GPT-4V *cannot* generate radiology report yet

Chacha Chen[1], Yuyang Jiang[1], Dang Nyugen[1]
Benjamin Mervak[2], Chenhao Tan[1]

[1] University of Chicago, [2] University of Michigan

Contact: chacha@uchicago

RSNA® 2023

LEADING THROUGH CHANGE

Preparing Radiology Trainees for AI and ChatGPT

RSNA

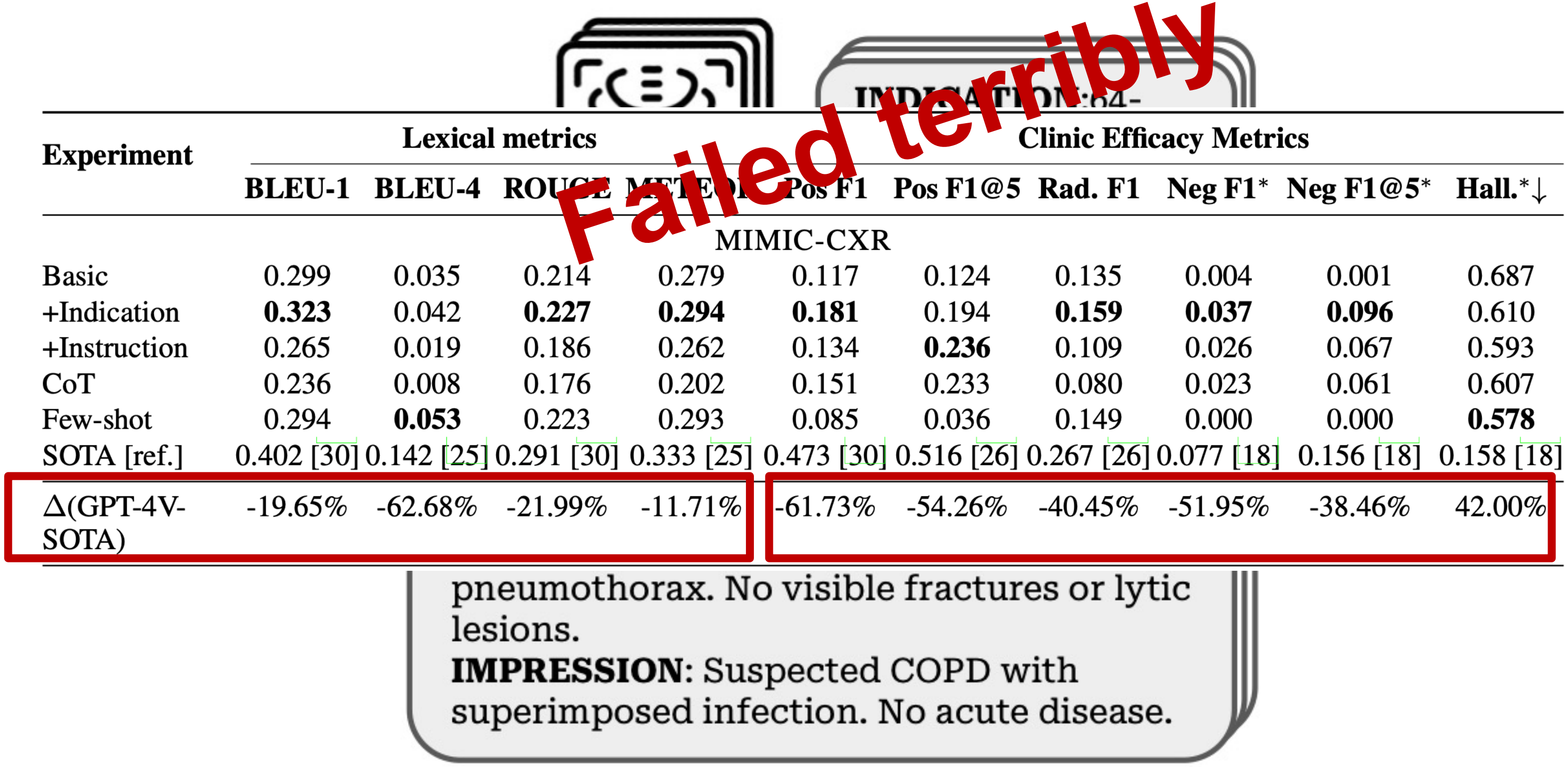
AI Replacing Radiologists... The Truth

Yasha Gupta, MD

Rad AI
Rad AI Reporting
Radiology Reporting Software with Unparalleled Efficiency and Accuracy
Nicholas Galante, MD
CMIO | RANT
2:24

!!!

Direct report generation



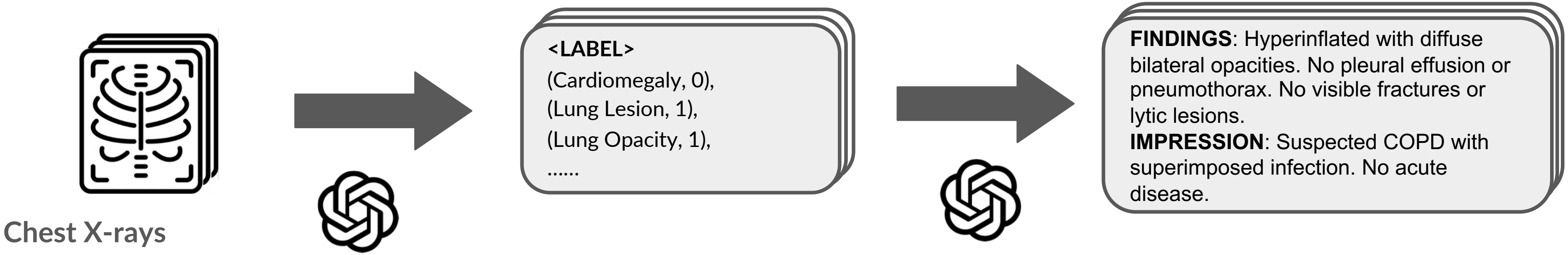
Failed terribly

INDICATION:64-

Experiment	Lexical metrics				Clinic Efficacy Metrics					
	BLEU-1	BLEU-4	ROUGE	METEOR	Pos F1	Pos F1@5	Rad. F1	Neg F1*	Neg F1@5*	Hall.*↓
					MIMIC-CXR					
Basic	0.299	0.035	0.214	0.279	0.117	0.124	0.135	0.004	0.001	0.687
+Indication	**0.323**	0.042	**0.227**	**0.294**	**0.181**	0.194	**0.159**	**0.037**	**0.096**	0.610
+Instruction	0.265	0.019	0.186	0.262	0.134	**0.236**	0.109	0.026	0.067	0.593
CoT	0.236	0.008	0.176	0.202	0.151	0.233	0.080	0.023	0.061	0.607
Few-shot	0.294	**0.053**	0.223	0.293	0.085	0.036	0.149	0.000	0.000	**0.578**
SOTA [ref.]	0.402 [30]	0.142 [25]	0.291 [30]	0.333 [25]	0.473 [30]	0.516 [26]	0.267 [26]	0.077 [18]	0.156 [18]	0.158 [18]
Δ(GPT-4V-SOTA)	-19.65%	-62.68%	-21.99%	-11.71%	-61.73%	-54.26%	-40.45%	-51.95%	-38.46%	42.00%

pneumothorax. No visible fractures or lytic lesions.
IMPRESSION: Suspected COPD with superimposed infection. No acute disease.

Report generation =



<LABEL>
(Cardiomegaly, 0),
(Lung Lesion, 1),
(Lung Opacity, 1),
……

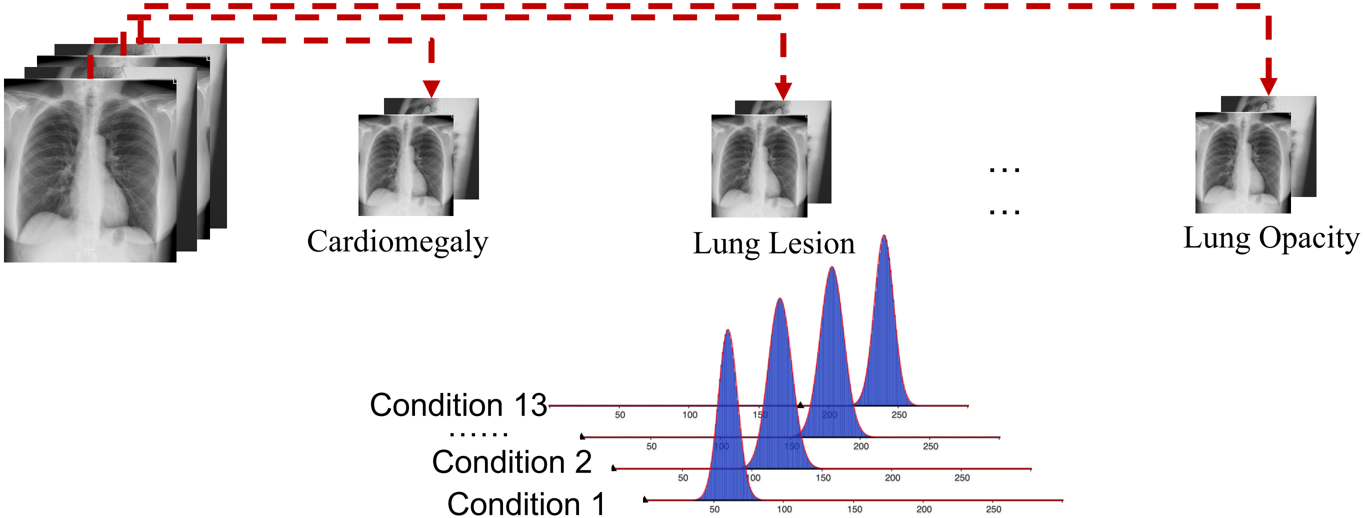
FINDINGS: Hyperinflated with diffuse bilateral opacities. No pleural effusion or pneumothorax. No visible fractures or lytic lesions.
IMPRESSION: Suspected COPD with superimposed infection. No acute disease.

Image reasoning report synthesis

Can GPT-4V interpret chest X-rays meaningfully?

Metric	MIMIC-CXR		IU X-RAY	
	Chain-of-Thought (1st Step)	Image Reasoning	Chain-of-Thought (1st Step)	Image Reasoning
Positive F1	0.166	0.146	0.072	0.049
Positive F1@5	0.261	0.208	0.095	0.056

The model performs poorly in identifying conditions from chest X-ray images across different prompting strategies.

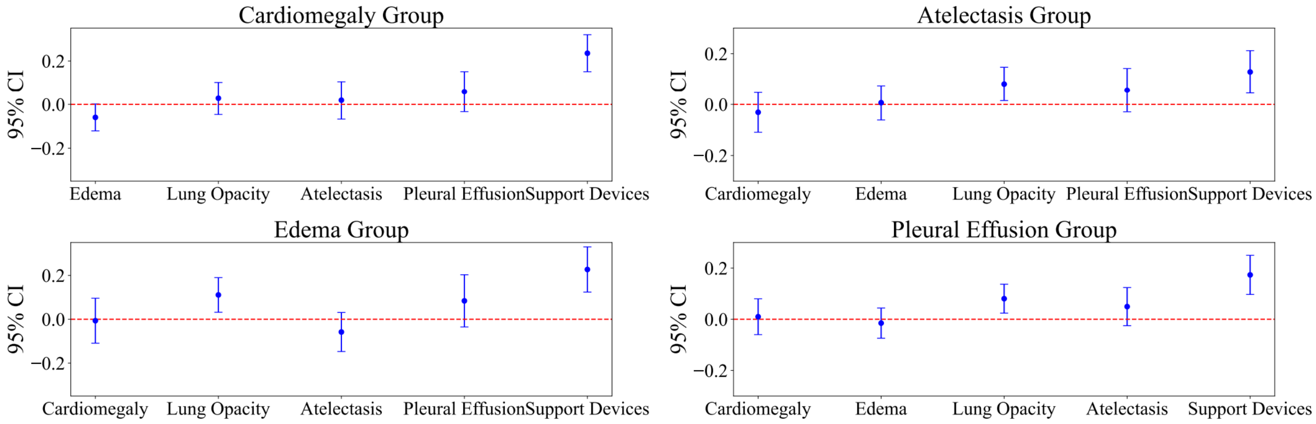


Sub-group label distribution?

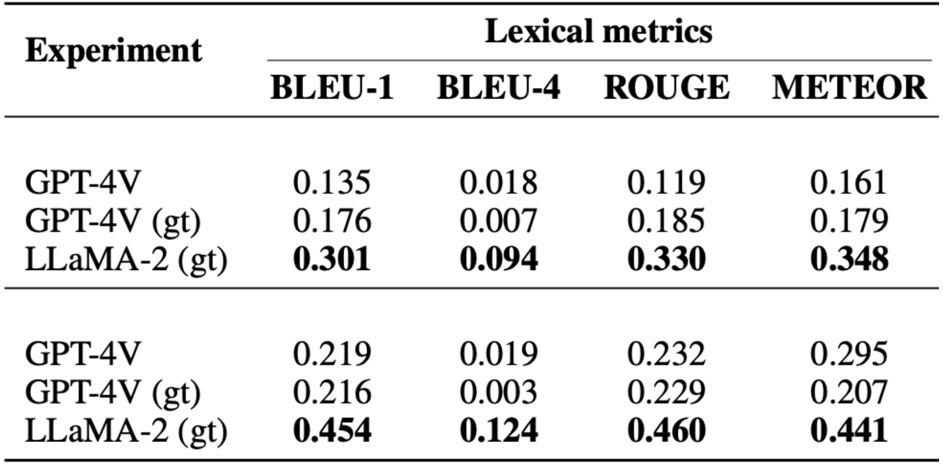
χ^2-test to test if GPT-4V follows the same distribution across different groups to identify positive conditions.

Statistics	**Overall**		**Top 6 Conditions**	
	Groundtruth	**GPT-4V**	**Groundtruth**	**GPT-4V**
χ^2 statistic	1770.38	74.25	317.86	6.11
p-value	$p < 0.0001$	1.0000	$p < 0.0001$	1.0000
df.	144	144	25	25

Bootstrap CI to test if GPT-4V labels one certain condition independently of the groundtruth condition group.



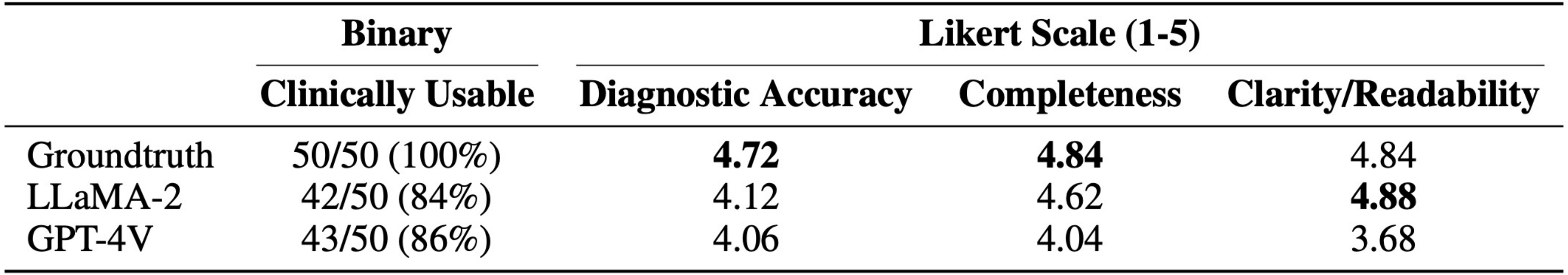
Report synthesis given groundtruth labels



Experiment	Lexical metrics			
	BLEU-1	BLEU-4	ROUGE	METEOR
GPT-4V	0.135	0.018	0.119	0.161
GPT-4V (gt)	0.176	0.007	0.185	0.179
LLaMA-2 (gt)	**0.301**	**0.094**	**0.330**	**0.348**
GPT-4V	0.219	0.019	0.232	0.295
GPT-4V (gt)	0.216	0.003	0.229	0.207
LLaMA-2 (gt)	**0.454**	**0.124**	**0.460**	**0.441**

✓ Significant **improvements**

Additional human evaluation by a **board certified radiologist**



	Binary	**Likert Scale (1-5)**		
	Clinically Usable	**Diagnostic Accuracy**	**Completeness**	**Clarity/Readability**
Groundtruth	50/50 (100%)	**4.72**	**4.84**	4.84
LLaMA-2	42/50 (84%)	4.12	4.62	**4.88**
GPT-4V	43/50 (86%)	4.06	4.04	3.68

1. Human written report are 100% usable, whereas even with groundtruth labels, model generated reports are still not perfect.
2. Human written reports contains richer and more nuanced information.
3. Model generated reports have the potential to have better clarity/readability.

❌ **Gap** to human written reports

Ongoing: building a **radiology foundation model**

- High quality medical data curation
- Extend LLM to MLLM, with medical image comprehension capability
- Effective training/finetuning recipe

Find us at the poster session!
or chacha@uchicago.edu