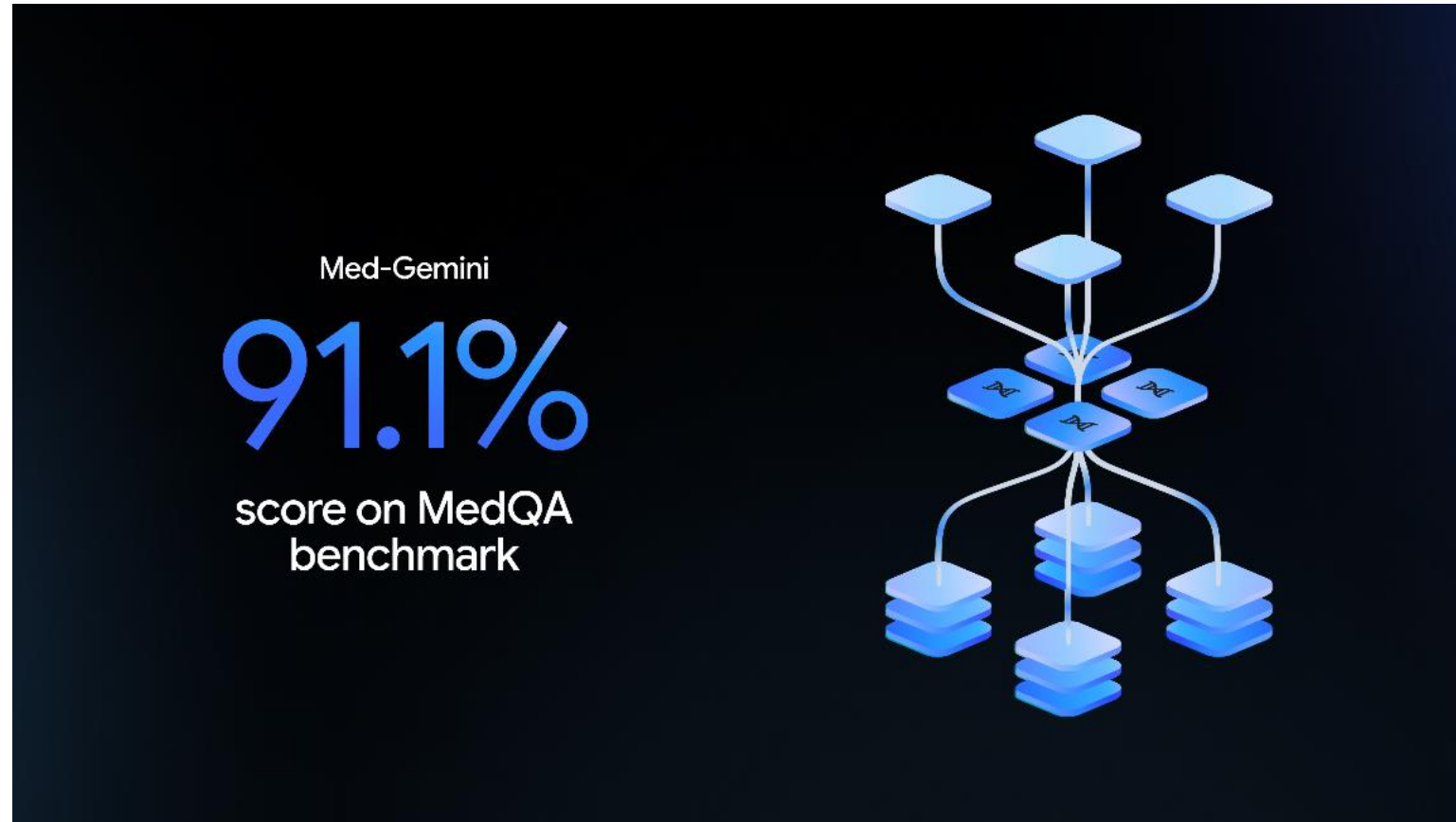


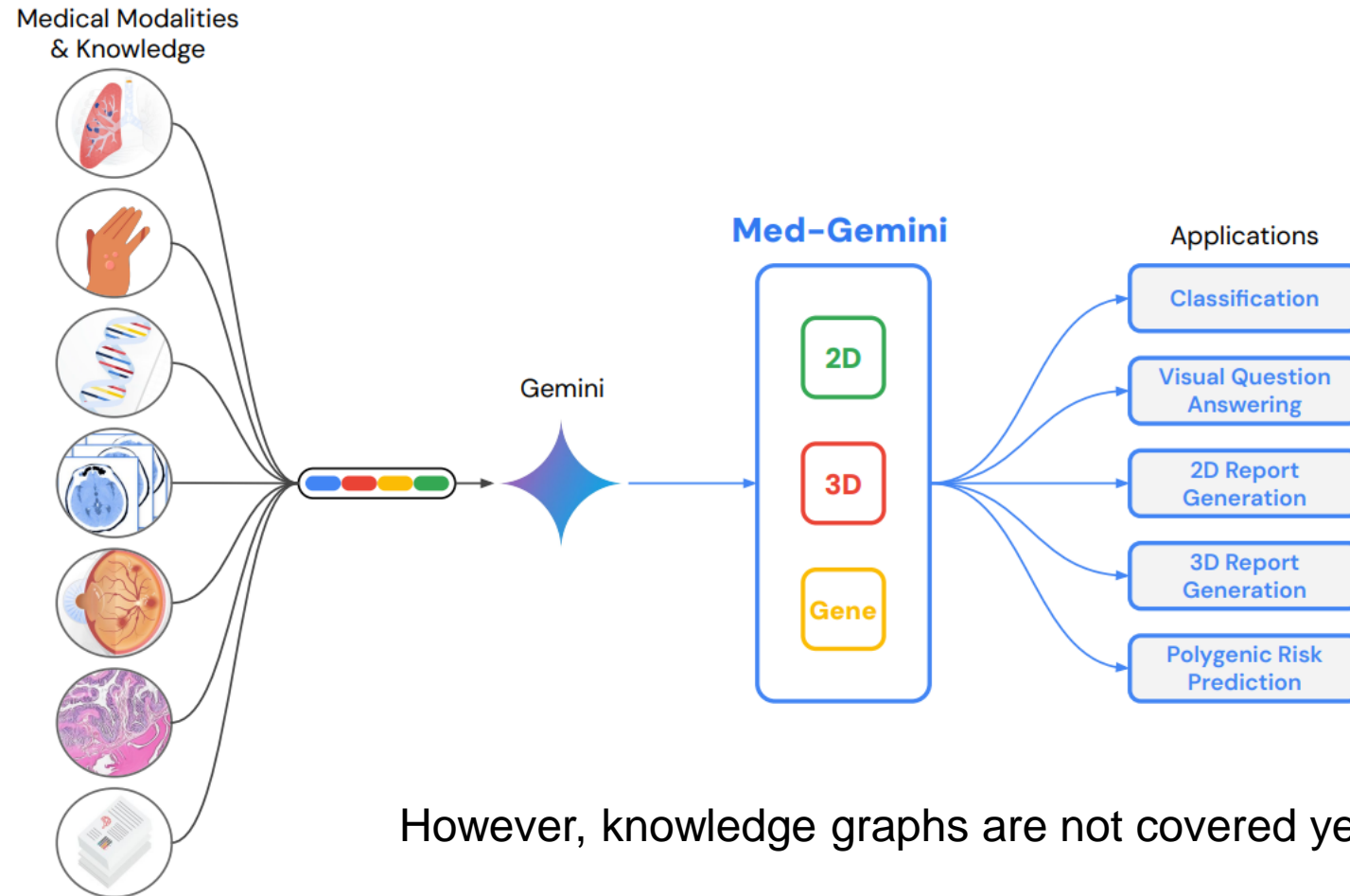
# Foundation Model in Medicine at IJCAI2024 and ACL2024

Yunsoo Kim

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



Med-Gemini can cover not only medical text but also medical images and gene sequence



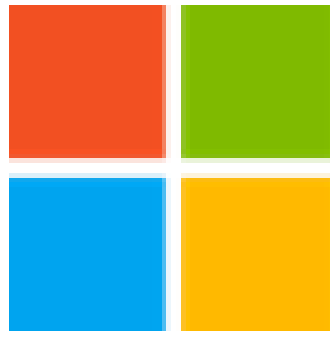
However, knowledge graphs are not covered yet!



# Proprietary models

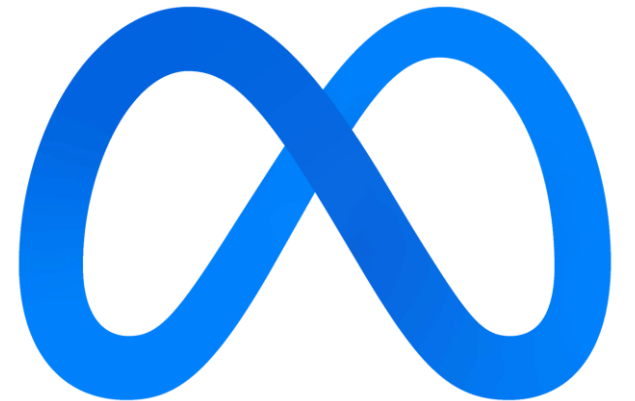
The privacy concerns with the patient data

July 10, 2024

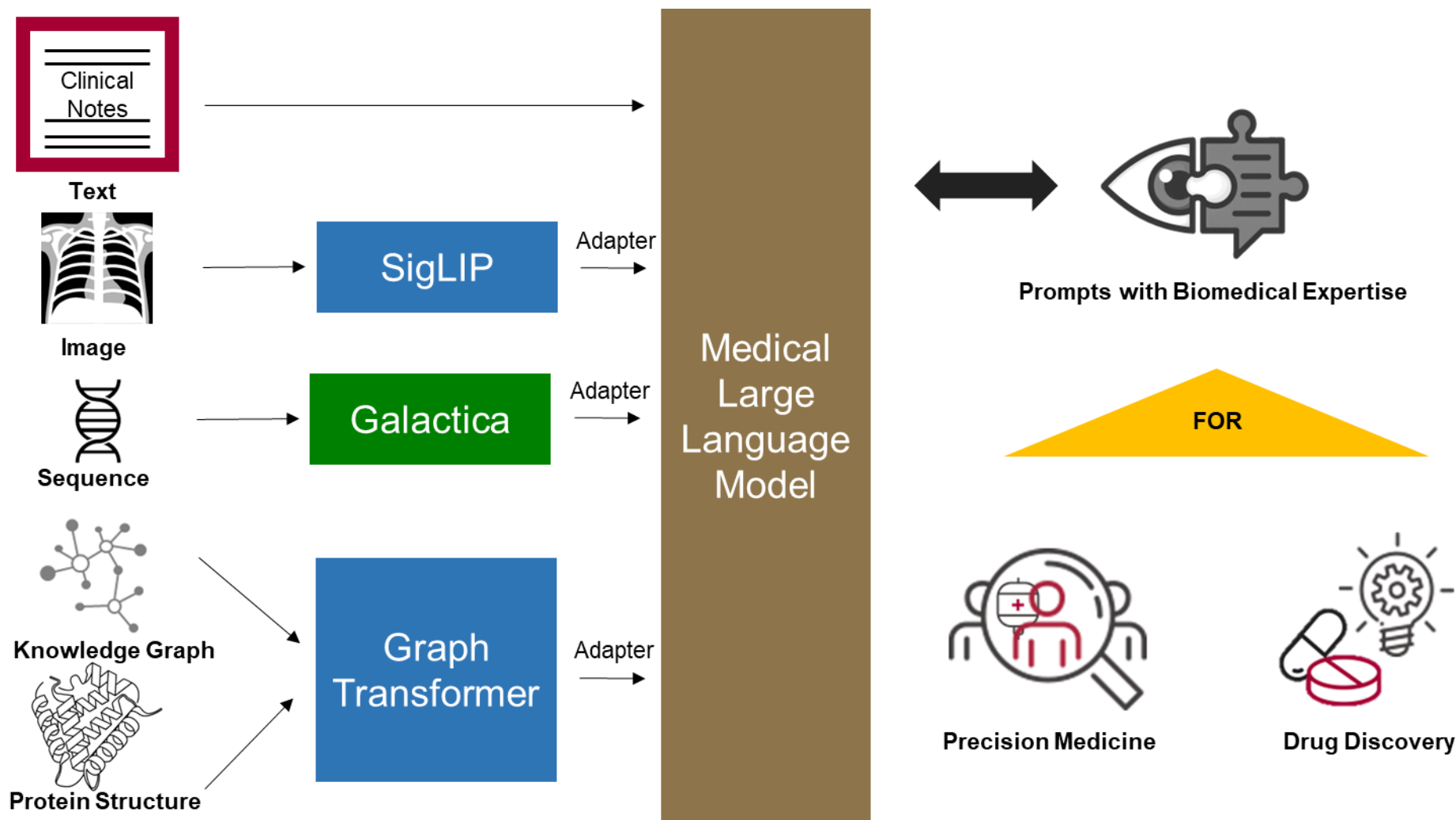


# Open-source LLMs

can be used in hospital settings



### Foundation Model with LLMs for Biomedical graphs



There were LLMs for Biomedical domain. The general consensus is that the models are not rigorously tested for the use in clinical settings.

## Asclepius 7B, 13B

- Supervised Finetuned on 158k synthetic EHR
- Evaluated with Generation of Discharge Summaries with MIMIC-III, MIMIC-IV, i2b2 data
- The evaluation of the generated summaries was done with GPT4 model as the Judge. The points given by GPT4 were compared with clinician's points.

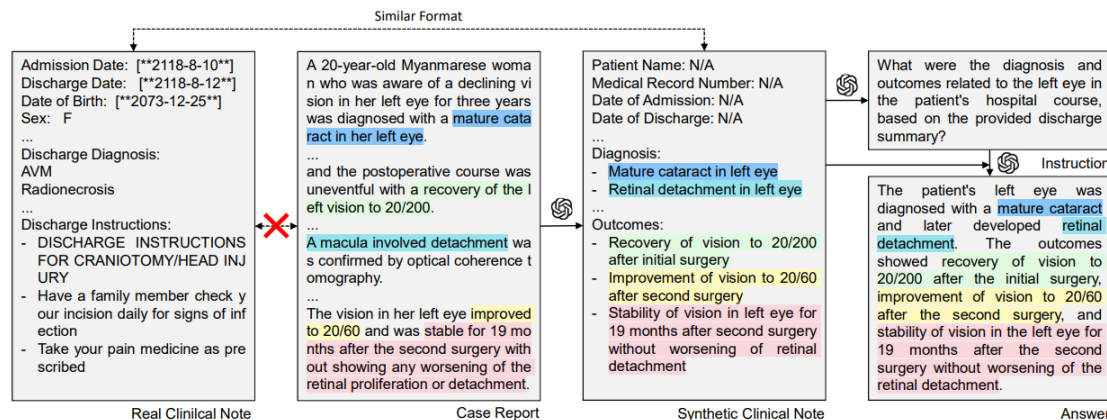


Figure 2: The first column (far left) is a part of the real discharge summary taken from MIMIC-III (Johnson et al., 2016). The second column is a case report from PMC-Patients (Zhao et al., 2023), and the third is the synthetic discharge summary created from this case report. Initially, the case report did not resemble the real clinical note in

## BioMistral 7B

- Pretrained with PMC Open Access Articles
  - Comprehensive experiments including model merging
  - Multilingual capabilities
  - Evaluation on MedQA, MMLU, MedExQA, and PubMedQA

Model Name	Base Model	Model Type	Sequence Length	Download
BioMistral-7B	<a href="#">Mistral-7B-Instruct-v0.1</a>	Further Pre-trained	2048	<a href="#">HuggingFace</a>
BioMistral-7B-DARE	<a href="#">Mistral-7B-Instruct-v0.1</a>	Merge DARE	2048	<a href="#">HuggingFace</a>
BioMistral-7B-TIES	<a href="#">Mistral-7B-Instruct-v0.1</a>	Merge TIES	2048	<a href="#">HuggingFace</a>
BioMistral-7B-SLERP	<a href="#">Mistral-7B-Instruct-v0.1</a>	Merge SLERP	2048	<a href="#">HuggingFace</a>

The gap in the current benchmarks for the comprehensive assessments of LLMs' ability

## MMLU

- Popular knowledge Intensive QA
  - 9 tasks related to medicine

In a genetic test of a newborn, a rare genetic disorder is found that has X-linked recessive transmission. Which of the following statements is likely true regarding the pedigree of this disorder?

- (A) All descendants on the maternal side will have the disorder.
- (B) Females will be approximately twice as affected as males in this family.
- (C) All daughters of an affected male will be affected.**
- (D) There will be equal distribution of males and females affected.

Figure 23: A College Medicine example.

Assessment is simplified to classification problem.

## 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**



Data contamination check using MIN-K% PROB score also confirms that open domain baseline models may have already been exposed to these evaluation datasets.

Dataset	LLaMA2	Mistral	Phi-3
MedQA	45	54	67
MMLU_MED	22	43	43
MedMCQA	14	26	30

Reference model is Phi-2 model. Higher → More likely to be contaminated

## Benchmarks for the comprehensive assessments of LLMs' ability

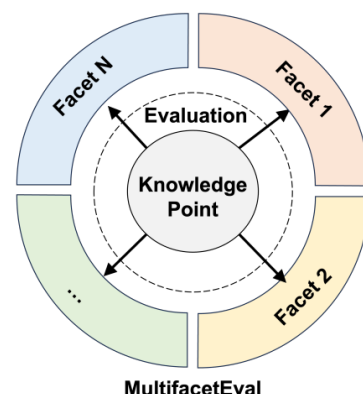
### MultifacetEval

- Revised MedQA to have multifaceted knowledge
  - Comparison: Original Multiple Choice
  - Rectification: Check correctness and revise if needed.
  - Discrimination: Multiple Answers QA
  - Verification: True or False QA

**Multiple-choice Question:**  
**Question:** A 25-year-old Hispanic male presents ... Which of the following would be consistent with this patient's disease?  
**Options:** A: Sympathetic underactivity B: Anti-thyroglobin antibodies  
**C: Exophthalmos** D: Increased TSH release E: Multinucleate giant cells present in the thyroid  
**Label: C**  
**ChatGPT's Answer:** The patient's ... **Therefore, the answer is: C**

↓ Rephrasing

**True-false Question:**  
**Question:** A 25-year-old Hispanic male presents ... Statement: "Exophthalmos would be consistent with this patient's disease.", is the statement above true or false? Please answer true/false.  
**Label: True**  
**ChatGPT's Answer:** Exophthalmos, ... so the statement is **false**



### MedExQA

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

Question	Which of the following is a correct description of a language learning disability?
	Correct Answer: A. It shows difficulties in reading and writing.
Asclepius-13B Score: 0	E. It is a type of specific learning disorder (SLD). ...
Meditron-70B Score: 0	The following is an open-ended question. You should answer it in your own words...
GPT3.5_1106 Score: 1	A. It shows difficulties in reading and writing.  Explanation: A language learning disability typically involves difficulties in reading and writing, such as dyslexia. ... While other options may also be associated with learning disabilities, the primary characteristic of a language learning disability is the difficulty in reading and writing.
MedPhi-2 Score: 1	The answer is A.... Additionally, language learning disorders are not typically caused by visual or hearing impairments, but rather by difficulties with processing and using linguistic information....

Mostly on report generation because of the shared task at the BioNLP workshop.

I also connected with the authors of RadFM, CheXagent, and CXR-LLaVA.

## CXRMate

- Top 1 model in the shared task
  - 150M parameters only
  - Uses the same techniques of LLaMA2
  - Supports multi views
  - Introduced a new loss based on entropy

## MAIRA

- Multimodal model for Chest X-ray Report Generation
  - Supports multi views
  - A model based on vicuna 13B model was 2<sup>nd</sup> place.
  - They are building a new model based on Phi-3 architecture, and told me the result is promising

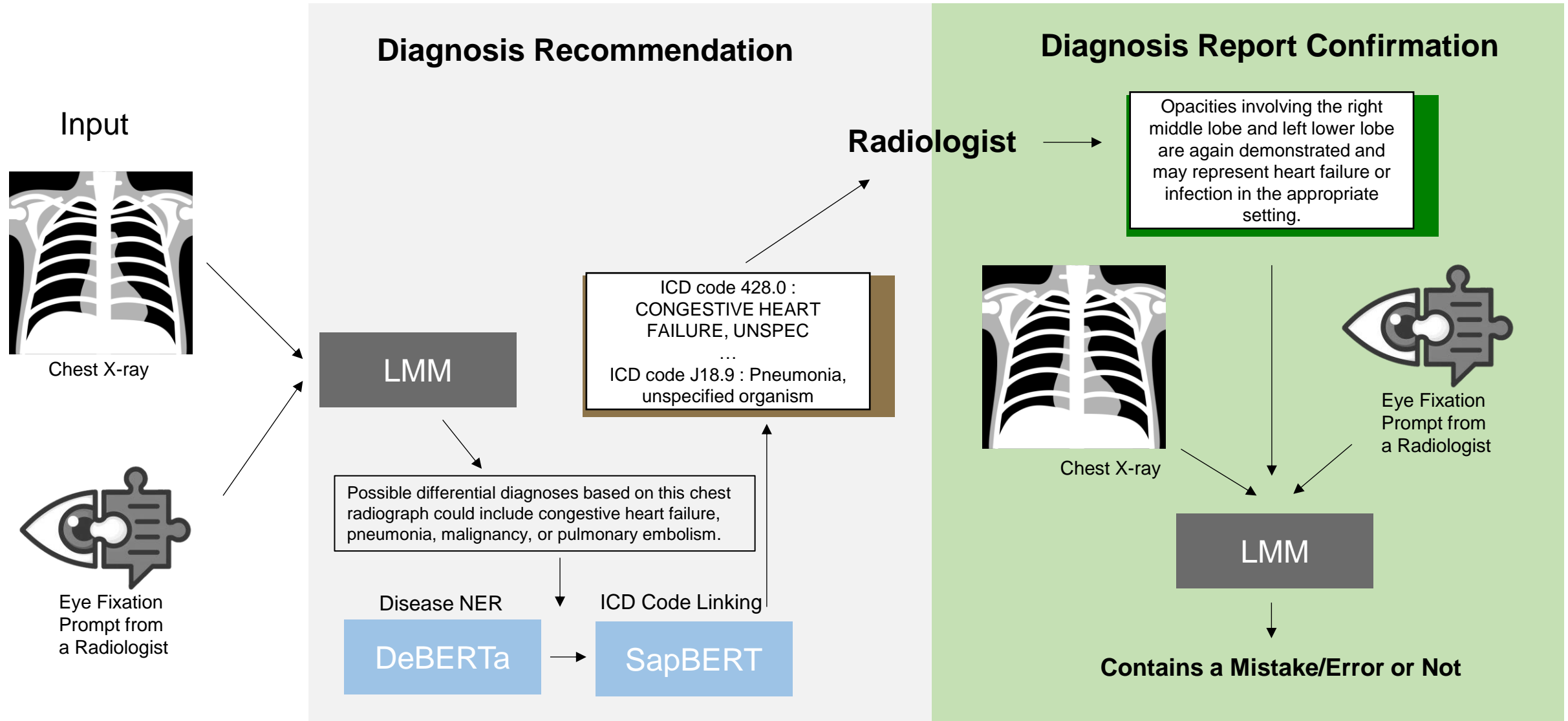
### EAST: Entropy-Augmented Self-Critical Sequence Training

$$L_{EAST}(\theta) = -(r(\mathbf{w}^s) - r(\mathbf{w}^b)) \cdot \log(\pi(\mathbf{w}^s | \mathbf{I}; \theta)) + \lambda \cdot H(\pi)$$

Reward      Radiologist report/ground truth      Sample/baseline report

$$r(\mathbf{w}^{s/b}) = RG_{ER}^{F_1}(\mathbf{w}, \mathbf{w}^{s/b})$$

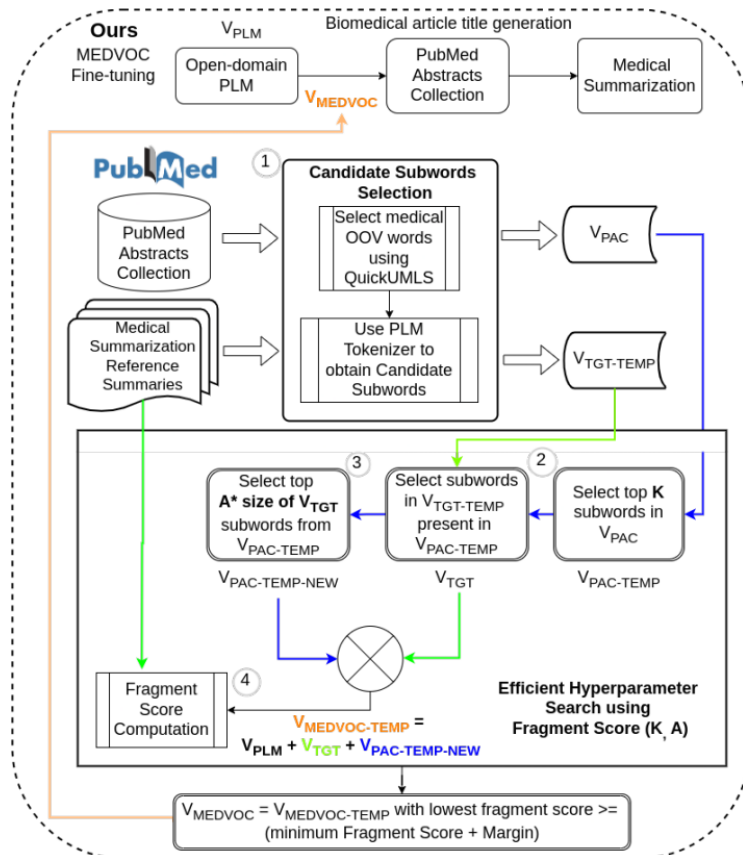
Team/Method	BLEU-4	ROUGE-L	BERTScore	CheXbert-F1	RadGraph-F1
<i>e-Health CSIRO</i>					
EAST	<b>11.68/12.33</b>	26.16/28.32	53.80/50.94	57.49/ <b>56.97</b>	<b>28.67/27.83</b>
SCST	10.25/10.95	26.10/27.34	53.88/50.07	55.78/54.79	27.29/24.97
TF	11.12/9.89	25.43/24.94	51.10/42.49	50.02/47.24	22.99/21.27
<i>Top three teams besides ours</i>					
maira	11.24/11.66	<b>26.58/28.48</b>	<b>54.22/51.62</b>	<b>57.87/53.27</b>	25.48/25.26
airi	9.97/10.91	25.82/27.46	52.42/49.55	54.25/52.32	25.29/24.67
gla-ai4biomedic	7.65/9.60	24.35/25.27	52.69/48.60	46.21/46.74	24.13/22.10



The gap in the current benchmarks for the comprehensive assessments of LLMs' ability

## MEDVOC

- Domain specific vocabulary based summarization



## MedDec

- 10 Medical Decisions annotated with Discharge Summaries
  - The dataset will be available in October.

Mr. [...] is a 61 y/oM with HIV and HCV and Hemophilia  
 [DEFINING PROBLEM] ... with suspicion for diffuse neoplastic process  
 of the liver [DEFINING PROBLEM] ...  
 admitted for biopsy [THERAPEUTIC PROCEDURE RELATED]  
 He was treated with interferon [DRUG RELATED]...  
 He has had multiple imaging that showed multiple focal lesions  
 that were not previously seen [EVALUATING TEST RESULT] ...  
 Past Medical History: Hemophilia [DEFINING PROBLEM] - followed  
 by Dr [...], Drs [...] and [...] \*\* [CONTACT RELATED]

# Thank you