis tasks

Coreference resolution tasks

...

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

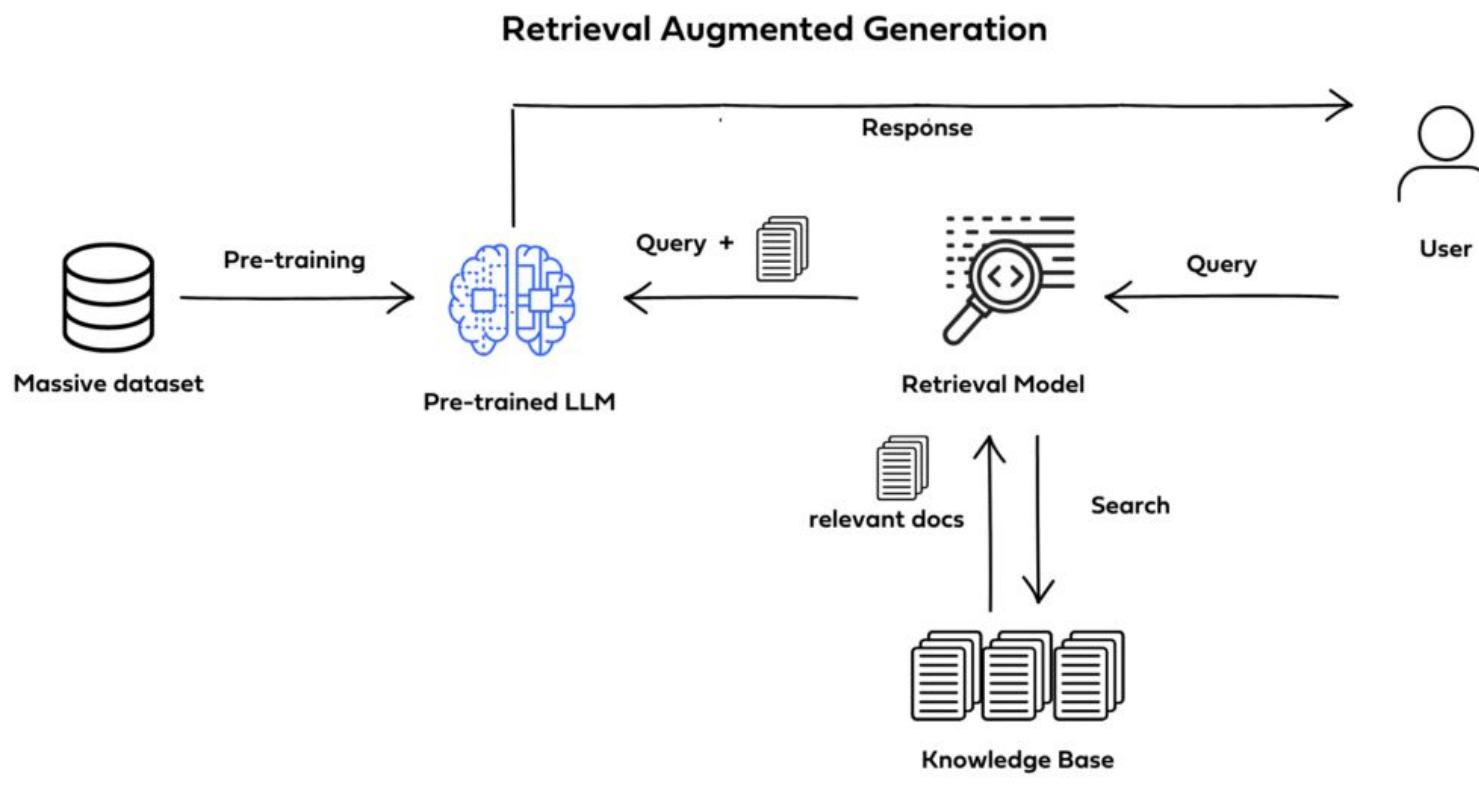
It is not possible to tell

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

I. Background - Prompting vs Instruction tuning

FLAN: Trained with 62 Datasets for 12 tasks.

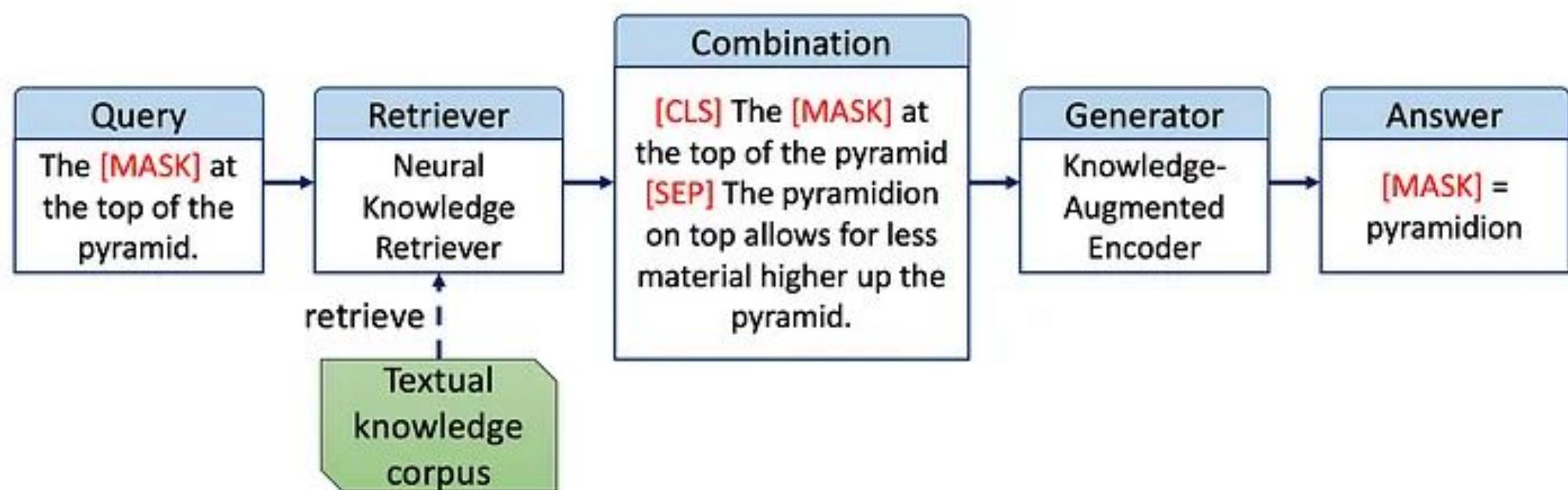
Natural language inference (7 datasets) ANLI (R1-R3) CB MNLI QNLI RTE SNLI WNLI	Commonsense (4 datasets) CoPA HellaSwag PiQA StoryCloze	Sentiment (4 datasets) IMDB Sent140 SST-2 Yelp	Paraphrase (4 datasets) MRPC QQP PAWS STS-B	Closed-book QA (3 datasets) ARC (easy/chal.) NQ TQA	Struct to text (4 datasets)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



RAG essentially couples information retrieval mechanisms with text generation models.

1. **Vector database:** external knowledge base and vectorization
2. **User query:** a question or statement that needs to be answered
3. **Retrieval component:** a retriever that scans the vector database and get relevant pieces to provide additional context to LLM
4. **Concatenation:** adding the retrieved documents into prompts
5. **Text generation:** produce the final output with augmented prompts

For example



Advantage of RAG:

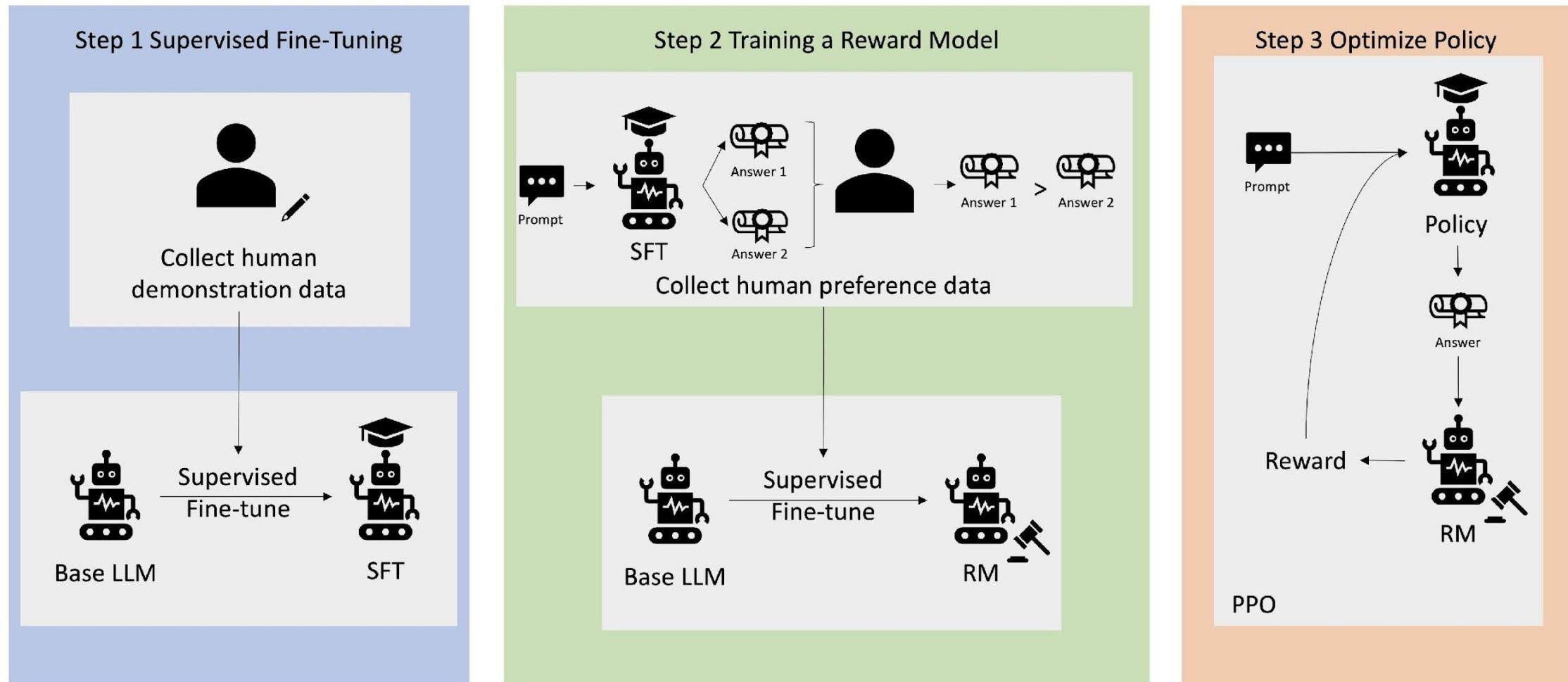
- Minimizes hallucinations
- Up-to-date knowledge
- Domain-specific responses
- Interpretable
- Cost effective

Disadvantage of RAG:

- Complexity on large-scale database
- Potential bias
- Handling ambiguity

I. Background – Alignment Tuning (post-training)

Aligns the LLM to respond in the preferred way. Often done with reinforcement learning.



I. Background – post-training approaches

Approach	Purpose	Key Concepts	🔗
SFT	Teach tasks, supervision	Instruction tuning, next-token loss	
RLHF	Align with human preferences	Reward model, PPO, rankings	
DPO-style	Align without RL	Preference pairs, stable training	
RLAIF	Scale alignment with AI judges	Synthetic preferences, low cost	
Continual / Domain Adaptation	Specialize or update	LoRA, PEFT, catastrophic forgetting	
Safety Training	Reduce harm	Refusal policies, red-teaming	
Tool-use	Enable external capabilities	Function calling, RAG	
MoE fine-tuning	Efficient scaling	Sparse experts, routing	

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-based thyroxine (T4) test is utilized to evaluate the activity of the thyroid gland by measuring the levels of T4 hormone it secretes. An overactive thyroid, or hyperthyroidism, is characterized by the excessive production of thyroid hormones such as T4. Thus, this test is especially pertinent for the diagnosis and management of hyperthyroidism.
Explanation B	The thyroxine (T4) test using radioactive isotopes is a diagnostic tool used to assess thyroid function. Thyroxine is a hormone produced by the thyroid gland, and its levels can be indicative of thyroid health. Hyperthyroidism is a condition where the thyroid gland is overactive and produces excessive amounts of thyroid hormones, including T4. Therefore, the test is particularly relevant for diagnosing and monitoring hyperthyroidism.

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
GPT-3.5
GPT-4
<u>Llama-2</u>
<u>Llama-2</u>
<u>Llama-3</u>
<u>Llama-3</u>
<u>LLaMA 4 (Sco</u>
<u>LLaMA 4 (Mave</u>
<u>Llama 4 (Behav</u>

Llama 4 Scout

Class-leading natively multimodal model that offers superior text and visual intelligence, single H100 GPU efficiency, and a 10M context window for seamless long document analysis.

Download



-Shot

Open Source GPT
OPT : 2022.05

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-4 : 2025.04

Size : 109B-2T(still training)

Native Multimodal / Multilingual / Reasoning

I. Background - Phi

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



Invested \$1B in OpenAI : 2019.07

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

Exclusive Right for OpenAI's model

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

↓ Trained their own model

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	78.9	81.4
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
Phi-2	2.8B	2023.12.	1.4T Tokens	30.9	53.4
MedPhi-2 (Our Own)	2.8B	2024.02.	110M Tokens	34.3	56.1
Phi-3	14B	2024.05.	4.8T Tokens	64.2	77.2
KnowMedPhi-3 (Our Own)	14B	2024.07.	126M Tokens	67.2	77.9
Phi-3	3.8B	2024.04.	3.3T Tokens	52.2	72.3
KnowMedPhi-3 (Our Own)	3.8B	2024.07.	126M Tokens	50.6	75.0
Phi-4	14B	2024.12	-	-	-

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

Phi-4 : 2024.12,

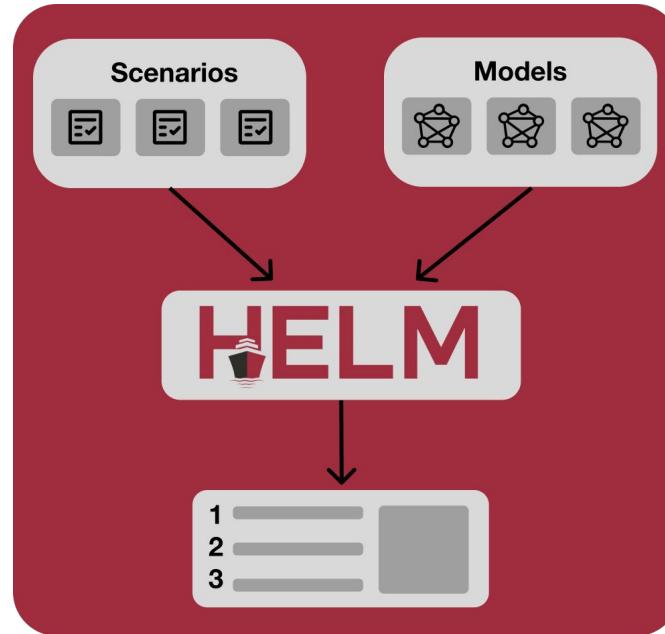
Size : 14B

reasoning/multimodal models in 2025



<https://chat.qwen.ai/>

Model	Size	Release Date	Pre-train Data Size	MedQA (official)	MedQA (community)
Qwen-1.0	8B-72B	Apr 2023	~2.2T tokens	Not published	None
Qwen-1.5 (7B/14B/72B)	0.5B–110B	Dec 2023	~3T tokens	72B: 48.9%	~50–55%
Qwen-2.0 (7B/57B/72B)	0.5B–72B	Jun 2024	~7T tokens	Not published	~60–65%
Qwen-2.0 Coder	0.5B–32B	Jun 2024	~7T tokens	Not published	~45–55%
Qwen-2.5 (0.5B–72B)	0.5B–72B	Nov 2024	9–15T tokens	Not published (medical exam only)	~68–72%
Qwen3-235B-A 22B (MoE)	235 Billion (22B active)	April 2025	Up to 36 Trillion		



Model	Mean win rate
DeepSeek R1	0.662 ↗
o3-mini (2025-01-31)	0.66 ↗
Claude 3.5 Sonnet (20241022)	0.632 ↗
Claude 3.7 Sonnet (20250219)	0.628 ↗
GPT-4o (2024-05-13)	0.579 ↗
Gemini 2.0 Flash	0.431 ↗
GPT-4o mini (2024-07-18)	0.372 ↗
Llama 3.3 Instruct (70B)	0.301 ↗
Gemini 1.5 Pro (001)	0.235 ↗

SEE MORE

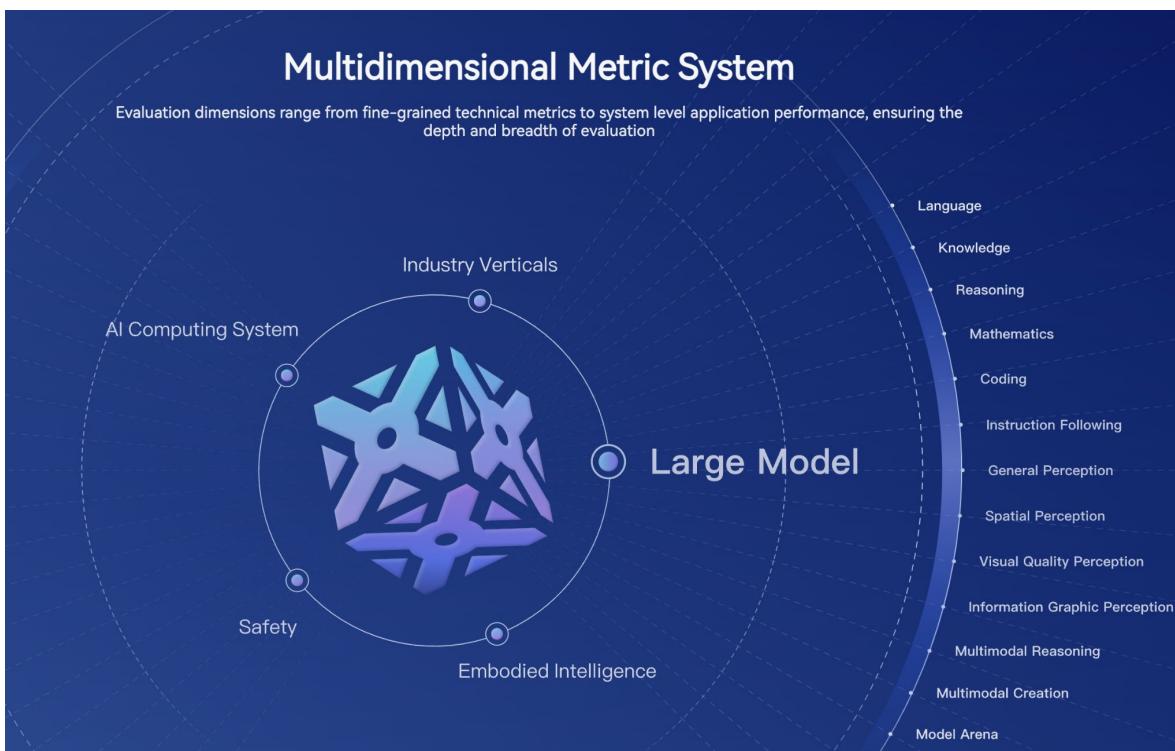
Holistic Evaluation of Large Language Models for Medical Tasks

MedHELM comprises a structured taxonomy with

- 5 categories,
- 22 subcategories
- 121 distinct clinical tasks
- 35 distinct benchmarks (14 private, 7 gated-access, and 14 public)



led by Shanghai Artificial Intelligence Laboratory



<https://opencompass.org.cn/home>

CompassBench Leaderboard		
	25-10 ▾	All ▾
Overall Score		
1	GPT-5-2025-08-07 (Thinking)	71.0
New	OpenAI	API
2	o3-2025-04-16 (high)	69.8
New	OpenAI	API
3	GLM-4.6	69.3
New	Zhipu AI	Weights
4	Qwen3-235B-A22B-Thinking-2507	68.9
▲ 1	Alibaba	Weights
5	Claude Sonnet 4.5 (Thinking)	68.4
New	Anthropic	API
6	o4-mini-2025-04-16 (high)	68.3
New	OpenAI	API
7	iFlytek-Spark-X1preview-Think-1025	67.9
New	iFLYTEK	API
8	DeepSeek-R1-0528	67.5
▼ 2	DeepSeek	Weights
9	Doubao-Seed-1.6-251015 (Thinking)	67.3
New	ByteDance	API
10	Grok-4	66.8
▼ 6	xAI	API

Error Detection and Fact Checking with Knowledge Graph

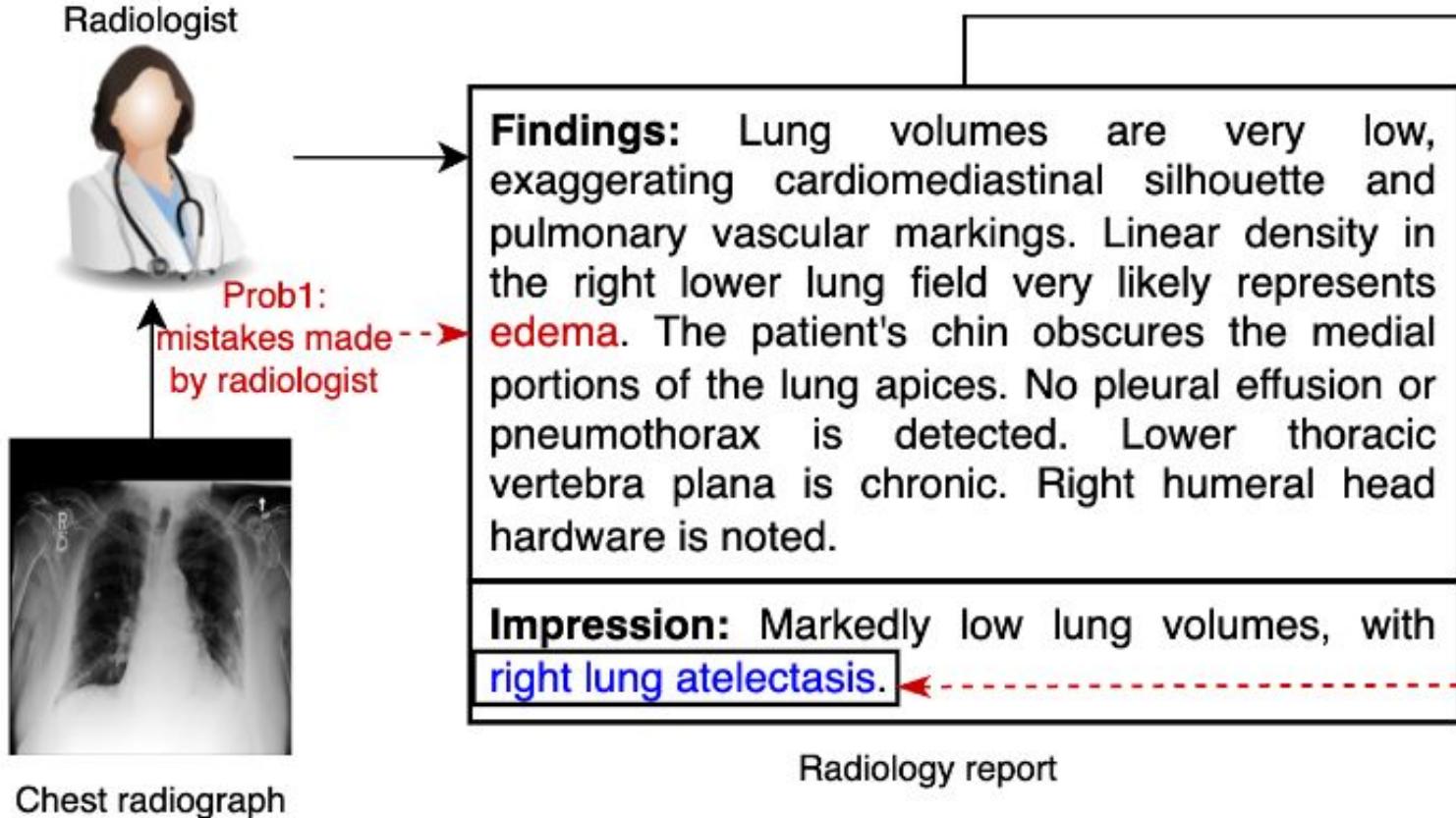
Honghan Wu

School of Health and Wellbeing, University of Glasgow

Institute of Health Informatics, University College London

24 November 2025

Motivation: the need for error detection in radiology



Prob 2: Inconsistent impression with findings



FINDINGS:
Possible right pleural effusion.
There is atelectasis at the right upper lung...

IMPRESSION:
More atelectasis at the right lower leg than previous radiograph.

Error detection



FINDINGS:
Possible **{right pleural effusion}**.
There is atelectasis at the **{right upper}** lung...

IMPRESSION:
More atelectasis at the **{right lower leg}** than previous radiograph.

Error correction



FINDINGS:
Possible **left pleural effusion**.
There is atelectasis at the **left lower** lung...

IMPRESSION:
More atelectasis at the **left lower lung** than previous radiograph.

Image-conditioned medical
autocorrection

MedAutoCorrect

Image-conditioned autocorrection in medical reporting

Arnold Caleb Asiimwe¹, Dídac Surís Coll-Vinent^{1,2}, Pranav Rajpurkar³, Carl Vondrick¹,

¹Columbia University, ²Meta AI, ³Harvard Medical School

Machine Learning for Health Conference (MLHC) 2024



Knowledge Graph

A W3C's vision – Semantic Web



The term “Semantic Web” refers to W3C’s vision of the **Web of linked data**.

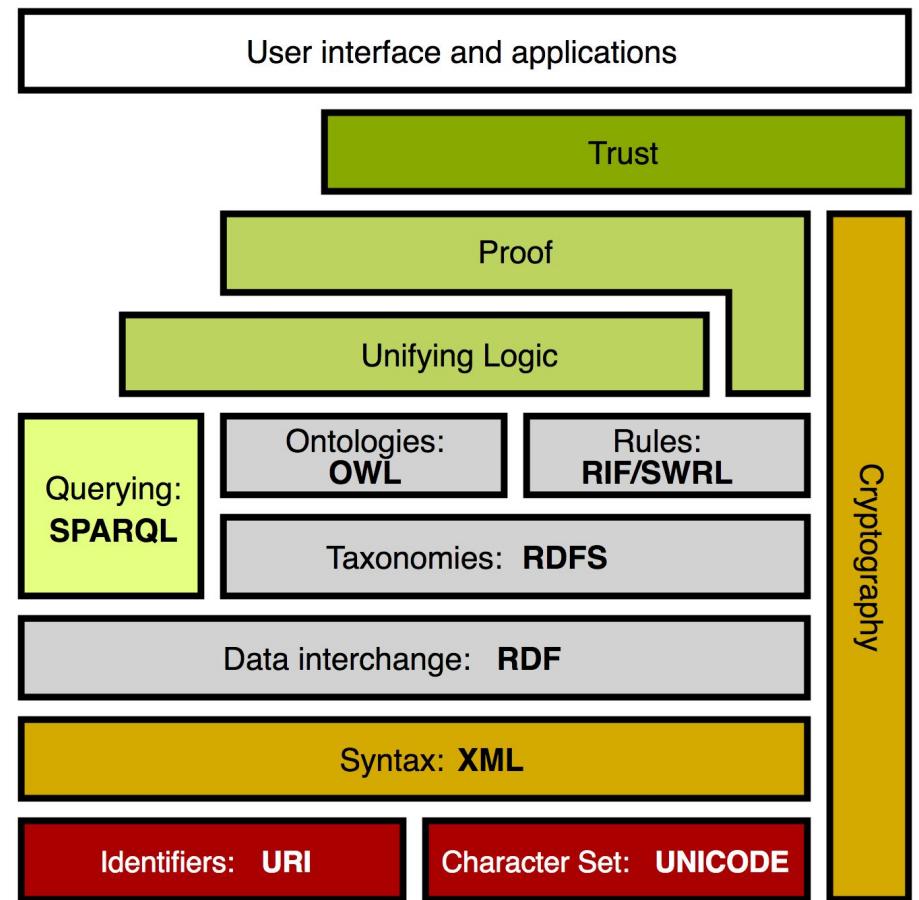
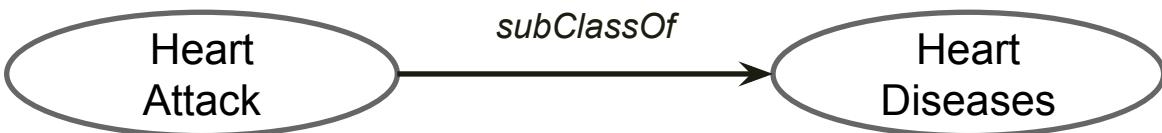
The ultimate goal of the Web of data is to enable computers to **do more useful work and to develop systems that can support trusted interactions** over the network.

Tim Berners-Lee: The next Web of open, linked data

318,726 views • 13 Mar 2009

Knowledge Representation at the core of Semantic Web

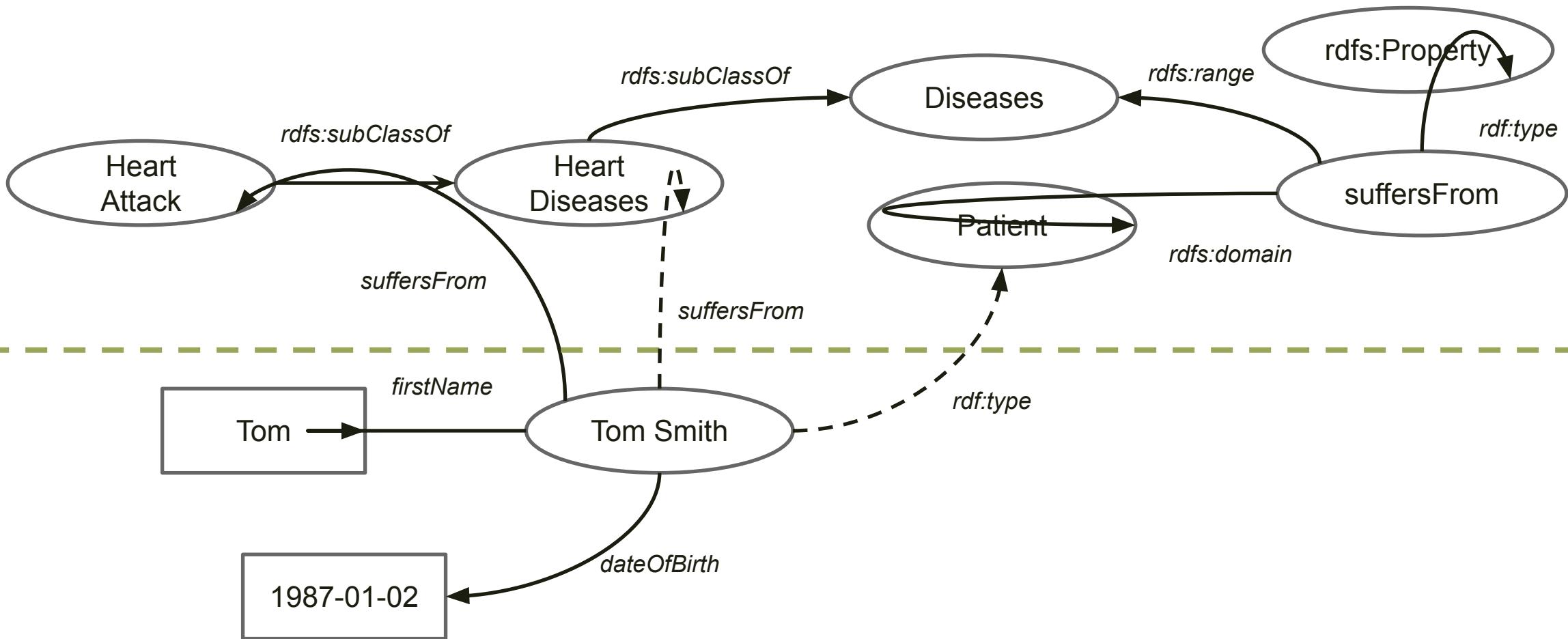
- Ontology: Describe the concepts and their relationships in a particular domain
- Computation: reasoning (set / rule)



Semantic Web Technology Stack

Knowledge Graph - a simple example

T-box



A-box

How can it help? - an example using semantic query

T-Box

(Valproic acid)
A1 - 13965000

(Myocardial infarction)
A2 - HP_0001658

(Myocardial infarction)
A3 - C0027051

subClassOf

(Mood stabilizing drug)
358927005

equivalentClass

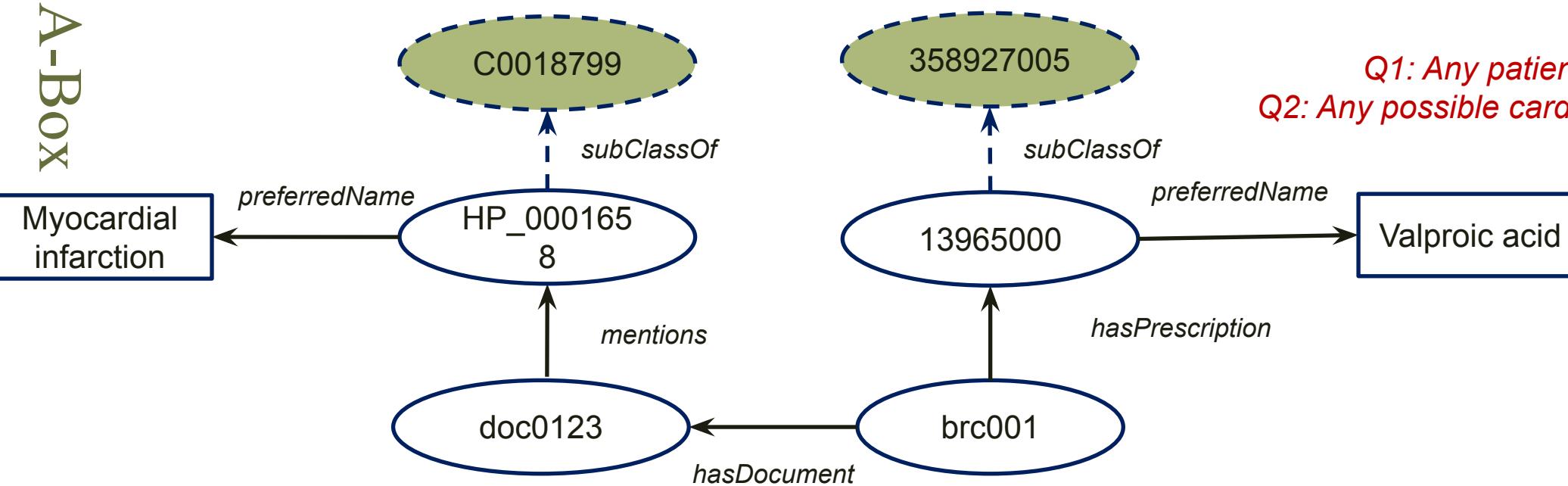
(Myocardial infarction)
UMLS:C0027051

SNOMED CT

NCBO

UMLS

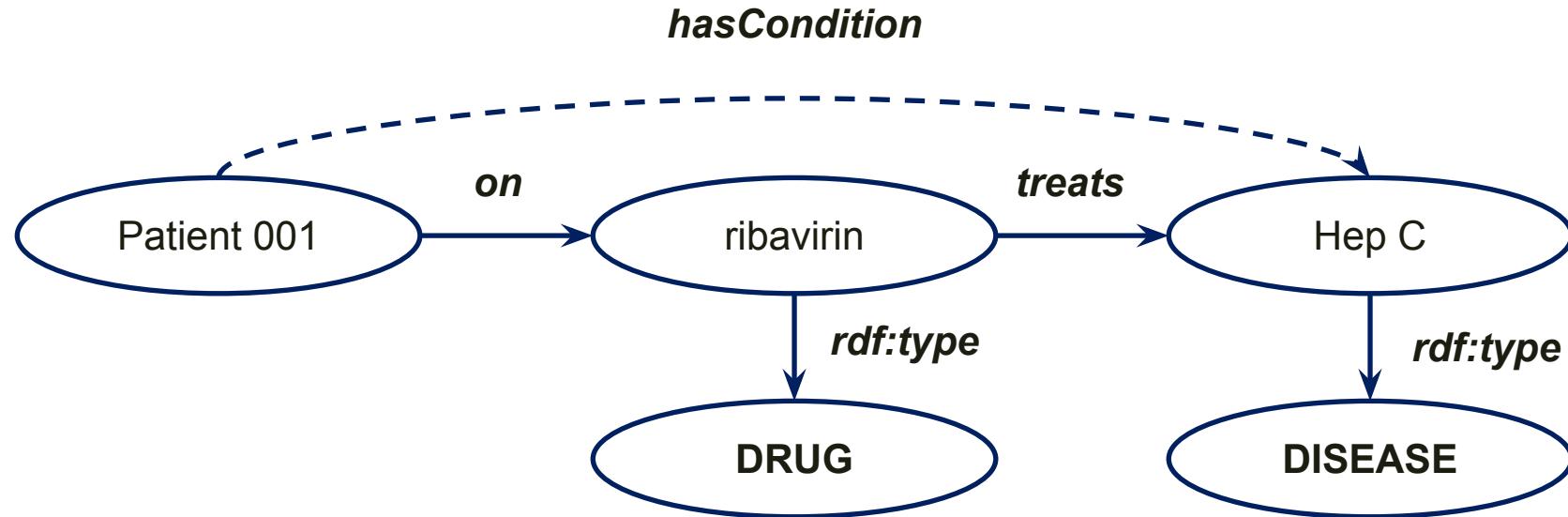
A-Box



Queries:

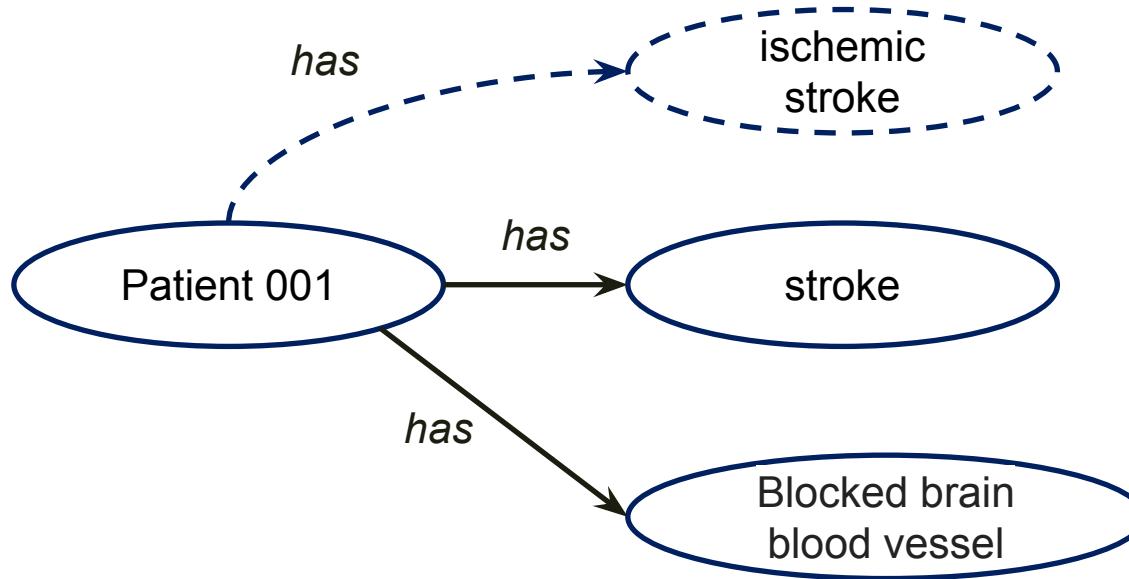
Q1: Any patients on mood stabilizer?
Q2: Any possible cardiovascular side effects
of mood stabilizers?

Beyond subClass inference...



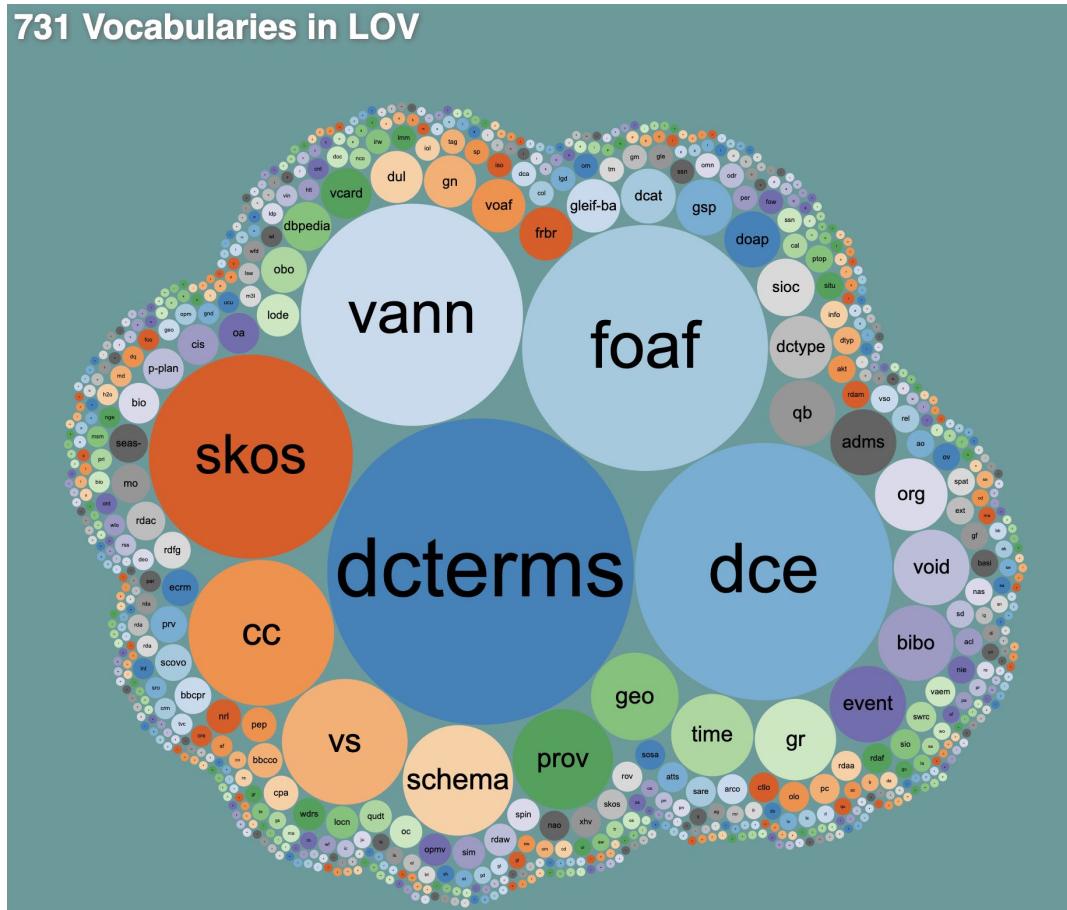
Rule1: if someone is on a DRUG that TREATS a condition, then he/she probably has that condition

Beyond subClass inference...



Rule2: if someone has a STROKE and also BLOCKED BRAIN BLOOD VESSEL, then the stroke is probably ischemic

Linked Open Vocabularies



Open and linked T-Box

<https://lov.linkeddata.es/dataset/lov>

BioPortal

BioPortal Ontologies Search Annotator Recommender Mappings

Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies

Search for a class

Enter a class, e.g. Melanoma

[Advanced search](#)

Find an ontology

Start typing ontology name, then choose from list

[Browse ontologies ▾](#)

Ontology visits (September 2024)

Ontology	Visits (September 2024)
MEDDRA	~22,000
SNOMEDCT	~15,000
RXNORM	~5,000
NDDF	~3,000
EDAM	~1,500

[More](#)

Statistics

Ontologies	1,151
Classes	15,610,783
Properties	36,286
Mappings	100,073,880

Biomedica
I T-Boxes

(as of Oct 2024)

Open A-Boxes

DBpedia

wiki.dbpedia.org

DBpedia is a project aiming to extract structured content from the information created in the Wikipedia project. This structured information is made available on the World Wide Web. [Wikipedia](#)

Operating system: Virtuoso Universal Server

Stable release: DBpedia 2016-10 / 4 July 2017

License: GNU General Public License

Written in: Scala, Java, Virtuoso Universal Server

Developer: Leipzig University, University of Mannheim, Sören Auer, Jens Lehmann, Georgi Kobilarov, Chris Bizer

Written in: Java, PHP, Scala

People also search for

View 10+ more

Wikidata Freebase AWS Google World Bank

Feedback

Wikidata

wikidata.org/wiki/Wikidata:Main_Page

Wikidata is a collaboratively edited multilingual knowledge graph hosted by the Wikimedia Foundation. It is a common source of open data that Wikimedia projects such as Wikipedia, and anyone else, can use under a public domain license. Wikidata is powered by the software Wikibase. [Wikipedia](#)

Founded: 29 October 2012; 8 years ago

Available in: Multiple languages

Editor: Wikidata editors

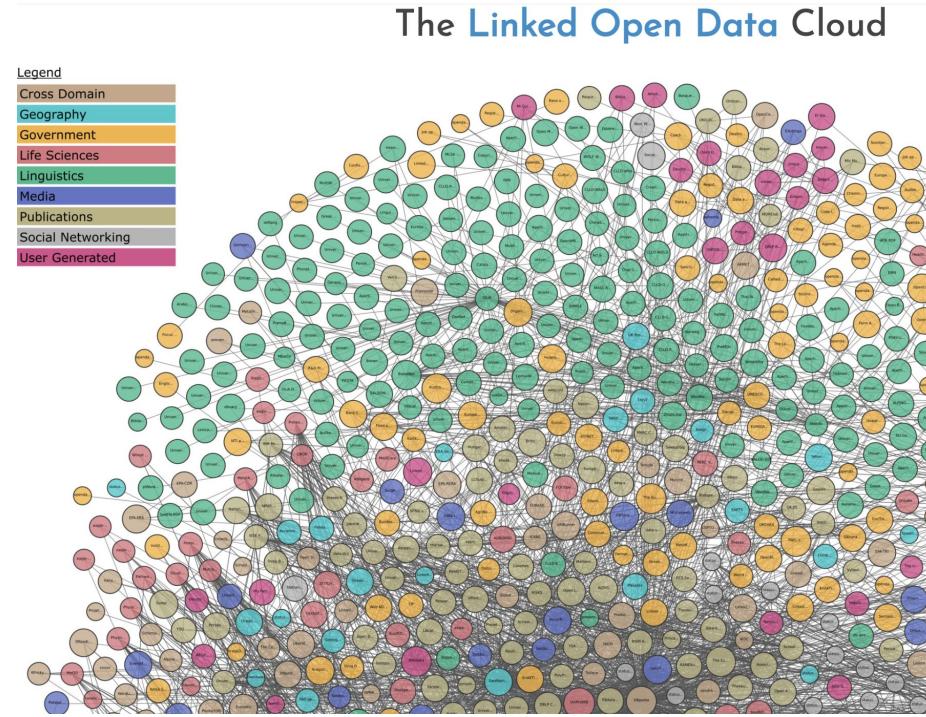
Types of site: Knowledge base, Wiki

People also search for

View 10+ more

Wikidata Freebase AWS Google World Bank

Feedback



The cloud currently contains 1346 datasets (as of Sep 2024)



“Error Correction in Radiology Reports: A Knowledge Distillation-Based Multi-Stage Framework”



Jinge Wu



Zhaolong Wu

Jinge Wu, Zhaolong Wu, Ruizhe Li, Tong Chen, Abul Hasan, Yunsoo Kim, Jason P.Y. Cheung, Teng Zhang, Honghan Wu

Accepted by AAAI 2025 Artificial Intelligence for Social Impact Track.

The knowledge graph from radiology report

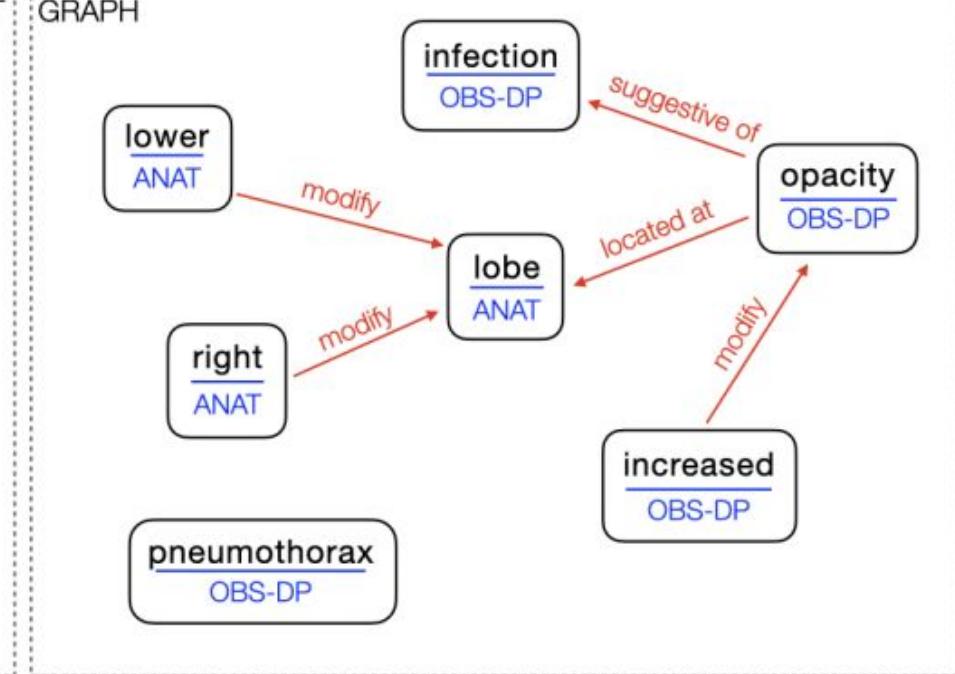
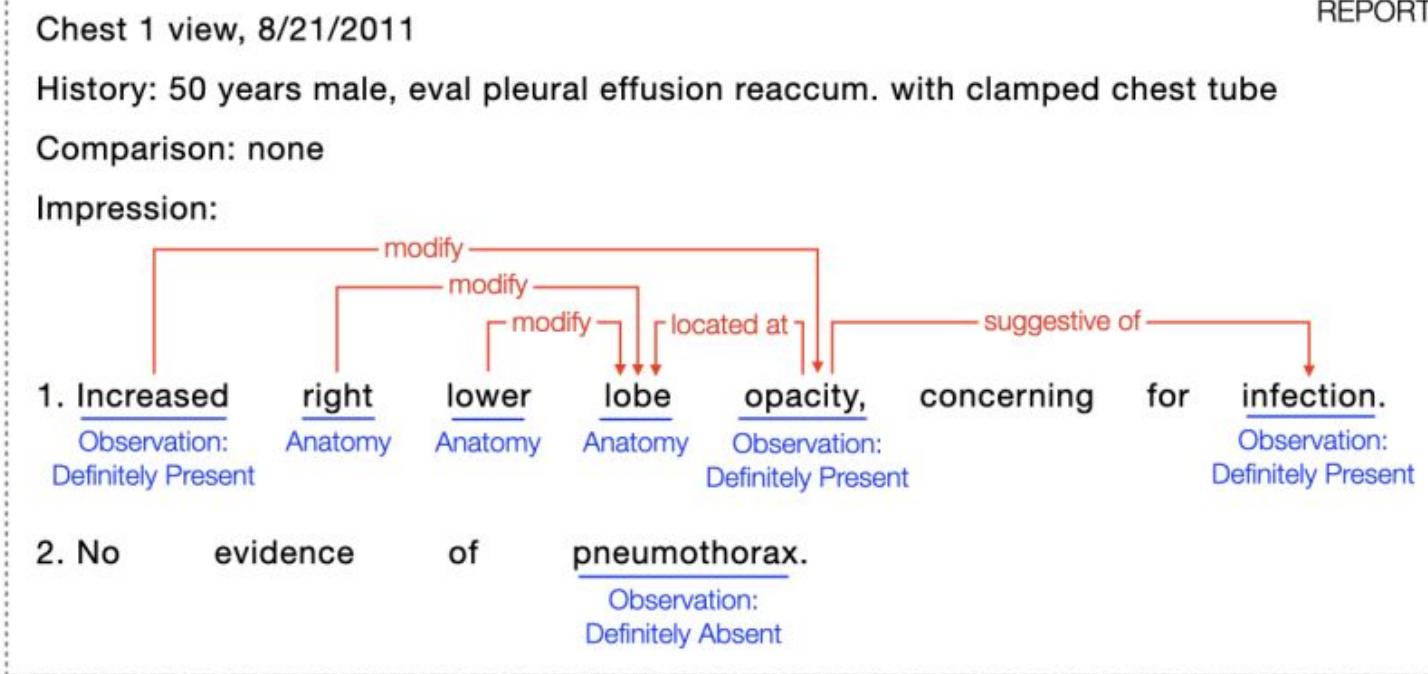
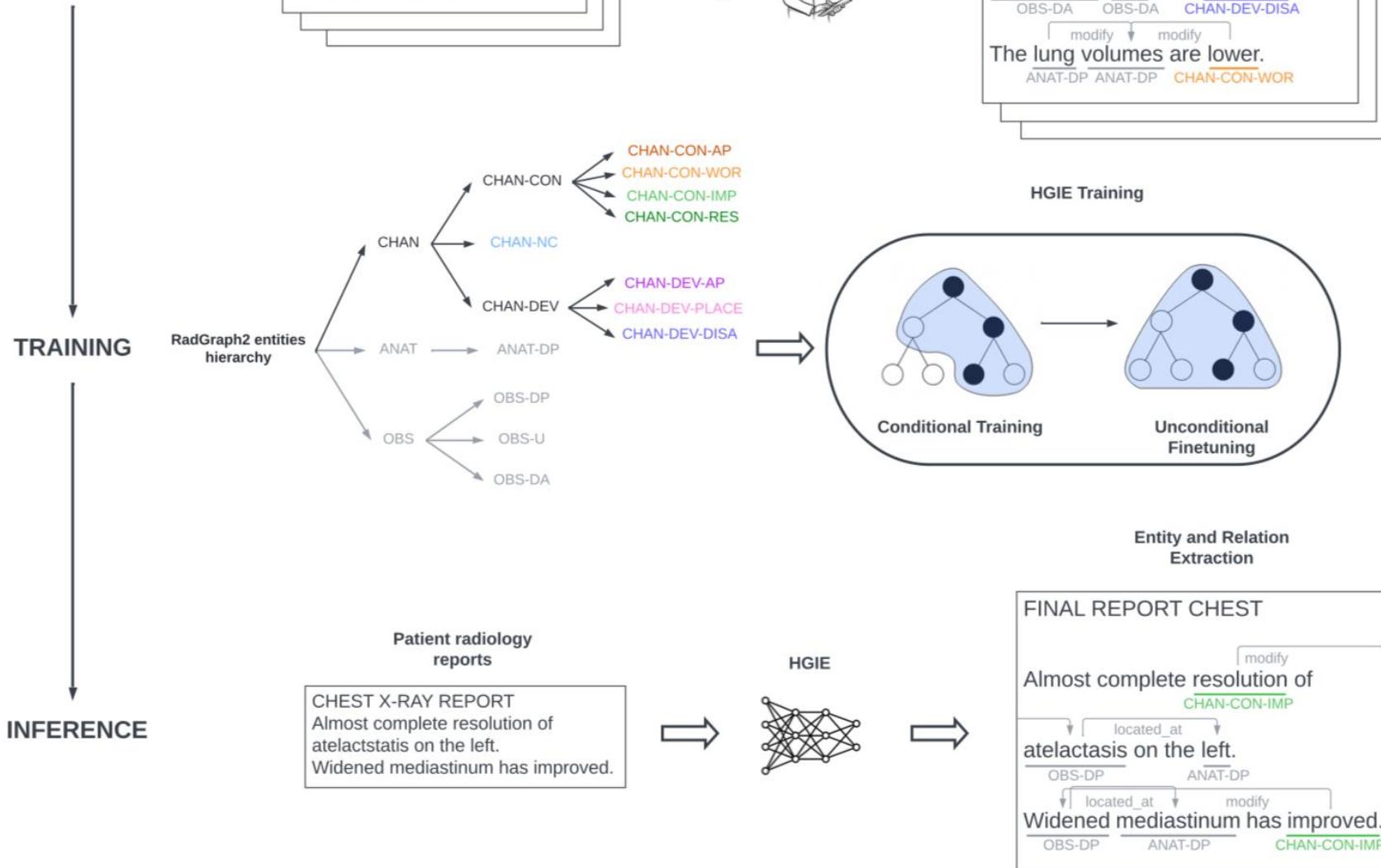


Figure 2: Sample report annotated according to the RadGraph schema (left) and the associated knowledge graph (right).

DATASET ANNOTATION



RadGraph2: Modeling Disease Progression in Radiology Reports via Hierarchical Information Extraction.

Sameer Khanna, et al.

Proceedings of the 8th Machine Learning for Healthcare Conference, PMLR 219:381-402, 2023.



Input report:

FINDINGS: PA and lateral views of the chest are provided. Left lower lobe opacity with volume loss and adjacent congestion. Air bronchogram is noted.

IMPRESSIONS: Consider pneumonia.

Here is the concise summary of the report formatted into natural language:

Left lower lobe opacity with volume loss and congestion

Air bronchogram suggests of pneumonia

Reference Material:

Below are example radiology reports similar to the input text. All these reports are correct and shall be used for reference:

Example 1:

Air bronchogram + Lower lobe → likely consolidation.

Volume loss + Opacity → consider consolidation/atelectasis.

Vascular markings + Hilar → likely congestion

Example 2: [...]

Example 3: [...]

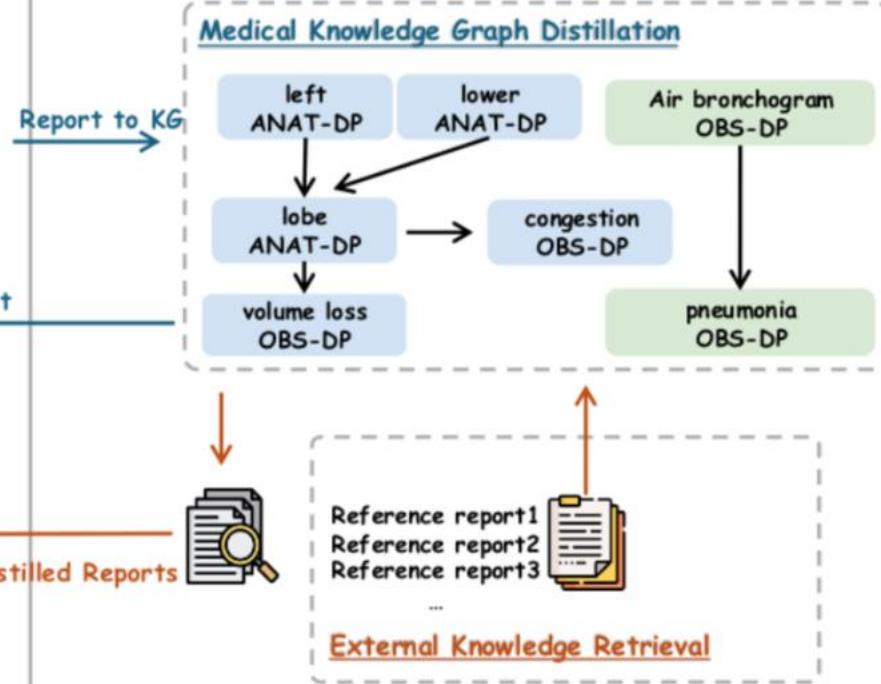


Figure 2: Illustration of our dual-knowledge infusion framework. Left: Input medical report with task description and reference examples. Right: MKGD transforms the report into a structured graph representation capturing anatomical entities (ANAT-DP) and observations (OBS-DP/DA) with their relationships (modify, located_at, suggestive_of), while EXKR provides relevant domain knowledge from reference reports to guide the correction process.

Results - KG retrieval vs baseline

Model	Error Detection (Acc %)		Error Localization (Acc %)		Error Correction (AggNLG)		
	Baseline (Staged)	Our Method (MKGD+EXKR)	Baseline (Staged)	Our Method (MKGD+EXKR)	End-to-End	Staged Base (Baseline)	Our Method (MKGD+EXKR)
<i>Medical Domain Models</i>							
MMedLM2	41.49	73.05 (+31.56)	30.94	46.05 (+15.11)	47.80	47.80	58.27 (+10.47)
Llama3-Aloe	45.31	67.26 (+21.95)	43.34	51.35 (+8.01)	63.33	63.33	90.17 (+26.84)
<i>General Purpose Models</i>							
Phi3-mini	67.26	73.06 (+5.80)	47.71	52.65 (+4.94)	74.36	74.36	74.08 (-0.28)
Phi3-small	79.03	80.21 (+1.18)	63.44	65.04 (+1.60)	80.03	80.03	86.57 (+6.53)
Phi3-medium	73.67	79.04 (+5.37)	69.73	63.44 (-6.29)	84.47	84.47	90.25 (+5.79)
Llama3-8B	37.79	62.27 (+24.48)	37.29	53.14 (+15.85)	84.34	84.34	94.29 (+9.95)
Average	57.43	72.48 (+15.05)	48.74	55.28 (+6.54)	72.39	72.39	82.27 (+9.88)

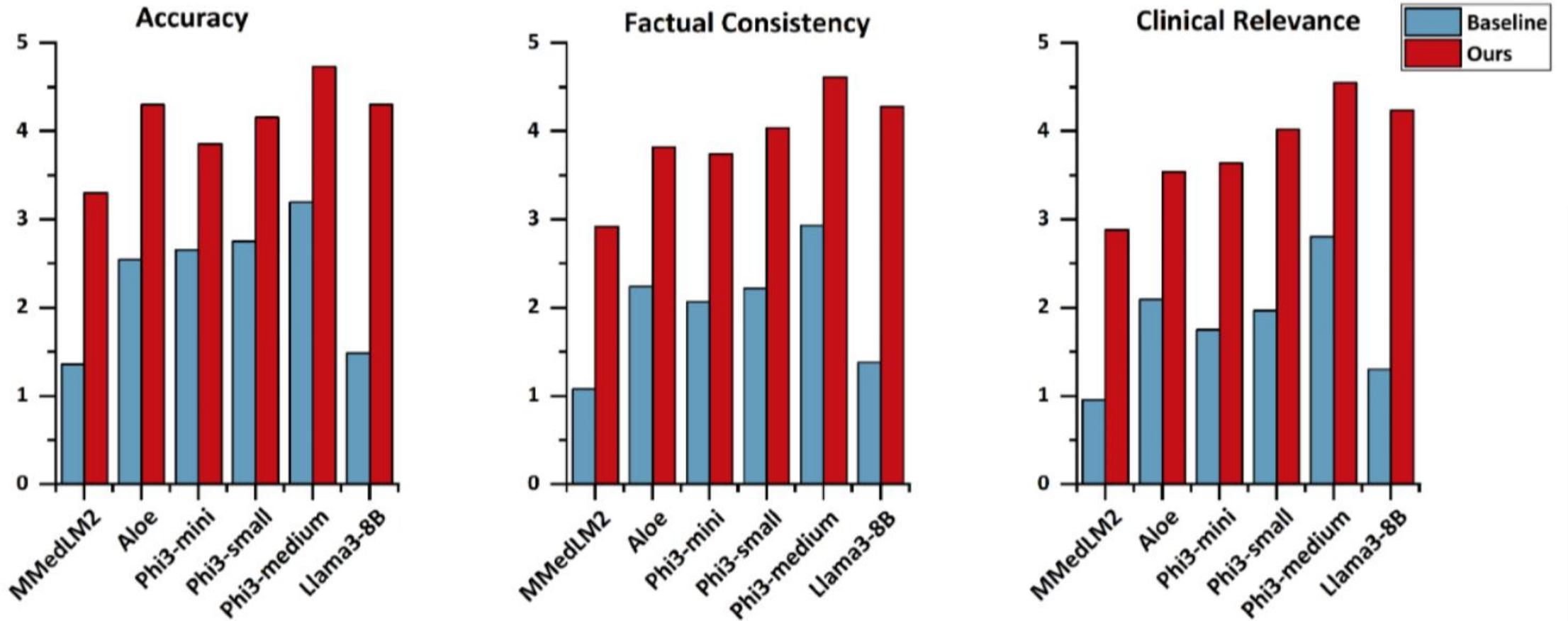


Figure 3: Human evaluation comparison of end-to-end error correction versus our proposed staged proofreading inference with dual-knowledge infusion framework.

Integrating LLMs and Knowledge Graphs for EHRs

Yusuf Abdulle

, King's College London

24 November 2025

Session 3

- S3.I – Introduction to EHRs, the challenges, (briefly) LLMs and KGs
- S3.II - Current research on LLMs and Knowledge Graphs in EHRs
- S3.III - Current research on LLMs in EHRs
- S3. IV - Current research on KGs in EHRs
- S3. V - Current research on LLMs and KGs in EHRs

Session 4

- S4.I – Overview on Multimodal Clinical and Genomic Data through Knowledge Graphs and LLMs
- S4.II – LLMs for Precision Medicine
- S4.III – MLLMs for Precision Medicine

- EHRs are really the narrative backbone of modern healthcare. They have long timelines, repeated encounters, cumulative detail
- They blend structured fields with free text
- The text carries a lot – nuance, uncertainty, reasoning, symptoms, progression thoughts etc
- The structured data carries quantitative values: lab values, codes, medications
- Together, they form a rich but unruly canvas — perfect for models that understand language and concepts

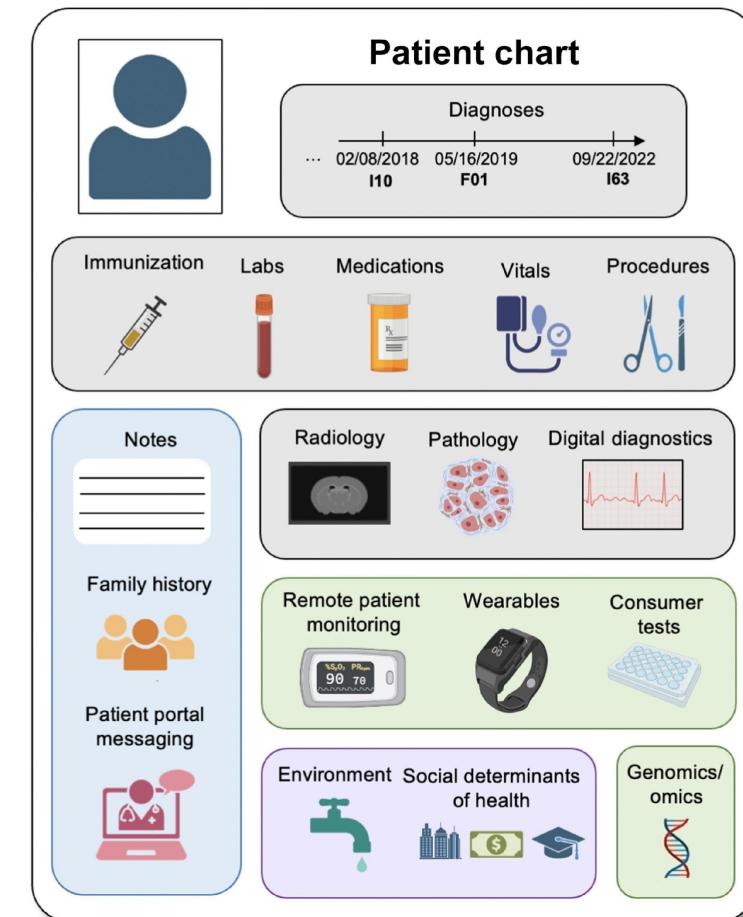
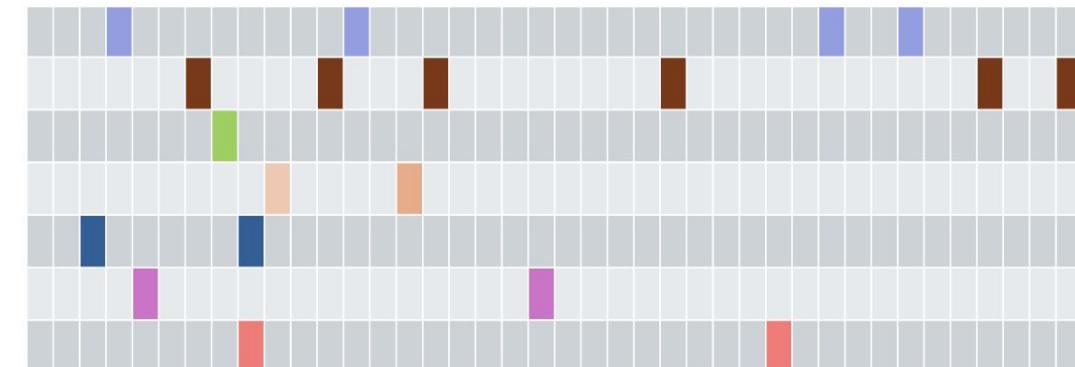


Figure adapted from Singhal, P., et al. "Opportunities and challenges for biomarker discovery using electronic health record data." Trends in molecular medicine 29.9 (2023): 765-776.

- Data is scattered across systems, formats and more
- Key clinical meaning often sits in text, but downstream models expect structured patterns
- Temporal gaps, inconsistent coding and sparse documentation limit purely statistical approaches
- So, what's the heart of the problem: clinical knowledge is implicit, distributed, and rarely codified in the EHR itself

DIAGNOSIS
CODES
LAB
RESULTS
MEDICATI
ONS
TEXT
NOTES
VITAL
SIGNS
IMAGI
NG
PROCEDU
RES



- LLMs can process clinical text
- They can extract entities, summarise narratives, interpret reasoning and anticipate clinical needs
- Unlike classical NLP models, LLMs capture patterns of clinical thought embedded in notes
- LLMs provide a linguistic analysis of free-text EHR, but they still lack explicit medical grounding

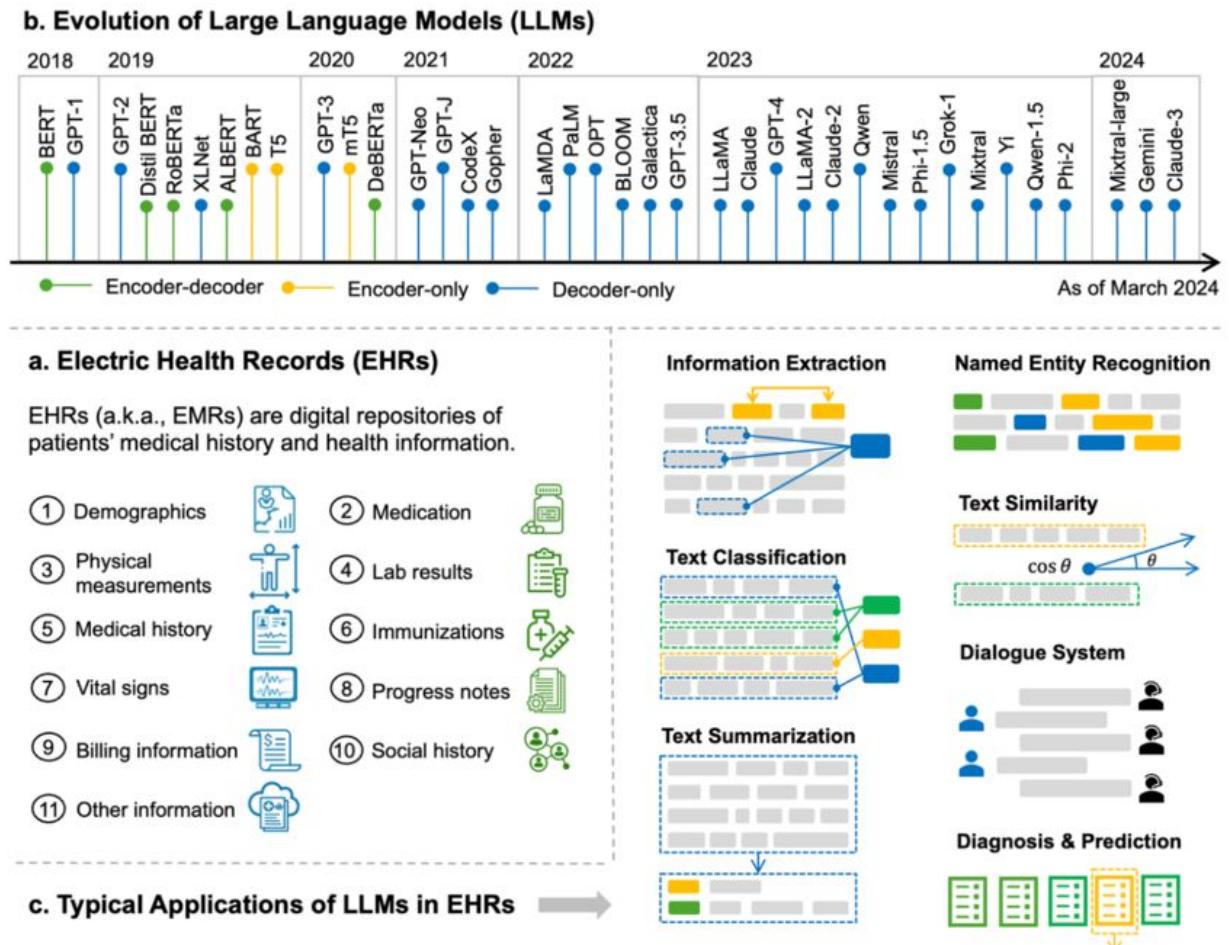


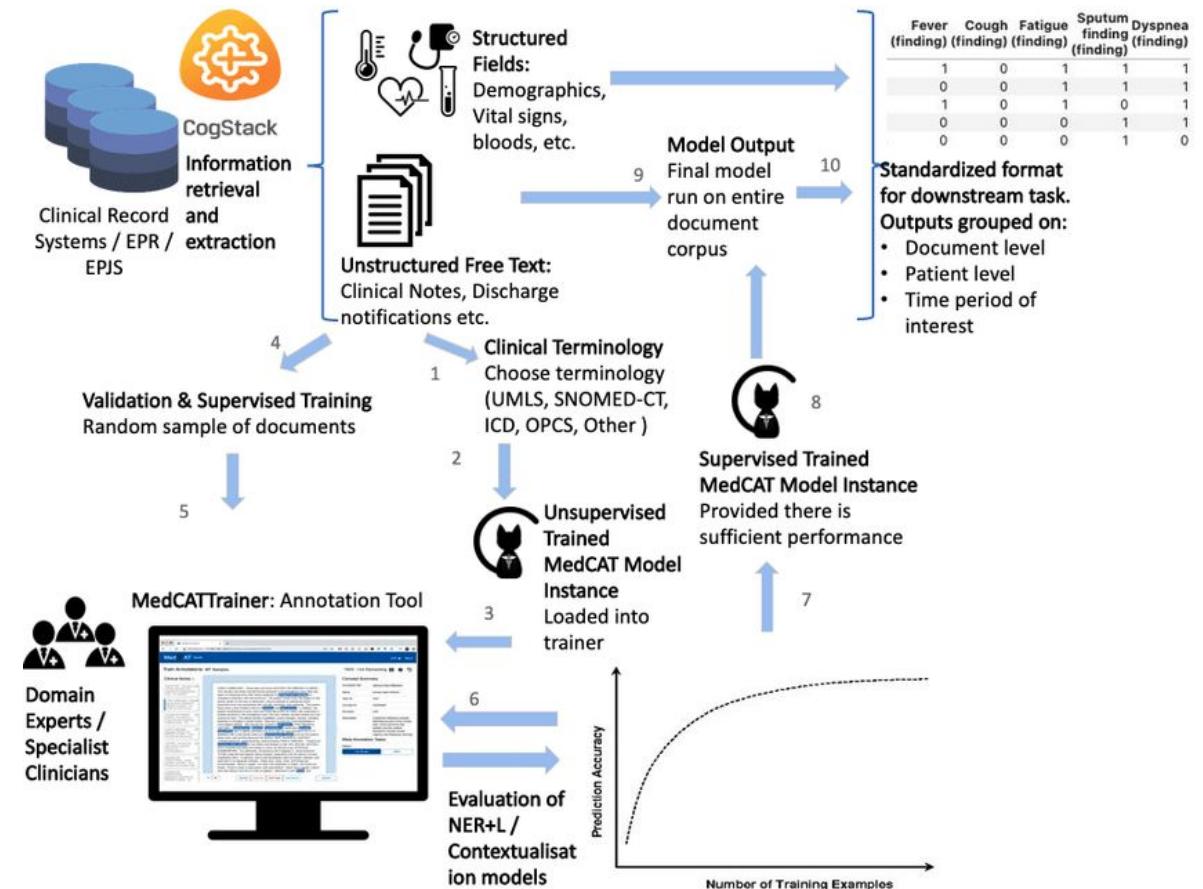
Figure adapted from Li, Lingyao, et al. "A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs)." arXiv preprint arXiv:2405.03066 (2024).

- Knowledge Graphs can have the structure of biology: diseases linked to genes, drugs, pathways, phenotypes
- Each node is a biomedical concept; each edge captures how those concepts interact biologically or clinically
- The structure can be built from curated sources like UMLS, SNOMED, DrugBank, GO and pathway databases etc
- Unlike LLMs, KGs are explicit, factual, mechanistic and traceable
- This explicit structure anchors models in verified biomedical knowledge
- They fix the grounding problem: LLMs hallucinate; KGs don't

- Clinical AI research is shifting from isolated models to systems that understand relationships, context and clinical reasoning
- LLMs bring strengths in interpreting narrative EHR text, capturing clinician intent and extracting complex clinical signals
- Knowledge graphs provide structured biomedical knowledge, linking diseases, genes, drugs, pathways and clinical concepts
- Combining LLMs with KGs is emerging as a powerful option: language understanding grounded by mechanistic, factual knowledge
- This integration pushes research toward more interpretable, reliable and biologically informed models of patient state

Information Extraction (IE)

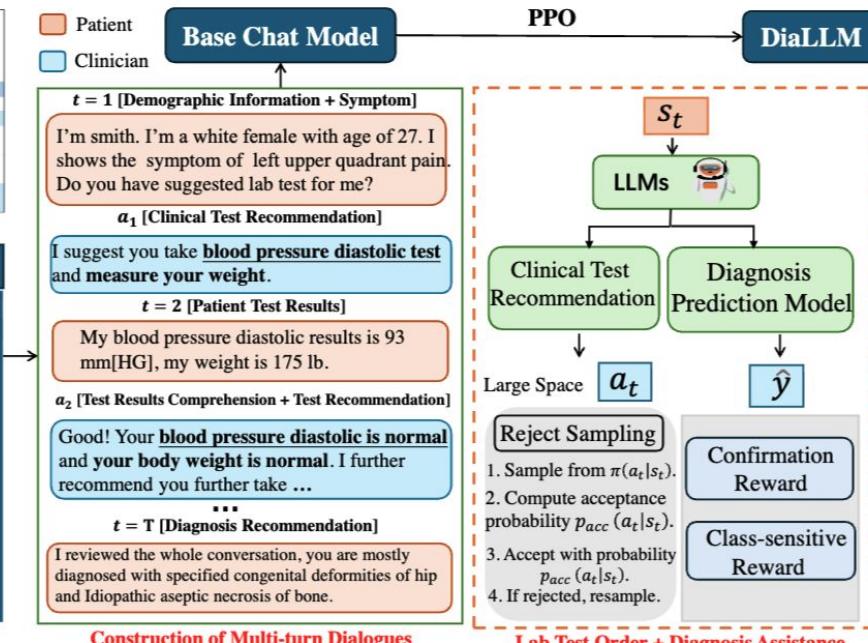
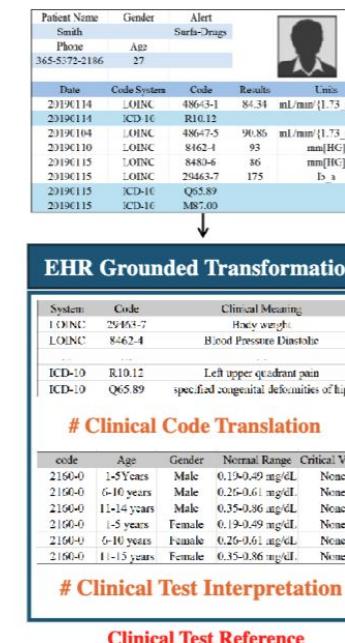
- Information extraction converts free text in EHRs into structured facts such as relations, events, and phenotypes.
- Most systems finetune BERT variants including ClinicalBERT, CancerBERT, and NeighBERT, while larger models such as GPT-5, LLaMA etc are used in zero or few shot settings.
- ClinicalBERT achieves an F1 of about 0.78 for adverse drug event extraction, and domain specific BERT models consistently outperform traditional machine learning baselines across relation and phenotype extraction tasks.



Li, Lingyao, et al. "A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs)." arXiv preprint arXiv:2405.03066 (2024).

Dialogue Systems

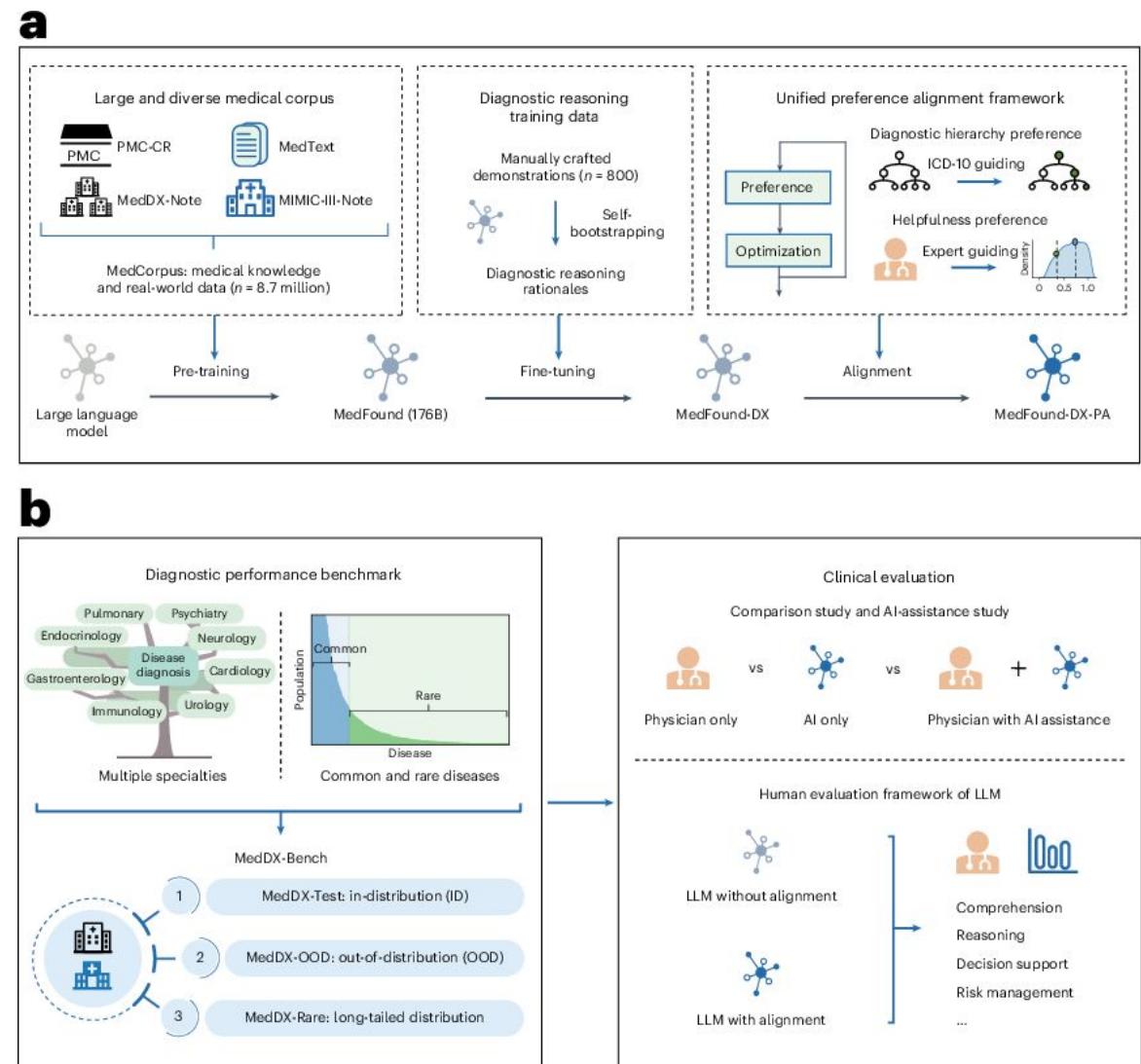
- Dialogue systems use LLMs to support patient messaging, question answering, and clinical communication, moving from simple keyword matching to fully conversational agents.
- Decoder only models such as GPT 4 dominate, generating draft replies, answering patient queries, and producing context aware clinical clarifications.
- These systems show strong reliability: GPT based responses reach a mean weighted F1 of about 0.61 against physician labels, and automated message drafting reduces clinician workload even with roughly 20 percent adoption.



Li, Lingyao, et al. "A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs)." arXiv preprint arXiv:2405.03066 (2024).

Diagnosis and Prediction

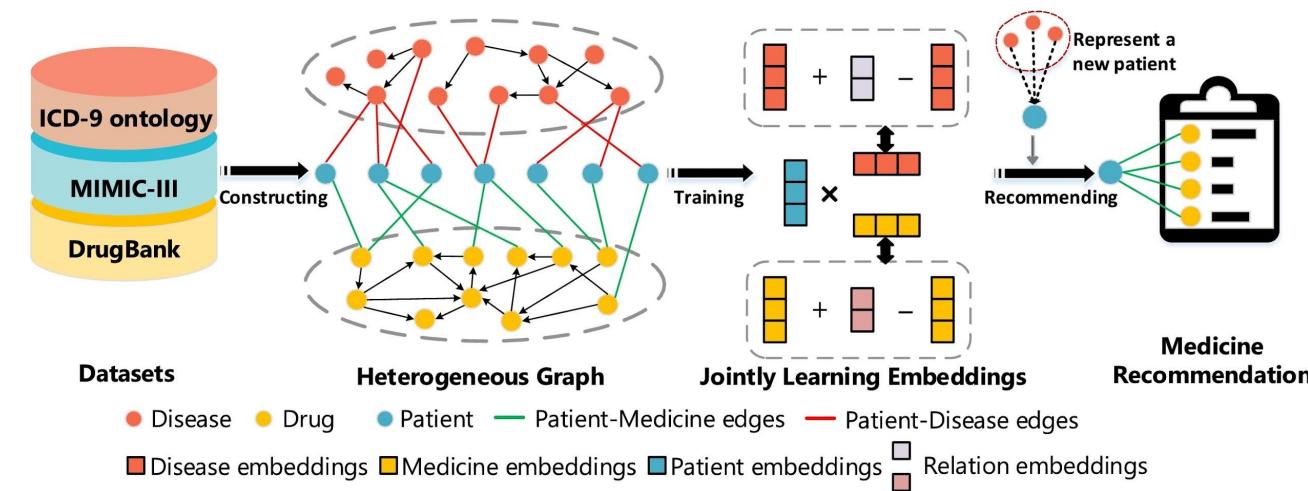
- LLMs are now used to predict diagnoses, risks, and future clinical events from EHR data, building on earlier work in clinical risk modelling.
- Studies test zero shot and few shot prompting with models such as GPT 4 and GPT 3.5, longitudinal prediction using sequential EHR histories, and multimodal fusion that combines structured fields with clinical text.
- GPT 4 exceeds traditional machine learning baselines for zero shot in hospital mortality prediction, few shot prompting boosts ICU admission and mortality forecasts, and longitudinal LLMs outperform RNNs, LSTMs, and gradient boosted trees across disease onset and prognosis tasks.



Li, Lingyao, et al. "A scoping review of using large language models (LLMs) to investigate electronic health records (EHRs)." arXiv preprint arXiv:2405.03066 (2024).

Medication Recommendation

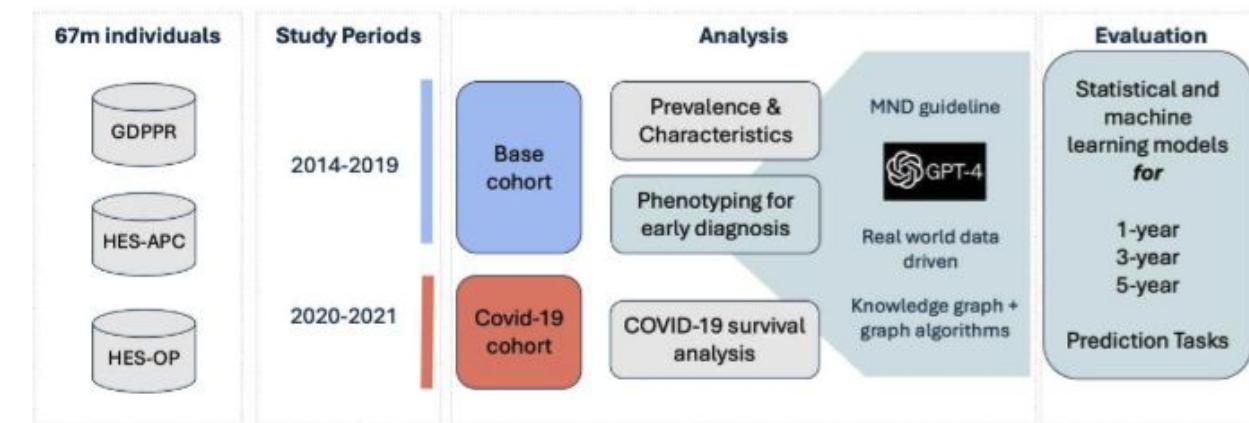
- EMR based medication recommenders struggle because they lack explicit facts such as drug–drug interactions, making safe personalised prescribing difficult.
- SMR builds a heterogeneous graph from MIMIC III, DrugBank, and ICD-9, learns patient, disease, and medication embeddings, and recommends drugs via link prediction that accounts for interaction risk.
- SMR outperforms rule based and multi label baselines in accuracy (0.6113 vs 0.5582) and achieves the lowest drug–drug interaction rate (0.17 percent).



Gong, Fan, et al. "SMR: medical knowledge graph embedding for safe medicine recommendation." Big Data Research 23 (2021): 100174.

Disease Diagnosis

- Early diagnosis of MND is difficult because guideline-based phenotypes such as the MNDA Red Flag List miss important predictors and have never been tested at population scale using national EHRs
- The study compared four phenotype sources — MNDA guidelines, knowledge graph derived phenotypes, GPT-4 generated phenotypes, and real-world data phenotypes — evaluated with models including logistic regression, SVMs, random forests, and LSTMs
- Real-world data phenotypes achieved the highest discrimination (AUROC 0.776), and the combined MNDA + KG + RWD ensemble reached the best overall performance (F1 0.557), outperforming MNDA guidelines alone (AUROC 0.703; F1 0.435).



Abdulle, Y., Wu, J., Budhdeo, S., Kim, Y., Shen, J., Sun, E., ... & Wu, H. (2025). Characteristics and Early Diagnosis of Motor Neuron Disease (MND) in 67 million individuals in England: a comparative study on phenotyping models derived by AI, Knowledge Graphs and the MND Association. medRxiv, 2025-07.

Question/Answering

- EHR QA is challenging because clinical data sit in complex relational schemas that require many SQL joins, making natural language to SQL translation brittle and poorly aligned with how clinical facts are stored.
- This work converts MIMIC III tables into a knowledge graph, builds NLQ to SPARQL datasets, and compares Seq2Seq models and the TREQS architecture on graph based versus SQL based query generation.
- Graph based SPARQL queries yield higher structural and logical form accuracy, with TREQS improving relation prediction by up to 5.1 percent and answer correctness by 3.6 percent, especially on complex multi join questions.

Patients

SUBJ_ID	NAME	DOB	Gender	...
12	"John"	2089-11-12	"M"	...
13	"Jane"	2117-12-25	"F"	...

Admissions

ADM_ID	SUBJ_ID	Age	Reason	...
231	12	81	"arrhythmia"	...
232	13	56	"cancer"	...

Diagnoses

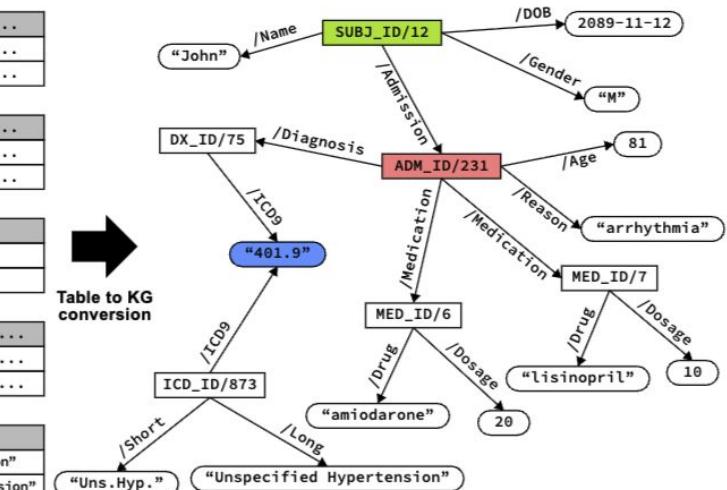
DX_ID	ADM_ID	ICD9	...
75	231	"401.9"	...
76	232	"162.9"	...

Medications

MED_ID	ADM_ID	Drug	Dosage	...
6	231	"amiodarone"	20	...
7	231	"lisinopril"	10	...

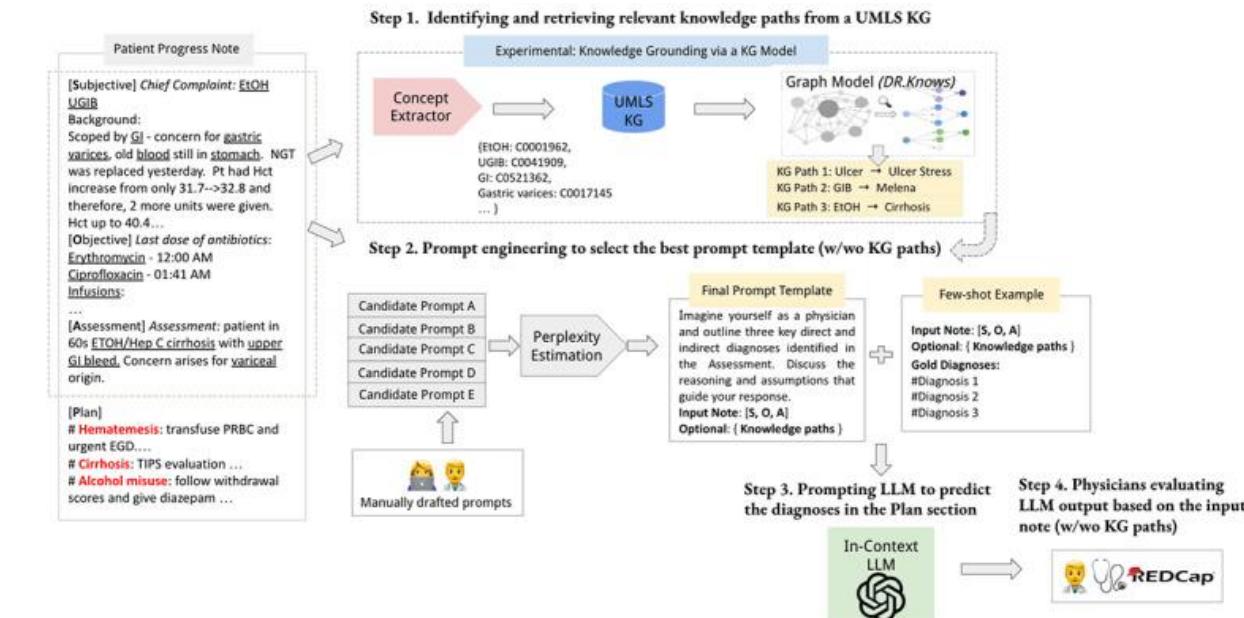
ICD9_Dx_Codes

ICD_ID	ICD9	Short	Long
872	"401.1"	"Ben.Hyp."	"Benign Hypertension"
873	"401.9"	"Uns.Hyp."	"Unspecified Hypertension"



DR-KNOWs

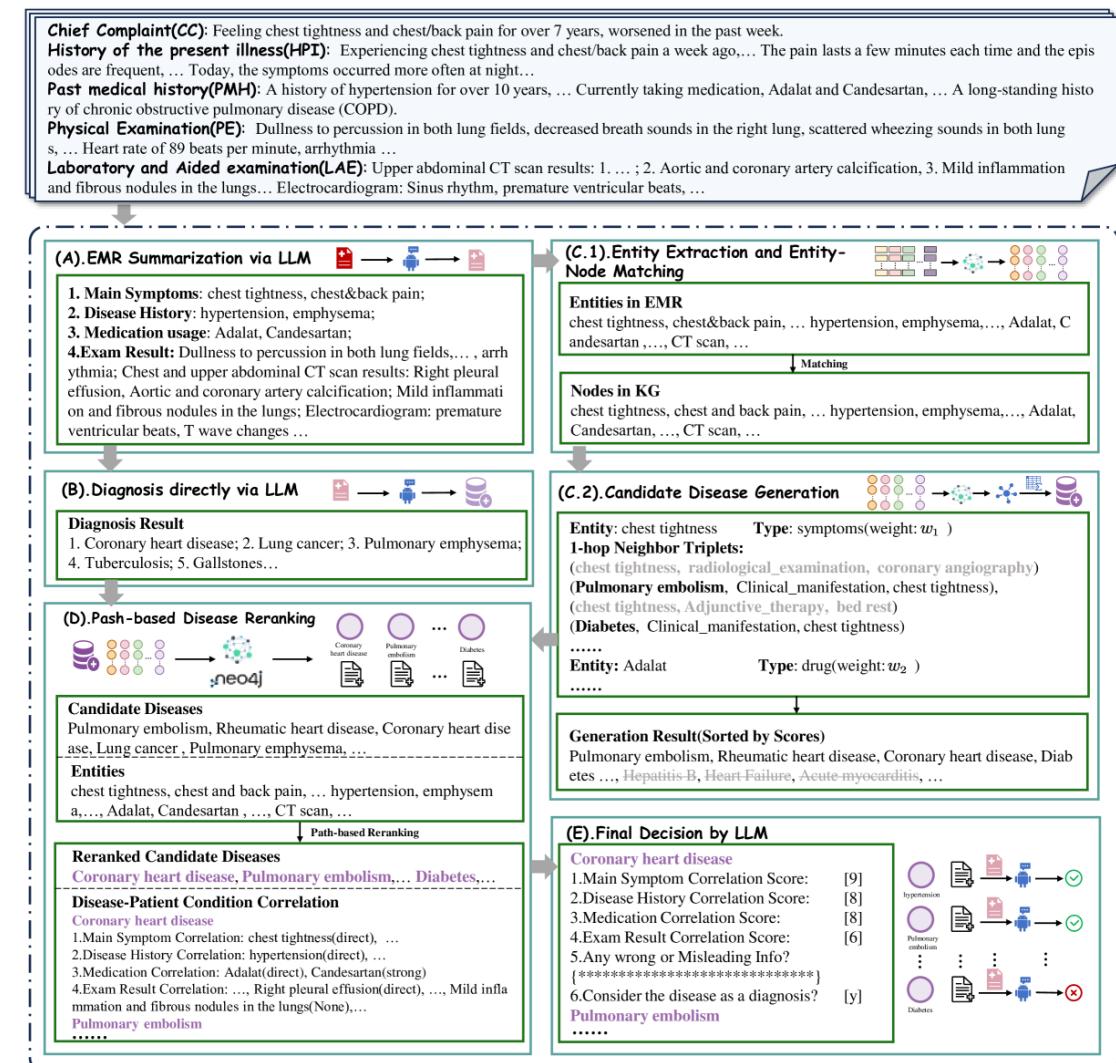
- LLMs can draft diagnoses from progress notes, but they often hallucinate or miss key clinical facts because the underlying diagnostic reasoning is not explicit in EHR text.
- DR.KNOWS retrieves patient specific diagnostic paths from a UMLS based knowledge graph and injects these ranked paths into prompts for models such as GPT 3.5.
- This improves concept extraction over QuickUMLS and significantly boosts diagnostic accuracy, with T5 plus knowledge paths achieving the best ROUGE L and CUI scores and human reviewers confirming clearer, more clinically correct reasoning.



Gao Y, Li R, Croxford E, Caskey J, Patterson B, Churpek M, Miller T, Dligach D, Afshar M
 Leveraging Medical Knowledge Graphs Into Large Language Models for Diagnosis Prediction: Design and Application Study

medIKAL

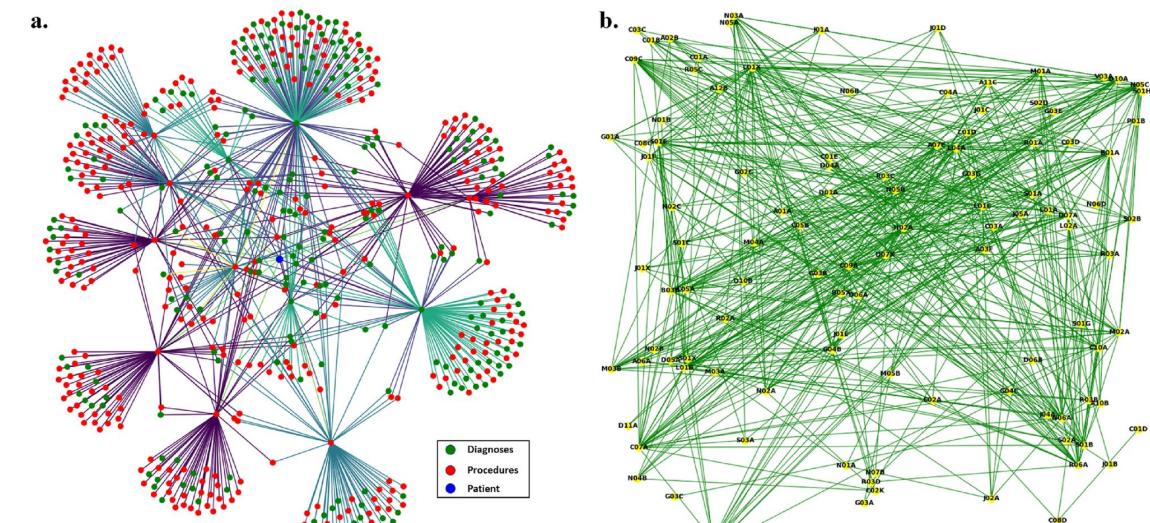
- The method addresses diagnostic complexity in noisy, redundant EHRs by pairing a large language model with a structured medical knowledge graph.
- It summarises EHR text with an LLM, weights extracted entities to guide knowledge graph retrieval, fuses LLM and KG signals through a ResNet like integration step, and reranks candidate diagnoses using shortest path correlations before producing a structured diagnostic output.
- Across multiple Chinese EHR datasets, it surpasses LLM only, KG augmented, and KG reasoning baselines, yielding higher diagnostic precision, recall, and overall stability.



Jia, Mingyi, et al. "medikal: Integrating knowledge graphs as assistants of llms for enhanced clinical diagnosis on emrs." Proceedings of the 31st International Conference on Computational Linguistics. 2025.

KGDNet

- Medication recommendation is difficult because models must use longitudinal EHR histories while also avoiding harmful drug–drug interactions, a balance many existing systems fail to achieve.
- KGDNet builds clinical and medication knowledge graphs, learns graph embeddings, models temporal history with GRUs, and fuses both streams before generating recommendations with multi-head attention.
- On MIMIC-IV, it outperforms all baselines in PR-AUC, F1, and Jaccard, while reducing DDI rates, providing more accurate and safer recommendations.



Mishra, R., Shridive, S. Knowledge graph driven medicine recommendation system using graph neural networks on longitudinal medical records. Sci Rep 14, 25449 (2024).
<https://doi.org/10.1038/s41598-024-75784-5>

Unifying Multimodal Clinical and Genomic Data through Knowledge Graphs and LLMs

Yusuf Abdulle

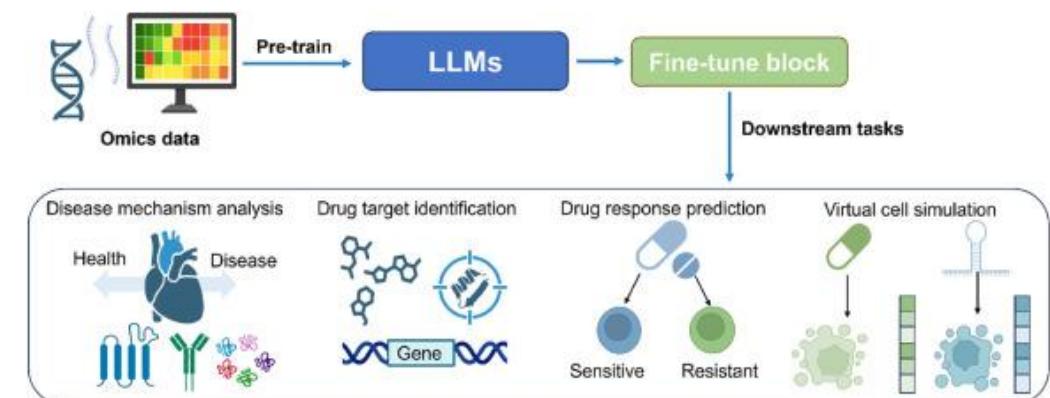
, King's College London

24 November 2025

- Modern healthcare data spans multiple modalities, including clinical text, structured EHR fields, imaging, laboratory values, and genomic profiles.
- Each modality captures a distinct layer of disease biology, but these data streams are often fragmented, sparse, and difficult to integrate.
- Knowledge graphs provide a unifying semantic framework by connecting molecular, phenotypic, and clinical concepts through explicit, interpretable relationships.
- Large language models offer a flexible reasoning layer capable of fusing diverse signals and generating clinically meaningful insights.
- Together, knowledge graphs and LLMs support holistic patient understanding, enabling more mechanistic, personalised, and context rich precision medicine.

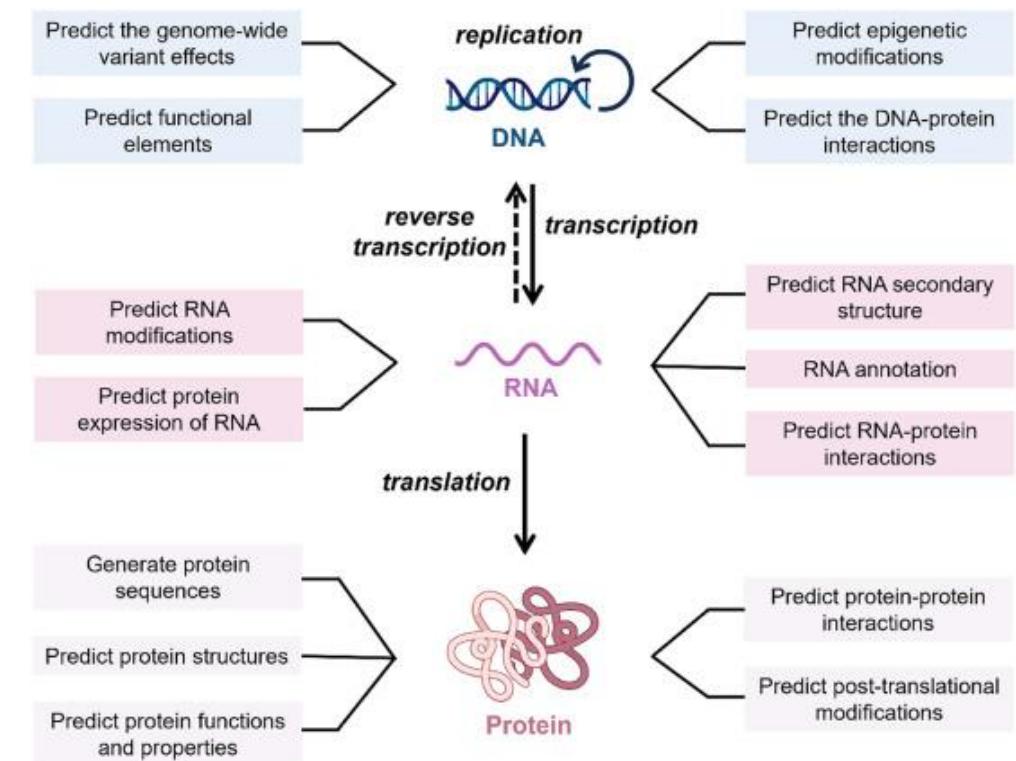
Drug Discovery

- Drug discovery spans chemistry, protein biology, omics, and clinical phenotypes, making single modality models insufficient for capturing cross domain therapeutic mechanisms.
- LLMs fuse SMILES and 3D structures, ESM protein embeddings, transcriptomic perturbation signatures, and knowledge graph drug–disease pathways through joint encoders and cross attention.
- They improve binding and target prediction, recover drug–disease links, and simulate gene and drug perturbations with greater mechanistic clarity than earlier methods.



Bioinformatics

- Bioinformatics data in genomics, transcriptomics, and proteomics are high dimensional, sparse, and heterogeneous, and older deep learning models often miss long range and cross modal biological structure.
- Transformer models are now adapted to biological sequences and multimodal omics, including DNABERT for DNA, RNA-FM and GeneRNA for RNA, ESM-2 for proteins, and multimodal systems such as AlphaGenome and ProLLM.
- These models outperform prior methods across promoter prediction, RNA structure modelling, protein interaction reasoning, and variant effect prediction, while offering stronger cross species generalisation and more interpretable biological insights.

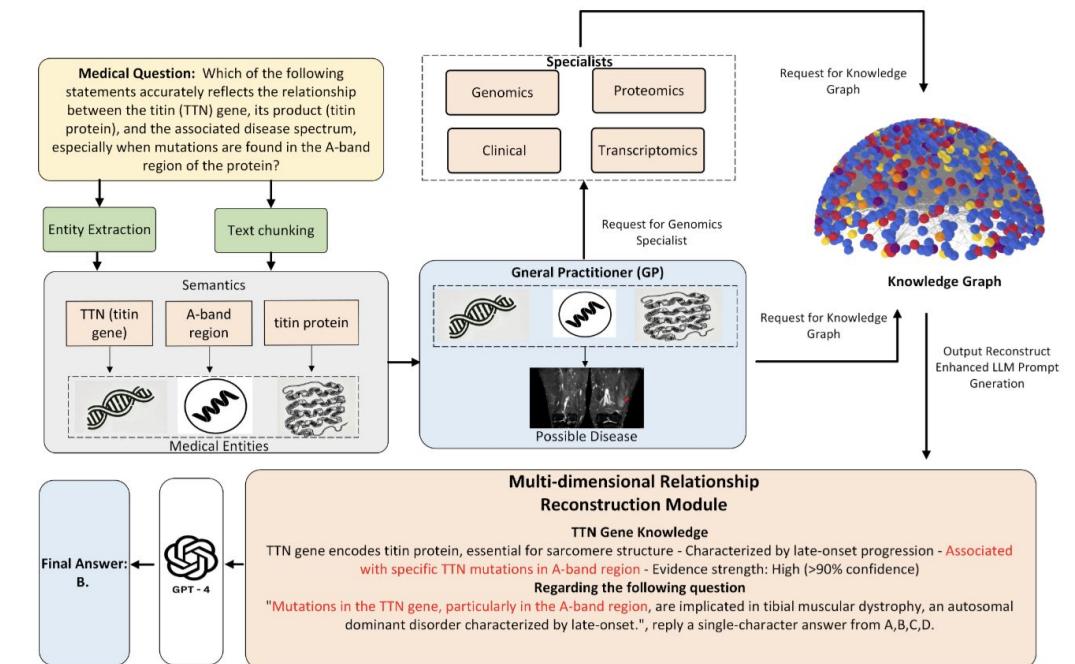


MultiCNKG

- The study unifies cognitive neuroscience, gene ontology, and disease ontology to overcome semantic gaps that limit reasoning and precision medicine.
- It uses LLM based entity normalisation, embedding similarity, relation unification, and iterative graph expansion to merge these sources into one multimodal knowledge graph.
- The resulting MultiCNKG (6.9 thousand nodes, 11.3 thousand edges) shows high expert rated precision and recall, improves link prediction over each source graph, and provides richer cross domain context for tasks such as variant interpretation and disease mechanism inference.

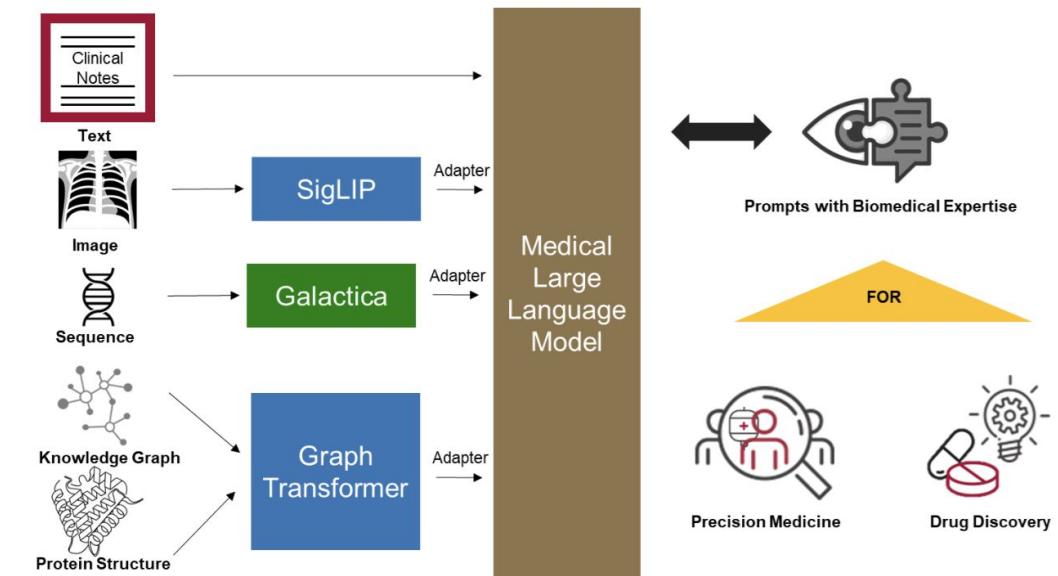
HEAL KGGen

- HEAL KGGen improves biomarker-based diagnosis by pairing a hierarchical multi-agent LLM system with an automatically constructed biomedical knowledge graph.
- Specialist agents handle genomics, transcriptomics, proteomics, and clinical queries, while a dynamic KG built from UMLS, MeSH, and SNOMED CT provides structured, context rich prompts.
- KGGen sharply boosts accuracy and reduces hallucinations across models, with Claude 3.5 Sonnet rising from 42.5 percent to 86.25 percent and similar gains for GPT-4, Gemini 1.5 Pro, and LLaMA 3.1.



Foundation Model for Biomedical Graphs

- Most biomedical LLMs handle text, images, or sequences but miss the structural information encoded in protein graphs, molecular graphs, and clinical knowledge graphs.
- The work introduces a multimodal foundation model that pairs a biomedical LLM with encoders for images, biological sequences, and a new graph transformer, all aligned through lightweight adapters in an LLaVA style framework.
- Early results show that continued pretraining on MedlinePlus boosts medical MCQ performance for smaller models, supporting the feasibility of the approach for tasks such as target identification and drug repurposing.



[Foundation Model for Biomedical Graphs: Integrating Knowledge Graphs and Protein Structures to Large Language Models](<https://aclanthology.org/2024.acl-srw.58/>) (Kim, ACL 2024)

Thank you

See you at 13:15PM after the lunch break.

Expanding Large Language Models to Vision: Multimodal LLMs

Yunsoo Kim

Institute of Health Informatics, University College London

24 November 2025

V. 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.



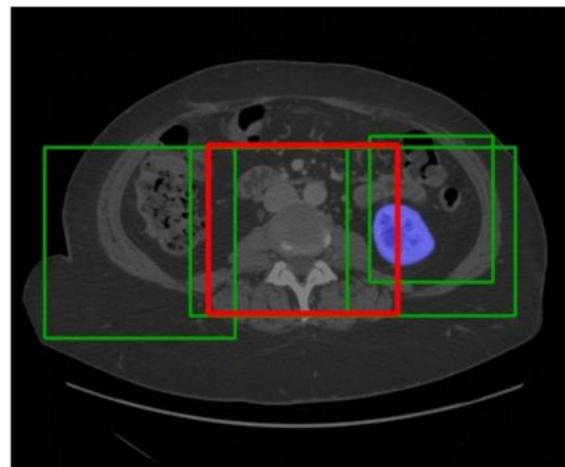
Prompt

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:

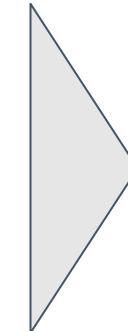
Top-left corner (x₁, y₁): Approximately (300, 150)

Bottom-right corner (x₂, y₂): Approximately (450, 300)

V. 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>

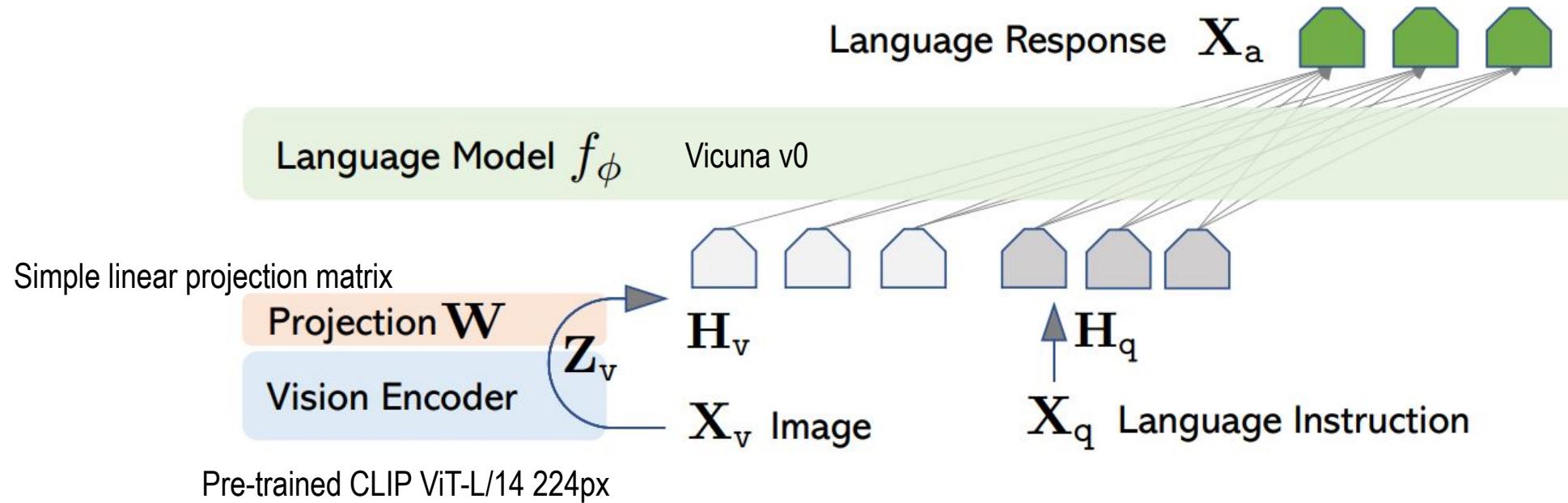
$$\mathbf{X}_{\text{instruct}}^t = \begin{cases} \text{Randomly choose } [\mathbf{X}_q^1, \mathbf{X}_v] \text{ or } [\mathbf{X}_v, \mathbf{X}_q^1], & \text{the first turn } t = 1 \\ \mathbf{X}_q^t, & \text{the remaining turns } t > 1 \end{cases}$$

158K visual instruction data

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

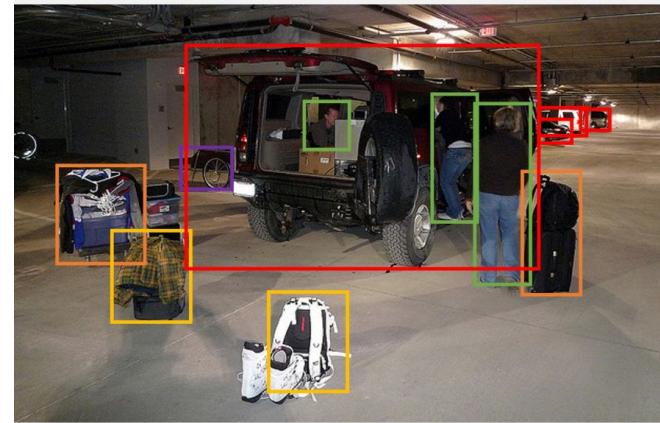
V. Expanding Large Language Models to Vision: LLaVA

LLaVA - Model Architecture



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.”

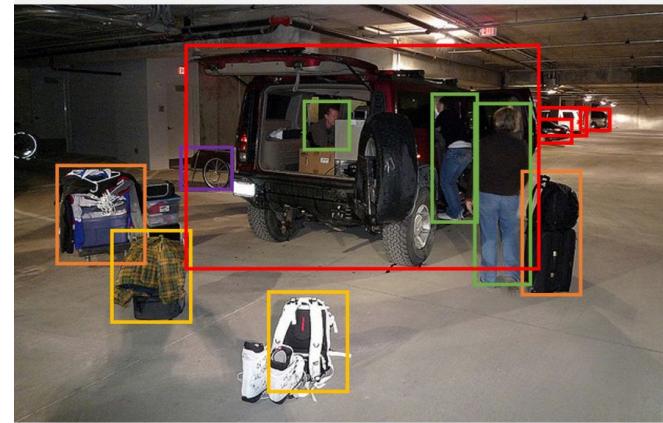


Ground Truth Answer is the captions

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.

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>

V. 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

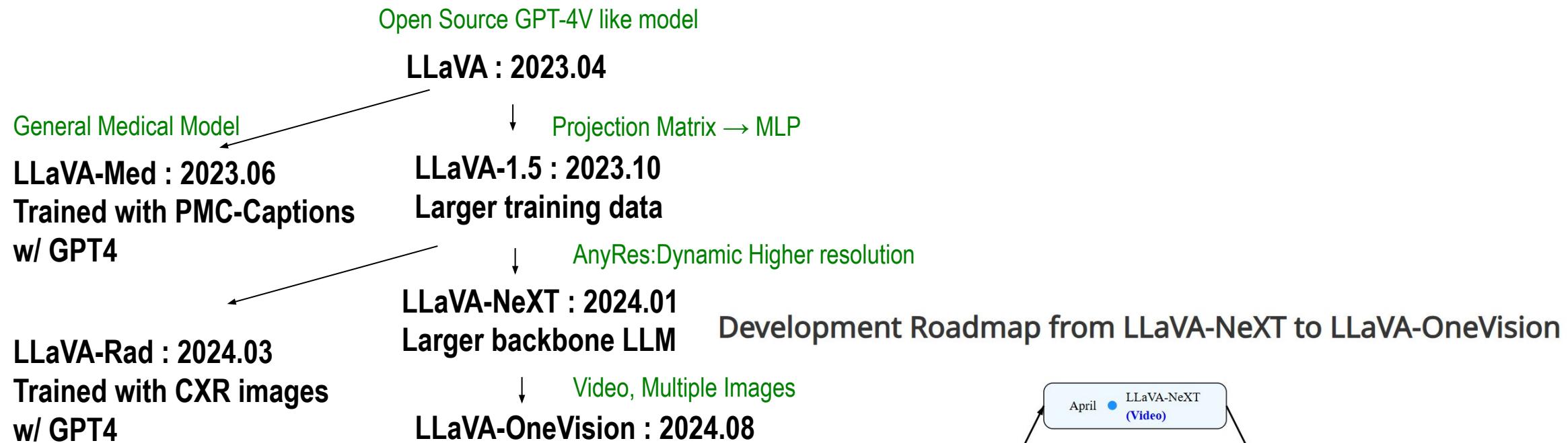


Source: <https://www.barnorama.com/wp-content/uploads/2016/12/03-Confusing-Pictures.jpg>

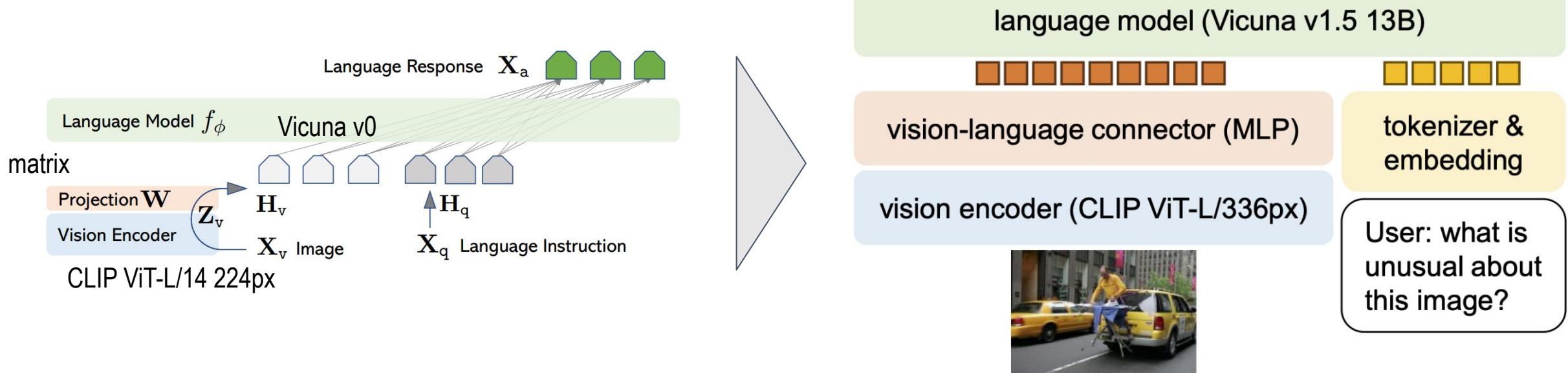
What is unusual about this image?

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.

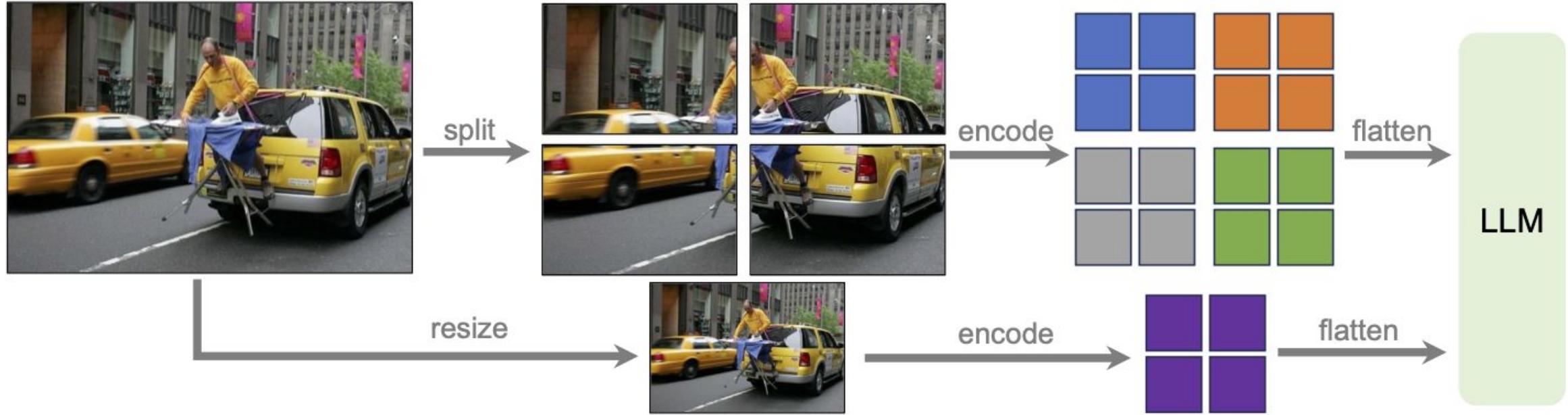
LLaVA-Family Models



LLaVA-1.5: Linear projection Matrix → Connector MLP
More train data : academic task-oriented data

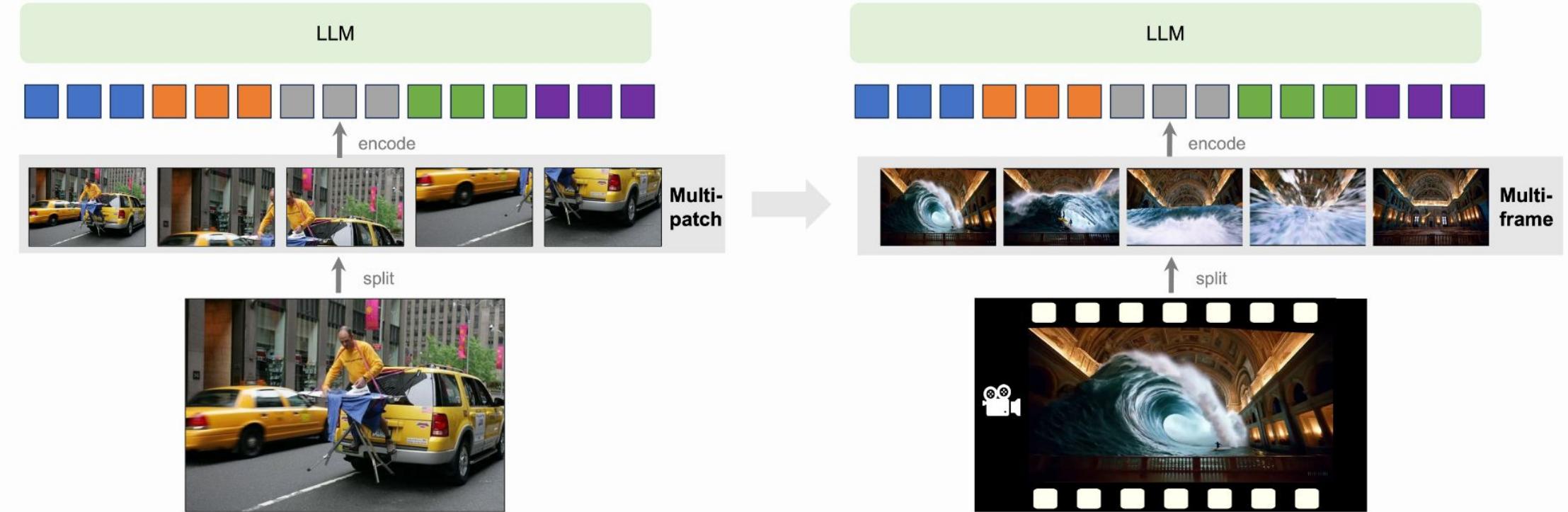


LLaVA-NeXT (also called LLaVA-1.6): Increased resolution with AnyRes

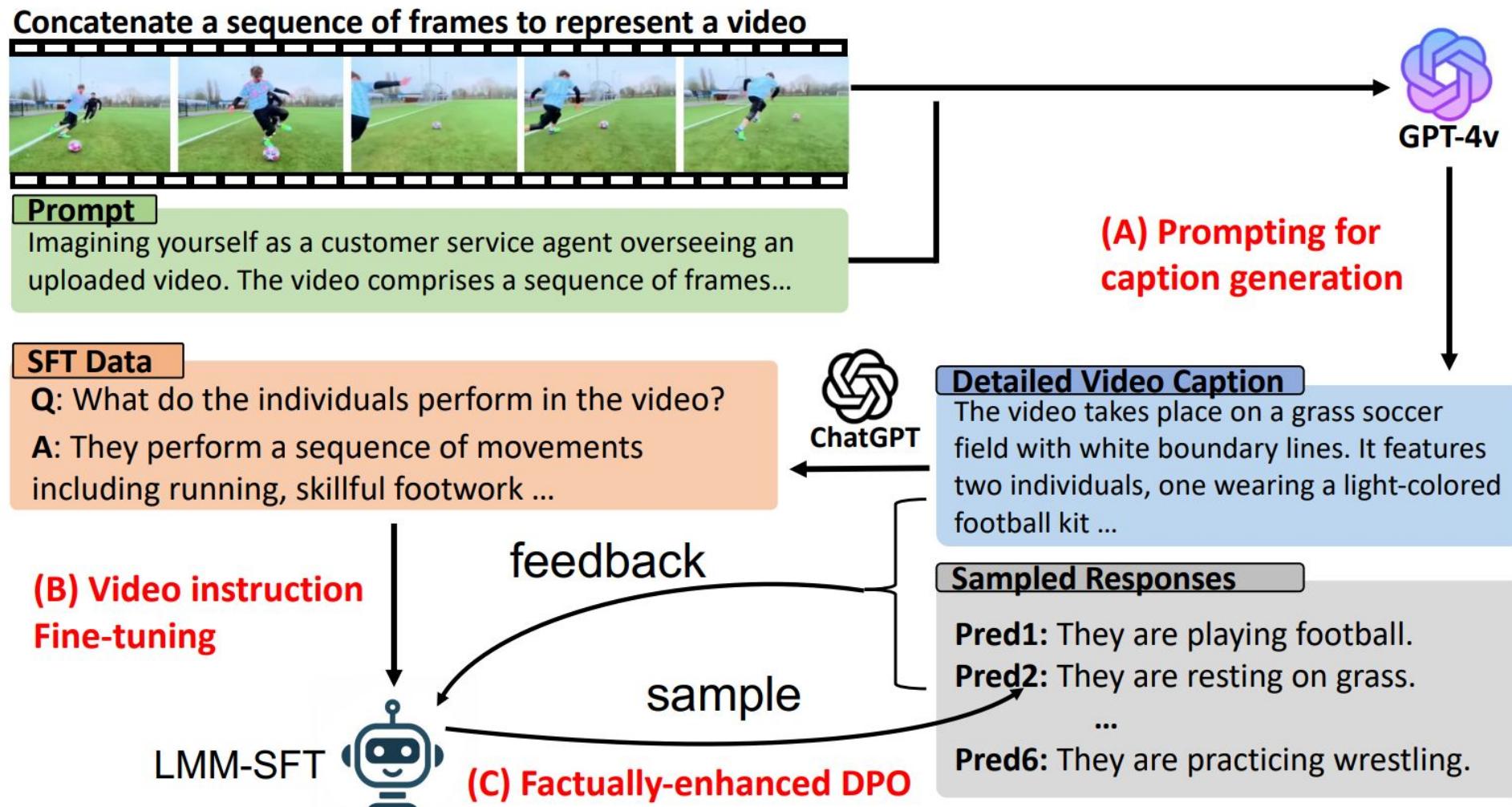


Other backbone LLMs: Mistral-7B and Yi-34B other than Vicuna

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



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

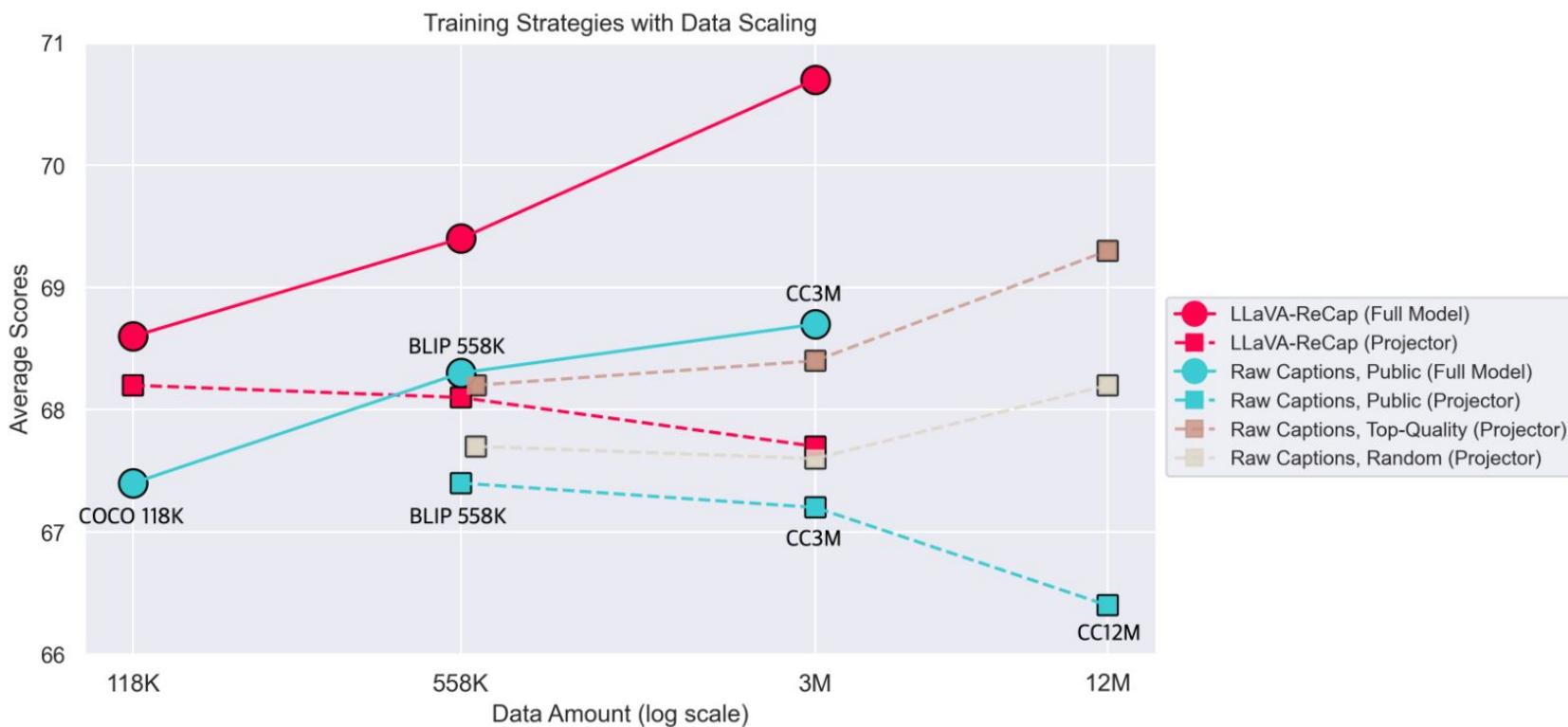


LLaVA-NeXT (Ablation) - Scaling model size effective with LLMs, but not with vision encoders

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

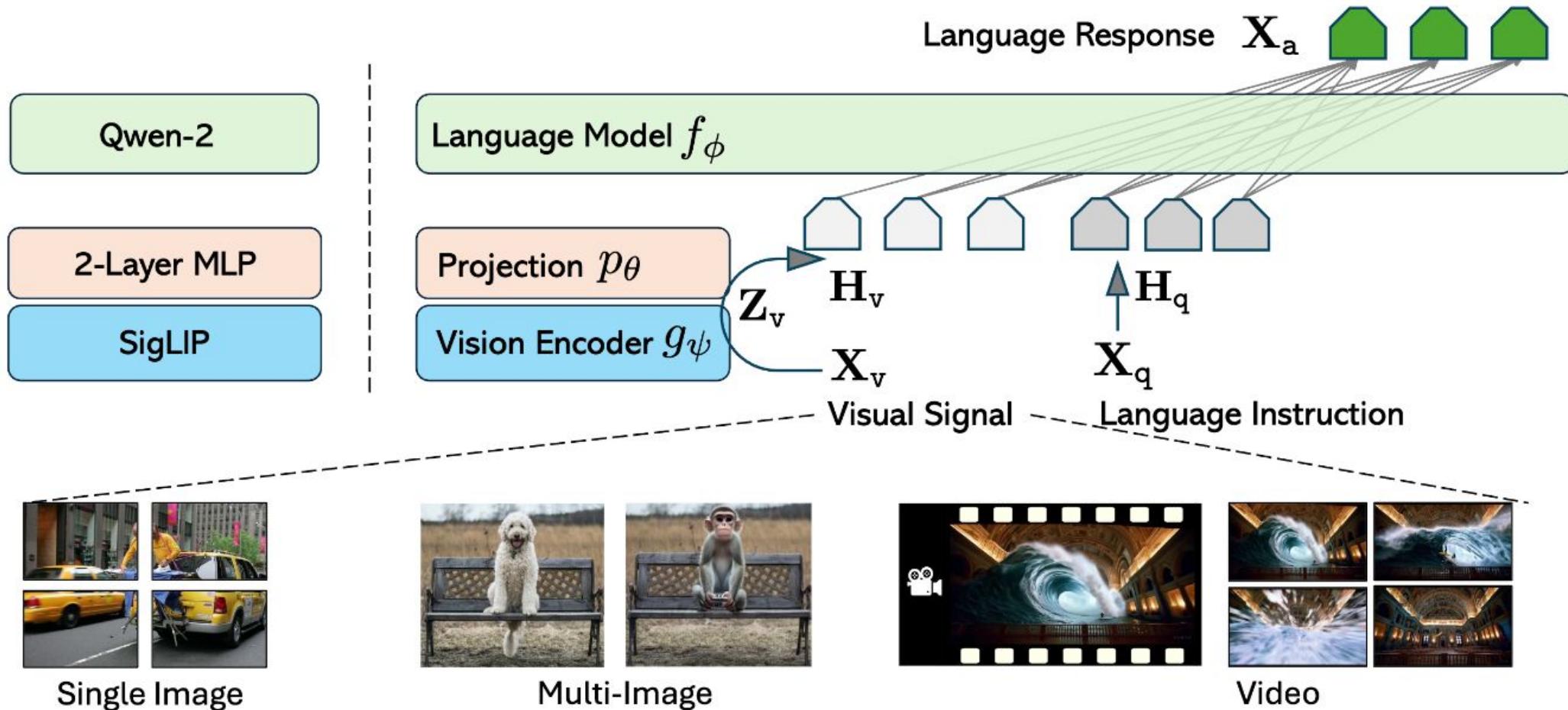
Vision Encoder	Model size	Avg.
CLIP-L	0.3B	63.4
CLIP-L	0.3B	65.3
EVA-02-E	4.7B	61.0
EVA-8B	8B	63.3
EVA-8B	8B	64.4
SO400M	0.4B	66.4

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



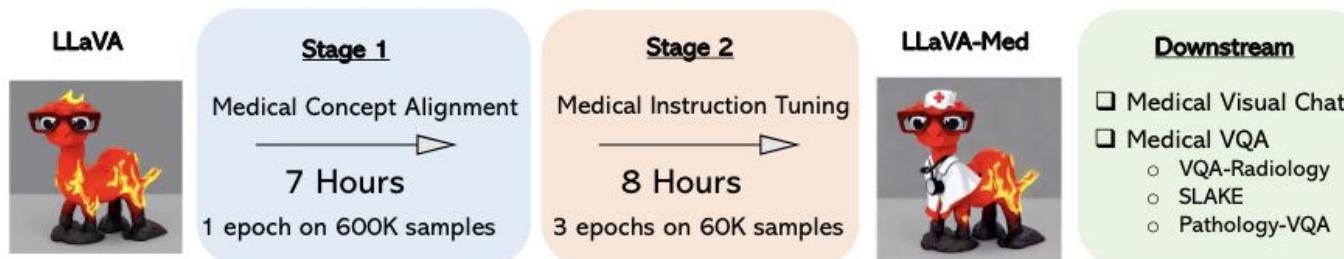
LLaVA-OneVision: Improved vision encoder and backbone LLM supporting single, multi images and video

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



V. Expanding Large Language Models to Vision: LLaVA-Med

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.

Multimodal LLMs for Radiology

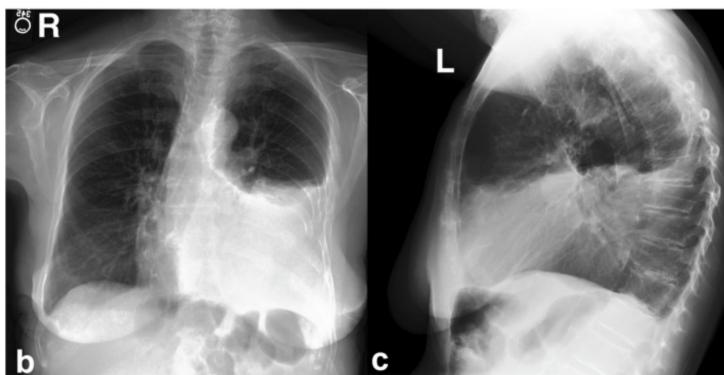
Jinge Wu

Institute of Health Informatics, University College London

24 November 2025

Data Construction

{Radiology Images} + {Textual Prompt} = {Textual Answer}



Can you generate
+ a report from the
X-ray images?

Input

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

Output

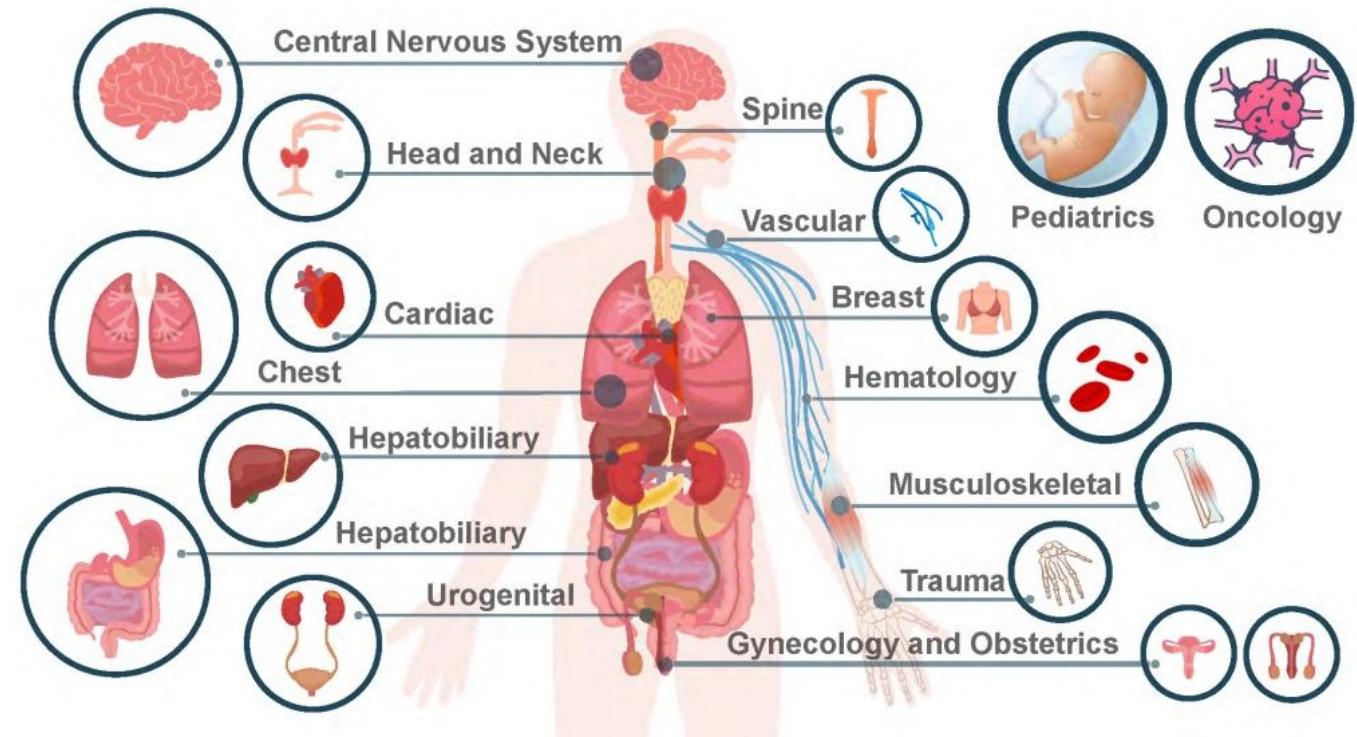
Image:

- Diverse modalities (X-ray, CT, MR ...)
- Across body systems (brain, chest ...)
- Different planes (axial, coronal, sagittal)
- 2D or 3D in different format:
 - DICOM/ NIfTI (common in 3D scans)
 - PNG/JPEG (often serve as 2D slices)
- One or multiple

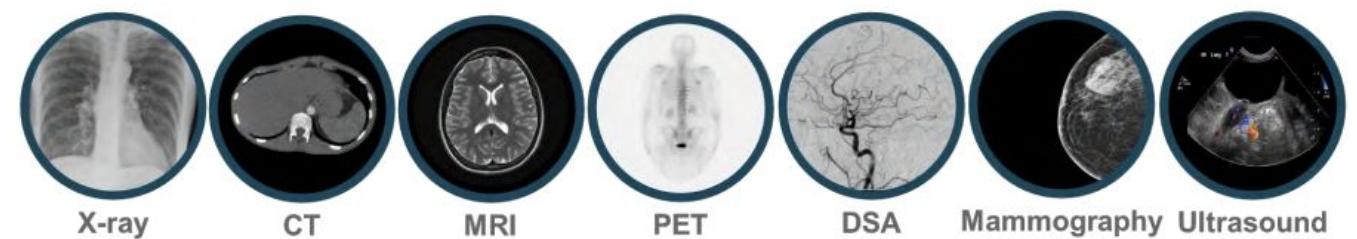
Text:

- Diverse task requirements (diagnosis, report generation, question-answering, treatment ...)
- Context-involved (other examination records, chief complaint, history of present illness ...)

a. Medical systems



b. Imaging modalities



From daily clinical report writing

MIMIC-CXR:

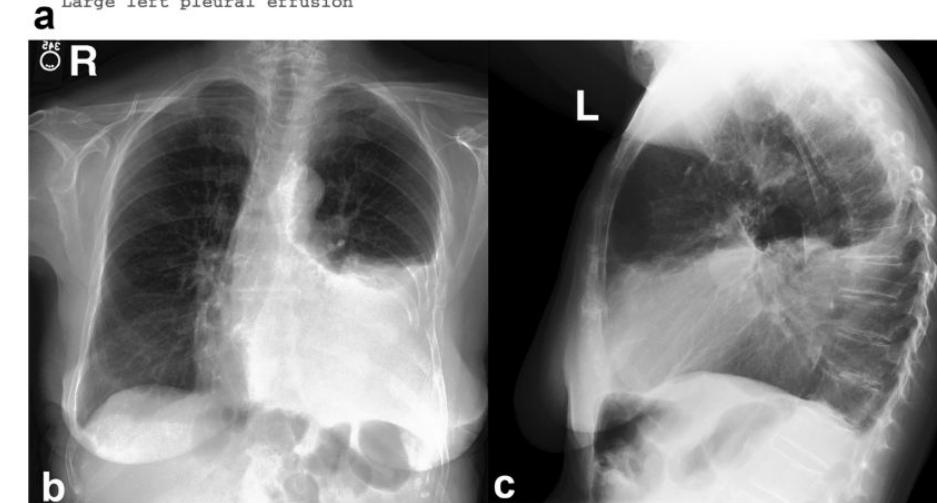
- **Volume:** 377,110 images; 227,835 studies; 65,379 distinct patients (2011–2016)
- **Paired Data:** Each study consists of the **X-ray image** linked to its corresponding **free-text radiology report** (Findings & Impression).
- **14 common pathologies** (e.g., Pneumonia, Cardiomegaly)
- **Integration with MIMIC-IV:** Unique ability to link imaging data to the full **Electronic Health Record (EHR)** database via subject_id.
- **Holistic Analysis:** Enables multimodal research combining pixel data with vital signs, laboratory results, medications, and patient outcomes.
- **Source:** Hosted on **PhysioNet** under a Credentialed Data Usage Agreement.

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



Example study contained in MIMIC-CXR. Above (a), the radiology report provides the interpretation of the image. PHI has been removed and replaced with three underscores (_ _ _). Below, the two chest radiographs for this study are shown: (b) the frontal view (left image) and (c) the lateral view (right image).

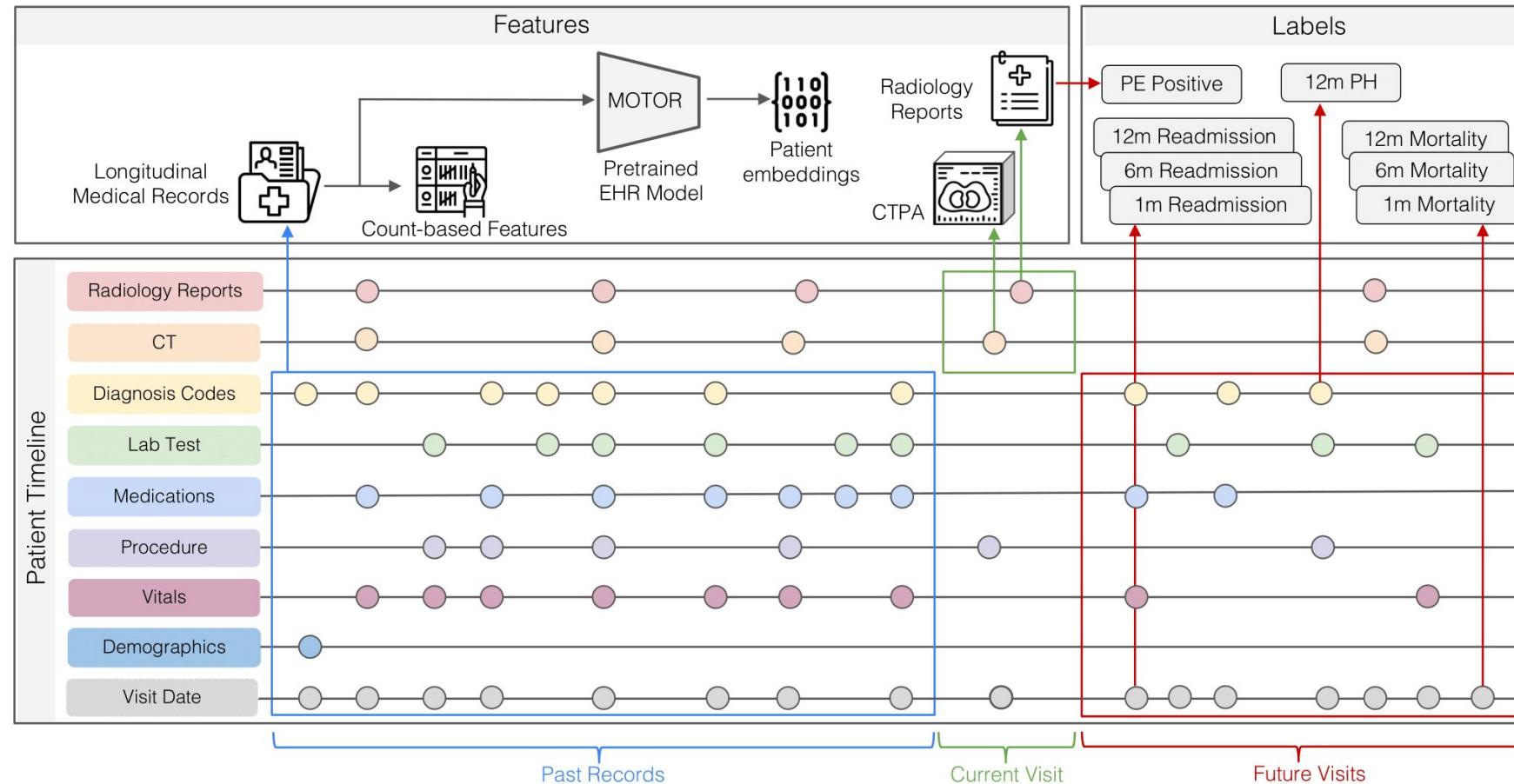
From daily clinical report writing

Medical-Diff-VQA: difference visual question answering simulated from **MIMIC-CXR** Chest X-Ray Images

<p>Main image</p>  <p>(a) Ground Truth Report: No change in the positioning of the left-sided chest tube. Re demonstration of fractures of the fifth and sixth left posterior ribs. A pleural line is not present on the current study. Residual left lung atelectasis is present. The right lung is clear. Heart size is normal. No focal consolidation. No pleural line is detected to indicate residual left pneumothorax. Left lower lobe atelectasis persists, in the setting of possible splinting given the posterior rib fractures.</p>	<p>(b) Text Mined Label from MIMIC Atelectasis Fracture Pneumothorax (incorrect)</p>
<p>Reference image</p>  <p>(e) Ground Truth Report: Comparison to ___, 07:47. Removal of the left chest tube. No evidence of pneumothorax. Mild basal atelectasis. Mild cardiomegaly. No pneumonia.</p>	<p>(c) Generated Report compared with the prior study, there is a new left apical pneumothorax with a small left apical pneumothorax. the left chest tube is in unchanged position. there is persistent elevation of the left hemidiaphragm and left basilar atelectasis. the right lung is clear. there is a small left pleural effusion (unrelated diseases). heart is normal in size.</p> <p>(d) ImageCLEF VQA-MED questions is there something wrong in the image? what is the primary abnormality in this image?</p> <p>(f) Our medical difference VQA questions What abnormality is seen in this image? Atelectasis, Fracture. Is there pneumothorax in this image? No Where is the fracture? Posterior ribs Where is the atelectasis? Left lower lobe What is the level of heart size? Normal which view is this image taken? AP view What has changed compared to the reference image? The main image has an additional finding of fracture. The main image is missing the finding of cardiomegaly.</p>

From daily clinical report writing

INSPECT: A Multimodal 3D-CT Dataset Integrating Imaging, Reports, and Longitudinal EHR



Focus: Pulmonary Embolism (PE) diagnosis

Modality: 3D Imaging + Text + Longitudinal EHR

Volume: 19,402 Patients; 23,248 CTPA Studies ; 23,248 Reports ; 225M+ Clinical Events

From daily clinical report writing

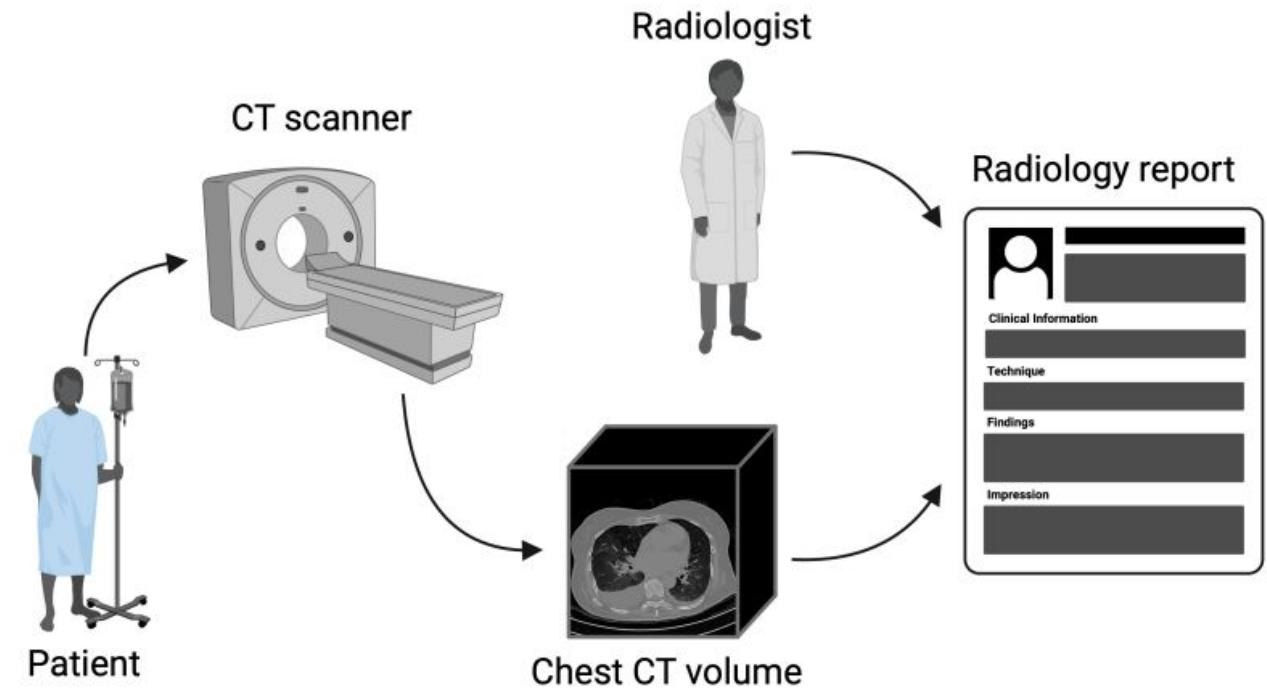
CT-RATE:

Data Sources.

- non-contrast 3D chest CT scans from hospital PACS with corresponding radiology reports from the EHR.

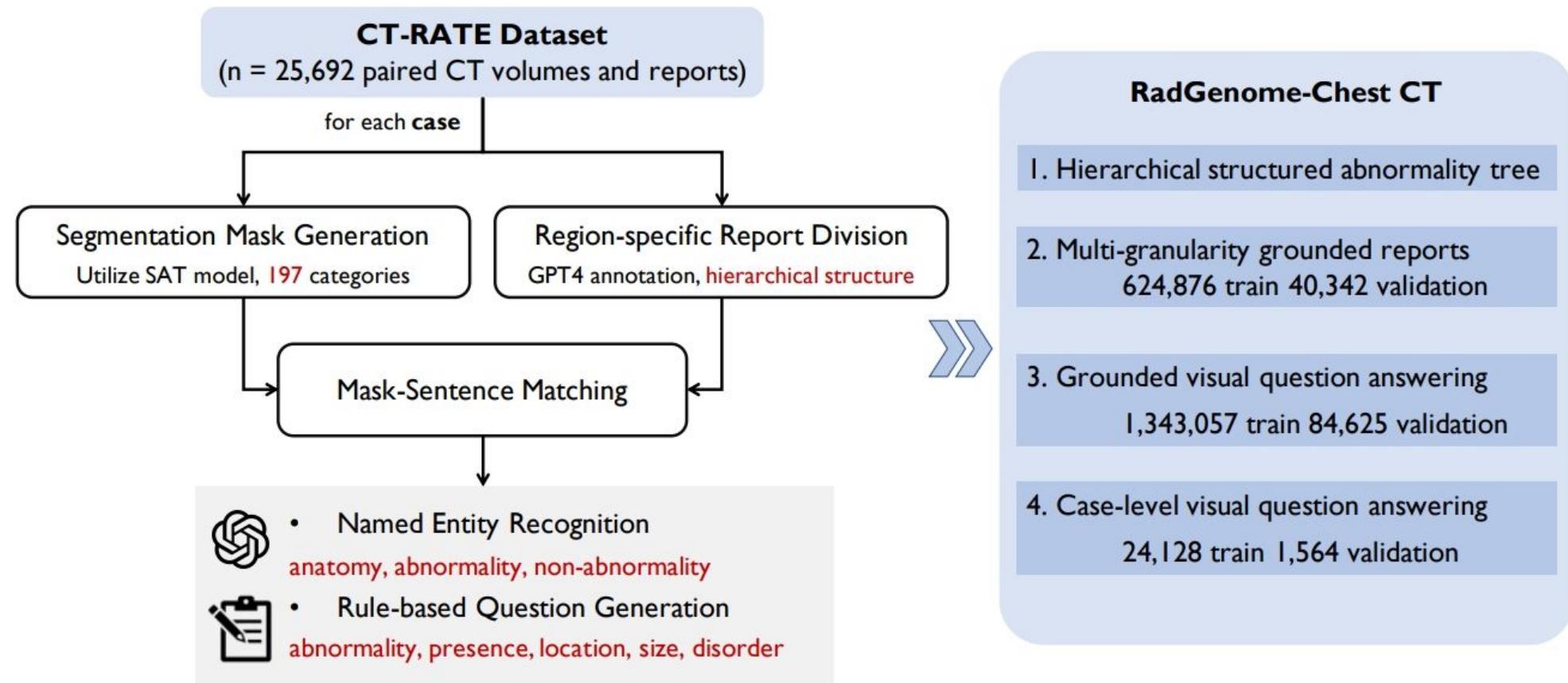
Final Structure.

- 25,692 scans from 21,304 patients, reconstructed into 50,188 volumes with over 14.3 million slices.
- Each volume is linked to its free-text report, forming the first large-scale 3D chest CT image–text dataset for multimodal research.



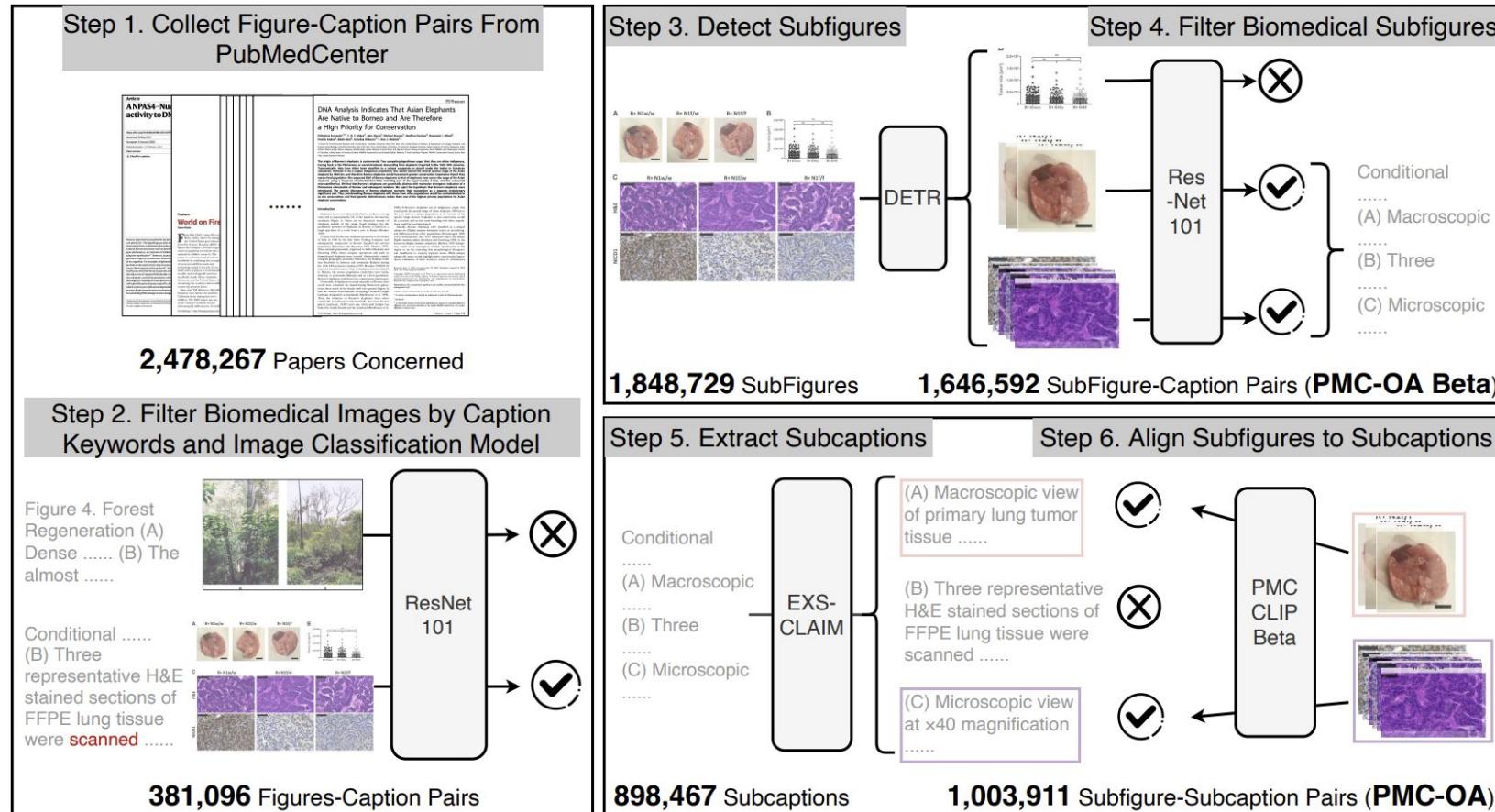
From daily clinical report writing

RadGenome-Abdomen: reprocess **CT-RATE** data into diverse VQA pairs.



From papers

PMC-OA: extracting medical image–text pairs from published papers



From papers

PMC-VQA: prompt the image-text pairs in **PMC-OA** into diverse VQA pairs

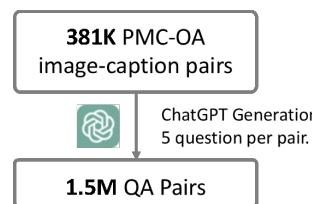


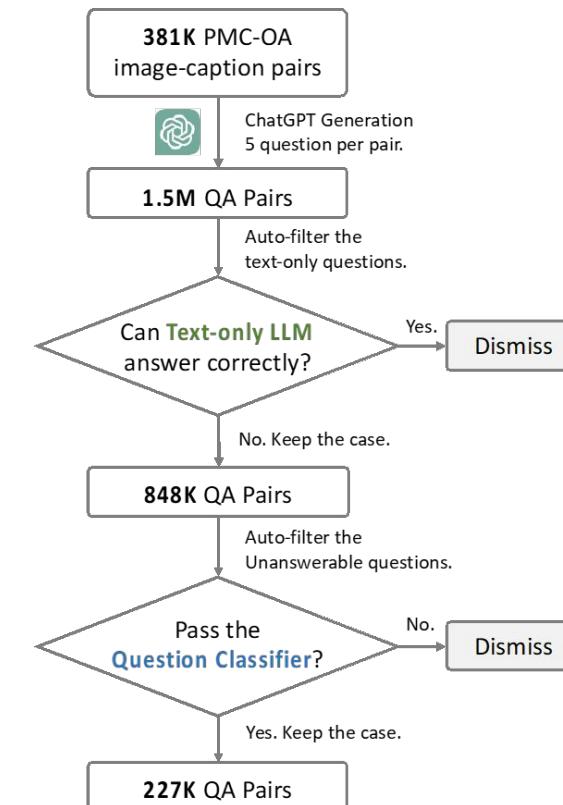
Image Caption: Anatomic and MRF data of subject excluded due to partial volume effects. a MRF T1-relaxometry map shows b Axial spin echo post-gadolinium T1WI shows typical homogeneous enhancement and extradural location of meningioma.



Question-Answer Generation

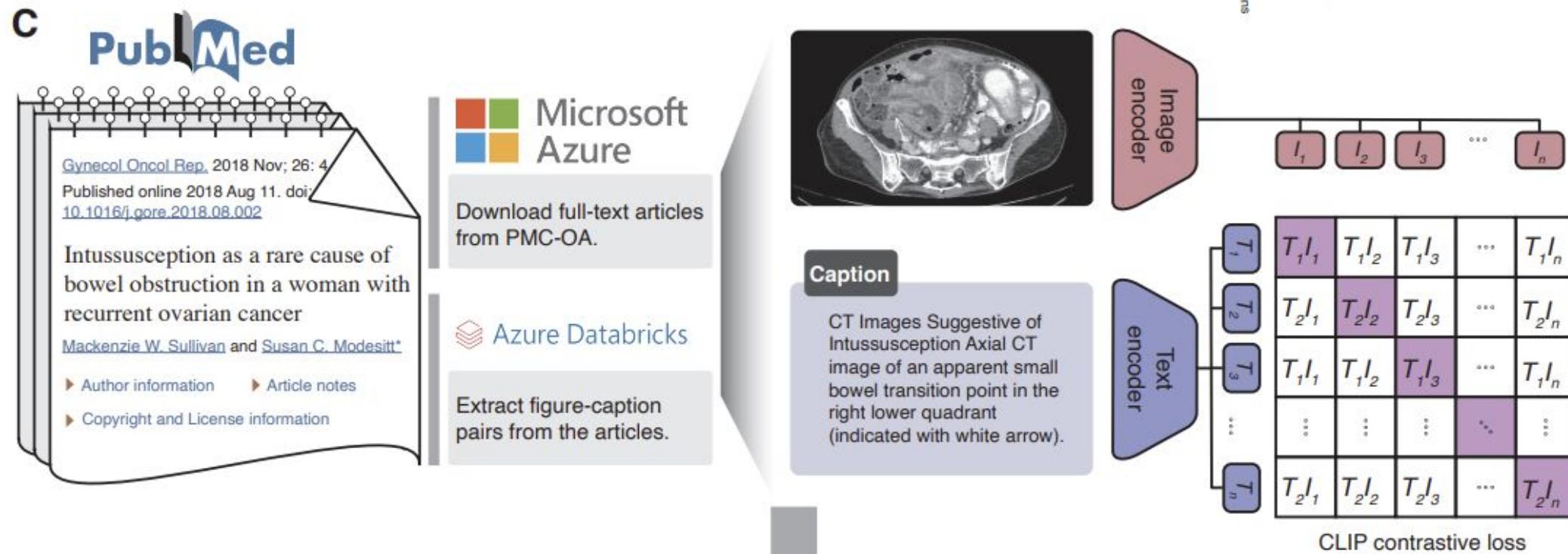
What did the axial spin echo post-gadolinium T1WI show in the patient?
A: Homogeneous enhancement of the brain tissue
B: Heterogeneous enhancement of the brain tissue
C: Homogeneous enhancement and extradural location of meningioma
D: Homogeneous enhancement and intradural location of meningioma

Prompts: Ask 5 questions about the content and generate four options for each question. The questions should be answerable with the information provided in the caption, and the four options should include one correct and three incorrect options, with the position of the correct option randomized. The output should use the following template: i:'the question index' question:'the generate question' choice: 'A:option content B:option content C:option content D:option content' answer: The correct option(A\B\C\D).



From papers

PMC-15M: extracting medical image–text pairs from published papers

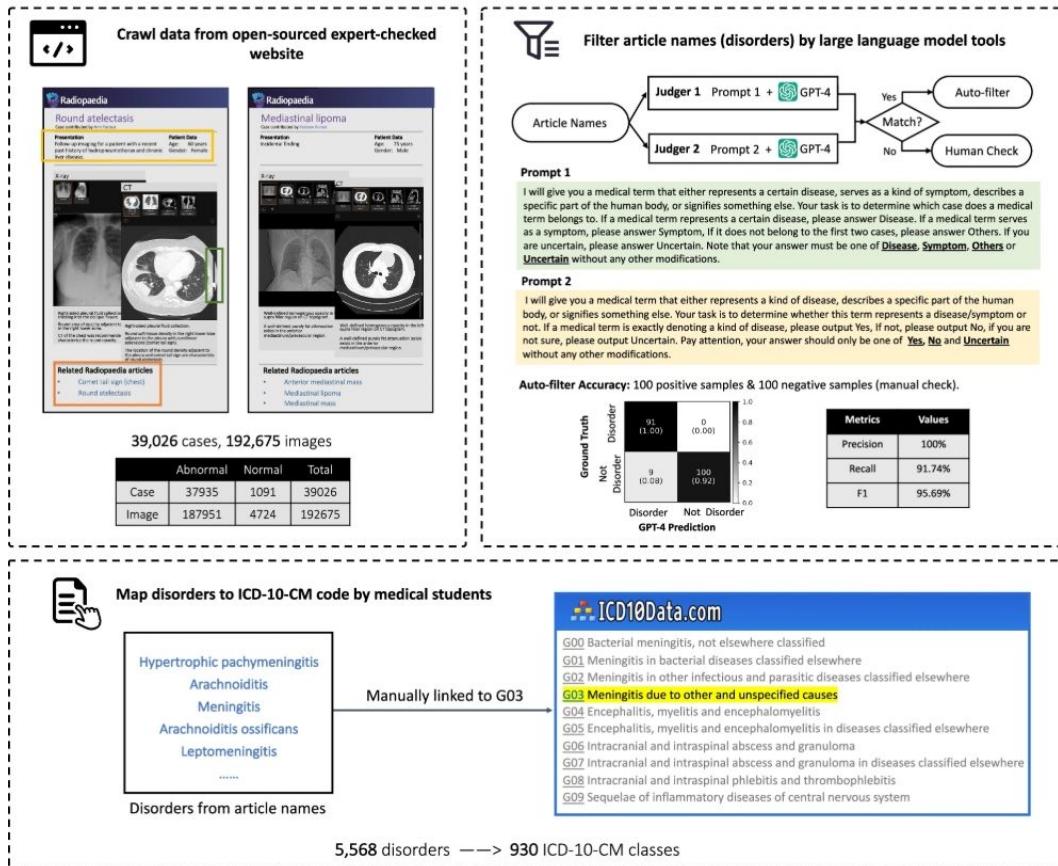


I. Where to access data

From websites

Radiopaedia: <https://radiopaedia.org/home>

Data Curation Pipeline



Crawl data from open-sourced expert checked website
↓
Filter article names (disorders) by large language models

Map disorders to ICD-10-CM

[1] Wu, Chaoyi, et al. "Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data." *Nature Communications* 16.1 (2025): 7866.

[2] Large-scale long-tailed disease diagnosis on radiology images. *Nat Commun* 15, 10147 (2024).

$$\{\text{Radiology Images}\} + \{\text{Textual Prompt}\} = \{\text{Textual Answer}\}$$

- Chest X-rays (CXR) dominate most existing radiology datasets.
- Public sources such as websites and publications provide valuable data but often introduce distribution shifts compared with real-world clinical practice.

 We need more data beyond CXR from clinical practices, like CT-RATE;

- Current prompts are often too simple, missing essential clinical context and failing to capture multi-turn interactions.
- Responses are usually plain, lacking reasoning, justification, and detailed explanation.

 We need more diverse question–answer pairs.

Visual Instruction Tuning

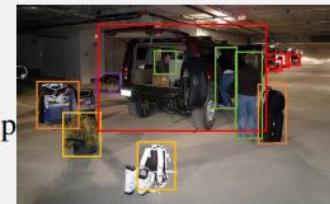
Llava: First Visual Instruction Tuning Attempt

Data:

- Use COCO images and generate three types of instruction-following data.
- Generates multimodal instruction-following data using language-only GPT-4 leveraging captions and bounding box.

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

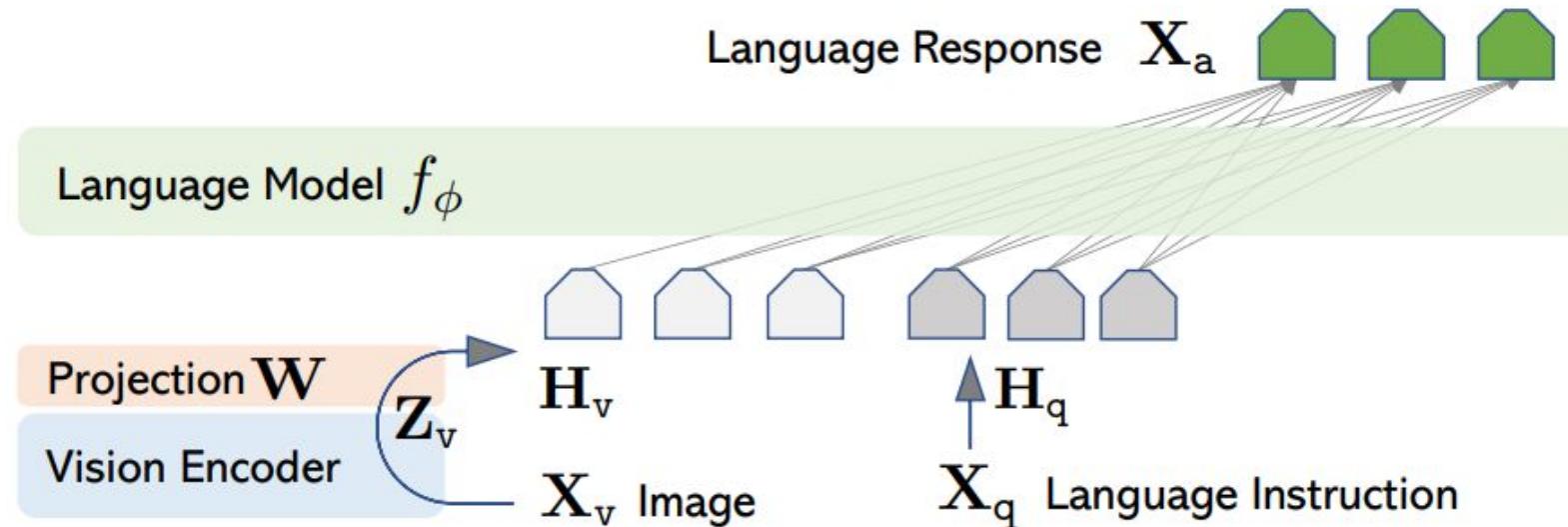
Question: What challenges do these people face?

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>

Llava: First Vison Instruction Tuning Attempt

Architecture (similar as BLIP-2):

- Connects a vision encoder with a large language model in an end-to-end framework.
- Self-attention based multimodal fusion, i.e,
 - View encoded images as “tokens” and concate with the text tokens directly.
 - Perform multimodal fusion, **reusing the self-attention layer** in language models to generate responses.



[1] Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2023): 34892-34916.

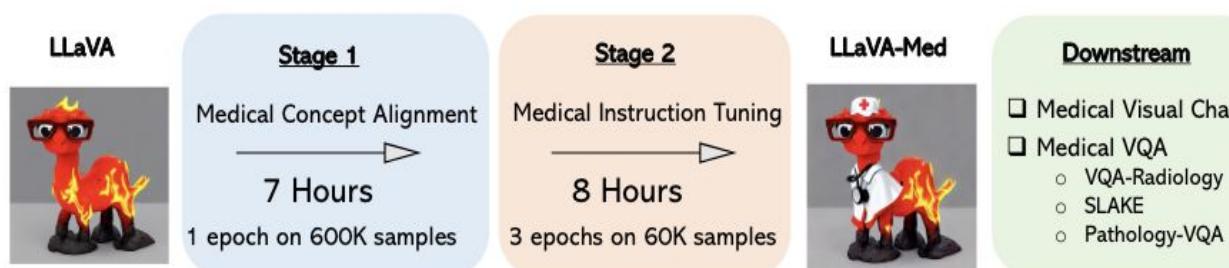
[2] Li, Junnan, et al. "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." *International conference on machine learning*. PMLR, 2023.

II. Visual Instruction Tuning

Llava-Med: First Medical MLLM

Model Training:

- LLM - Vicuna, Vision Encoder - CLIPViT-L 224px;
- 600K image-text pairs generated from PMC-15M (papers) using GPT-4;
- 2 staged training:
 - Train the projection layer with 600K pairs to align the image embedding space and text embedding space;
 - Tune the whole network with 60K high-quality GPT-4 augmented instruction 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.

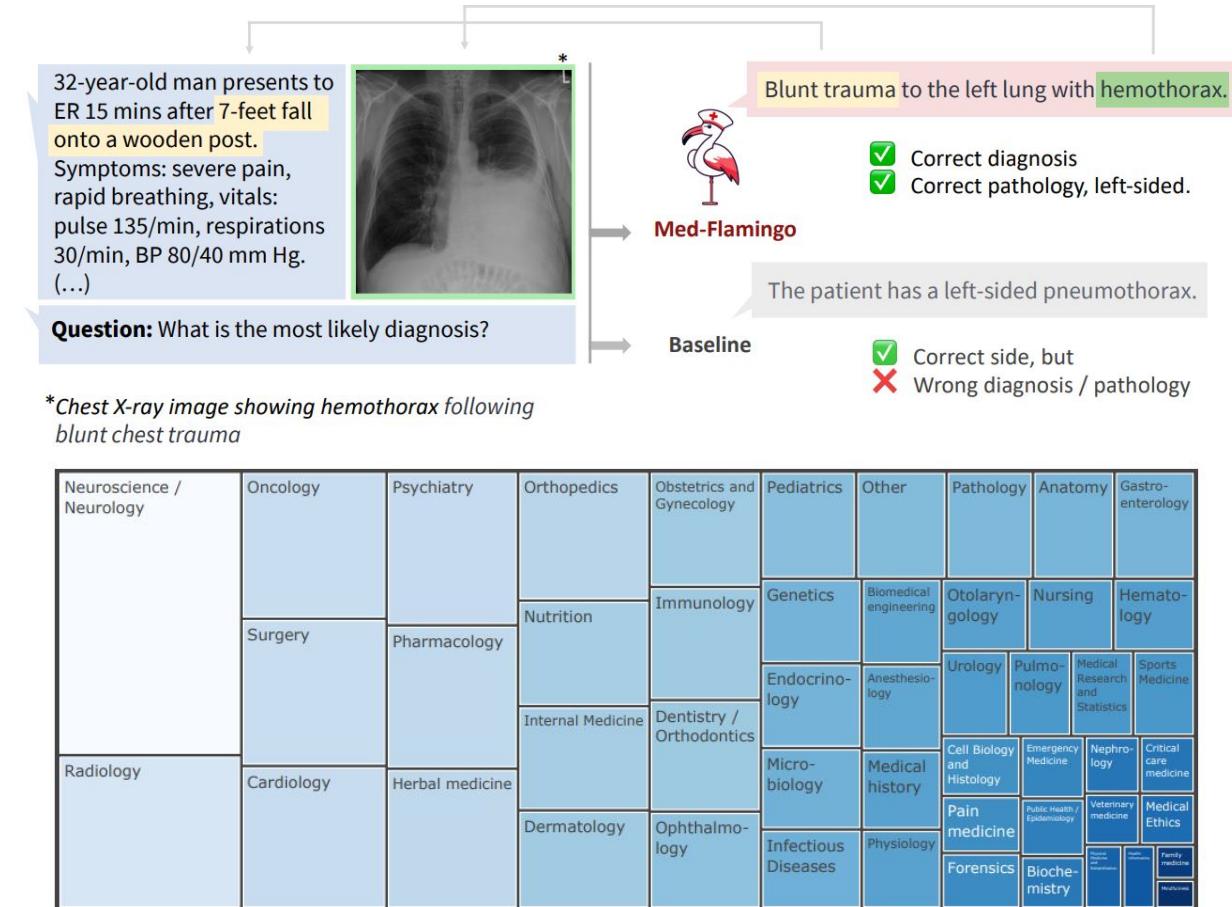
User: Can you explain more about the cavitating lesion?

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.

MedFlamingo: First Medical MLLMs based on Open-Flamingo

Data:

- Built a new interleaved image-text dataset from 4,721 medical textbooks, with robust cleaning and segmentation
- Paired data sourced from PMC-OA: 1.6M biomedical image-caption pairs collected from PubMedCentral



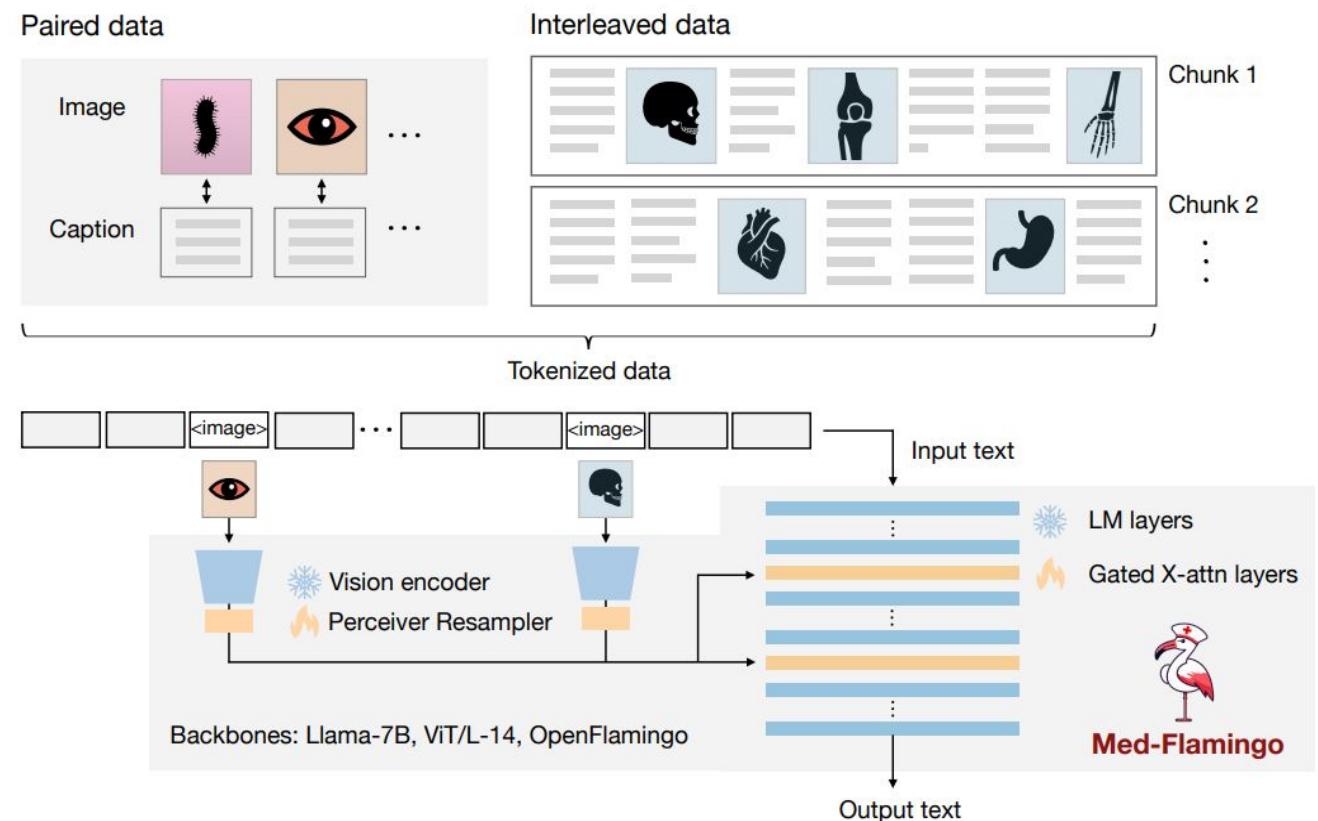
[1] Moor, Michael, et al. "Med-flamingo: a multimodal medical few-shot learner." *Machine Learning for Health (ML4H)*. PMLR, 2023.

[2] PMC-OA: Lin, Weixiong, et al. "Pmc-clip: Contrastive language-image pre-training using biomedical documents." *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Cham: Springer Nature Switzerland, 2023.

MedFlamingo: First Medical VLM

Architecture (Similar as Flamingo):

- Project images into tokens
- Fuse image tokens and image tokens with **newly added cross-attention layers** across each layer in the LLMs.
- Since this design disrupts the original integrity of the LLM, only a few studies have continued to adopt this architecture. Instead, the self-attention design has remained the mainstream choice.

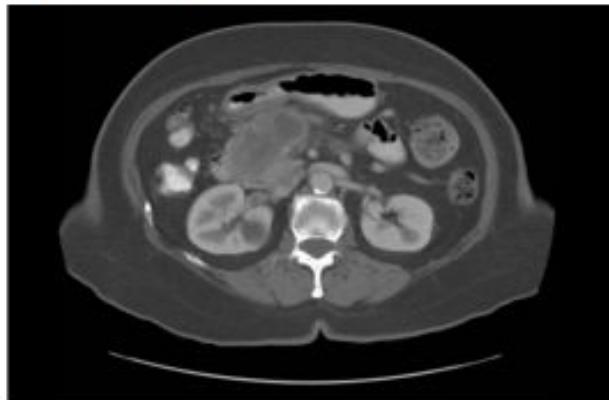


MedFlamingo: First Medical VLM

Model Training:

- Pre-training Med-Flamingo on large-scale, curated medical image-text data across multiple disciplines
- Instruction tuning using **few-shot** generative visual question answering.

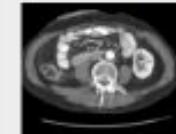
Problem



Question: What do the small white lesions in the aorta mean?

Multimodal few-shot prompt

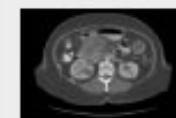
Instruction



Question,
Answer



Question,
Answer



Question

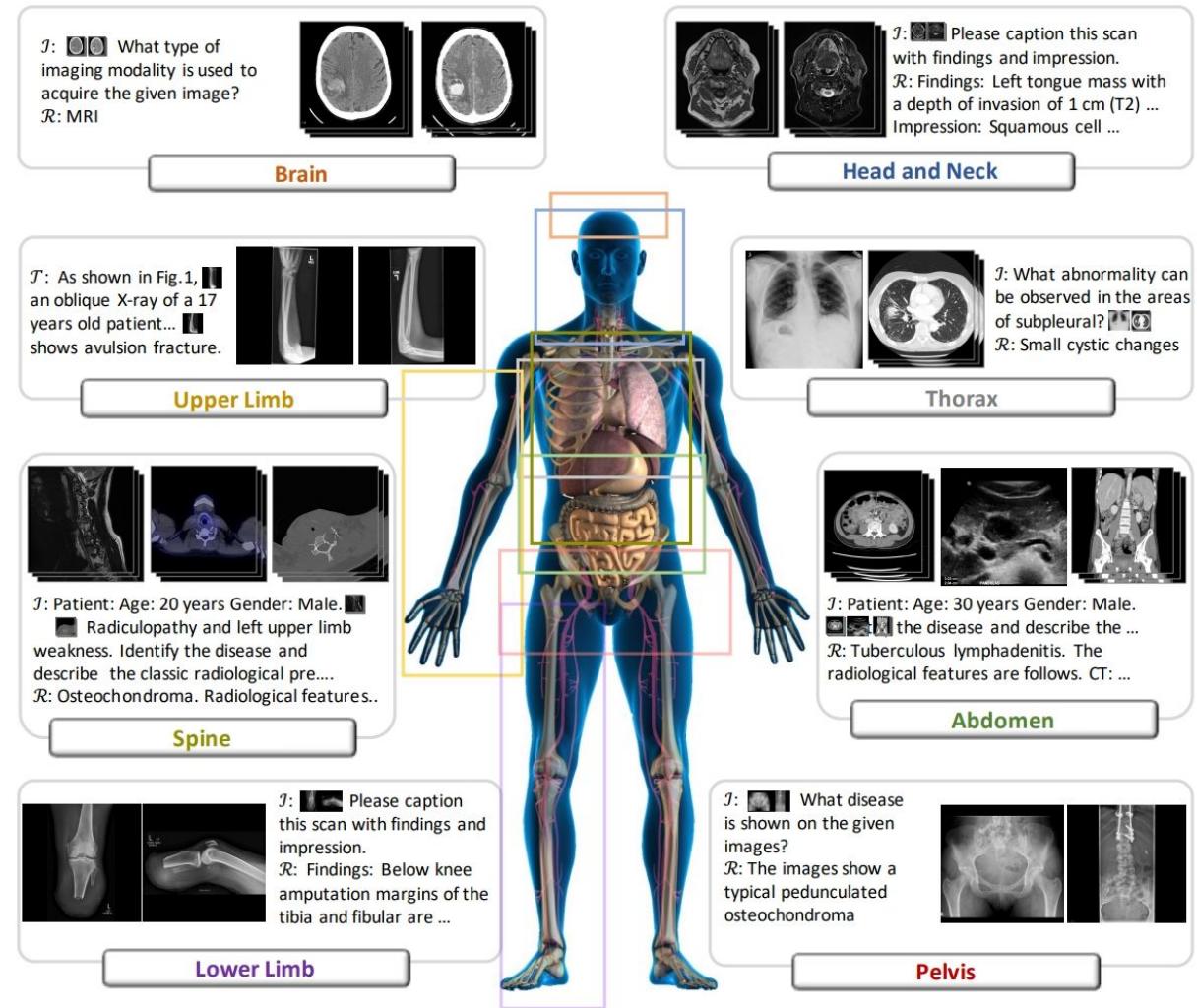
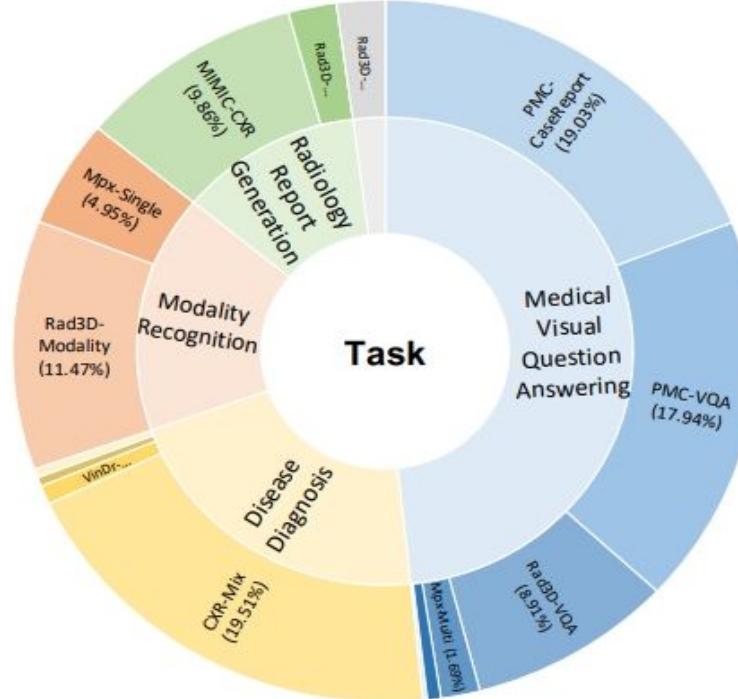
Answer: Calcification of the aortic wall.



RadFM: Foundation Model in Radiology

Data

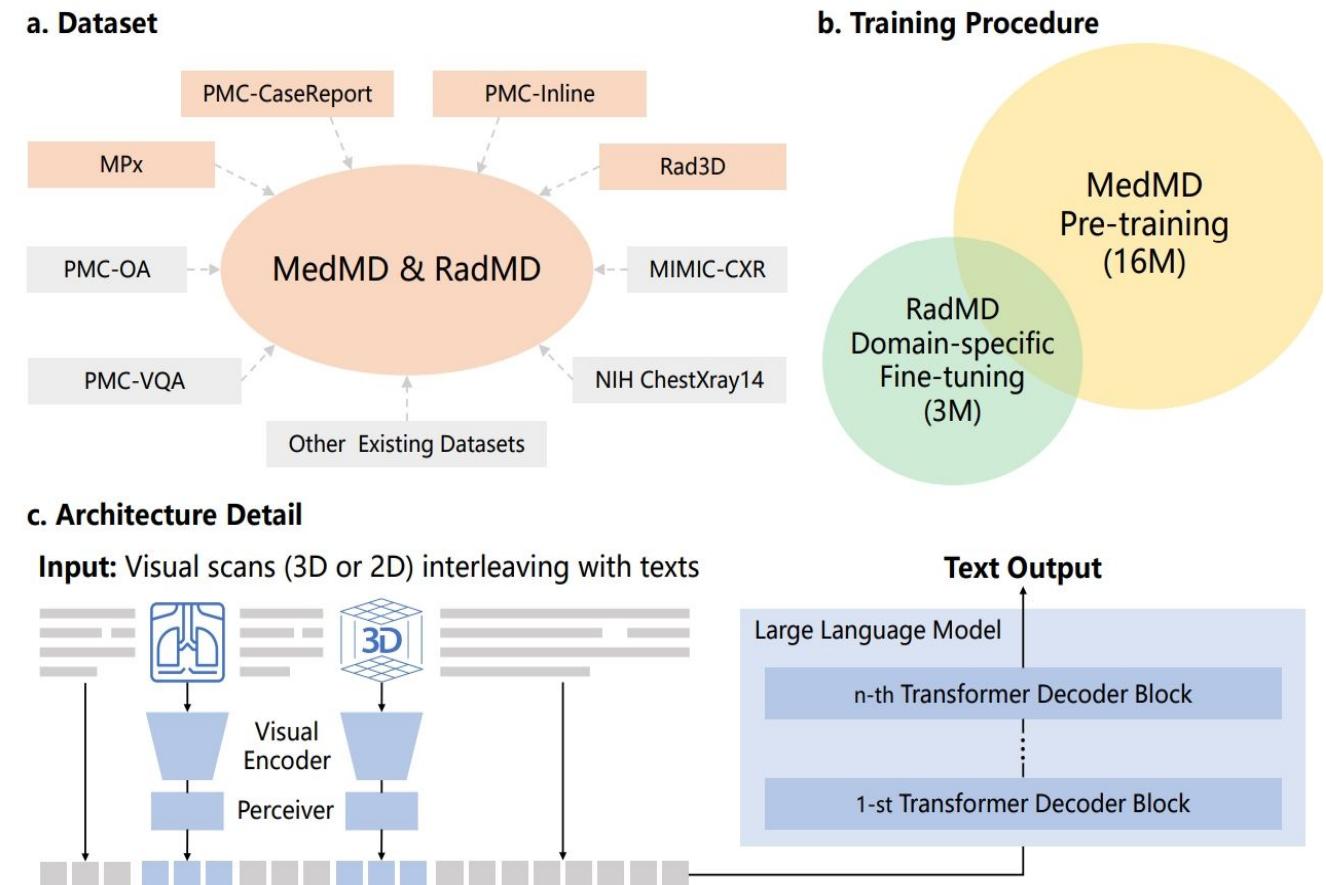
- Diverse tasks including VQA, Diagnosis, and Report Generation;
- Beyond paper data, 16M 2D and 3D scans.



RadFM: Foundation Model in Radiology

Architecture and Training

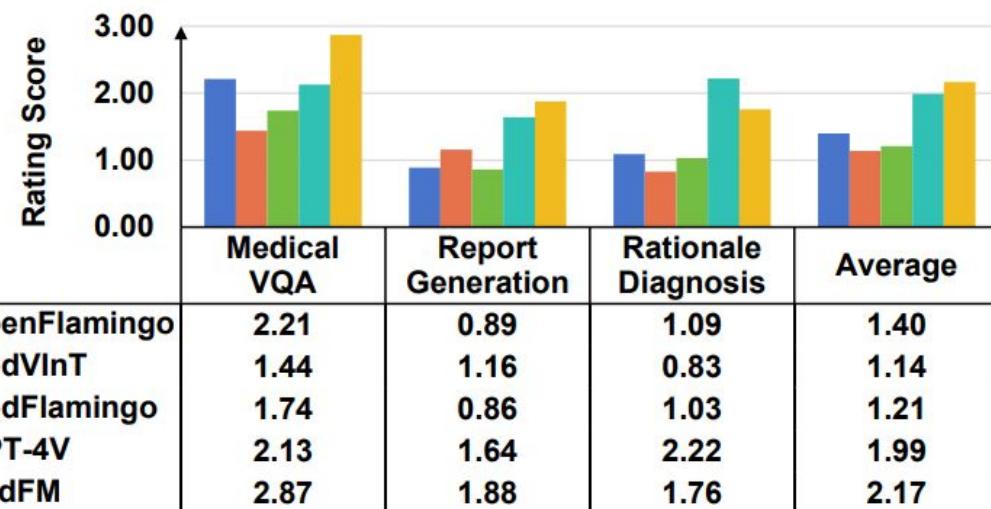
- **Architecture slightly different due to 3D**
 - MedLLaMA-13B, medical version of LLaMA model trained with PMC
 - **3D ViT + Perceiver** for Vision Encoder
 - Interleaved images with text with texts.
- **2-staged training**
 - Auto-regressive Pretraining (Predict next token)
 - Visual Instruction Tuning (Domain-specific Finetuning)
 - The main difference falls in what tokens are used for loss calculation.



RadFM: Foundation Model in Radiology

Results

- Outperform SoTA on various Benchmark
- A comprehensive human rating also shows RadFM's superior performance.

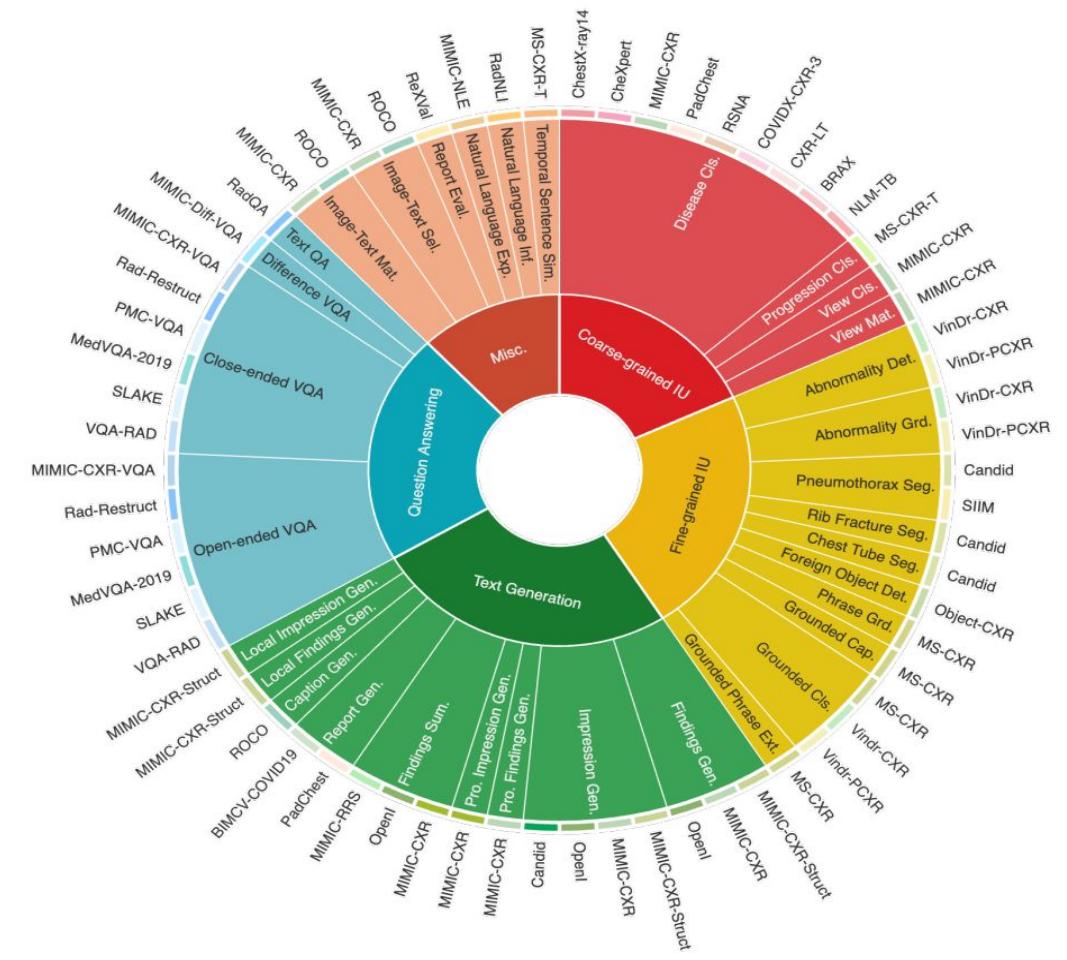
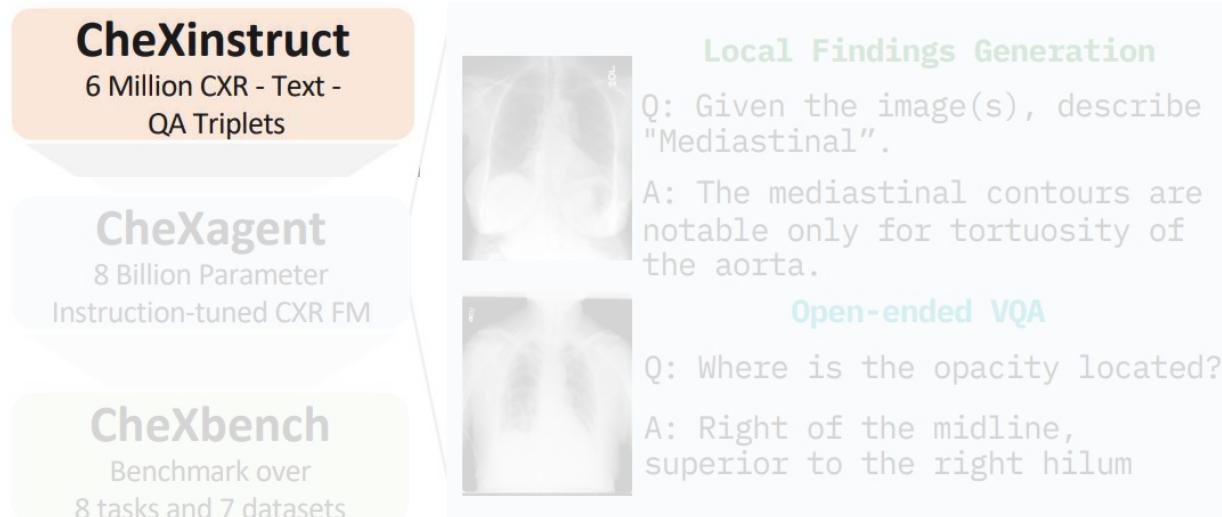


Dataset	Modality	Metric	SOTA	RadFM
Disease Diagnosis				
VinDr-Mammo	2D Mammography	Macro-AUC	64.5 [47]	64.76 (64.23, 65.88)
		Macro-F1	N/A	39.42 (39.37, 39.59)
CXR14	2D X-ray	Macro-AUC	80.1 [49]	81.13 (81.07, 81.18)
		Macro-F1	N/A	30.20 (30.17, 30.22)
LDCT	3D CT	Macro-AUC	82.1 [47]	83.23 (81.97, 85.85)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

CheXagent: First MLLMs Specialized to Chest X-ray Application

Data

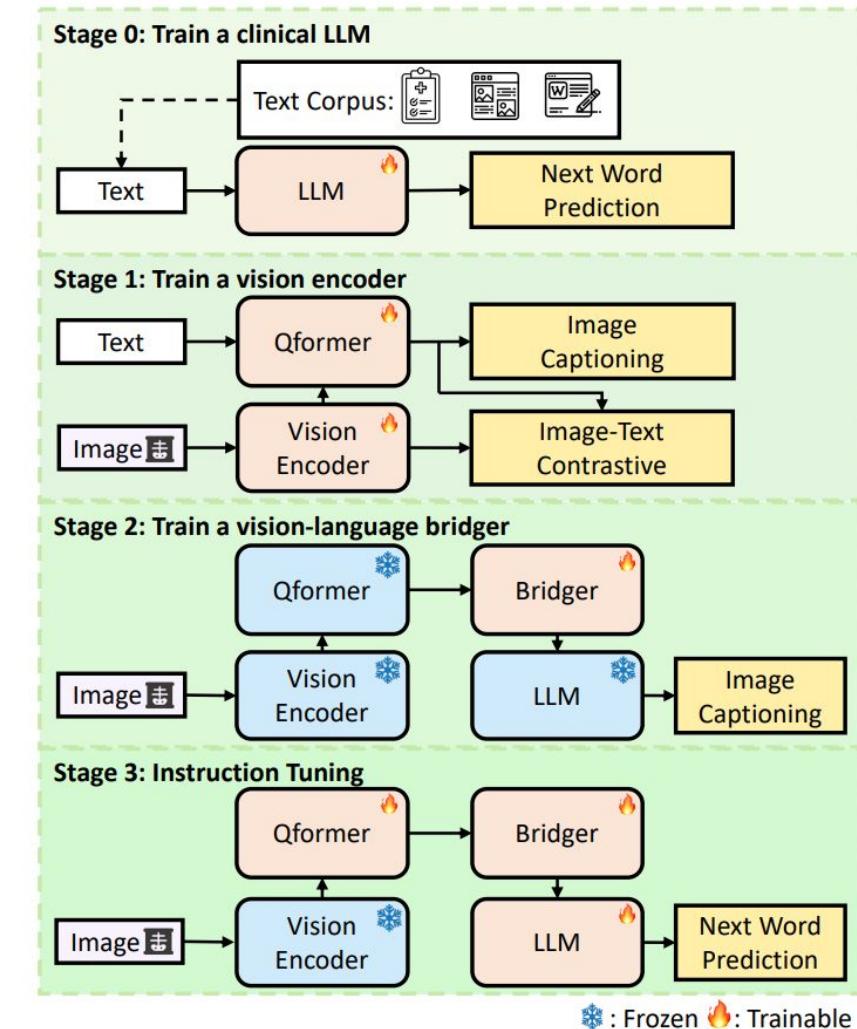
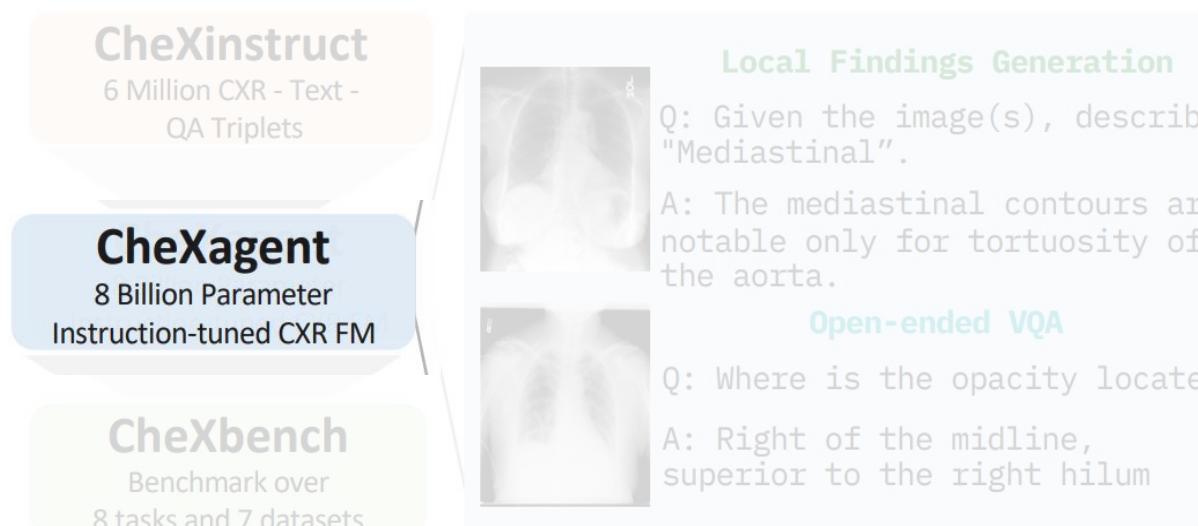
- More careful collection, filtering, and organization of data in the CXR domain with **6M CXR QA pairs**.
 - Diverse tasks such as VQA, NLI, Diagnosis ...



CheXagent: First Chest X-ray Application

Model Training:

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



CheXagent: First Chest X-ray Application

Results:

- Two Evaluation Method
 - Automated metrics
 - GPT-4 as Judge
- Outperforms state-of-the-art general and medical foundation model in CXR.

Model	Size	BERT-S	Private Dataset		MIMIC-CXR	
			CheXbert-S	RadGraph-S	BERT-S	CheXbert-S
MedFlamingo	8B	8.5	2.7	1.7	10.4	3.2
LLaVA-Med	8B	12.5	17.0	4.2	6.2	17.5
RadFM	14B	35.7	12.7	5.1	45.7	17.5
XrayGPT	8B	40.1	23.4	9.0	44.0	24.2
CheXagent	8B	46.6	23.7	14.6	50.4	24.9
						18.6

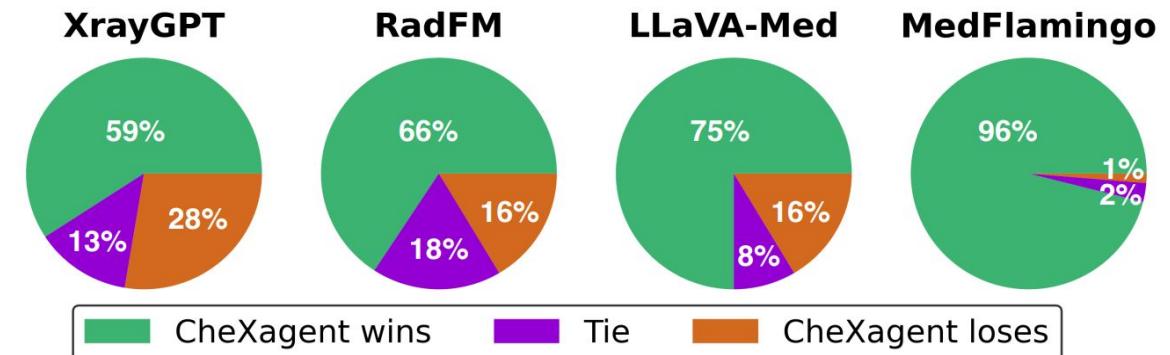
CheXinstruct
6 Million CXR - Text - QA Triplets



Local Findings Generation
Q: Given the image(s), describe "Mediastinal".
A: The mediastinal contours are notable only for tortuosity of the aorta.

Open-ended VQA
Q: Where is the opacity located?
A: Right of the midline, superior to the right hilum



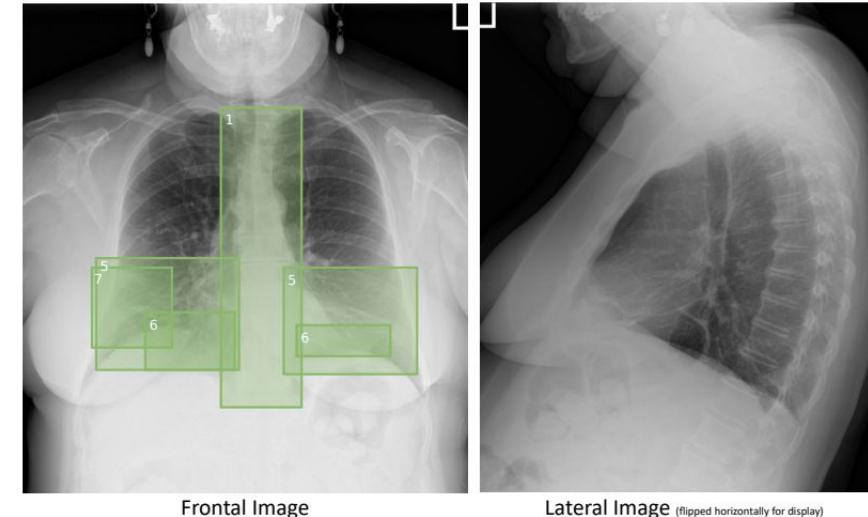


Visual Instruction Tuning

MAIRA-2

Data:

- Each sample includes images, detailed findings, and extra context like indication, technique, and comparison sections when available.
- Combines public (MIMIC-CXR, PadChest, IU-Xray) and private (USMix) chest X-ray report generation datasets.



Indication: Cough and wheezing for 5 months.
Technique: PA and lateral views of the chest were obtained.
Comparison: None
Prior Report: None

Findings:

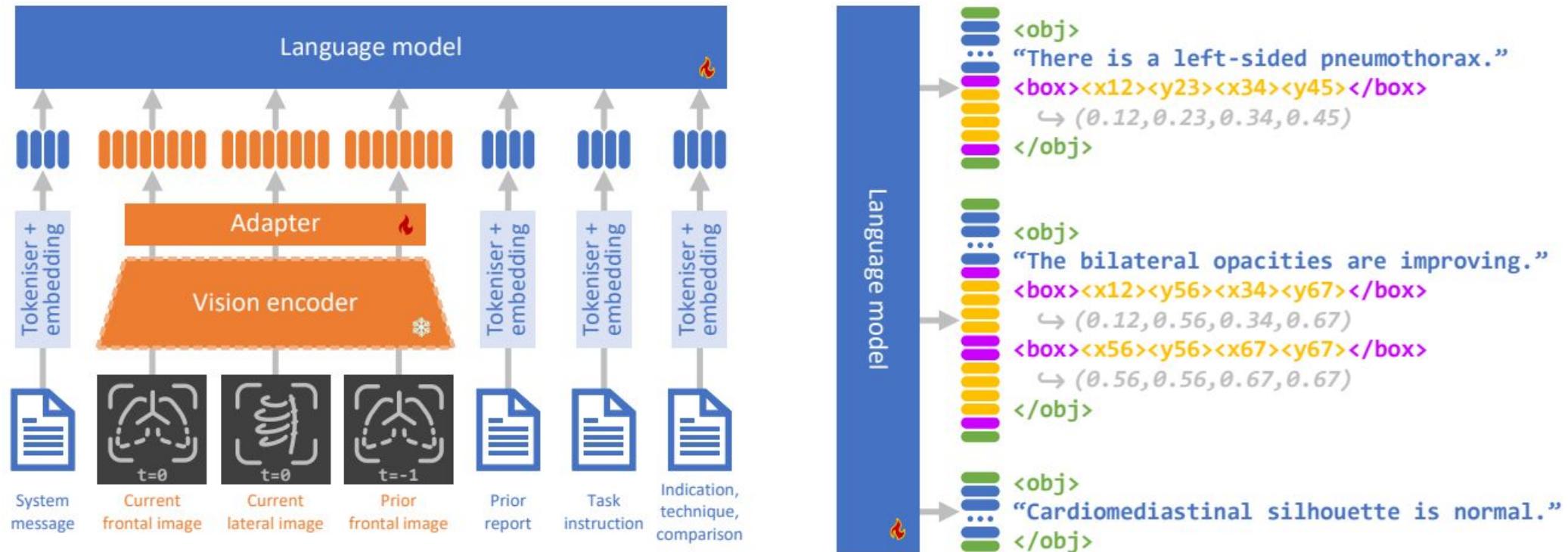
1. Degenerative changes seen in the dorsal spine.
2. The lungs show no mass.
3. The lungs show no effusion.
4. Diaphragms are sharp.
5. Prominent interstitial lung markings at lung bases more the right than the left with some consolidation of markings.
6. Atelectasis is noted.
7. Infiltrate in the right middle lobe laterally.
8. Cardiac size is normal.
9. There is no hilar adenopathy.
10. There is no mediastinal adenopathy.

Data source	Subset	Task	# Patients		# Samples		% Has Lateral		% Has Prior	
			Train	Test	Train (%)	Test	Train	Test	Train	Test
MIMIC-CXR	All	FindGen	55 218	285	158 555 (31%)	2461	60.6	45.3	64.2	88.6
	MS-CXR	PhraseGround	595	128	817 (0.2%)	176	0	0	0	0
PadChest	All	FindGen	52 828	1559	85 598 (17%)	2925	46.0	50.4	38.3	48.1
	PadChest-GR	GroundRep	3122	893	3183 (0.6%)	915	44.7	45.7	32.3	31.7
USMix	All	FindGen	118 031	–	193 652 (38%)	–	51.7	–	0	–
	GR-1	GroundRep	45 155	–	60 463 (12%)	–	48.0	–	0	–
	GR-Bench	GroundRep	8458	1199	8580 (1.7%)	1231	81.2	79.8	0	0
IU-Xray	All	FindGen	–	3198*	–	3306	–	92.1	–	0
Total			226 077	–	510 848 (100%)	–	53.4	–	26.5	–

MAIRA-2

Architecture:

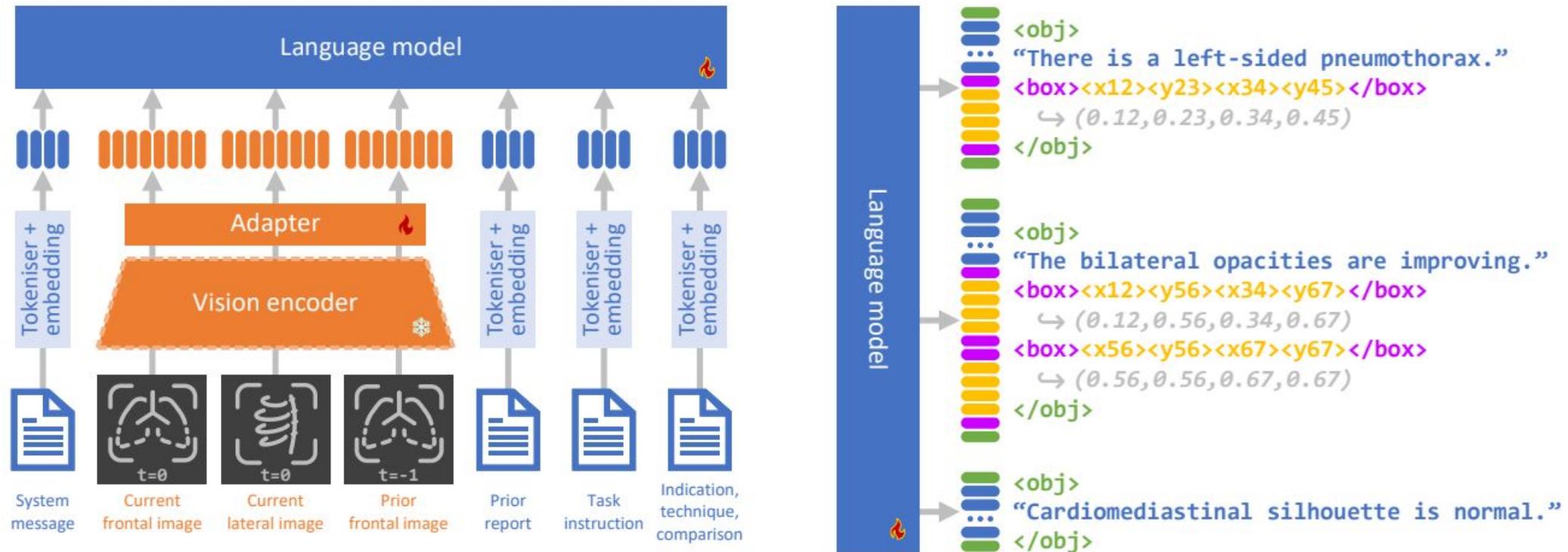
- Encode multi-image with a frozen vision encoder (Rad-DINO-MAIRA-2) and training an adapter
- Self-attention based fusion design with LLMs



MAIRA-2

Model Training:

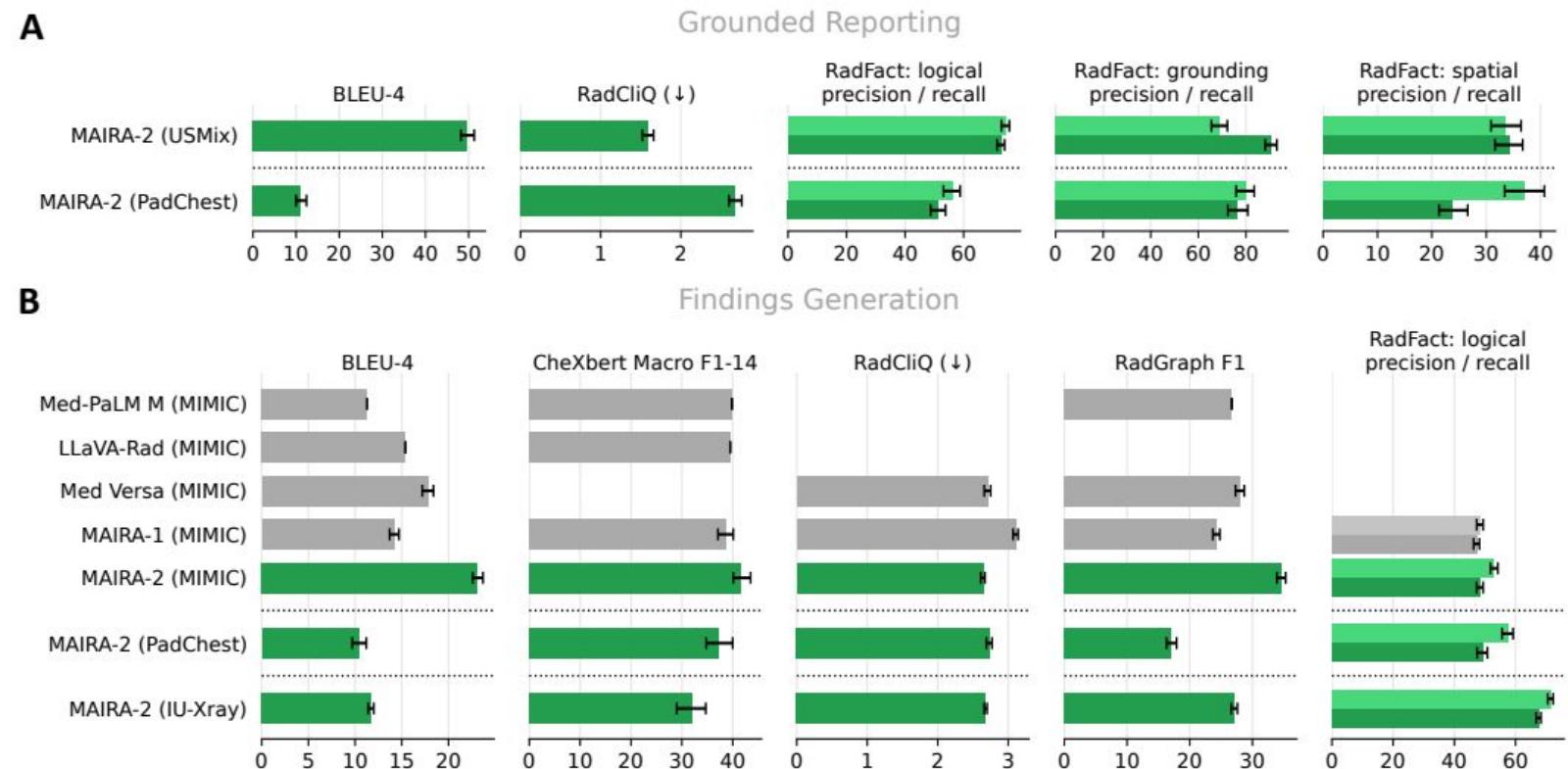
- Visual Instruction-tuning with **grounded report generation**: generates findings with linked bounding boxes in texts to localize abnormalities on the image.



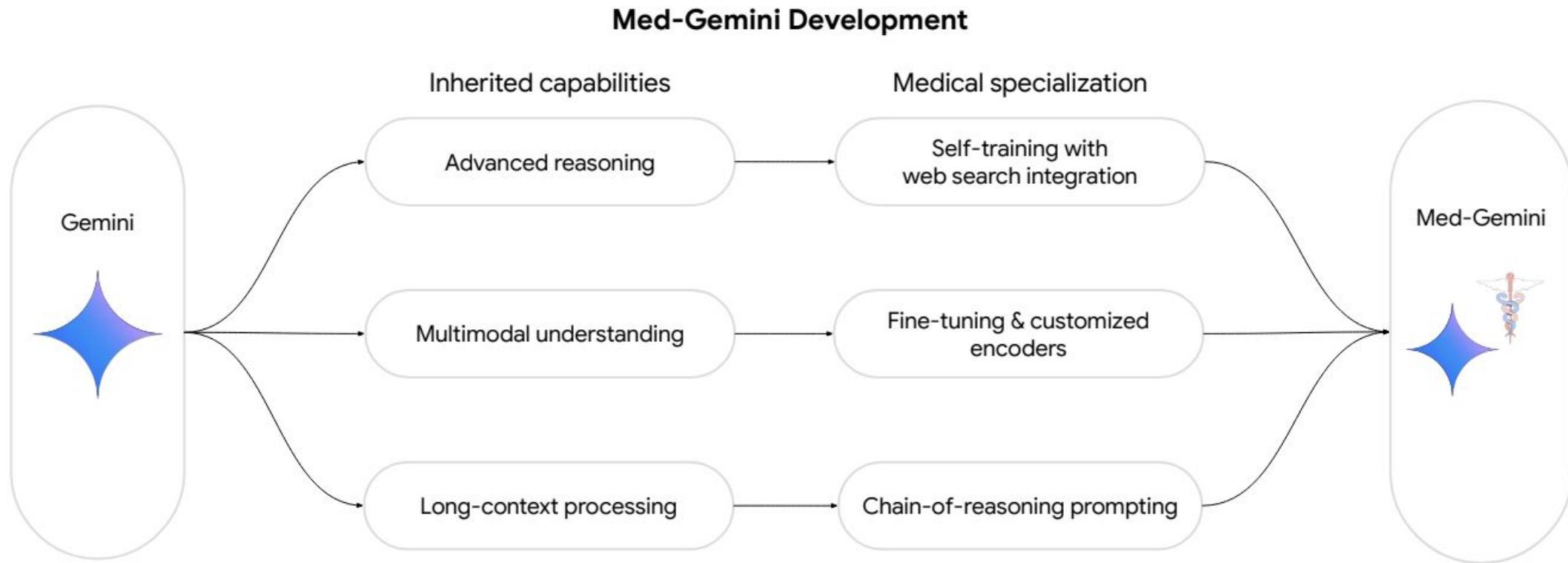
MAIRA-2

Results:

- MAIRA-2 achieves top logical precision (>50%) and high grounding precision on new grounded report tasks.
- Sets new SoTA on traditional report generation benchmarks, outperforming prior models and generalizing well to unseen data.



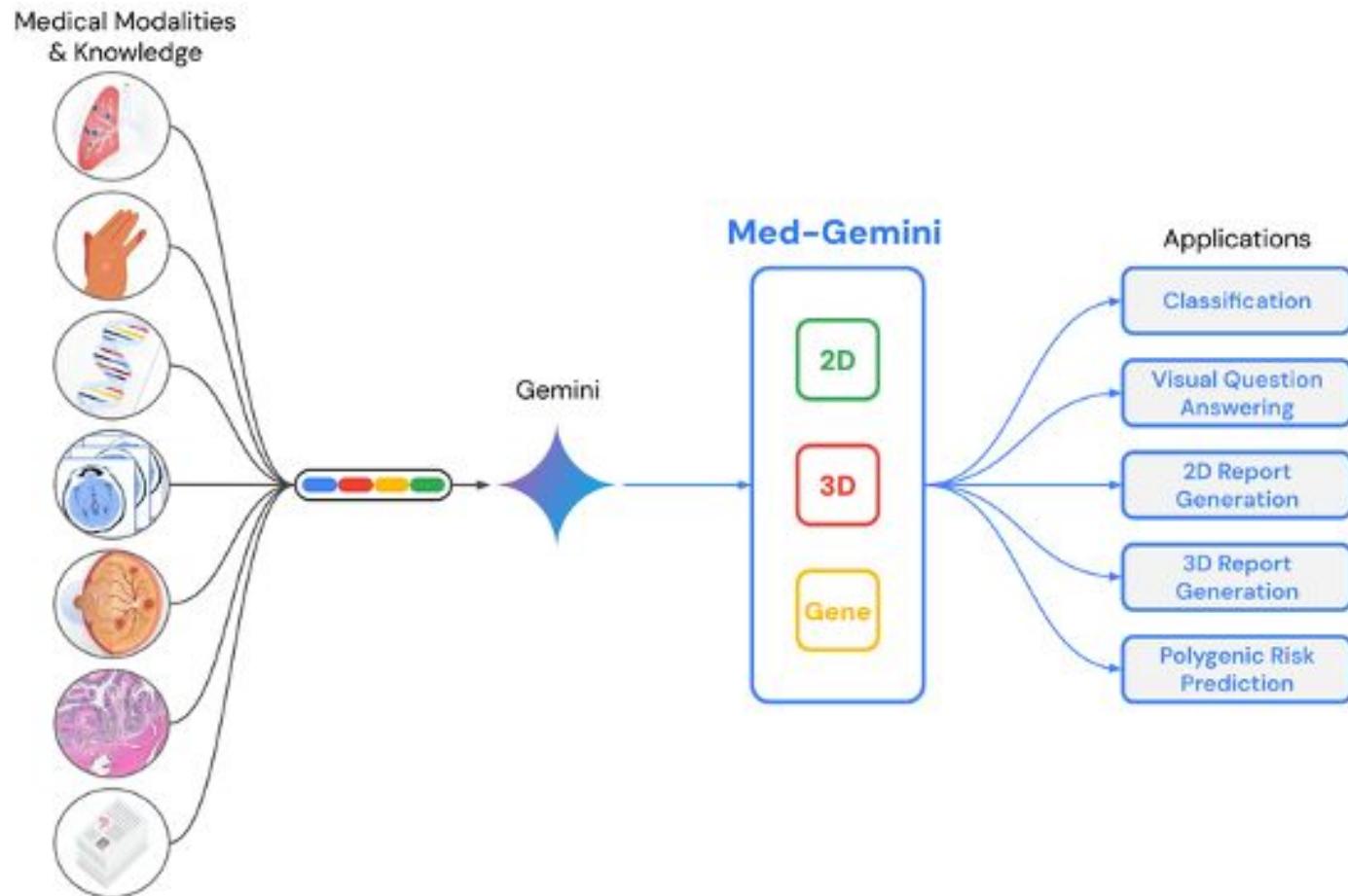
MedGemini



[1] Saab, Khaled, et al. "Capabilities of gemini models in medicine." *arXiv preprint arXiv:2404.18416* (2024).

[2] Yang, Lin, et al. "Advancing multimodal medical capabilities of Gemini." *arXiv preprint arXiv:2405.03162* (2024).

MedGemini



[1] Saab, Khaled, et al. "Capabilities of gemini models in medicine." *arXiv preprint arXiv:2404.18416* (2024).

[2] Yang, Lin, et al. "Advancing multimodal medical capabilities of Gemini." *arXiv preprint arXiv:2405.03162* (2024).

MedGemini

Data:

More than **7 million data samples**
from 3.7 million medical images
and cases is used for fine-tuning
 and further instruction-tuning of
 Gemini for medical applications in
 Med-Gemini

Modality	Dataset	No. examples	No. Images	Description
Radiology (2D)	Slake-VQA	4,919	450	Radiology images & QA pairs
	MIMIC-CXR	2,142,892	231,483	Radiology images & free-form reports
	Digital Knee X-ray	1,469	1,469	Knee X-ray images & labels
	CXR-US2	132,680	132,680	Radiology images & free-form reports
	NLST	2,199	2,199	2D CT slices & free-form reports
	CT-US1	3,207	3,207	2D CT slices & free-form reports
Radiology (3D)	CT-US1	657,719	657,719	3D CT images & free-form reports
Pathology	PathVQA	19,654	2,599	Pathology images & QA pairs
	Histopathology	1,550,976	207,603	Histopathology images, captions, & QA pairs
Dermatology	PAD-UFES-20	2,047	2,047	Skin lesion images & labels
Ophthalmology	EyePACS	14,406	14,406	Fundus images & labels
Medical VQA	PMC	2,246,656	2,246,656	PubMed Central images & caption pairs
	MedVQA	12,664	3,168	Medical images & QA pairs
Genomics	UK Biobank	259,225	259,225	Genomic data & disease outcomes

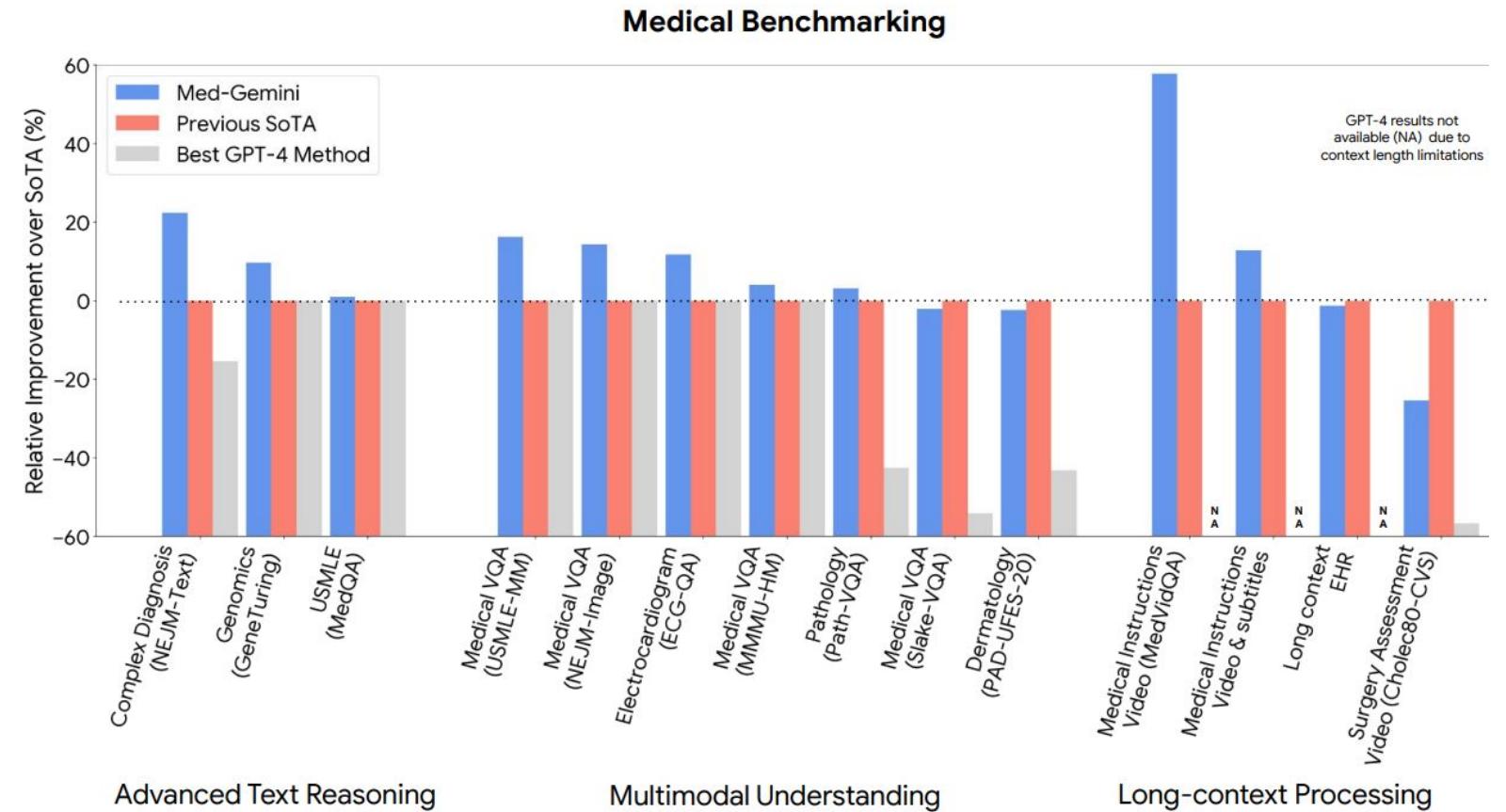
[1] Saab, Khaled, et al. "Capabilities of gemini models in medicine." *arXiv preprint arXiv:2404.18416* (2024).

[2] Yang, Lin, et al. "Advancing multimodal medical capabilities of Gemini." *arXiv preprint arXiv:2405.03162* (2024).

MedGemini

Results:

- Evaluation of advanced reasoning on text-based tasks.
- Evaluation of multimodal capabilities- 2D, 3D, Signal QA.
- Evaluation of long-context capabilities on video and EHR tasks



[1] Saab, Khaled, et al. "Capabilities of gemini models in medicine." *arXiv preprint arXiv:2404.18416* (2024).

[2] Yang, Lin, et al. "Advancing multimodal medical capabilities of Gemini." *arXiv preprint arXiv:2405.03162* (2024).

II. Visual Instruction Tuning

MedGemini

Example of Med-Gemini's long-context capabilities on a Chest X-ray image.

Here, Med-Gemini helps interpret a patient's chest X-ray. It produces a report noting a normal heart, clear lungs, and mild age-related spine changes. When asked, it explains these findings in detail and emphasizes that they may not fully account for the patient's back pain. It then rephrases the results in simple, patient-friendly language. The official radiology report—unseen by the model—closely matches its output, showing Med-Gemini's accuracy and flexibility in both clinical and patient communication.



Hello, I am a primary care physician and this is a chest x-ray for a patient of mine. The formal radiology report is still pending and I would like some help to understand the x-ray. Please write a radiology report for me:

FINDINGS: The heart is at the upper limits of normal size. The mediastinal and hilar contours appear within normal limits. The lungs appear clear. There are no pleural effusions or pneumothorax. Mild degenerative changes are noted along the lower thoracic spine. IMPRESSION: No evidence of acute disease.

What evidence do you see for degenerative changes?

Mild degenerative changes are noted along the lower thoracic spine. This can be seen by the slight narrowing of the intervertebral disc spaces and the presence of small osteophytes (bone spurs) along the anterior margins of the vertebral bodies.

My patient has a history of back pain. Might this explain the back pain?

While the degenerative changes seen on the x-ray could potentially contribute to your patient's back pain, it's important to note that correlation doesn't equal causation. Many people have degenerative changes in their spine without experiencing pain, and conversely, back pain can have numerous causes beyond just degenerative changes.

Therefore, while the x-ray findings are worth noting, they don't definitively explain your patient's back pain. A comprehensive evaluation, including a detailed history, physical examination, and potentially additional imaging or tests, would be necessary to determine the exact cause of their pain.

Please let me know if you have any other questions.

To help me communicate all of this to my patient, please rewrite the report without medical terminology:

Chest X-ray Results:

- Heart: Normal size.
- Lungs: Clear, no signs of infection or fluid buildup.
- Bones: Minor wear and tear in the lower spine, common with age.

Overall: No signs of serious problems in the chest. The wear and tear in the lower spine might be related to your back pain, but other factors could also be involved. We'll need to do more tests to figure out the exact cause of your pain.

Thank you. That is useful while I wait for the radiologists report.

You're welcome! I'm glad I could help. If you have any further questions, feel free to ask.

Ground truth report (not provided to model)

FINDINGS: The cardiomediastinal and hilar contours are normal. The lungs are well expanded and clear, without focal consolidation, pleural effusion or pneumothorax. Mild degenerative changes are seen in the thoracic spine. IMPRESSION: No acute cardiopulmonary pathology.

[1] Saab, Khaled, et al. "Capabilities of gemini models in medicine." *arXiv preprint arXiv:2404.18416* (2024).

[2] Yang, Lin, et al. "Advancing multimodal medical capabilities of Gemini." *arXiv preprint arXiv:2405.03162* (2024).

SLaVA-CXR: three-stage visual instruction tuning

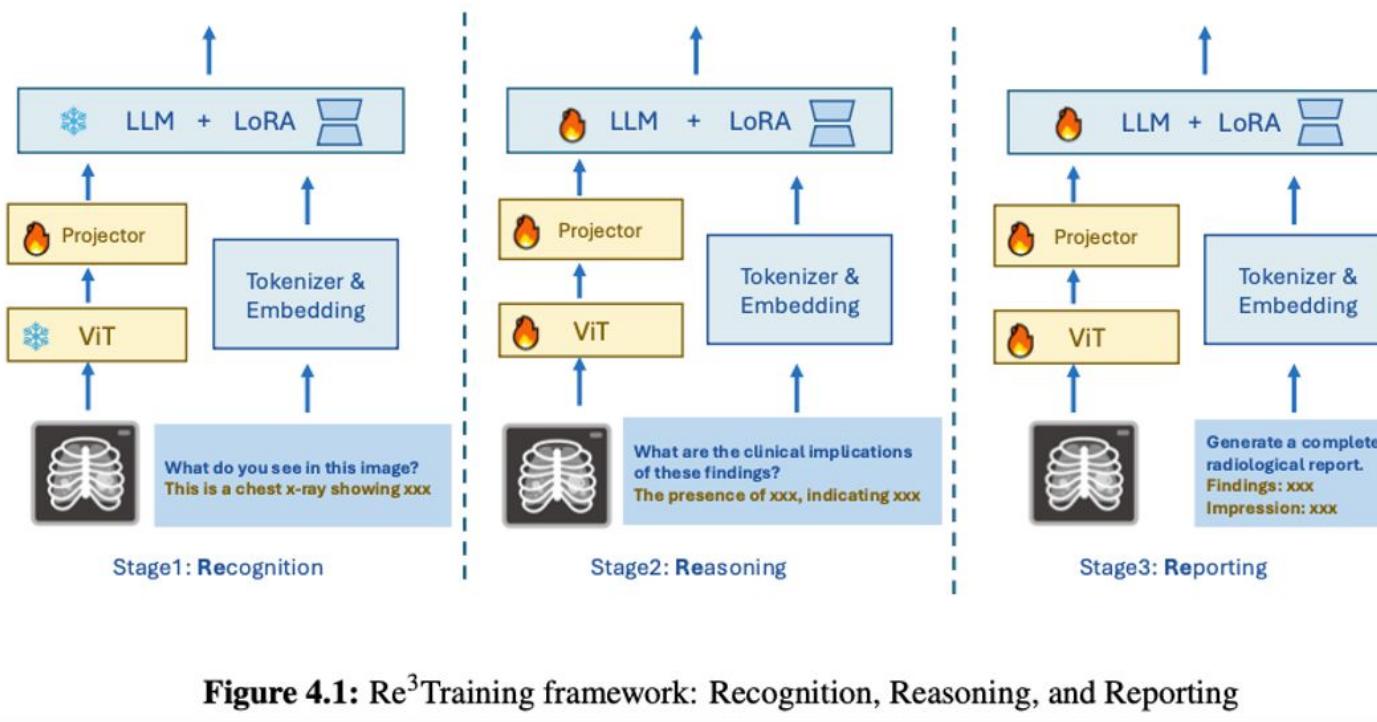


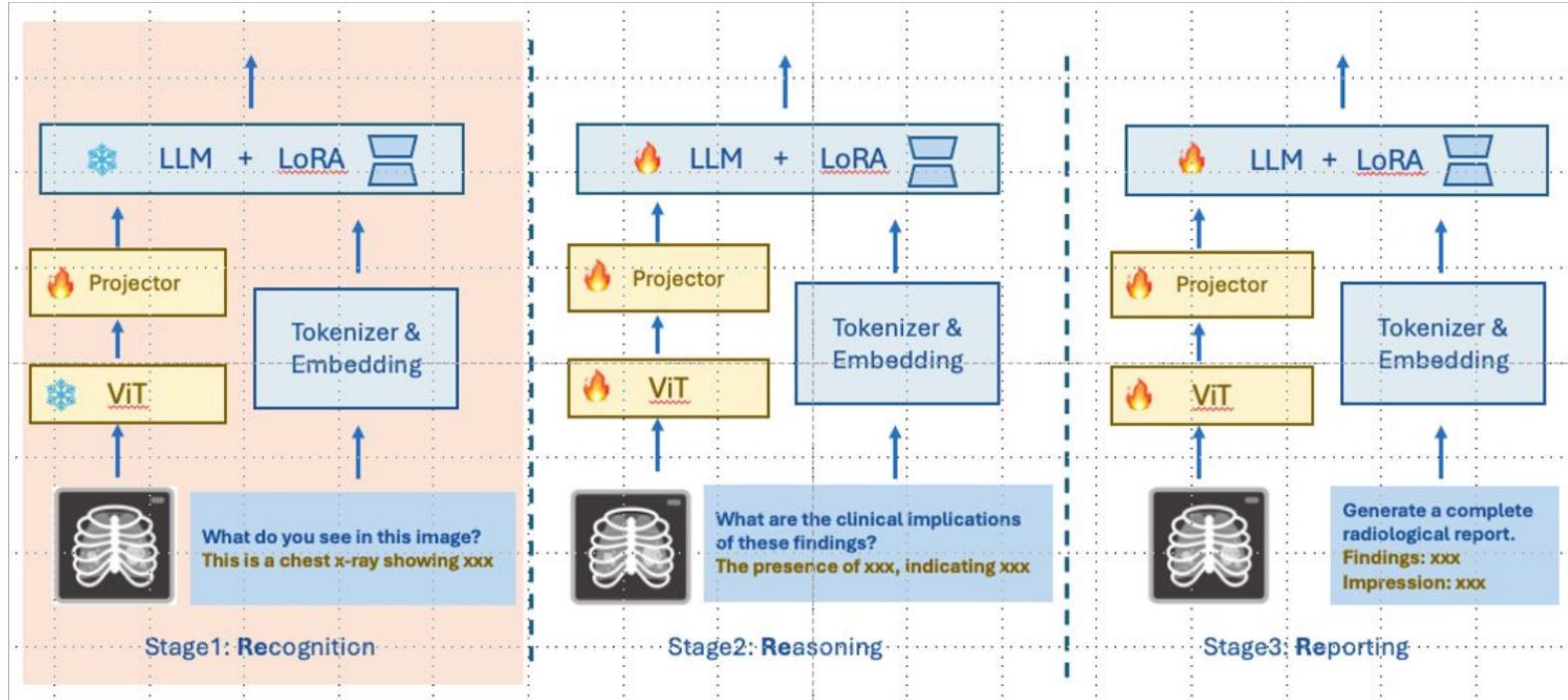
Figure 4.1: Re³Training framework: Recognition, Reasoning, and Reporting

Re³Training: Progressive Learning Framework

- **Recognition (Foundation Building)**
- **Reasoning (Diagnostic Development)**
- **Reporting (Clinical Integration)**

II. Visual Instruction Tuning

SLaVA-CXR: three-stage visual instruction tuning

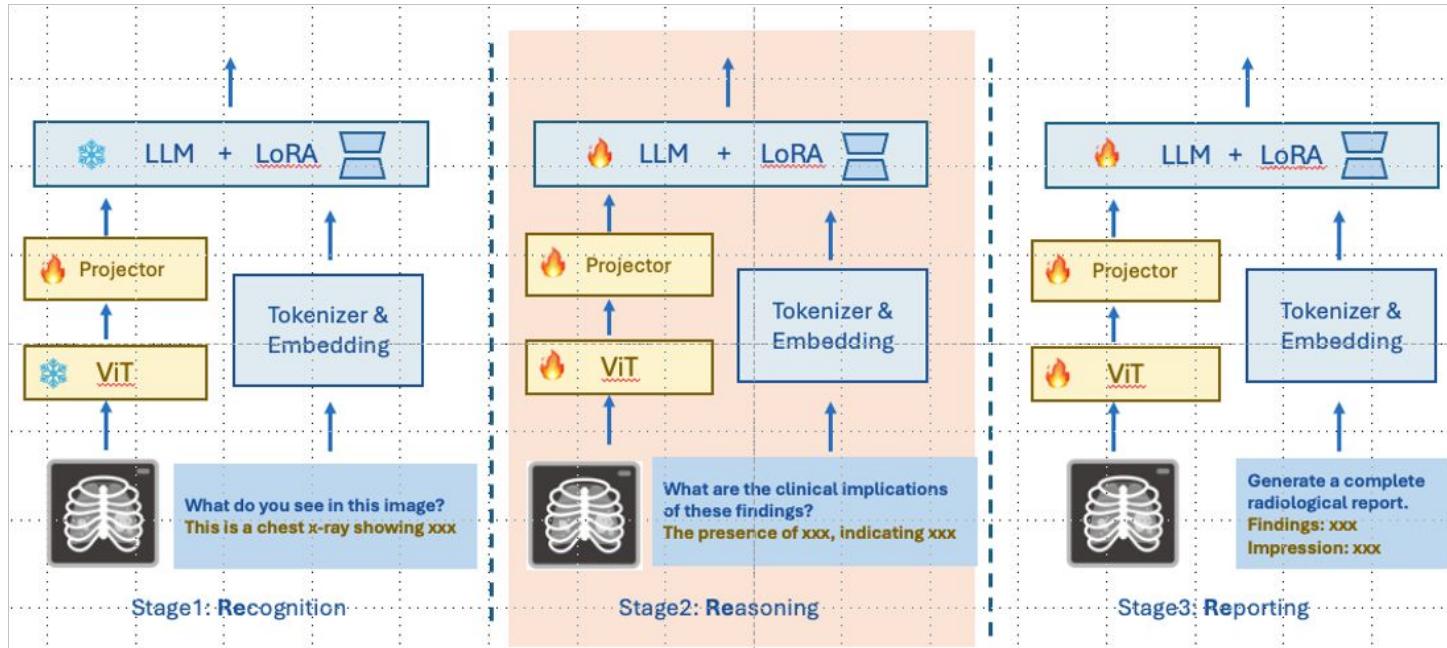


Re³Training: Progressive Learning Framework

Stage 1 - Recognition (Foundation Building)

- Dataset:** 560k general + 1,436 CXR-specific image-text pairs
- Objective:** Basic visual-semantic alignment (projector training only)
- Learning:** "What type of medical image is this?" → "This is a chest X-ray"

SLaVA-CXR: three-stage visual instruction tuning

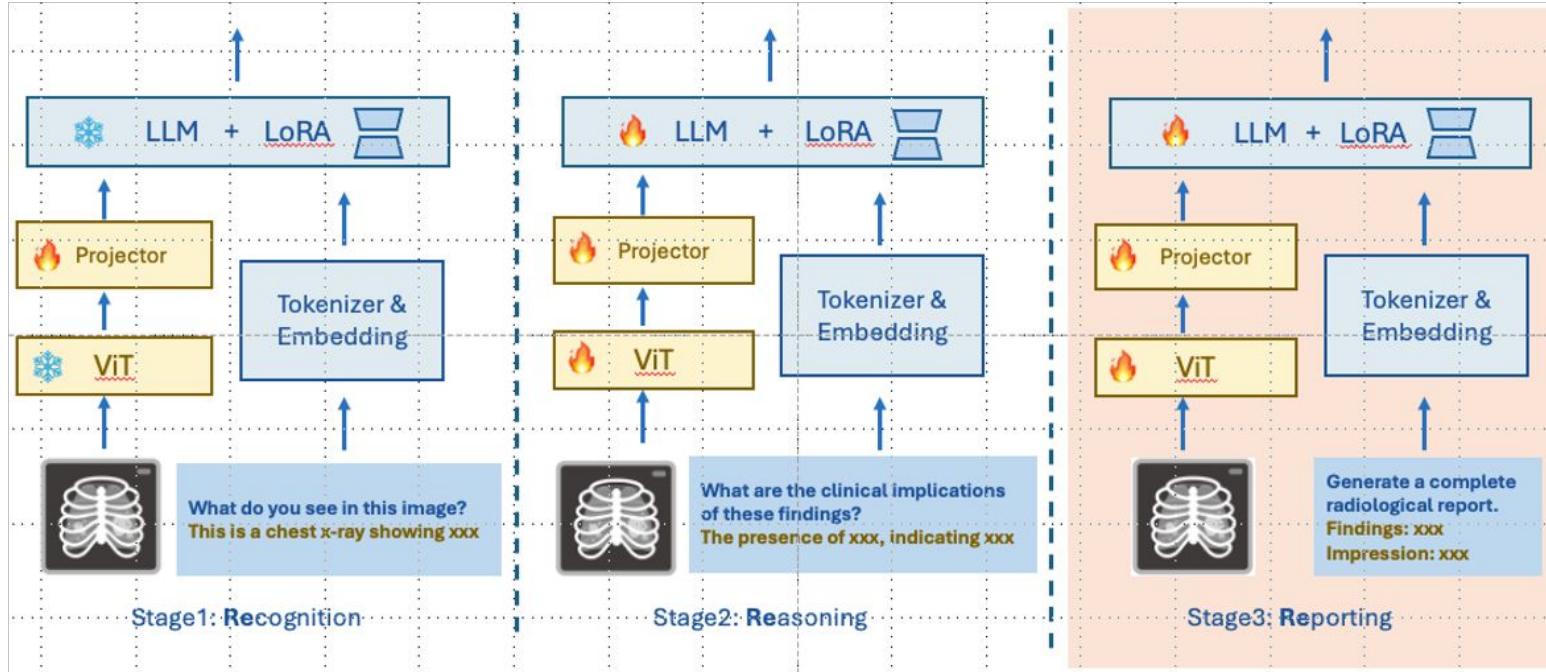


Re³Training: Progressive Learning Framework

Stage 2 - Reasoning (Diagnostic Development)

- **Dataset:** 632k samples with diagnostic reasoning patterns
- **Objective:** Multi-step clinical interpretation
- **Learning:** "What does cardiomegaly + pulmonary congestion suggest?" → "Heart failure"

SLaVA-CXR: three-stage visual instruction tuning



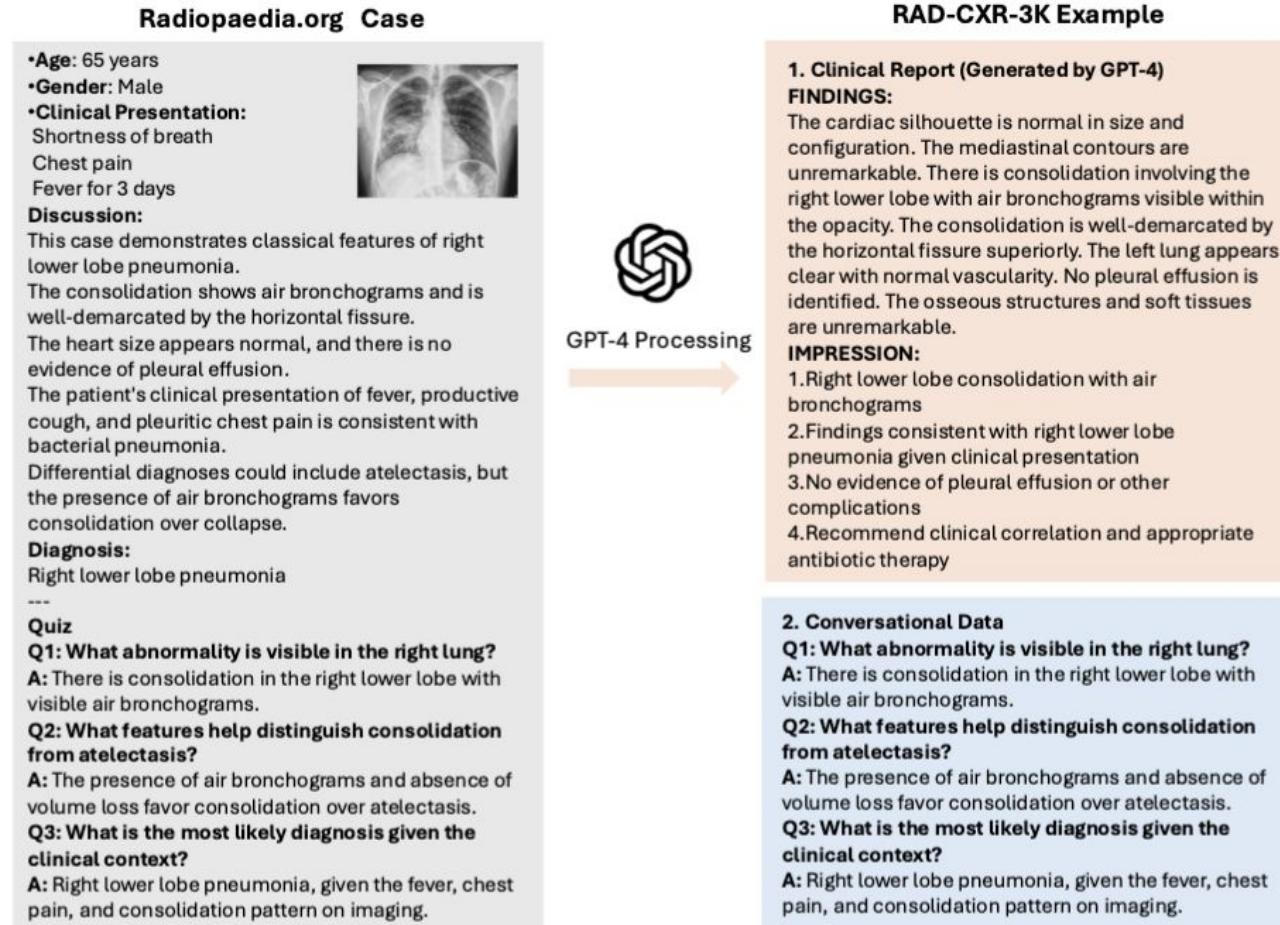
Re³Training: Progressive Learning Framework

Stage 3 - Reporting (Clinical Integration)

- **Dataset:** RAD-CXR-3K high-quality educational cases
- **Objective:** Comprehensive structured report generation
- **Learning:** Image → Full clinical report with findings + impression

II. Visual Instruction Tuning

SLaVA-CXR: Small and efficient VLM for radiology



Only train from 3000+ Radiopaedia case studies
AI-enhanced data construction
VQA and reporting supportive

Instruction following data

SLaVA-CXR: Small and efficient VLM for radiology

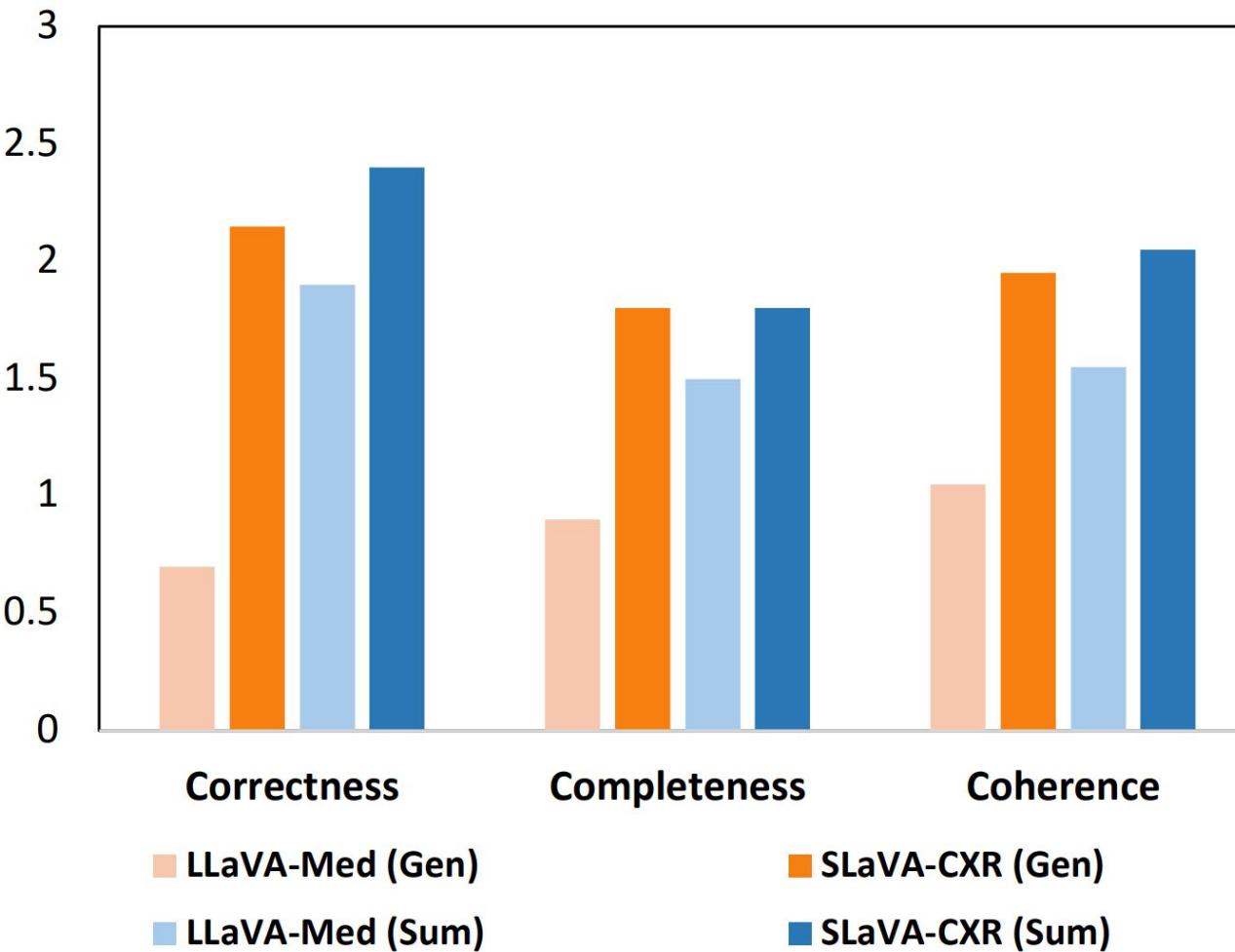
Methods	No Finding	Enlarged Cardiomediastinum	Cardiomegaly	Lung Lesion	Lung Opacity	Edema	Consolidation
LLaVAv0 (Liu et al., 2023c)	51.71	50.12	49.29	50.11	50.17	51.01	50.25
LLaVA-Med (Li et al., 2023)	51.66	50.00	53.96	54.59	50.16	52.80	52.89
LLaVAv1.5 (Liu et al., 2023b)	49.98	50.00	50.59	50.59	50.14	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)	58.87	50.53	57.30	59.15	59.34	59.68	58.15

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	55.47
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)	55.62	52.21	58.85	55.54	50.54	54.91	55.11

The model performs well across 12 abnormalities.

II. Visual Instruction Tuning

SLaVA-CXR: Small and efficient VLM for radiology



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	3.32	4.45	1.26	1.10	2.53

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

II. Visual Instruction Tuning

SLaVA-CXR: Small and efficient VLM for radiology

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.

SLaVA-CXR: Small and efficient VLM for radiology

Evaluation

Generation; Summarization; Classification

Key Results:

- 60% performance improvement vs LLaVA-Med (ROUGE-L: 13.77% vs 8.60%)
- 6.1× faster inference (2.53s vs 15.50s average processing time)
- Superior efficiency: 2.7B parameters vs 7B+ competitors

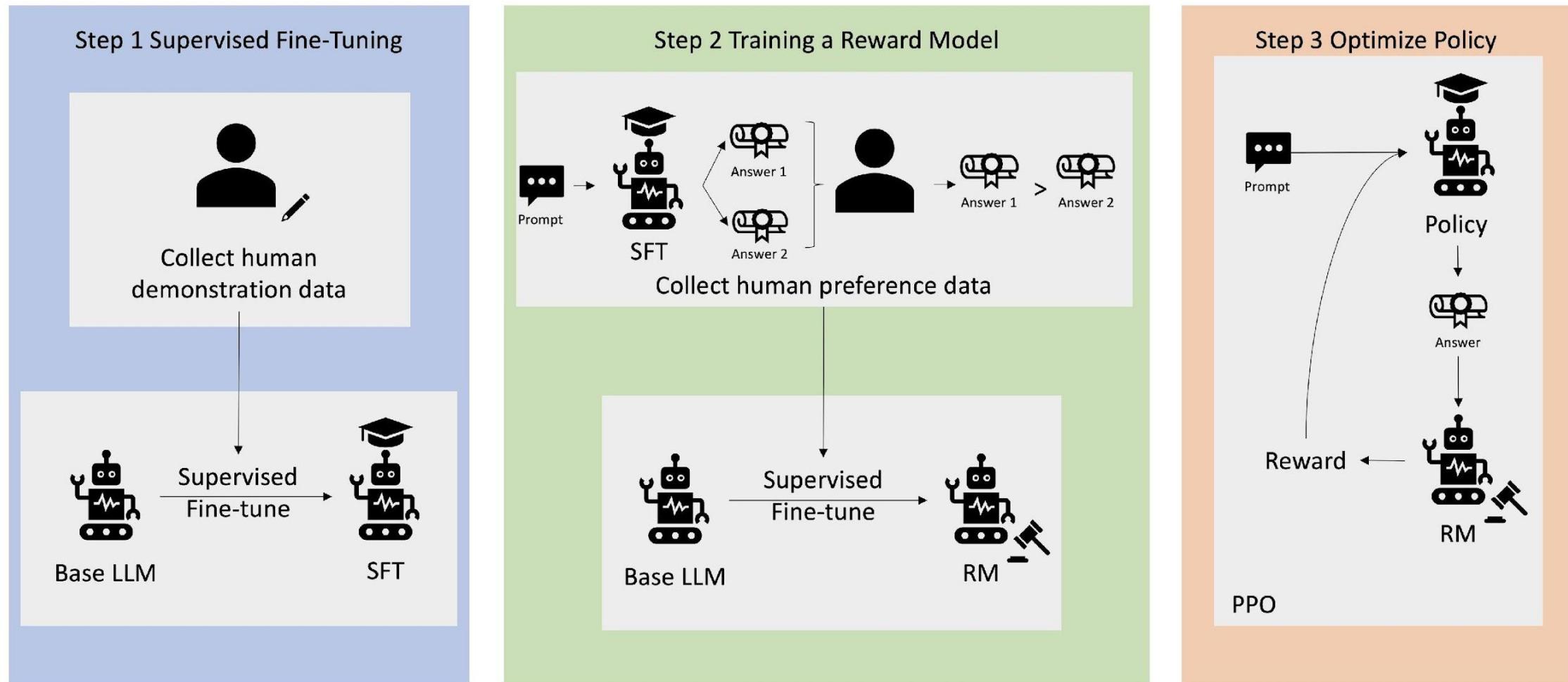
Smaller, Faster, Smarter!

Contributions:

- Proved quality over quantity: small curated datasets outperform large ones
- Lightweight architecture suitable for resource-constrained clinical environments
- Progressive training methodology for better multimodal alignment

II. Visual Instruction Tuning

Aligns the LLM to respond in the preferred way. Often done with reinforcement learning.



Med-Gemini: Gemini Models in Medicine

Modality	Dataset	No. examples	No. Images	Description
Radiology (2D)	Slake-VQA	4,919	450	Radiology images & QA pairs
	MIMIC-CXR	2,142,892	231,483	Radiology images & free-form reports
	Digital Knee X-ray	1,469	1,469	Knee X-ray images & labels
	CXR-US2	132,680	132,680	Radiology images & free-form reports
	NLST	2,199	2,199	2D CT slices & free-form reports
	CT-US1	3,207	3,207	2D CT slices & free-form reports
Radiology (3D)	CT-US1	657,719	657,719	3D CT images & free-form reports
Pathology	PathVQA	19,654	2,599	Pathology images & QA pairs
	Histopathology	1,550,976	207,603	Histopathology images, captions, & QA pairs
Dermatology	PAD-UFES-20	2,047	2,047	Skin lesion images & labels
Ophthalmology	EyePACS	14,406	14,406	Fundus images & labels
Medical VQA	PMC	2,246,656	2,246,656	PubMed Central images & caption pairs
	MedVQA	12,664	3,168	Medical images & QA pairs
Genomics	UK Biobank	259,225	259,225	Genomic data & disease outcomes

Signal QA | ECG-QA (cardiology) | 159,306 | Close-ended signal QA of electrocardiograms

More than 7 million data samples from 3.7 million medical images and cases is used for fine-tuning and further instruction-tuning of Gemini for medical applications in Med-Gemini

Med-Gemini: Gemini Models in Medicine

Evaluation:

1. Evaluation of advanced reasoning on text-based tasks - QA, generation, summarization, simplification, diagnostics
2. Evaluation of multimodal capabilities- 2D, 3D, Signal QA
3. Evaluation of long-context capabilities on video and EHR tasks
4. Evaluation of Genomics “images” (polygenic risk scores (PRS) projected into 2D)

Med-Gemini: Gemini Models in Medicine

Prompt



...



...



You are a helpful medical video assistant.

Task:

Assess whether Critical View of Safety (CVS) is fully achieved in the provided cholecystectomy video.

The Critical View of Safety (CVS) is fully achieved if the following three criteria are met:

- C1: Clear view of 2 tubular structures connected to the gallbladder.
- C2: A carefully dissected hepatocystic triangle presenting an unimpeded view of only the 2 cystic structures and the cystic plate.
- C3: The lower third of the gallbladder is dissected off the cystic plate.

Instructions:

1. Watch the video carefully, and answer which of the Critical View of Safety (CVS) criteria are met.

2. Output Format:

- CVS Assessment:
 - C1 - achieved / partially achieved / not achieved (Optional explanation)
 - C2 - achieved / partially achieved / not achieved (Optional explanation)
 - C3 - achieved / partially achieved / not achieved (Optional explanation)

Response

CVS Assessment:

C1 - partially achieved: Two tubular structures are visible but the view is not entirely clear due to some bleeding obscuring the area.

C2 - not achieved: The hepatocystic triangle is not clearly dissected, and there is not an unimpeded view of only the two cystic structures and the cystic plate.

C3 - not achieved: The lower third of the gallbladder is not dissected off the cystic plate.

Example of Med-Gemini's long-context capabilities on a surgical video.

It analyzes a video clip from the Cholec80 dataset to assess achievement of the Critical View of Safety (CVS) during a laparoscopic cholecystectomy (a keyhole operation to remove the gallbladder). The model assesses whether the three criteria defining the achievement of the CVS are met, with a detailed explanation for each criterion separately.

Reasoning Enhancement

III. Reasoning Enhancement

Original Clinical Visual Question:



Does this chest X-ray show cardiomegaly?

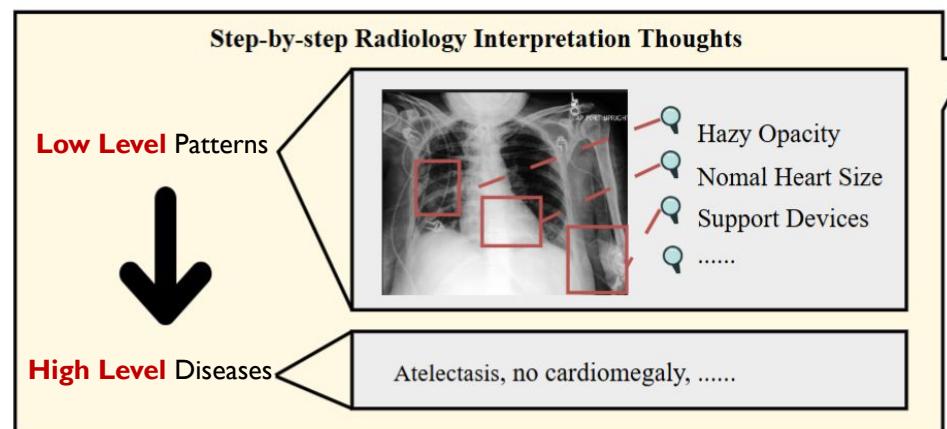
- Options:
A) Yes
B) No

Previous Model Output:

Output: A) Yes

Result-only

Reasoning-enhanced Model Output:



Output:

Reasoning Processes:

To determine if the chest X-ray shows cardiomegaly, it's crucial to analyze the size of the heart. The key indicator is the cardiothoracic ratio, which is the ratio of the maximal horizontal cardiac diameter to the maximal horizontal thoracic diameter. In a normal PA chest X-ray, this ratio is expected to be less than 0.50 ...

Therefore, the answer is A) Yes

Data: Combined Multiple-choice (MCQ) VQA datasets

- VQA-RAD [18], SLAKE [22], PathVQA [14], OmniMedVQA [15], and PMC-VQA [34].
- In total, 17,300 MCQs linked to images, with 2–6 possible choices per question.
- Focus on radiology modalities: CT, MRI, and X-ray.
 - 600 MRI image-question pairs for training.
 - 300 MRI (in-domain), 300 CT (out-of-domain), and 300 X-ray (out-of-domain) pairs for testing.
- Original data **has no reasoning supervision and the “Think” part is inferred by model itself**

Problem 1: What can be observed in this image?

- A) Lungs B) Bladder C) Brain D) Heart

MedVLM-R1 Output:

<think>

The image is a chest X-ray, which is a type of radiographic image used to visualize the internal structures of the body, particularly the lungs and bones. The presence of lung markings and the ribcage are characteristic features of a chest X-ray.

</think>

<answer>A</answer> **Groundtruth Answer: A**



Problem 1: What can be observed in this image?

- A) Lungs B) Bladder C) Brain D) Heart

MedVLM-R1 Output:

<think>

The image is a chest X-ray, which is a type of radiographic image used to visualize the internal structures of the body, particularly the lungs and bones. The presence of lung markings and the ribcage are characteristic features of a chest X-ray.

</think>

<answer>A</answer> **Groundtruth Answer: A**

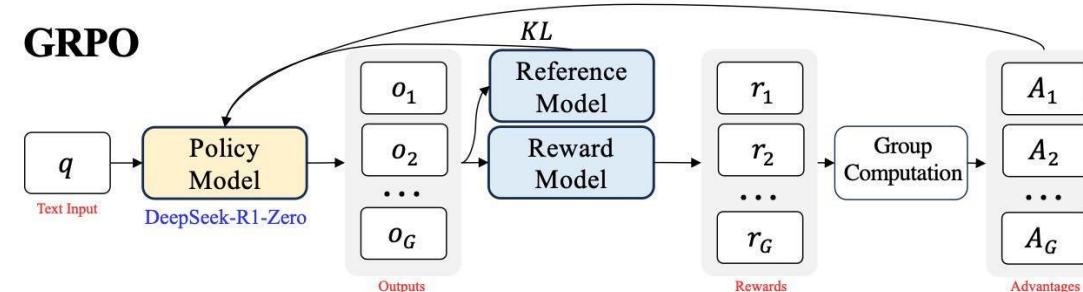


MedvIm-r1

Model Training:

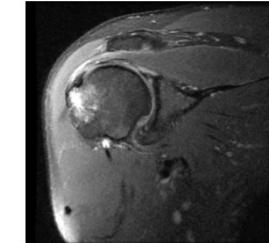
Outcome-based Reinforcement Training with GRPO
(similar as DeepSeek-R1[2])

- Two rewards (similar as DeepSeek-R1[2]):
 - Format reward: output organized in “**<think>...</think>**” tag and the final answer in the “**<answer>...</answer>**” tag.
 - Accuracy reward: output the correct choice (A / B / C / D in the final answer tag)



Question Template

What can be observed in this image?
 A) Cartilage degeneration B) Labral pathology
 C) Bone fracture D) Tendonitis



Model Input Prompt

{Question} Your task:

1. Think through the question step by step, enclose your reasoning process in **<think>...</think>** tags.
2. Then provide the correct single-letter choice (A, B, C, D,...) inside **<answer>...</answer>** tags.
3. No extra information or text outside of these tags.

Model Output Examples

<think> The image is a grayscale MRI image of an upper arm joint. The bicondylar humeral head of the humerus is visible. There is a well-defined ... **</think>**
<answer>B, there is no clear indication of ... **</answer>**

Format Reward = 1 due to the present of all tags and no content outside the tags

Accuracy Reward = 0.5 due to extra explanation appended after the answer

[1] Pan, Jiazen, et al. "MedvIm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning." arXiv preprint arXiv:2502.19634 (2025)

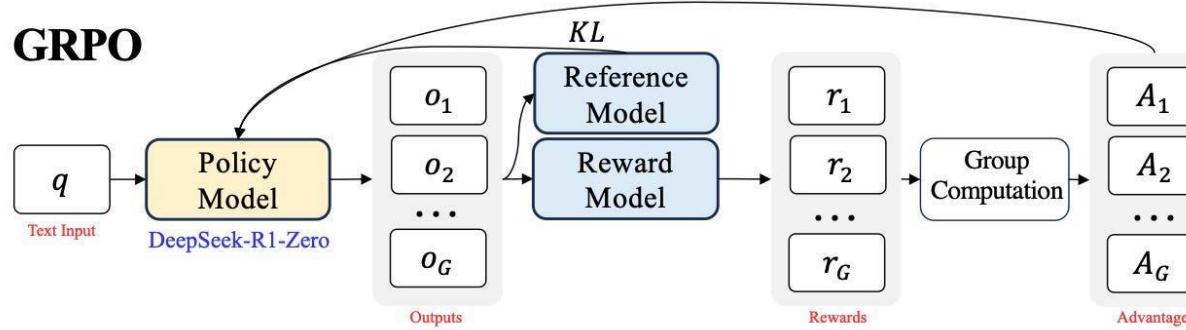
[2] Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." arXiv preprint arXiv:2501.12948 (2025).

Medvlm-r1

Model Training:

The RL algorithm simply optimizes the MLLM network to maximize the reward functions.

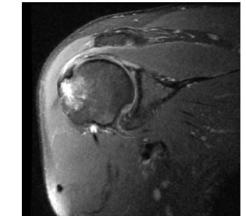
- Start from Qwen2-VL-2B
- Expect the model to learn the reasoning ability through self-exploration **without extra instruction-wise supervisions.**



Question Template

What can be observed in this image?

- A) Cartilage degeneration B) Labral pathology
C) Bone fracture D) Tendonitis



Model Input Prompt

{Question} Your task:

1. Think through the question step by step, enclose your reasoning process in <think>...</think> tags.
2. Then provide the correct single-letter choice (A, B, C, D,...) inside <answer>...</answer> tags.
3. No extra information or text outside of these tags.

Model Output Examples

<think> The image is a grayscale MRI image of an upper arm joint. The bicondylar humeral head of the humerus is visible. There is a well-defined ... </think>
<answer>B, there is no clear indication of ... </answer>

Format Reward = 1 due to the present of all tags and no content outside the tags

Accuracy Reward = 0.5 due to extra explanation appended after the answer

[1] Pan, Jiazen, et al. "Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning." *arXiv preprint arXiv:2502.19634* (2025).

[2] Guo, Daya, et al. "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning." *arXiv preprint arXiv:2501.12948* (2025).

MedvIm-r1

Main Results:

- RL surpasses SFT significantly
- Strong generalization abilities

Table 1: Results of VQA-VLMs on MRI (in-domain), and CT and X-Ray (out-of-domain) modalities. "-2B" indicates the model has 2 billion parameters, etc.

Method	Num. of Seen Medical Sample	In-Domain / Out-of-Domain			Average
		(MRI→MRI)	(MRI→CT)	(MRI→X-ray)	
Random Guess	/	25.00	30.25	26.00	27.08
<i>Zero-shot VLM</i>					
Qwen2-VL-2B	/	61.67	50.67	53.00	55.11
Qwen2-VL-7B	/	72.33	68.67	66.63	69.21
Qwen2-VL-72B	/	68.67	60.67	72.33	67.22
<i>Zero-shot Medical VLM</i>					
Huatuo-GPT-vision-7B	1,294,062	71.00	63.00	73.66	69.22
<i>MRI fine-tuned VLM</i>					
Qwen2-VL-2B (SFT)	600	94.00	54.33	34.00	59.44
Ours-2B (GRPO)	600	95.33	70.33	69.00	78.22

III. Reasoning Enhancement

MedVlm-r1

Main Results:

- RL Surpasses SFT significantly
- Strong generalization abilities
- Model learns to fill the <think> part.

Problem 1: What can be observed in this image?

- A) Lungs B) Bladder C) Brain D) Heart

MedVLM-R1 Output:

<think>
The image is a chest X-ray, which is a type of radiographic image used to visualize the internal structures of the body, particularly the lungs and bones. The presence of lung markings and the ribcage are characteristic features of a chest X-ray.
</think>



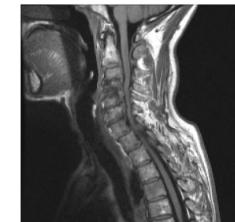
<answer>A</answer> **Groundtruth Answer: A**

Problem 3: What is the finding in this image?

- A) Stomach ulcer, B) Nerve compression,
C) Cord pathology, D) Liver cirrhosis

MedVLM-R1 Output:

<think> To determine the correct answer, we need to identify the finding in the image provided. The MRI scan in question is likely a view of a spine, not the stomach, or the liver. Stomach ulcers, liver cirrhosis, and spinal cord pathology can be diagnosed with MRI, but the image in question shows a spinal view, not a gastrointestinal or abdominal view. Therefore, the only reasonable options left are nerve compression and cord pathology.
</think>



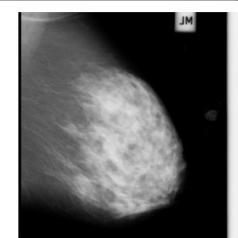
<answer>C</answer> **Groundtruth Answer: C**

Problem 2: What content appears in this image?

- A) Cardiac tissue, B) Breast tissue, C) Liver tissue, D) Skin tissue

MedVLM-R1 Output:

<think> The image appears to be a mammogram, which is a type of X-ray used to detect abnormalities in the breast tissue. The image shows the breast tissue with various densities and patterns, which are typical of mammograms.
</think>



<answer>B</answer> **Groundtruth Answer: B**

Problem 4: What is the specific abnormality present in this image?

- A) Gastritis, B) Arthritis, C) Appendicitis, D) Urolithiasis

MedVLM-R1 Output:

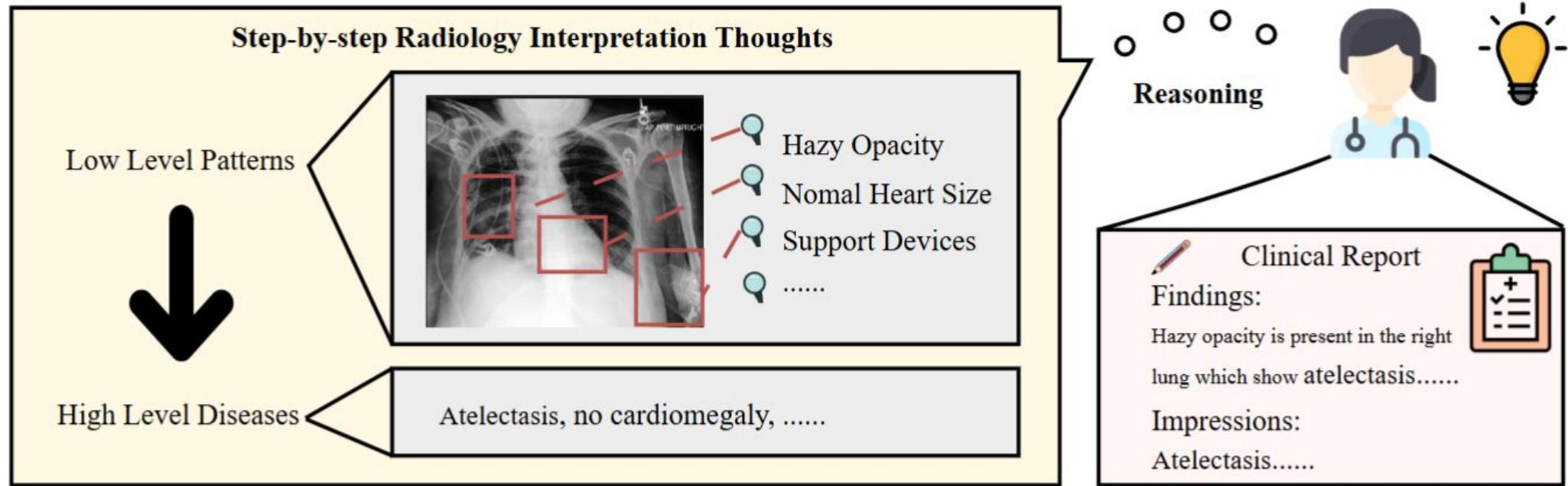
<think> The image presented is a transverse CT scan of the abdomen and pelvis. The presence of calculi (urines filled with stones or grit) in the pelvic organs is a consistent finding in urolithiasis.
</think>



<answer>D</answer> **Groundtruth Answer: D**

III. Reasoning Enhancement

Is that true that “radiology reasoning is hard to obtain?”



Key insight:

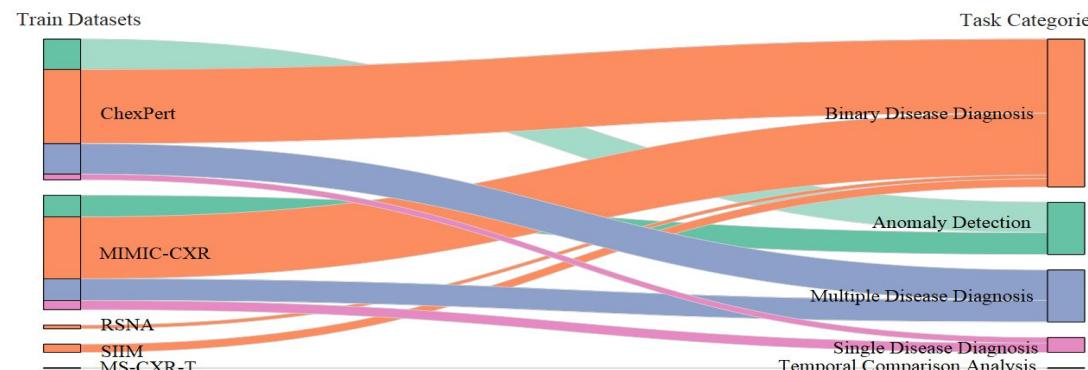
- Radiology interpretation tasks naturally follow medical reasoning thoughts from finding **low-level patterns to synthesizing high-level diseases.**
- The reasoning processes are **maintained in daily clinical reports.**

ChestX-Reasoner:

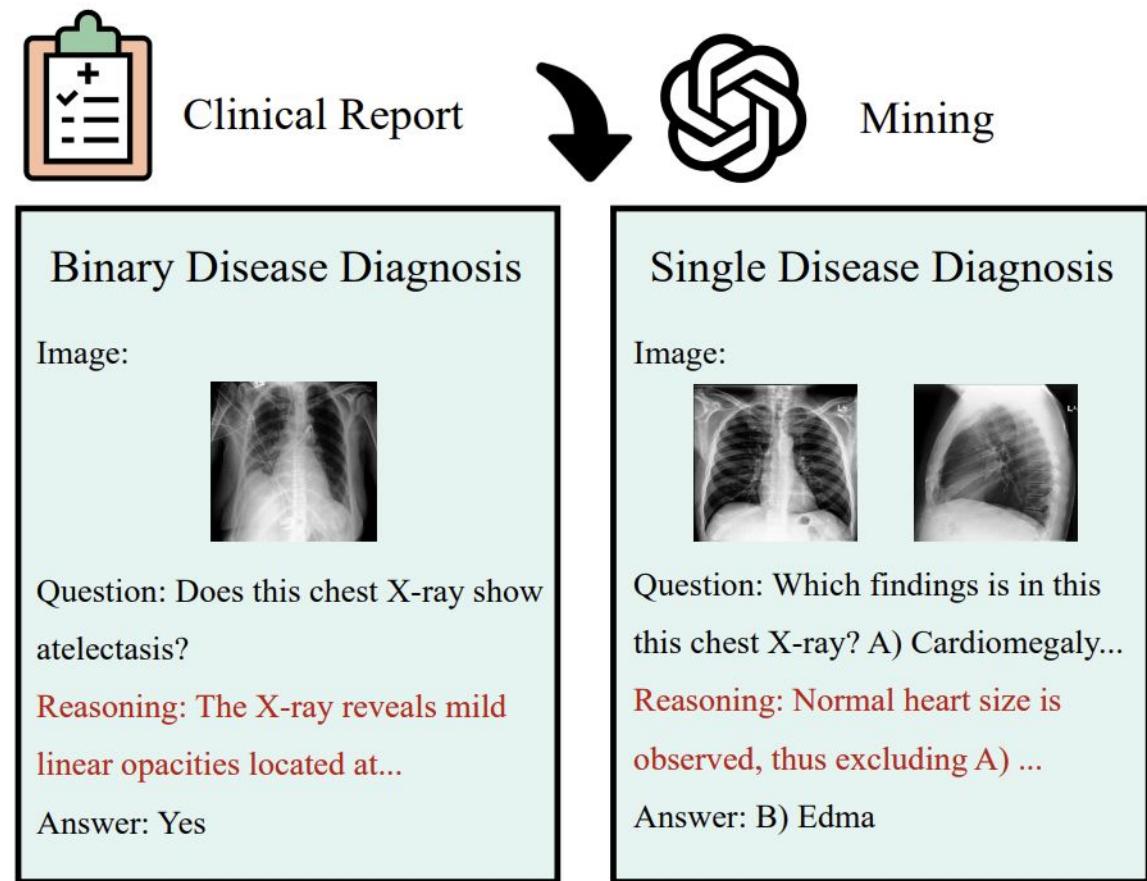
Data construction:

- Adopt radiology reports as the basic references;
- Adopt diverse case as task seeds;
- Prompt GPT-4 to generate the case with reasoning based on reports.

Data and task distribution:



Mined reasoning case in CXR (Beyond MCQs)



ChestX-Reasoner:

Metrics on Reasoning (RadRScore)

- Effectiveness:** Proportion of efficient reasoning steps.
- Factuality:** Proportion of valid reasoning steps.
- Completeness:** Recall of referenced reasoning steps.

Effectiveness

Model Reasoning Steps

- Find opacification ...
- Find enlarged heart ...
- Find mild rib fracture ...
- ...



GT Reasoning Steps

- Find opacification in left lower lung, showing ...
- Find no evidence of cardiomegaly ...

Factuality

Entities in Model Reasoning

- left lung opacification
- enlarged heart size
- mild rib fracture
- ...



Radiology Report

... show **opacification** in left lower lung which represents new **pneumonia** ...No evidence of **cardiomegaly**



Completeness

Entities in GT Reasoning

- pneumonia
- left lung opacification
- no cardiomegaly
- ...



Entities in Model Reasoning

- left lung opacification
- enlarged heart size
- mild rib fracture
- ...

III. Reasoning Enhancement

ChestX-Reasoner:

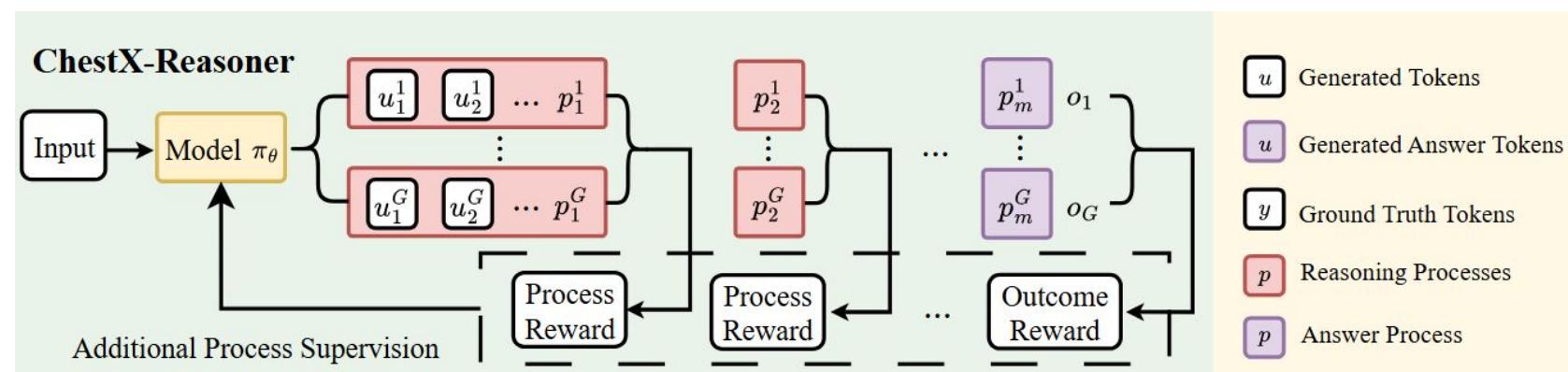
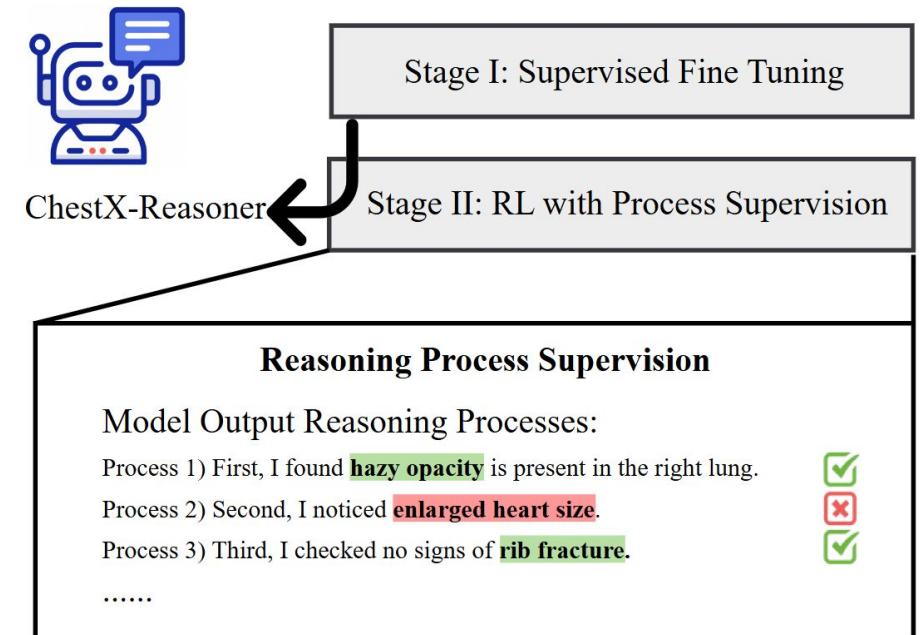
Model Training:

Involves two stages of training:

- Supervised fine-tuning
- Reinforcement learning

Designs three reward functions:

- Outcome-based rewards: format and accuracy
- Process factuality reward: correctness of generated reasoning steps

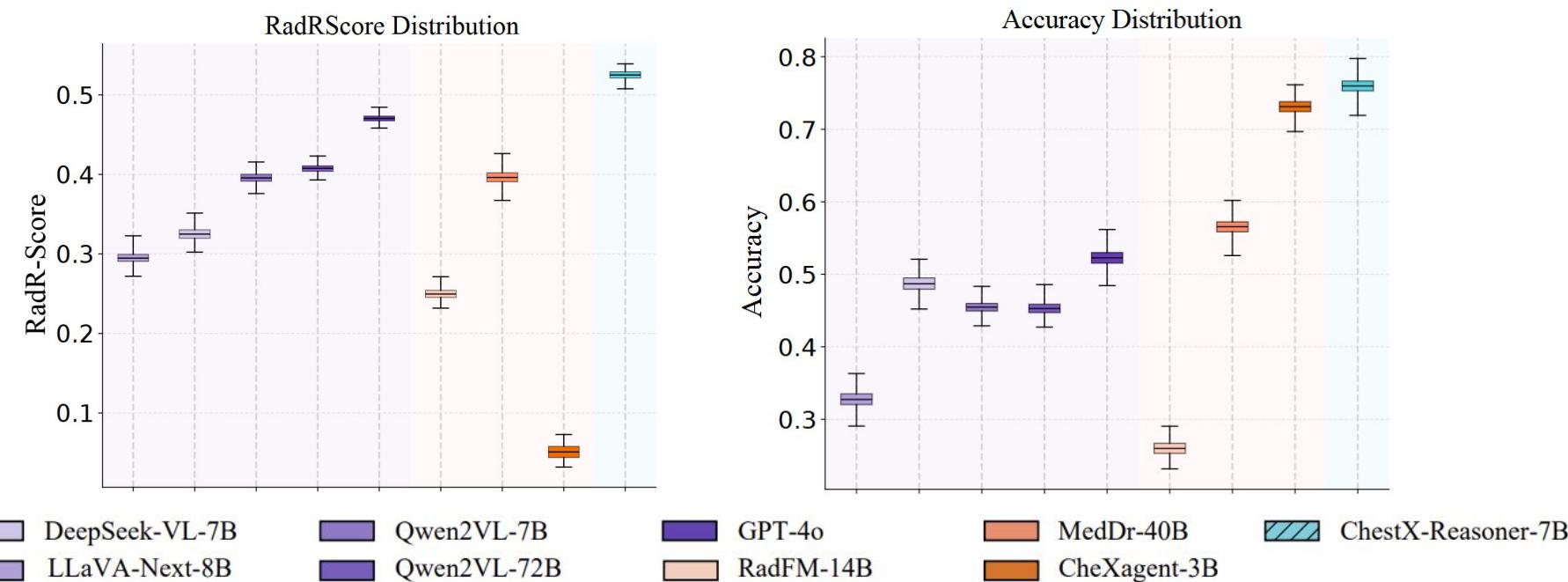


III. Reasoning Enhancement

ChestX-Reasoner:

Results: Naturally, process supervision enhances both final accuracy and, **more notably, reasoning reliability.**

- **16% and 3.3% improvement** on reasoning and outcome accuracy compared to medical MLLMs (MedDR etc).
- **5.9% and 24% improvement** on reasoning ability and outcome accuracy compared to commercial MLLMs (GPT4o etc).
- **18% and 27% improvement** on reasoning ability and outcome accuracy compared to general MLLMs(Qwen2VL-7B).



Human – guided reasoning

Traditional medical AI systems lack understanding of how radiologists actually think and diagnose

Radiologist eye fixations are NOT random—they reflect diagnostic strategies, clinical experience, and pathology detection logic

Two Key Pillars:

1.Human-Centered AI

1. Capture expert cognitive processes through eye tracking
2. Use visual attention as clinical prior knowledge
3. Bridge the gap between AI predictions and human reasoning

2.Reasoning Enhancement for MLLMs

1. Move beyond simple pattern recognition
2. Simulate expert diagnostic reasoning workflows
3. Learn *how* radiologists think, not just *what* they see

Human – guided reasoning

The REFLACX Dataset

- 109 radiologists + 3,032 chest X-rays
- Eye tracking data recorded during diagnostic reading
- Includes:

- Temporal eye fixation coordinates
- Diagnostic text reports
- Disease labels

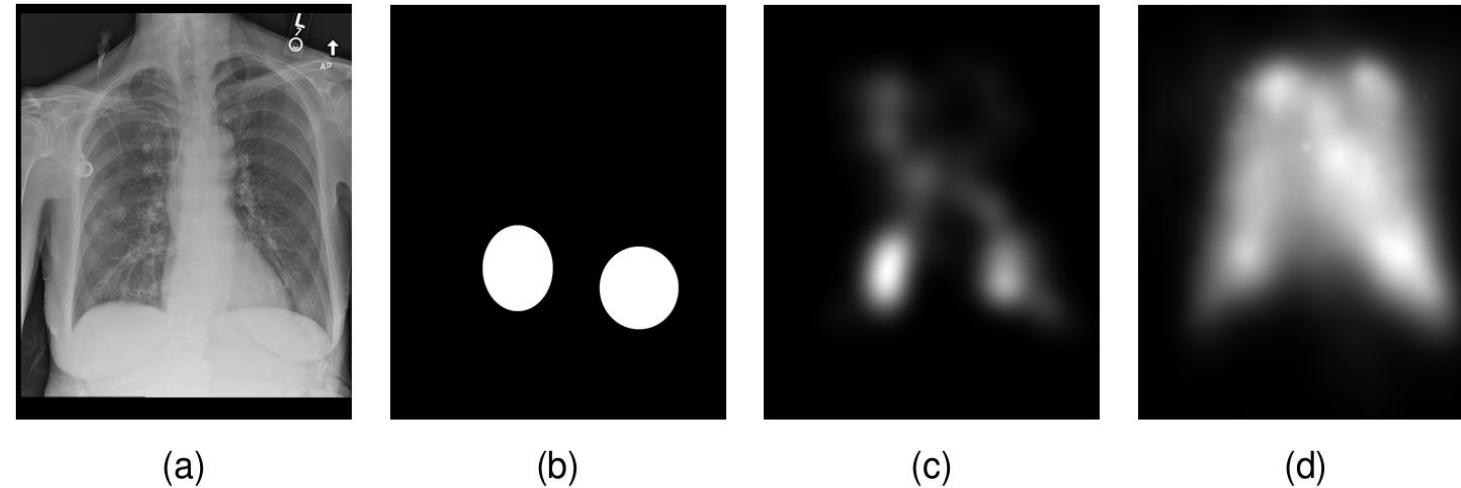
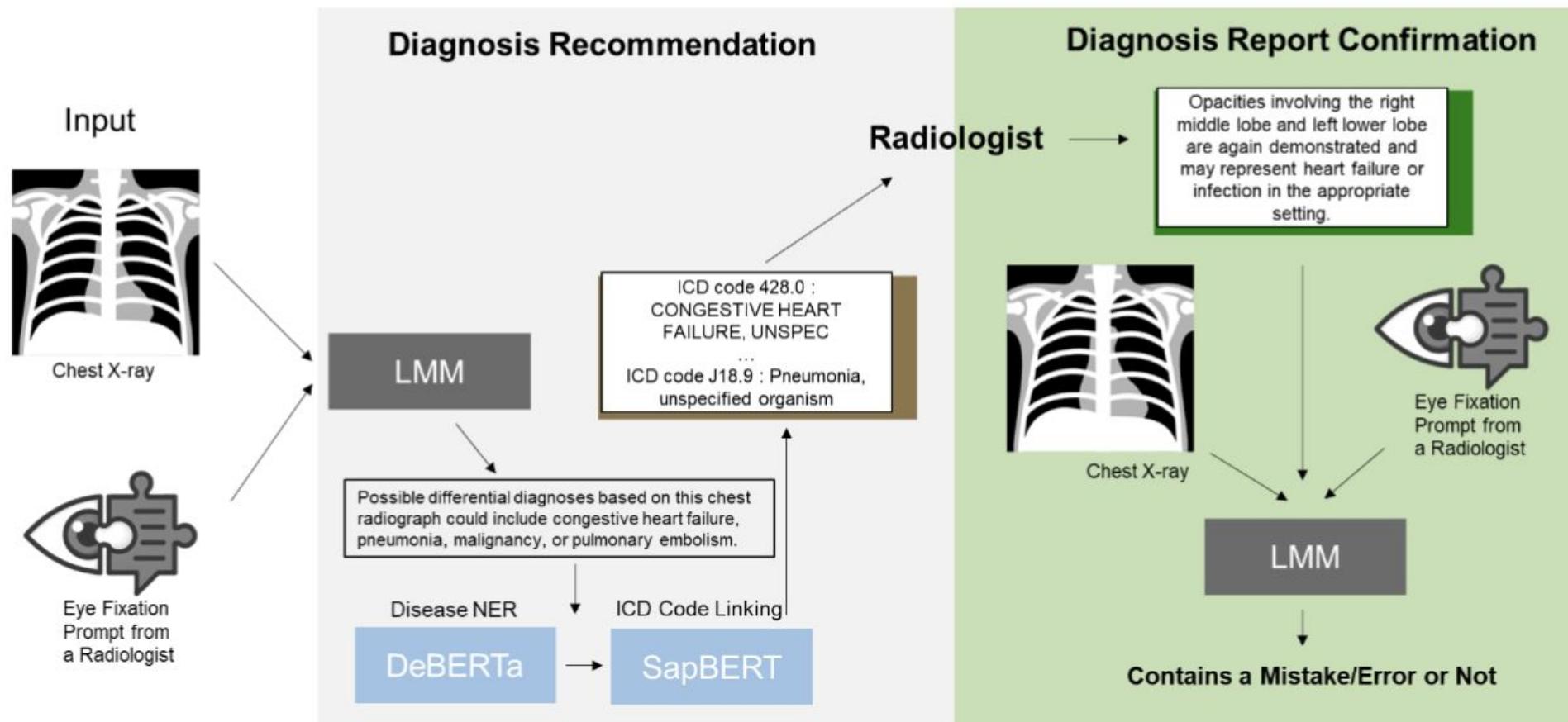


Figure 3. Example of the localization information provided by the eye-tracking data and how it was validated. (a) CXR read by the radiologist. (b) Union of the abnormality ellipses selected by radiologists used to compare against heatmaps. (c) Heatmap generated by the fixations made by the radiologist while dictating the report. (d) Average heatmap for all radiologists and CXRs read in phases 1 and 2, normalized to the location of lung and heart of the CXR.

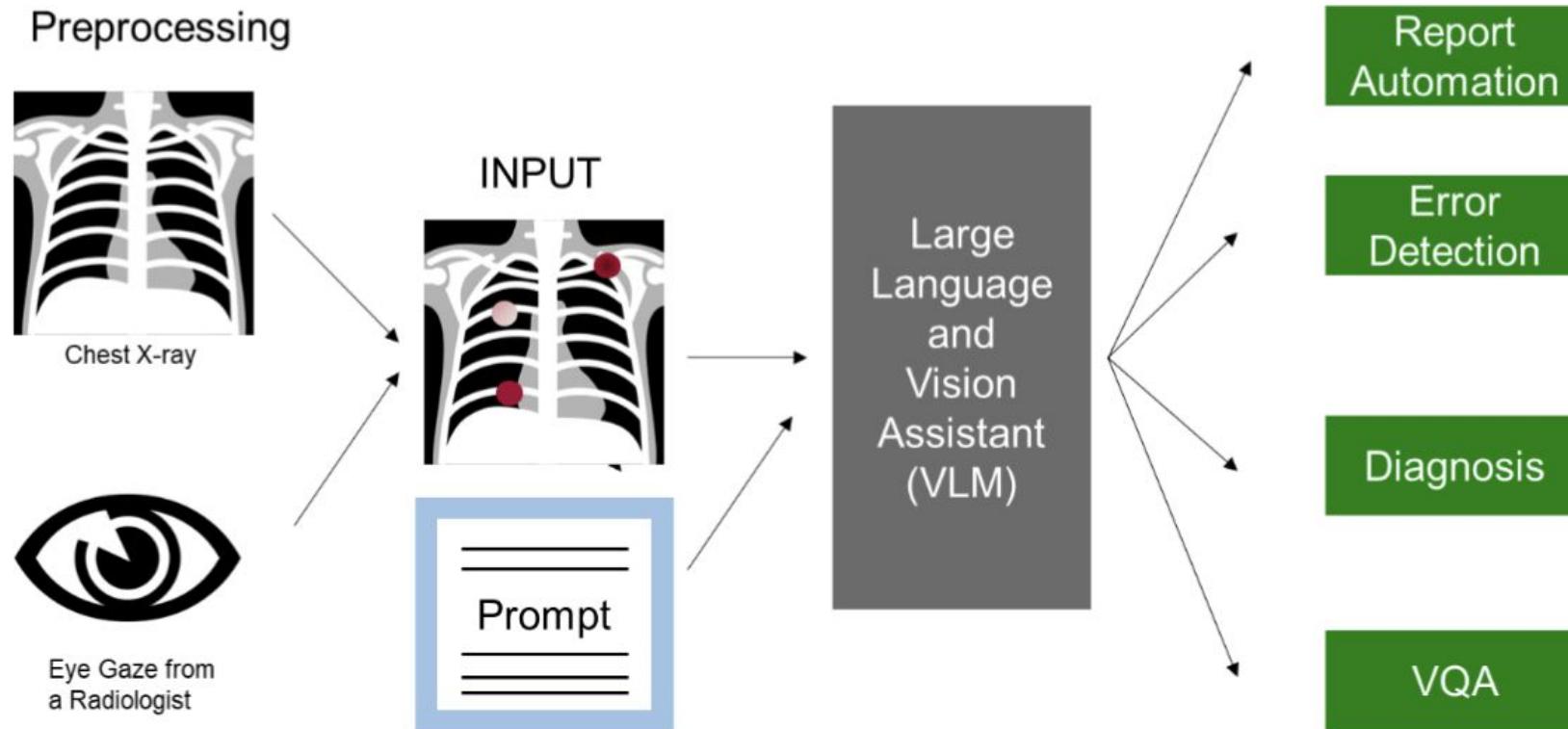
Eye-Gaze – vI



Integrate the eye fixation information into textual prompts

III. Reasoning Enhancement

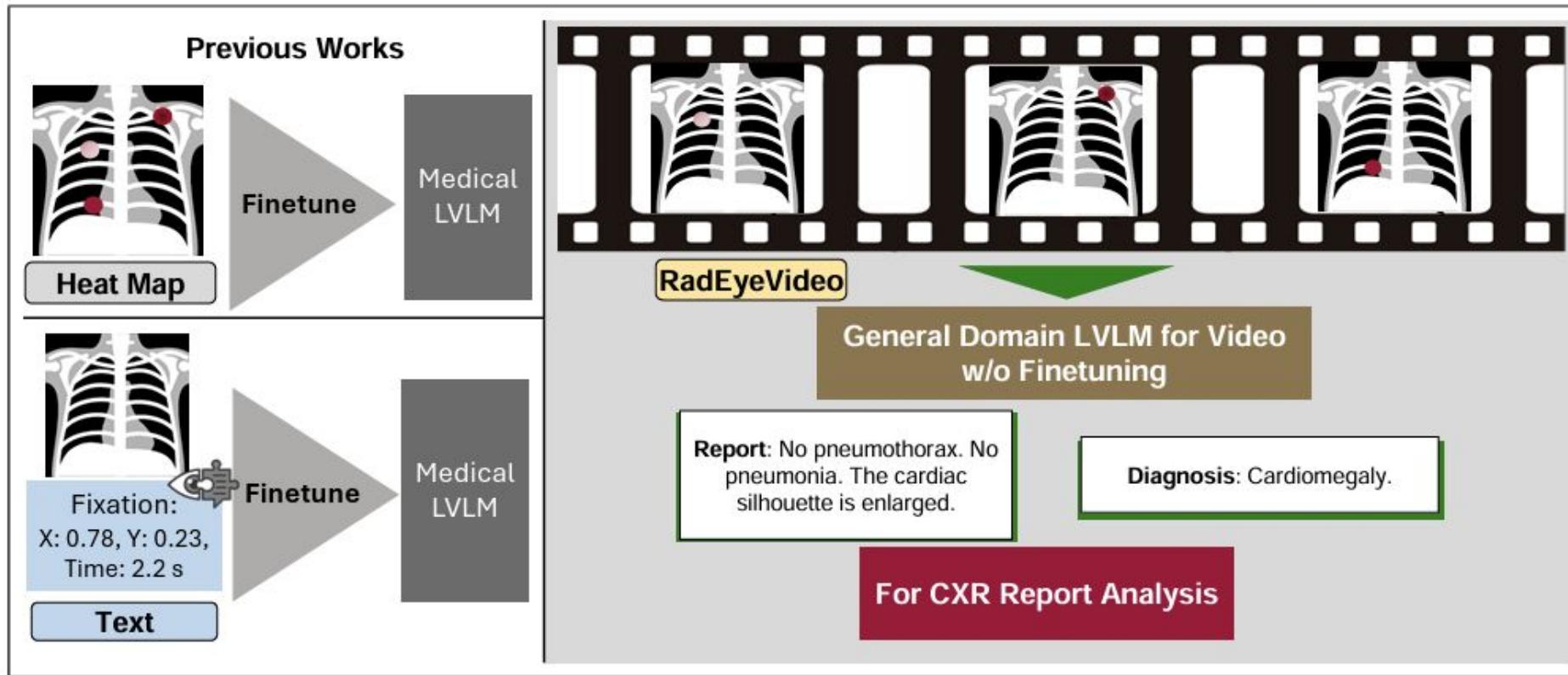
Eye-Gaze – v2:



Heatmaps with VLMs for clinical applications

Highlight the precise focal points and duration of a radiologist's attention when analysing a CXR

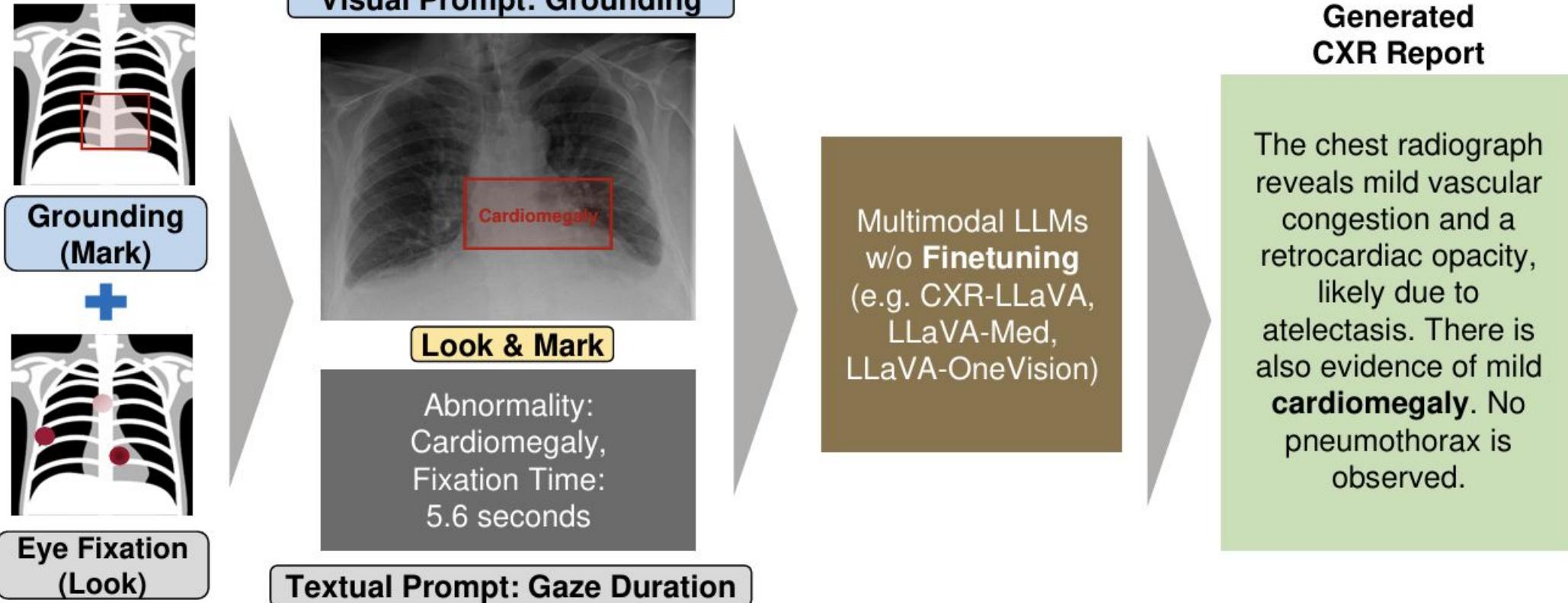
Eye-Gaze – v3:



Dynamic Eye Gaze with Video for CXR Analysis

- leverage both spatial and temporal aspects of gaze patterns

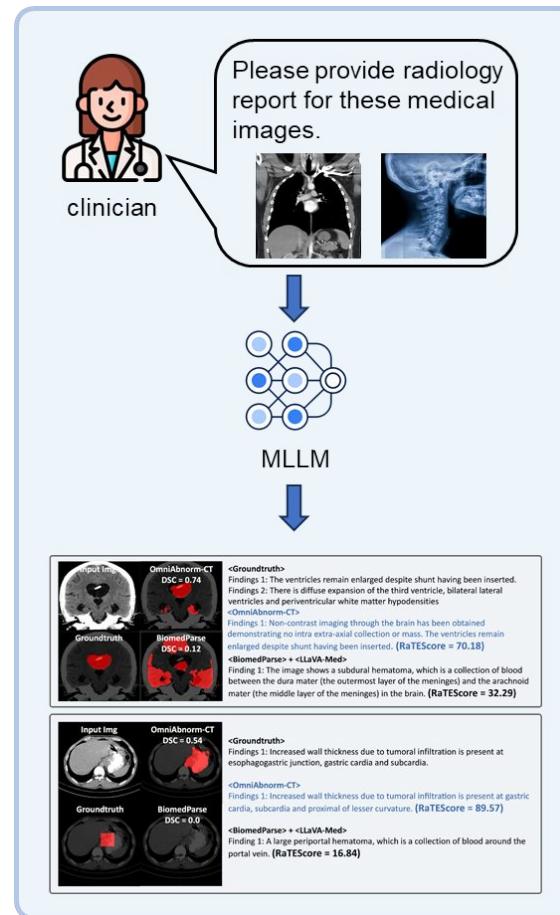
Eye-Gaze – v4:



Combine eye gaze + bounding boxes for multi-level visual guidance

Applications

Report Generation



Prompt

Given the image, what are the radiology findings?

Response

FINDINGS: mild cortical atrophy and periventricular white matter changes are noted compatible with microvasculopathy. there is proportionate ventriculomegaly. no bleed, mass effect, midline shift or abnormal intracranial fluid is seen. no evidence of acute intracranial hemorrhage, territorial infarct, abnormal fluid or mass effect.no abnormal extra-axial fluid collection is identified. no definitive evidence of intracranial hemorrhage or large territorial infarct within the limits of ct imaging. **old left basilar ganglia infarct**. no intracranial mass, swelling, hemorrhage, infarction or extra-axial collection is seen. visualized mastoid air cells: well pneumatized. visualized paranasal sinuses: well pneumatized. atherosclerotic calcification of vasculature at the skull base.

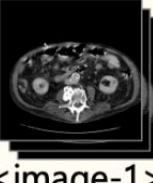
Radiologist Written Report

FINDINGS: the fourth ventricle is midline. the remaining ventricular system is within normal limits. there is mild chronic white matter microangiopathic disease and moderate generalized cortical atrophy. no bleed, mass effect, midline shift or abnormal intracranial fluid is seen. bone windows demonstrate an intact skull. the remaining visualized paranasal sinuses and mastoids are clear. chronic white matter microangiopathic disease, mild generalized cortical atrophy, not significantly changed since 2015. no evidence of acute intracranial hemorrhage, territorial infarct, abnormal fluid or mass effect.

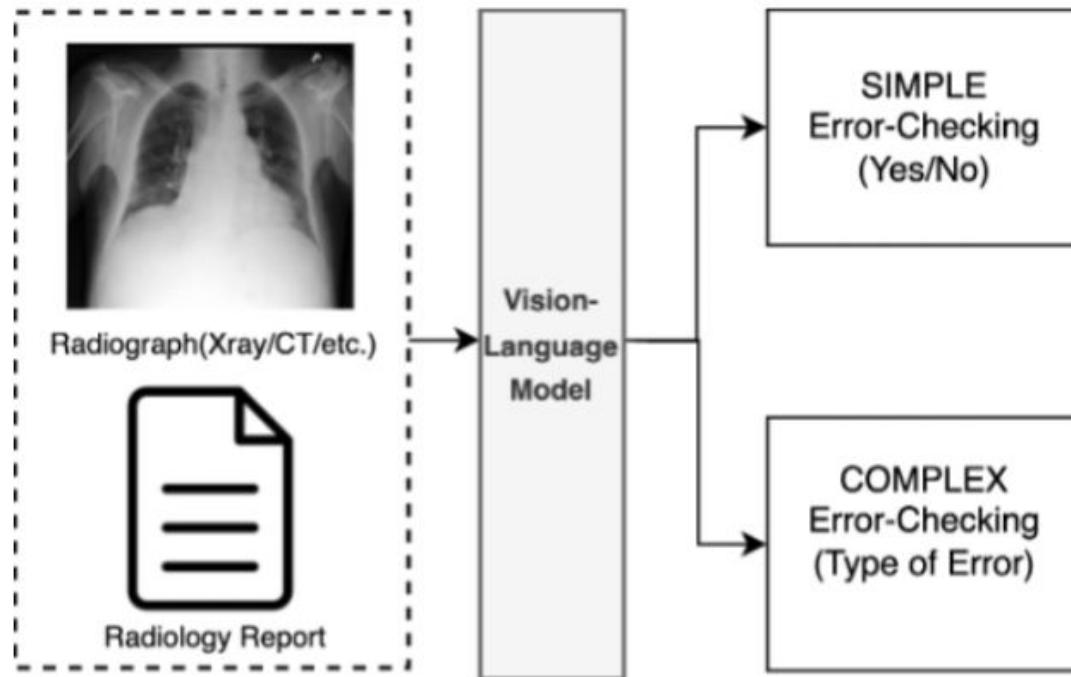
[1] Saab, K. et al. Saab, Khaled, et al. "Capabilities of gemini models in medicine." *arXiv preprint arXiv:2404.18416* (2024).

[2] Yang, L. et al. Yang, Lin, et al. "Advancing multimodal medical capabilities of Gemini." *arXiv preprint arXiv:2405.03162* (2024).. 2024.

Diverse VQA and diagnosis

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

Error Checking



Error Checking

Systematic Error Taxonomy & Dataset Construction

Three Clinically-Relevant Error Types

1. False Positive (FP): Template/voice recognition artifacts

1. Example: Report mentions "pleural effusion" not visible in image
2. Clinical cause: Auto-populated findings from previous studies

2. Missed Findings (MF): Satisfaction of search bias

1. Example: Obvious fracture documented, subtle pneumothorax missed
2. Clinical cause: Cognitive bias after finding major abnormality

3. Terminology Confusion (TC): Dictation/transcription errors

1. Example: "Right lower lobe" → "Left lower lobe"
2. Clinical cause: Voice recognition errors, fatigue-induced mistakes

Expert-Guided Dataset Creation

- Entity curation: 20 common chest radiography findings (expert-validated)
- Synthetic generation: Systematic error introduction with clinical realism
- Distribution: 75% correct / 25% error (reflects real-world rates)
- Scale: 17,000 training + 2,000 evaluation samples
- Cross-validation: MIMIC-CXR + IU X-ray for generalization testing



FINDINGS:
There is consolidation in the right lower lobe consistent with pneumonia. The cardiac silhouette is enlarged suggesting mild cardiomegaly. No pleural effusion or pneumothorax is identified. Bilateral costophrenic angles are sharp.

IMPRESSION:
Right lower lobe pneumonia. Mild cardiomegaly.

Extracted Entities: consolidation, pneumonia, cardiomegaly, right lower lobe

Generated Error Examples

False Positive (FP)

FINDINGS:
There is consolidation in the right lower lobe consistent with pneumonia. The cardiac silhouette is enlarged suggesting mild cardiomegaly. No pleural effusion or pneumothorax is identified. The left lung appears clear.

IMPRESSION:
Right lower lobe pneumonia. Mild cardiomegaly. Small pleural effusion is noted bilaterally.

Missed Findings (MF)

FINDINGS:
There is consolidation in the right lower lobe consistent with pneumonia. The cardiac silhouette is enlarged suggesting mild cardiomegaly. No pleural effusion or pneumothorax is identified. The left lung appears clear.

IMPRESSION:
Right lower lobe pneumonia. Mild cardiomegaly.

Terminology Confusion (TC)

FINDINGS:
There is consolidation in the left lower lobe consistent with pneumonia. The cardiac silhouette is enlarged suggesting mild cardiomegaly. No pleural effusion or pneumothorax is identified. The left lung appears clear.

IMPRESSION:
Right lower lobe pneumonia. Mild cardiomegaly.

Error Checking

Experimental Analysis

Zero-shot Results:

- Task-specific models dominate:** LLaVA-CXR-ES achieves 83.18% F1 (SIMPLE)
- Scale paradox:** LLaVA-1.5-13B fails completely (0.00% F1) vs smaller models
- Complexity barrier:** COMPLEX task proves significantly harder (max 15.66% F1)

Few-shot Learning Discovery:

- Counter-intuitive patterns:** Specialized models degrade with examples
- General models improve:** LLaVA-1.5-13B recovers (0% → 33.33% F1)
- Optimization interference:** Fine-tuning conflicts with in-context learning

Shot Number Optimization Analysis

- Non-monotonic relationship:** Performance peaks at different shot numbers
- Stability trade-off:** Higher shots → higher variance (up to ±12.82 F1)
- Task dependency:** SIMPLE (2-3 shots optimal) vs COMPLEX (variable)

Model	MIMIC-CXR		IU X-ray	
	SIMPLE	COMPLEX	SIMPLE	COMPLEX
BiomedCLIP	33.52	12.04	38.48	12.84
LLaVA-1.5-13B	0.00	0.00	0.00	0.00
LLaVA-1.0	12.91	3.36	8.61	0.00
LLaVA-Med	33.52	0.40	29.64	4.91
LLaVA-1.5-7B	35.79	0.00	40.60	0.00
LLaVA-CXR-EC	39.16	11.27	32.36	15.66
LLaVA-CXR-ES	83.18	0.00	66.00	0.00

Table 6.5: Zero-shot evaluation performance measured by F1-scores across tasks. Optimal results are highlighted in bold. LLaVA-CXR-ES and LLaVA-CXR-EC denote task-specific adaptations of LLaVA-1.5-7B for SIMPLE and COMPLEX error detection scenarios, respectively.

Model	MIMIC-CXR		IU X-ray	
	SIMPLE (F1)	COMPLEX (F1)	SIMPLE (F1)	COMPLEX (F1)
LLaVA-1.5-13B	9.32(+9.32)	13.19(+13.19)	33.33(+33.33)	14.04(+14.04)
LLaVA-1.0	8.53(-4.38)	3.36(-2.15)	38.16(+29.55)	0.00(+0.00)
LLaVA-Med	3.04(-30.48)	0.20(-0.20)	19.91(-9.73)	0.74(-4.91)
LLaVA-1.5-7B	40.10(+4.31)	0.00(+0.00)	47.39(+6.79)	4.53(+4.53)
LLaVA-CXR-EC	0.00(-39.16)	7.47(-3.80)	7.04(-25.32)	14.05(-1.61)
LLaVA-CXR-ES	83.29(+0.11)	0.00(+0.00)	40.89(-25.11)	0.00(+0.00)

Table 6.6: In-context learning evaluation results. 2-shot and 4-shot examples were provided for SIMPLE and COMPLEX tasks respectively. Performance change from zero-shot evaluation is shown in parentheses. LLaVA-CXR-ES and LLaVA-CXR-EC are task-specific fine-tuned version of LLaVA-1.5-7B on SIMPLE and COMPLEX task, respectively.

Error Checking

Clinical Validation

2 board-certified radiologists, 60 cases (48 with errors, 12 correct)

Key Findings

- **75-78% accuracy** even for trained radiologists
- **35% of cases challenging** - at least one expert made errors
- **LLaVA-CXR-ES:** 76.7% accuracy (close to human performance on training domain)

AI as Quality Assurance Tool: Rather than replacing radiologists, AI serves as an additional safety net for challenging cases where human experts are prone to errors.

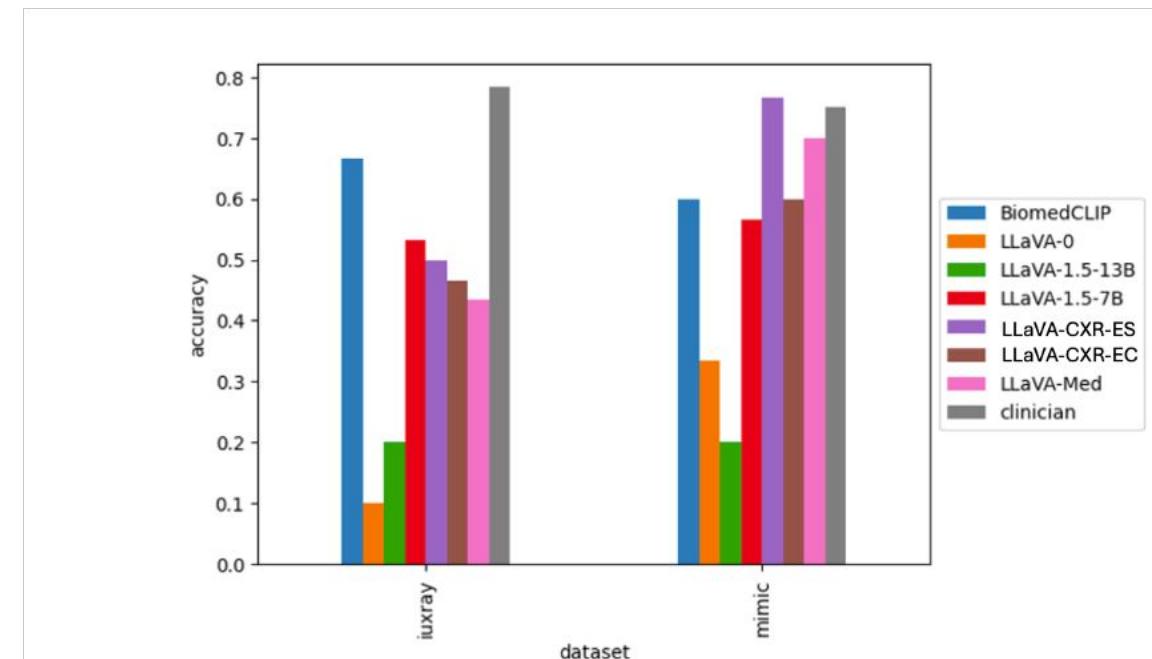


Figure 6.3: Human evaluation accuracy comparison for the SIMPLE task. Results show performance of clinical experts versus different model configurations.

Error Checking

Human-AI Complementarity

2 board-certified radiologists, 60 cases (48 with errors, 12 correct)

Key Findings from Pie Chart

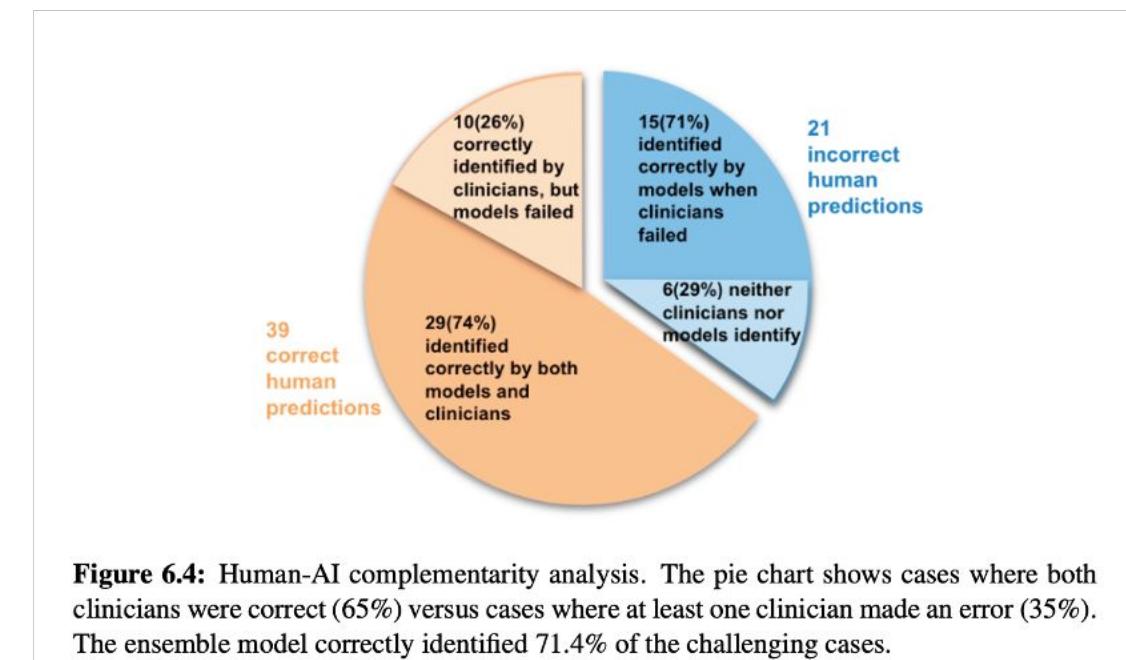
- **65% of cases:** Both clinicians correct (in orange)
- **35% of cases:** At least one expert made errors (in blue)

AI Assistance Value

- **71.4% accuracy** on difficult cases where human experts struggled
- AI catches errors humans miss

Practical Impact

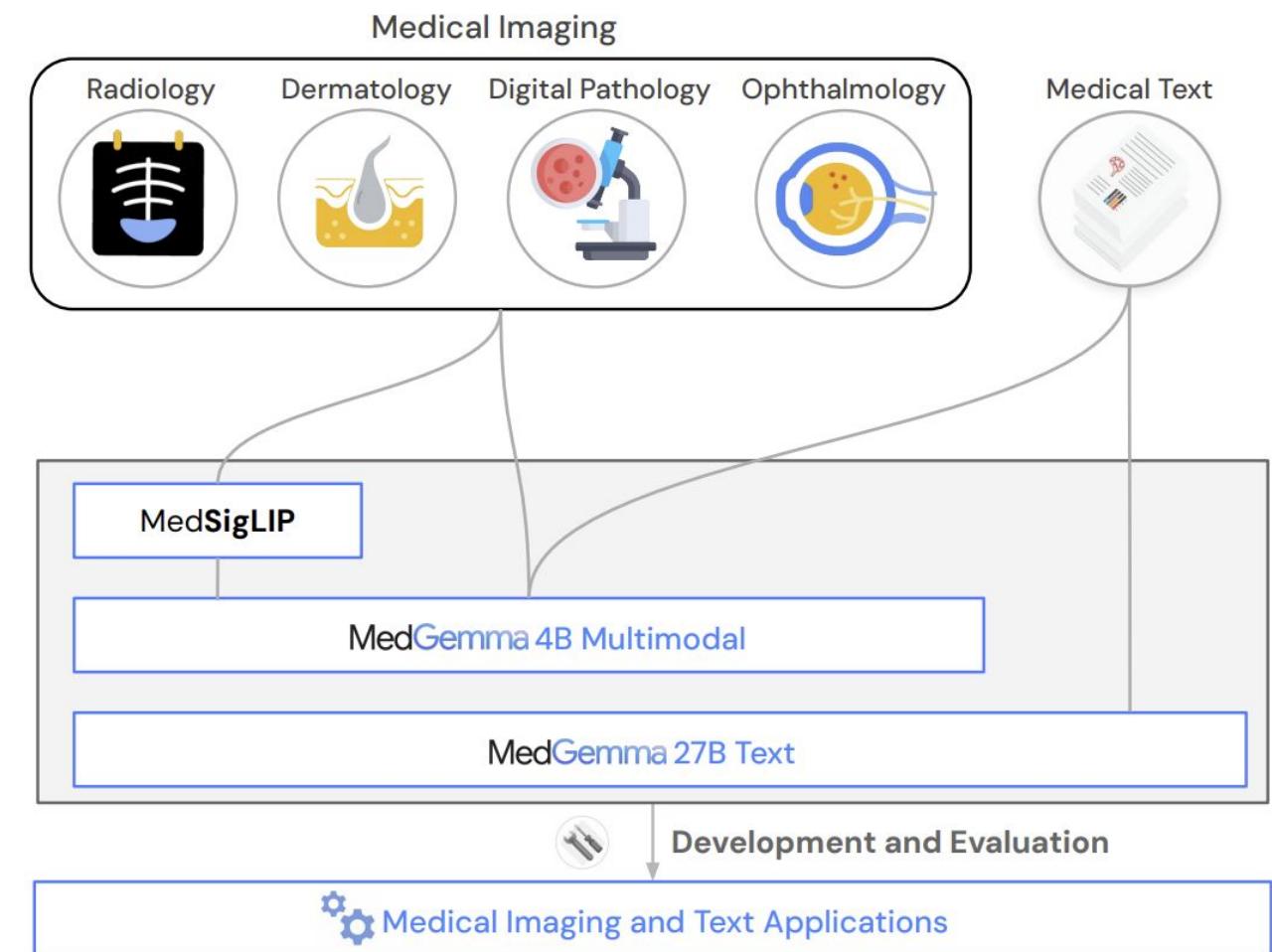
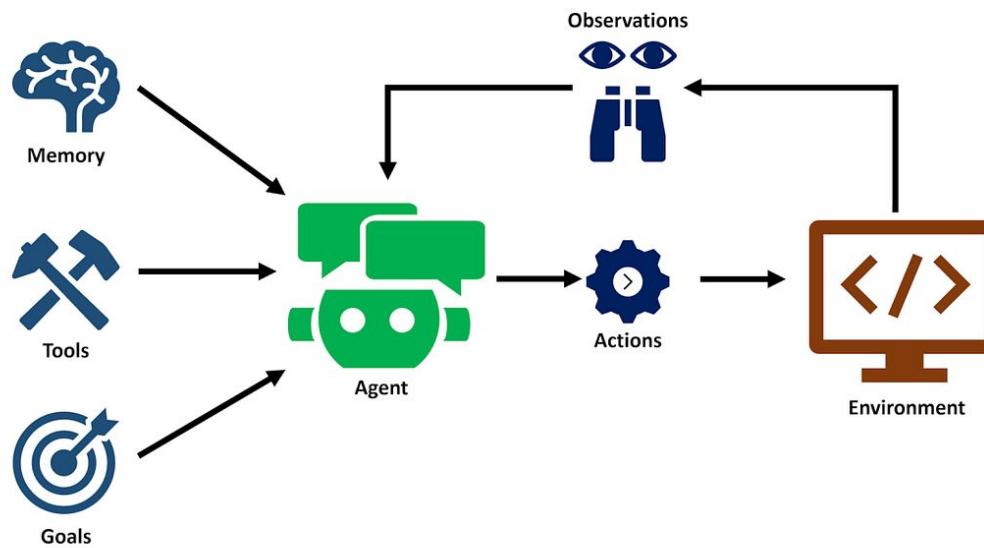
- ✓ **Human expertise** handles routine cases effectively
- ✓ **AI assistance** provides backup for challenging cases
- ✓ **Combined approach** achieves better outcomes than either alone



MedGemma (Latest Medical MLLMs):

- 4B for multimodal; 27B for text-only
- Application: **agentic orchestration**

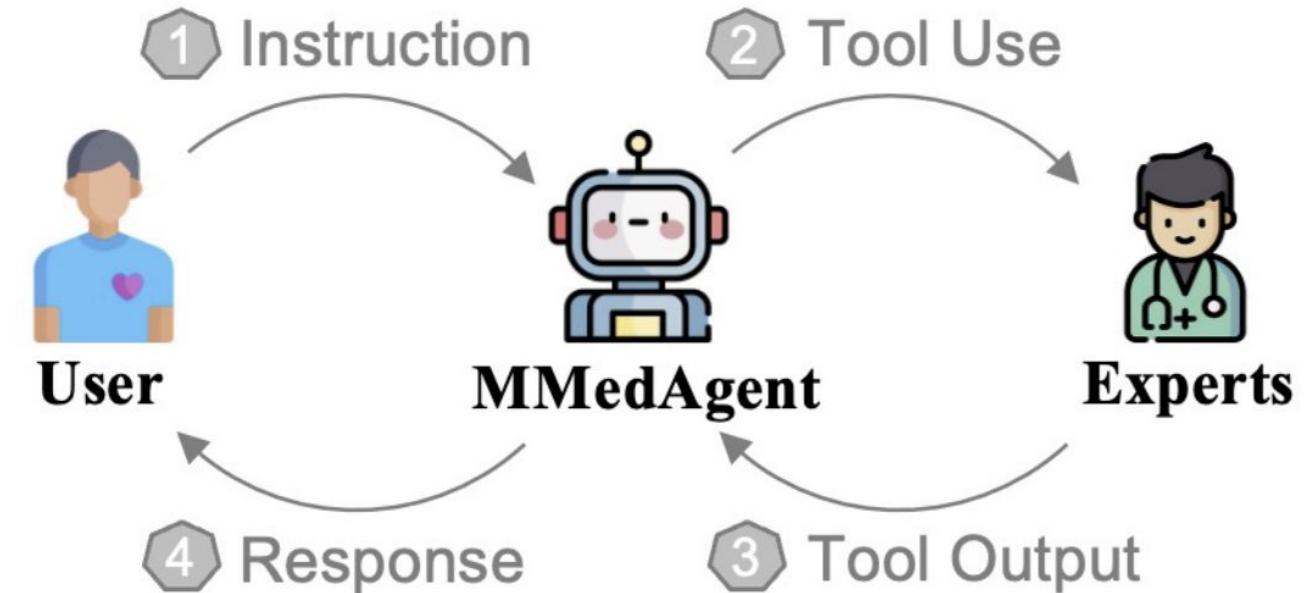
Tools designed to integrate with LLM-based agents, enabling them to perceive visual inputs.



Mmedagent:

Agentic system to orchestrate MLLMs

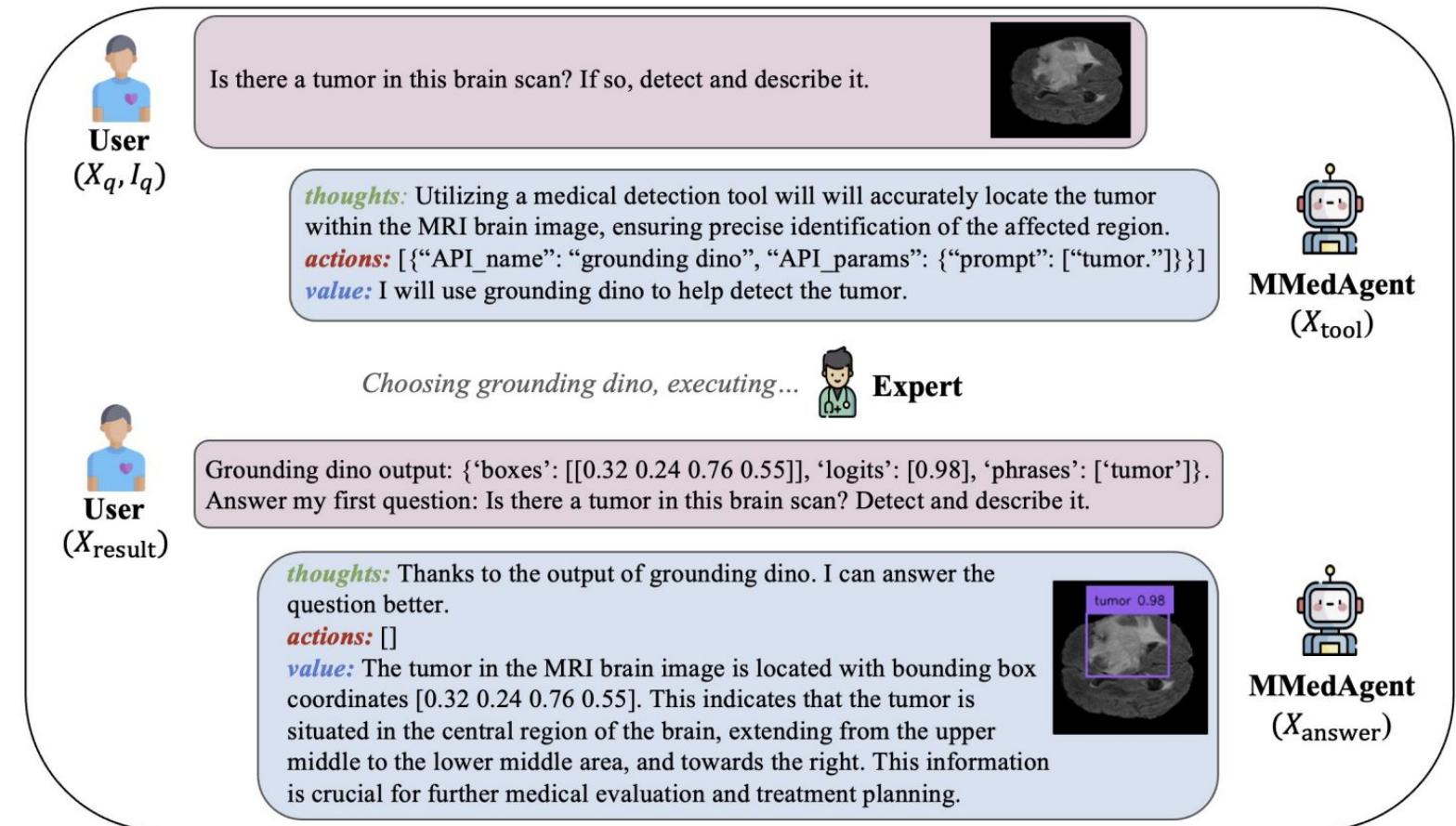
- LLMs as the planner (Agentic Core/Brain)
- MLLMs as a perception model to read the images (Eyes)
- Enhance the LLMs with multimodal ability while maintaining their superior, flexible reasoning ability.



Mmedagent:

Typical workflow:

- LLMs understand the user queries
- Call the related multimodal/visual APIs.
- Transform the visual output into textual information. (For MLLMs, their outputs are inherently texts).
- LLMs combine the information to obtain the final answer.



Mmedagent:

MLLMs' role in agentic systems:

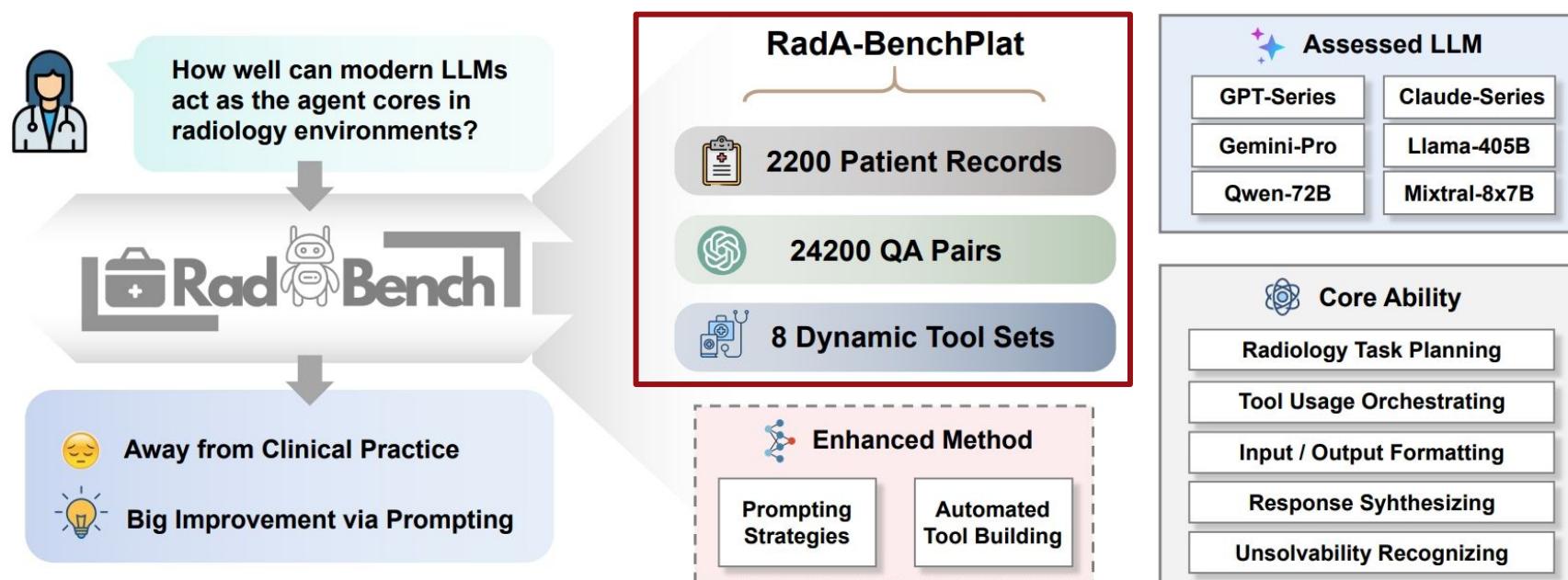
- MLLMs interpret free-text prompts, enabling natural integration with enhanced LLMs as sub-tools.
- They act as a flexible tool, complementing and extending the specialist tool collection in radiology.
- Other specialized tools—such as grounding models or diagnosis models—can make certain queries more precise.

Task	Tool	Data Source	Imaging Modality
VQA	LLaVA-Med (Li et al., 2023)	PMC article <i>60K-IM</i> (Li et al., 2023)	MRI, CT, X-ray, Histology, Gross
Classification	BiomedCLIP (Zhang et al., 2024a)	PMC article <i>60K-IM</i>	MRI, CT, X-ray, Histology, Gross
Grounding	Grounding DINO (Liu et al., 2023c)	WORD, etc.*	MRI, CT, X-ray, Histology
Segmentation	MedSAM (Ma et al., 2024a)	WORD, etc.*	MRI, CT, X-ray, Histology, Gross
G-Seg.	Grounding DINO + MedSAM	WORD, etc.*	MRI, CT, X-ray, Histology
MRG	ChatCAD (Wang et al., 2023b)	MIMIC-CXR (Johnson et al., 2019)	X-ray
RAG	ChatCAD+ (Zhao et al., 2024b)	Merck Manual (Porter and Kaplan, 2011)	—

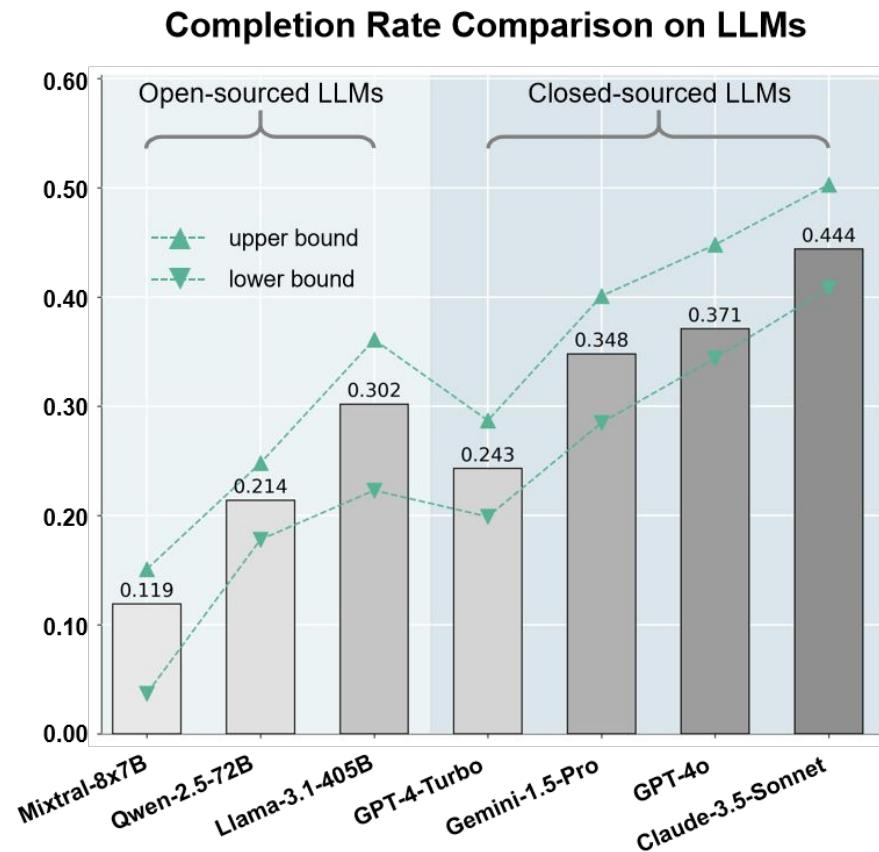
Before scaling, we investigate ...

How effectively can LLMs understand multimodal medical tools and plan its agentic workflow chain in radiological settings?

A simulated evaluation dataset tailed for assessing LLM-based agent cores' ability in radiology



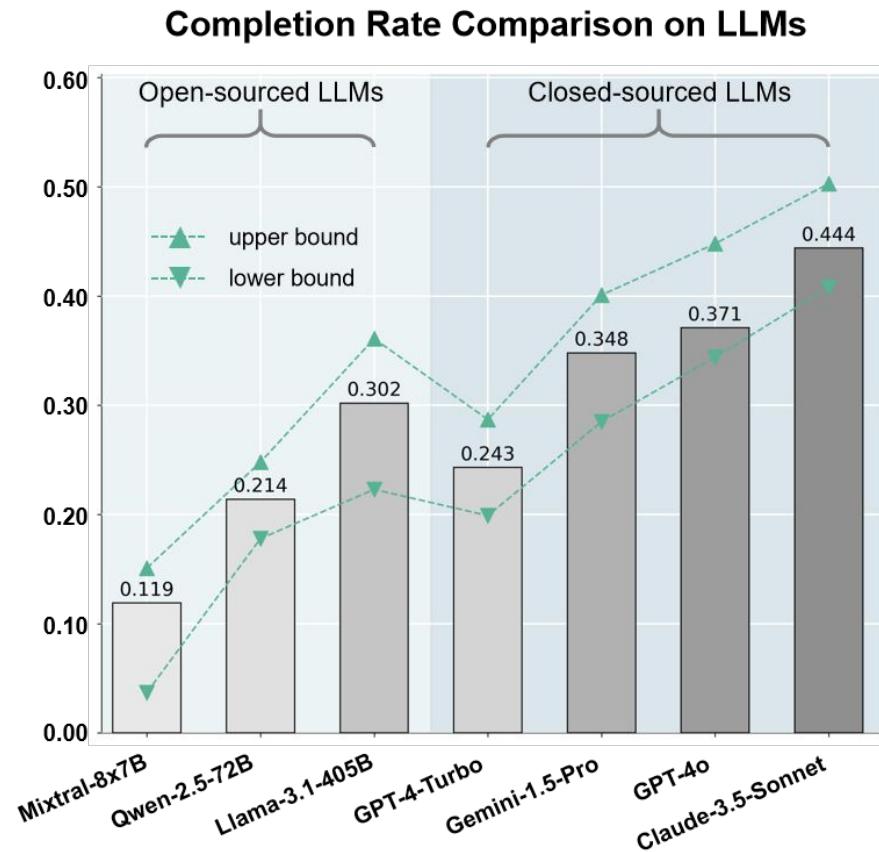
RadABench:



Findings:

- **Tool Misunderstanding:** Challenges in understanding complex or too many external tool contexts.
- **Context Loss:** Inefficiencies in synthesizing multi-round information.
- **Hallucination Risk:** LLMs are prone to significant “tool incomplete hallucinations”.
- **IO Misalignment:** LLMs struggle with organizing strict IO formats for successive tools.
- **Poor Tool Selection:** LLMs often fail to select the most appropriate tools based on their performance.

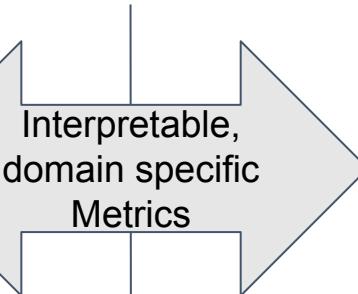
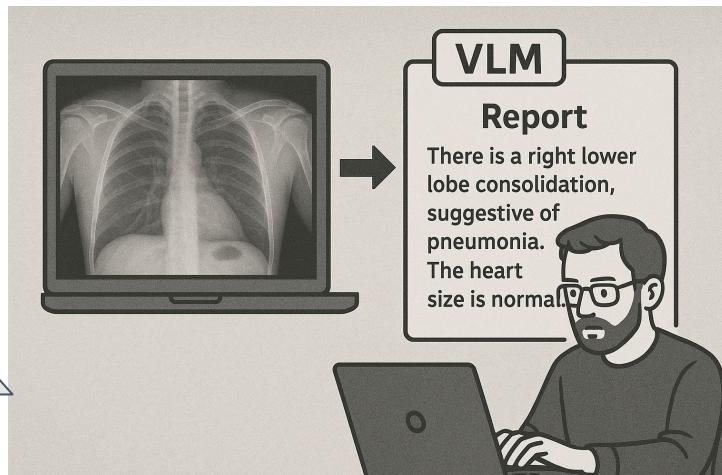
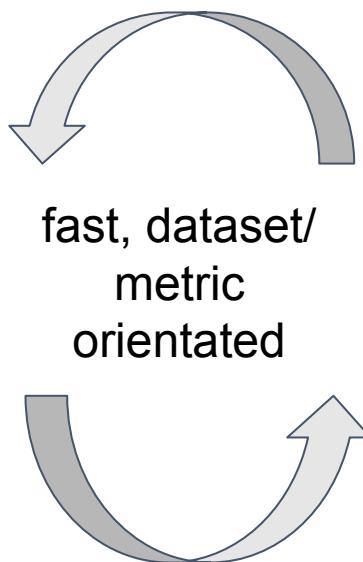
RadABench:



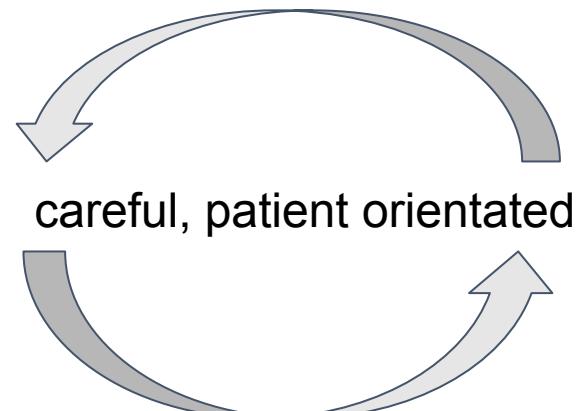
It remains essential to expand the capabilities of MLLMs in radiology—broadening task coverage, thereby reducing tool-context complexity, and potentially even fully replacing the LLM core.

Evaluation Metrics

Research: Model Development for Radiology Report Generation



Clinical Practice: Human Radiology Report Generation



IV. Evaluation Example

Ground Truth (Reference)

"The lungs are clear. **No pneumothorax or pleural effusion.**"

Evaluating two hypothetical AI models against one reference.

Model A ("The Hallucinator")

"The lungs are clear.
Pneumothorax is present."

CRITICAL ERROR False Positive

Model B ("The Stutterer")

"Clear lungs. No effusion.
No pneumothorax."

CORRECT DIAGNOSIS Poor Grammar

1. BLEU Score

(NLG Metric)

N-Gram Precision

The Theory

- Logic: "Word overlap with reference."
- Formula: Matches / Total Words

Applying to "The Hallucinator"

Gen: "...**Pneumothorax** is present."
Ref: "...No **pneumothorax**..."

~0.65 (High)

⚠ Verdict: BLEU rewards hallucinations if keywords match. It ignores "No".

2. ROUGE-L (NLG Metric)

Longest Common Subsequence

The Theory

- Logic: "Longest chain of words in same order."
- Role: Measures Fluency & Structure.

Applying to "The Stutterer"

Ref: "The **lungs** are **clear**..."
Gen: "**Clear lungs**..."

LOW

Sequence broken. Matches content but fails on structure.

3. CIDEr (NLG Metric)

TF-IDF Weighting

The Theory

- ⚖️ Logic: "Common words are cheap. Rare words are expensive."

"The" → Weight: 0.01

"Pneumothorax" → Weight: 10.0

Applying to "The Hallucinator"

Gen: "...**Pneumothorax** is present."

VERY HIGH

It successfully generated the rare, high-value keyword "Pneumothorax". CIDEr is great for topic matching but bad at Positive/Negative distinction.

4. CheXpert F1 (Clinical Efficacy)

GOLD STANDARD

The Theory

- Logic: "Ignore style. Extract diagnosis (0 or 1)."
- Process: Text → Labeler → Vector → F1 Score

The Moment of Truth

Ref
0

Model A
1 (FP)

Model B
0 (TN)

Model A Score:

0.0

Model B Score:

1.0

5. BERTScore (Semantic Metric)

Vector Similarity

The Theory

- 🧠 Logic: "Semantic Equivalence."
- ▣ Method: Contextual Embeddings (BERT/Roberta).

The "Stutterer" Advantage

Ref: "The lungs are **clear.**"
Gen: "No **opacity.**"

BLEU

0.0

BERTSCORE

High

BERTScore understands "Clear" and "No opacity" are clinically identical, unlike BLEU.

6. LLM-Based Metrics (GREEN & MRScore)

The New Frontier

GREEN

GENERATIVE ENTAILMENT (STANFORD 2024)

-  **Logic:** Uses an LLM to judge if Candidate entails Reference.
-  **Strength:** Detects contradictions (e.g., "No pneumothorax" vs "Pneumothorax present") better than any overlap metric.

Result for Model A: **Contradiction (Score: 0)**

MRScore

LLM-BASED REWARD SYSTEM

-  **Mechanism:** An Reward to score reports on LLM Model Accuracy, acts as a Completeness, and Conciseness.
-  **Dual Use:** Can be used for Evaluation AND RLHF . Use: as a Reward Function for training

Verdict: Scores align 95% with radiologists.

MICCAI 2024

7. Advanced Clinical Metrics

STRUCTURE & COMPOSITE

RadGraph

Uses Knowledge Graphs instead of simple labels.

```
Nodule --(located_at)--> Right  
Upper Lobe
```

✓ Captures Anatomy & Modifiers.

RadCliQ

A Composite Metric optimized for radiologist preference.

$$\text{RadCliQ} = w1 * \text{BLEU} + w2 * \text{CheXpert} + w3 * \text{BERTScore}$$

✓ Combines NLG fluency with Clinical accuracy.

8. Human Evaluation

THE ULTIMATE JUDGE



Accuracy

No false positives/negatives.



Completeness

All abnormalities mentioned.



Fluency

Professional language.

Method A: Likert Scale

Unacceptable (1)  Perfect (5)

Method B: Error Counting

Major Errors (Risk):

Count

Minor Errors:

Count



LLM-as-a-Judge

Using GPT-4 to evaluate reports (G-Eval).



Patient-Centered

Evaluating "Layman Summaries".



Visual Grounding

Checking text-to-image alignment.

Multimodal LLMs for Pathology

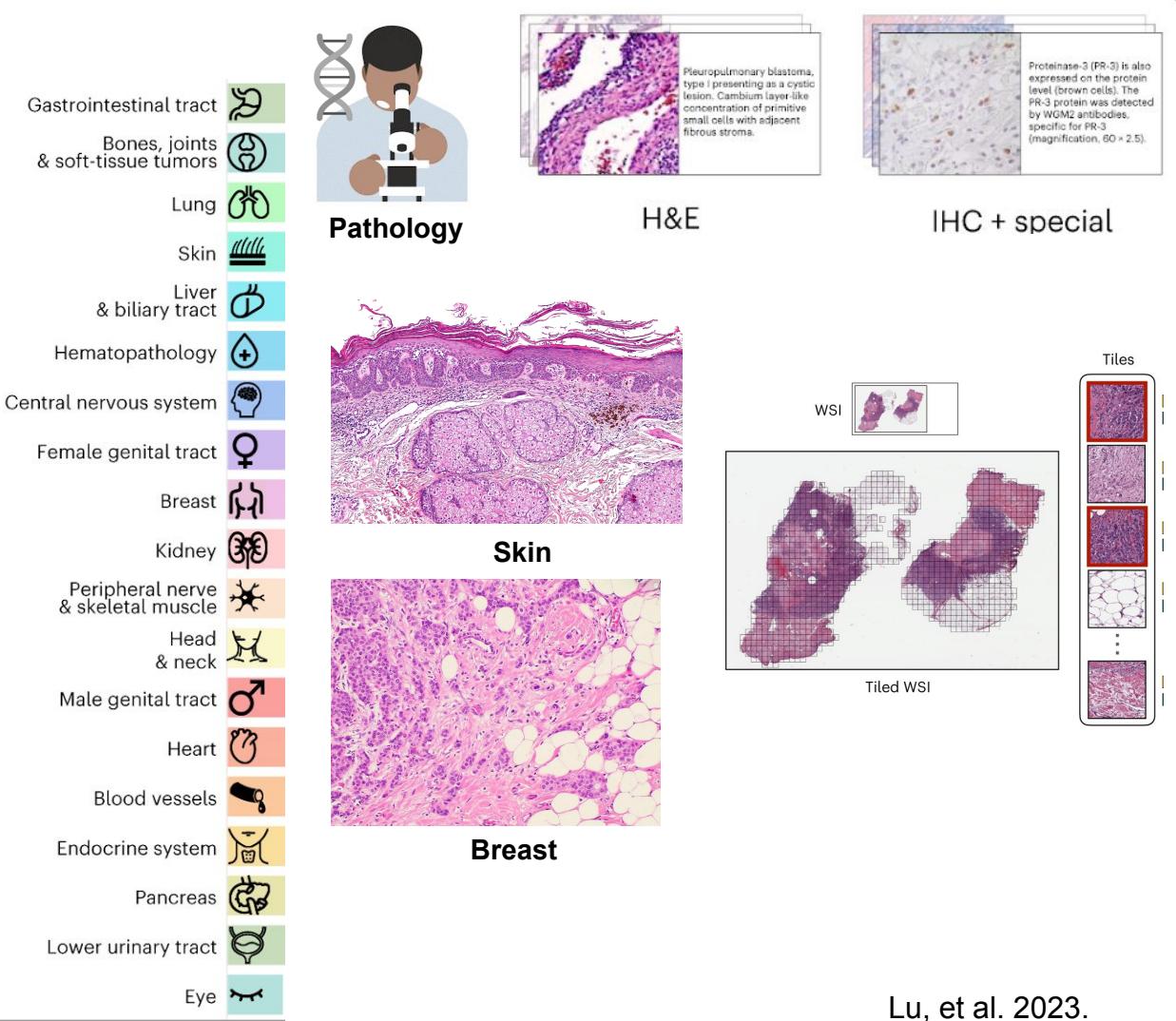
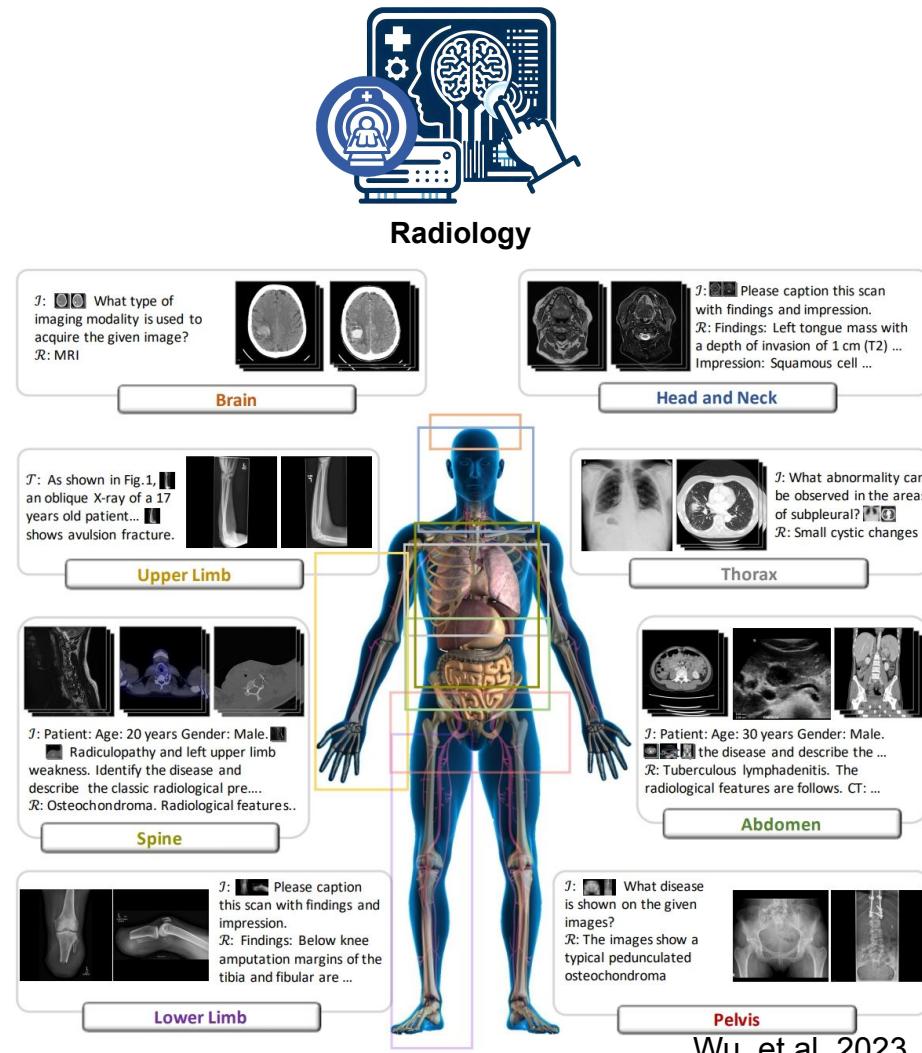
Yunsoo Kim

Institute of Health Informatics, University College London

24 November 2025

VII. Radiology vs Pathology

How Radiology is different from Pathology/Histopathology



How Radiology is different from Pathology/Histopathology

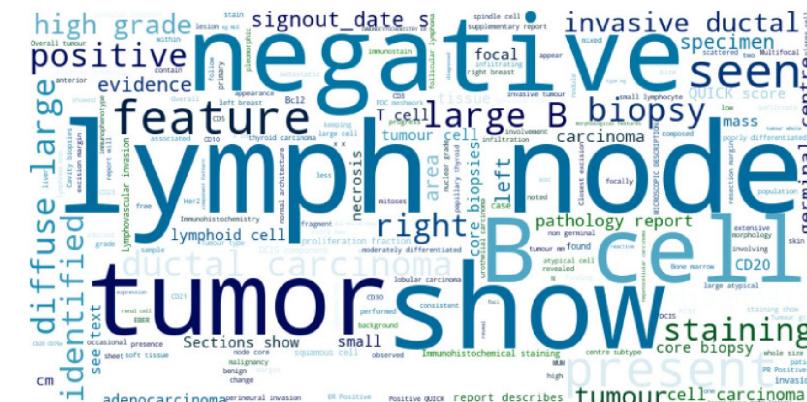


Reports composed of:
Findings
Impression



Pathology

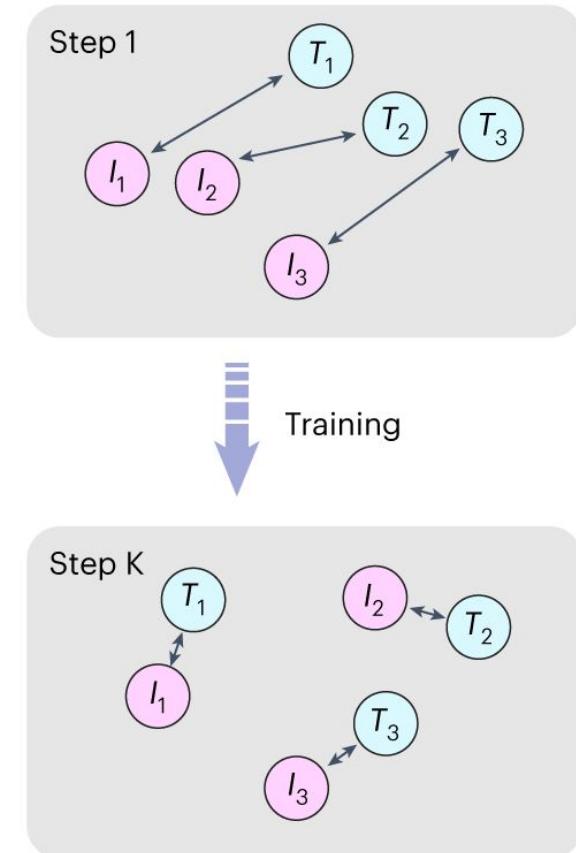
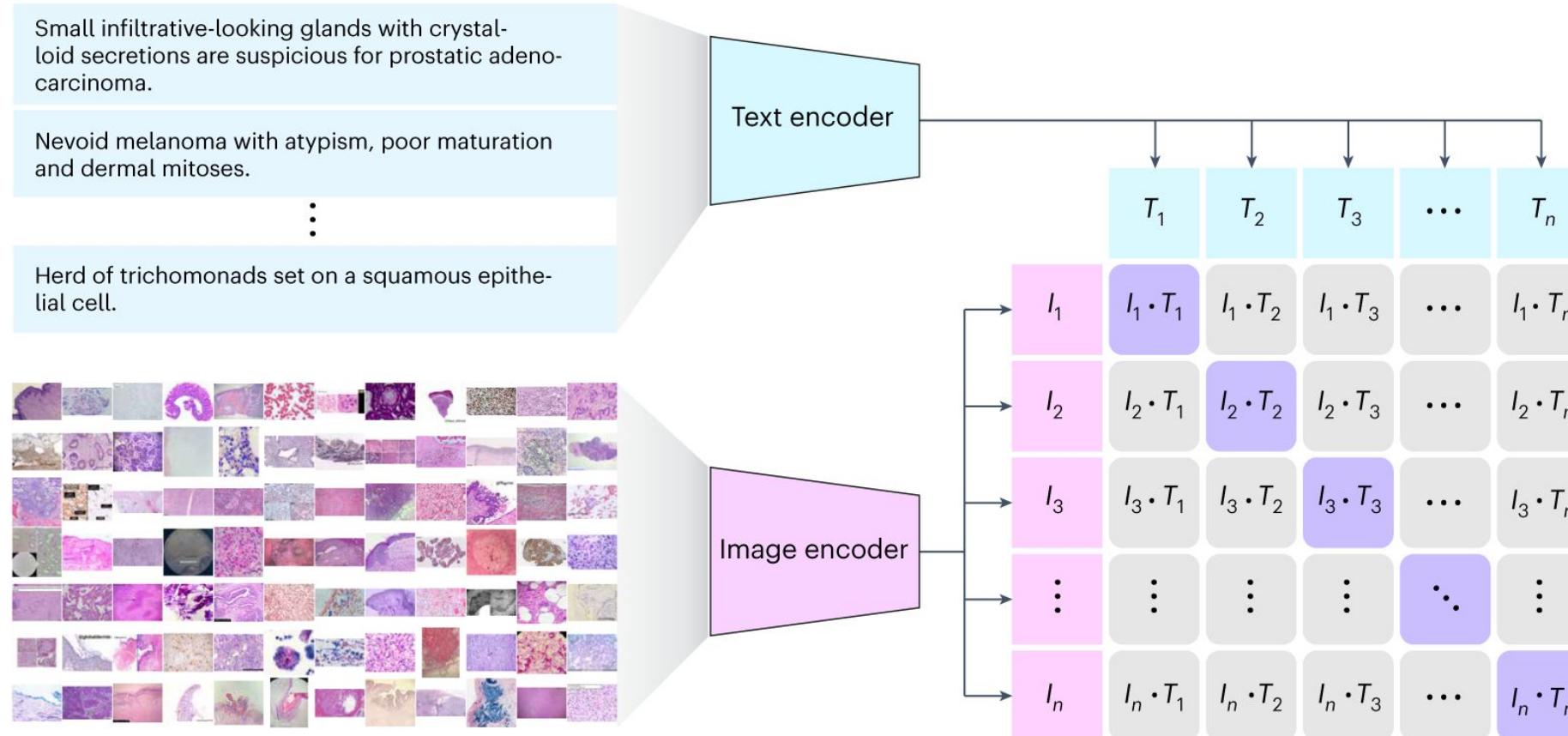
Reports composed of:
Specimen
Macroscopic description
Microscopic description
Final Diagnosis



VLMs in Pathology

VII. Vision Language Model in Pathology

PLIP first visual-language foundation model (CLIP-based) in Pathology Contrastive Learning



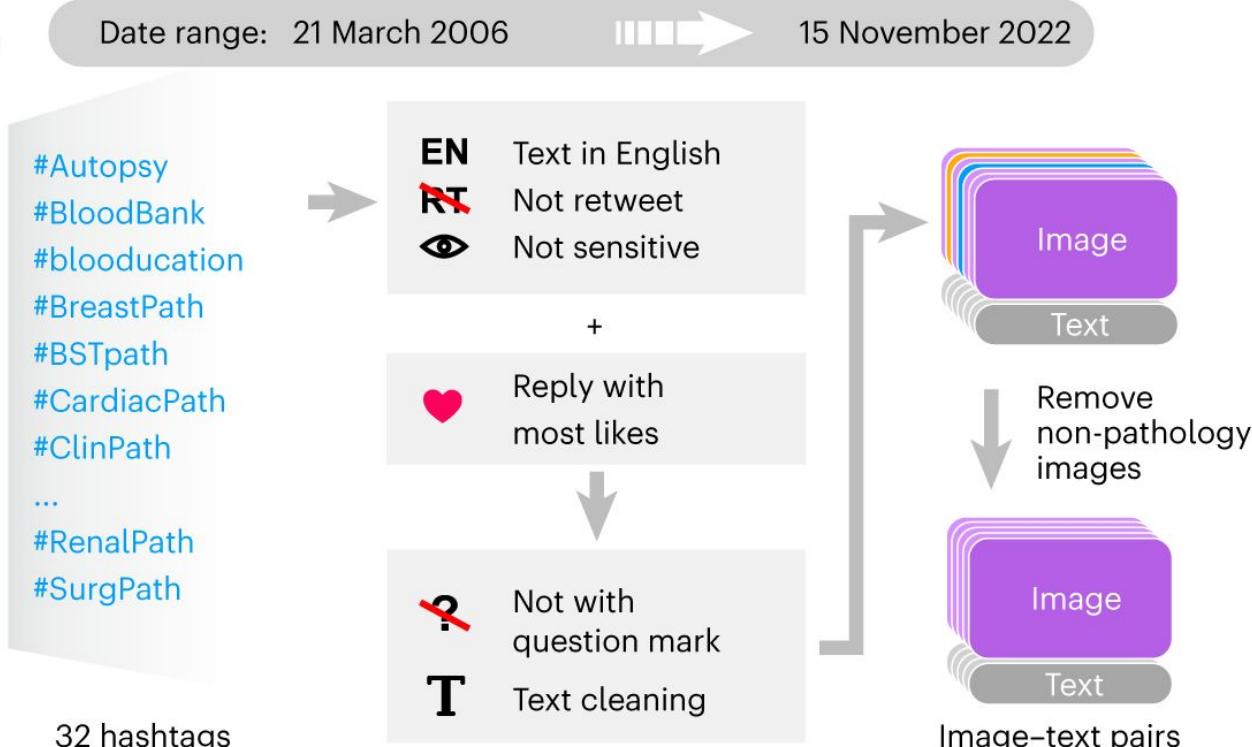
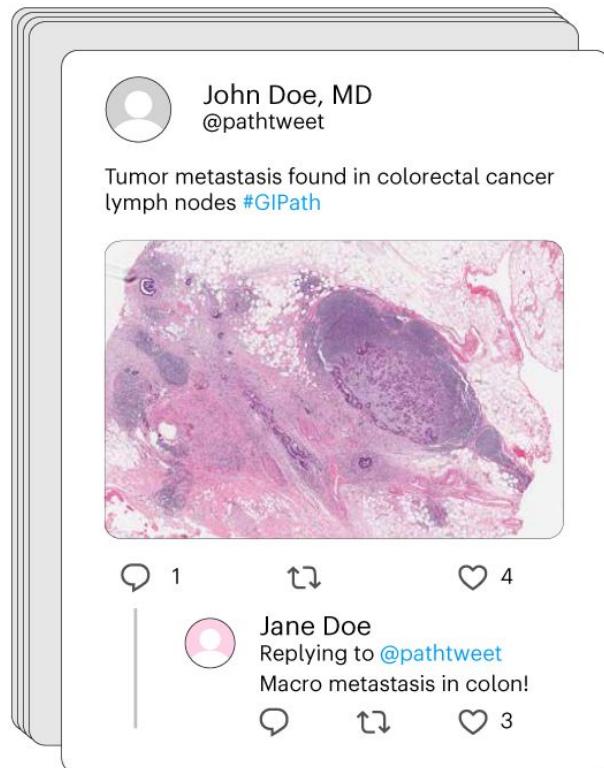
VII. Vision Language Model in Pathology

PLIP trained using 200K tile-caption datasets from Twitter

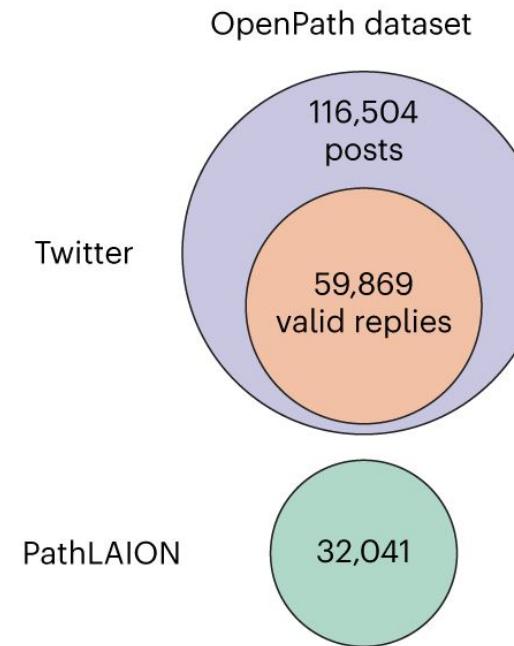
Tile-Caption from Twitter (176K) and LAION dataset (32K)

- For Twitter: HashTags and Replies were used to construct the data

a

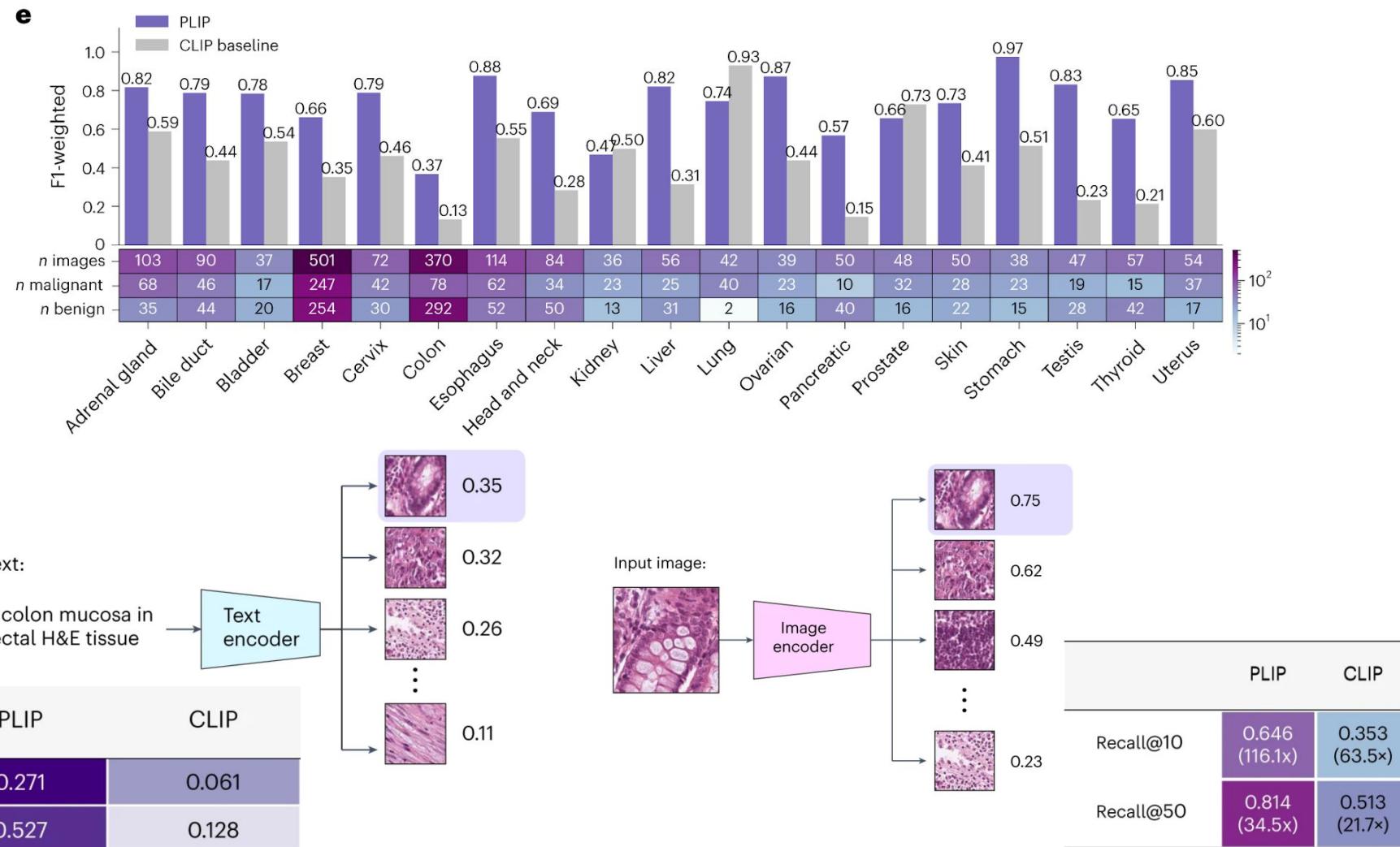


b



VII. Vision Language Model in Pathology

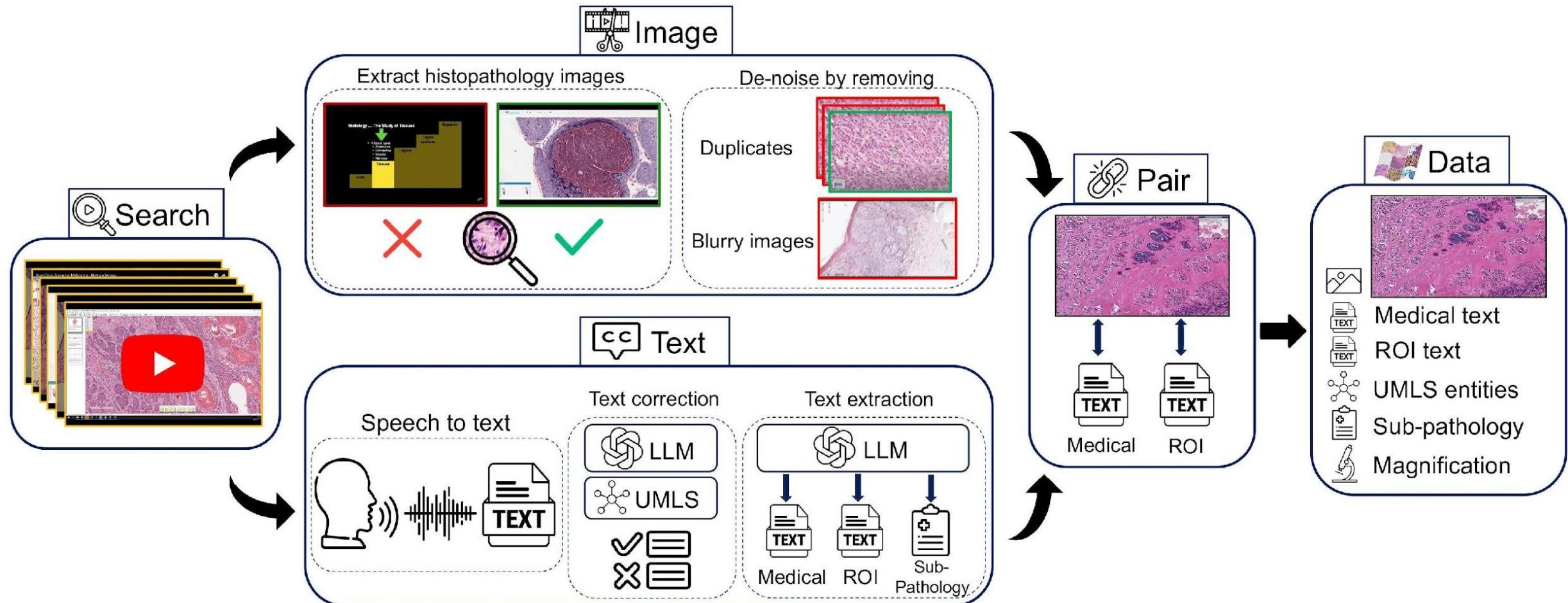
PLIP was tested on zero-shot classification, image-image retrieval, and text-image retrieval



VII. Vision Language Model in Pathology

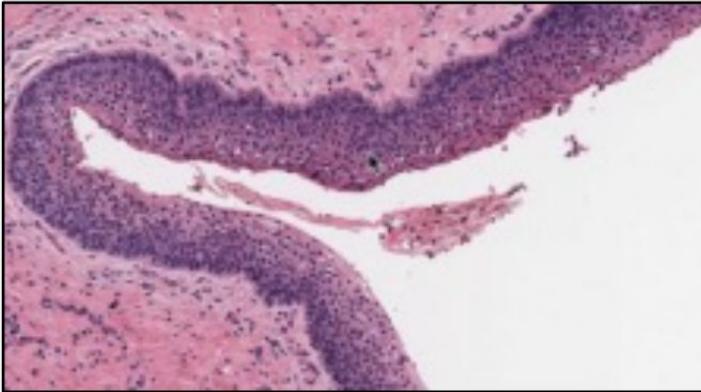
QuiltNet is a CLIP model trained with a dataset from Youtube narrative-style pathology videos

1. Extracted histopathology images (438K) from videos using keyframe histopathology image ensemble classifier
2. Converted speech to text and corrected the text (802K) using an LLM (GPT-3.5) and UMLS



VII. Vision Language Model in Pathology

Quilt Dataset : Region of Interest Text – further extracted for phrases with “here” or “look”



INPUT: "...so here we have a huge lumen and we have the lining epithelium and then we have the underlying connective tissue capsule. let's see the features in detail one by one. if you see this lumen lumen has sparse amount of keratin. so definitely this is a keratinous cyst. so we have a cyst that is actually producing keratin. let's move into the epithelium and the epithelium if you carefully observe most of the areas the epithelial appears uniform thickness right from here.

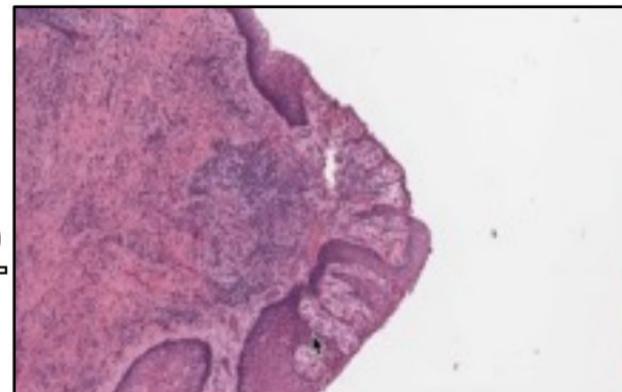
OUTPUT:

{'MED':["The cyst is producing keratin, indicating it is a keratinous cyst."],

"The epithelial lining appears to have uniform thickness"]

ROI:

"Sparse amount of keratin in the lumen.",
"uniform thickness of the epithelial lining"]}



INPUT: "...tend to see some inflammation here yes. when you notice this there is a dense amount of inflammation here yeah you see this tiny blue dots these are the lymphocytes and these cells are your inflammatory cells and they are kind of having an effect on the epithelium. if you see this epithelium because of the inflammation the epithelium starts "

OUTPUT:

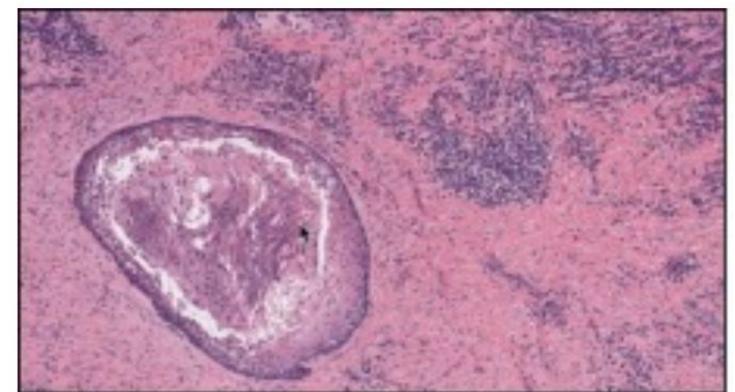
{"MED":

"Dense inflammation is present in the **region** with lymphocytes and other inflammatory cells having an effect on the epithelial cells"]

'ROI':["dense inflammation **here**",

"lymphocytes",

"inflammatory cells"]}}



INPUT: "...of a daughter cyst where you see so this was an epithelial island with showing central keratinization the cells producing lot of keratin and at the same time you have the epithelium with the lumen. so this forms a tiny daughter cyst and this is one of the most important trees in why this keratocyst has a high amount of recurrence rate. one more important feature of odontogenic keratocyst is lies within the epithelium."

OUTPUT:

{"MED":["The presence of daughter cysts is one of the reasons for the high recurrence rate of keratocysts",

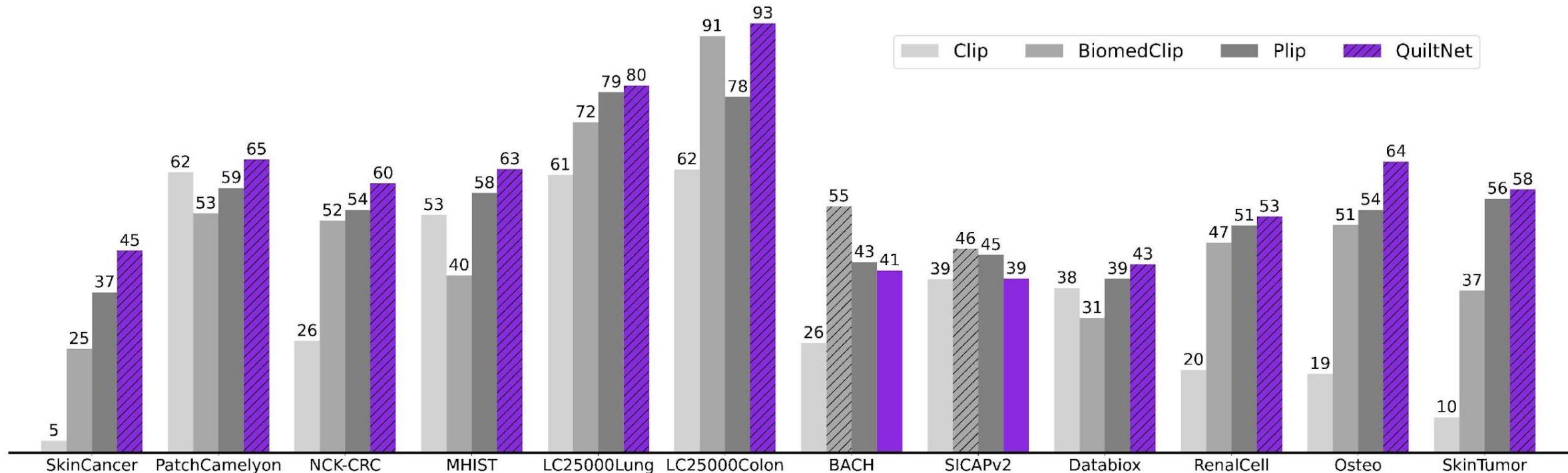
"Epithelial islands with central keratinization and high keratin production are seen in keratocysts"]}

'ROI':["epithelial island",

"central keratinization", "daughter cyst"]}}

VII. Vision Language Model in Pathology

QuiltNet test on mostly classification and image-text retrieval. Outperforming PLIP.

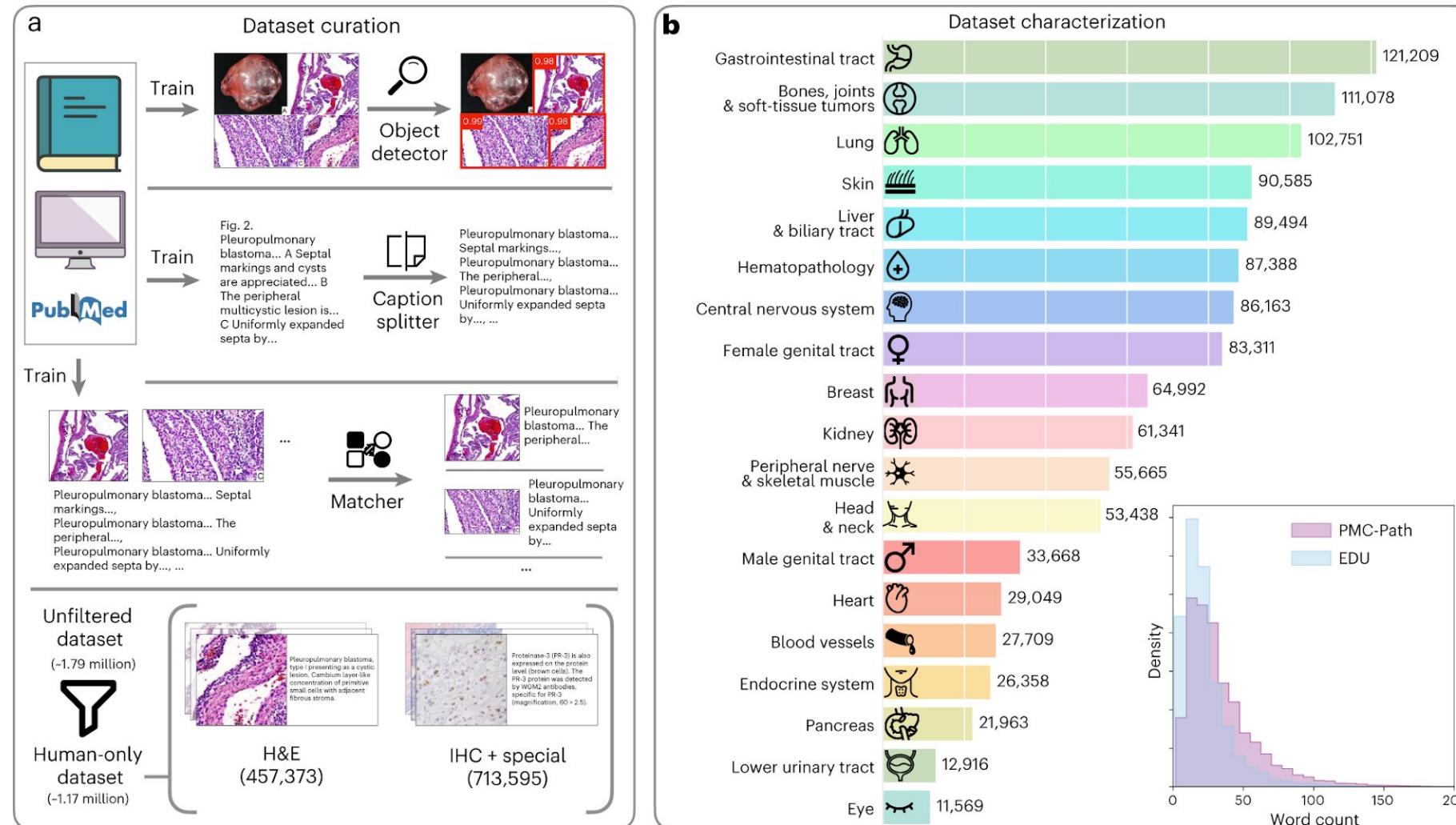


model	config	Text-to-Image (%)			Image-to-Text (%)		
		R@1	R@50	R@200	R@1	R@50	R@200
CLIP	ViT-B/32 GPT/77	0.49/0.07	4.73/2.42	10.15/7.21	0.39/0.05	3.99/2.52	8.80/7.22
PLIP	ViT-B/32 GPT/77	1.05/0.56	10.79/13.10	21.80/29.85	0.87/0.74	11.04/13.75	21.63/29.46
QUILTNET	ViT-B/32 GPT/77	1.17/1.41	16.31/19.87	31.99/39.13	1.24/1.35	14.89/19.20	28.97/38.57

VII. Vision Language Model in Pathology

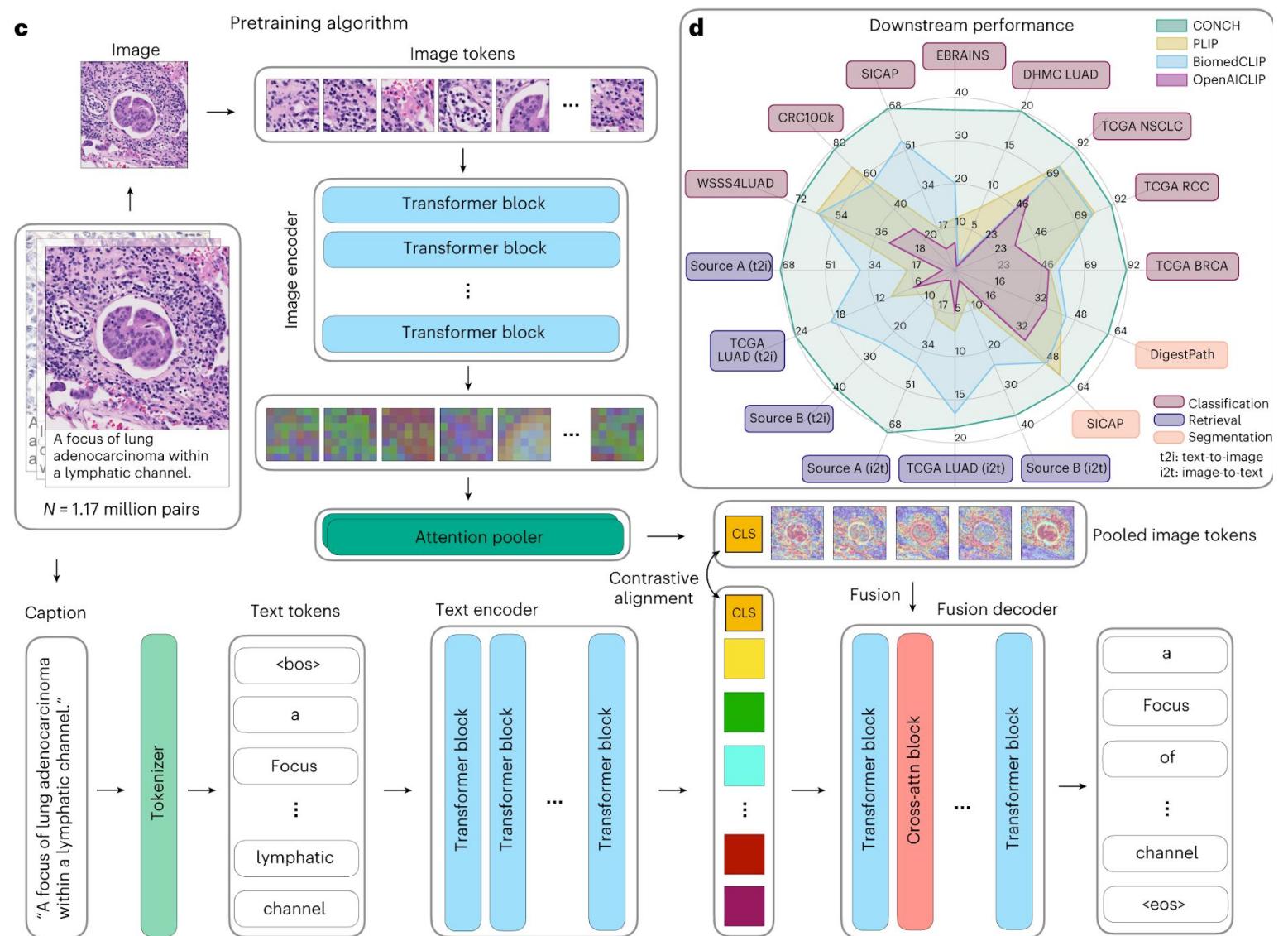
CONCH – Foundation Model trained with PubMed OA (1.17M Tile-Caption Pairs)

Captions split using BioGPT



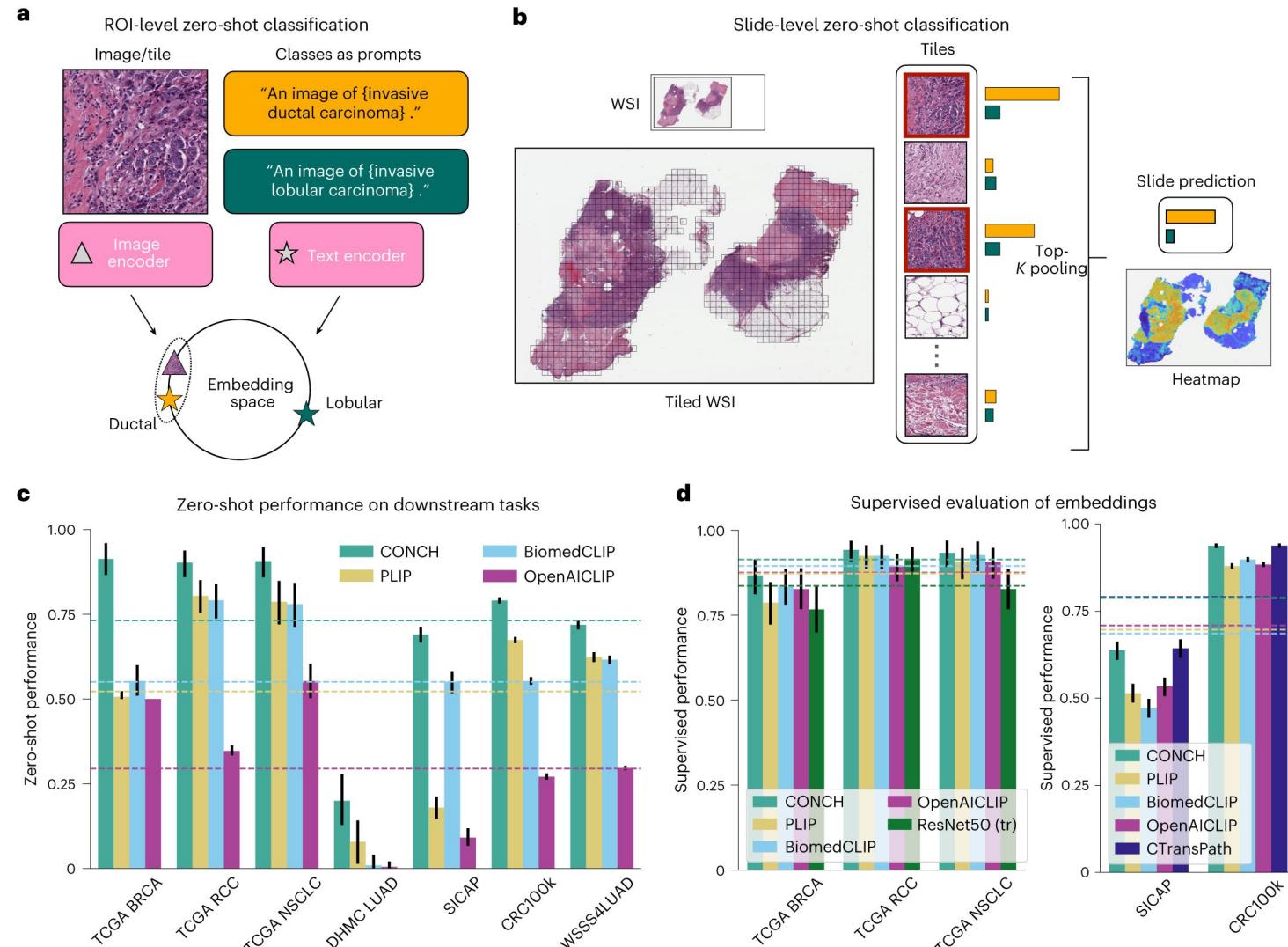
VII. Vision Language Model in Pathology

CONCH – CoCa Approach

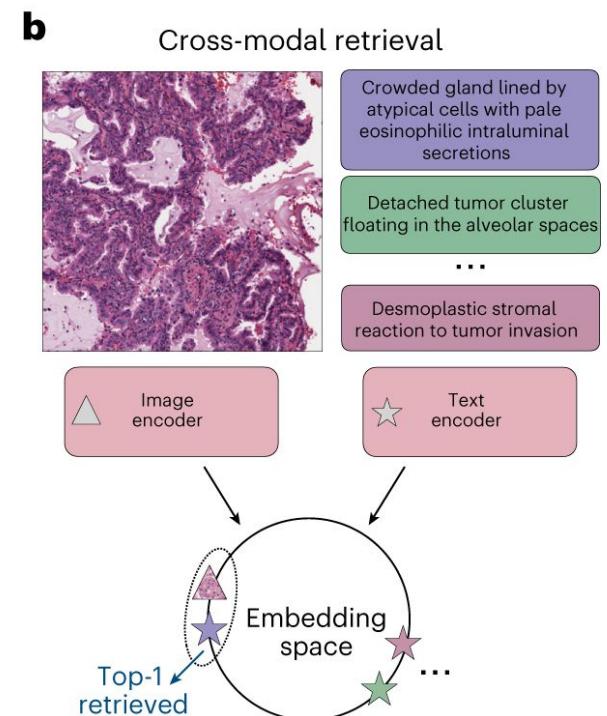
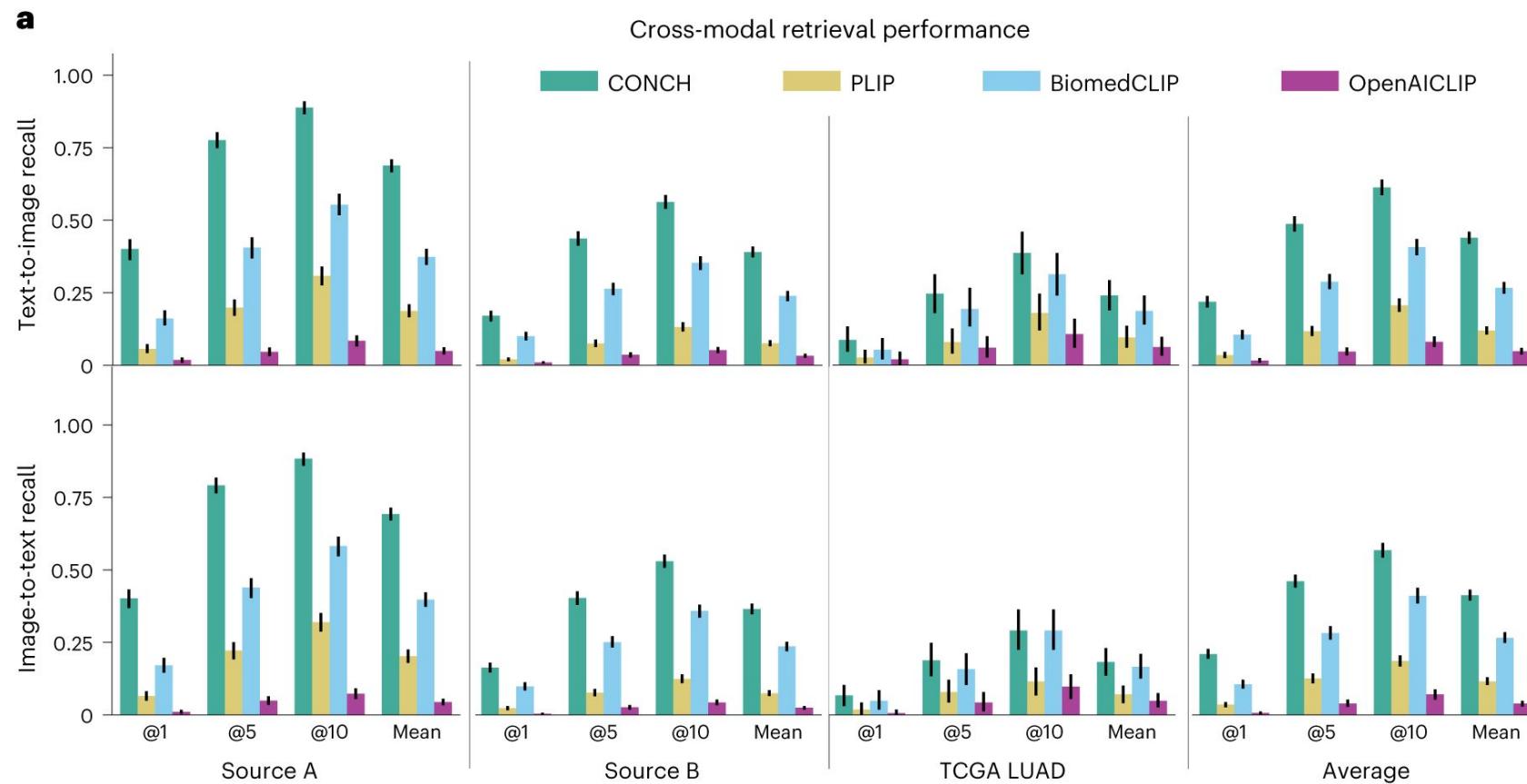


VII. Vision Language Model in Pathology

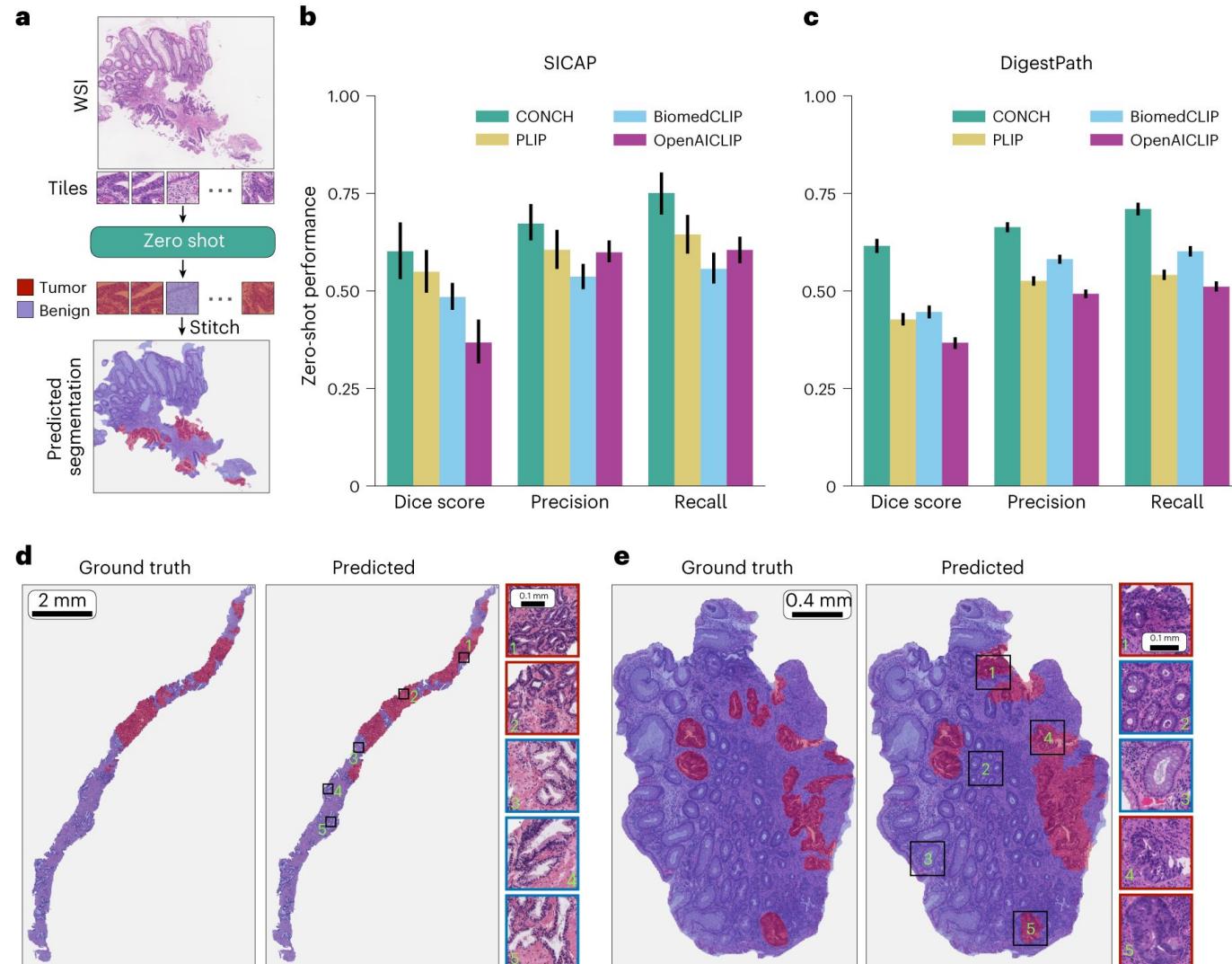
CONCH – outperforming PLIP in Zero-shot classification



CONCH – outperforming PLIP in retrieval



CONCH extended to segmentation



VQA Chat Assistant Reports generation

Infiltrating duct carcinoma measuring 2.5 cm x 2.0 cm x 1.5 cm, nuclear and architectural grade 2. Eight out of thirteen lymph nodes with metastatic carcinoma and extracapsular extension.

Lymphovascular space involvement present.

Ductal carcinoma in situ, predominantly solid type, 30% of tumor volume.

- Report from TCGA

VLMs downstream tasks are limited by small number of input tokens

Context length for input text prompt is much shorter than a question and report

Multimodal LLMs in Pathology



Proprietary models leading...

The privacy concerns with the patient data
Still, we have way around this.

A Hallucination Benchmark of Medical Images including Pathology

Modified datasets

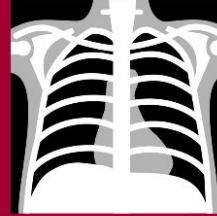
PMC-VQA¹



PathVQA²



VQA-RAD³

<p>1) FAKE question - Hallucinations from Q Fake question generated by GPT3.5-Turbo</p>	<p>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?</p>
<p>2) NONE of the above - Hallucinations from A Correct answer is replaced with NOTA</p>	<p>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</p>
<p>3) Image SWAP - Hallucinations from I Image swapped with unrelated one</p>	 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	77.90	0	8.70	0	79.71	0	55.44	0.0
GPT-4-turbo-vision	72.93	43	44.40	44	72.37	40	63.23	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



For report generation, even GPT-4o Vision performs poorly



Histopathologist scored 600 reports generated with GPT-4o vision based on the following criteria:

- 5** - Perfect match with ground truth
- 4** – Correct diagnosis with at least one minor error in histopathological description
- 3** – Correct Diagnosis but completely wrong in histopathological description
- 2** – Correct Disease Category
- 1** – Incorrect Diagnosis with some description matching
- 0** – Incorrect diagnosis with no histological description matching

Score	Count
0	369
1	71
2	90
3	62
4	8
5	0



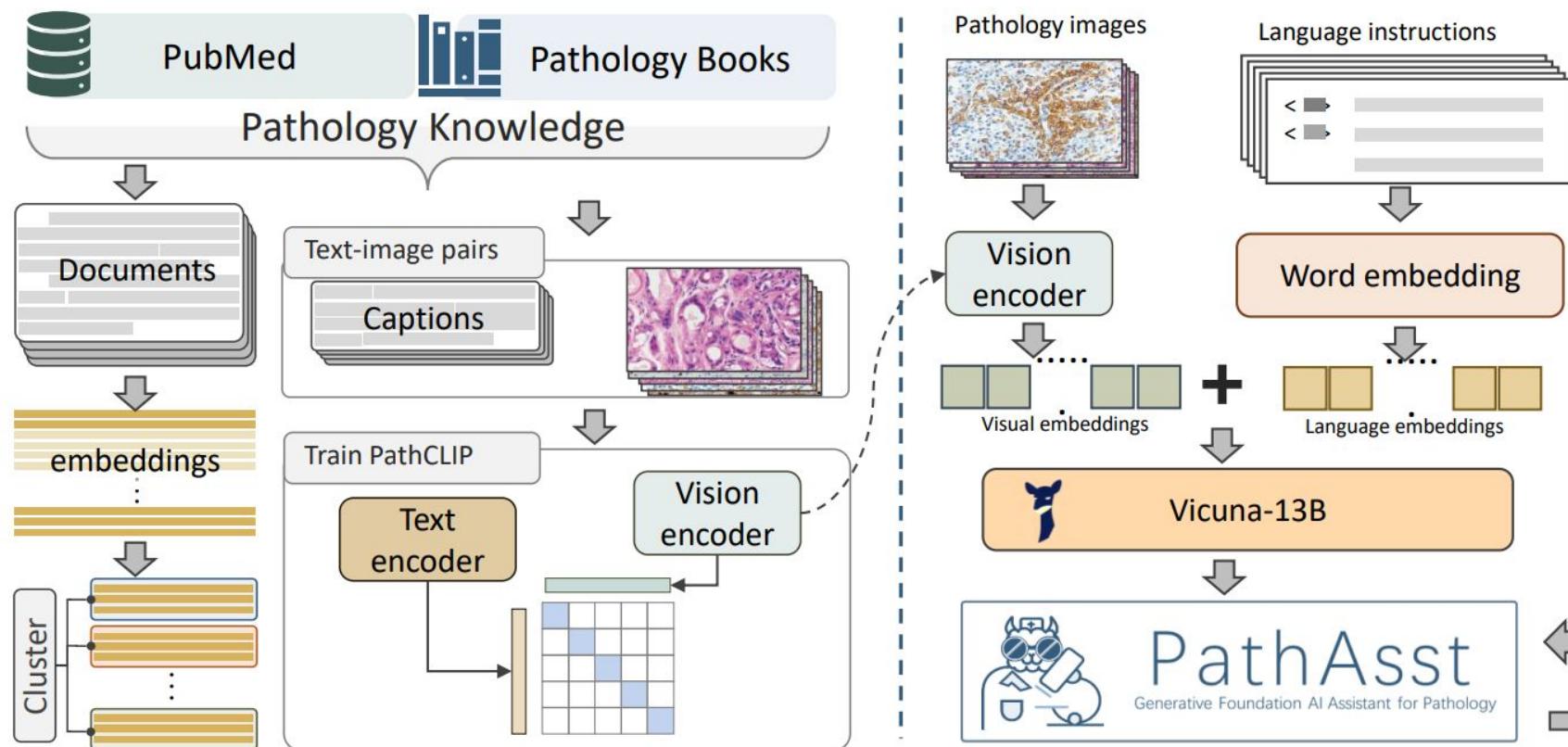
We are not far
behind the
proprietary models.

Domain adaptation with high
quality pathology dataset

VII. Multimodal LLMs in Pathology

PathAsst first multimodal LLMs in Pathology based on Vicuna

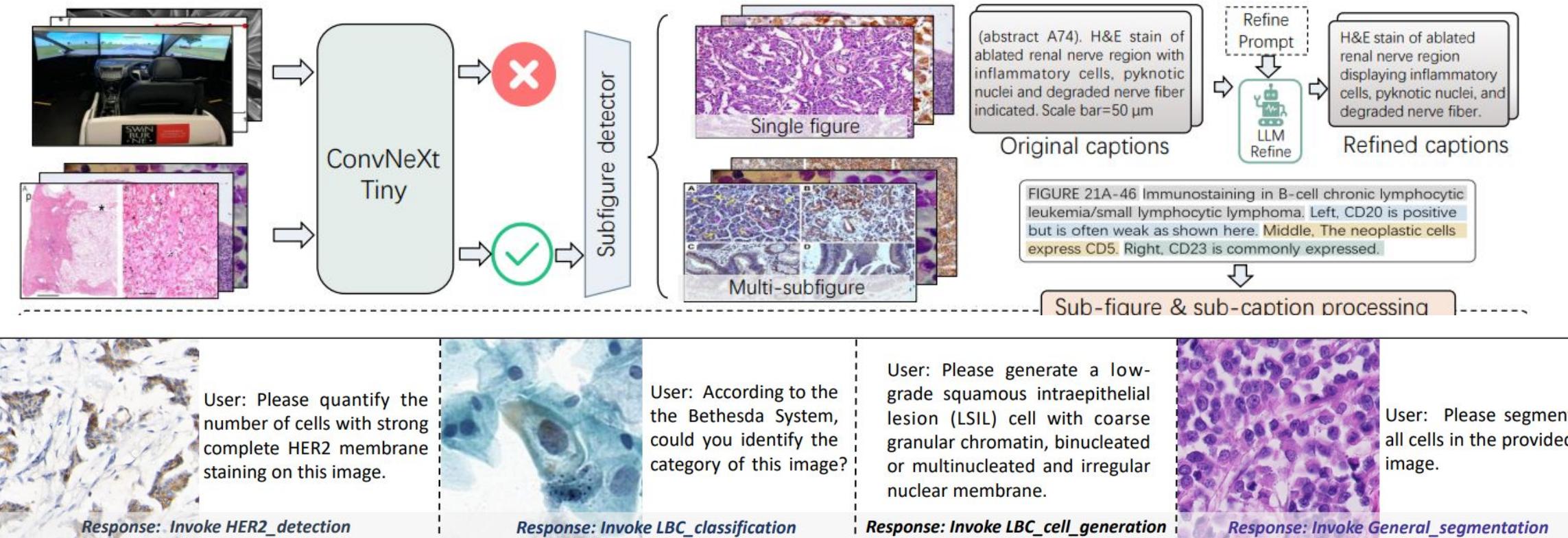
1. PathCLIP - Image Encoder Newly Trained from ViT
2. FC layer used to connect vision and language embeddings



VII. Multimodal LLMs in Pathology

Path-Asst uses LLaVA approach but backbone LLM is Vicuna

- 1 stage training with 207K sourced from PubMed for captioning
- 2 stage training with 180K pathology instruction tuning dataset : 4 tasks (Generated by ChatGPT)



VII. Multimodal LLMs in Pathology

PathCLIP outperforms PLIP in classification and retrieval and PathAsst outperforms LLaVA in VQA

Model	CRC	WSSS4LUAD	LC-lung	LC-colon
OpenAI CLIP	22.2	61.6	31.5	75.7
PLIP	53.1	69.5	86.0	87.0
PathCLIP	54.2	81.1	88.7	94.3

PathCLIP – Zero-Shot Classification

Metric	OpenAI CLIP	PLIP	PathCLIP
R@10	3.1	3	33.2
R@50	8	8.9	56.9

PathCLIP – Image Retrieval

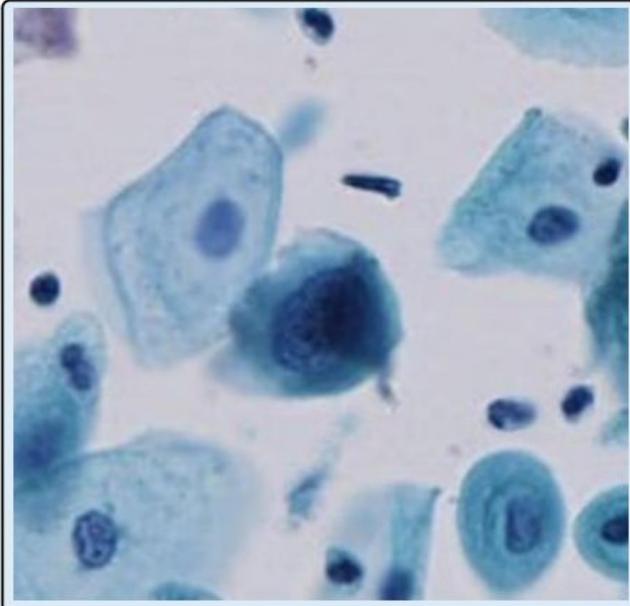
Method	PathVQA	
	Closed	Open
M2I2 (Li et al. 2022)	88.0	36.3
CLIP-ViT w/ GPT2 (van Sonsbeek et al. 2023)	87.0	40.0
MMQ (Do et al. 2021)	84.0	13.4
LLaVA (Liu et al. 2023a)	81.0	19.2
BLIP-2 Flan-T5 XXL (Li et al. 2023)	80.1	34.1
PathAsst (w/ CLIP)	89.7	37.6
PathAsst (w/ PathCLIP)	90.9	38.4

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)

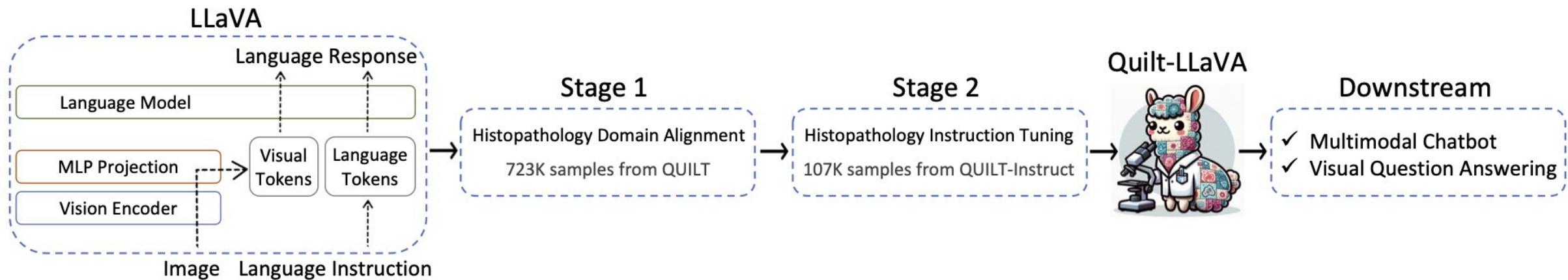
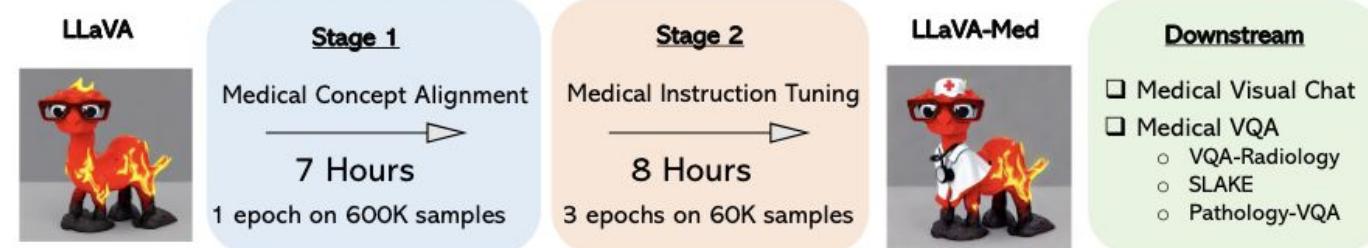
Based on these observations. I would recommend further testing to confirm the presence of abnormal cells. An abnormal cervical Pap test result may indicate a high-risk for cervical cancer. It's important to follow up with your healthcare provider for further evaluation and treatment.

Llava

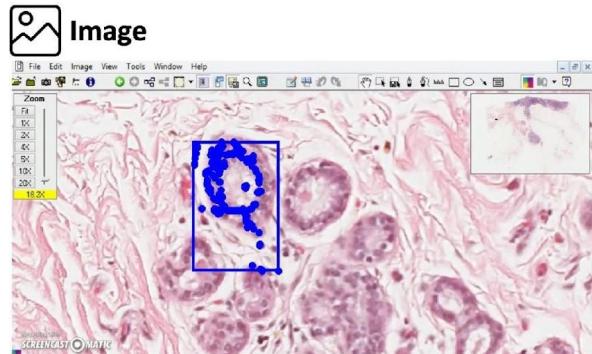
The image is a close-up view of a blue substance, possibly ink or a dye, dispersed in water. The blue particles are evenly distributed throughout the water, creating a visually striking scene. The blue substance appears to be in the form of small spheres or shapes, which are spread across the entire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

VII. Multimodal LLMs in Pathology

Quilt LLaVA : further training LLaVA with 2 stages : domain alignment and visual instruction tuning
Very similar approach to LLaVA-Med, but using high-quality histopathology specific dataset



For the instruction tuning, created the dataset using the Quilt Youtube videos
Using mouse cursor movements to highlight the ROIs



Description

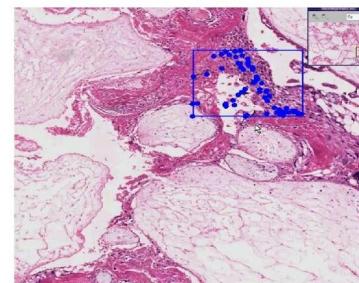
look at the lobule, they're usually made up of small rounded glandular formations. And classically, the benign lobule has got two layers. There is an outer myoepithelial layer you can see here, one row of outer nuclei and then one inner row of nuclei. These are the luminal cells and of course that's the lumen. So this bilayered appearance is indicative of benign breast parenchyma and you can see this nice lobular architecture, [0.31, 0.27, 0.46, 0.71]

Detailed Description

User: In a few words, what does the histopathology image depict?

Assistant: In the central to lower right portion of the histopathology image, there is a clear depiction of a lobule, a characteristic structure in benign breast parenchyma. This lobule is composed of small, rounded glandular formations, demonstrating a classic bilayered appearance. The outer layer of the lobule is made up of myoepithelial cells, which can be identified by a single row of nuclei. Adjacent to this, there is an inner layer composed of luminal cells, also identifiable by a single row of nuclei. The space between these luminal cells forms the lumen. This bilayered structure, with its distinct outer myoepithelial layer and inner luminal cell layer surrounding the lumen, is a hallmark of benign breast tissue. The lobular architecture is well-preserved and clearly visible in this region of the image, further supporting the benign nature of the tissue.

Image



Description

to take this trophoblast out here and out here. And you say, how much does it invade? So if you were to find trophoblasts like this deep within the myometrium, even if it looked as nice as it does over here, you can probably call it a choriocarcinoma. If, on the other hand, it's entirely limited to tissue that is not within the myometrium and is strictly in the placenta, so you take out the placenta. [0.5, 0.15, 0.81, 0.39]

Conversation

User: Can you describe the main features visible in this histopathology image?

Assistant: The image primarily shows the presence of trophoblasts, which are cells forming the outer layer of a blastocyst. These cells are seen in two distinct areas. There's also a portion of the image on the upper right that appears to represent placental tissue.

User: Are the trophoblasts invading the myometrium or are they limited to the placental tissue?

Assistant: Based on the spatial layout, it seems that the trophoblasts are primarily confined to the placental tissue, which is located from the middle to the upper right part of the image. There's no clear indication of deep invasion into the myometrium in the visible field.

User: What could be the potential implications of these findings?

Assistant: If trophoblasts are found deep within the myometrium, it could indicate a condition like choriocarcinoma, a rare and aggressive form of cancer. However, in this case, the trophoblasts appear to be limited to the placental tissue. This could suggest a condition like placenta accreta, where the placenta attaches too deeply in the uterine wall but does not penetrate it. However, a definitive diagnosis would require further investigation and clinical correlation.

VII. Multimodal LLMs in Pathology

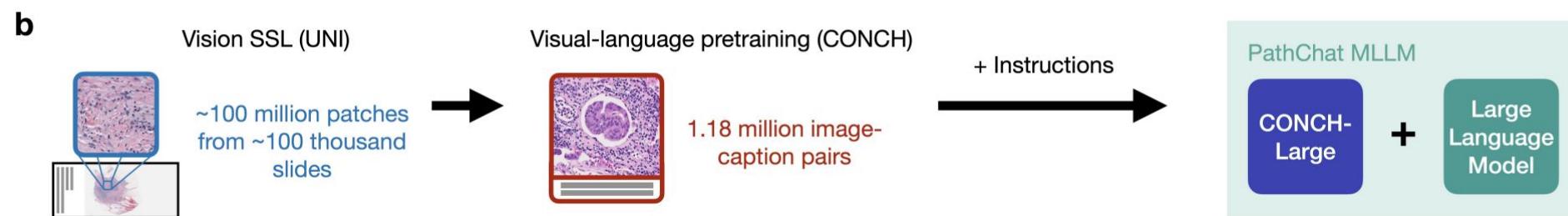
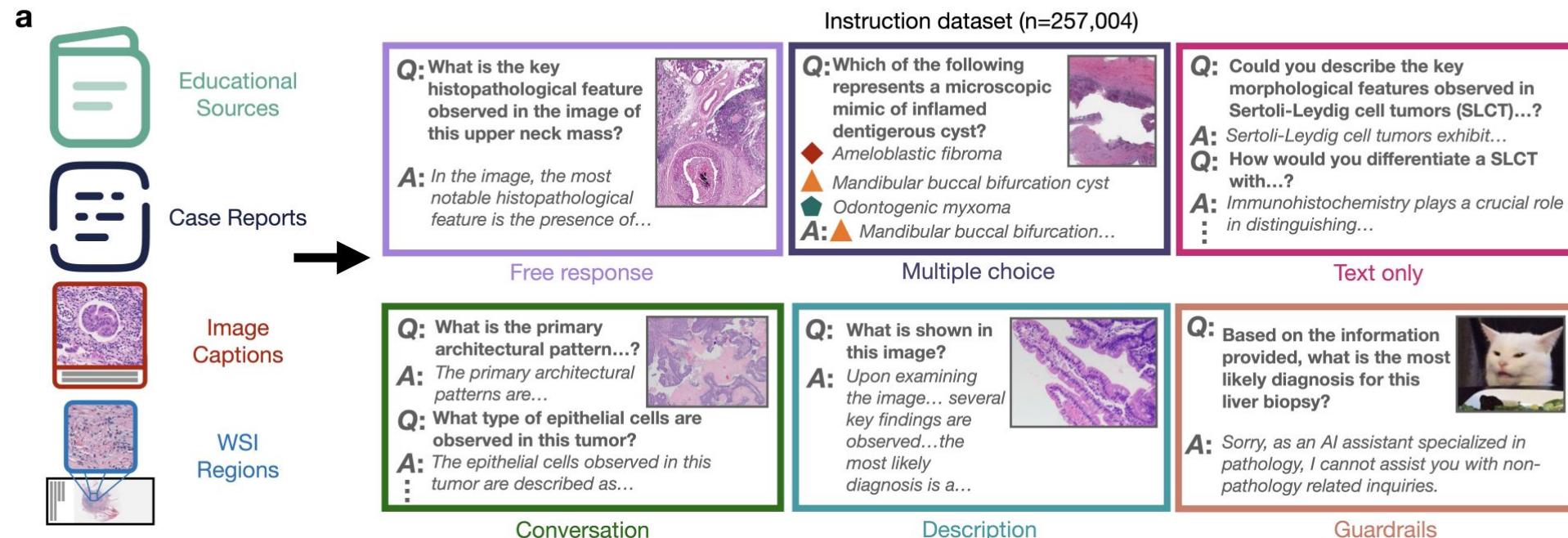
Tested on VQAs including their own benchmark, Quilt-VQA (generated using videos and GPT-4)

QUILT-LLAVA Model Variants			PathVQA		PMC-VQA-Subset		QUILT-VQA		QUILT-VQA ○		Average	
Instruct	Stage 1	Stage 2	Open	Closed	Closed	Open	Closed	Open (w/o ● w/)	Closed (w/o ● w/)	Open	Closed	
<i>QUILTNET ViT-B-32 Vision Encoder [9], 7B Language Model</i>												
107K	1	0	14.34	53.78	27.05	47.69	56.56	49.62 ● 54.13	55.56 ● 58.33	41.45	50.26	
107K	1	1	14.24	58.42	19.63	59.82	64.43	58.81 ● 61.08	68.52 ● 70.37	48.49	56.27	
107K	1	3	12.79	56.30	17.21	57.62	63.55	56.21 ● 58.32	65.74 ● 69.44	46.24	54.45	
107K	3	1	15.30	54.93	16.01	60.97	60.64	59.24 ● 64.06	56.48 ● 59.26	49.89	49.46	
<i>LLAVA [16] checkpoint, 7B Language Model</i>												
107K	1	0	11.65	54.03	33.91	55.80	58.02	54.77 ● 59.56	51.85 ● 60.19	45.45	51.60	
107K	1	1	15.06	58.68	28.56	55.39	68.81	54.24 ● 59.83	71.30 ● 75.00	46.13	60.47	
<i>Baselines</i>												
LLaVA-Med [14] 7B	0	0	11.97	56.15	1.34	54.81	61.22	52.58 ● 53.97	69.44 ● 64.81	43.33	50.59	
LLaVA [16] 7B	0	0	11.65	54.02	33.91	55.81	57.73	54.74 ● 59.96	51.85 ● 60.19	45.54	51.54	

VII. Multimodal LLMs in Pathology

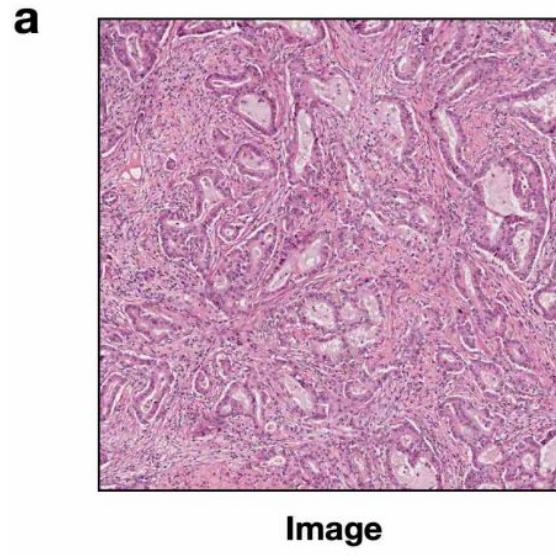
PathChat : Diverse tasks covered with more data (257K) than PathAsst.

Using CONCH dataset + instructions to train Llama-2 13B backbone LLM.



VII. Multimodal LLMs in Pathology

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



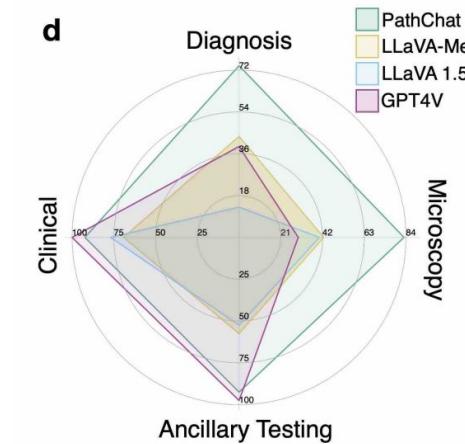
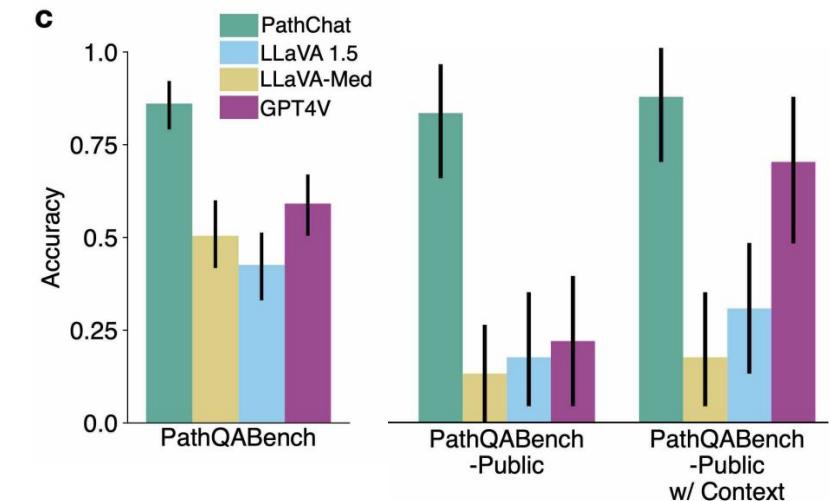
A 63-year-old male presents with chronic cough and unintentional weight loss over the past 5 months. Chest X-ray shows a dense, spiculated 3 cm mass.

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

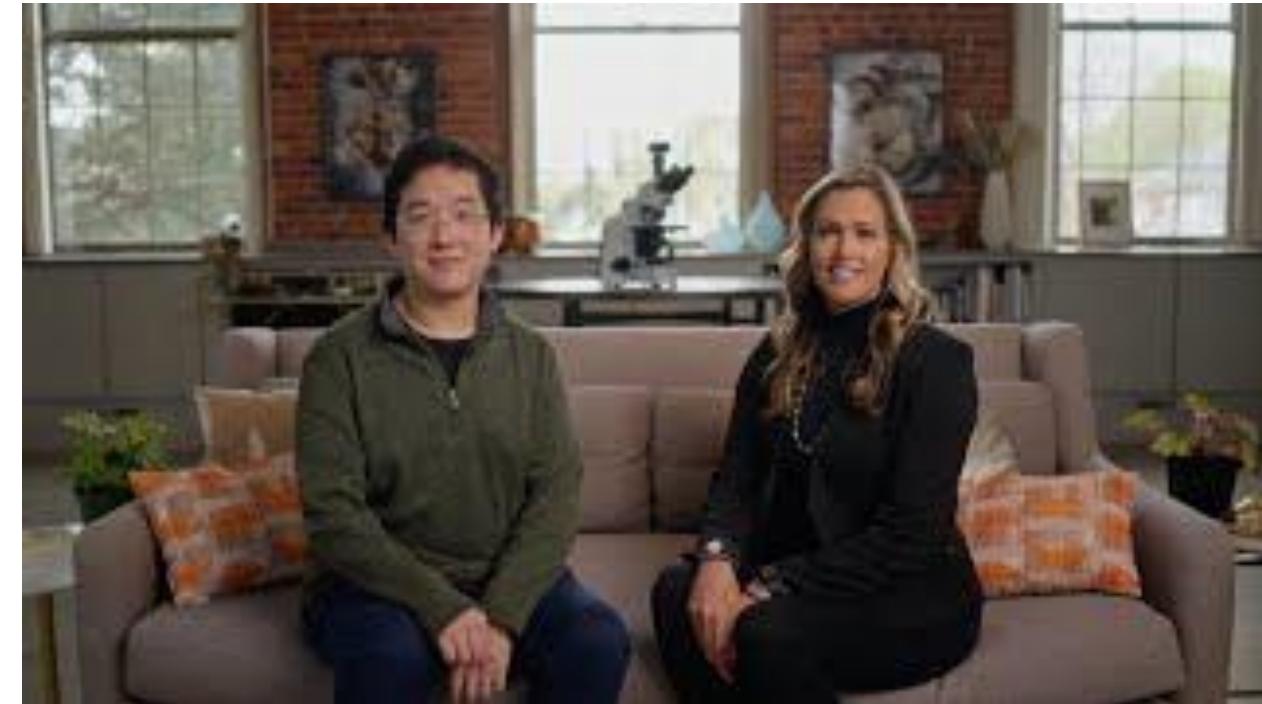
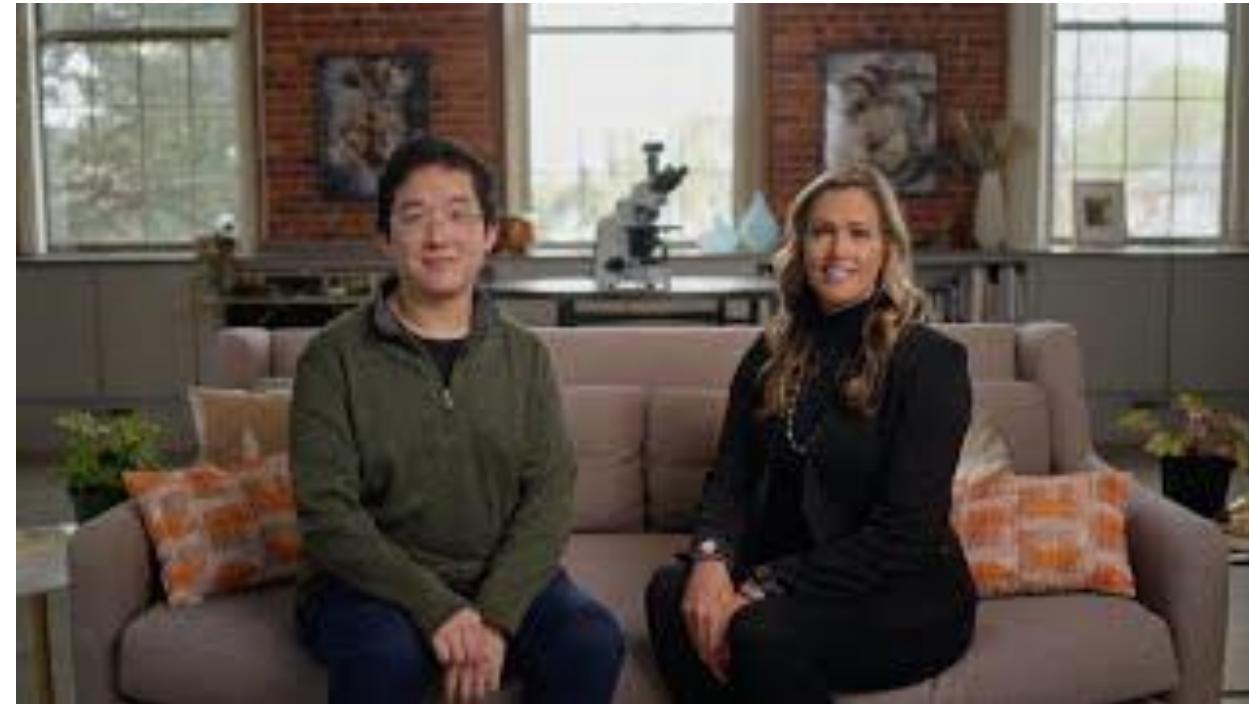
Answer with the option's letter from the given choices directly.

Context
Prompt



VII. Multimodal LLMs in Pathology

PathChat : Chat Assistant capable of summarizing the WSI or tiles.



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

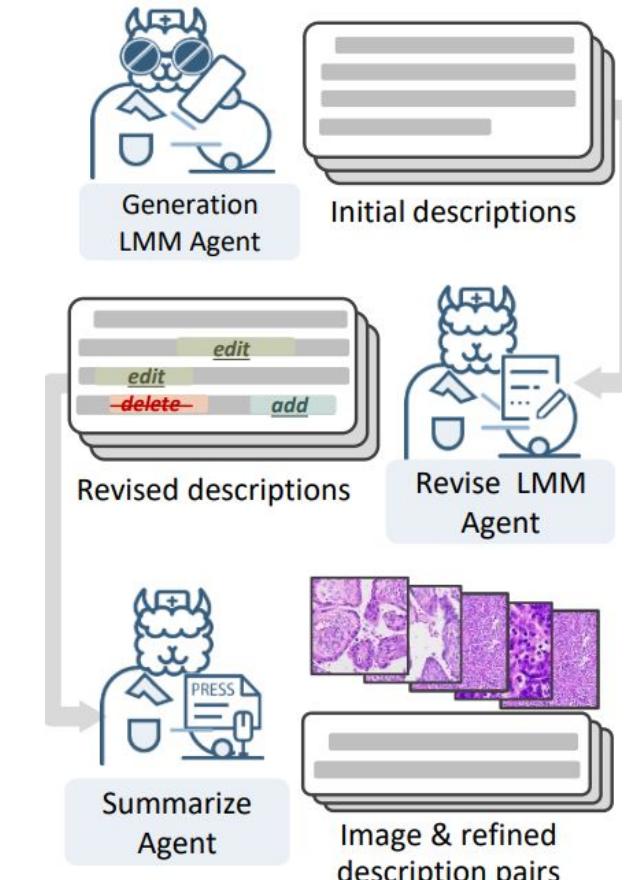
Replaced LLaVA 1.5's CLIP with PathGenCLIP which was trained with 1.6M Pathology Image-text Pairs (patch level) generated using GPT-4V

1st stage training - Feature Alignment using 1.6M

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

- 95K pathology MCQs
- 105K pathology conversation

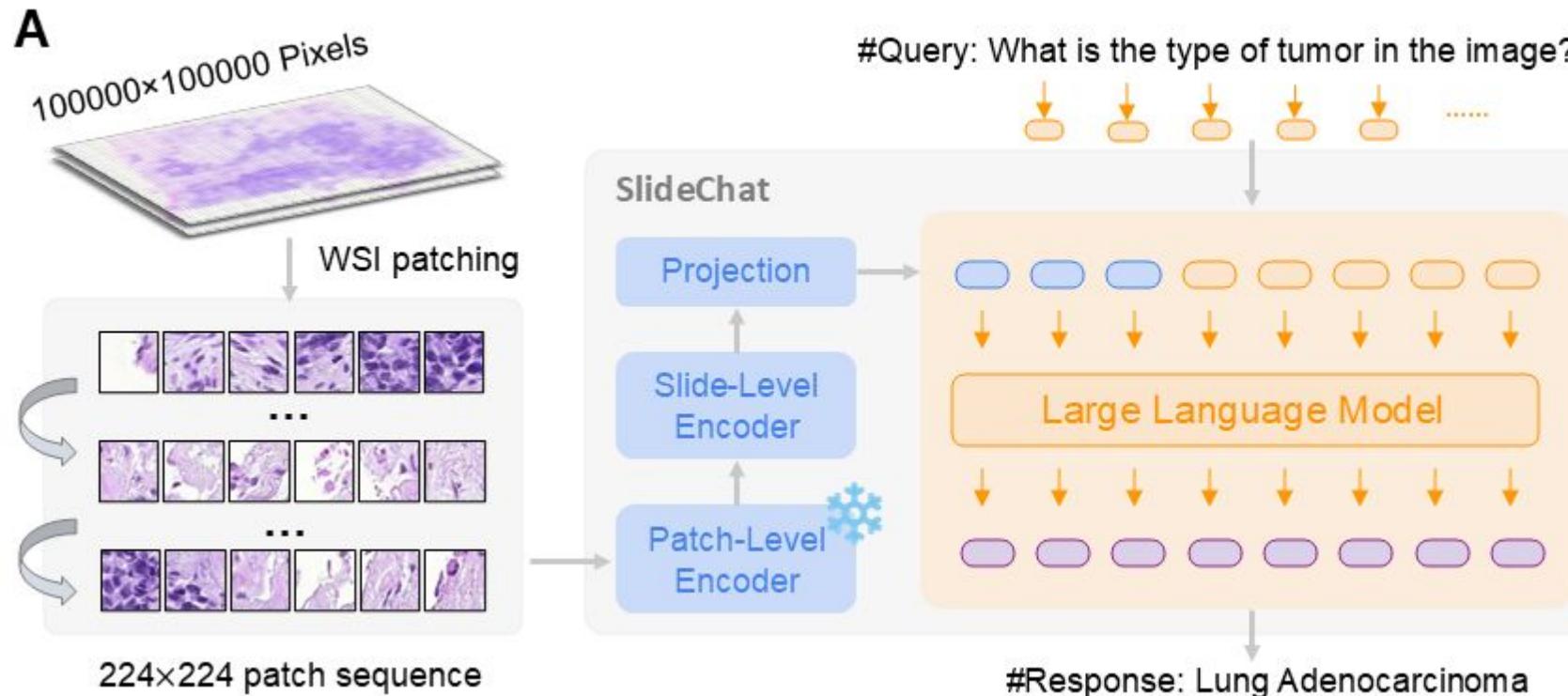
	Test Overall		PubMed		SocialPath		EduContent		Atlas		PathCLS	
	Tiny (1156)	All (9677)	Tiny (281)	All (3068)	Tiny (235)	All (1855)	Tiny (255)	All (1938)	Tiny (208)	All (1007)	Tiny (177)	All (1809)
Random Choice	22.1	23.7	22.1	25.1	25.5	26.5	25.5	26.0	19.7	23.0	15.3	16.3
Frequent Choice	27.7	25.5	28.8	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
Pathology-specific Large Multimodal Models												
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	60.1	58.4	60.1	60.1	60.9	58.8	60.8	60.7	63.5	64.9	54.2	48.9



VII. Multimodal LLMs in Pathology

SlideChat extended to WSI level

SlideChat uses Qwen2.5-7B and Tile level as well as WSI encoder



B Stage1: Domain Alignment

4.2K WSI-caption pairs

LLM Projection

Slide-level Encoder

Stage2: Instruction Tuning

176K WSI VQAs

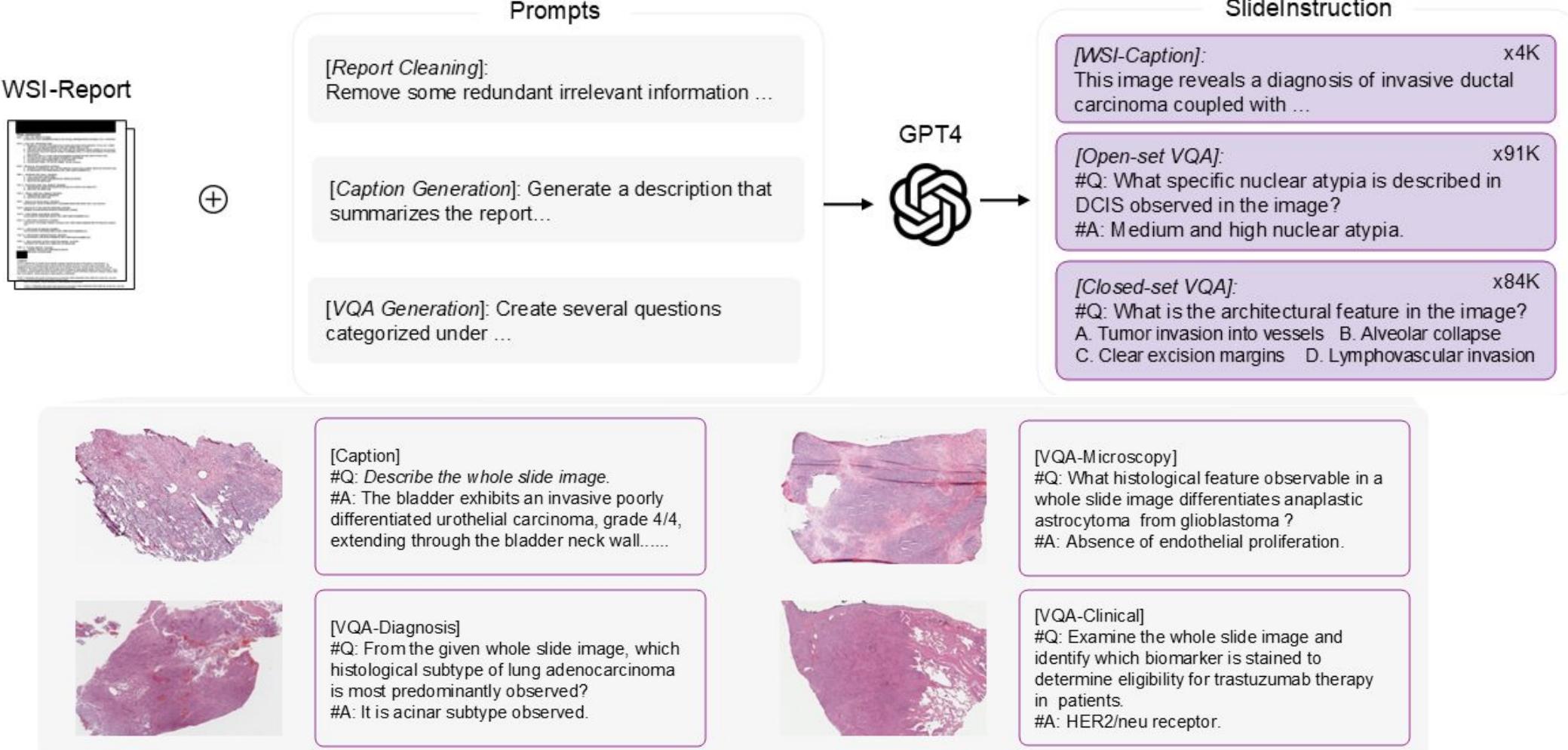
LLM Projection

Slide-level Encoder

VII. Multimodal LLMs in Pathology

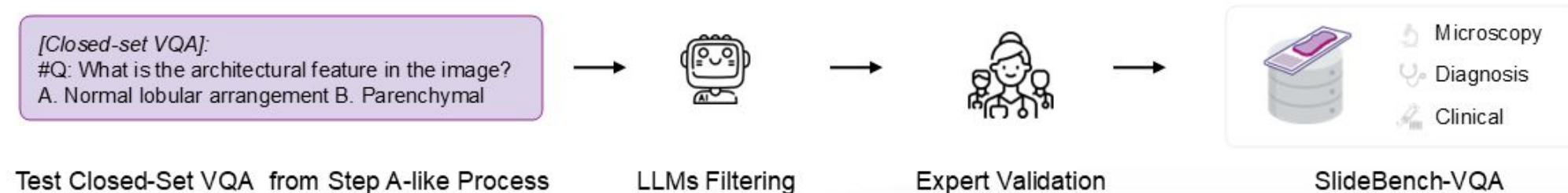
SlideChat used GPT4 to generate slide level Instruction Tuning dataset. 4.2K WSI-Report pairs.

A

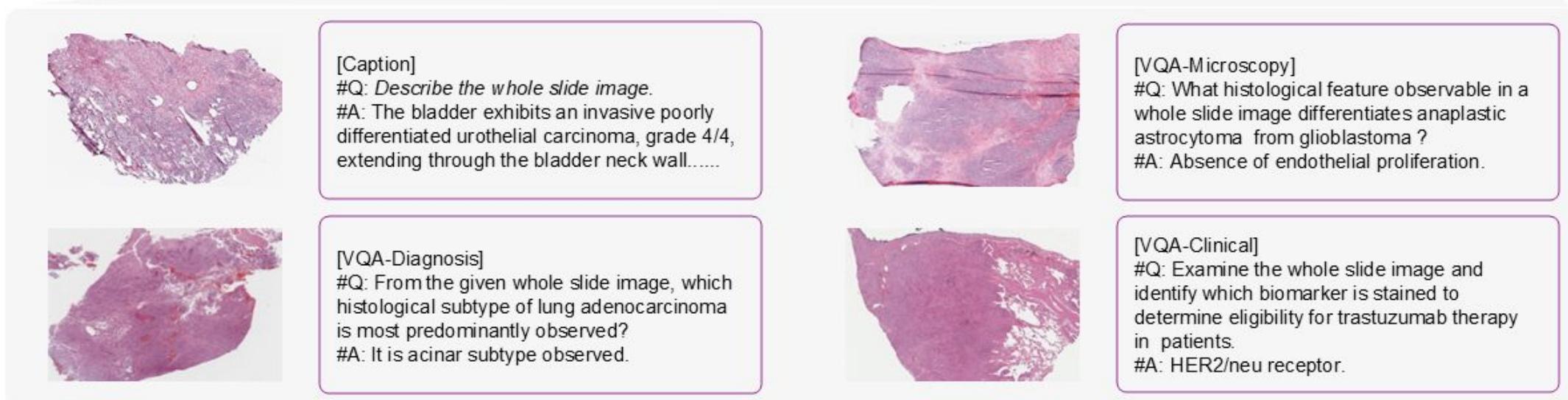


VII. Multimodal LLMs in Pathology

SlideBench – 734 WSI captions and closed-set VQAs. 7.8K TCGA and 7.2K (Early Breast Cancer Core-Needle Biopsy).



C



[Caption]
#Q: *Describe the whole slide image.*
#A: The bladder exhibits an invasive poorly differentiated urothelial carcinoma, grade 4/4, extending through the bladder neck wall.....

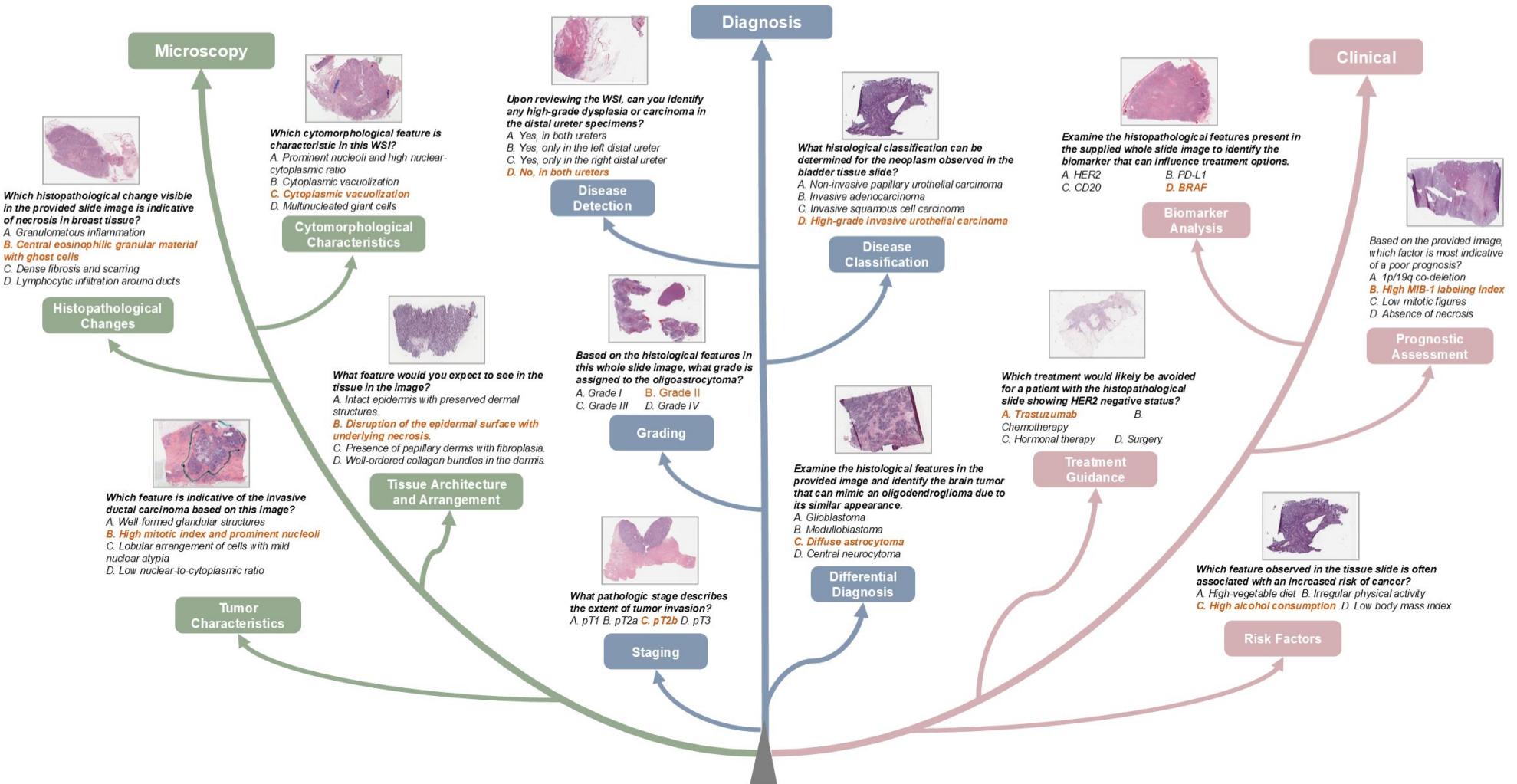
[VQA-Diagnosis]
#Q: From the given whole slide image, which histological subtype of lung adenocarcinoma is most predominantly observed?
#A: It is acinar subtype observed.

[VQA-Microscopy]
#Q: What histological feature observable in a whole slide image differentiates anaplastic astrocytoma from glioblastoma ?
#A: Absence of endothelial proliferation.

[VQA-Clinical]
#Q: Examine the whole slide image and identify which biomarker is stained to determine eligibility for trastuzumab therapy in patients.
#A: HER2/neu receptor.

VII. Multimodal LLMs in Pathology

SlideChat covers various tasks in VQA.



SlideChat outperforms GPT-4 and GPT-4o.

Methods	Input	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Rouge-L	GPT score
GPT-4o	Patch	0.16	0.03	0.01	0.01	0.13	1.54
GPT-4o	Slide (T)	0.10	0.03	0.01	0.01	0.11	1.03
Quilt-LLaVA	Patch	0.30	0.19	0.11	0.05	0.18	3.87
Quilt-LLaVA	Slide (T)	0.23	0.09	0.04	0.01	0.16	1.89
MI-Gen	Slide	0.37	0.24	0.15	0.10	0.25	4.14
SlideChat	Slide	0.37	0.21	0.12	0.08	0.24	4.11

SlideChat outperforms GPT-4 and GPT-4o.

Methods	Input	SlideBench-VQA (TCGA)			
		Microscopy	Diagnosis	Clinical	Overall
Random	Text	24.44	24.91	26.44	25.02
GPT-4		38.28	29.09	45.00	37.25
GPT-4o	Patch	62.89	46.69	66.77	57.91
MedDr		73.30	57.78	74.25	67.70
LLaVA-Med	Slide (T)	47.34	32.78	47.96	42.00
Quilt-LLaVA		57.76	35.96	53.07	48.07
GPT-4o	Slide	38.28	23.10	43.42	34.07
MedDr		70.48	52.47	72.80	64.25
LLaVA-Med	Slide	45.82	27.58	40.84	37.39
Quilt-LLaVA		49.12	26.97	44.75	39.39
SlideChat	Slide	87.64 (+14.34)	73.27 (+15.49)	84.26 (+10.01)	81.17 (+13.47)

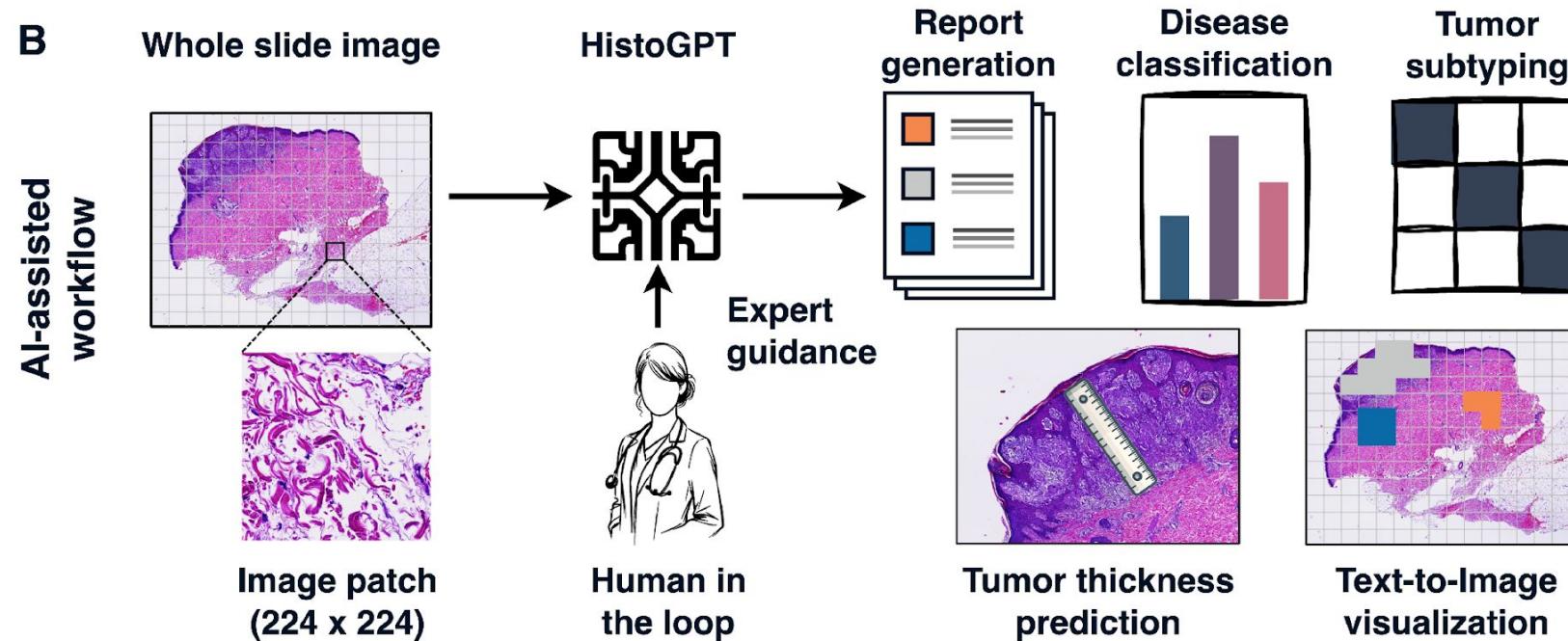
VII. Multimodal LLMs in Pathology

HistGen – Great source of publicly available report generation dataset. Used GPT-4 to clean TCGA reports.

Whole Slide Image	HistGen Generated Results	Ground Truth
	<p>a. renal cell carcinoma conventional clear cell type <u>fuhrman nuclear grade 3 of 4</u>. b. the neoplasm is confined within the renal capsule c. no invasion of the renal vein is identified . d . no evidence of angiolympathic invasion is identified. e. all surgical margins are free of the neoplasm. f. non-neoplastic kidney with no evidence of malignancy . g. the adrenal gland is not involved. no evidence of malignancy</p>	<p>a. renal cell carcinoma conventional clear cell type. b. <u>fuhrman nuclear grade is 2 of 4</u> . c . <u>the greatest diameter of the neoplasm is 45 cm</u>. d . the neoplasm is confined within the renal capsule. e. no angiolympathic invasion is identified. f. all final assessed surgical resection margins are free of neoplasm. g. pathologic tnm stage pt1b nx mx</p>
	<p>Specimen type gastrectomy, with tumor site stomach. Tumor features ulcerated histologic type adenocarcinoma histologic grade poorly differentiated tumor extent adjacent structures specify - <u>lesser omentum</u>, lymph nodes positive for metastasis intraabdominal. lymphatic invasion not specified. venous invasion not specified. perineural invasion not specified. margins uninvolved. evidence of neo-adjuvant treatment not specified. additional pathologic findings not specified</p>	<p>Specimen type gastrectomy. Tumor site stomach. tumor features ulcerated. histologic type adenocarcinoma. histologic grade poorly differentiated. tumor extent adjacent structures specify - <u>lesser and greater omentum</u>, lymph nodes positive for metastasis intraabdominal. lymphatic invasion not specified . venous invasion not specified . perineural invasion not specified . margins not specified. evidence of neoadjuvant treatment not specified. additional pathologic findings not specified .</p>
	<p>diagnosis a left kidney renal cell carcinoma conventional clear cell type fuhrman nuclear grade 4. tumor confined to the kidney. vascular ureteral and soft tissue margins of resection free of tumor. adrenal gland no tumor present. gross description a left kidney and adrenal gland - a nephrectomy specimen including the renal sinus adipose tissue. the renal vein is present. <u>tumor invading blood vessel in renal hilum</u>.</p>	<p>diagnosis a left kidney renal cell carcinoma conventional clear cell type fuhrman nuclear grade 4. tumor extending focally into perirenal adipose tissue. tumor invading blood vessel in renal hilum. vascular ureteral and soft tissue margins of resection free of tumor. adrenal gland no tumor present. comment grossly no renal vein invasion was identified in one of the sections from the tumor in relation to the renal hilar region parts a7 and a8. <u>there is tumor invading the lumen and partially the wall of a thick-walled vessel</u>.</p>

VII. Multimodal LLMs in Pathology

HistoGPT – Patch Level. Training data not released.



HistoGPT – extended to thickness, subtypes prediction. Also, overlays attention heatmap on WSI.

C

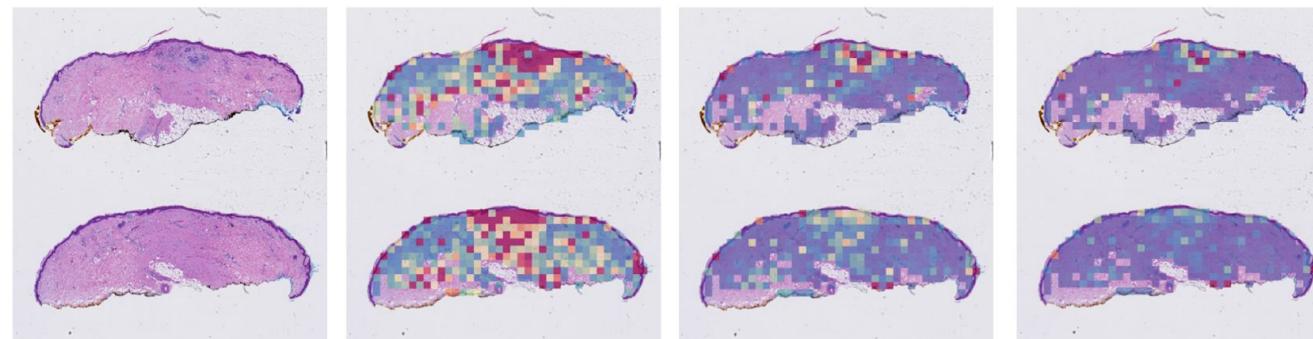
*This is clearly a **basal cell carcinoma**!*

Let me inform HistoGPT about the diagnosis and ask the model to write me a first draft for a report.

Example report
and visualization



Final diagnosis: **Basal cell carcinoma**. Critical findings: A superficial basal cell carcinoma is present, with a **tumor thickness of 0.7 mm**, in association with a cell-rich scar. The cut edges are clear. Microscopic findings: A wide punch biopsy specimen is provided. The epidermis is atrophically flattened, with a predominantly basket-weave stratum corneum. From the epidermis, there is a proliferation of basaloid tumor cell clusters into the upper dermis. Characteristic palisades are positioned in the peripheral area, with contraction artifacts and peritumoral stroma induction. A dense, plasma-rich **lymphocytic inflammatory infiltrate** is observed peritumorally.



Input whole slide image
(serial section)

Gradient-attention for
basal cell carcinoma

Gradient-attention for
tumor thickness

Gradient-attention for
lym. inflam. infilt.

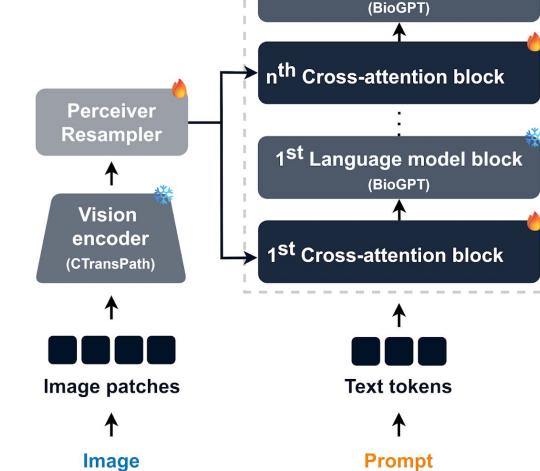
VII. Multimodal LLMs in Pathology

HistoGPT – Image encoder is CTransPath. BioGPT as language model.

A Model architecture

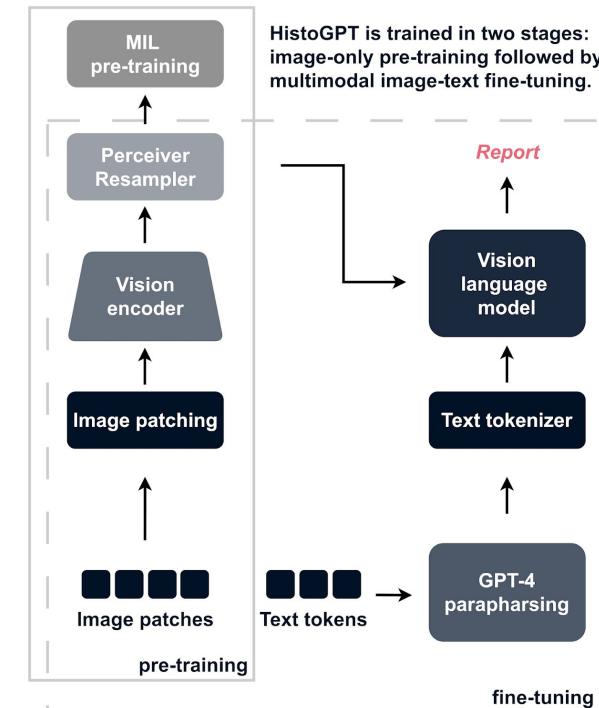
Microscopic description: Malignant tumor cells can be found in the upper layer of... Final diagnosis: Basal cell carcinoma.

- Pretrained and frozen
- Trained from scratch

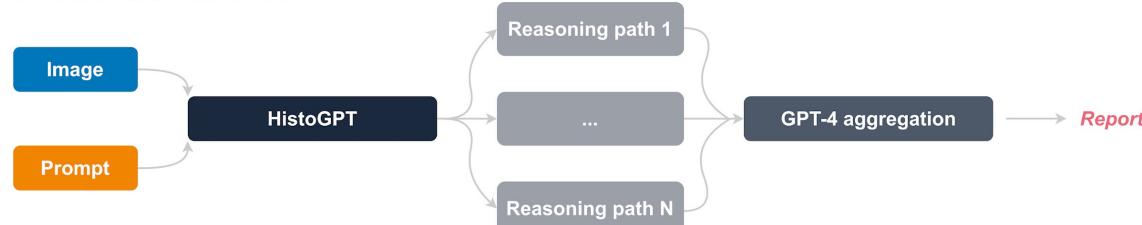


B Model training

HistoGPT is trained in two stages:
image-only pre-training followed by
multimodal image-text fine-tuning.

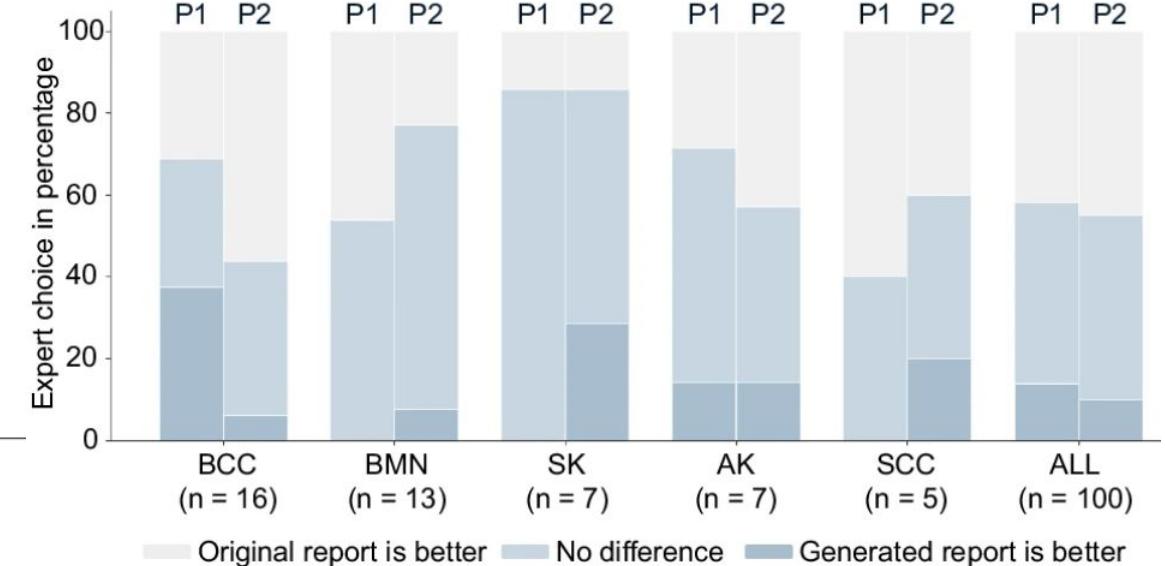
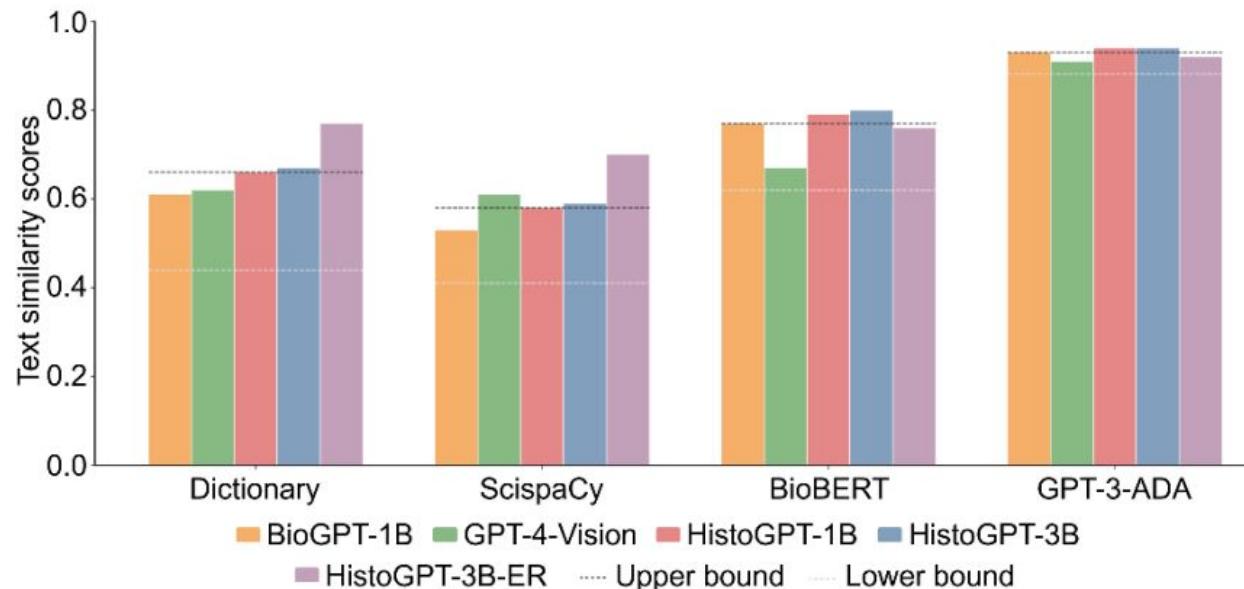


C Ensemble refinement

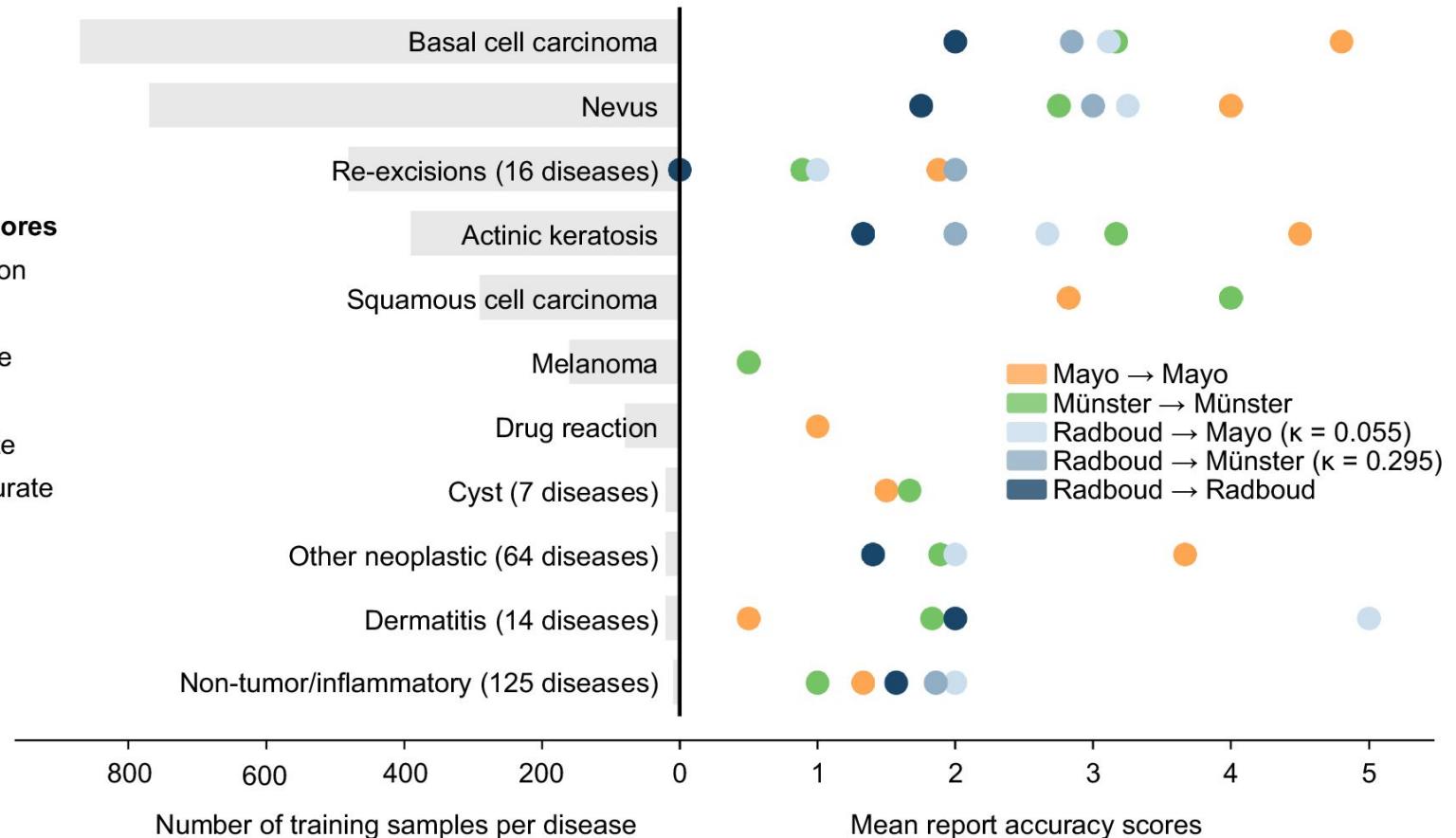
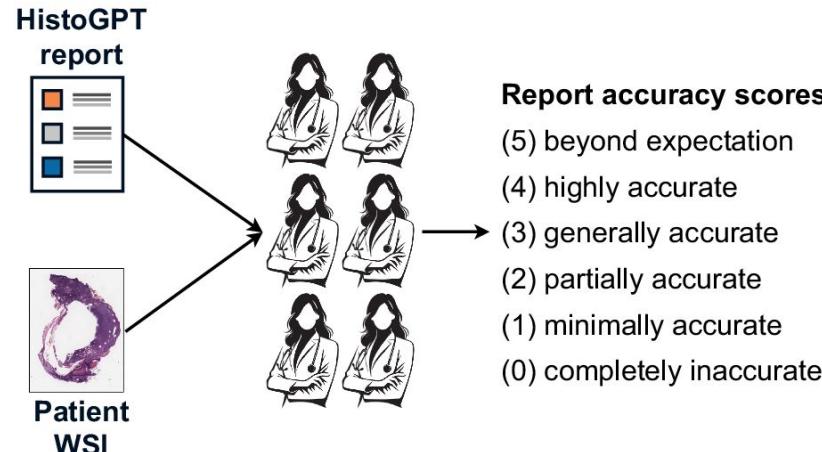


VII. Multimodal LLMs in Pathology

Report generation outperforming GPT-4-Vision.
Sometimes generated reports are better than the ground truth.

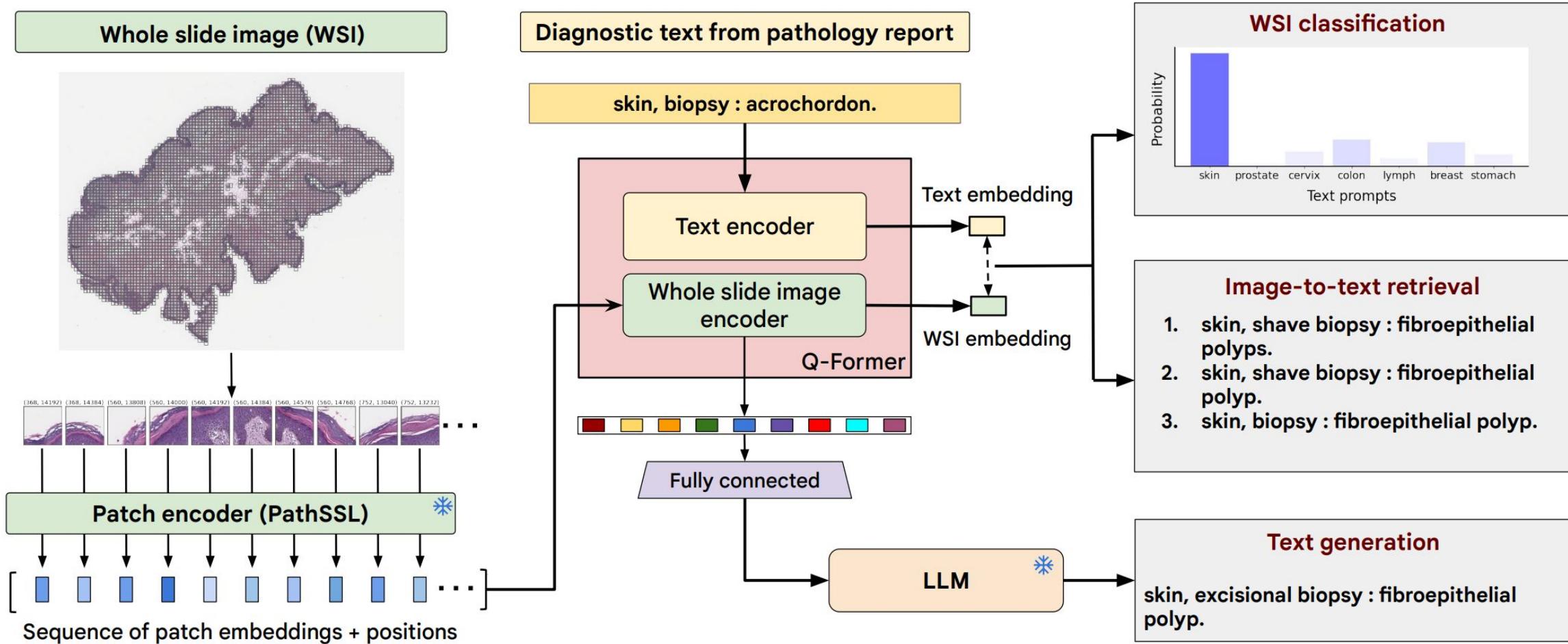


HistoGPT – Histopathologist scoring still varies.

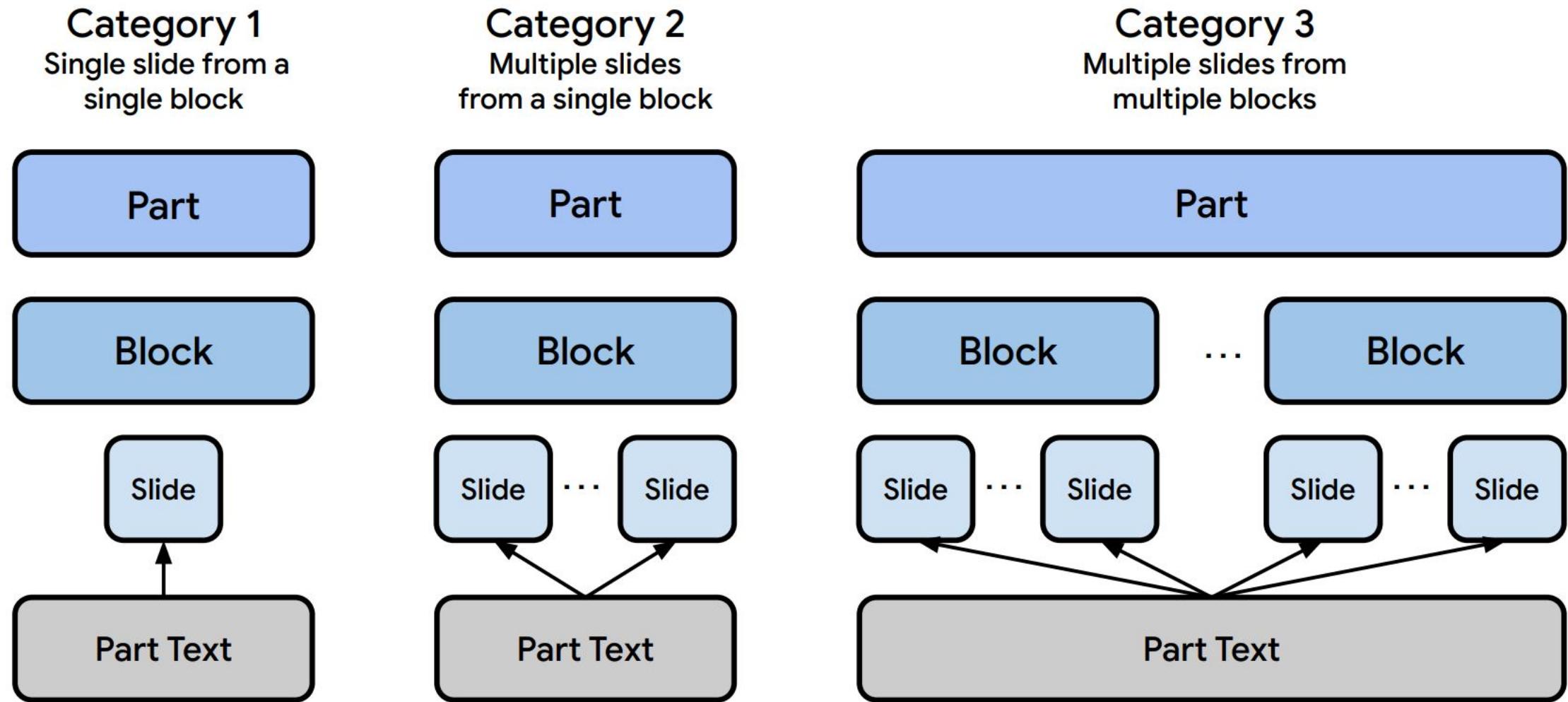


VII. Multimodal LLMs in Pathology

PathAlign – WSI encoder for report generation. BLIP-2 approach. PathAlign-R. PathAlign-G with PaLM-2.

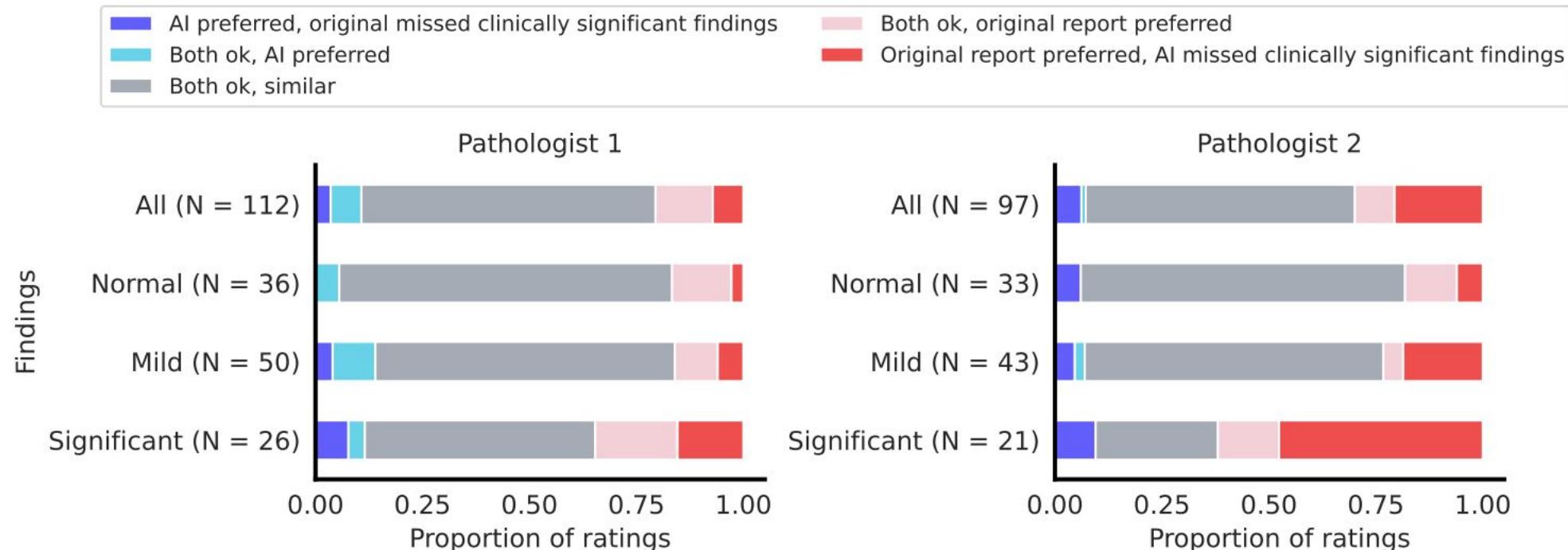


PathAlign – uses in-house WSIs with reports. Aligns text to the right WSI



VII. Multimodal LLMs in Pathology

PathAlign – generated reports largely similar to the ground truth



PathAlign – relative high lexical metrics

- Higher score doesn't necessary guarantee the correctness of the reports.

Dataset	ROUGE-L (F-Measure)	METEOR
DS1 ($N = 4,876$)	0.579	0.612

Text i	Text j	Score $_{ij}$
soft tissue, supraclavicular region, biopsy : classical hodgkin lymphoma.	cervix : biopsy: high grade squamous intraepithelial lesion (cinii).	0.619
colon, cecum, polypectomy : fragments of tubulovillous adenoma.	colon, polypectomy : tubular adenoma.	0.681
cervical polyp biopsy : low-grade squamous intraepithelial lesion (cin i).	cervix, biopsy : benign squamous mucosa, no transformation zone identified.	0.771
skin, excisional biopsy : neurofibroma.	skin, excisional biopsy : dermatofibroma.	0.953
colon, biopsy : hyperplastic polyp.	colon, biopsy : adenomatous polyp.	0.957
cervix, biopsy : high-grade squamous intraepithelial lesion.	cervix, biopsy : low-grade squamous intraepithelial lesion.	0.970
uterine cervix, biopsy : benign cervical tissue, no dysplasia identified.	cervix : biopsy: benign cervical tissue, no dysplasia identified.	0.978

Histopathology Reports Evaluation Metric

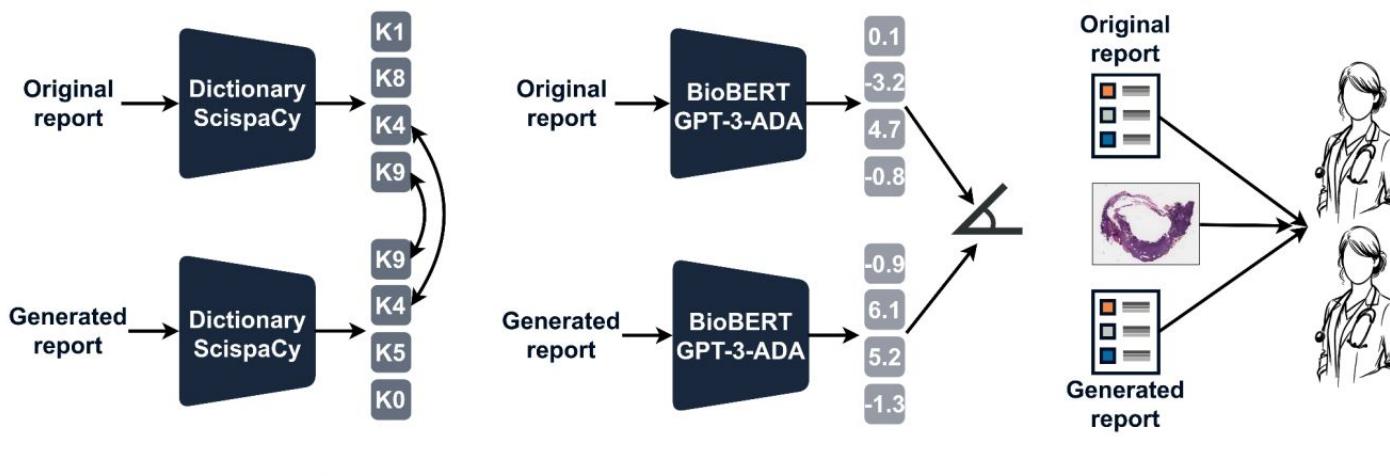
VII. Evaluation Metric for Pathology Reports - Motivation

No evaluation metrics designed for automated pathology reports.

Table 1. Comparisons of the proposed **HistGen** with other SOTA models on WSI report generation w.r.t NLG metrics.

Feature Extractor	Methods	Metric					
		BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
	Show&Tell [30]	0.249	0.099	0.047	0.025	0.086	0.165

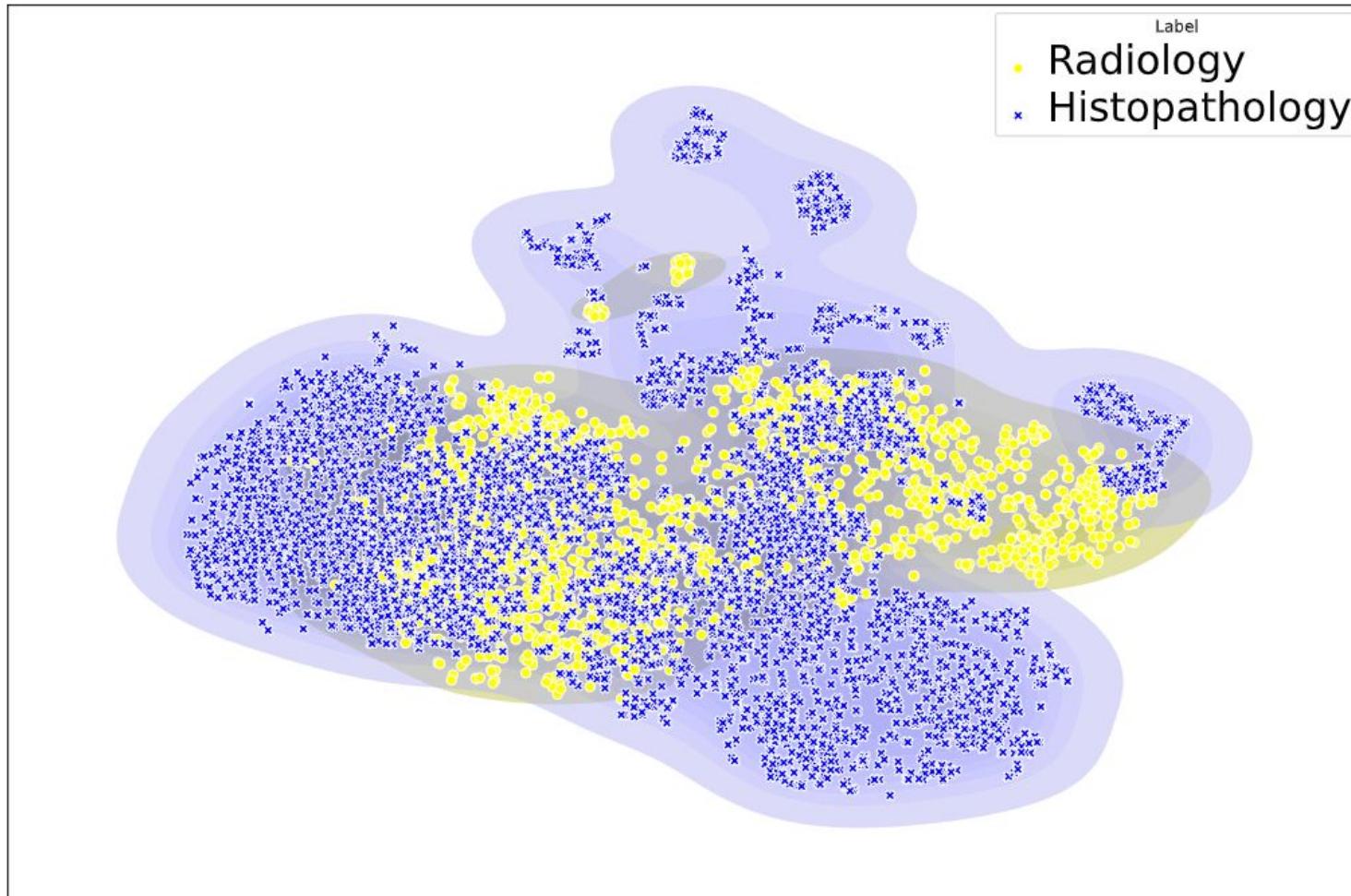
Guo, et al. 2024.



Tran, et al. 2024.

VII. Evaluation Metric for Pathology Reports - Motivation

There are medical report evaluation metrics, but mostly on Radiology.



We proposed histopathology entity and relation centric evaluation metric

Entity Recognition and Relation Extraction



HARE-NER



HARE-RE

- Pathological Diagnosis**
- Anatomical Site**
- Diagnosis Descriptor**
- IHC Modifier**
- Immunohistochemistry**

Reference

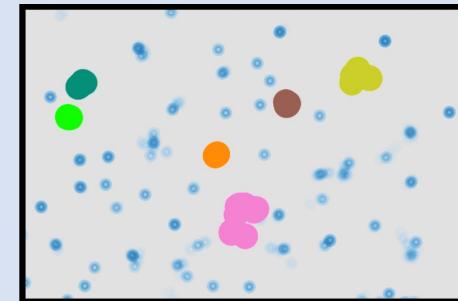
Right Breast: Grade 3 **Invasive ductal carcinoma**.
ER – **weak** positive, **GATA3** - positive, **HER2** - negative.

Candidate

Specimen Site: Right **Breast**.
Histological Findings: Presence of atypical ductal structures. Areas of necrosis.
Microscopic Description: High-grade **ductal** carcinoma in situ (DCIS) with comedo necrosis.
Final Diagnosis: **Invasive Ductal** Carcinoma, **Grade III**.

Entity Embedding and Alignment

Cosine similarity (Invasive ductal carcinoma, invasive ductal) = 0.9054



GatorTronS-UMLS



UMLS Concepts

HARE Score

NER_F1 = 0.6492
RE_F1 = 0.6667



HARE-Score

1.3159

Expert-Score

4/5: Correct diagnosis with at least one incorrect or incomplete description.

We collected and annotated 854 reports from Local Hospital and 652 reports from TCGA for training.

Pathological Diagnosis	Anatomical Site	Diagnosis Descriptor	IHC Modifier	Immunohistochemistry
Diagnosis: Left kidney, partial nephrectomy due to a 5.1cm unifocal renal cell carcinoma, chromophobe type, Fuhrman grade 3/4. No sarcomatoid features identified, and no extra-capsular invasion. Light microscope examination shows solid sheets of eosinophilic and clear cells with focal punched-out perinuclear clearing, consistent with the diagnosis of chromophobe carcinoma. Immunostains show reactivity for CK7, no reactivity for c-Kit or RCC, supporting a diagnosis of chromophobe carcinoma.				

Type	Hospital	TCGA
IHC Markers	6,945	180
IHC Modifier	1,401	173
Pathological Diagnosis	941	882
Anatomical Site	786	794
Diagnosis Descriptor	264	475
Relations	1,826	653

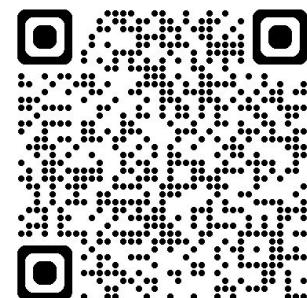
Model	NER	RE	Overall
PathologyBERT	0.771	0.798	0.785
BERT-base	0.833	0.798	0.816
BERT-large	0.825	0.798	0.811
DeBERTa-large	0.841	0.798	0.820
BiomedBERT-large	0.843	0.798	0.820
DeBERTa-xsmall	0.794	0.962	0.878
SapBERT	0.835	0.970	0.903
BiomedBERT-base	0.844	0.962	0.903
GatorTronS	0.854	0.977	0.915

VII. HARE – Alignment with Pathologist Score

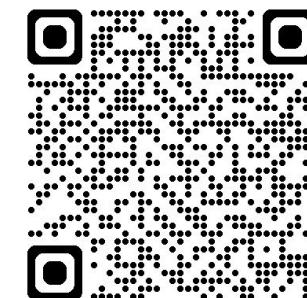
HARE has the best alignment with the scores of the 600 reports

Method	r	r p-val	ρ	ρ p-val	τ	τ p-val	R^2	RMSE
ROUGE-L	0.048	0.470	0.030	0.647	0.025	0.616	0.002	0.169
BLEU	0.077	0.241	0.106	0.108	0.099	0.107	0.006	0.168
GPT-4.1	0.177	0.007	0.173	0.008	0.146	0.008	0.031	0.166
BERTScore	0.203	0.002	0.180	0.006	0.141	0.005	0.041	0.165
METEOR	0.265	4.51e-05	0.179	0.006	0.136	0.007	0.070	0.163
RaTEScore	0.372	5.36e-09	0.350	4.81e-08	0.276	4.60e-08	0.138	0.157
RadGraph-XL	0.427	1.22e-11	0.425	1.43e-11	0.351	8.51e-11	0.182	0.153
GREEN	0.438	2.90e-12	0.482	7.58e-15	0.410	2.18e-13	0.192	0.152
HARE (Ours)	0.606	1.39e-24	0.643	2.62e-28	0.533	1.51e-24	0.368	0.134

Paper



Code and model

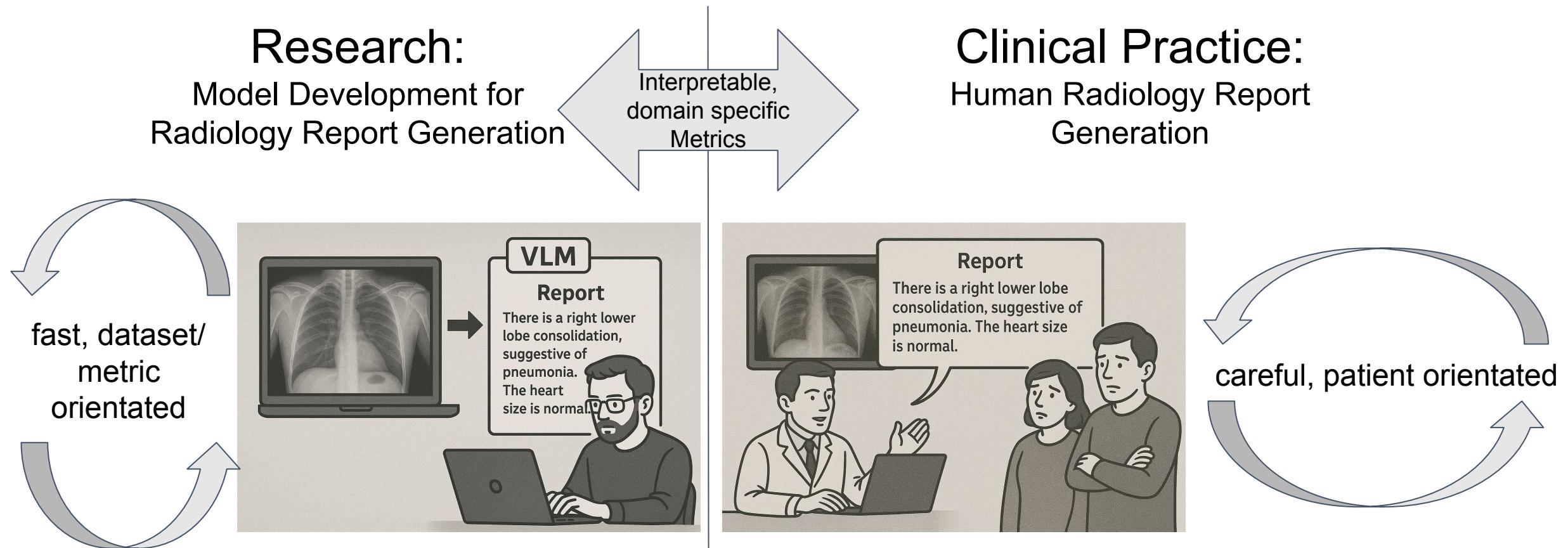


Thank you

Overview

1. Report evaluation problem setting
2. Evaluation metrics (Bleu, ROUGE, BertScore, F1RadGraph, GREEN, RadFact, RaTEScore)
3. Practical Demonstration: How to use LLMs to evaluate generated reports as well as error detection

1. Report evaluation problem setting



1. Report evaluation problem setting

Reference : “Pleural effusion present.”

Candidate 1:
pleural effusion **is** present.

Candidate 2:
pleural effusion **not** present.

2. Evaluation metrics: Human

Reference : “Pleural effusion present.”

Score between 0 and 1?

Human

Candidate 1: pleural effusion **is** present. Candidate 2: pleural effusion **not** present.

1.0

0.0

Explanation:
“is present and
present are the
same”

Explanation:
“pleural effusion is
mentioned in the
reference report”

2. Evaluation metrics: Bleu-2

Reference : “Pleural effusion present.”

BLEU-2 Evaluation

Candidate 1:
pleural effusion **is** present.

Candidate 2:
pleural effusion **not** present.



=

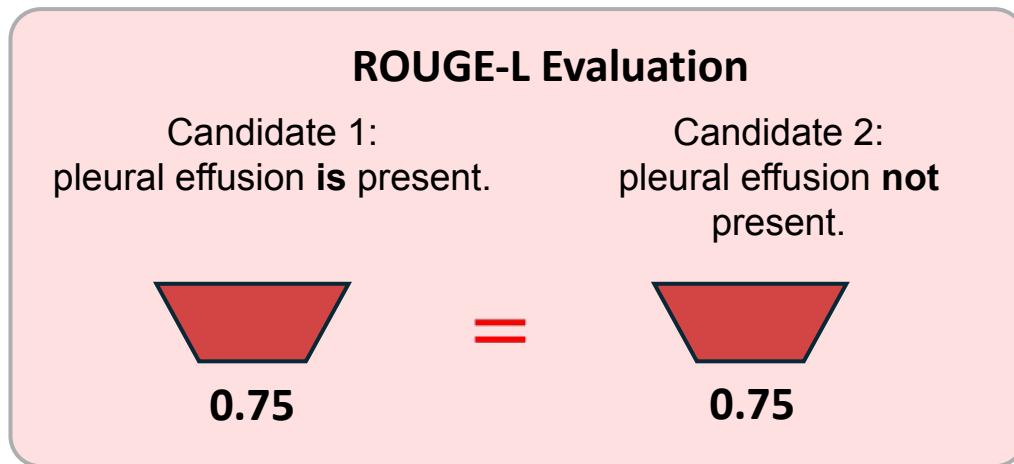


0.50

0.50

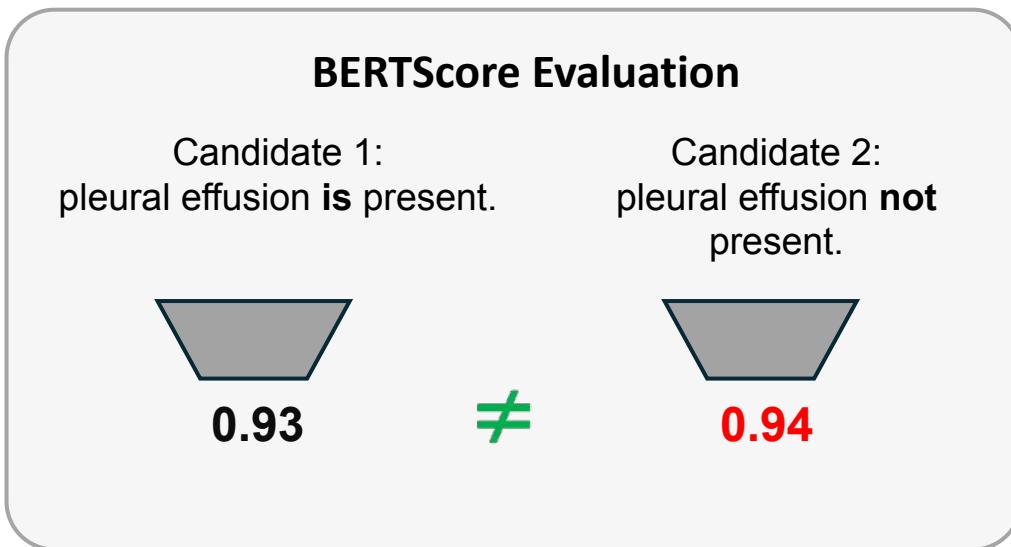
2. Evaluation metrics: ROUGE-L

Reference : “Pleural effusion present.”



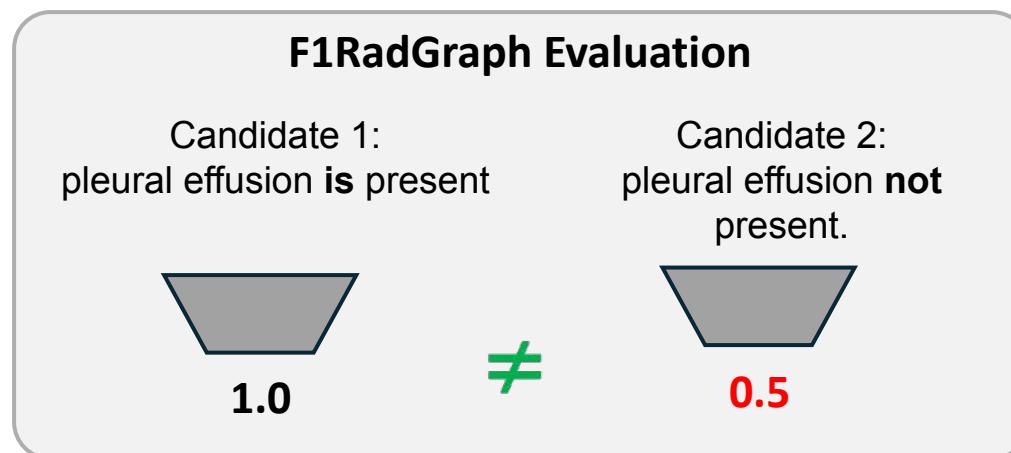
2. Evaluation metrics: BERTScore

Reference : “Pleural effusion present.”



2. Evaluation metrics: F1RadGraph

Reference : “Pleural effusion present.”



2. Evaluation metrics: CheXbertF1

Reference : “Pleural effusion present.”

CheXbertF1

Candidate 1:
pleural effusion **is** present.



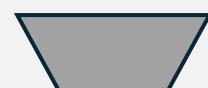
1.0

Candidate 2:
pleural effusion **not**
present.



0.0

Candidate 3:
pleural effusion **disappeared**.



1.0

2. Evaluation metrics: GREEN

Reference : “Pleural effusion present.”

Candidate 1:
pleural effusion **is** present.



1.0

≠

Our GREEN Evaluation

Candidate 2:
pleural effusion **not** present.



0.0

≠

Candidate 3:
pleural effusion **disappeared**.



0.0

Generative Explanation: Pleural effusion is marked as positive in both reference and candidate reports.

Error Notation: Clinically significant errors: 0.
Matched Findings: 1. pleural effusion is present.

Generative Explanation: Pleural effusion is marked as positive in reference but negative in candidate.

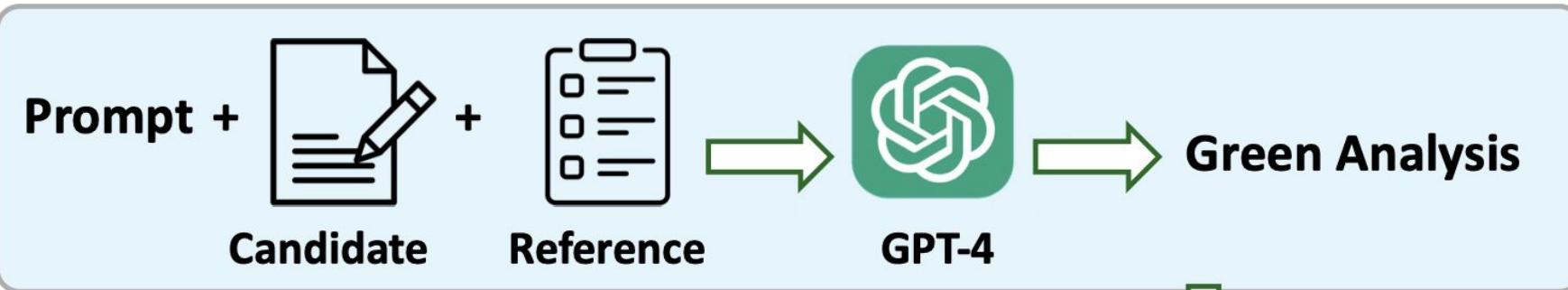
Error Notation: Clinically significant errors: 1. pleural effusion should be present.
Matched Findings: 0.

Generative Explanation: Pleural effusion is marked as positive in reference but disappeared in candidate.

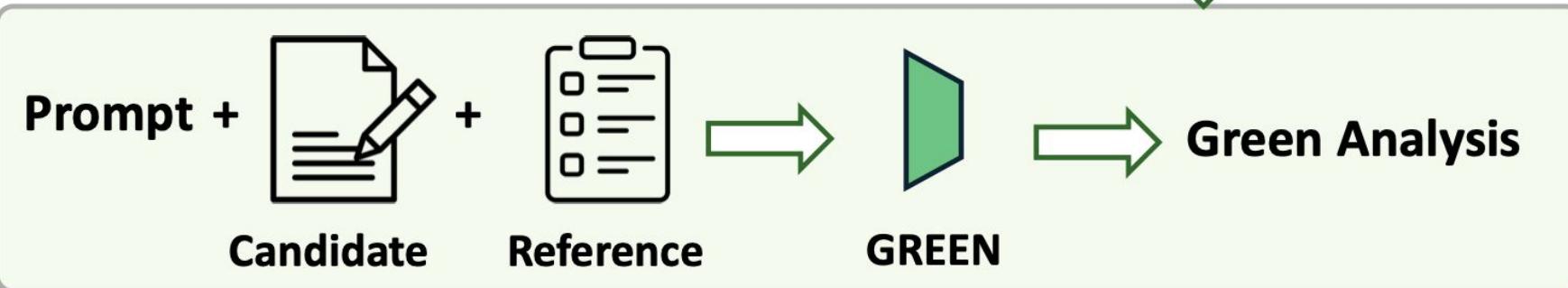
Error Notation: Clinically significant errors: 1. pleural effusion should be present. Matched Findings: 0.

2. Evaluation metrics: GREEN

Step1: Dataset Generation w/ GPT-4



Step 2: Training (Distilling the knowledge to a small LLM)



GPT-4 Prompt

2. Evaluation

Objective:

Evaluate the accuracy of a candidate radiology report in comparison to a reference radiology report composed by expert radiologists.

Process Overview:

You will be presented with:

1. The criteria for making a judgment.
2. The reference radiology report.
3. The candidate radiology report.
4. The desired format for your assessment.

1. Criteria for Judgment:

For each candidate report, determine:

- The count of clinically significant errors.
- The count of clinically insignificant errors.

Errors can fall into one of these categories:

- a) False report of a finding in the candidate.
- b) Missing a finding present in the reference.
- c) Misidentification of a finding's anatomic location/position.
- d) Misassessment of the severity of a finding.
- e) Mentioning a comparison that isn't in the reference.
- f) Omitting a comparison detailing a change from a prior study.

Note: Concentrate on the clinical findings rather than the report's writing style.
Evaluate only the findings that appear in both reports.

2. Reference Report:

****Reference Report****

3. Candidate Report:

****Candidate Report****

4. Reporting Your Assessment:

Follow this specific format for your output, even if no errors are found:
```

[Explanation]:

<Explanation>

[Clinically Significant Errors]:

(a) <Error Type>: <The number of errors>. <Error 1>; <Error 2>; ...; <Error n>

...

(f) <Error Type>: <The number of errors>. <Error 1>; <Error 2>; ...; <Error n>

[Clinically Insignificant Errors]:

(a) <Error Type>: <The number of errors>. <Error 1>; <Error 2>; ...; <Error n>

...

(f) <Error Type>: <The number of errors>. <Error 1>; <Error 2>; ...; <Error n>

[Matched Findings]:

<The number of matched findings>. <Finding 1>; <Finding 2>; ...; <Finding n>

### GPT-4 Response

## 2. Evaluation

[Explanation]:

The candidate report misidentifies the anatomic location of the infiltrates. The reference report mentions "upper middle right field" while the candidate report mentions "upper dorsal right field".

[Clinically Significant Errors]:

(c) Misidentification of a finding's anatomic location/position: 1. The infiltrates are in the upper middle right field, not the upper dorsal right field.

[Clinically Insignificant Errors]:

- (a) False report of a finding in the candidate: 0.
- (b) Missing a finding present in the reference: 0.
- (c) Misidentification of a finding's anatomic location/position: 0.
- (d) Misassessment of the severity of a finding: 0.
- (e) Mentioning a comparison that isn't in the reference: 0.
- (f) Omitting a comparison detailing a change from a prior study: 0.

[Matched Findings]:

3. Doubtful retrocardiac suggestive of respiratory infection;  
Costophrenic sinuses are clear; No other notable findings.  
```

$$\text{TOTAL ERROR} = \sum_{i=(a)}^{(f)} \# \text{ error sig.}_{,i} + \sum_{i=(a)}^{(f)} \# \text{ error insig.}_{,i} =]$$

$$\text{GREEN} = \frac{\# \text{ matched findings}}{\# \text{ matched findings} + \sum_{i=(a)}^{(f)} \# \text{ error}_{\text{sig.},i}} = \frac{3}{3+1} = 0.75$$

2. Evaluation metrics: GREEN

GREEN Summary

[Summary]:

Green score: mean 0.23 std 0.04

[Clinically Significant Errors]:

(a) False report of a finding in the candidate: 0.9

[Small right pleural effusion]

(b) Missing a finding present in the reference: 0.7

[Underlying chronic upper lobe scarring.]

(c) Misidentification of a finding's anatomic location/position: 0.4

[The opacity is in the right lower lobe, not the right upper lobe.]

(d) Misassessment of the severity of a finding: 0.8

[Bilateral pleural effusion]

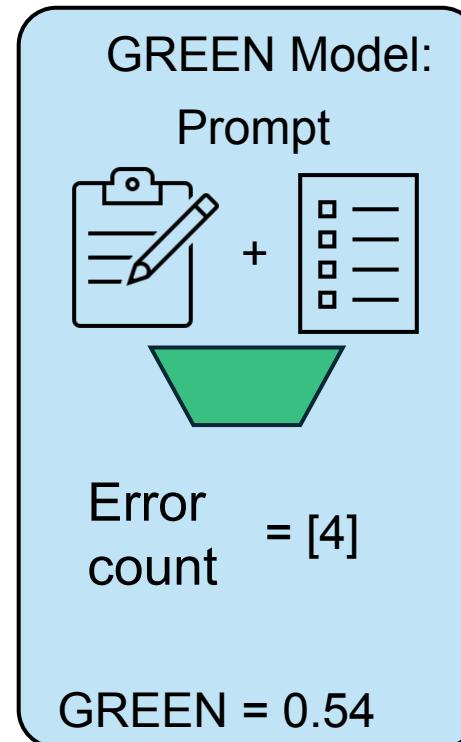
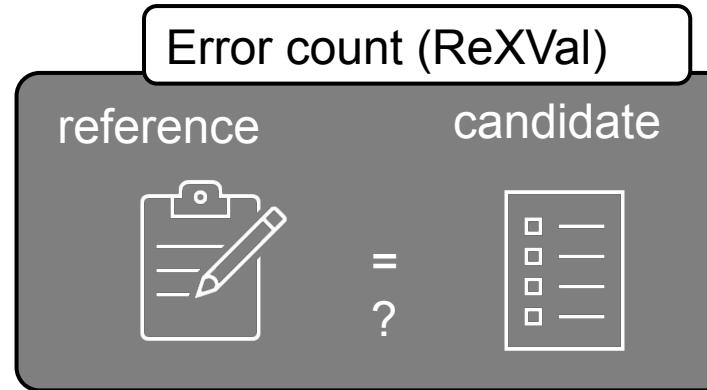
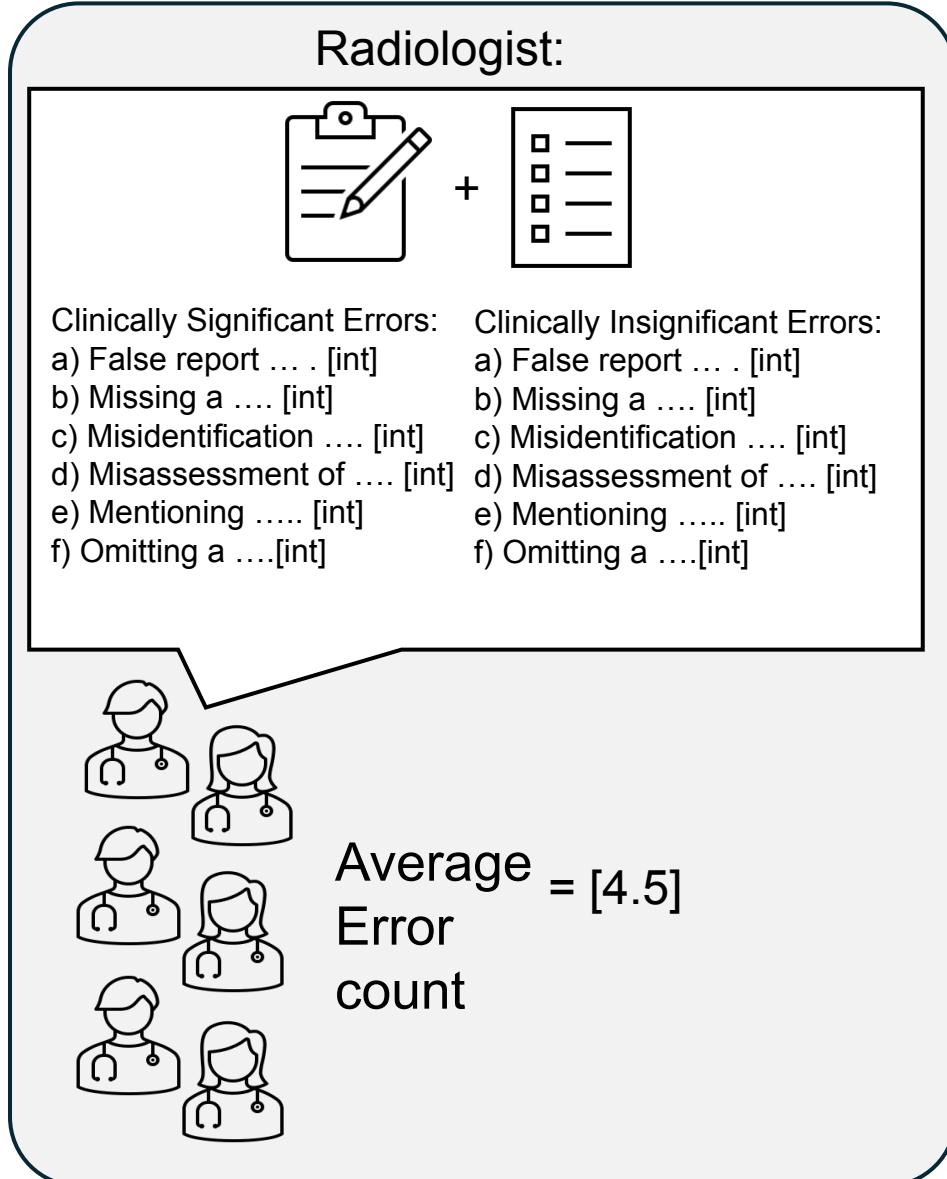
(e) Mentioning a comparison that isn't in the reference: 0.7

[The candidate report mentions a discussion between doctors, which is not present in the reference report]

(f) Omitting a comparison detailing a change from a prior study: 0.5

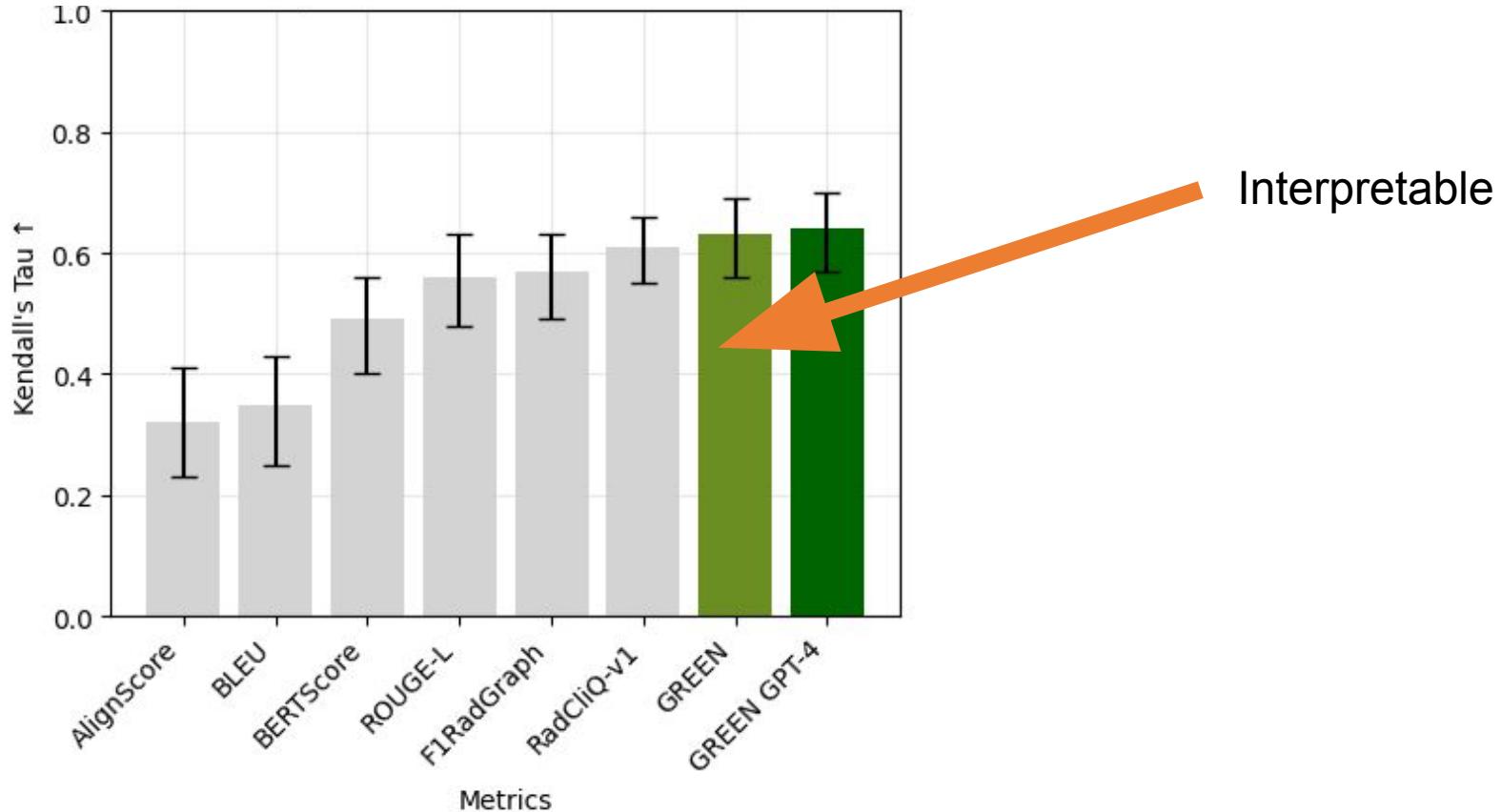
[The candidate report does not mention the absence of disease progression]

2. Evaluation metrics: GREEN

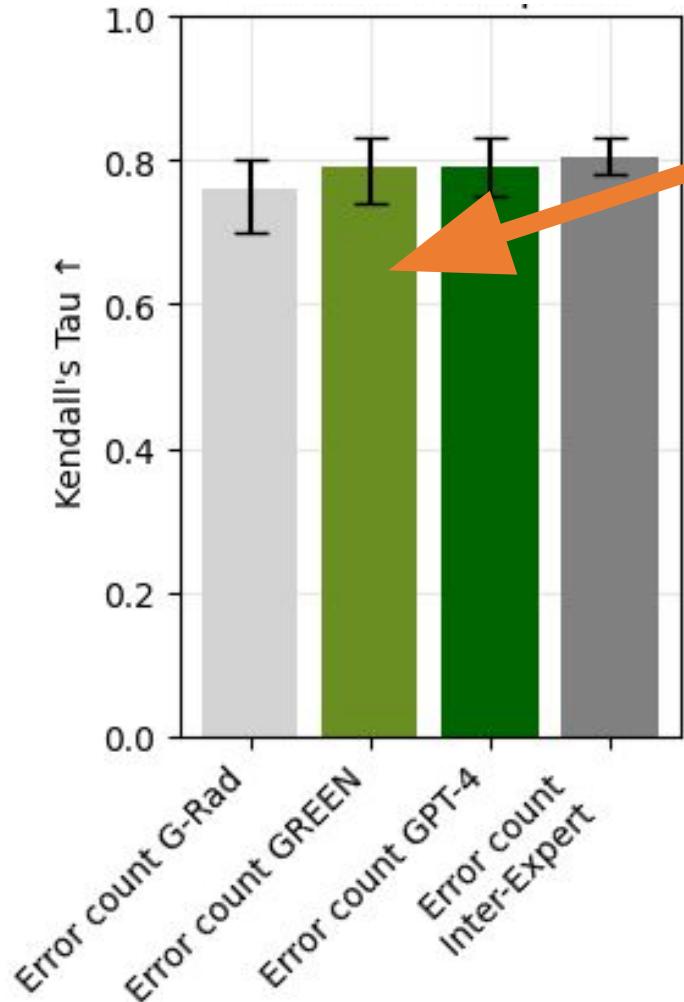


Yu, Feiyang, et al. "Radiology Report Expert Evaluation (ReXVal) Dataset." (2023).

2. Evaluation metrics: GREEN



2. Evaluation metrics: GREEN



Just 8B Parameter and
open-source

Limitations:

- currently chest xrays
- current implementation does not support vllm

2. Evaluation metrics: RadFact, RaTEScore

MAIRA-2: Grounded Radiology Report Generation

Shruthi Bannur^{*1}, Kenza Bouzid^{*1}, Daniel C. Ilse¹, Fernando Pérez-García¹, Valentina Salvat Srivastav², Julia Gong³, Fabian Falck¹, Ozan Wetscherek^{1,5}, Javier Alvarez-Valle^{*1}, and Ste



RaTEScore: A Metric for Radiology Report Generation

¹Microsoft Research Health Futures

²Microsoft Research India

³Microsoft Azure AI

⁴Microsoft Health and Life Sciences

⁵Department of Radiology, Addenbrooke's Hos

Weike Zhao^{1,2}, Chaoyi Wu^{1,2}, Xiaoman Zhang^{1,2},
Ya Zhang^{1,2}, Yanfeng Wang^{1,2,†}, Weidi Xie^{1,2,†},

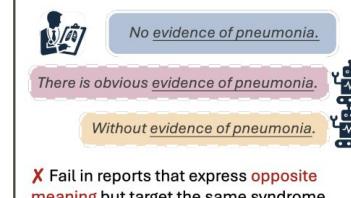
¹Shanghai Jiao Tong University, ²Shanghai AI Laboratory

<https://angelakeke.github.io/RaTEScore/>

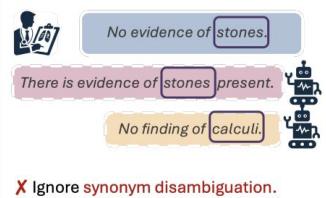
Abstract

This paper introduces a novel, entity-aware metric, termed as **Radiological Report (Text) Evaluation (RaTEScore)**, to assess the quality of medical reports generated by AI models. RaTEScore emphasizes crucial medical enti-

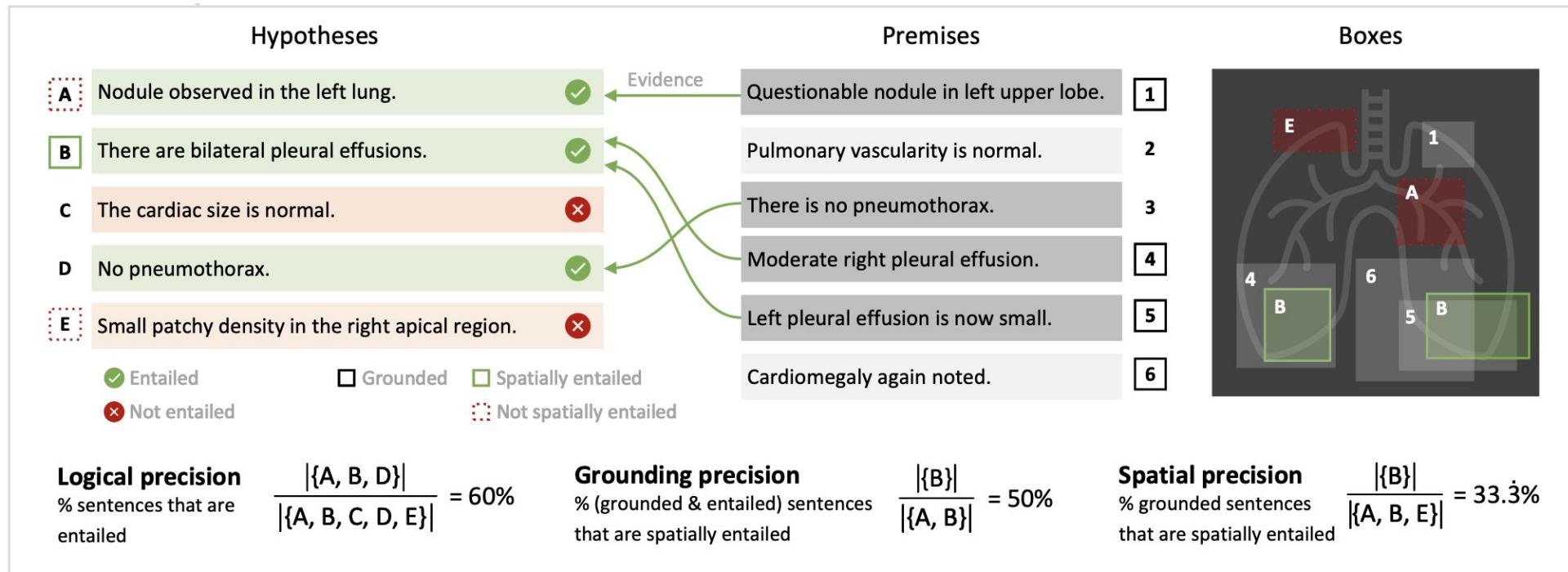
Word Overlap Metrics :



NER-F1 Metrics :

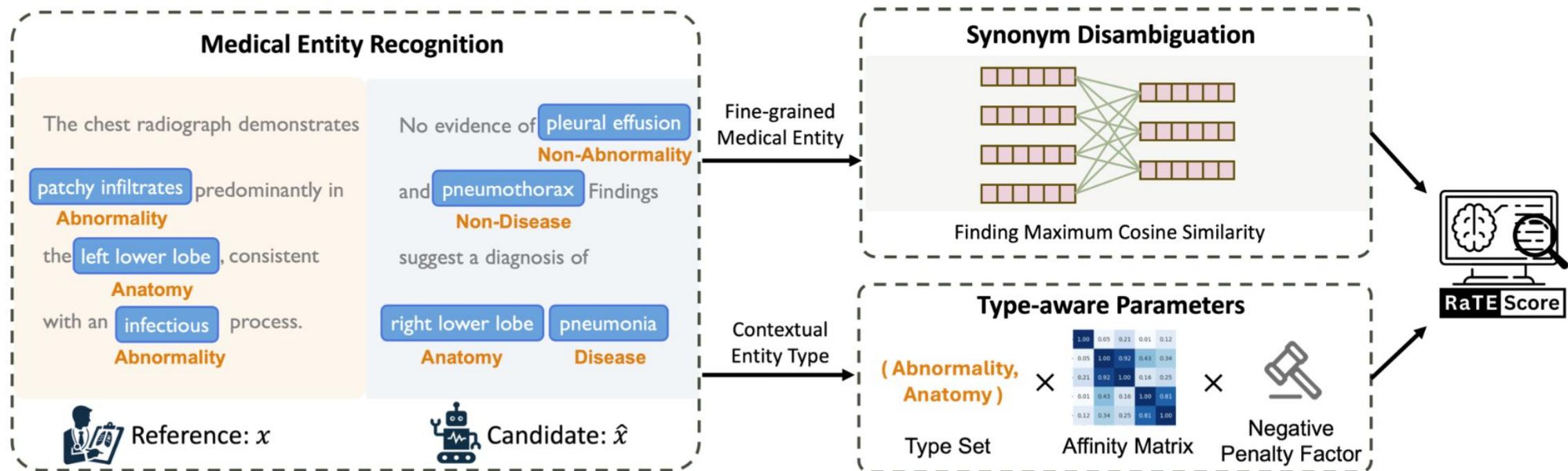


2. Evaluation metrics: RadFact



Llama3-70B-Instruct, no direct human metric evaluation

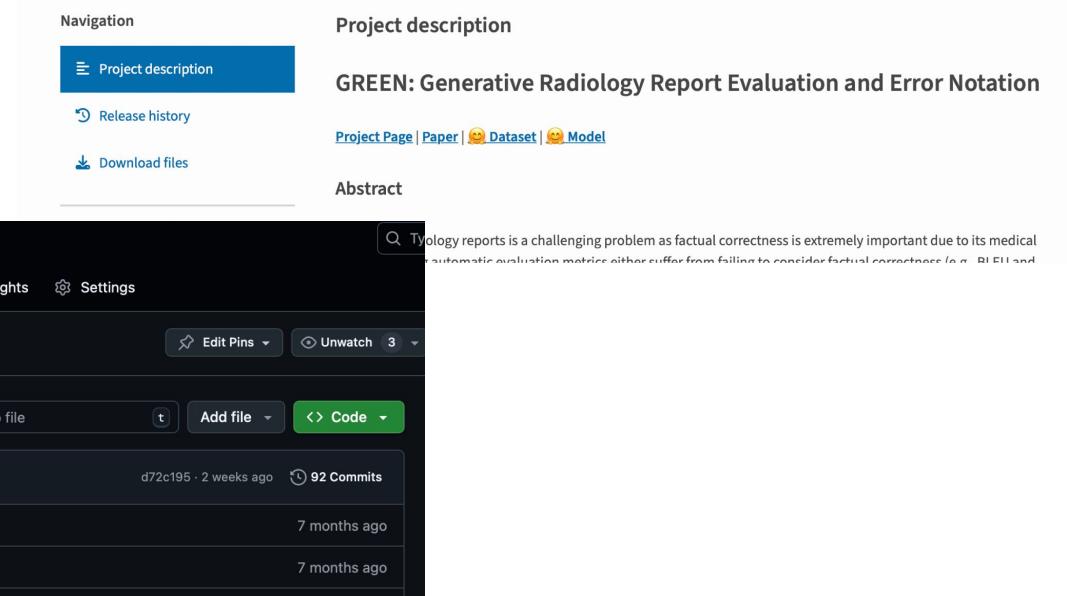
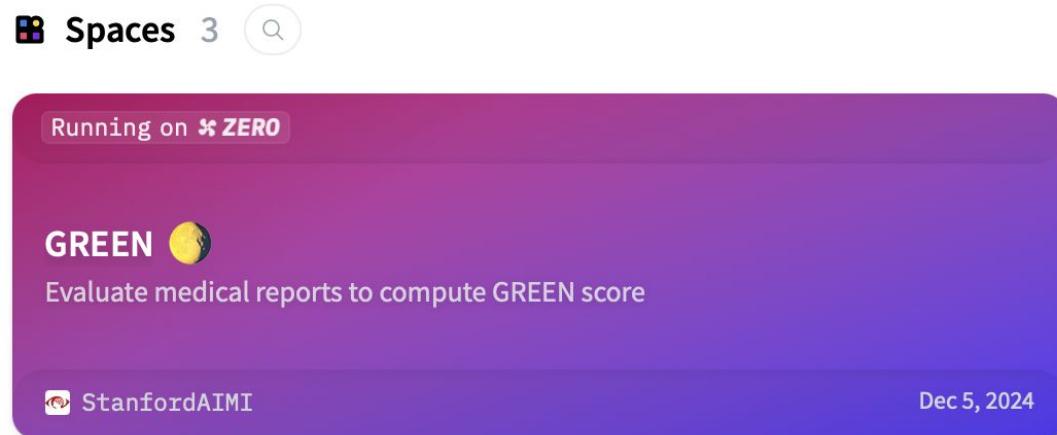
2. Evaluation metrics: RaTEScore



modality agnostic, efficient
not directly interpretable numeric score, lower performance

3. Practical Demonstration:

How to use LLMs to evaluate generated reports as well as error detection?



3. Practical Demonstration:

Spaces | StanfordAIMI/GREEN | like 0 | Running on ZERO | ⏪ | ⏴ | ⏵

GREEN Score Evaluation Demo

This demo evaluates a single pair of reference and hypothesis reports to compute the GREEN score.

Choose Input Type

Custom

Example 1

Example 2

Example 3

Reference Report

The heart size is normal. Lungs are clear without any infiltrates.

Hypothesis Report

The heart size is mildly enlarged. Mild infiltrates in the left upper lobe.

Model Name

StanfordAIMI/GREEN-radllama2-7b

GREEN Analysis

infiltrates, while the reference report states that the lungs are clear without any infiltrates.

[Clinically Significant Errors]:

- (a) False report of a finding: 2. The heart size is reported as mildly enlarged; Mild infiltrates are reported in the left upper lobe.
- (b) Missing a finding present in the reference: 0.
- (c) Misidentification of a finding's anatomic location/position: 0.
- (d) Misassessment of the severity of a finding: 0.
- (e) Mentioning a comparison that isn't in the reference: 0.
- (f) Omitting a comparison detailing a change from a prior study: 0.

[Clinically Insignificant Errors]:

- (a) False report of a finding: 0.
- (b) Missing a finding present in the reference: 0.
- (c) Misidentification of a finding's anatomic location/position: 0.
- (d) Misassessment of the severity of a finding: 0.
- (e) Mentioning a comparison that isn't in the reference: 0.
- (f) Omitting a comparison detailing a change from a prior study: 0.

[Matched Findings]:

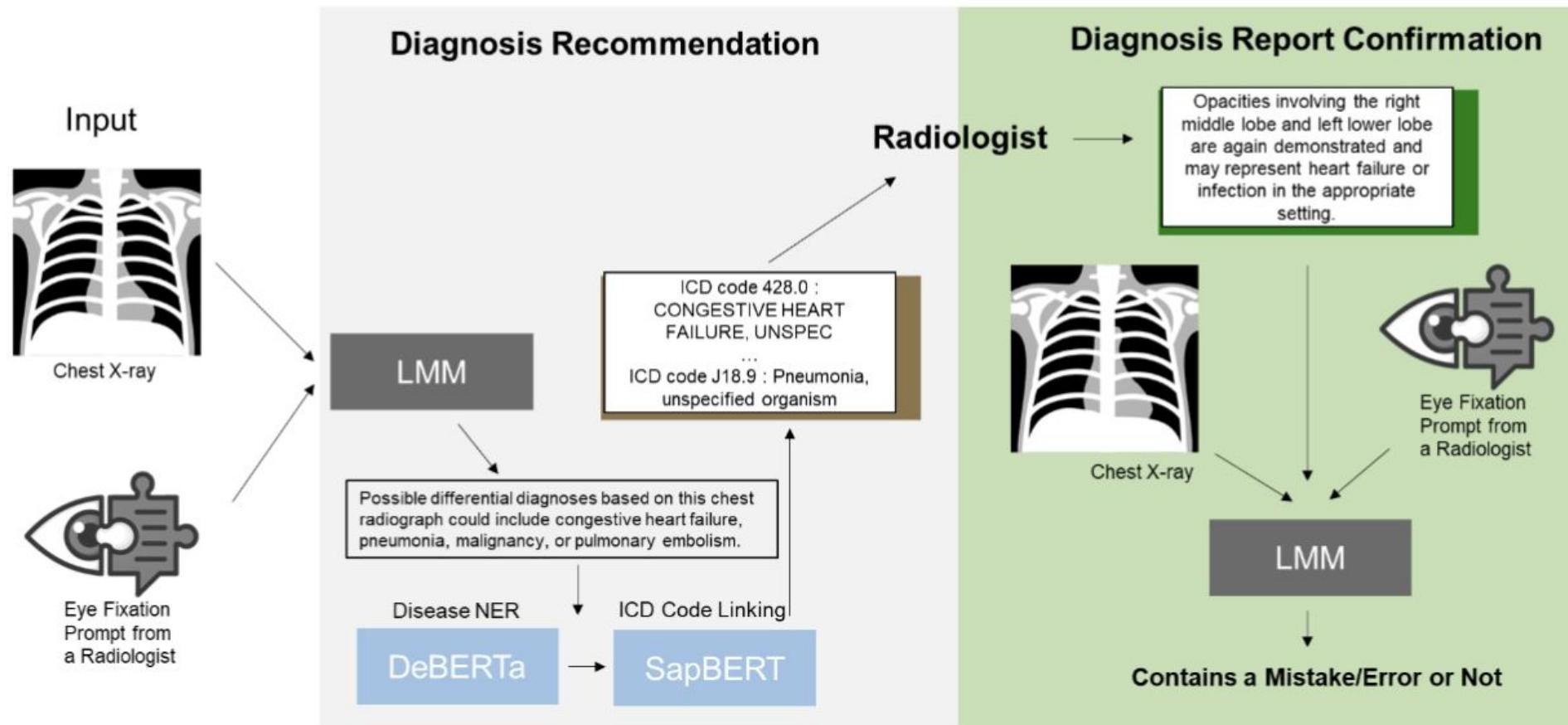
0.

Scores and Errors

Metric	Value
green_score	0
(a) False report of a finding in the candidate	2
(b) Missing a finding present in the reference	0
(c) Misidentification of a finding's anatomic location/position	0
(d) Misassessment of the severity of a finding	0
(e) Mentioning a comparison that isn't in the reference	0
(f) Omitting a comparison detailing a change from a prior study	0
Matched Findings	0

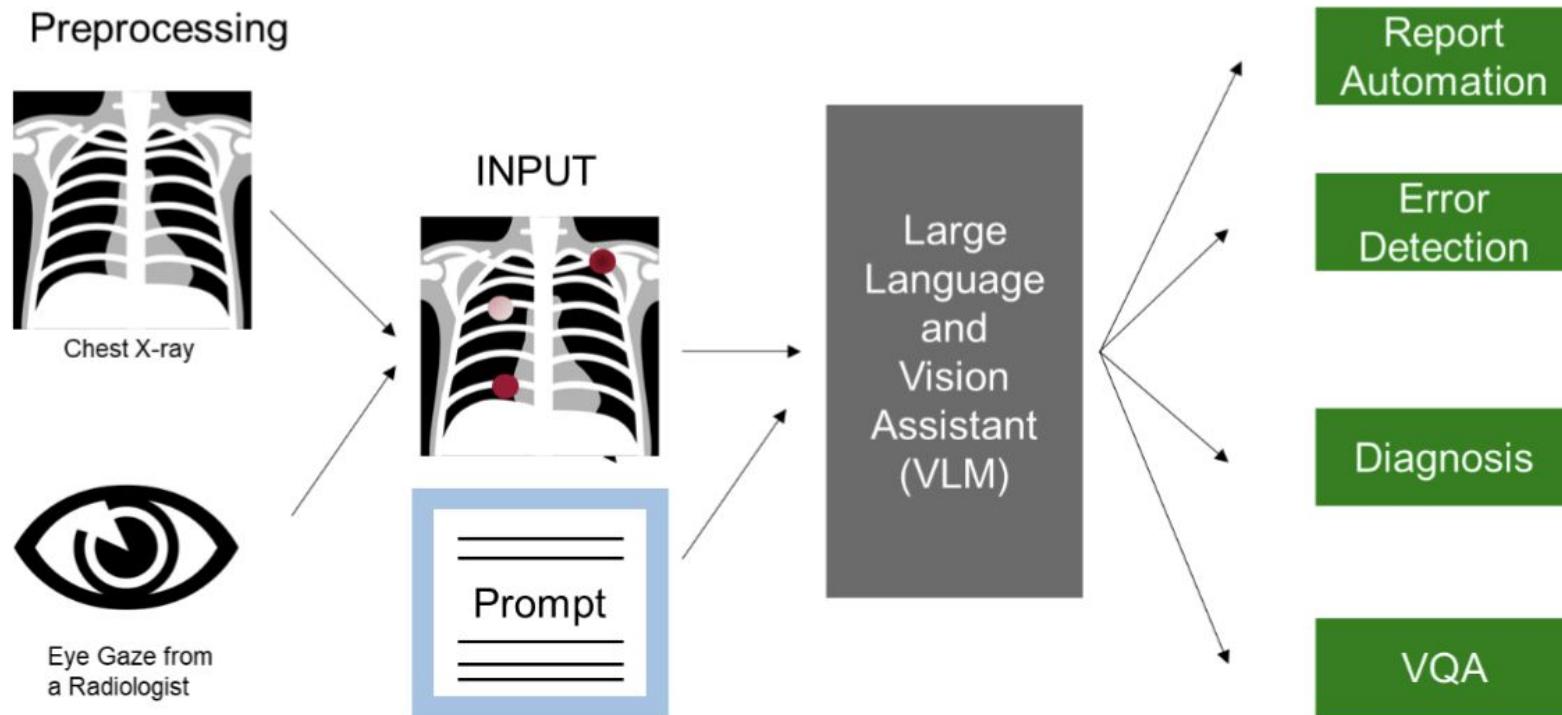


- ▶ Using metric for **direct alignment algorithms** → “ChexAlign”, ACL 2025
- ▶ **Generalization of GREEN** using self-supervised, data-efficient distillation → “MedVal”, in submission
- ▶ **Structuring Reports** for better evaluation → “Automated Structured Radiology Report Generation”, ACL 2025
- ▶ **General Internal Signals of LLMs** → “Head Entropy”, Coming soon 🚀



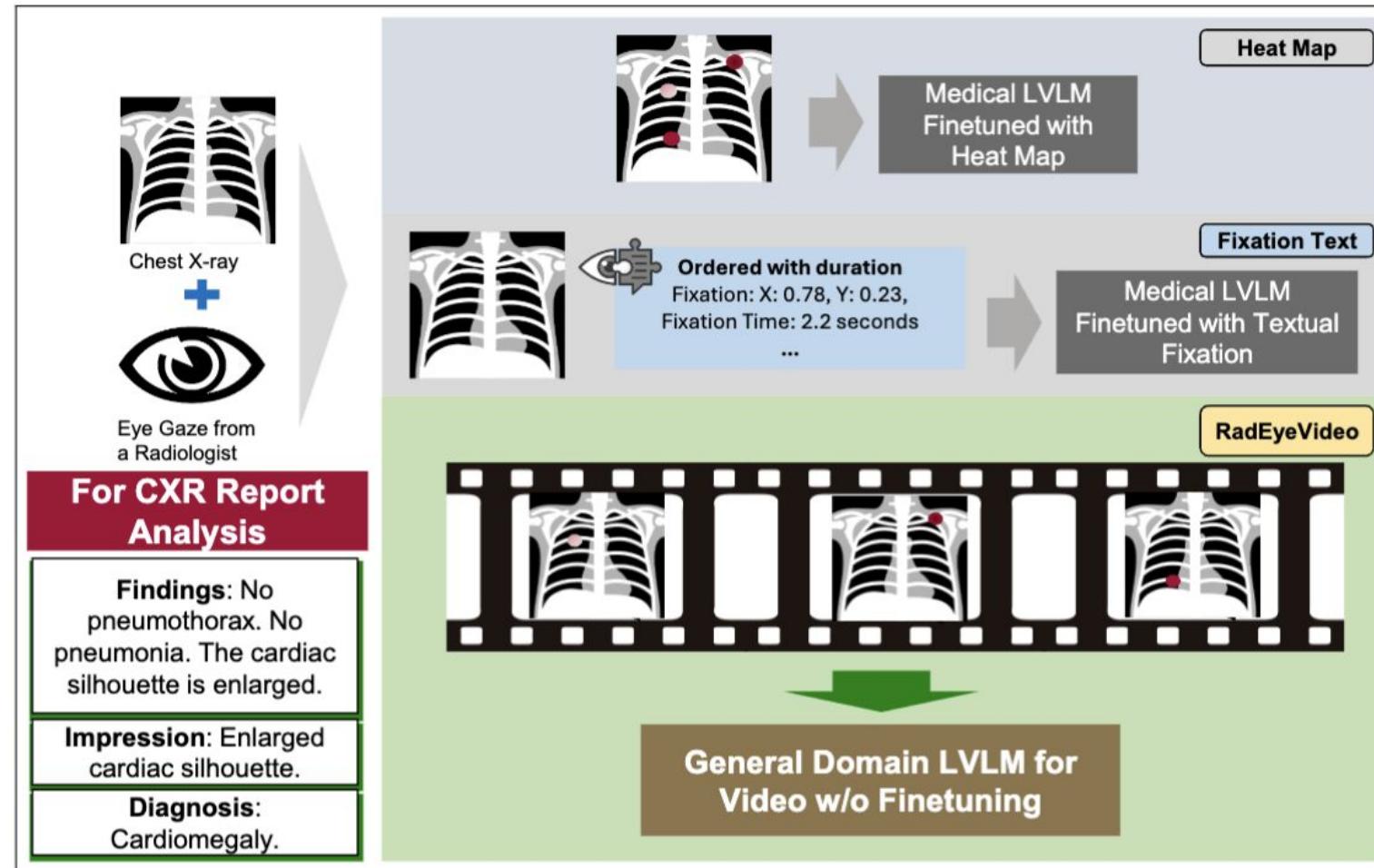
Integrate the eye fixation information into textual prompts

VI. Enhancing LVLMs utilization: Human Computer Interaction



Heatmaps with VLMs for clinical applications

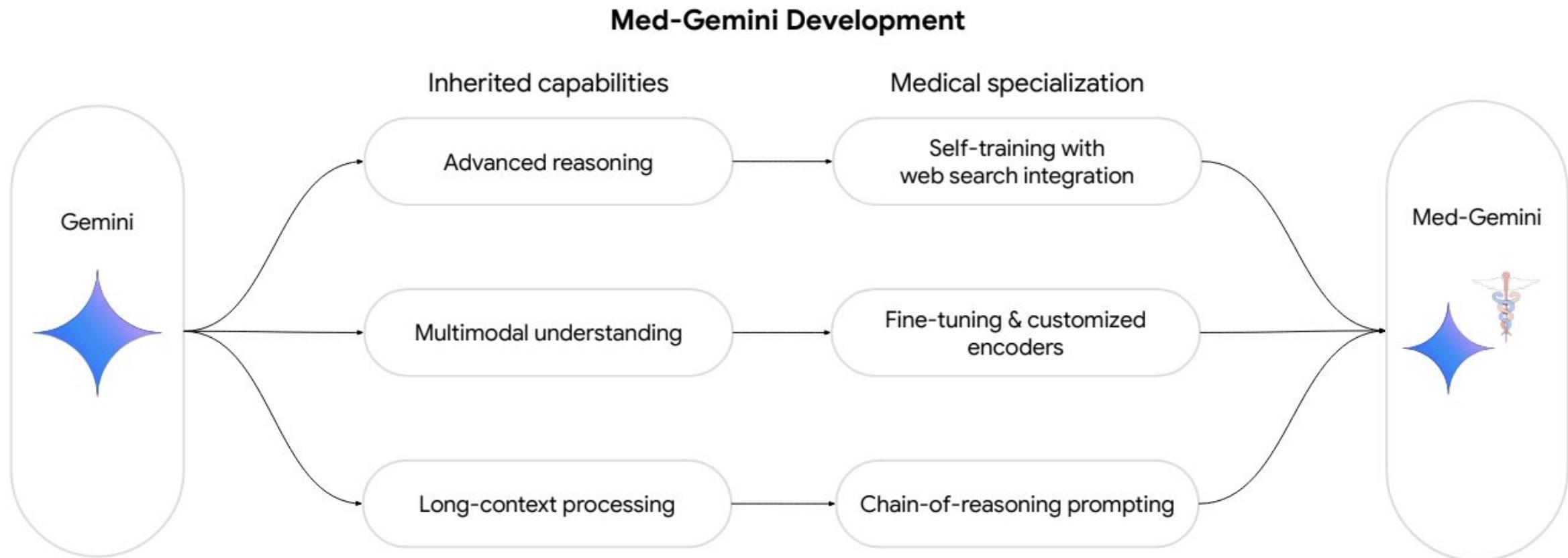
Highlight the precise focal points and duration of a radiologist's attention when analysing a CXR



Dynamic Eye Gaze with Video for CXR Analysis

- leverage both spatial and temporal aspects of gaze patterns

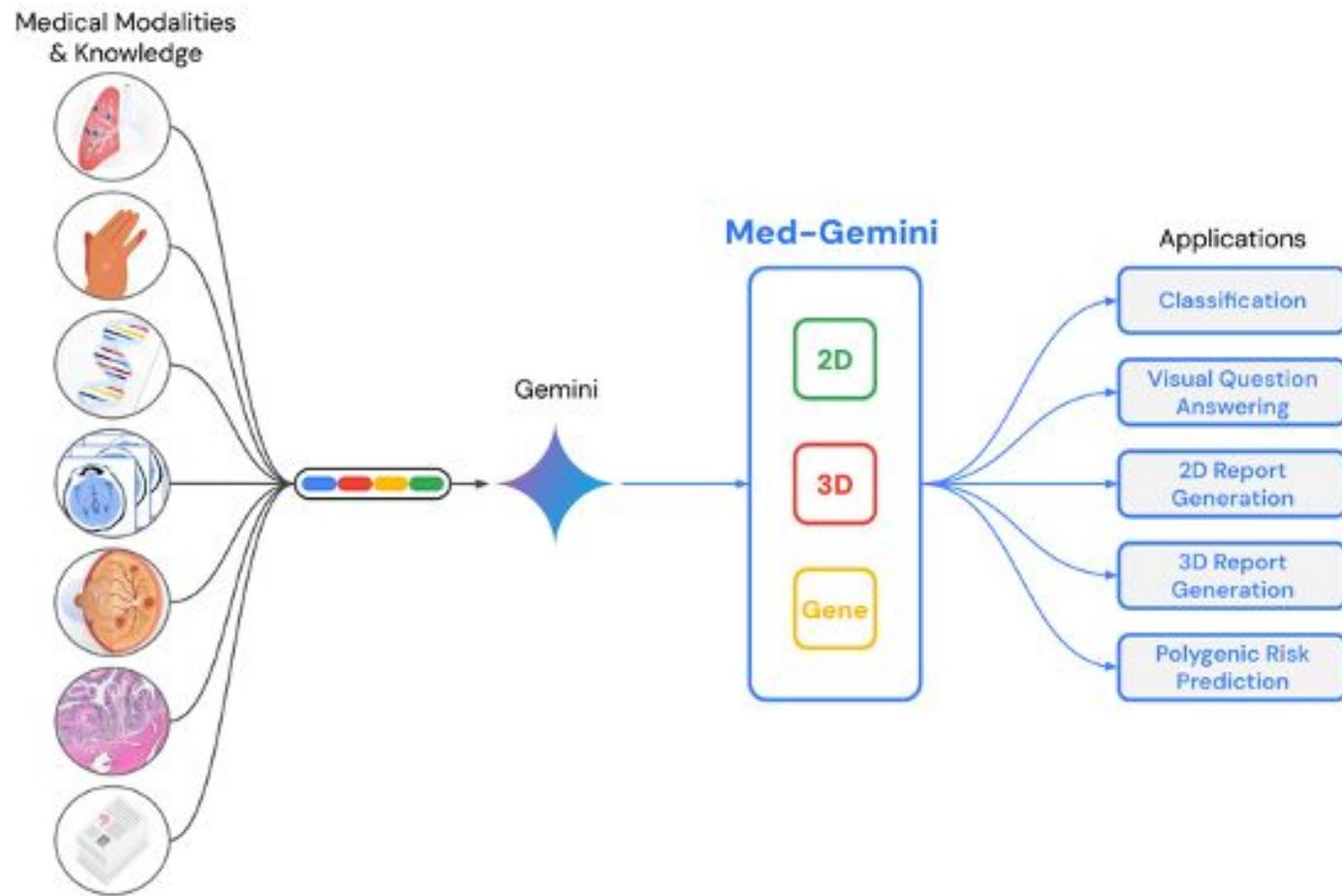
Med-Gemini: Gemini Models in Medicine



Saab, K. et al. *Capabilities of Gemini Models in Medicine*. 2024.

Yang, L. et al. *Advancing Multimodal Medical Capabilities of Gemini*. 2024.

Med-Gemini: Gemini Models in Medicine



Saab, K. et al. *Capabilities of Gemini Models in Medicine*. 2024.

Yang, L. et al. *Advancing Multimodal Medical Capabilities of Gemini*. 2024.

Med-Gemini: Gemini Models in Medicine

Modality	Dataset	No. examples	No. Images	Description
Radiology (2D)	Slake-VQA	4,919	450	Radiology images & QA pairs
	MIMIC-CXR	2,142,892	231,483	Radiology images & free-form reports
	Digital Knee X-ray	1,469	1,469	Knee X-ray images & labels
	CXR-US2	132,680	132,680	Radiology images & free-form reports
	NLST	2,199	2,199	2D CT slices & free-form reports
	CT-US1	3,207	3,207	2D CT slices & free-form reports
Radiology (3D)	CT-US1	657,719	657,719	3D CT images & free-form reports
Pathology	PathVQA	19,654	2,599	Pathology images & QA pairs
	Histopathology	1,550,976	207,603	Histopathology images, captions, & QA pairs
Dermatology	PAD-UFES-20	2,047	2,047	Skin lesion images & labels
Ophthalmology	EyePACS	14,406	14,406	Fundus images & labels
Medical VQA	PMC	2,246,656	2,246,656	PubMed Central images & caption pairs
	MedVQA	12,664	3,168	Medical images & QA pairs
Genomics	UK Biobank	259,225	259,225	Genomic data & disease outcomes

Signal QA | ECG-QA (cardiology) | 159,306 | Close-ended signal QA of electrocardiograms

More than 7 million data samples from 3.7 million medical images and cases is used for fine-tuning and further instruction-tuning of Gemini for medical applications in Med-Gemini

Saab, K. et al. *Capabilities of Gemini Models in Medicine*. 2024.

Yang, L. et al. *Advancing Multimodal Medical Capabilities of Gemini*. 2024.

Med-Gemini: Gemini Models in Medicine

Evaluation:

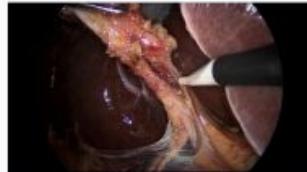
1. Evaluation of advanced reasoning on text-based tasks - QA, generation, summarization, simplification, diagnostics
2. Evaluation of multimodal capabilities- 2D, 3D, Signal QA
3. Evaluation of long-context capabilities on video and EHR tasks
4. Evaluation of Genomics “images” (polygenic risk scores (PRS) projected into 2D)

Saab, K. et al. *Capabilities of Gemini Models in Medicine*. 2024.

Yang, L. et al. *Advancing Multimodal Medical Capabilities of Gemini*. 2024.

Med-Gemini: Gemini Models in Medicine

Prompt



...



...



You are a helpful medical video assistant.

Task:

Assess whether Critical View of Safety (CVS) is fully achieved in the provided cholecystectomy video.

The Critical View of Safety (CVS) is fully achieved if the following three criteria are met:

- C1: Clear view of 2 tubular structures connected to the gallbladder.
- C2: A carefully dissected hepatocystic triangle presenting an unimpeded view of only the 2 cystic structures and the cystic plate.
- C3: The lower third of the gallbladder is dissected off the cystic plate.

Instructions:

1. Watch the video carefully, and answer which of the Critical View of Safety (CVS) criteria are met.

2. Output Format:

- CVS Assessment:
 - C1 - achieved / partially achieved / not achieved (Optional explanation)
 - C2 - achieved / partially achieved / not achieved (Optional explanation)
 - C3 - achieved / partially achieved / not achieved (Optional explanation)

Response

CVS Assessment:

C1 - partially achieved: Two tubular structures are visible but the view is not entirely clear due to some bleeding obscuring the area.

C2 - not achieved: The hepatocystic triangle is not clearly dissected, and there is not an unimpeded view of only the two cystic structures and the cystic plate.

C3 - not achieved: The lower third of the gallbladder is not dissected off the cystic plate.

Example of Med-Gemini's long-context capabilities on a surgical video.

It analyzes a video clip from the Cholec80 dataset to assess achievement of the Critical View of Safety (CVS) during a laparoscopic cholecystectomy (a keyhole operation to remove the gallbladder). The model assesses whether the three criteria defining the achievement of the CVS are met, with a detailed explanation for each criterion separately.

Saab, K. et al. *Capabilities of Gemini Models in Medicine*. 2024.

Yang, L. et al. *Advancing Multimodal Medical Capabilities of Gemini*. 2024.

Overview



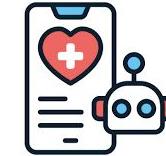
Auto-regressive Training



Instruction Fine-tuning



RL Fine-tuning

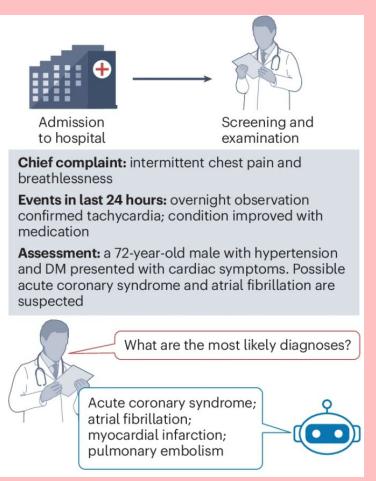


Clinical LLMs

Applications:

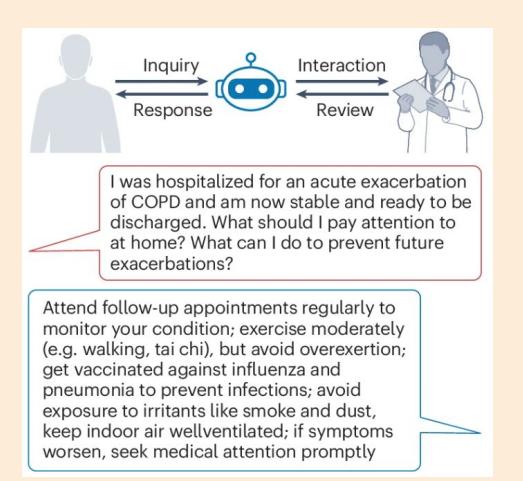
• Medical Decision-Making:

- Diagnosis
- Treatment Planning
- Risk Prediction
- Mortality
- Prognosis
- ...



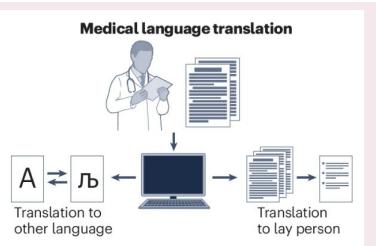
• Medical Inquiry and Response (Chatbot):

- Conversational Consultation
- Patient Education
- Virtual Coach
- Spiritual Care
- Follow-up Reminder
- ...



• Medical Translation:

- Multilingual
- Lay Translation
- ...



Others

- Research Support
- Robotic Instruction
- Clinical Trial Design

• Mental Health Support

- Research Support
- Workflow Assistant
- Note Generation

- Medical Education
- Lab Test Interpretation
- ... (so many!!)

The Pre-training Era (2019 - 2025)

- Technical focus: improving base model by **injecting medical knowledge**, i.e., collecting medical-specific data.
- Applications: **Multiple choice question (MCQ) medical exams** (knowledge-wise question-answering) and Chatbot

- Architecture: **Encoder-only** (BERT)
- Training stage: **Pre-training**
- Training objective: **Masked token completion**
- Applications: Specific **fine-tuning**
- Data: EHR+Medical Literature

- Architecture: **Decoder-only** (GPT)
- Training stage: **Pre-training and Instruction Fine-tuning**
- Training Objective: **Auto-regressive**
- Applications: MCQ exams and Chatbot (**zero-shot**)
- Data: **Web-scale Corpus + Instruction Data (MCQs & dialogues)**

With the advancements by various medical LLMs, we have finally **surpassed expert-level performance in medical exams !**

2019-2023

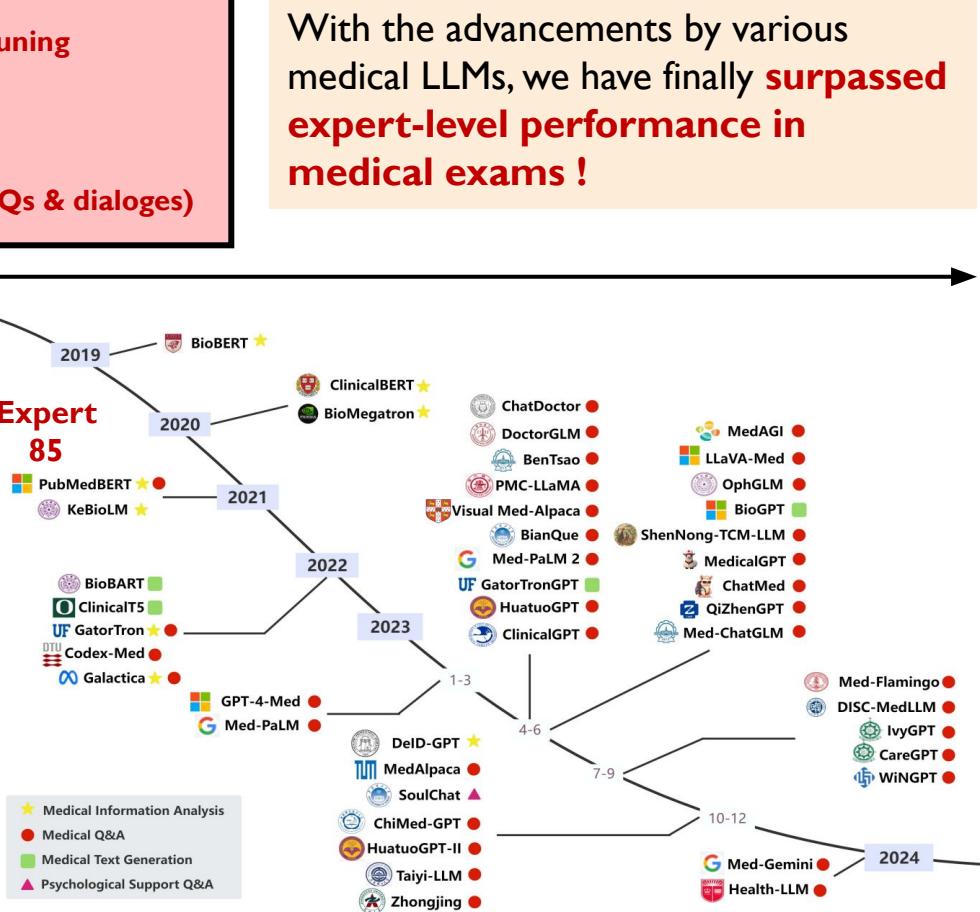
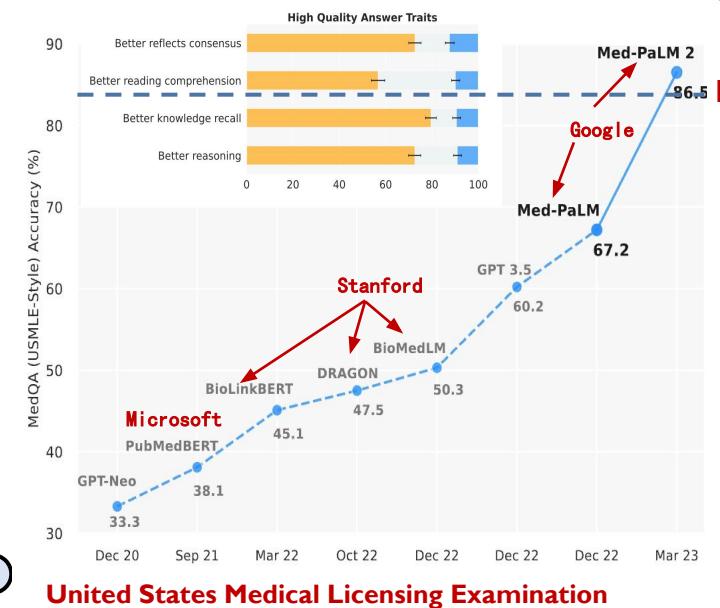
BioBERT (2019.01)
Journal: *Bioinformatics*, Size: **110M**
Data: **18B tokens**, PubMed + PMC

ClinicalBERT (2019.04)
Journal: *arXiv*, Size: **110M**
Data: **112k clinical notes**, MIMIC-IV

PubMedBERT (2020.07)
Conference: *ACL*, Size: **110M/340M**
Data: **3.2B tokens**, PubMed + PMC

...
GatorTron (2022.12)
Journal: *npj Digit. Med.*, Size: **8.9B**
Data: **>82B tokens + 6B tokens**, EHRs + PubMed

2023-2025



The Post-training Era (2025 - now)

- Technical focus: improving the model's reasoning ability.
- Applications: **serious and practical clinical tasks**.

Method:

Unknown:

2025.05: MedGemma

MedGemma 27B has been trained exclusively on medical text and optimized for inference-time computation.

MedGemma

A Gemma 3 variant optimized for medical text and image comprehension

m1: Unleash the Potential of Test-Time Scaling for Medical Reasoning with Large Language Models

Xiaoke Huang¹, Juncheng Wu¹, Hui Liu², Xianfeng Tang², Yuyin Zhou¹

¹ UC Santa Cruz ² Amazon Research

<https://github.com/UCSC-VLAA/m1>

2025.05: m1 (test-time scaling)

Prompt the model to perform test-time scaling through thinking step-by-step.
Build up reasoning MCQ data through prompting general LLMs and filtering.

HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs

Junying Chen¹, Zhenyang Cai¹, Ke Ji¹, Xidong Wang¹, Wanlong Liu¹
Rongsheng Wang¹, Jianye Hou¹, Benyou Wang^{1,2*}

¹ The Chinese University of Hong Kong, Shenzhen

² Shenzhen Research Institute of Big Data

2024.12: HuatuoGPT-o1 (instruction + RL)

Distill general reasoning LLMs + medical verifier by prompting GPT-4o.
RL through verifiable medical questions
Construct a 40K medical verification dataset from MCQs

MedReason: Eliciting Factual Medical Reasoning Steps in LLMs via Knowledge Graphs

Juncheng Wu^{1,*}, Wenlong Deng^{2,8,*}, Xingxuan Li³, Sheng Liu⁴, Taomian Mi²,
Yifan Peng⁵, Ziyang Xu⁶, Yi Liu⁵, Hyunjin Cho⁷, Chang-In Choi⁹, Yihan Cao¹⁰,
Hui Ren¹⁰, Xiang Li¹⁰, Xiaoxiao Li^{2,8,†}, Yuyin Zhou^{1,†}

2025.05: MedReason (instruction + RL)

Utilize a structured medical knowledge graph (KG) to convert clinical MCQ pairs into logical chains of reasoning.
Trace connections from question elements to answers via KG entities.

The Post-training Era (2025 - now)

Diagnosis:

2025.01

Develop a 176B medical **diagnosis LLM** through pre-training and SFT

nature medicine

Explore content ▾ About the journal ▾ Publish with us ▾

nature > nature medicine > articles > article

Article | Published: 08 January 2025

A generalist medical language model for disease diagnosis assistance

Xiaohong Liu, Haili Liu, Guoxing Yang, Zeyu Jiang, Shuguang Cui, Zhaoze Zhang, Huan Wang, Liyuan Tao, Yongchang Sun, Zhu Song, Tianpei Hong, Jin Yang, Tianran Gao, Jiangjiang Zhang, Xiaohu Li, Jing Zhang, Ye Sang, Zhao Yang, Kanmin Xue, Song Wu, Ping Zhang, Jian Yang✉, Chunli Song✉ & Guangyu Wang✉

2025.04

Articulate Medical Intelligence Explorer (AMIE), language model, optimized for diagnostic reasoning.

nature

Explore content ▾ About the journal ▾ Publish with us ▾

nature > articles > article

Article | [Open access](#) | Published: 09 April 2025

Towards accurate differential diagnosis with large language models

Daniel McDuff✉, Mike Schaeckermann✉, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, ... Vivek Natarajan✉ + Show authors

2025.03

Benchmark how the **lab results affect the DDx prediction for LLMs**

npj | digital medicine

Explore content ▾ About the journal ▾ Publish with us ▾

nature > npj digital medicine > articles > article

Article | [Open access](#) | Published: 18 March 2025

Preliminary analysis of the impact of lab results on large language model generated differential diagnoses

Balu Basuram, Qiao Jin, Yuzhang Xie, Carl Yang, Karim Hanna, Jennifer Costa, Cindy Shavor, Wenshan Han, Zhiyong Lu & Zhe He✉

2025.04

Conversational diagnosis by simulating patients and doctors to generate instruction data, with a rule-based reasoning chain and multi-turn interaction.

nature

Explore content ▾ About the journal ▾ Publish with us ▾

nature > articles > article

Article | [Open access](#) | Published: 09 April 2025

Towards conversational diagnostic artificial intelligence

Tao Tu✉, Mike Schaeckermann✉, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webster, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, ... Vivek Natarajan✉ + Show authors

Clinical Decision-making:

2024.07

Evaluation **LLM Decision-making ability leveraging EHR data.**

nature medicine

Explore content ▾ About the journal ▾ Publish with us ▾

nature > nature medicine > articles > article

Article | [Open access](#) | Published: 04 July 2024

Evaluation and mitigation of the limitations of large language models in clinical decision-making

Paul Hager✉, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Bräuer, Georgios Kaassis & Daniel Rueckert

2024.10

Evaluation **LLM clinical recommendation ability leveraging EHR data** (10,000 Emergency Department visits from MIMIC-ED).

nature communications

Explore content ▾ About the journal ▾ Publish with us ▾

nature > nature communications > articles > article

Article | [Open access](#) | Published: 08 October 2024

Evaluating the use of large language models to provide clinical recommendations in the Emergency Department

2025.05

Evaluate **LLMs' decision-making ability for triage and referral, and diagnosis, with EHR data (MIMIC-ED).**

npj | digital medicine

Explore content ▾ About the journal ▾ Publish with us ▾

nature > npj digital medicine > articles > article

Article | [Open access](#) | Published: 09 May 2025

Evaluating large language model workflows in clinical decision support for triage and referral and diagnosis

Sarah Sandmann, Stefan Hegelmann, Michael Fajarski, Lucas Bickmann, Benjamin Wild, Roland Eils✉ & Julian Varghese

Farieda Gaber, Maqsood Shaik, Fabio Allega, Agnes Julia Bilecz, Felix Busch, Kelsey Goon, Vedran Franke & Altuna Akalin✉

The Post-training Era (2025 - now)

Holistic Clinical Task Collection:

npj | digital medicine

Explore content ▾ About the journal ▾ Publish with us ▾

nature > npj digital medicine > articles > article

Article | [Open access](#) | Published: 27 January 2025

Towards evaluating and building versatile large language models for medicine

2025.01

Systematically **curate different existing bioNLP tasks** and build up a new wide clinical task coverage instruction dataset.

2025.03

Build up Me-LLaMA through pre-training and instruction tuning with **multiple clinical text analysis tasks**.

npj | digital medicine

Explore content ▾ About the journal ▾ Publish with us ▾

nature > npj digital medicine > articles > article

Article | [Open access](#) | Published: 05 March 2025

Medical foundation large language models for comprehensive text analysis and beyond

Explore content ▾ About the journal ▾ Publish with us ▾

nature > npj digital medicine > articles > article

Article | [Open access](#) | Published: 05 March 2025

Medical foundation large language models for comprehensive text analysis and beyond

Qianqian Xie, Qingyu Chen, Aokun Chen, Cheng Peng, Yan Hu, Fongci Lin, Xueqing Peng, Jimin Huang, Jeffrey Zhang, Vipina Keloth, Xinyu Zhou, Lingfei Qian, Huan He, Dennis Shung, Lucila Ohno-Machado, Yonghui Wu, Hua Xu & Jiang Bian

Clinical Trial:

2024.11

Leverage LLMs to **match patients with clinical trials with prompt engineering**.

nature communications

Explore content ▾ About the journal ▾ Publish with us ▾

nature > nature communications > articles > article

Article | [Open access](#) | Published: 18 November 2024

Matching patients to clinical trials with large language models

Qiao Jin, Zifeng Wang, Charalampos S. Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun & Zhiyong Lu

Patient Interaction:

2024.11

Build up a **conversational interaction evaluation frame**, reflecting how LLMs can interact with patients work to evaluate.

nature communications

Explore content ▾ About the journal ▾ Publish with us ▾

nature > nature communications > articles > article

Article | [Open access](#) | Published: 18 November 2024

Matching patients to clinical trials with large language models

Qiao Jin, Zifeng Wang, Charalampos S. Floudas, Fangyuan Chen, Changlin Gong, Dara Bracken-Clarke, Elisabetta Xue, Yifan Yang, Jimeng Sun & Zhiyong Lu

The Post-training Era (2025 - now)

- Technical focus: improving the model's reasoning ability.
- Applications: **serious and practical clinical tasks**.

Evaluating final accuracy on reasoning-wise questions:

2025.01:

Integrate metacognitive tasks **into** multiple-choice medical questions.

Current models exhibit a critical disconnect between perceived and actual capabilities in medical reasoning.

The screenshot shows the header of a Nature Communications article. The title is 'Large Language Models lack essential metacognition for reliable medical reasoning'. It includes the names of the authors: Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, & Demet Yüksel. The article was published on 14 January 2025.

2025.01:
Organize a new expert-level reasoning MCQ dataset

2025.05:
Train a **BERT** classifier to disentangle the reasoning-wise MCQs from MCQ benchmarks.

MedXpertQA: Benchmarking Expert-Level Medical Reasoning and Understanding

Yuxin Zuo^{*1}, Shang Qu^{*1,2}, Yifei Li¹, Zhangren Chen¹, Xuekai Zhu¹, Ermo Hua¹, Kaiyan Zhang¹, Ning Ding^{1,2}, Bowen Zhou^{1,2}

Disentangling Reasoning and Knowledge in Medical Large Language Models

Rahul Thapa^{1,2,*}, Qingyang Wu²,
Kevin Wu¹, Harrison Zhang¹, Angela Zhang³, Eric Wu¹,
Haotian Ye¹, Suhana Bedi¹, Nevin Aresh¹, Joseph Boen¹,
Shriya Reddy¹, Ben Athiwaratkun², Shuaiwen Leon Song², James Zou^{1,2}

Reasoning quality comparison:

2025.04:

Answering questions from the **United States Medical Licensing Examination (USMLE)**, interpreting the basis of text-based diagnostic and management cases, for tumor classification, and report summary.
Manually rate the reasoning quality

The screenshot shows the header of a Nature Medicine article. The title is 'Comparative benchmarking of the DeepSeek large language model on medical tasks and clinical reasoning'. It includes the names of the authors: Mickael Tordjman, Zelong Liu, Murat Yuce, Valentin Fauveau, Yunhao Mei, Jerome Hadjadj, Ian Bolger, Haidara Almansour, Carolyn Horst, Ashwin Singh Parihar, Amine Geahchan, Anis Meribout, Nader Yatim, Nicole Ng, Phillip Robson, Alexander Zhou, Sara Lewis, Mingjian Huang, Timothy Deyer, Bachir Taouli, Hao-Chih Lee, Zahi A. Fayad, & Xueyan Mei. The article was published on 23 April 2025.

2025.02:

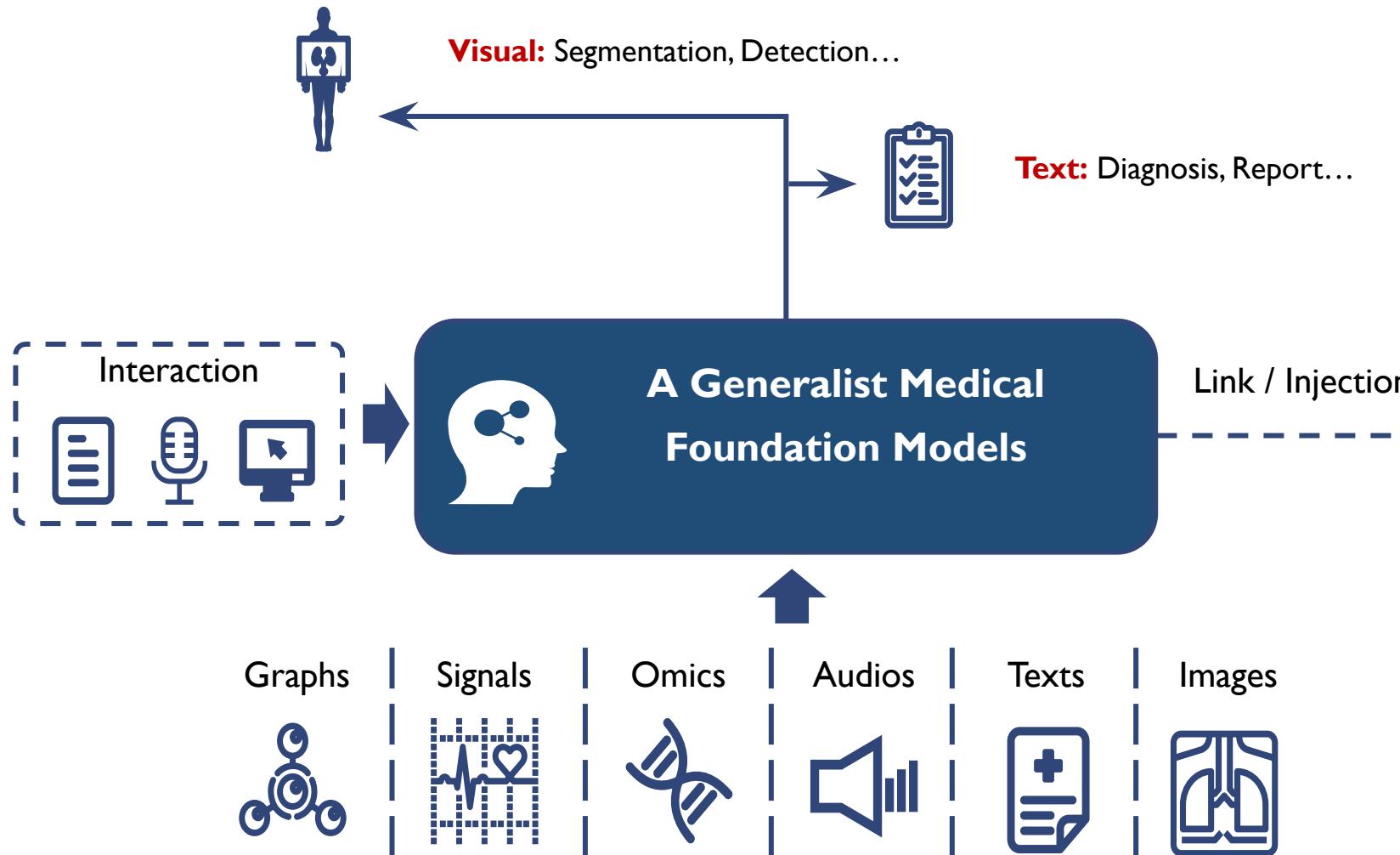
Quantifying the Reasoning Abilities of LLMs on Real-world Clinical Cases
Pengcheng Qiu^{1,2,*}, Chaoyi Wu^{1,2,*}, Shuyu Liu¹, Weike Zhao^{1,2}, Zhuoxia Chen³, Hongfei Gu³, Chuanjin Peng³, Ya Zhang^{1,2}, Yanfeng Wang^{1,2,†} and Weidi Xie^{1,2,‡}
Leverage real case report to benchmark the clinical decision-making for diagnosis, treatment, and examination recommendations
Quantify reasoning quality using agentic auto-metrics enhanced by web search.

Generative Visual-Language Models in AI4healthcare

Credits to @Chaoyi Wu, Jinghao Feng, Ziqing Fan, Haoning Wu

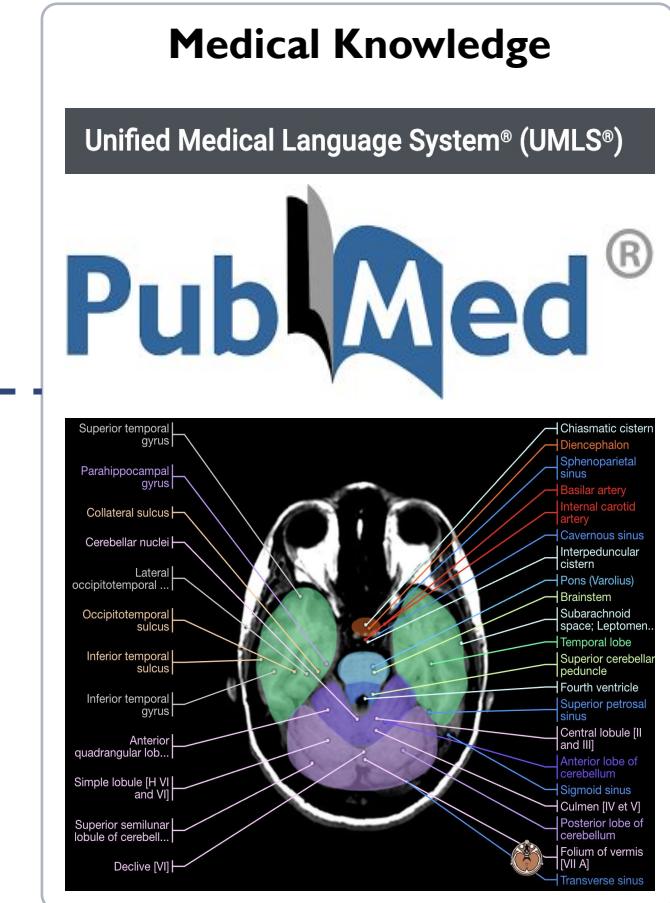


Medical MLLMs-Preliminary Goal



One modal for all

Strong generalization



Comprehensive knowledge

Trend

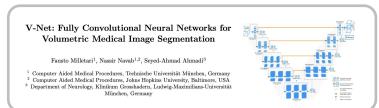
Task-specific small models



- 2015.5. U-Net, MICCAI



- 2016.6. V-Net, 3DV



- 2017.11 CheXNet



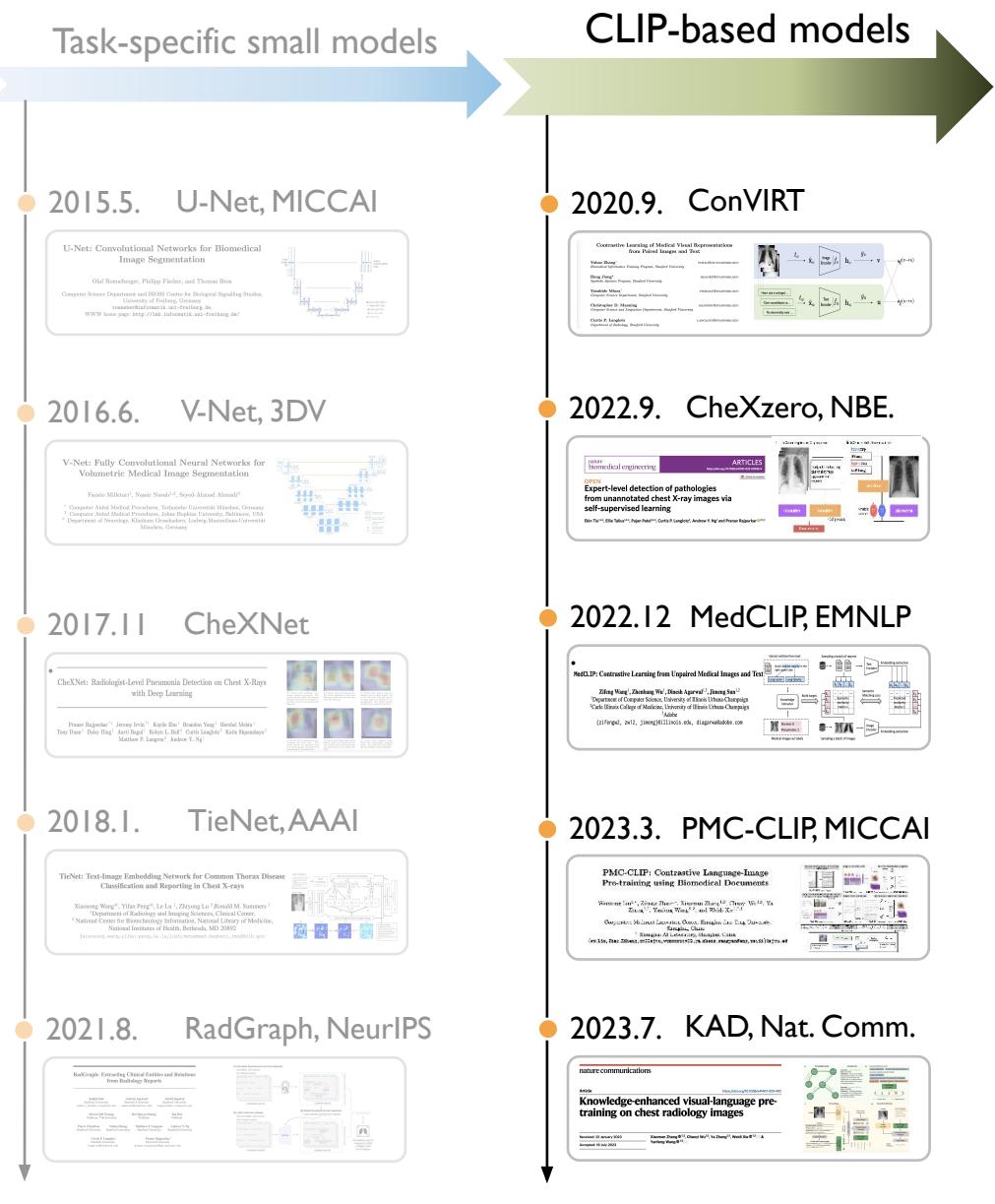
- 2018.I. TieNet,AAAI



- 2021.8. RadGraph, NeurIPS



Trend

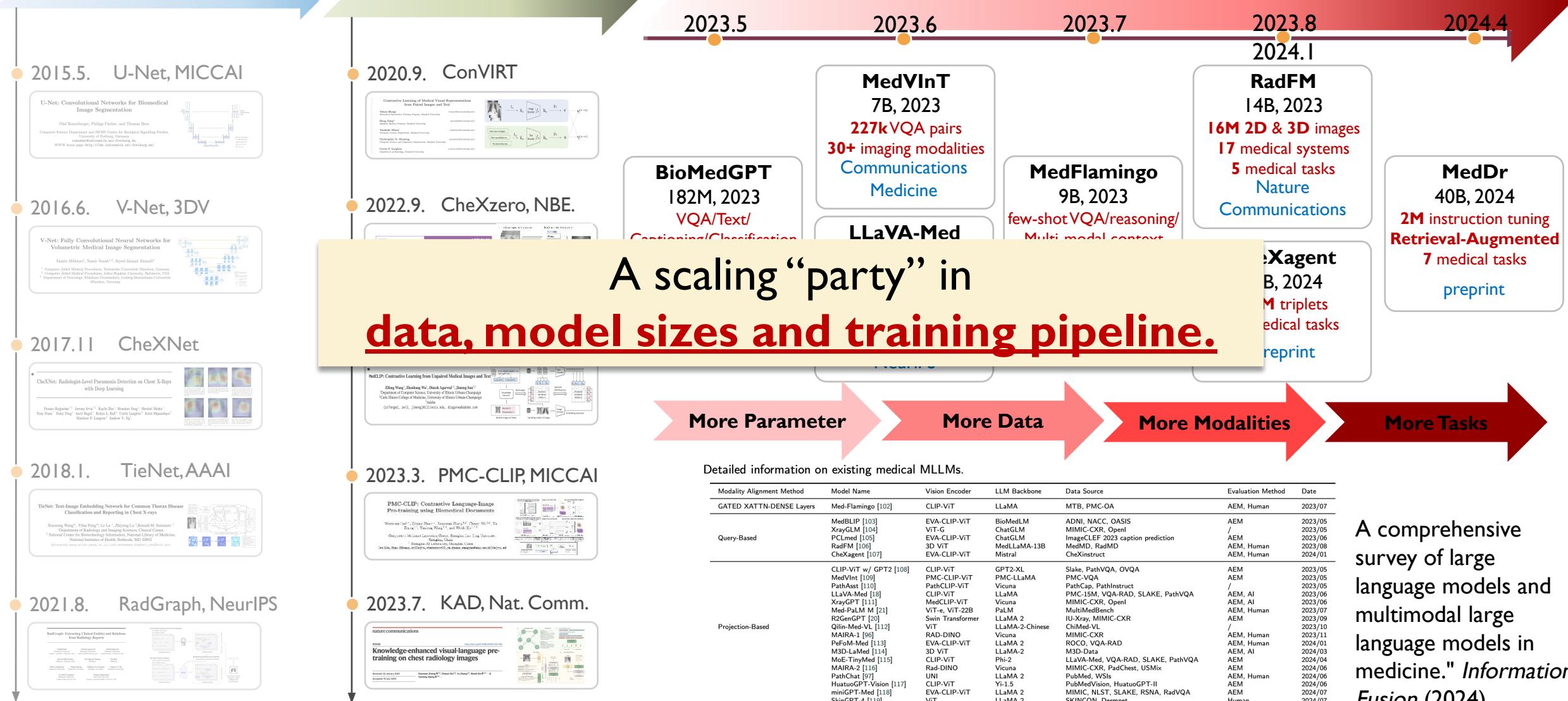


Trend

Task-specific small models

CLIP-based models

Large Foundation Models



A comprehensive survey of large language models and multimodal large language models in medicine." *Information Fusion* (2024)

Recent Progress and Future Directions

Reasoning Enhancement

Original Clinical Visual Question:



Does this chest X-ray show cardiomegaly?

Options:

A) Yes

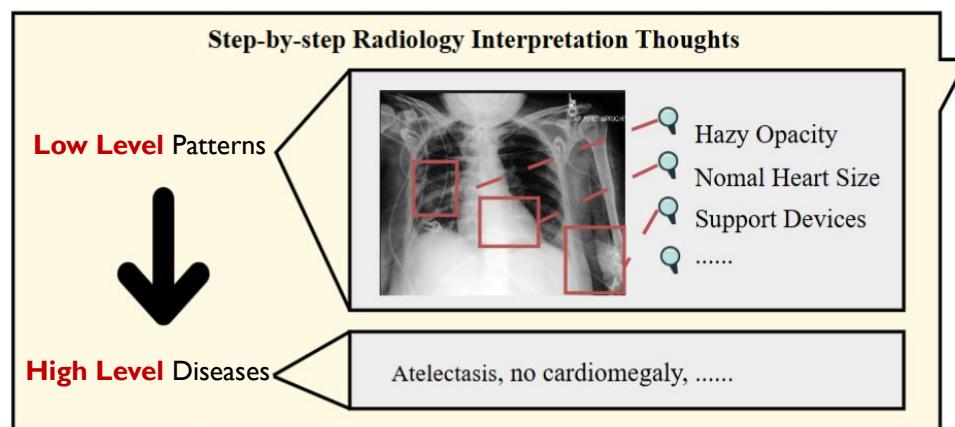
B) No

Previous Model Output:

Output: A) Yes

Result-only

Reasoning-enhanced Model Output:



Output:

Reasoning Processes:

To determine if the chest X-ray shows cardiomegaly, it's crucial to analyze the size of the heart. The key indicator is the cardiothoracic ratio, which is the ratio of the maximal horizontal cardiac diameter to the maximal horizontal thoracic diameter. In a normal PA chest X-ray, this ratio is expected to be less than 0.50 ...

Therefore, the answer is A) Yes

Recent Progress and Future Directions

Reasoning Enhancement

MLLM for reasoning with outcome RL[1-3]

Radiology VQA [1-2] (2025.02 & 05)

MedVLM-R1: Incentivizing Medical Reasoning Capability of Vision-Language Models (VLMs) via Reinforcement Learning

Jiazen Pan^{1,2*}, Che Liu^{3*}, Junde Wu², Fenglin Liu², Jiayuan Zhu², Hongwei Bran Li⁴, Chen Chen^{5,6}, Cheng Ouyang^{2,6†}, Daniel Rueckert^{1,6†}

Question Template
What can be observed in this image?
A) Tissue degeneration B) Labral pathology
C) Bone fracture D) Tendonitis

Model Input Prompt
{Question} Your task:
1. Think through the question step by step, enclose your reasoning process in <think>...</think> tags.
2. Then provide the correct single-letter choice (A, B, C, D,...) inside <answer>...</answer> tags.
3. No extra information or text outside of these tags.

Model Output Examples
<think> The image is a grayscale MRI image of an upper arm joint. The bicondylar humeral head of the humerus is visible. There is a well-defined ... </think>
<answer>B</answer>, there is no clear indication of ... </answer>

Format Reward = 1 due to the present of all tags and no content outside the tags
Accuracy Reward = 0.5 due to extra explanation appended after the answer

Med-R1: Reinforcement Learning for Generalizable Medical Reasoning in Vision-Language Models

Yuxiang Lai^{1*} Jike Zhong^{2*} Ming Li^{3*} Shitian Zhao⁴ Xiaofeng Yang^{1,5,6†}

¹Department of Computer Science and Informatics, Emory University

²Department of Computer Science, University of Southern California

³Department of Computer Science, University of Tokyo

⁴Department of Computer Science, Johns Hopkins University

⁵Department of Biomedical Engineering, Georgia Institute of Technology and Emory University

⁶Department of Radiation Oncology and Winship Cancer Institute, Emory University

Pathology VQA [3] (2025.05)

Patho-R1: A Multimodal Reinforcement Learning-Based Pathology Expert Reasoner

Wenchuan Zhang^{1,2*} Penghao Zhang^{3*} Jingru Guo^{4*} Tao Cheng⁵ Jie Chen²
Shuwan Zhang⁶ Zhang Zhang¹ Yuhan Yi^{1,2†} Hong Bu^{1,2}

¹Department of Pathology, West China Hospital, Sichuan University

²Institute of Clinical Pathology, West China Hospital, Sichuan University

³Independent Researcher ⁴University of Toronto ⁵Business School, Sichuan University

⁶Department of Pathology, Shengjing Hospital of China Medical University
zhangwenchuan@stu.scu.edu.cn, yuhanyi@scu.edu.cn

What is the condition of the interstitial in the image? *Ground truth: The interstitial in the image looks normal.* Sample source: Quilt-VQA

think> To determine what the image shows, I will follow a structured approach: 1. Examine the image: I will carefully observe the image to identify any visible features, such as tissue types, cellular structures, or any recognizable patterns. 2. Identify Key Features: I will look for distinctive characteristics like the presence of glands, stroma, or any specific tissue arrangement that might suggest a particular type of tissue or condition. 3. Consider Contextual Clues: I will consider the context of the question and any additional information that might be provided, such as clinical history or imaging findings. **answer>** The image shows a section of tissue that appears to have glands and surrounding stroma. The presence of these glands suggests the image is from a glandular organ system. Given the context of the question, the image is likely a representation of a breast tissue section, which often contains lobules and stroma. The appearance of these elements is consistent with a normal or benign breast tissue sample. **answer>**

Llama-3-ZV-11B-cot

think> After deciding that an image shows a borderline ovarian tumor, the next consideration should be to assess the specific features of the tumor that are characteristic of a borderline ovarian tumor. This includes evaluating the degree of atypia, the presence of stromal hyperplasia, and any architectural features. Additionally, it is important to consider the clinical context, including age, symptoms, and family history, as these can influence the management and prognosis of the tumor. The next step would be to perform a thorough histopathological examination to confirm the diagnosis and to determine the appropriate management plan, which may include surveillance, surgery, or chemotherapy depending on the specific characteristics of the tumor. **answer>** The next consideration should be to perform a thorough histopathological examination to confirm the diagnosis and to determine the appropriate management plan, which may include surveillance, surgery, or chemotherapy depending on the specific characteristics of the tumor. **answer>**

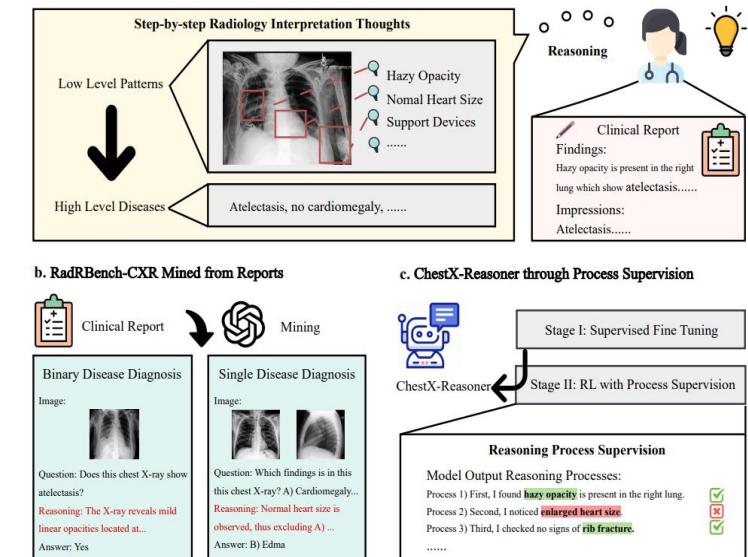
InternVL3-8B **DeepSeek-VL2-Tiny** **Patho-R1-7B**

MLLM for reasoning with process RL [4]

Chest X-ray VQA+Interpretation [4] (2025.05)

ChestX-Reasoner: Advancing Radiology Foundation Models with Reasoning through Step-by-Step Verification

Ziqing Fan^{1,2,*}, Cheng Liang^{2,*}, Chaoyi Wu^{1,2},
Ya Zhang^{1,2}, Yanfeng Wang^{1,2} and Weidi Xie^{1,2}



[1] Lai, Yuxiang, et al. "Med-r1: Reinforcement learning for generalizable medical reasoning in vision-language models." arXiv preprint arXiv:2503.13939 (2025).

[2] Pan, Jiazen, et al. "Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning." arXiv preprint arXiv:2502.19634 (2025).

[3] Zhang, Wenchuan, et al. "Patho-R1: A Multimodal Reinforcement Learning-Based Pathology Expert Reasoner." arXiv preprint arXiv:2505.11404 (2025).

[4] Fan, Ziqing, et al. "ChestX-Reasoner: Advancing Radiology Foundation Models with Reasoning through Step-by-Step Verification." arXiv preprint arXiv:2504.20930 (2025).