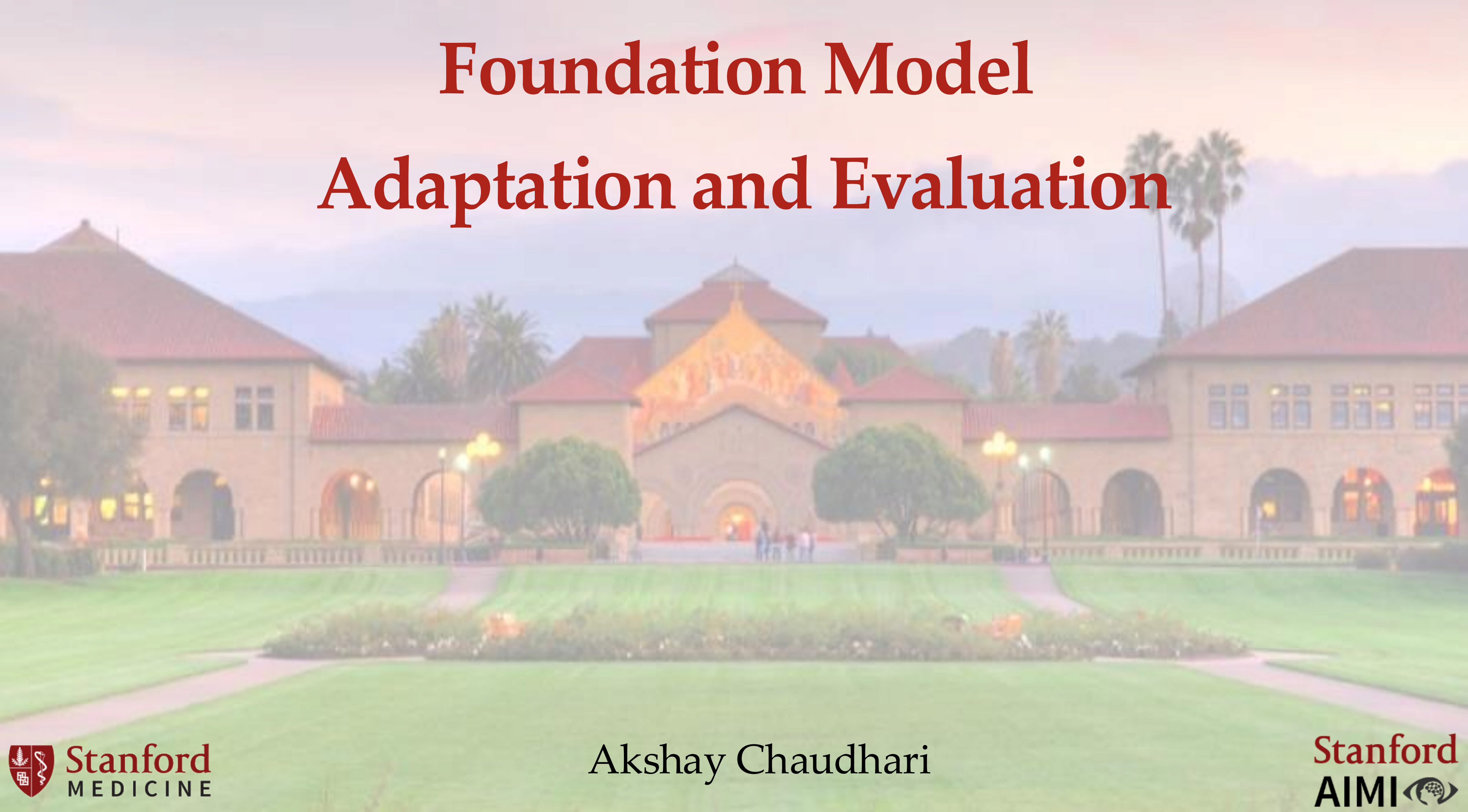
Foundation Model Adaptation and Evaluation

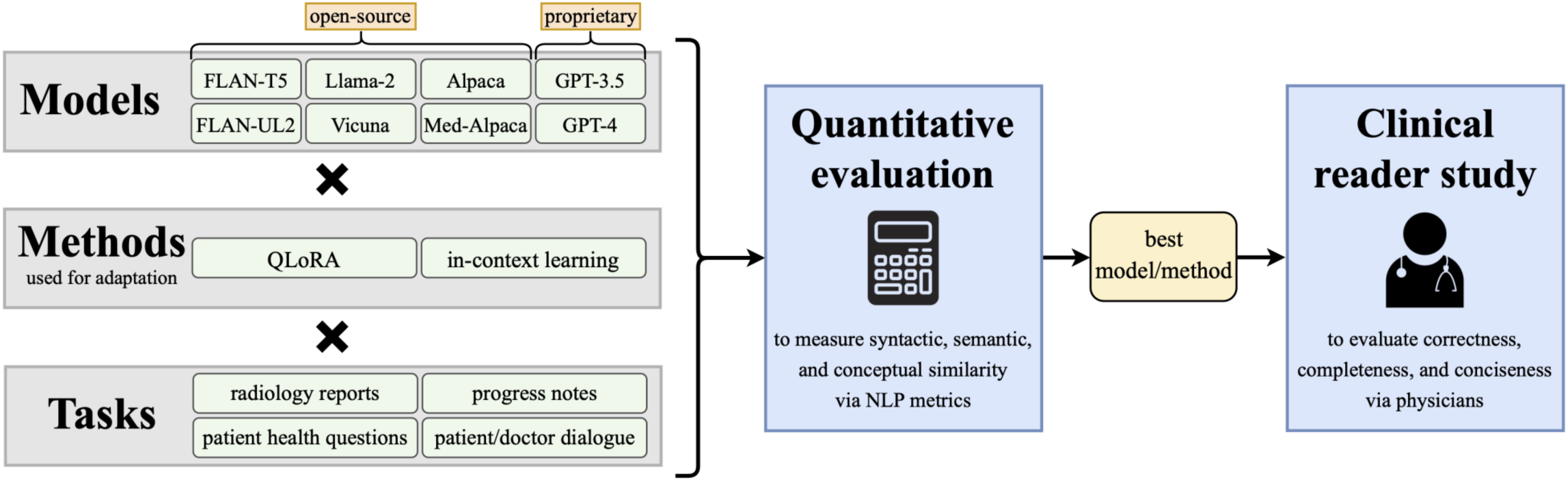


Akshay Chaudhari

Stanford MEDICINE

Stanford AIMI

LLM to Summarize Medical Text



Van Veen et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. Nature Medicine, 2024.

Example Datasets

Radiology Report Findings → Report Impressions

The patient is s/p left frontal craniotomy. A small amount of intracranial gas is seen posterior to the surgical intervention, which could represent postoperative changes. Extensive edema is seen in the left frontal lobe at the site of presumed surgery. Additionally multiple foci of hemorrhage are seen in the region of the left frontal lobe. Midline shift to the right is seen in the frontal region. The ventricles, cisterns, and sulci are unremarkable, without effacement. Comparison with prior studies from outside institution would be helpful in further evaluation of these findings.

1. Left frontal craniotomy.
2. Frontal midline shift to the right.
3. Extensive left frontal lobe edema.
4. Multiple foci of hemorrhage in the right frontal lobe.

Van Veen et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. Nature Medicine, 2024.

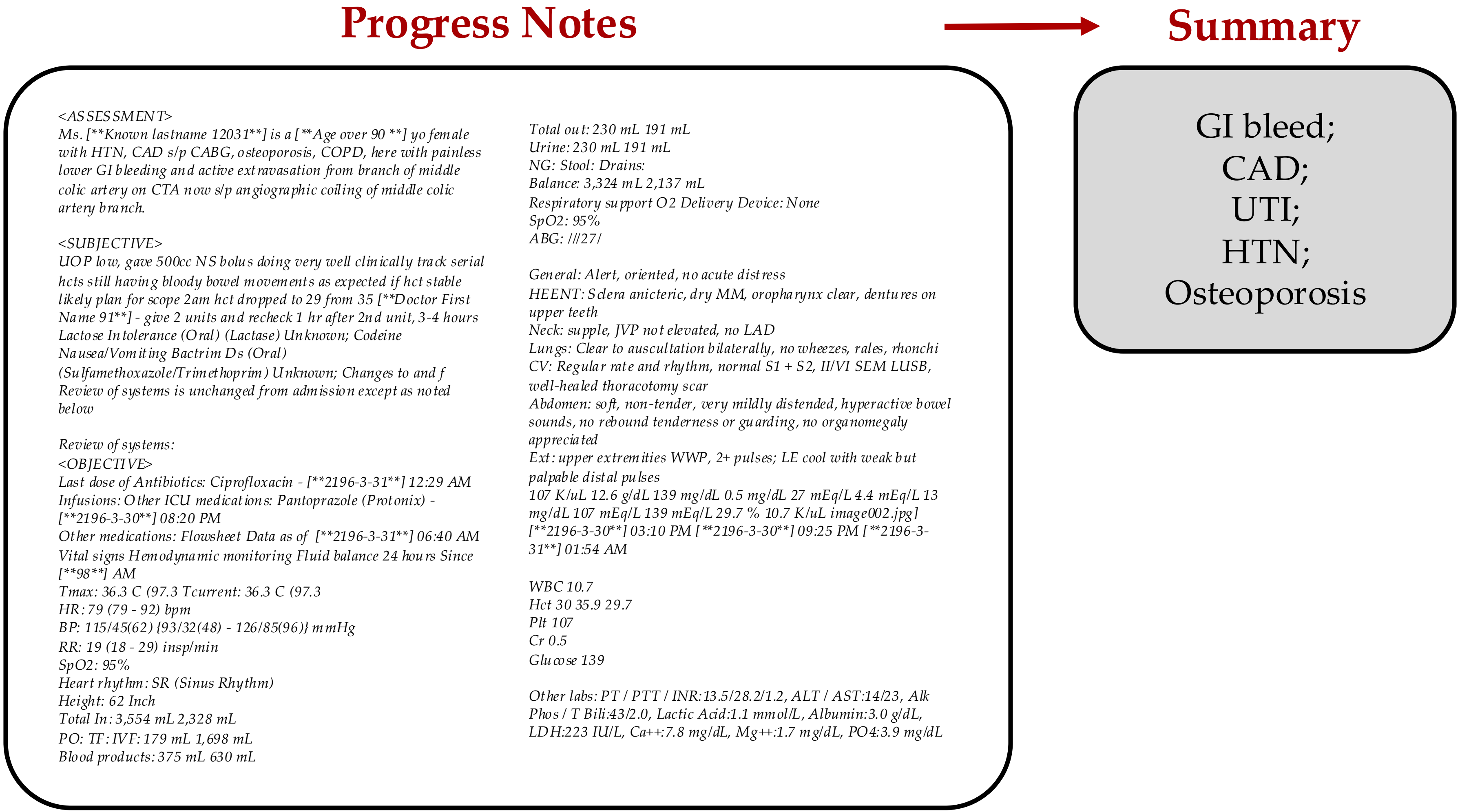
Example Datasets

Patient Questions → Summary

Hello, I have been dealing with trimethylaminuria since I was a child. I have done some of my own research and it looks like not much can be done for this condition. I do not have it all over my body it's only in my armpits. In the past I've gone to doctors and dermatologist they gave me no answers until I looked online today and finally found out what I have. I don't know maybe I'm wrong. But this disease isn't even consider common because no one has done anything about it. I'm sure they're thousands of women with it... Can I be tested for it and help in some kind of way to finding a cure or something? What testing is done for this? And where? Thank you

How can I get tested and treated for trimethylaminuria?

Van Veen et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. Nature Medicine, 2024.



Progress Notes

<ASSESSMENT>
Ms. [**Known lastname 12031**] is a [**Age over 90 **] yo female with HTN, CAD s/p CABG, osteoporosis, COPD, here with painless lower GI bleeding and active extravasation from branch of middle colic artery on CTA now s/p angiographic coiling of middle colic artery branch.

<SUBJECTIVE>
UOP low, gave 500cc NS bolus doing very well clinically track serial hcts still having bloody bowel movements as expected if hct stable likely plan for scope 2am hct dropped to 29 from 35 [**Doctor First Name 91**] - give 2 units and recheck 1 hr after 2nd unit, 3-4 hours
Lactose Intolerance (Oral) (Lactase) Unknown; Codeine
Nausea/Vomiting Bactrim Ds (Oral)
(Sulfamethoxazole/Trimethoprim) Unknown; Changes to and f
Review of systems is unchanged from admission except as noted below

Review of systems:
<OBJECTIVE>
Last dose of Antibiotics: Ciprofloxacin - [**2196-3-31**] 12:29 AM
Infusions: Other ICU medications: Pantoprazole (Protonix) - [**2196-3-30**] 08:20 PM
Other medications: Flowsheet Data as of [**2196-3-31**] 06:40 AM
Vital signs Hemodynamic monitoring Fluid balance 24 hours Since [**98**] AM
Tmax: 36.3 C (97.3 Tcurrent: 36.3 C (97.3
HR: 79 (79 - 92) bpm
BP: 115/45(62) {93/32(48) - 126/85(96)} mmHg
RR: 19 (18 - 29) insp/min
SpO2: 95%
Heart rhythm: SR (Sinus Rhythm)
Height: 62 Inch
Total In: 3,554 mL 2,328 mL
PO: TF: IVF: 179 mL 1,698 mL
Blood products: 375 mL 630 mL
Total out: 230 mL 191 mL
Urine: 230 mL 191 mL
NG: Stool: Drains:
Balance: 3,324 mL 2,137 mL
Respiratory support O2 Delivery Device: None
SpO2: 95%
ABG: ///27/

General: Alert, oriented, no acute distress
HEENT: Sclera anicteric, dry MM, oropharynx clear, dentures on upper teeth
Neck: supple, JVP not elevated, no LAD
Lungs: Clear to auscultation bilaterally, no wheezes, rales, rhonchi
CV: Regular rate and rhythm, normal S1 + S2, II/VI SEM LUSB, well-healed thoracotomy scar
Abdomen: soft, non-tender, very mildly distended, hyperactive bowel sounds, no rebound tenderness or guarding, no organomegaly appreciated
Ext: upper extremities WWP, 2+ pulses; LE cool with weak but palpable distal pulses
107 K/uL 12.6 g/dL 139 mg/dL 0.5 mg/dL 27 mEq/L 4.4 mEq/L 13 mg/dL 107 mEq/L 139 mEq/L 29.7 % 10.7 K/uL image002.jpg] [**2196-3-30**] 03:10 PM [**2196-3-30**] 09:25 PM [**2196-3-31**] 01:54 AM

WBC 10.7
Hct 30 35.9 29.7
Plt 107
Cr 0.5
Glucose 139

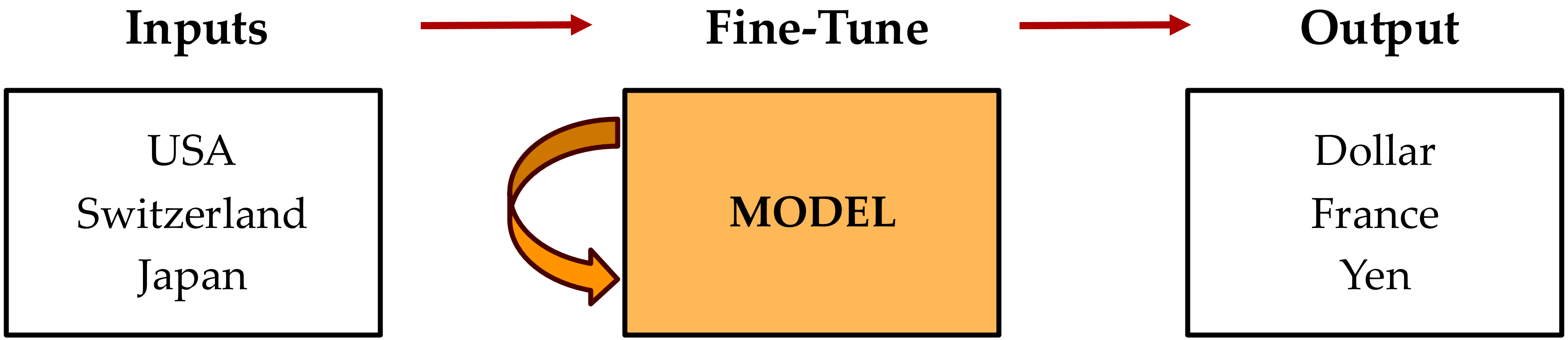
Other labs: PT / PTT / INR:13.5/28.2/1.2, ALT / AST:14/23, Alk Phos / T Bili:43/2.0, Lactic Acid:1.1 mmol/L, Albumin:3.0 g/dL, LDH:223 IU/L, Ca++:7.8 mg/dL, Mg++:1.7 mg/dL, PO4:3.9 mg/dL

Summary

GI bleed;
CAD;
UTI;
HTN;
Osteoporosis

Van Veen et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. Nature Medicine, 2024.

Supervised Finetuning of LLMs



Inputs

USA
Switzerland
Japan

Fine-Tune

MODEL

Output

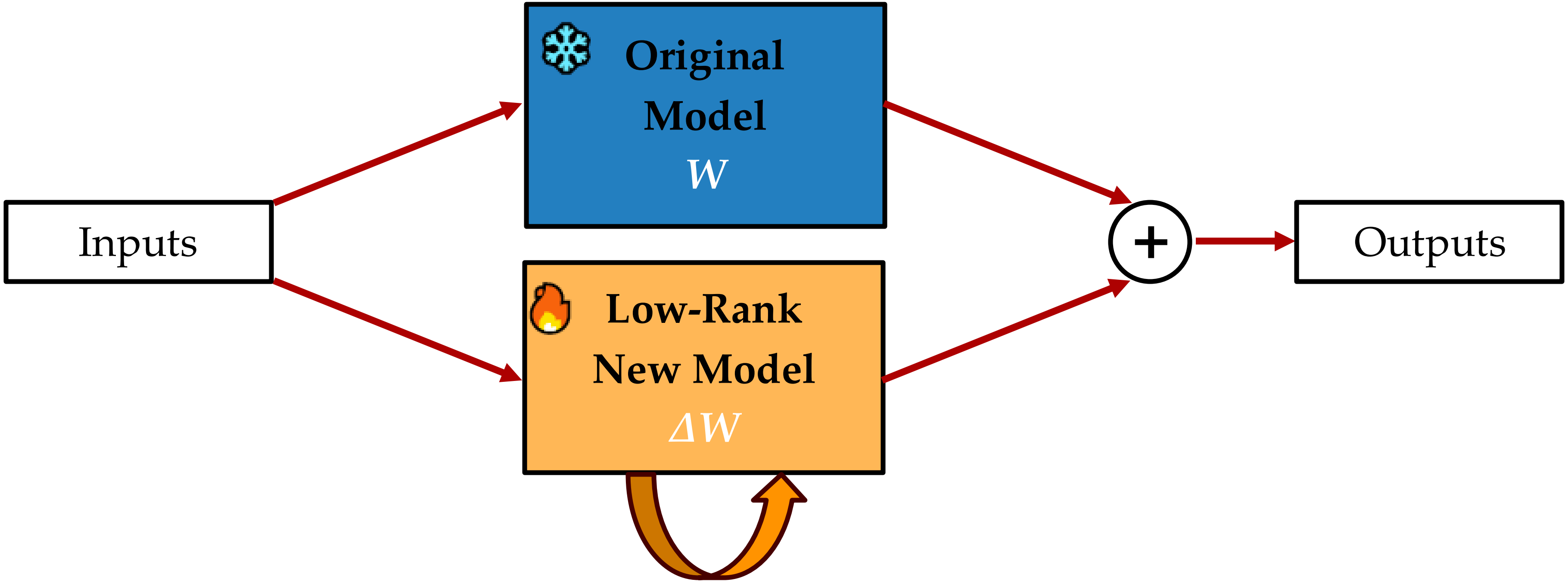
Dollar
France
Yen

Model Weights $= W$

Fine-Tuning $= W + \Delta W$

It is challenging to fine-tune billion+ parameter LLMs!

Low Rank Adaptation



Hu et al. LoRA: Low-Rank Adaptation of Large Language Models. ICLR 2022

In Context Learning

- Simply pass training examples as inputs in prompts

Complete this sentence

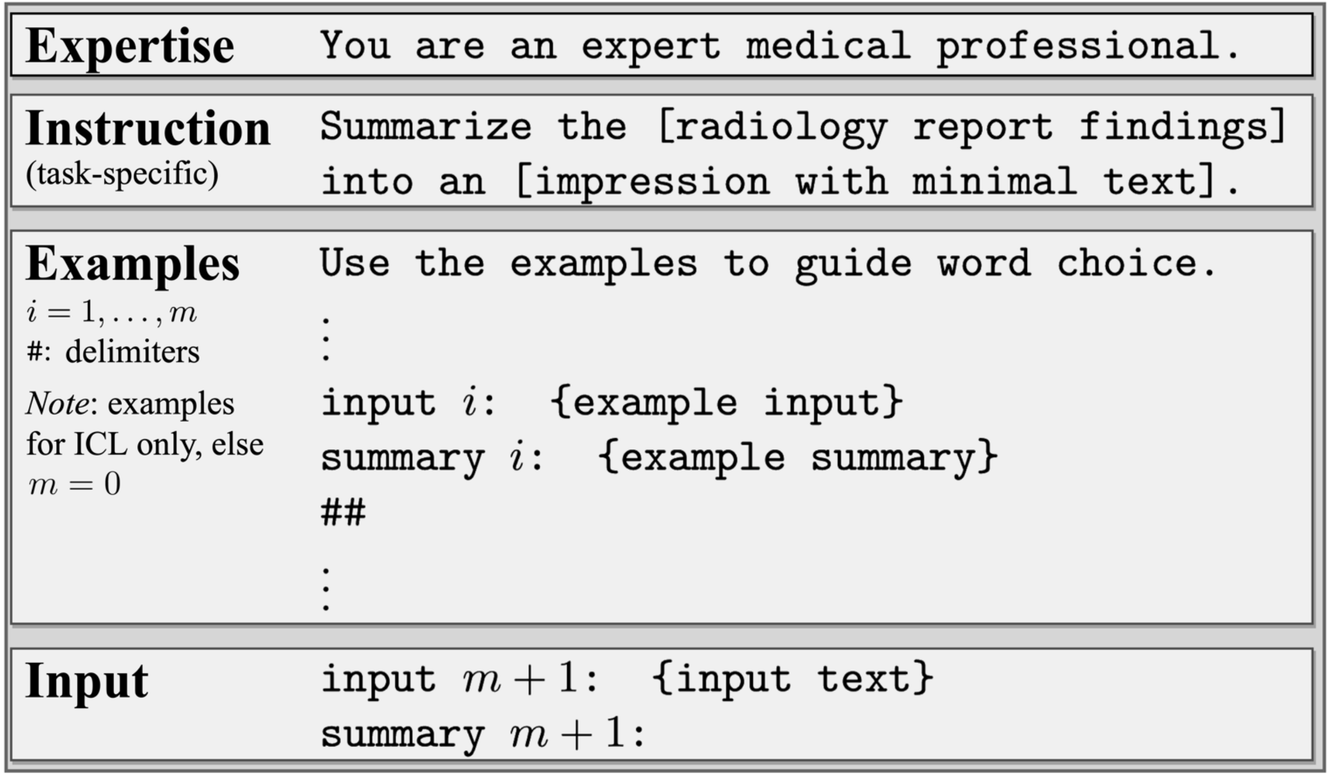
USA: Dollar,

Switzerland: Franc,

Japan: Yen,

Denmark: _____

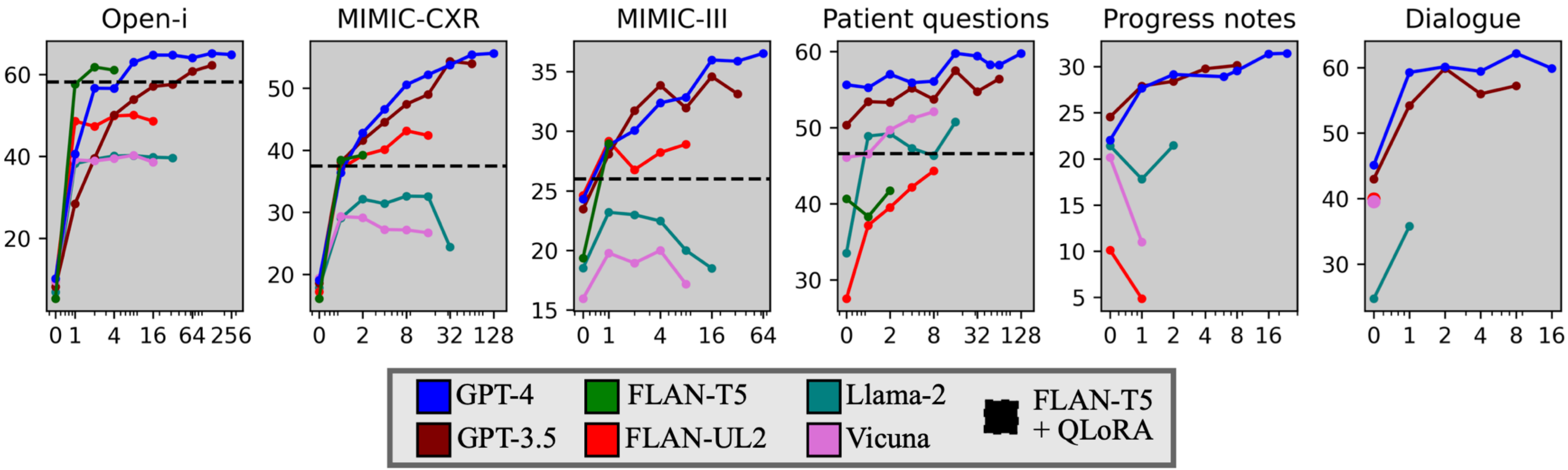
Prompt Anatomy



Expertise	You are an expert medical professional.
Instruction (task-specific)	Summarize the [radiology report findings] into an [impression with minimal text].
Examples $i = 1, \ldots, m$ #: delimiters *Note*: examples for ICL only, else $m = 0$	Use the examples to guide word choice. ⋮ input i: {example input} summary i: {example summary} ## ⋮
Input	input $m+1$: {input text} summary $m+1$:

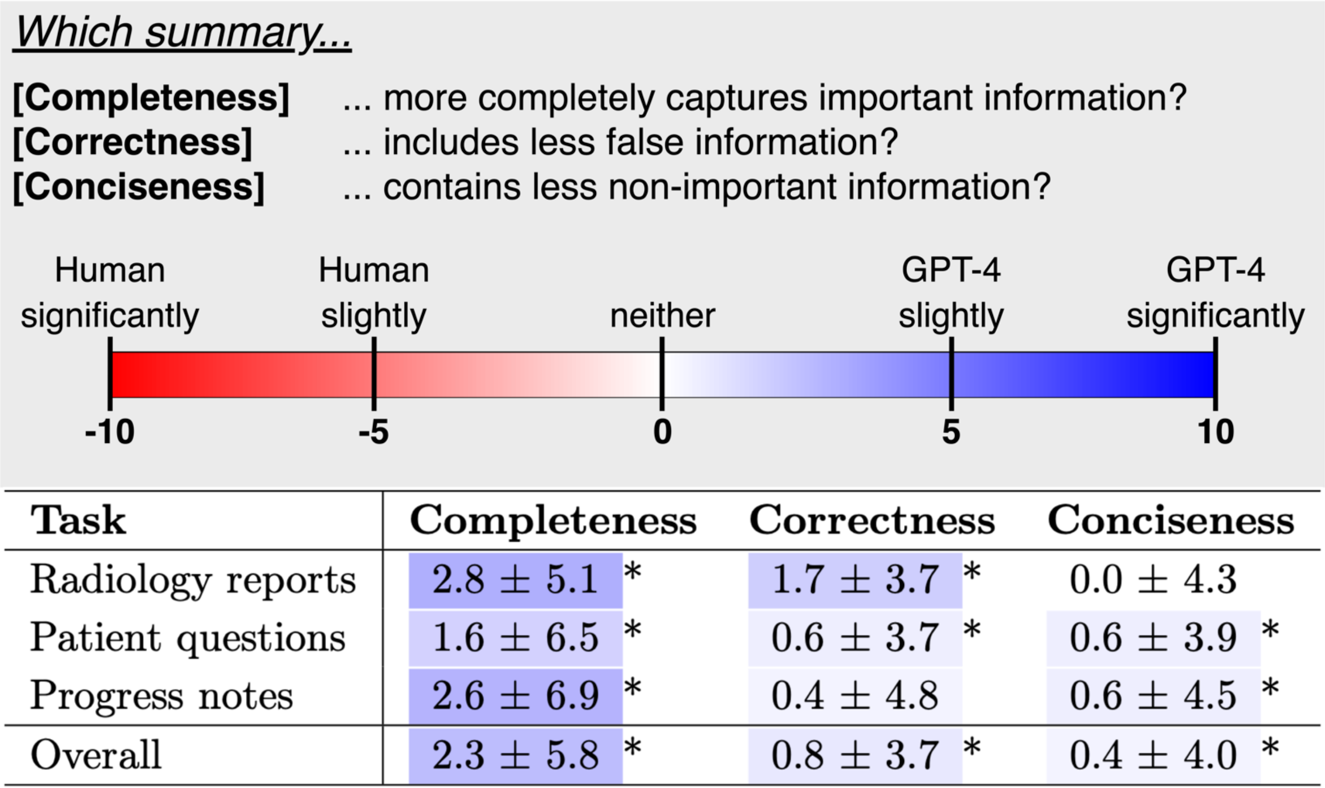
Van Veen et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. Nature Medicine, 2024.

In Context Learning



Van Veen et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. Nature Medicine (accepted), 2024.

Reader Study Results



Which summary...

[Completeness] ... more completely captures important information?
[Correctness] ... includes less false information?
[Conciseness] ... contains less non-important information?

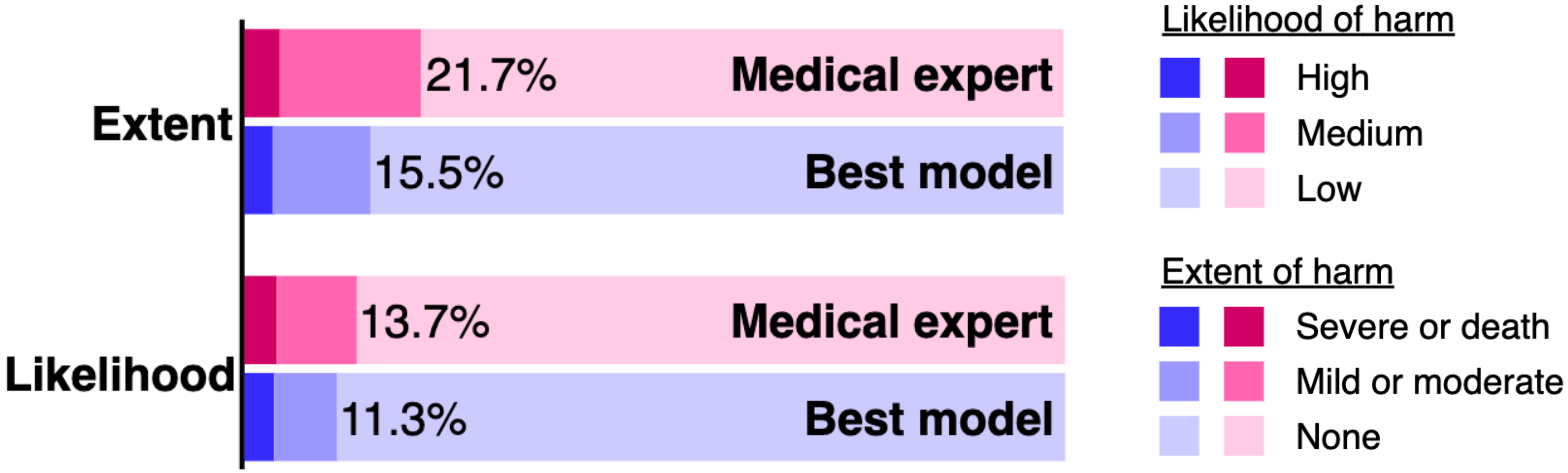
Human significantly
Human slightly
neither
GPT-4 slightly
GPT-4 significantly
-10
-5
0
5
10

Task	Completeness	Correctness	Conciseness
Radiology reports	2.8 ± 5.1 *	1.7 ± 3.7 *	0.0 ± 4.3
Patient questions	1.6 ± 6.5 *	0.6 ± 3.7 *	0.6 ± 3.9 *
Progress notes	2.6 ± 6.9 *	0.4 ± 4.8	0.6 ± 4.5 *
Overall	2.3 ± 5.8 *	0.8 ± 3.7 *	0.4 ± 4.0 *

Van Veen et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. Nature Medicine (accepted), 2024.

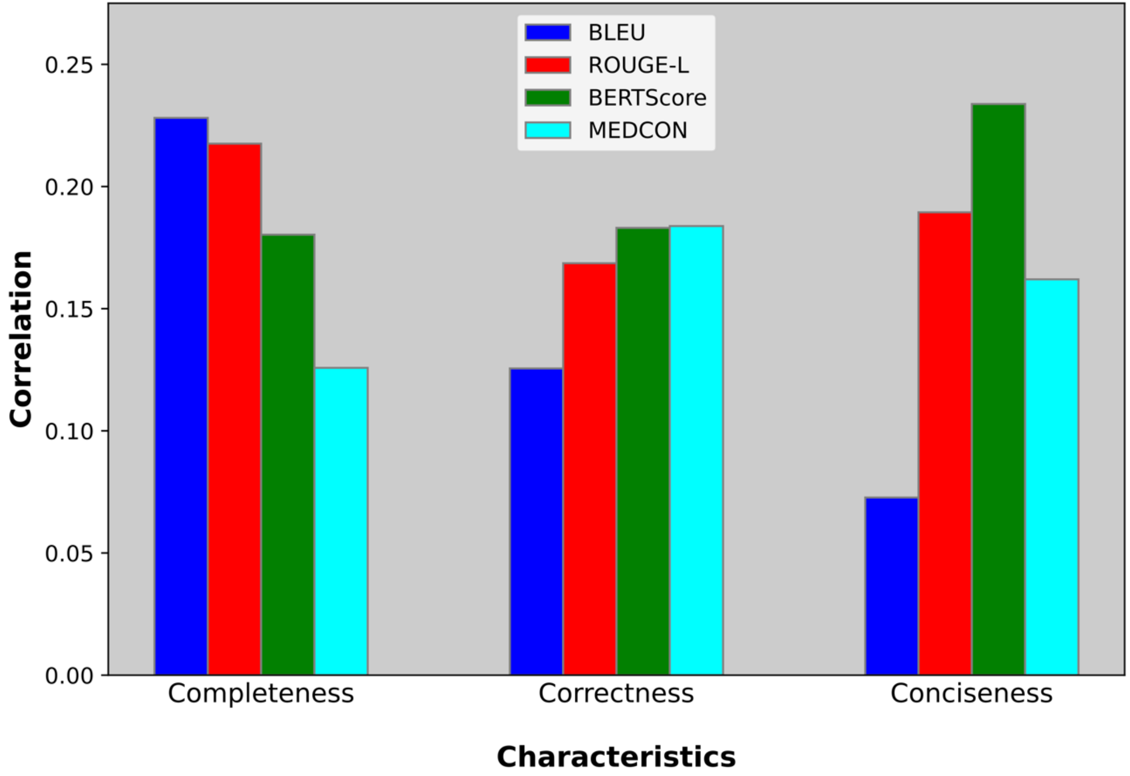
Reader Study Results

- Where either the human/LLM output was preferred, if the other inferior summary were to be used…



Van Veen et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. Nature Medicine (accepted), 2024.

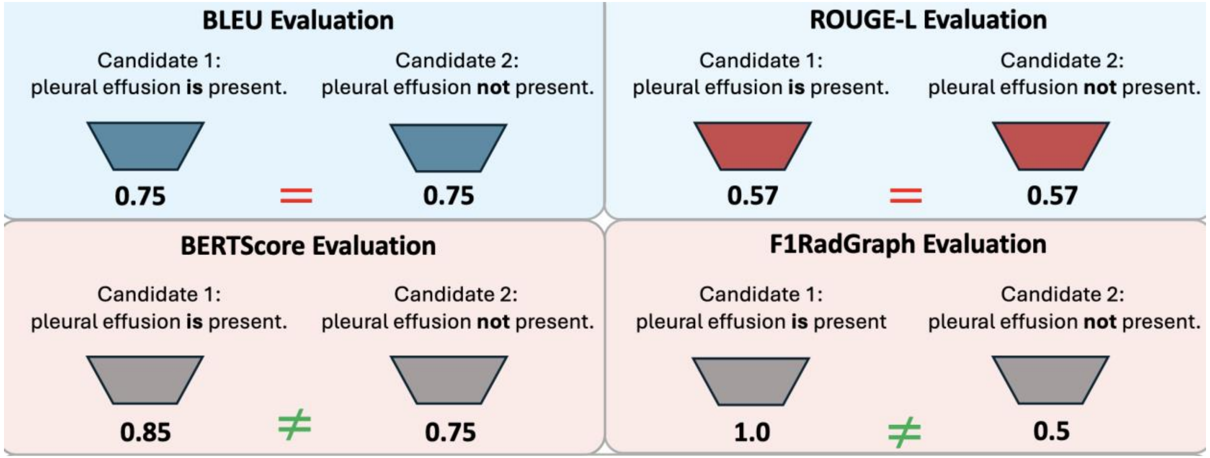
Metrics Correlation



Van Veen et al. Clinical Text Summarization: Adapting Large Language Models Can Outperform Human Experts. Nature Medicine (accepted), 2024.

Imperfect Metrics

- Reference statement: *“Pleural effusion present”*
- Compared to two candidate outputs


BLEU Evaluation
Candidate 1: pleural effusion **is** present.
Candidate 2: pleural effusion **not** present.
0.75
=
0.75
ROUGE-L Evaluation
Candidate 1: pleural effusion **is** present.
Candidate 2: pleural effusion **not** present.
0.57
=
0.57
BERTScore Evaluation
Candidate 1: pleural effusion **is** present.
Candidate 2: pleural effusion **not** present.
0.85
≠
0.75
F1RadGraph Evaluation
Candidate 1: pleural effusion **is** present
Candidate 2: pleural effusion **not** present.
1.0
≠
0.5

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.*

Ostmeier et al. GREEN: Generative Radiology Report Evaluation and Error Notation. EMNLP Findings 2024.

LLM Training Reminder

Self-supervised Pretraining

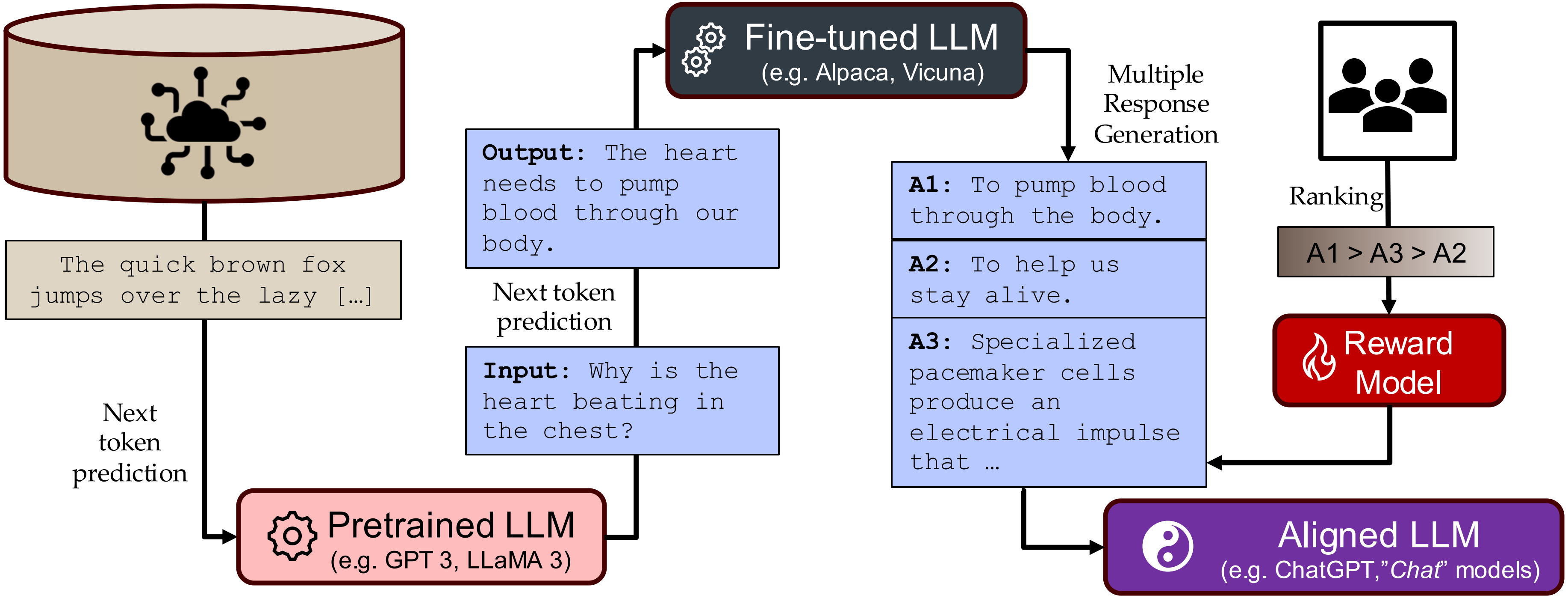
Low-quality data
> 15 trillion tokens

Instruction Tuning

High-quality data
10k-100k pairs/triplets

Alignment Tuning

Preference data
100k-1M ratings



```
The quick brown fox
jumps over the lazy […]
```

Next token prediction

Pretrained LLM
(e.g. GPT 3, LLaMA 3)

```
Input: Why is the
heart beating in
the chest?
```

Next token prediction

```
Output: The heart
needs to pump
blood through our
body.
```

Fine-tuned LLM
(e.g. Alpaca, Vicuna)

Multiple Response Generation

```
A1: To pump blood
through the body.
```

```
A2: To help us
stay alive.
```

```
A3: Specialized
pacemaker cells
produce an
electrical impulse
that …
```

Ranking

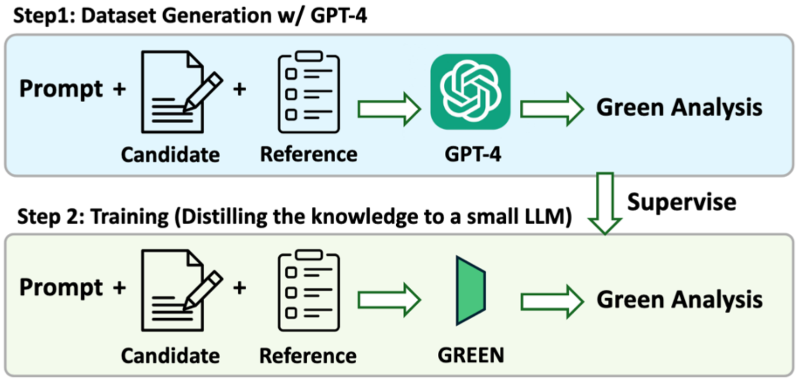
A1 > A3 > A2

Reward Model

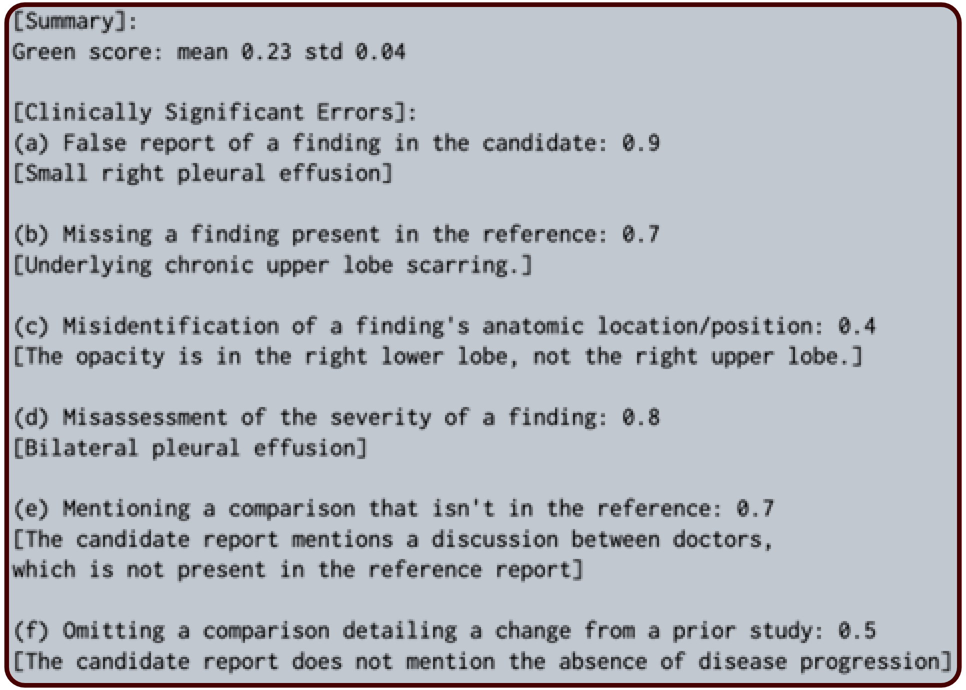
Aligned LLM
(e.g. ChatGPT,"*Chat*" models)

GREEN Metric

- Quantitative metrics + error summary



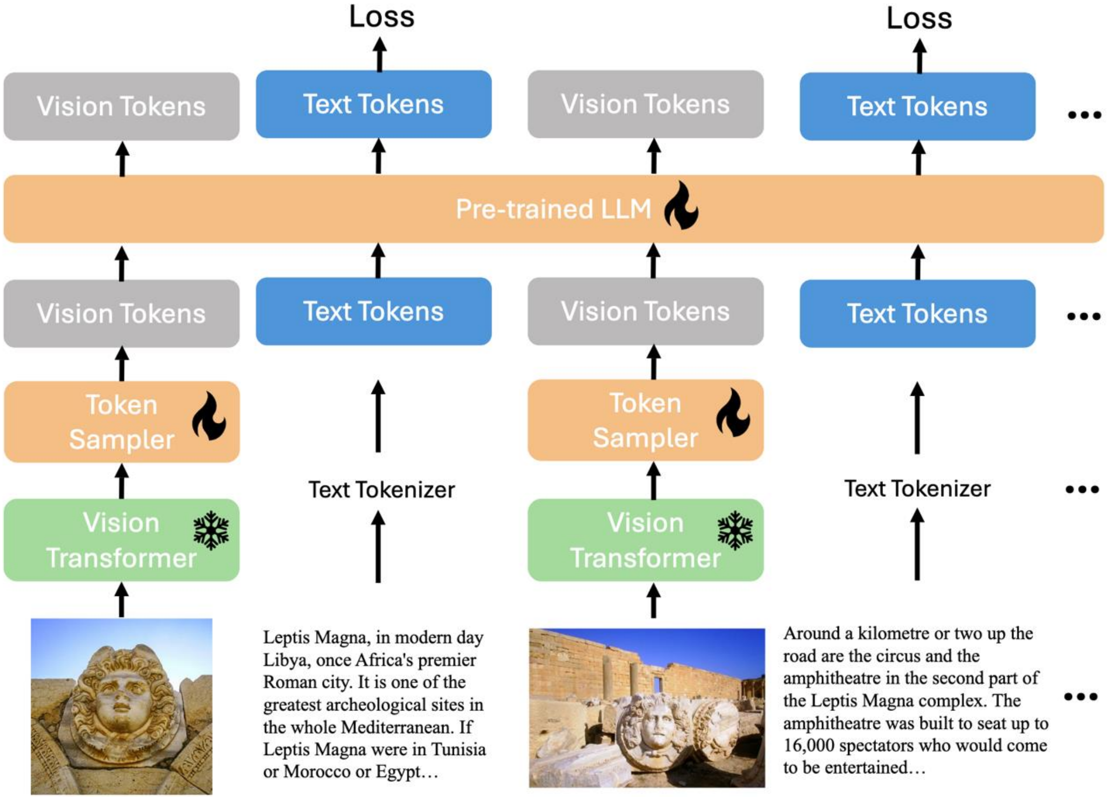
$$\text{GREEN} = \frac{\#\text{ matched findings}}{\#\text{ matched findings} + \sum_{i=(a)}^{(f)} \#\text{ error}_{\text{sig.},i}}$$



```
[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]
```

Ostmeier et al. GREEN: Generative Radiology Report Evaluation and Error Notation. EMNLP Findings 2024.

Visual Instruction Tuning

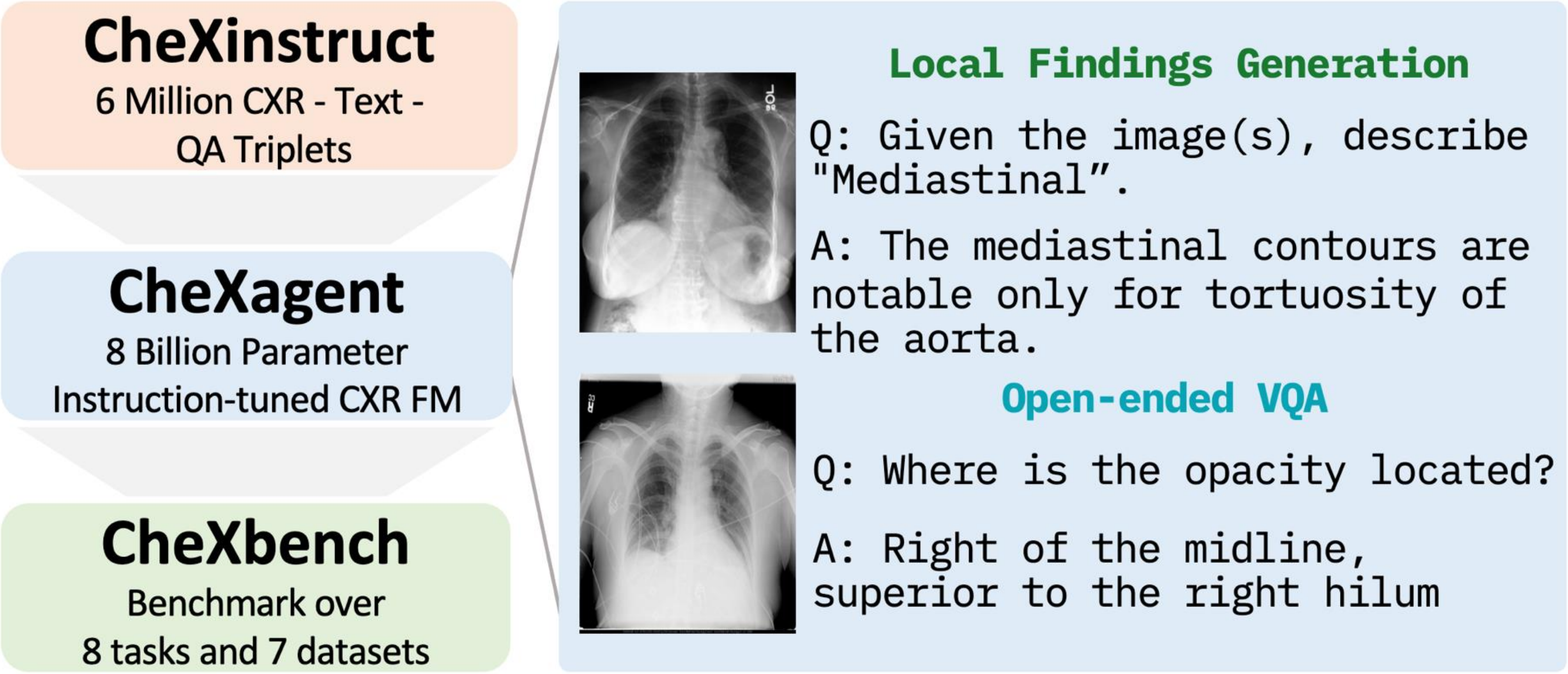


Leptis Magna, in modern day Libya, once Africa's premier Roman city. It is one of the greatest archeological sites in the whole Mediterranean. If Leptis Magna were in Tunisia or Morocco or Egypt…

Around a kilometre or two up the road are the circus and the amphitheatre in the second part of the Leptis Magna complex. The amphitheatre was built to seat up to 16,000 spectators who would come to be entertained…

Xue et al. xGen-MM (BLIP-3): A Family of Open Large Multimodal Models, 2024.

CheXagent Radiology FM



CheXinstruct

6 Million CXR - Text - QA Triplets

CheXagent

8 Billion Parameter Instruction-tuned CXR FM

CheXbench

Benchmark over 8 tasks and 7 datasets

Local Findings Generation

Q: Given the image(s), describe "Mediastinal”.

A: The mediastinal contours are notable only for tortuosity of the aorta.

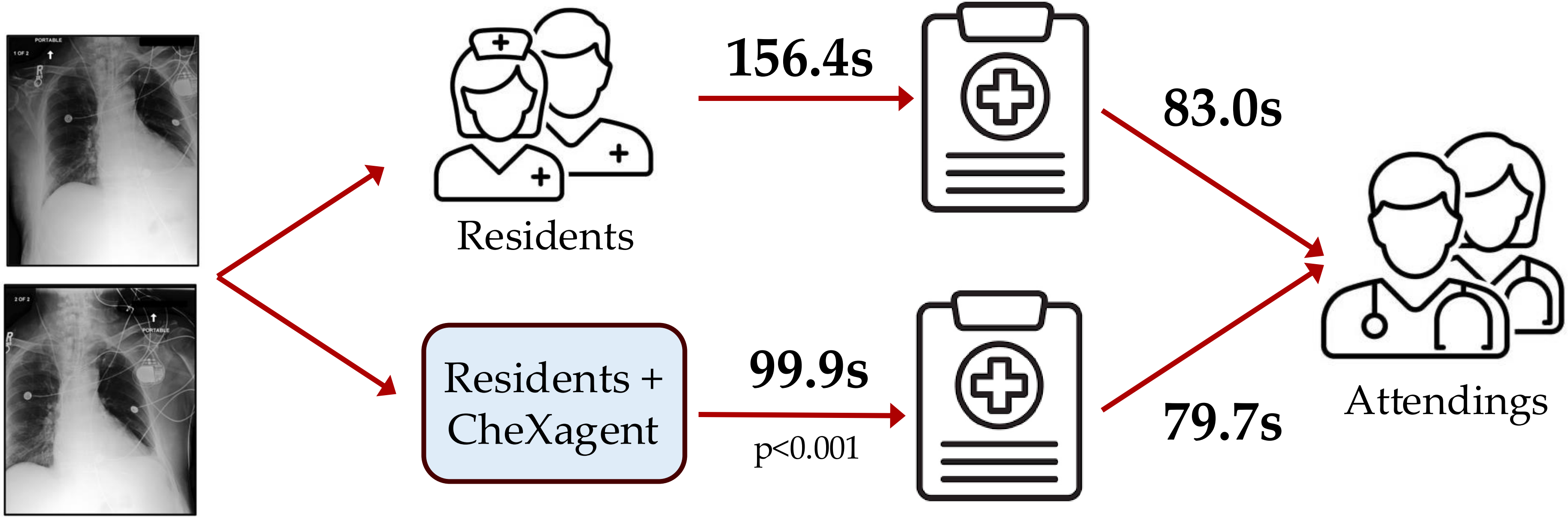
Open-ended VQA

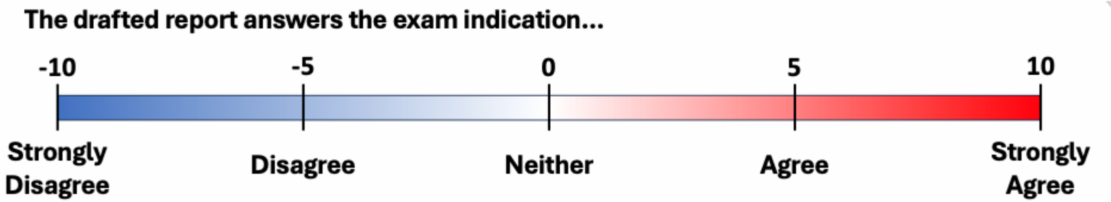
Q: Where is the opacity located?

A: Right of the midline, superior to the right hilum

Chen Z*, Varma M*, Delbrouck JB* et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024

CheXagent: Radiology Report Generation




The drafted report answers the exam indication...

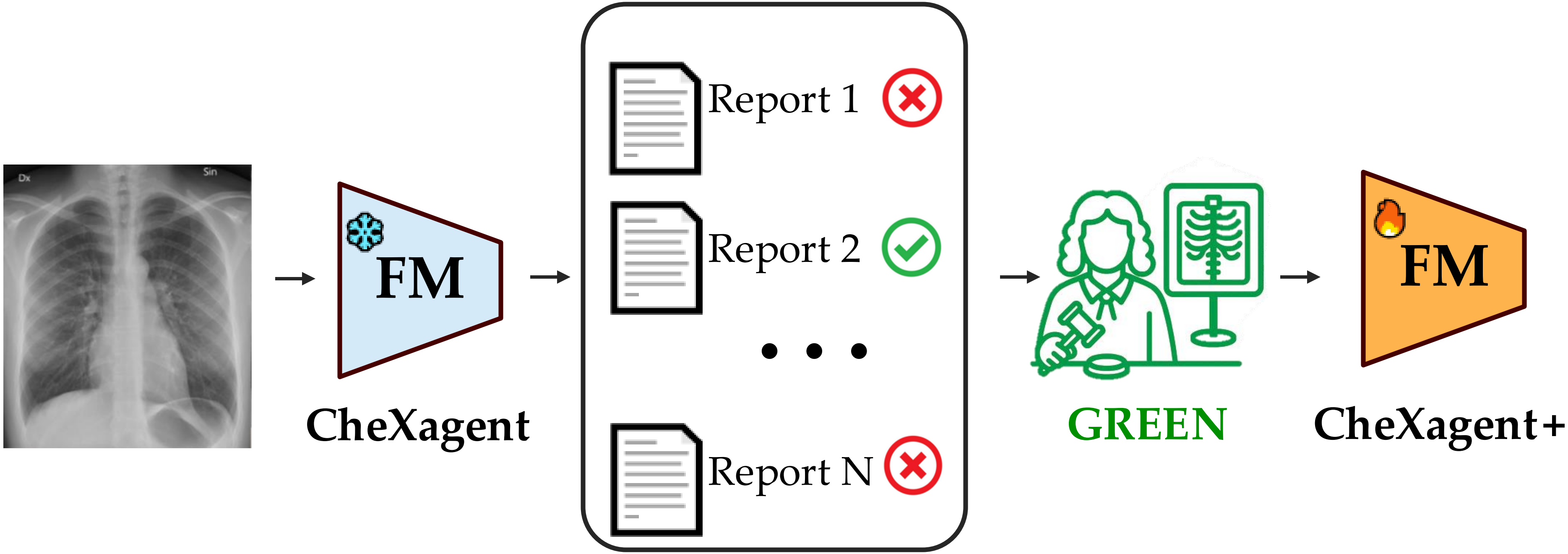
Residents Rating CheXagent: **5.3** ± 6.0

Attendings rating CheXagent: **4.6** ± 5.9

Attendings rating Residents: **5.6** ± 5.4

Chen Z*, Varma M*, Delbrouck JB* et al. CheXagent: Towards a Foundation Model for Chest X-Ray Interpretation. arXiv 2024

Reinforcement Learning w AI Feedback



FM

CheXagent

Report 1 ✗

Report 2 ✓

• • •

Report N ✗

GREEN

FM

CheXagent+

10% improvement without any radiologist feedback

Hein et al. CheXalign: Preference fine-tuning in chest X-ray interpretation models without human feedback. ACL 2025

Questions?

akshaysc@stanford.edu