Beyond Generative AI: The Rise of Verification for Clinical Safety in Radiology AI

Tanveer Syeda-Mahmood
IBM Fellow & Adjunct Professor
Stanford University

Automated Radiology Report Generation (RRG) using Generative AI

AI Read

Preliminary report

Lung Findings : Left and right lung are clear.
No evidence of pleural effusion, pneumothorax or pulmonary edema.

Mediastinum findings :
Cardio-mediastinal silhouette is normal.

Impression: No Consolidation
No Pleural Effusion
No Pneumothorax
No pulmonary edema.

Radiologist signoff

- **Expedites clinical workflow operations**
  - Helps when experts are not available
  - Reduces radiologists' workload
    - Saves radiologists from dictation
    - Expedited review
  - Improves consistency

- **Responsible radiology reporting:**
  - Report present and relevant absent findings
  - Describe at high granularity covering location, laterality, severity, etc.

Problems with Generative AI in Radiology

- Limited number of findings covered
- Diagnostic accuracy not sufficient for clinical use
  - **hallucinations** (fabricating or missing key findings), **laterality errors** (confusing left and right), and **diagnostic inaccuracies**
- Performance dominated by prevalent findings. Harder to catch rarer findings
- Opaque reasoning – not very explainable
- Much of the source of hallucination is due to auto-regressive language generation

Auto-read from AI has errors

RRG models

Match to GT | Mismatches

Present findings: enlarged cardiac silhouette, pacer, calcifications, sternotomy wires, mediastinal clips, clear lungs
Absent findings: pleural effusion, vascular congestion

Ground truth report

MEDICAL CONDITION:
78 year old woman with chest pain, hx of chf
REASON FOR THIS EXAMINATION:
please evaluate for edema
FINAL REPORT
CHEST, TWO VIEWS:
HISTORY: 78-year-old female with chest pain. History of congestive failur
Question pulmonary edema.

FINDINGS: PA and lateral views of the chest are compared to previous exam from .
Compared with prior, there has been no significant interval change. The lungs remain clear. There is no pleural effusion. There is no pulmonary vascular engorgement. Cardiac silhouette is enlarged, but stable in configuration. Biventricular pacing device again seen with multiple leads stable positions. Atherosclerotic calcifications seen throughout the aorta Median sternotomy wires and mediastinal clips again noted.

IMPRESSION: No acute cardiopulmonary process.

X-ray GPT:

In this patient who has an upper lobe nodule and associated pain complaints, there is no indication of pleural effusion and cardiac failure. However, edema can be present, which should be evaluated to exclude a pulmonary disease. The patient has mild pulmonary fibrosis and may have underlying COPD which should also be considered. The findings suggest cardiac failure and the patient requires diuretics for treatment. The ECG appears normal and the dual lead AID pacing device remains intact.

RGRG:

The lungs are clear. No acute cardiopulmonary process. The cardiomediastinal silhouette and hilar contours are normal. There is no focal consolidation, effusion, or pneumothorax. There are no acute osseous abnormalities. Moderate cardiomegaly is unchanged. Left-sided pacemaker leads terminate in the right atrium and right ventricle.

GPT4o:
FINDINGS:
Lines/Tubes: Pacemaker leads are in place.
Lungs: Mild interstitial markings observed; no acute infiltrates.
Pleura: No significant pleural effusion detected.
Heart and Mediastinum: Enlarged cardiac silhouette consistent with known CHF.
Bones and soft tissues: No acute fractures or bone abnormalities detected.

- **Errors could be due to :**
  - Missed findings
  - Incorrect or hallucinated findings
  - Incorrect descriptions of findings
  - Reversed findings (presence -> absence or vice versa)

IEEE ISBI 2026

How to address such errors?

- Existing methods are mostly at training time
  - Alignment methods such as PPO, DPO
    - Used at training time
    - Human input
- Inference-time methods
  - Reinforcement learning methods: GRPO
    - Require generation of multiple outputs for reinforcement learning
    - Consensus could still present hallucinations
  - LLM as a judge
    - Still needs a ground truth
    - Suffers from hallucination as well
- Radiology needs:
  - Error detection and correction should be at inference time
  - Such detection and correction should work for any RRG model
  - No ground truth can be expected for correction
  - Inference-time operations should be lightweight for performance

Verification AI

- A class of designed to detect, measure, correct, and ultimately prevent errors in AI-generated clinical reports
- But how is this possible?
  - Since the output of generative AI could be a very large number of possibilities depending on the model, training prep, etc.
- This is possible in domain-specific contexts such as RRG
  - Reports describe known findings in the image.
  - Description syntax follows some common norms (e.g. presence or absence, core findings, location, laterality, etc.)
  - Errors made by RRG models are in the description of the findings
  - Could be generalizable for other domain-specific contests
- Such models could be discriminative or generative themselves

Verification AI models

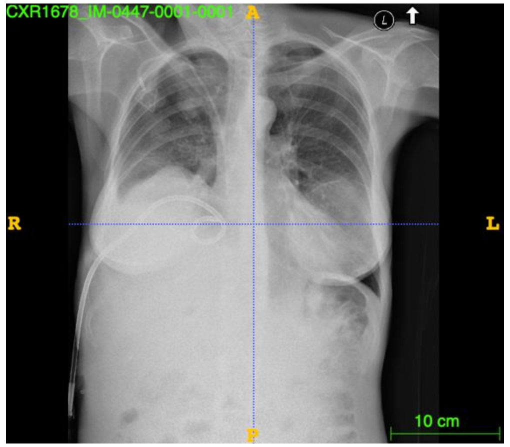
- Fact-checking discriminative models
  - Detects errors in the output produced by generative AI
- Report correction models
  - Corrects for the errors detected
- Non-RL Inference-time alignment models
  - Lightweight alignment of RRG models onsite using corrective models
- Retrieval-augmented verification
  - Trains LLM as judges through retrieval-augmented verification based on patient similarity

Modeling the space of errors in radiology reports

Types of errors found in reports

- Correct Finding (No error)
- Incorrect Finding
  - Irrelevant finding
  - Reverse finding
  - Exchange finding
- Incorrect location
- Missing or other Findings

Chest X-ray



Associated GT report

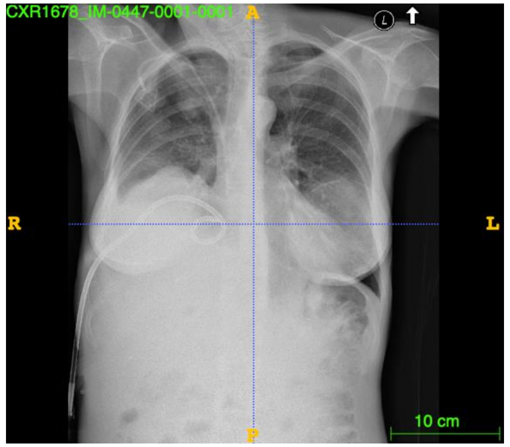
The heart is normal in size. Right chest xxxx tip is again seen at the cavoatrial junction. There is no pneumothorax. There is again elevation of right hemidiaphragm with right-sided pleural effusion. Vague opacities are noted in the right upper lobe, xxxx from prior study. These may be related to overlying rib lesions versus true pulmonary nodules. The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant.

Modeling the space of errors in radiology reports

- Correct Finding (No error)
- Incorrect Finding
  - Irrelevant finding
  - Reverse finding
  - Exchange finding
- Incorrect location
- Missing or other Findings

Irrelevant finding

Chest X-ray



Associated GT report

The heart is normal in size. Right chest xxxx tip is again seen at the cavoatrial junction. There is no pneumothorax. There is again elevation of right hemidiaphragm with right-sided pleural effusion. Vague opacities are noted in the right upper lobe, xxxx from prior study. These may be related to overlying rib lesions versus true pulmonary nodules. The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant.

Automated report

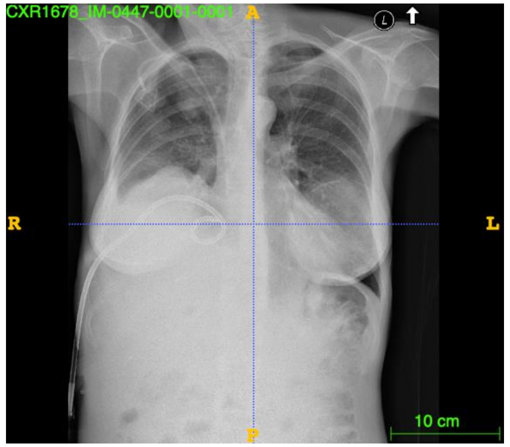
The heart is normal in size. Right chest xxxx tip is again seen at the cavoatrial junction. There is no pneumothorax. There is again elevation of right hemidiaphragm with right-sided pleural effusion. Vague opacities are noted in the right upper lobe, xxxx from prior study. These may be related to overlying rib lesions versus true pulmonary nodules. The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant. **There are diffuse increased interstitial markings, suggestive of pulmonary fibrosis in bilateral lung.**

Modeling the space of errors in radiology reports

- Correct Finding (No error)
- Incorrect Finding
  - Irrelevant finding
  - Exchange finding
  - Reverse finding
- Incorrect location
- Missing or other Findings

Reverse finding

Chest X-ray



Associated GT report

The heart is normal in size. Right chest xxxx tip is again seen at the cavoatrial junction. There is no pneumothorax. There is again elevation of right hemidiaphragm with right-sided pleural effusion. Vague opacities are noted in the right upper lobe, xxxx from prior study. These may be related to overlying rib lesions versus true pulmonary nodules. The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant.

Automated report

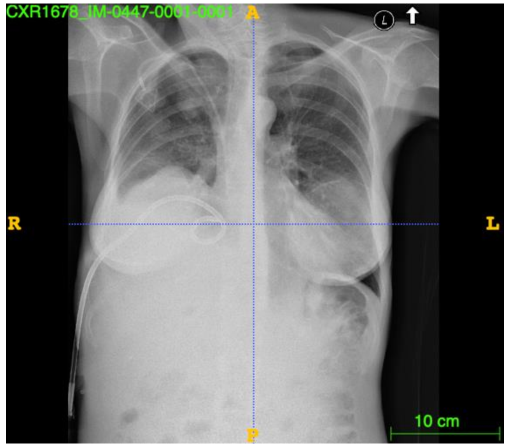
The heart is normal in size. Right chest xxxx tip is again seen at the cavoatrial junction. **There is pneumothorax**. There is again elevation of right hemidiaphragm with right-sided pleural effusion. Vague opacities are noted in the right upper lobe, xxxx from prior study. These may be related to overlying rib lesions versus true pulmonary nodules. The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant.

Modeling the space of errors in radiology reports

- Correct Finding (No error)
- Incorrect Finding
  - Irrelevant finding
  - Reverse finding
  - Exchange finding
- Incorrect location
- Missing or other Findings

Incorrect location

Chest X-ray



Associated GT report

The heart is normal in size. Right chest xxxx tip is again seen at the cavoatrial junction. There is no pneumothorax. There is again elevation of right hemidiaphragm with right-sided pleural effusion. Vague opacities are noted in the right upper lobe, xxxx from prior study. These may be related to overlying rib lesions versus true pulmonary nodules. The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant.

Automated report

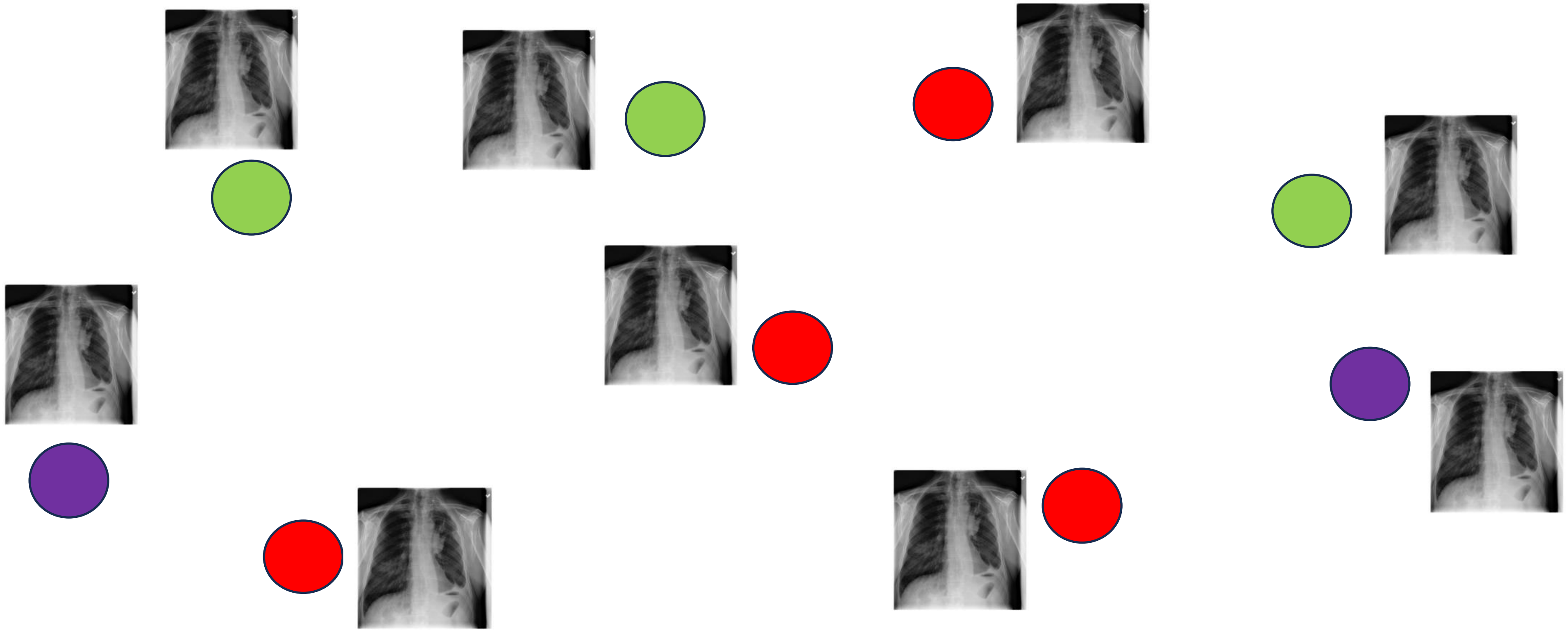
The heart is normal in size. Right chest xxxx tip is again seen at the cavoatrial junction. There is no pneumothorax. There is again elevation of the **left hemidiaphragm with left-sided pleural effusion**. Vague opacities are noted in the right upper lobe, xxxx from prior study. These may be related to overlying rib lesions versus true pulmonary nodules. The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant.

Modeling the space of errors made by generative AI

Collect labeled image-finding pairs simulating errors

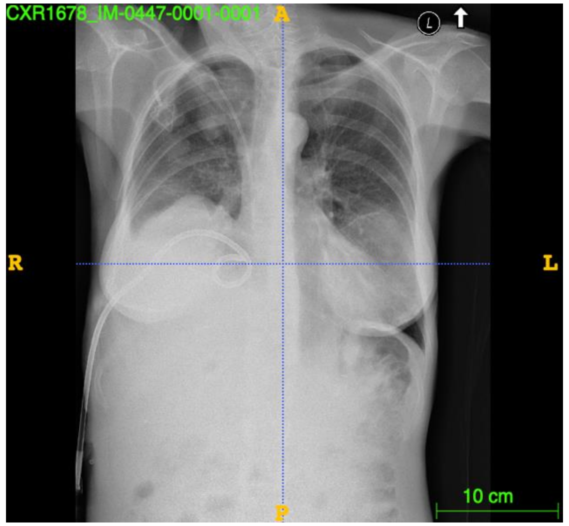
Synthesize Image-Text Data

- Representative of typical errors in automated reports
- Realistic and possible combinations verifiable through ground truth
- Robust to written variations



Synthesizing image-text pairs

Synthetic Data Generation



- *Many many combinations possible!*
- *Not all of them are realistic.*
- *How to filter?*
  - *Based on findings provided*
  - *Avoid repetitions*
  - *Remove Contradictions*

R1

The heart is normal in size. There is no pneumothorax. There is again elevation of right hemidiaphragm with right-sided pleural effusion.

- Normal heart
- No pneumothorax
- Elevated hemidiaphragm
- Pleural effusion

R2

The left lung appears grossly clear. Drainage catheter seen overlying the right upper quadrant.

- Clear lung
- Chest tube

R3

There are diffuse increased interstitial markings, suggestive of pulmonary fibrosis in bilateral lung. There is a small left pleural effusion/pneumothorax.

- ILD
- Pulmonary fibrosis
- Pleural effusion
- Pneumothorax.

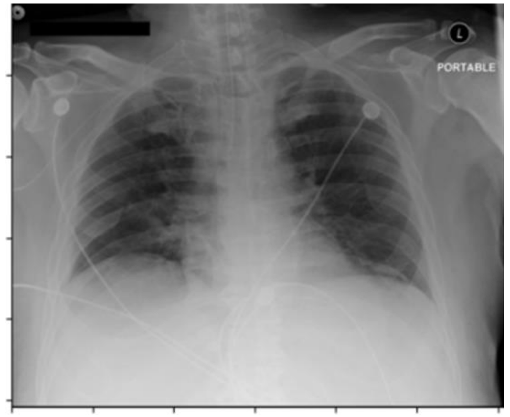
Real

Fake

Drop pleural effusion
Drop pneumothorax

Synthesizing Data – Labeling findings

Synthetic Data Generation



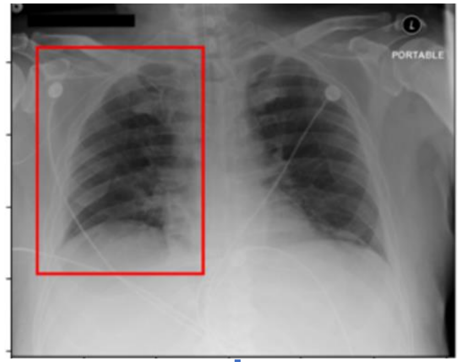
Mild basilar {atelectatic changes} without evidence of acute pneumonia or vascular congestion. There may well be a small right pleural effusion.

There may well be a small right pleural effusion

FFL: Anatomical finding|yes|pleural effusion|lung|right

Short FFL Yes|Pleural effusion

Left lung



Anatomical location <177,395,1146,1569>

Normalized location <.06,0.15,0.37,0.62>

INPUT

Yes|Pleural effusion

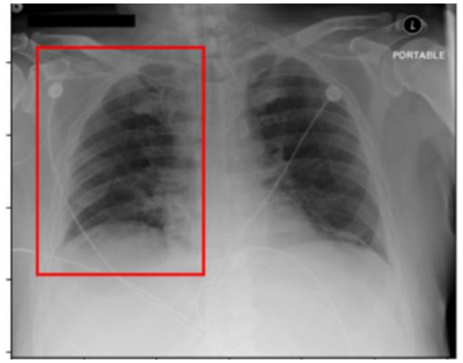
LABEL

<.06,0.15,0.37,0.62,1>

1=Real, 0=Fake

Synthesizing Data – Perturbing presence findings

Synthetic Data Generation

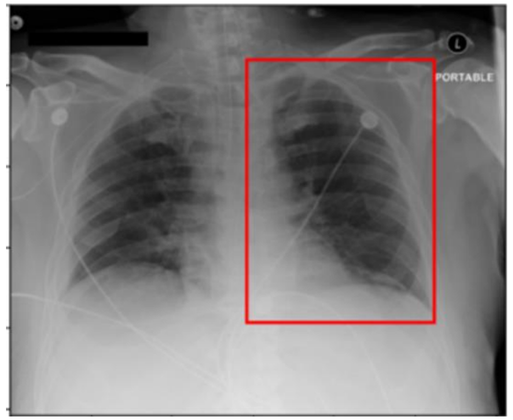


Yes|Pleural effusion

<.06,0.15,0.37,0.62,1>

There may well be a small right pleural effusion.

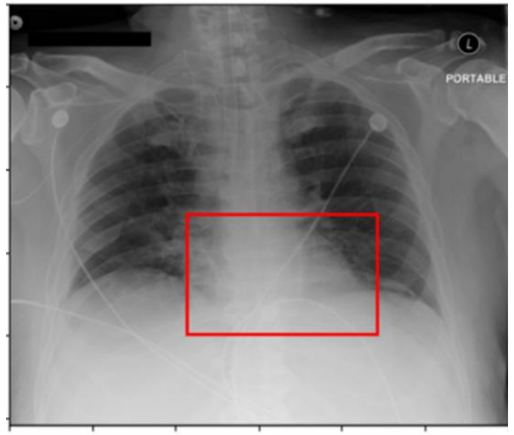
Finding is correct but location is incorrect



Yes|Pleural effusion

<0.47, 0.13, 0.38,0.64,0>

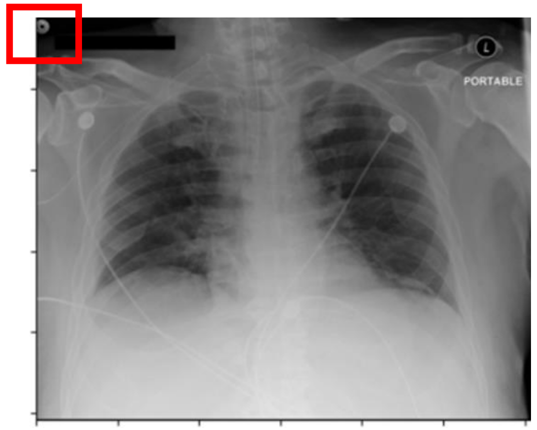
Exchange Finding



Yes|Cardiomegaly

<0.35, 0.5,0.37,0.28,0>

Reverse Finding



No| Pleural effusion

<0,0,0,0,0>

Error modeling requires finding descriptors

Error modeling

Fine-grained Finding Label (FFL) Pattern

Drainage catheter seen overlying the left lower quadrant.

Tubes and lines|yes|chest tube|lower lung lobe|left

Type of finding

Present/ absent

Finding name

Anatomical location

Laterality

Syeda-Mahmood et al. MICCAI 2020

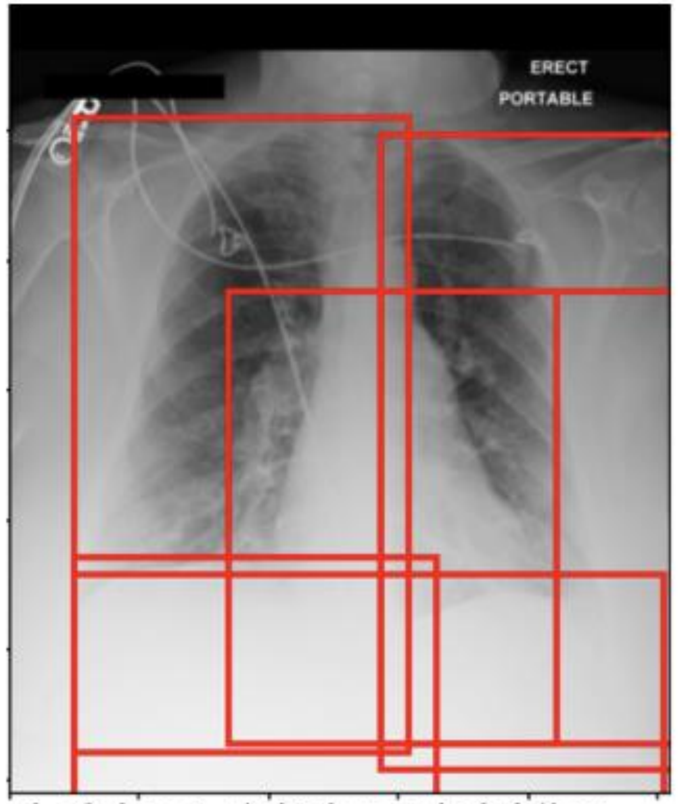
Extracted FFL patterns

No	Sentence	FFL
1.	There is enlargement of the cardiac silhouette, accentuated by the low lung volumes.	anatomical finding|yes|enlarged cardiac silhouette|heart
2.	FINDINGS: The heart appears mildly enlarged.	anatomical finding|yes|enlarged cardiac silhouette|heart
3.	This accentuates the size of the cardiac silhouette which is likely mildly enlarged but unchanged.	anatomical finding|yes|enlarged cardiac silhouette|heart
4.	Cardiac size is slightly enlarged allowing for limitations of this AP view.	anatomical finding|yes|enlarged cardiac silhouette|heart
5.	Persistent mild-to-moderate cardiomegaly.	anatomical finding|yes|enlarged cardiac silhouette|heart
6.	Nasogastric tube terminates in the stomach.	tubes and lines|yes|enteric tube|stomach
7.	Enteric tube seen passing below the field of view, partially coiled in the stomach.	tubes and lines|yes|enteric tube|stomach
8.	Pleural vasculature is not engorged and the patient has moderate pulmonary edema on the right.	anatomicalfinding|no|vascular congestion|lung anatomicalfinding|yes|pumonary edema|lung|right

Error modeling requires finding descriptors

Error modeling

Anatomical localization



Left lung, right lung, left hilar structures, right hilar structures, right hemidiaphragm, abdomen

Bounding boxes for all major(40) anatomical regions available

<x,y,w,h>

Anatomical regions localized

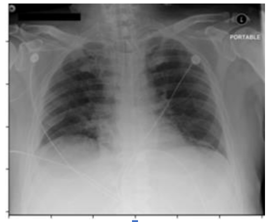
S.No.	Anatomical Region	S.No.	Anatomical Region	S. No.	Anatomical Region
1.	left mid lung zone	14.	mediastinum	27.	right lower lung zone
2.	left costophrenic angle	15.	left apical zone	28.	left upper lung zone
3.	abdomen	16.	left chest wall	29.	neck
4.	left breast	17.	right atrium	30.	trachea
5.	cavoatrial junction	18.	right breast	31.	aortic arch
6.	right costophrenic angle	19.	right chest wall	32.	left lung
7.	right arm	20.	carina	33.	right apical zone
8.	left shoulder	21.	left hemidiaphragm	34.	right mid lung zone
9.	spine	22.	right hemidiaphragm	35.	left arm
10.	left clavicle	23.	left hilar structures	36.	right shoulder
11.	left lower lung zone	24.	right hilar structures	37.	right clavicle
12.	svc	25.	right upper lung zone	39.	cardiac silhouette
13.	right lung	26.	upper mediastinum	40.	unknown

Wu et al. ISBI 2020

Error modeling requires finding localization

Synthetic Data Generation

How to spatially locate a finding?



FFL extraction

Extract location description

Anatomical mapping

Anatomical localization

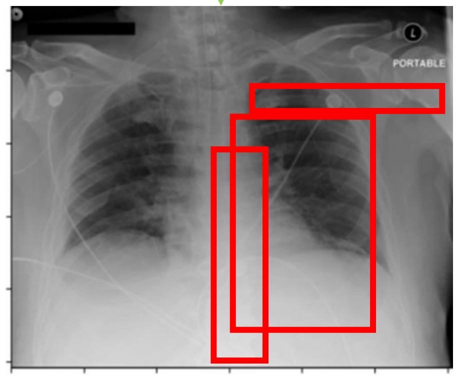
Merge Bboxes

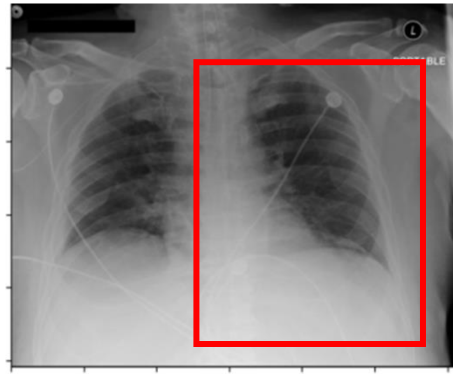
<x,y,w,h>

Sentence

Extract core finding

F





Anatomical finding|no|bone lesion|chest wall|left

Sentence

Left chest wall

['left lung', 'left clavicle', 'spine’]

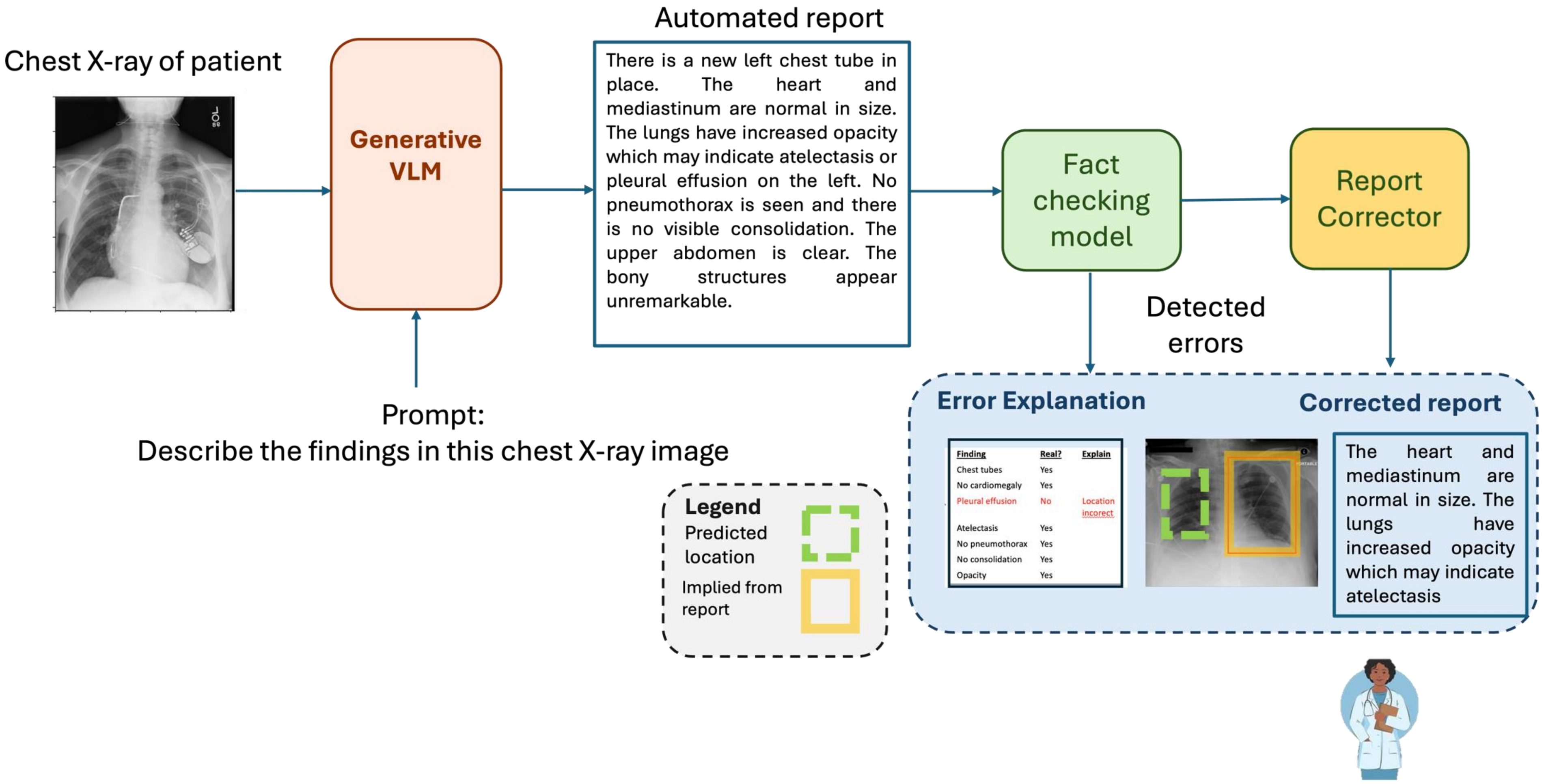
No|bone lesion

Bony structures are unremarkable aside for mild leftward convex curvature.

Fact-Checking Models

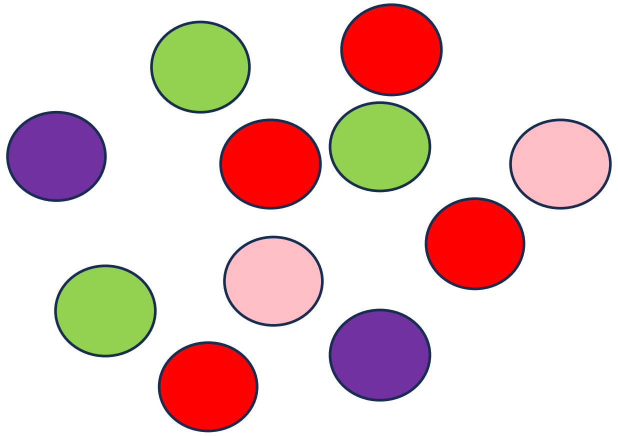
Fact-checking Model for RRG

- Discriminative model
- Trained on synthetic dataset modeling correct and error-prone data
  - Missing findings
  - Incorrect findings
  - Irrelevant findings
  - Incorrect descriptions
- Agnostic to RRG models



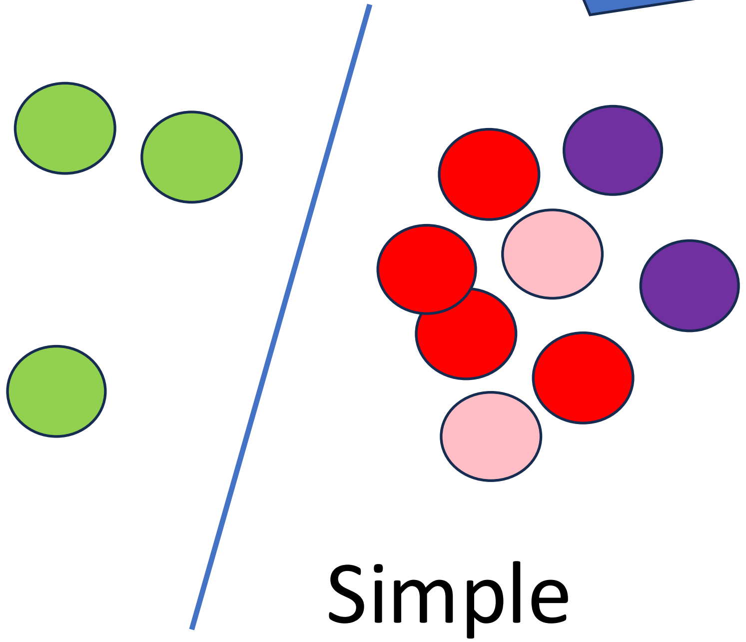
Fact-checking model – Key idea

Finding errors in automated reports

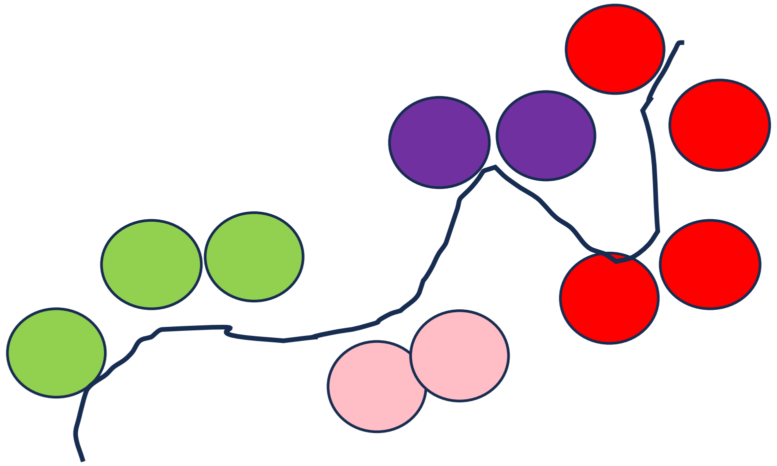


- Correct Finding
- Incorrect location of finding
- Incorrect Finding
- Other errors

Learn the manifold separating the correct findings from others



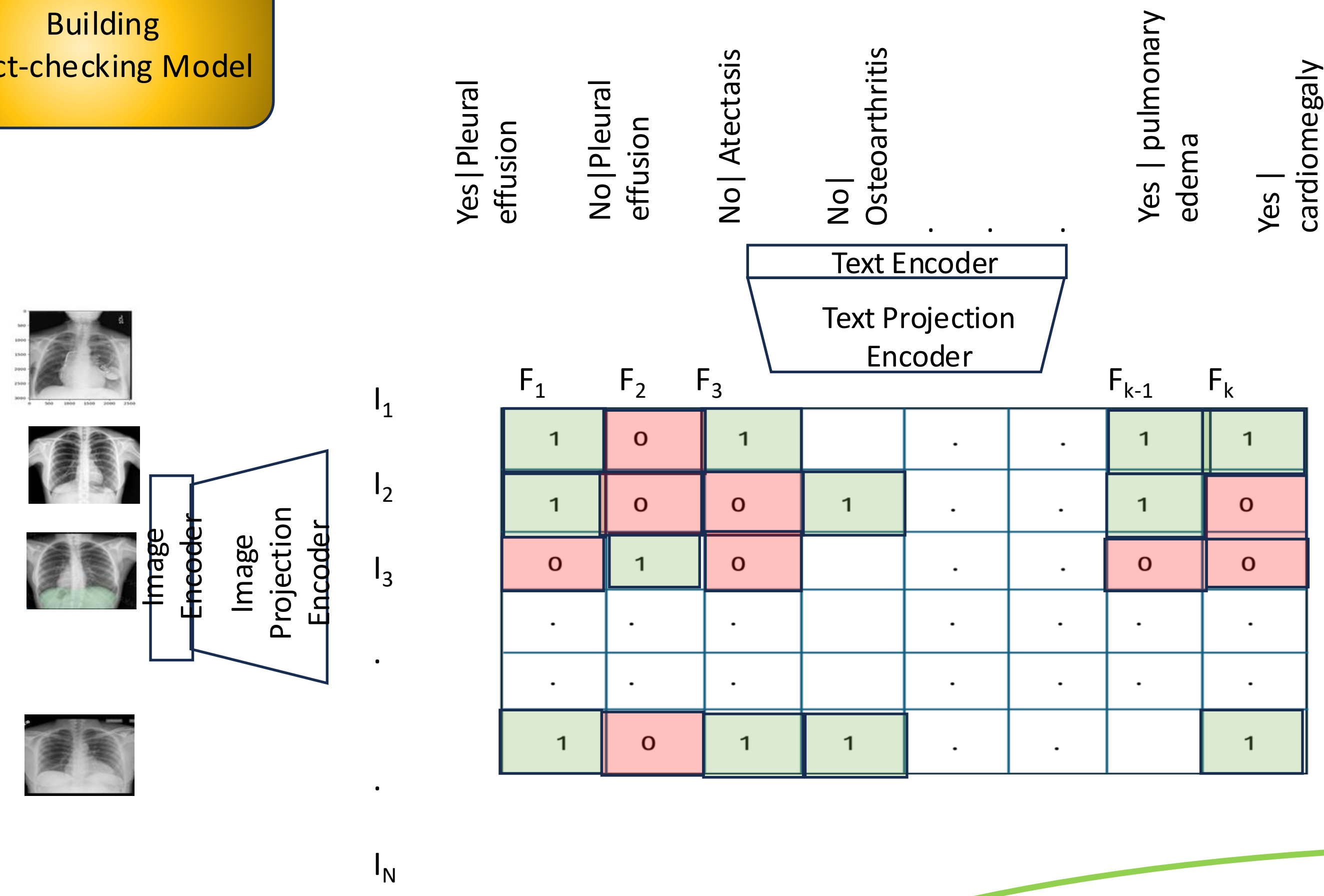
Simple



Advanced

Multi-label cross-modal supervised contrastive feature learning

Building
Fact-checking Model



Supervised contrastive CLIP (SC-CLIP)

- Sparse similarity matrix. Easy to train
- Pre-trained image and text encoders from ChexZero
- Projection encoders fresh-trained using new loss function
  - 768x512 for image
  - 512x512 for text
- 151,277,313 parameters

$$s_{if_{ij}} = z_i \cdot z_{f_{ij}}$$ Latent features: real findings

$$s_{ia_{ik}} = z_i \cdot z_{a_{ik}}$$ Latent features: fake findings

Real findings

Cross-modal supervised contrastive loss

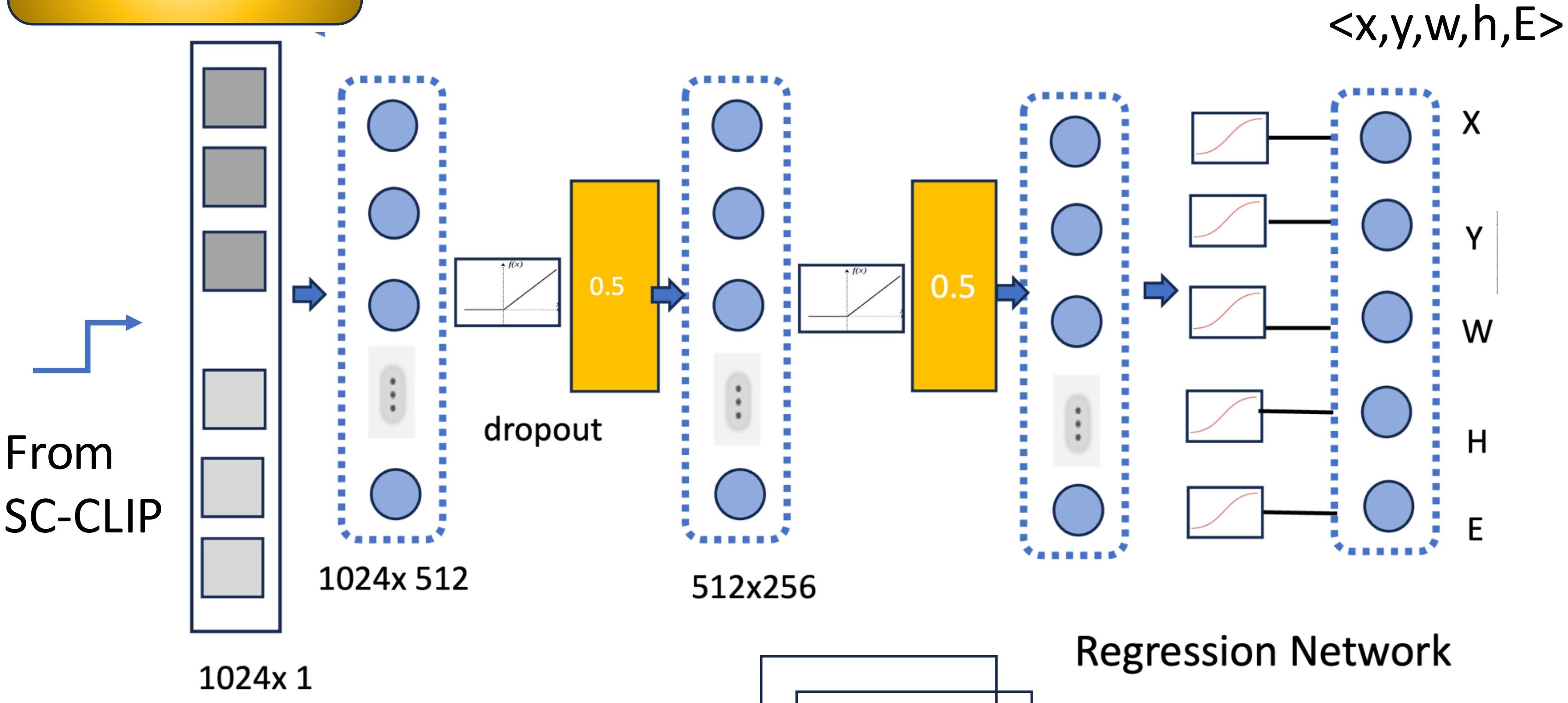
$$\mathcal{L}_{SupC} = \sum_{i \in I} \frac{-1}{|F_{iReal}|} \sum_{f_{ij} \in F_{iReal}} log \frac{e^{s_i f_{ij}/\tau}}{\sum_{a_{ik} \in F_{iFake}} e^{s_{ia_{ik}}/\tau}}$$

Fake findings

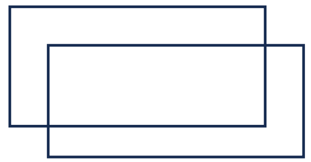
Unlike SCL, the denominator is over the fake findings per image

Cross-modal regression network

Building Fact-checking Model



- Ground truth labels:
  $Y_g =< Y_{1g}, Y_{2g} >$
  where $Y_{1g} =< x_g, y_g, w_g, h_g >$ and $Y_{2g} = E_g = 1/0$
- Predicted labels:
  $Y_p =< Y_{1p}, Y_{2p} >$
  where $Y_{1p} =< x_p, y_p, w_p, h_p >$ and $Y_{2p} = E_p = 1/0$
- Convex hull of bounding boxes Y_{1g}, Y_{1p} $\quad C_{Y_{1p},Y_{1g}}$
- 657,413 parameters



$$\mathcal{L}_{Reg_i} = \underbrace{|Y_{1p} - Y_{1g}|}_{\mathcal{L}_1(Y_{1p},Y_{1g})} + \underbrace{\frac{|Y_{1p} \cap Y_{1g}|}{|Y_{1p} \cup Y_{1g}|} - \frac{|C_{Y_{1p},Y_{1g}} \backslash Y_{1p} \cap Y_{1g}|}{|C_{Y_{1p},Y_{1g}}|}}_{\mathcal{L}_{\mathrm{giou}}(Y_{1p},Y_{1g})} + \underbrace{|Y_{1p} - Y_{1g}|^2}_{\mathcal{L}_{\mathrm{mse}}(Y_{1p},Y_{1g})} - \underbrace{[Y_{2g}\mathbf{log}(Y_{2p}) + (1 - Y_{2g})\mathbf{log}(1 - Y_{2p})]}_{\mathcal{L}_{BCE}(Y_{2p},Y_{2g})}$$

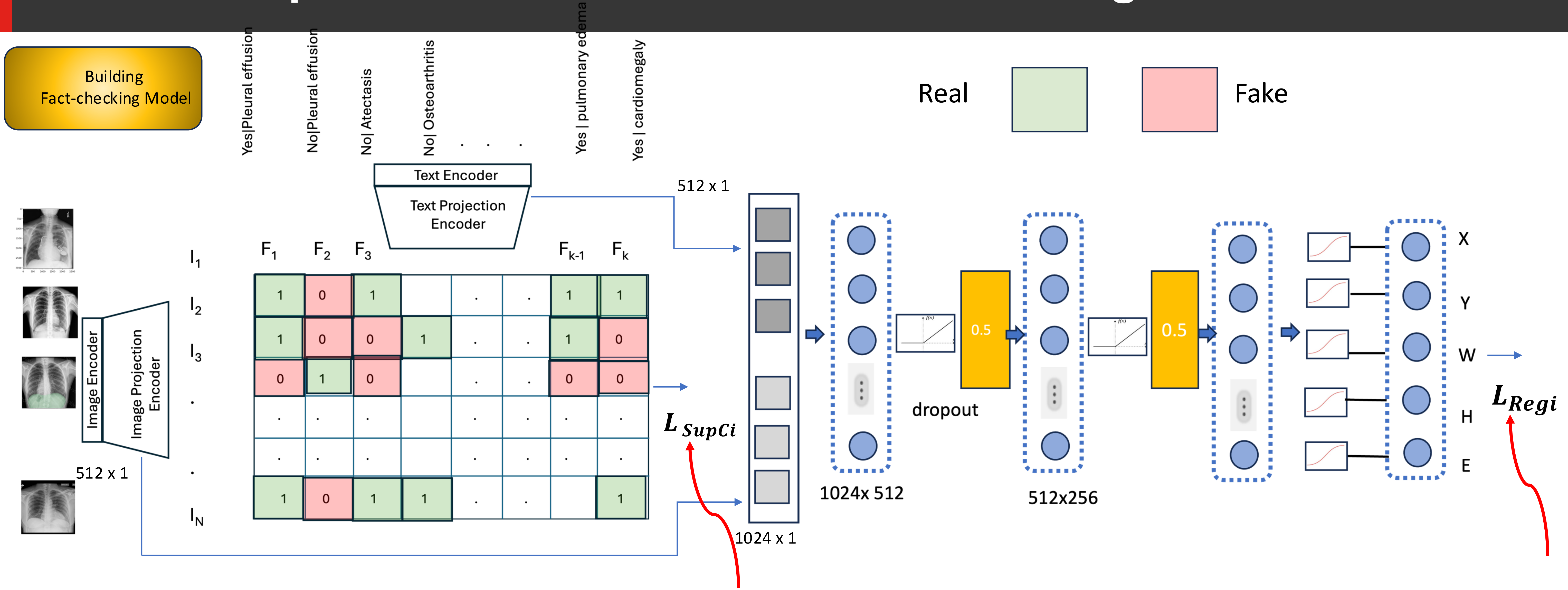
L1 loss for each spatial coordinate

Bbox region overlap GIOU loss

MSE to account for absence findings <0,0,0,0> coords

BCE loss for the finding veracity label E

FCModel: Supervised Cross-modal Contrastive Regression Network



- Built as an end-to-end trainable network
  - image and text projection encoders
  - concatenated image-text representation
  - Regression on contrasted features
- Findings can be presence or absence findings

- Novel formulation differs from:
  - Contrastive learning
  - Supervised contrastive learning
  - Cross-modal CLIP-style learning
  - Multi-label contrastive learning

Explainable fact-checking Example – Correct case

Building Fact-checking Model

Case: Finding and location are correct

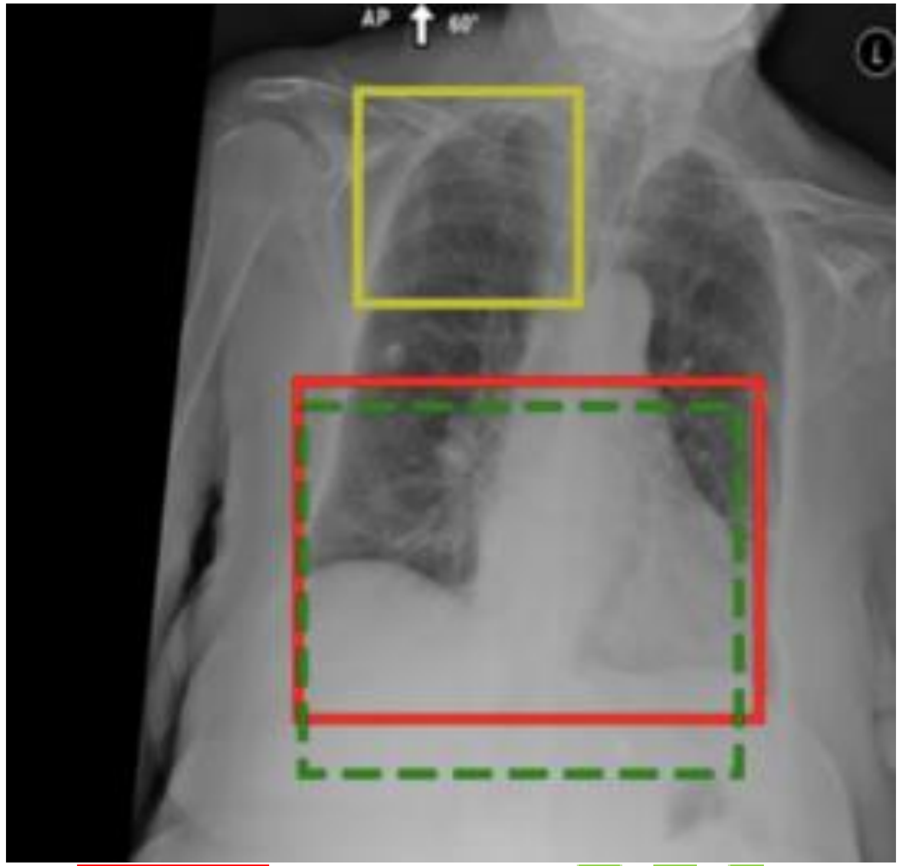
Ground truth sentence	... pronounced alveolar opacities in a perihilar distribution, right greater than left, likely representing pulmonary edema
Automated report sentence	The pulmonary vasculature is not engorged, and the patient has moderate pulmonary edema on the right.
Target finding	Yes | Pulmonary edema
Implied merged anatomical locations	['right hilar structures', 'right lung']
GT location	[0.120, 0.125, 0.366 0.586]
Predicted location	[0.151, 0.100, 0.351, 0.538, 0.428]
Predicted label	1

GT Predicted Indicated

Explainable fact-checking Example – Incorrect case

Building
Fact-checking Model

Case: Finding is correct but location is wrong



Ground truth sentence	Opacities are present at the lung bases bilaterally.
Automated report sentence	Subtle opacity in the right upper lung.
Target finding	Yes|Opacity
Implied merged anatomical locations	[right upper lung]
GT location	[0.328, 0.419, 0.498, 0.363]
Predicted location	[0.332, 0.442, 0.470, 0.396]
Predicted label	0

GT Predicted Indicated

Evaluation

- Raw datasets for forming the synthetic data should show demographics variety
- List of findings to address should be large enough
- RRG models to be corrected need to be of different styles

Experimental results - Datasets

Dataset	Patients Train/Val/Test	Images	Findings	Regions	Real/Synth.
CImagenomeS[79]	44,133/6274/12,538	243,311	49	922,295	1.616M/27.047M
CImaGenomeG[79]	288/33/69	461	35	5,477	4,063/23,463
MS-CXR[24]	478/54/114	925	8	2,254	2,247/24,338
ChestXray8[71]	457/51/109	880	8	1,571	1,571/10,137
VinDr-CXR[43]	9,450/1,050/2,250	15,000	23	69,052	47,973/132,632

- All datasets clinician validated
- ChestImagenomeS (Silver) used for training the FCModel only
- ChestImagenomeG (Gold) used for other evaluations (comparisons, report correction, etc.)

Findings (Presence and Absence)

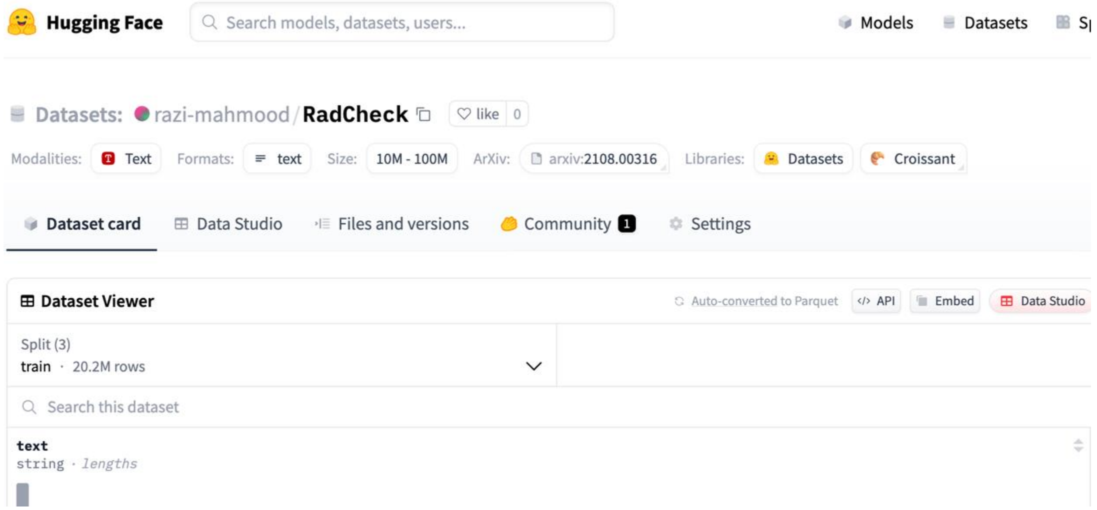
1.	Aortic enlargement	10.	Atelectasis	18.	Calcification
2.	Cardiomegaly	11.	Clavicle fracture	19.	Consolidation
3.	Edema	12.	Emphysema	20.	Enlarged pa
4.	ILD	13.	Infiltration	21.	Infiltrate
5.	Lung cavity	14.	Lung cyst	22.	Lung opacity
6.	Mediastinal shift	15.	Nodule/mass	23.	Other lesion
7.	Pleural effusion	16.	Pleural thickening	24.	Pneumothorax
8.	Pulmonary fibrosis	17.	Rib fracture	25.	Pneumonia
9.	No finding				

Largest set of findings considered!

Synthetic dataset creation details

Partition	Patients	Studies	Images	Raw Findings	Raw Data	Synthetic Data
Train	44133	149252	166521	1,740,616	2,040,902	18,400,427
Test	12538	42498	47393	498,940	523,233	5,254,091
Validate	6274	21479	23953	253,972	262,518	2,640,780
Total	62945	213229	230229	2,493,528	2,826,653	23,919,298

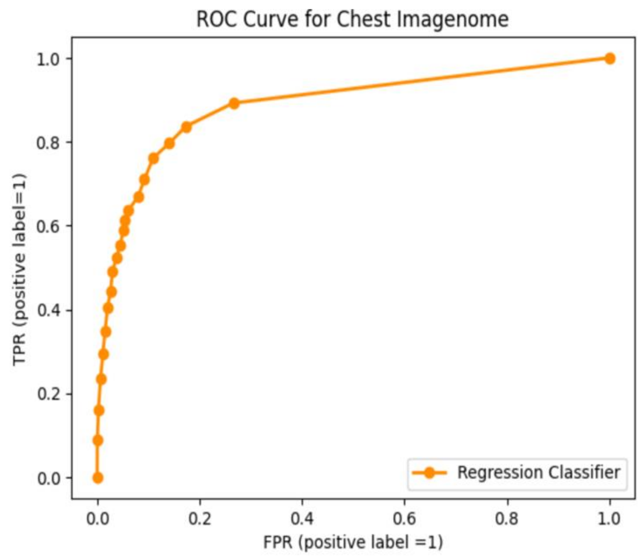
Perturbation Type	Number of Instances
Replace finding and location	40,806,571
Reverse finding	1,546,418
Change Finding location	5,263,608
Replace finding	7,725,552
Reverse finding and change location	4,453,177
Original (Real) finding	1,547,425

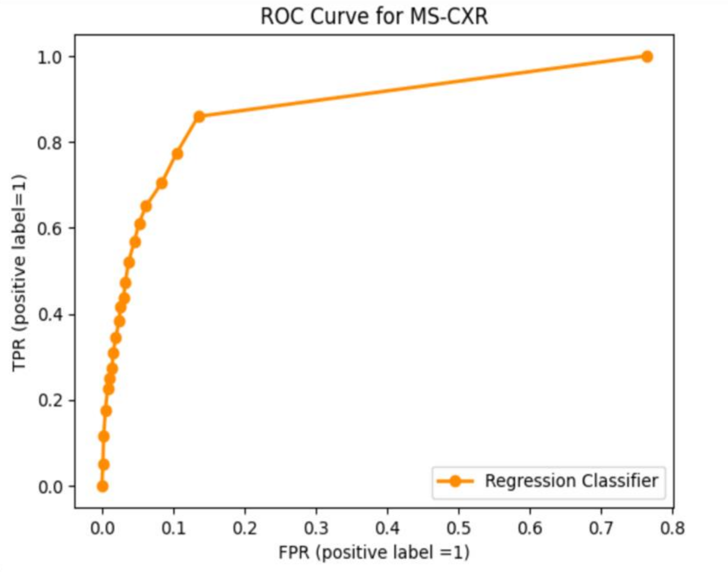


Nearly 24 million samples created. This dataset RADCHECK now contributed to open source on HuggingFace

FC Model performance – Real/Fake classification

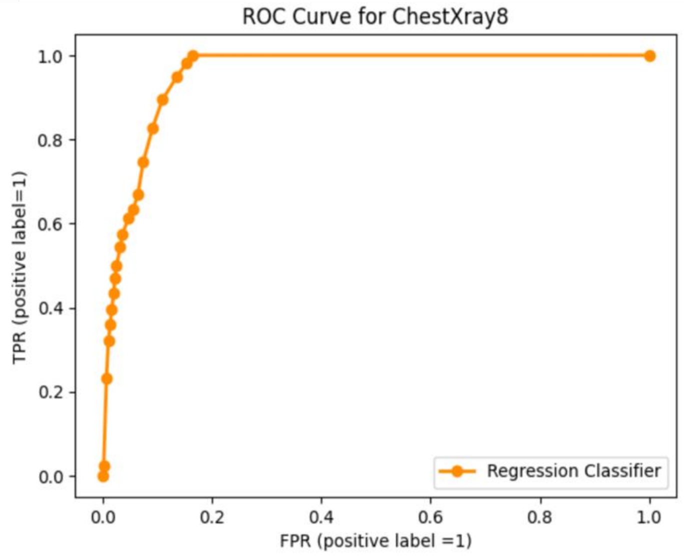
ROC curves

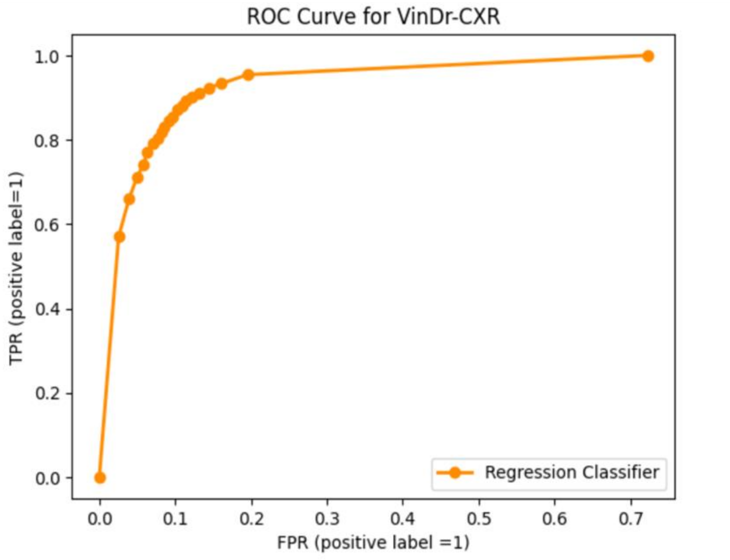




Chest ImaGenome

MS-CXR





ChestXray8

VinDr-CXR

FCModel performance – Classification and localization accuracy

Fact-checking Accuracy (Veracity classification)

Method	CImagenomeG	MS-CXR	ChestXray8	VinDR-CXR
FCModel	0.92	0.94	0.92	0.90
FCSimple	0.84	0.78	0.81	0.83

Mean IOU (Spatial localization accuracy)

Method	CImagenomeG	MS-CXR	ChestXray8	VinDR-CXR
FCModel	0.54	0.53	0.57	0.49
Med-RPG[7]	0.23	0.32	0.28	0.38
Maira-2[4]	0.39	0.48	0.51	0.42

FC Model outperforms in both finding veracity detection and localization

Report Correction Models

Report correction problem

Ground Truth Report

Mild basilar {atelectatic changes} without evidence of acute pneumonia or vascular congestion There may well be a small right pleural effusion. Mild basilar atelectatic changes without evidence of acute pneumonia or vascular congestion. There is an endotracheal tube in place with its tip approximately 3 cm above the carina. Nasogastric tube extends well into the stomach.

Automated Report

There is a new left chest tube in place. The heart and mediastinum are normal in size. The lungs have increased opacity which may indicate atelectasis or pleural effusion on the left. No pneumothorax is seen and there is no visible consolidation. The upper abdomen is clear. The bony structures appear unremarkable.

Corrected Report

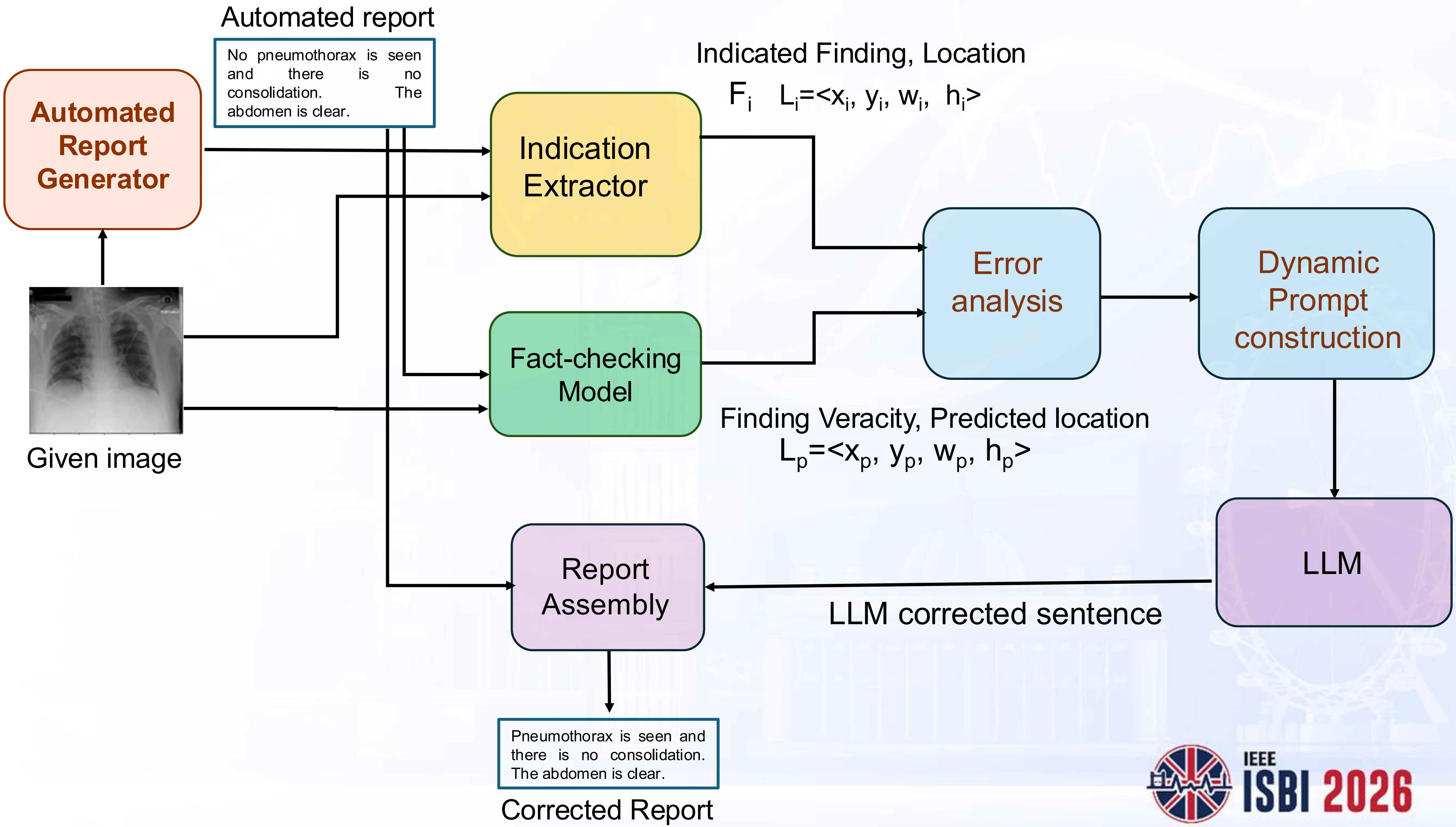
There is a new left chest tube in place. The heart and mediastinum are normal in size. The lungs have increased opacity, which may indicate atelectasis on the left and pleural effusion in the right lung. No pneumothorax is seen and there is no visible consolidation. The upper abdomen is clear. The bony structures appear unremarkable.

Even if the fact-checking model can indicate an error occurred, how to correct it?

Report correction – An agentic workflow

- Involves analyzing the discrepancy between the indicated findings and predicted findings
  - Indicated findings are from generative reports, predictive findings from fact-checking models
- Plan the correction based on discrepancy analysis
- Use a controlled LLM to do the correction
- Fact-checking model output serves as surrogate ground truth

Report correction model



IEEE ISBI 2026

Dynamic Template Generation

Dynamic Prompt Generation

- 6 possible error combinations
  - Allows for FC model error as well
  - Quantizes the error values:
    - Location:
      - Large or small
    - Finding occurrence:
      - Presence or absence
    - Finding correctness:
      - Correct
      - Incorrect
- 5 possible corrections
- Actual sentence and finding filled in the template for dynamic generation

Prompt Templates

l_p	E_p	F_p	$\overline{IOU_{pi}}$	Interpretation	Corrective Action	Prompt
≈ 0	1	Absence	$<= \Gamma$	Both finding and location are correct.	Do nothing as it is correct.	None
> 0	0	Absence	$> \Gamma$	Finding is present as per FC.	Flip the finding from absence to presence. Leave the location unspecified.	Remove "no $< F_i >$ and add "yes $< F_i >$" in the sentence: $< S_i >$
≈ 0	0	Presence	$<= \Gamma$	FC Model is saying finding is absent	Flip the finding from present to absent. Leave the location unspecified as it is either close or unspecified already.	Remove "yes $< F_i >$ and add "no $< F_i >$" in the sentence: $< S_i >$
≈ 0	0	Presence	$> \Gamma$	FC model is saying finding is absent	Flip the finding from present to absent. Remove location hint since the location is far away.	Remove "yes $< F_i >$, add "no $< F_i >$", and remove location $< A_i >$ from the sentence: $< S_i >$
> 0	1	Presence	$<= \Gamma$	Both finding and location are correct. Finding is a presence finding	Do nothing as it is correct.	None
> 0	1	Presence	$> \Gamma$	Finding is correct and present but location is wrong	Drop location only. Keep the finding.	Remove location $< A_i >$ from the sentence: $< S_i >$
All other combinations.				Either E_p or l_p is incorrect.	Do Nothing as FC Model itself is incorrect.	None

IEEE ISBI 2026

Report correction

Report Correction

- Through prompt engineering:
  - <system prompt>
  - <dynamically generated error prompt>
- System prompt:
  - You are correcting a radiology report sentence. Follow the instruction and make a reasonable corrected sentence.
- Many LLMs possible
  - Llama3.2
- Limited correction
  - Partial correction of the sentence for the affected finding
  - Conservative in choice of correction to cause the least harm

System prompt:

You are correcting a radiology report sentence. Follow the instruction and make a reasonable corrected sentence.

Original Sentence	**Incorrect finding**	**LLM Prompt (w/o sentence)**	**LLM-Corrected Sentence**
Left-sided pleural effusion found and the right atelectasis still remains.	yes|pleural effusion	Remove ”pleural effusion” from sentence:	Right atelectasis still remains.
The chest x ray image shows no focal consolidation, pulmonary edema, pleural effusion or pneumothorax	no|pneumothorax	Remove ”no pneumothorax” from the sentence:	The chest X-ray image shows no focal consolidation, pulmonary edema, pleural effusion.
The chest x ray image shows no focal consolidation, pulmonary edema, pleural effusion or pneumothorax	no|pneumothorax	Remove ”no pneumothorax” and add ”yes pneumothorax” in the sentence:	The chest X-ray image shows no focal consolidation, pulmonary edema, or pleural effusion, but shows pneumothorax.
There is left pleural effusion and pneumothorax	yes|pleural effusion	Remove ”left pleural effusion” but keep the ”left pneumothorax” in the sentence:	There is left pneumothorax.
The chest x ray image shows no left focal consolidation.	no|consolidation	Remove ”left” from the sentence:	The chest x ray image shows no focal consolidation.

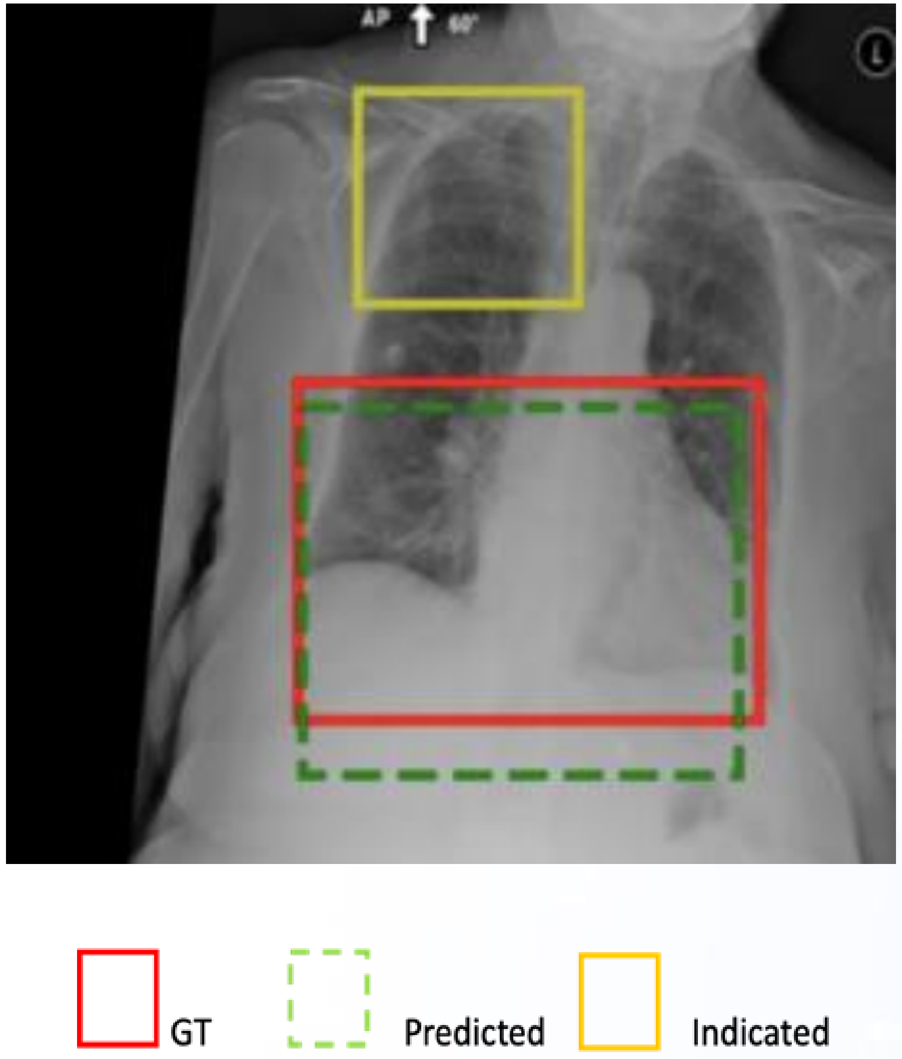
Report correction example – A no error case

Case: Nothing to correct

Ground truth sentence	... pronounced alveolar opacities in a perihilar distribution, right greater than left, likely representing pulmonary edema
Automated report sentence	The pulmonary vasculature is not engorged, and the patient has moderate pulmonary edema on the right.
Target finding	Yes|Pulmonary edema
Indicated location	['right hilar structures', 'right lung’] [0.142, 0.120, 0.343, 0.542]
GT location	[0.120, 0.125, 0.366 0.586]
Predicted location	[0.151, 0.100, 0.351, 0.538]
Predicted veracity label	1 (True)
Dynamic prompt	None
Corrected sentence	The pulmonary vasculature is not engorged, and the patient has moderate pulmonary edema on the right.

IEEE ISBI 2026

Report correction example – An error case



Case: Finding is correct but location is wrong

Ground truth sentence	Opacities are present at the lung bases bilaterally.
Automated report sentence	Subtle opacity in the right upper lung.
Target finding	Yes|Opacity
Indicated anatomical locations	[right upper lung]
GT location	[0.328, 0.419, 0.498, 0.363]
Predicted location	[0.332, 0.442, 0.470, 0.396]
Veracity label	0 (False)
System prompt	You are correcting a radiology report sentence. Follow the instruction and make a reasonable corrected sentence.
Dynamic prompt	Remove location " right upper lung" from the sentence: "Subtle opacity in the right upper lung."
LLM Corrected sentence	**Subtle opacity is noted.**

IEEE ISBI 2026

Report correction evaluation

Clinician Validated Datasets

RADCHECK on HuggingFace

Dataset	Patients Train/Val/Test	Images	Findings	Regions
RadCheck[24]	44,133/6,274/12,538	243,311	49	922,295
CImaGenomeG[20]	288/33/69	461	35	5,477
MS-CXR[25]	478/54/114	925	8	2,254
ChestXray8[26]	457/51/109	880	8	1,571
VinDr-CXR[27]	9,450/1,050/2,250	15,000	23	69,052

Findings (Presence and Absence)

1.	Aortic enlargement	10.	Atelectasis	18.	Calcification
2.	Cardiomegaly	11.	Clavicle fracture	19.	Consolidation
3.	Edema	12.	Emphysema	20.	Enlarged pa
4.	ILD	13.	Infiltration	21.	Infiltrate
5.	Lung cavity	14.	Lung cyst	22.	Lung opacity
6.	Mediastinal shift	15.	Nodule/mass	23.	Other lesion
7.	Pleural effusion	16.	Pleural thickening	24.	Pneumothorax
8.	Pulmonary fibrosis	17.	Rib fracture	25.	Pneumonia
9.	No finding				

Largest set of findings considered!

Report Generators Evaluated

- GPT style:
  - XRayGPT
  - GPT4-0
  - R2GenGPT
- Distillation GPT
  - CV2DistillGPT2
- Phrasal-grounded Llava-style VLMs
  - RGRG
- RAG style:
  - CheXRepair
- Multi-view and longitudinal
  - Maira-2

IEEE ISBI 2026

Measuring report quality

F1-score based on FFL patterns

$$W_j(F_i) = T_i|N_i|C_i|M_1|...|M_j$$

: $W_j(F_{Gi})$ and $W_j(F_{Ak})$ are matching prefixes of corresponding FFLs

$$t_p = |W_j(F_{Gi})|, \ \text{s.t.} W_j(F_{Gi}) = W_j(F_{Ak})$$

$$f_p = |W_j(F_{Ak})| - t_{Gj}, \ f_n = |W_j(F_{Gi})| - t_{Gj}$$

$$F1_{A,G} = \frac{2t_p}{2t_p + f_p + f_n}$$

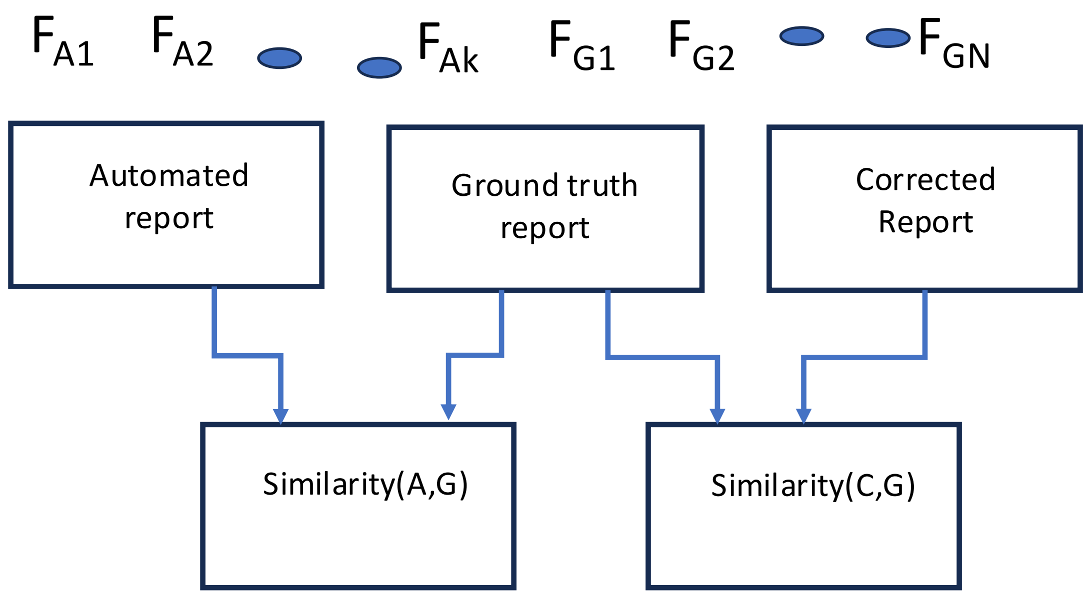
MIOU based on Spatial overlap

$$l_{Ak} = < x_{Ak}, y_{Ak}, w_{Ak}, h_{Ak} > ; \ l_{Gi} = < x_{Gi}, y_{Gi}, w_{Gi}, h_{Gi} >$$

$$I_{ki} = \sum_i \frac{|l_{Ak} \cap l_{Gi}|}{|l_{Ak} \cup l_{Gi}|} \qquad I_{AG} = \sum_i I_{AiGj}$$

$$\boxed{MIOU(A,G) = \frac{2 * I_{AG}}{|F_A| + |F_G|}}$$

How to compare reports?



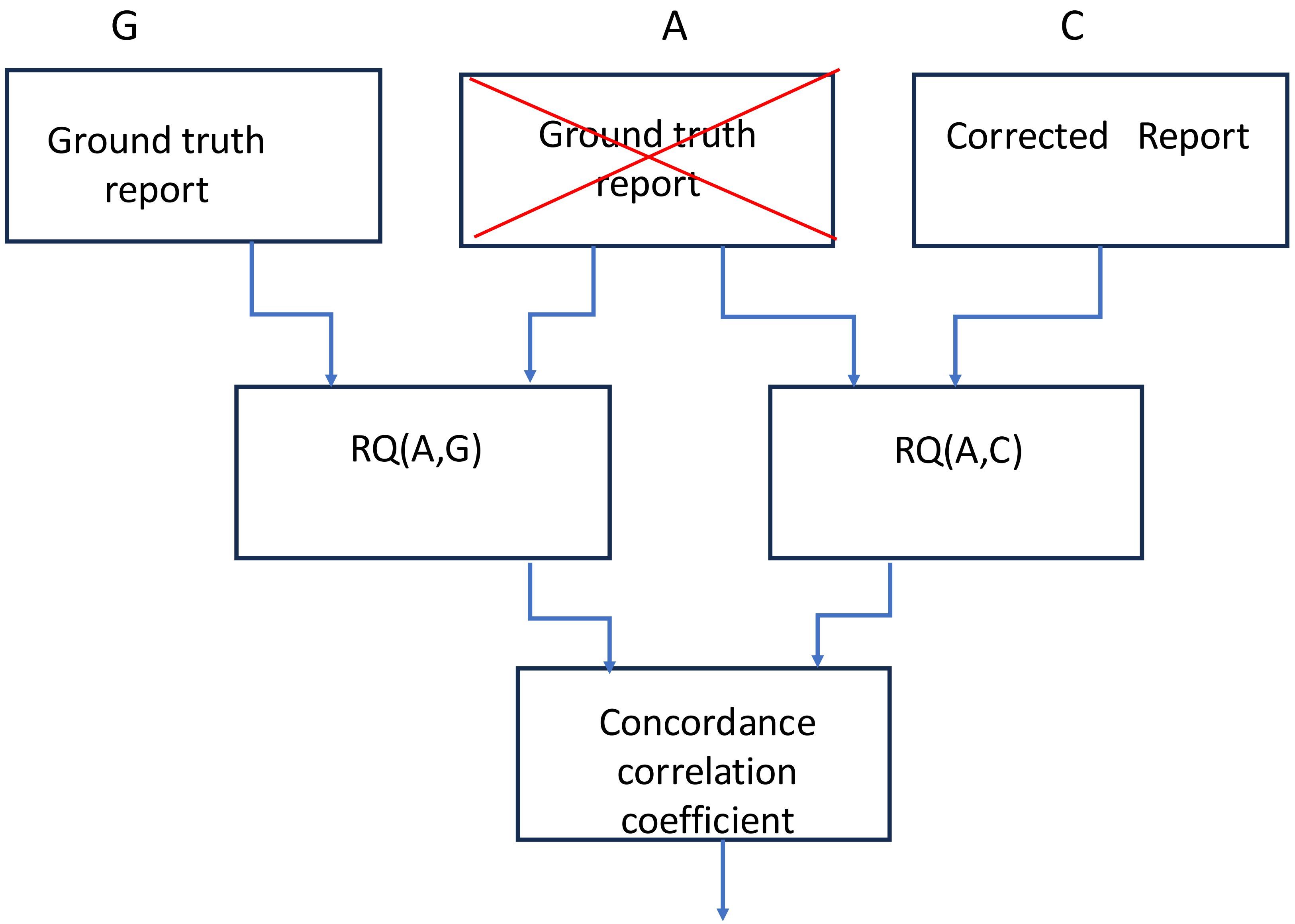
Report Quality Score:

$$\boxed{RQ(A,G) = F1(A,G) + MIOU(A,G)}$$

RQ(C,G) can be computed similarly

Measuring quality improvement during inference

During inference in clinical workflows, the ground truth reports are not available



- Agreement between RQ(A,G) and RQ(A,C) measured by **concordance correlation coefficient (CCC)**:

$$\rho_c = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2}$$

Mean μ_x , μ_y σ_x^2 σ_y^2 Covariance

ρ Pearson's correlation

- If CCC is high => FC model and its correction can serve as a surrogate ground truth during inference!

Report correction performance

All 7 report generators evaluated against the test partitions of 4 datasets

Report Generator	CImaGenomeG *RQ*		MS-CXR *RQ*		ChestX-ray8 *FCScore*		VinDR-CXR *RQ*	
	(A,P)	(A,G)	(A,P)	(A,G)	(A,P)	(A,G)	(A,P)	(A,G)
RGRG[65]	0.541	0.537	0.329	0.308	0.305	0.298	0.549	0.537
XrayGPT[66]	0.622	0.626	0.388	0.391	0.377	0.355	0.618	0.609
GPT4-inhouse	0.658	0.653	0.433	0.426	0.399	0.408	0.636	0.630
R2GenGPT[73]	0.587	0.585	0.377	0.374	0.346	0.333	0.581	0.579
CV2DistillGPT2[45]	0.576	0.573	0.439	0.433	0.427	0.420	0.588	0.6
CheXRepair[54]	0.744	0.733	0.466	0.461	0.439	0.432	0.709	0.714
Maira-2[4]	0.619	0.633	0.423	0.425	0.412	0.419	0.578	0.569

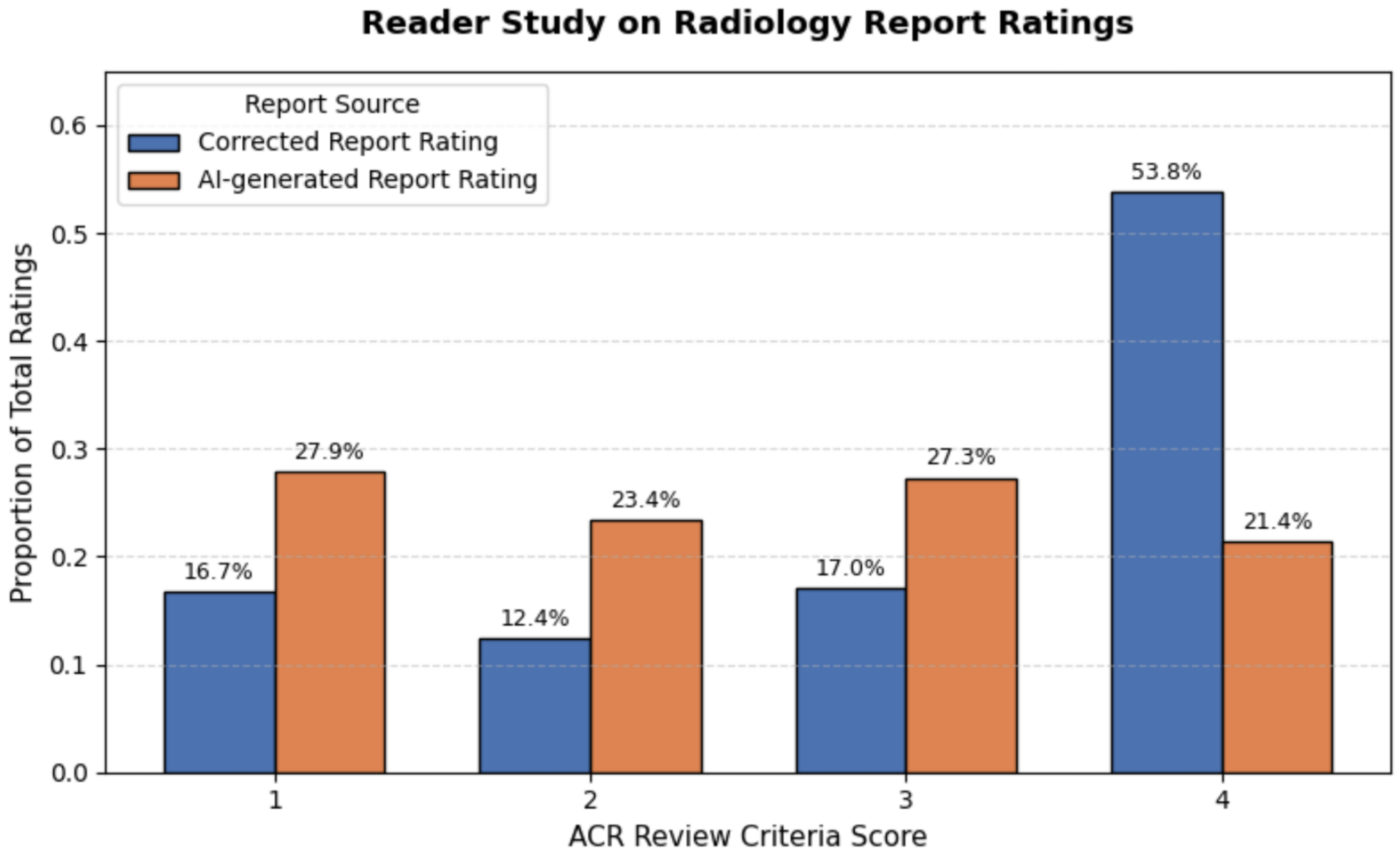
Generator	RadF1		RQ		BLEU		SBERT	
	(A,G)	(C,G)	(A,G)	(C,G)	(A,G)	(C,G)	(A,G)	(C,G)
RGRG[16]	0.52	0.67	0.46	0.52	0.24	0.29	0.33	0.43
XrayGPT[17]	0.39	0.45	0.37	0.48	0.14	0.24	0.26	0.38
GPT4-in	0.43	0.51	0.35	0.47	0.11	0.19	0.09	0.14
R2GenGPT[18]	0.54	0.58	0.37	0.49	0.19	0.27	0.38	0.47
CV2GPT2[19]	0.41	0.49	0.38	0.48	0.14	0.24	0.43	0.54
CheXRepair[3]	0.38	0.43	0.36	0.44	0.21	0.28	0.39	0.46
Maira-2[2]	0.58	0.63	0.52	0.59	0.20	0.26	0.43	0.51
Avg.Improv.	13.5%		27%		48.2%		32.5%	

Model	CCC (A,P)(A,G)
FCModel	0.997
FCSimple	0.831

FC model has high CCC!

MIMC Reader Study

- Compare qualitative difference:
  - corrected and generative AI report
- Test dataset:
  - 160 images from CG-Gold
- Report generators evaluated:
  - XrayGPT
  - GPT4-o
  - ChexAgent
- 3 board-certified clinicians
- Blinded to report source
- ACR Review criteria



Reader study

Reader Study Comparing Radiology Reports

Clinician ID

2

Current case

Case_0

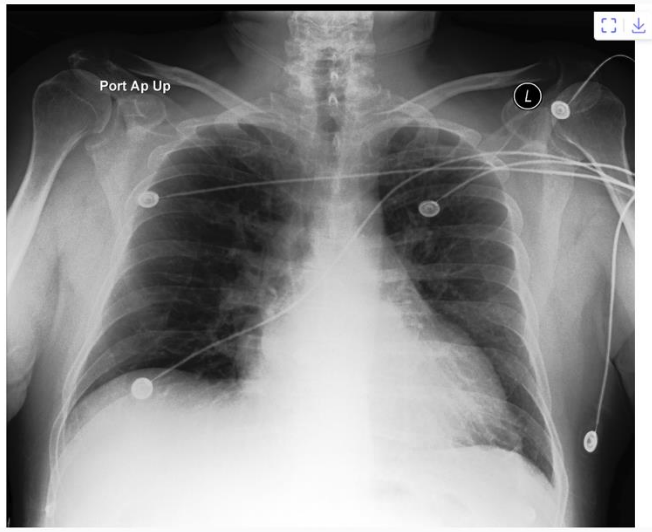
Status

Starting afresh: Case_0...0 out of 299 completed

Report A

No abnormality is found in either the
left lung or the right lung.
No consolidation is noted in either the
left or right lung.
No pleural effusions are found at the
left lung and left costophrenic angles.
Lung opacities are found in the left
lung and right hilar structures.
No pneumothorax is noted.
There is a mild enlargement of the
cardiac silhouette.
No lung opacities are found in either
lung.
No vascular congestion is noted.
No abnormal findings are noted in the
left lung or hilar regions of either lung.
No abnormality is found in either lung.
No atelectasis is found in the left lung
and right lower lung zones.
No mild cardiac silhouette
abnormality is found.

Chest X-ray study



Report B

no evidence of pneumothorax
pulmonary vascular congestion with
increased upper lobe vascular
markings
compressive atelectasis in the lower
lung zones
bilateral pleural effusions with
blunting of costophrenic angles
mild cardiomegaly, unchanged
no focal consolidation
normal hilar and mediastinal contours

Rate each report

Excellent -> Excellent quality report

Average -> Average quality report, i.e. clinically insignificant misinterpretation of difficult non-urgent findings

Below Average -> Below average quality report, i.e. clinically significant misinterpretation of non-urgent findings

Poor -> Poor quality report, i.e. clinically significant misinterpretation of urgent/critical findings

1. Report A Quality Level

- (a) Excellent
- (b) Average
- (c) Below Average
- (d) Poor

2. Report B Quality Level

- (a) Excellent
- (b) Average
- (c) Below Average
- (d) Poor

3. Which report variant do you clinically prefer for active workflow integration?

- Report A (Left Column)
- Report B (Right Column)
- No Preference (Strictly Equal)

4. Which report contains fewer incorrect or unsupported findings relative to the reference?

- Report A (Left Column)
- Report B (Right Column)
- No Preference (About the same)

5. Which report better captures all findings present in the reference?

- Report A (Left Column)
- Report B (Right Column)
- No Preference (About the same)

6. Which report is better written stylewise?

- Report A (Left Column)
- Report B (Right Column)
- No Preference (About the same)

Inference-time alignment models

Using verification models for inference-time alignment

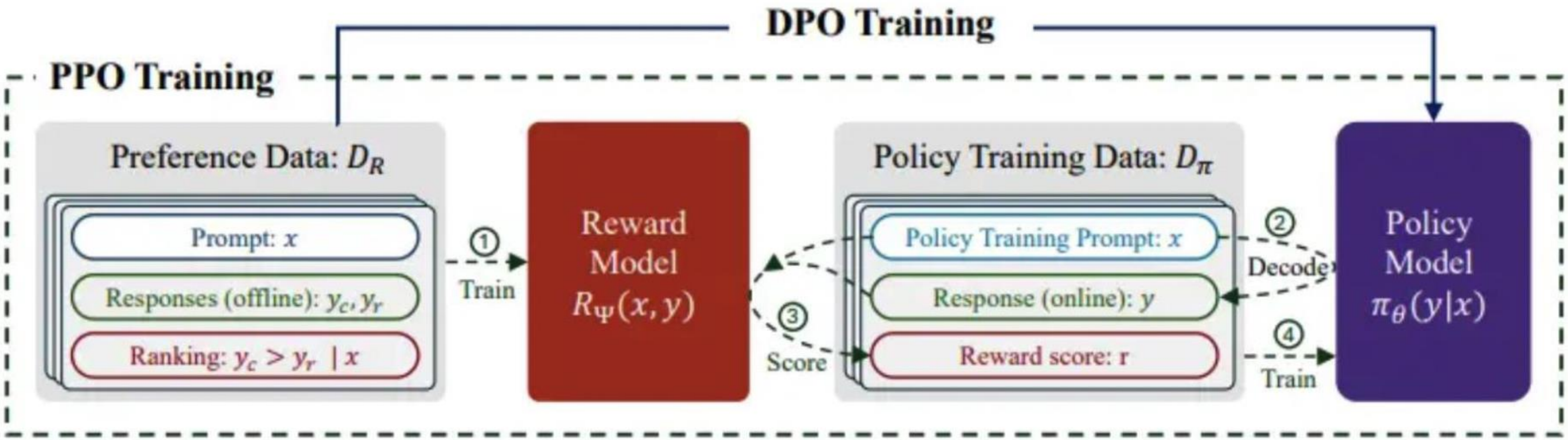
- Improve VLM models through alignment post-training
  - PPO, DPO, GRPO
  - Already being done using manual labels or ground truth reports
- But if these models are already deployed?
  - Can lightweight onsite fine-tuning help?
  - Yes:
    - But no ground truth data is available to correct the output
    - GRPO then relies on consensus leading to false negatives
    - LLM-as-judge are also a problem since none is a 'correct' VLM

Inference-time alignment

- Based on findings
  - They are key to determining if a report is correct
  - Also used in EditGRPO approach
  - Our approach extracts both names and geometric locations of findings
- Using a fact-checking model supplies surrogate ground truth
- Forms a corrected report for updating the policy
  - The policy update rule requires textual response as the preferred response.
- Similar approach to direct preference optimization (DPO) done at training time

Direct policy optimization (DPO)

- No separate reward model needed
- Implicit reward included in the policy update rule.
  - Analytical mapping derived from reward functions to optimal policies, which transform a loss function over reward functions into a loss function over policies
- Human feedback still taken for preferences after SFT of the LLM
- The LLM model is optimized without using RL by minimizing the L_{DPO} loss.



[Direct Preference Optimization: Your Language Model is Secretly a Reward Model, NeurIPS'23]

Direct preference optimization

- Not an RL method

- a **frozen reference policy** π_{ref}
- a **trainable policy** π_θ

Human-curated preference training data

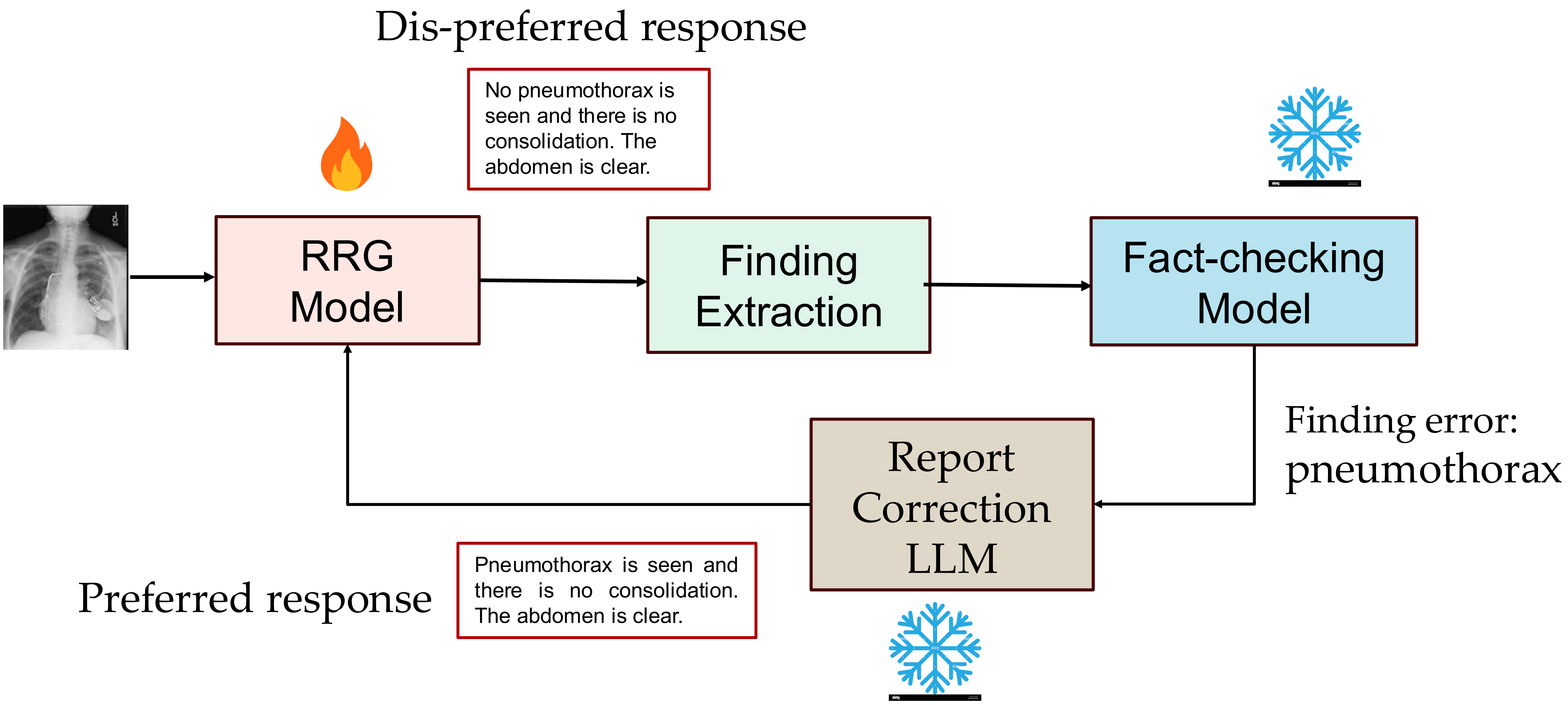
$$(x, y^+, y^-)$$

Maximize

$$\log \sigma\left(\beta\left[\log \pi_\theta(y^+|x) - \log \pi_\theta(y^-|x) - (\log \pi_{\text{ref}}(y^+|x) - \log \pi_{\text{ref}}(y^-|x))\right]\right)$$

- Increase probability of preferred answer
- Decrease probability of rejected answer
- Relative to the reference policy
- offline, supervised optimization on preference pairs.
- Still needs preference pairs to be assembled

Automatic preference optimization (APO) – A new inference-time alignment



Preference alignment method

Create a phrase-grounded preference dataset from RRG model output

$$D = \{x^i, y_w^i, L_w^i, y_l^i, L_l^i\}_{i=1}^{N}$$

$$L_{APO}(\pi_\theta, \pi_{ref}) = -\mathbb{E}(x, y_w, y_l)\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)} + \gamma\log\frac{O_\theta}{O_{ref}}\right)\right]$$

KL-Divergence of preferred response with reference

KL-Divergence of dispreferred response with reference

Phrasal grounding loss

π_{ref} : Original policy

π_θ : Updated policy

$x^{\{i\}}$ = initial prompt + Image

$y_{\{w\}}^{\{i\}}$ = preferred response

$y_{\{l\}}^{\{i\}}$ = dis-preferred response

M_w = Preferred response findings

M_l = Dispreferred response findings

M_{ref} = Original response findings

$$L_l = \{l_j\}_{j=1}^{M_l} = \{< u_j, v_j, w_j, h_j >\}_{j=1}^{M_l}$$

$$L_w = \{l_j\}_{j=1}^{M_w} = \{< u_j, v_j, w_j, h_j >\}_{j=1}^{M_w}$$

$$O_{ref} = \frac{1}{M_{ref}}\sum_j \frac{|l_{wref\,j} \cap l_{lref\,j}|}{|l_{wref\,j} \cup l_{lref\,j}|}$$

$$O_\theta = \frac{1}{M_\theta}\sum_j \frac{|l_{w\theta j} \cap l_{l\theta j}|}{|l_{w\theta j} \cup l_{l\theta j}|}$$

APO updates policy:

- Without deviating too much from the original response
- Keeping the corrected findings
- Increasing the overlap between the locations of corresponding findings

Experiments

- Conducted over:
  - Multiple datasets
    - comprehensive finding coverage
  - Multiple RRG models
  - Multiple error metrics to judge quality improvement
- Compared with DPO, PPO, GRP)

Data	Patients Train/Test	Images	Labels	Regions
RadCheck[14]	44,133/12,538	243,311	49	922,295
CImaGenomeG[39]	288/69	461	35	5,477
MS-CXR[10]	478/114	925	8	2,254
ChestXray8[35]	457/109	880	8	1,571
VinDr-CXR[19]	9,450/2,250	15,000	23	69,052

RRG Models

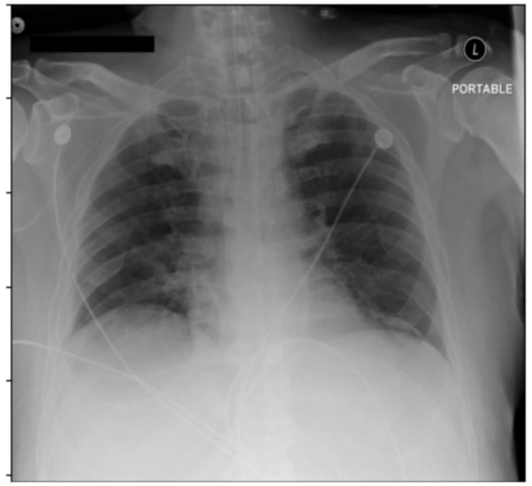
RGRG
XraGPT
GPT-4o
R2GenGPT
CV2GPT
CheXRepair
Maira-2

Quality metrics

BLEU
SBERT
Radgraph F1
RQ
Green

APO training results

Chest X-ray image



Ground Truth Report

Mild basilar {atelectatic changes} without evidence of acute pneumonia or vascular congestion There may well be a small right pleural effusion. Mild basilar atelectatic changes without evidence of acute pneumonia or vascular congestion. There is an endotracheal tube in place with its tip approximately 3 cm above the carina. Nasogastric tube extends well into the stomach.

Dis-preferred Response

There is a new left chest tube in place. The heart and mediastinum are normal in size. The lungs have increased opacity which may indicate atelectasis or pleural effusion on the left. No pneumothorax is seen and there is no visible consolidation. The upper abdomen is clear. The bony structures appear unremarkable.

Preferred Response

There is a new left chest tube in place. The heart and mediastinum are normal in size. The lungs have increased opacity, which may indicate atelectasis on the left. No pneumothorax is seen and there is no visible consolidation. The upper abdomen is clear. The bony structures appear unremarkable.

APO Improved Response

There is a new left chest tube in place. The heart and mediastinum are normal in size. The lungs have increased opacity, indicating left atelectasis. Small right pleural effusion cannot be ruled out. No pneumothorax is seen and there is no visible consolidation. The upper abdomen is clear. The bony structures appear unremarkable

Results

- **17-40% improvement in quality**
- Multiple RRG models
- Multiple error metrics to judge quality improvement
- Comparison with DPO, PPO, GRPO
- Comprehensive coverage of findings

Generator	RadgraphF1			RQ			BLEU			SBERT			Green		
	(A,G)	(P,G)	(C,G)	(A,G)	(P,G)	(C,G)	(A,G)	(P,G)	(C,G)	(A,G)	(P,G)	(C,G)	(A,G)	(P,G)	(C,G)
RGRG[33]	0.52	0.67	0.69	0.46	0.52	0.56	0.24	0.29	0.34	0.33	0.43	0.46	0.39	0.43	0.45
XrayGPT[34]	0.39	0.45	0.47	0.37	0.48	0.52	0.14	0.24	0.28	0.26	0.38	0.43	0.28	0.32	0.36
GPT4-in	0.43	0.51	0.54	0.35	0.47	0.51	0.11	0.19	0.24	0.09	0.14	0.19	0.33	0.36	0.45
R2GenGPT[36]	0.54	0.58	0.59	0.37	0.49	0.53	0.19	0.27	0.34	0.38	0.47	0.50	0.29	0.34	0.37
CV2GPT2[20]	0.41	0.49	0.51	0.38	0.48	0.54	0.14	0.24	0.29	0.43	0.54	0.56	0.30	0.34	0.37
CheXRepair[27]	0.38	0.43	0.44	0.36	0.44	0.49	0.21	0.28	0.34	0.39	0.46	0.49	0.32	0.34	0.39
Maira-2[1]	0.58	0.63	0.66	0.52	0.59	0.61	0.20	0.26	0.29	0.43	0.51	0.54	0.43	0.47	0.49
Avg.Improv.	13.5%		20.2%	27%		35.0%	48.2%		78.9%	32.5%		42%	11.4%		23.8%

Generator	MCXR			C8			Vindr		
	RadF1	SBERT	Green	RadF1	SBERT	Green	RadF1	SBERT	Green
RGRG	0.46/0.49	0.38/0.45	0.40/0.47	0.32/0.43	0.36/0.45	0.36/0.42	0.41/0.52	0.39/0.45	0.41/0.48
XrayGPT	0.37/0.40	0.32/0.41	0.34/0.39	0.38/0.43	0.31/0.40	0.28/0.36	0.32/0.40	0.38/0.46	0.39/0.45
GPT4-in	0.43/0.48	0.39/0.44	0.41/0.46	0.33/0.42	0.37/0.45	0.35/0.43	0.42/0.54	0.40/0.47	0.40/0.49
R2GenGPT	0.49/0.55	0.40/0.47	0.41/0.48	0.39/0.43	0.40/0.47	0.42/0.50	0.40/0.48	0.42/0.47	0.42/0.49
CV2GPT2	0.40/0.44	0.37/0.43	0.39/0.43	0.33/0.40	0.34/0.40	0.35/0.45	0.42/0.48	0.39/0.46	0.41/0.49
CheXRepair	0.37/0.40	0.34/0.41	0.38/0.41	0.32/0.39	0.32/0.41	0.32/0.40	0.38/0.45	0.31/0.44	0.41/0.48
Maira-2	0.49/0.56	0.43/0.59	0.51/0.57	0.39/0.47	0.49/0.56	0.51/0.58	0.39/0.48	0.45/0.56	0.42/0.54
Avg. Impv	8.9%	16.8%	13.1%	21.2%	10.7%	10.1%	13.3%	11.1%	9.5%

Generator	CG				MCXR				C8				VinDr				ChexPert			
	RQ				*RQ*				*RQ*				*RQ*				*RQ*			
	(A,G)	(C,G)	(D,G)	(R,G)	(A,G)	(C,G)	(D,G)	(R,G)	(A,G)	(C,G)	(D,G)	(R,G)	(A,G)	(C,G)	(D,G)	(R,G)	(A,G)	(C,G)	(D,G)	(R,G)
RGRG[33]	0.46	0.56	0.51	0.46	0.51	0.62	0.57	0.52	0.38	0.49	0.43	0.39	0.51	0.63	0.57	0.51	0.43	0.54	0.50	0.44
XrayGPT[34]	0.37	0.52	0.46	0.36	0.45	0.49	0.46	0.44	0.35	0.42	0.38	0.40	0.46	0.54	0.49	0.45	0.49	0.54	0.51	0.48
GPT4-inhouse	0.35	0.51	0.46	0.38	0.46	0.54	0.51	0.46	0.41	0.48	0.43	0.42	0.51	0.58	0.54	0.51	0.47	0.54	0.51	0.47
R2GenGPT[36]	0.37	0.53	0.43	0.38	0.44	0.54	0.48	0.45	0.38	0.47	0.42	0.39	0.51	0.57	0.54	0.52	0.51	0.61	0.55	0.52
CV2DisGPT2[20]	0.38	0.54	0.45	0.38	0.39	0.49	0.43	0.38	0.41	0.47	0.44	0.42	0.52	0.6	0.56	0.53	0.48	0.59	0.55	0.47
CheXRepair[27]	0.36	0.49	0.43	0.35	0.45	0.51	0.48	0.45	0.43	0.49	0.46	0.44	0.51	0.59	0.56	0.50	0.49	0.59	0.52	0.49
Maira-2[1]	0.52	0.61	0.57	0.51	0.47	0.58	0.53	0.48	0.41	0.49	0.45	0.42	0.50	0.61	0.57	0.51	0.52	0.61	0.56	0.53
Avg. Impv.	35.2%		18.6%	0.5%	18.9%		9.1%	0.22%	19.6%		9.1%	4.17%	17.1%		8.8%	2.5%	18.7%		9.3%	2.9%

Summary

- Verification AI is about discriminative and generative models focused on detecting, correcting and fine-tuning generative AI models at inference time.
- Current generative AI offers no guarantees of performance
- Using Verification AI methods in combination with generative AI could lead towards more trustworthy AI solutions in future.
- RadCheck – A synthetic data set of 27 million pairs of image and finding in HuggingFace