Phrase-grounded Fact-checking for Automatically Generated Chest X-ray Reports

Razi Mahmood[1], Diego Machado-Reyes[1], Joy Wu[2,3], Parisa Kaviani[4],
Ken C.L. Wong[2], Niharika D'Souza[2], Mannudeep Kalra[4], Ge Wang[1],
Pingkun Yan[1], Tanveer Syeda-Mahmood[2,3]

[1] Rensselaer Polytechnic Institute, NY, USA ,
[2] IBM Research, Almaden, CA, USA
[3] Stanford University, CA, USA,
[4] Massachusetts General Hospital (MGH), Boston, USA
mahmor@rpi.edu, stf@us.ibm.com

Abstract. With the emergence of large-scale vision language models (VLM), it is now possible to produce realistic-looking radiology reports for chest X-ray images. However, their clinical translation has been hampered by the factual errors and hallucinations in the produced descriptions during inference. In this paper, we present a novel phrase-grounded fact-checking model (FC model) that detects errors in findings and their indicated locations in automatically generated chest radiology reports. Specifically, we simulate the errors in reports through a large synthetic dataset derived by perturbing findings and their locations in ground truth reports to form real and fake findings-location pairs with images. A new multi-label cross-modal contrastive regression network is then trained on this dataset. We present results demonstrating the robustness of our method in terms of accuracy of finding veracity prediction and localization on multiple X-ray datasets. We also show its effectiveness for error detection in reports of SOTA report generators on multiple datasets achieving a concordance correlation coefficient of 0.997 with ground truth-based verification, thus pointing to its utility during clinical inference in radiology workflows.

Keywords: Fact-checking · report generation · vision language models.

1 Introduction

Preliminary radiology reports generated by automated report generation models are valuable in emergency room settings, where radiologists may not be immediately available and rapid interpretation is required[22]. Current methods of report generation are predominantly based on vision language models (VLM)[1, 14, 4, 19, 18] which still suffer from hallucinations and factual errors that limit their clinical applicability[3]. Strategies to correct such models exist such as direct policy optimization (DPO)[17, 5, 15, 33] or proximal policy optimization (PPO)[32] with reward models[34] directly tap into the generative decoder parameters to compute hallucination risk scores [5]. However, they are applicable

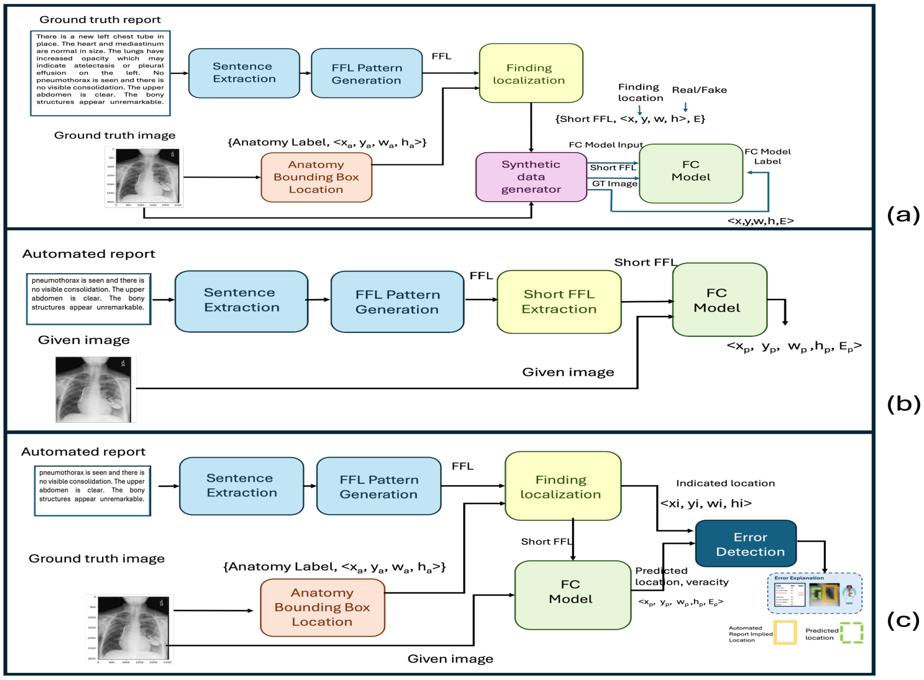
Ground truth report
There is a new left chest tube in place. The heart and mediastinum are normal in size. The lungs have increased opacity which may indicate atelectasis or pleural effusion on the left. No pneumothorax is seen and there is no visible consolidation. The upper abdomen is clear. The bony structures appear unremarkable.
Sentence Extraction
FFL Pattern Generation
FFL
Finding localization
Finding location
Real/Fake
{Short FFL, <x, y, w, h>, E}
Ground truth image
{Anatomy Label, <x_a, y_a, w_a, h_a>}
Anatomy Bounding Box Location
FC Model Input
Synthetic data generator
Short FFL
GT Image
FC Model
FC Model Label
<x,y,w,h,E>
(a)
Automated report
pneumothorax is seen and there is no visible consolidation. The upper abdomen is clear. The bony structures appear unremarkable.
Sentence Extraction
FFL Pattern Generation
FFL
Short FFL Extraction
Short FFL
FC Model
Given image
<x_p, y_p, w_p, h_p, E_p>
Given image
(b)
Automated report
pneumothorax is seen and there is no visible consolidation. The upper abdomen is clear. The bony structures appear unremarkable.
Sentence Extraction
FFL Pattern Generation
FFL
Finding localization
Indicated location
<xi, yi, wi, hi>
Error Detection
Ground truth image
{Anatomy Label, <x_a, y_a, w_a, h_a>}
Anatomy Bounding Box Location
Short FFL
FC Model
Predicted location, veracity
<x_p, y_p, w_p, h_p, E_p>
Error Explanation
Automated Report Implied Location
Predicted location
Given image
(c)

Fig. 1. Illustration of workflow for FC model training (red lines), inference (green lines), and error detection and explanation (orange lines) using common modules.

during training or fine-tuning stages. Methods of fact-checking at inference time exist but they often consult external knowledge [15, 8, 20] to spot the factual errors which are unsuitable for radiology reports as the report needs to be specific to the patient-image. Similarly, methods that use generic large language models (LLMs) as judges to verify the radiology report text [1, 18, 31] are not suitable either, since they themselves have hallucinations, and may not corroborate their deductions with the patient-specific image.

Thus, there is a need to develop an independent fact-checking method for the clinical inference phase to bootstrap radiology report generation. Realizing this, we had earlier attempted to build a simple fact-checking model using a pre-trained vision-language model (CLIP) and a binary SVM to classify sentences as real or fake in automated reports[10]. However, being based on full sentences, it was sensitive to writing styles in reports. Further, it used a frozen encoder and did not offer any explanation of the errors nor perform a phrasal grounding of the findings in images.

In this paper, we introduce an innovative method of fact-checking chest X-ray radiology reports during clinical inference with the following novel contributions. First, we derive a large synthetic dataset of over 27 million images paired with real and fake findings to simulate errors in reports through perturbation of their identities and location descriptions in ground truth reports. This dataset is now being contributed to open source. We also develop a new multi-label contrastive regression model for fact-checking (FC model) that is trained to discriminate and anatomically ground the real and fake findings. Through extensive testing, we show that the FC model can detect radiology report errors with a concordance correlation coefficient of 0.997 with ground truth-based verification making it a potential surrogate for ground truth during clinical inference.

2 Method

Our overall approach to factual error detection in automated radiology reports consists of training a fact-checking model (Figure 1a), using it in inference mode on automated reports to record predicted findings and their locations (Figure 1b), and recording the deviations of implied findings from automated reports from predictions as shown in Figure 1c. As can be seen from Figure 1, the workflows use common pre-processing modules of sentence extraction, finding extraction and anatomy localization, but are fed different image-text pairs during training and inference. Specifically, the training workflow depicted in Figure 1a involves: (i) finding localization, (ii) synthetic data generation and (iii) FC model training. Step (i) extracts anatomical locations (L) from ground truth images (I), findings (F) from their reports, and collates to generate bounding boxes $< x, y, w, h >$ for findings. In step (ii) synthetic perturbations are applied to generate real/fake pairs $< F, I, x, y, w, h, E >$ where E is the veracity label. In step (iii) the FC model is trained using $< F, I >$ as input and $< x, y, w, h, E >$ as output. During inference, the FC model is given findings extracted from automated reports and their image as input, to predict the output $< x_p, y_p, w_p, h_p, E_p >$ where $< x_p, y_p, w_p, h_p >$ is the bounding box and Ep is the predicted real/fake label as shown in Figure 1b. Finally, the error detection and quantification workflow shown in Figure 1c recovers the indicated location from automated report and compares it to the predicted finding and location using an error measure that results in a visual explanation.

2.1 Training dataset generation

For training data generation, we use prior work on finding extraction[22] and anatomical region detection [27, 21] as pre-processing. To make our fact-checking approach agnostic to sentence writing styles, we abstract the described findings in sentences into a simplified structured form called FFL (fine-grained finding labels) using the method described in [22] and as illustrated in Table 1. Each finding is normalized to a standard vocabulary (*e.g.*, pulmonary vasculature engorged -> vascular congestion) using a comprehensive clinician-curated chest X-ray lexicon of 101,088 distinct FFL [27, 21] which are sufficient to capture the variety seen in automatically generated reports. The FFL extraction algorithm reported in [22] had a 97% accuracy and was seen as sufficient for our pre-processing. In addition to finding descriptions, we also use the anatomical location algorithm described in [28, 29] to locate bounding boxes in any frontal chest X-ray image for 36 anatomical regions cataloged in the chest X-ray lexicon [27, 21]. Its accuracy was previously assessed at 0.896 precision and 0.881 recall, and was used to generate the Chest ImaGenome benchmark dataset[29].

Table 1. Illustration of FFL.

Sentence	**Simplified FFL**
Pleural vasculature is not engorged and the patient has moderate pulmonary edema on the right	anatomicalfinding | no | vascular congestion | lung anatomicalfinding | yes | pumonary edema|right lung

Table 2. Illustration of synthetic perturbations to produce the training dataset for the FC model. Only the core finding in column 2 for simplicity.
'E' : (0=non-existent finding, 1=existing finding)

Synthetic Perturbation	Generated Finding	Label (<xy,w,h,E>)
Original	yes|edema	$< 0.14, 0.13, 0.72, 0.56, 1 >$
Reversal	no|edema	$< 0, 0, 0, 0, 0 >$
Relocate	yes|edema	$< 0.85, 0.74, 0.10, 0.21, 0 >$
Relocate	yes|edema	$< 0.90, 0.70, 0.10, 0.20, 0 >$
Substitution	yes|lung cyst	$< 0.02, 0.48, 0.10, 0.14, 0 >$

Let $< I, R >$ be the sample set of ground truth image-report pairs in a gold dataset D. Let $F = \{F_j\}$ be the total list of possible findings in chest X-ray datasets. The set of real Finding-location (FL) pairs extracted by the pre-processing per sample $D_i =< I_i, R_i >\in D$ can be denoted by $FL_{iReal} = \{fl_{ij}\} = \{< f_{ij}, l_{ij} >\}$ where:

$$f_{ij} =< T_{ij}|N_{ij}|C_{ij} >, l_{ij} =< x_{ij}, y_{ij}, w_{ij}, h_{ij} > . \tag{1}$$

Here $f_{ij} \in F_{iReal}$ is the jth real finding in report R_i and l_{ij} is the bounding box for the finding f_{ij} in image I_i starting at (x_{ij}, y_{ij}) of width w_{ij} and height h_{ij} in normalized coordinates ranging from 0 to 1.

Let $L_j = \{l_{ij}\}$ be the list of all normalized locations accumulated across all images of D for a finding F_j. With normalized coordinates, and since we pick among the valid finding locations, any synthetic location generated for F_j will be valid for some image in the dataset.

The errors found in generated reports are known to include false predictions, incorrect finding locations, omissions, or incorrect severity assessments[30]. We focus on the first two errors so that given a real finding f_{ij} at location l_{ij} for a sample D_i, we create 3 variants to reflect (a) reversal of polarity (b) relocation of the finding (c) and substitution with appropriate relocation as $FL_{iFake} = \{< \overline{fl_{ij}}, fl_{ik}, fl_{mn} >\}$, where $\overline{fl_{ij}}$ is the reversed finding, fl_{ik} is finding f_{ij} relocated to a random new position $l_k \in L_j$, and fl_{mj} is obtained by randomly substituting finding f_j with $f_m \notin F_i$ at location $l_n \in L_m$ taking care to avoid repeats and contradictions. Table 2 shows synthetic perturbations created from an original finding "yes|edema" based on the operations above.

2.2 Building the FC model

The end-to-end architecture of the FC model is illustrated in Figure 2. We use the encodings of images and FFL text to learn a joint embedding space that is designed to separate the real FFL labels from fake labels using supervised contrastive learning[7]. The embeddings from real and fake text-image pairs are then concatenated to learn an inner regression network to predict both the location and veracity of the finding.

Let z_i be the vision projection encoder output, and let $z_{f_{ij}}$ be the text encodings of findings for each sample $D_i = (I_i, F_i)$ where $f_{ij} \in F_i = F_{iReal} \cup F_{iFake}$ are the real and fake labels per sample. We define a multi-label cross-modal supervised contrastive loss per sample as:

$$\mathcal{L}_{SupC_i} = \frac{-1}{|F_{iReal}|} \sum_{f_{ij} \in F_{iReal}} log \frac{e^{s_i f_{ij}/\tau}}{\sum_{a_{ik} \in F_{iFake}} e^{s_{ia_{ik}}/\tau}} \tag{2}$$

where $s_{if_{ij}} = z_i \cdot z_{f_{ij}}$ is the pairwise cosine similarity between image and textual embedding vectors from the real findings $f_{ij} \in F_{iReal}$, and $s_{ia_{ik}} = z_i \cdot z_{a_{ik}}$ is with the fake findings where $a_{ik} \in F_{iFake}$. The overall loss is obtained by averaging across all the samples in the batch. Here τ is the temperature parameter. This formulation results in a non-diagonal similarity matrix as shown in Figure 2 and differs significantly from existing VLM contrastive encoders based on CLIP who all assume a diagonal similarity matrix[16, 10, 18] and are self-supervised. It also differs from supervised contrastive learning which was previously unimodal and used for image classification from augmented version of images treated as positive samples[7]. *To our knowledge, the supervised contrastive learning formulation has not been used to develop vision-language encoders with real and fake labels.*

Next, the inner regression network takes the projected joint embeddings $T_{ijReal} = [z_i|z_{f_{ij}}]$ of image I_i paired with real finding label $f_{ij} \in F_{iReal}$ or fake labels $T_{ijFake} = [zi|z_{a_{ik}}]$ where $a_{ik} \in F_{iFake}$ and the corresponding supervision label $Y_g =< Y_{1g}, Y_{2g} >$ where $Y_{1g} =< x, y, w, h >$ and $Y_{2g} = E = 1$ for the real finding and 0 otherwise. Using $Y_p =< Y_{1p}, Y_{2p} >$ as the prediction from the network, we can express the regression loss per sample as

$$\begin{aligned}\mathcal{L}_{Reg_i} = &\underbrace{|Y_{1p} - Y_{1g}|}_{\mathcal{L}_1(Y_{1p},Y_{1g})} + \underbrace{\frac{|Y_{1p} \cap Y_{1g}|}{|Y_{1p} \cup Y_{1g}|} - \frac{|C_{Y_{1p},Y_{1g}} \backslash Y_{1p} \cap Y_{1g}|}{|C_{Y_{1p},Y_{1g}}|}}_{\mathcal{L}_{\text{giou}}(Y_{1p},Y_{1g})} \\ &+ \underbrace{|Y_{1p} - Y_{1g}|^2}_{\mathcal{L}_{\text{mse}}(Y_{1p},Y_{1g})} - \underbrace{[Y_{2g}\mathbf{log}(Y_{2p}) + (1 - Y_{2g})\mathbf{log}(1 - Y_{2p})]}_{\mathcal{L}_{BCE}(Y_{2p},Y_{2g})}\end{aligned} \tag{3}$$

where $C_{Y_{1p},Y_{1g}}$ is the convex hull of the bounding boxes defined by Y_{1p} and Y_{1g}.

The loss function reflects the dual attributes being optimized, namely, the location and the veracity of the finding. The L1 loss and generalized IOU loss have previously been used for regression[2]. However, since in our case, the negative findings have bounding box coordinates as $< 0, 0, 0, 0 >$ which poses a problem for generalized IOU when the prediction error is small. For this reason, and to ensure smooth convergence, we added the mean square penalty. Finally, for the veracity indicator variable E, we use the binary cross entropy loss.

Implementation details: We used a chest X-ray pre-trained CLIP encoder (151,277,313 parameters) [18] and retained its image encoder (ViT-B/32) and

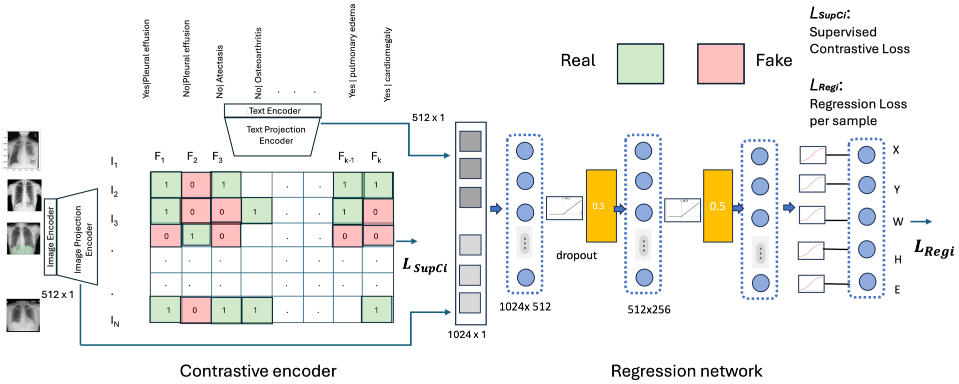

Fig. 2. Illustration of the architecture of our FC model. The real FFL are taken as positive and the fake FFL as negative in the contrastive formulation.

Table 3. Details of datasets used in experiments. Here CImagenomeS stands for Chest ImagGenome silver dataset.

Dataset	**Patients**	**Images**	**Findings**	**Regions**	**Real/Synth.**
	Train/Val/Test				
CImagenomeS[29]	44,133/6274/12,538	243,311	49	922,295	1.616M/27.047M
CImaGenomeG[29]	288/33/69	461	35	5,477	4,063/23,463
MS-CXR[6]	478/54/114	925	8	2,254	2,247/24,338
ChestXray8[25]	457/51/109	880	8	1,571	1,571/10,137
VinDr-CXR[12]	9,450/1,050/2,250	15,000	23	69,052	47,973/132,632

text encoder (masked self-attention Transformer). The joint embedding projection layers of CLIP (768x512 for image and 512x512 for text) were, however, fresh-trained using the new supervised contrastive formulation derived from real-fake labels. The regression network (657,413 parameters) consisted two linear layers, two drop out layers with RELU for intermediate layers and separate sigmoidal functions for producing the output regression vectors as shown in Figure 2. To train this network in an end-to-end fashion, the losses defined in Equations 2 and 3 were applied at the respective heads shown in Figure 2. The FC model was trained for 100 epochs using the AdamW optimizer on an NVIDIA A100 GPU with 40GB of memory and a batch size of 32. The cosine annealing learning rate scheduler was used with the maximum learning rate of 1e-5 and 50 steps for warm up.

2.3 Error detection using the FC model

The FFL extracted from an automated report and the image are used by the FC model to predict the veracity of the finding label and its location as shown in Figure 1b. To quantify the error, we use a phrase-grounded error measure called the FC score[11] which was shown to outperform other evaluation measures such as Radgraph F1, SBERT, and BLEU score. Specifically, we calculate the error

Table 4. This table illustrates multiple aspects of the FC model evaluation. The FC model performance under different ablation architecture configurations across multiple datasets are rows in the first 4 rows. The last two rows show comparison of our FC model's phrasal grounding and real/fake classification performance against SOTA methods.

Method	CImaGenomeG		MS-CXR		ChestX-ray8		VinDR-CXR	
	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU	Accuracy	MIOU
FCRegComb.	**0.92**	**0.54**	**0.94**	**0.53**	**0.92**	**0.57**	**0.90**	**0.49**
FCRegBCE	0.88	0.49	0.92	0.46	0.90	0.53	0.88	0.45
FCRegDual	0.87	0.51	0.89	0.49	0.87	0.51	0.86	0.47
FCRegSep	0.89	0.38	0.89	0.39	**0.92**	0.42	0.89	0.37
Med-RPG[2]	-	0.23	-	0.32	-	0.28	-	0.38
Maira-2[1]	-	0.39	-	0.48	-	0.51	-	0.42
R/F Model[10]	0.84	-	0.78	-	0.81	-	0.83	-

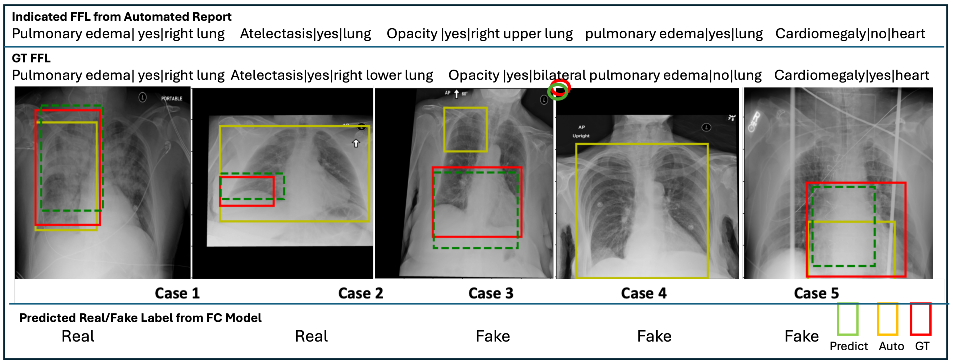

Fig. 3. Illustration of error detection and localization for 5 sentences from reports generated by X-rayGPT[24]. The legend for bounding boxes: predicted finding location: Green, indicated finding: orange, ground truth finding: red.

detected by FC model as $RQ(A,P) = 1 - \text{FCScore(A,P)}$ as :

$$RQ(A,P) == 1 - \frac{1}{2}\left(\frac{|E_{pj}=1|}{\sum_{E_{pj}\in E_p} E_{pj}} + \frac{1}{|L_p|}\sum_j \frac{|L_{Aj}\cap L_{pj}|}{2|L_{Aj}\cup L_{pj}|}\right) \qquad (4)$$

Here E_{pj} is a predicted veracity for an indicated label $F_{Aj} \in F_A$ in the automated report, and L_{Aj}, L_{pj} are the indicated locations from automated reports and the predicted locations from the FC model respectively computed as shown in Figure 1c.

3 Results

We conducted several experiments using chest X-ray datasets with location and finding annotations shown in Table 3. Of these, Chest ImaGenome gold (CImagenomeG)[29] dataset was set aside for error detection evaluation as it had a

Table 5. Illustrating the effectiveness of the FC model in assessing errors in generated reports. High concordance can be seen between error detection using ground truth (A,G) and error detection using FC model (A,P) in all cases.

Report Generator	**CImaGenomeG**		**MS-CXR**		**ChestX-ray8**		**VinDR-CXR**	
	RQ		*RQ*		*RQ*		*RQ*	
	(A,P)	(A,G)	(A,P)	(A,G)	(A,P)	(A,G)	(A,P)	(A,G)
RGRG[23]	0.541	0.537	0.329	0.308	0.305	0.298	0.549	0.537
XrayGPT[24]	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[26]	0.587	0.585	0.377	0.374	0.346	0.333	0.581	0.579
CV2DistillGPT2[13]	0.576	0.573	0.439	0.433	0.427	0.420	0.588	0.6
CheXRepair[18]	0.744	0.733	0.466	0.461	0.439	0.432	0.709	0.714
Maira-2[1]	0.619	0.633	0.423	0.425	0.412	0.419	0.578	0.569

complete set of ground truth reports, clinician-verified findings and their locations[29]. The training partitions of the rest of the datasets were used for the generation of the synthetic dataset yielding over 27 million samples as shown in Table 3.

Automated report generators evaluated: We selected several SOTA report generators whose code was freely available as shown in Table 5. All report generators were given the same prompt, and automated reports collected for the 439 images of the (CImagenomeG) were retained for error analysis.

Real/Fake classification performance: We evaluated the accuracy of FC model's in FFL veracity prediction using the test partitions of the datasets shown in Table 3. The model consistently yielded an accuracy over 90% for real/fake classification, as shown in Table 4. By using 10 fold cross-validation in the generation of the (70-10-20) splits for the datasets, the average accuracy of the test sets lay in the range 0.92 ± 0.12.

Anatomical grounding performance: Figure 3 illustrates sample explainable error detection by the FC model on XrayGPT-generated reports[24]). By comparing the bounding box locations and predicted labels to ground truth FFLs, we observed that the FC model correctly flags errors and localizes findings with greater overlap with ground truth. In fact, the mean IOU with the ground truth bounding boxes ranged from 0.49-0.57 as shown in Table 4 (rows 1-4), across various model architectures.

Comparison to related methods: With no prior work on fact-checking with phrasal grounding for chest X-ray reports, we compared to the nearest methods that either do phrasal grounding (MED-RPG[2],Maira-2[1]) or real/fake classification (the R/F Model from [10]). The results are shown in Table 4 in the last three rows recording the relevant numbers for a regressor or classifier respectively. In comparison to pure phrase grounding or real/fake classification only, our method predicts both veracity and location of findings, and outperforms these methods across all the datasets.

Ablation studies: We conducted ablation studies using 4 different architectures, namely, (a) end-to-end training as shown in Figure 2 (FCRegComb), (b) replacing supervised contrastive loss with BCE loss (FCRegBCE), (c) using a generic pre-built CLIP encoder with regressor (FCRegSep), and (d) using a dual head regressor with separate loss functions for regression and classification (FCRegDual). The results of real/fake classification and phrasal grounding shown in Table 4 indicate that combining the contrastive encoder with the regressor in an end-to-end fashion gave the best performance.

Fact-checking report performance: Fact-checking involves computing the error between the indicated (A) and predicted FL pairs (P) as $RQ(A, P)$ and comparing it to $RQ(A, G)$ of indicated FL pairs with the ground truth G. These results are summarized in Table 5 averaged across all images for each report generator tested. As can be seen, the $RQ(A, P)$ has good correlation with $RQ(A, G)$ and the overall concordance correlation coefficient[9] with the ground truth at 0.997. In comparison, using the real/fake classifier model[10], the concordance correlation coefficient was lower at 0.831 since the location errors could not be verified. These results show the potential of fact-checking models for error detection during inference in clinical workflows even when no ground truth is available.

4 Conclusions

In this paper, we have presented a new fact-checking model for chest X-ray reports that detects errors in findings and their reported locations. The model has a high concordance coefficient with the ground truth for error estimation pointing to its utility as a surrogate for ground truth during inference. Future work on the FC model will address findings omitted from reports, and explore ways of incorporating it during the training phases to further improve report generators.

Disclosure of Interests. The authors have no competing interests.

References

1. Bannur, S., et al.: Maira-2: Grounded radiology report generation (2024), https://arxiv.org/abs/2406.04449
2. Chen, Z., Zhou, Y., Tran, A., Zhao, J., Wan, L., Ooi, G., Cheng, L.E., Thng, C., Xu, X., Liu, Y., Fu, H.: Medical phrase grounding with region-phrase context contrastive alignment. In: MICCAI (2023)
3. Gale, W., Oakden-Rayner, L., Carneiro, G., Bradley, A.P., Palmer, L.J.: Producing radiologist-quality reports for interpretable artificial intelligence. arXiv preprint arXiv:1806.00340 (2018)
4. Gao, D., Kong, M., Zhao, Y., Huang, J., Huang, Z., Kuang, K., Wu, F., Zhu, Q.: Simulating doctors' thinking logic for chest x-ray report generation via transformer-based semantic query learning. Medical Image Analysis **91**, 102982 (1 2024). https://doi.org/10.1016/J.MEDIA.2023.102982
5. Hardy, R., Kim, S.E., Ro, D.H., Rajpurkar, P.: Rextrust: A model for fine-grained hallucination detection in ai-generated radiology reports (12 2024), https://arxiv.org/abs/2412.15264v3

6. Johnson, A.E.W., et al.: Mimic-cxr: A large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
7. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv preprint arXiv:2004.11362 (2020)
8. Lab, N.J.: Ai will start fact-checking. we may not like the results., https://www.niemanlab.org/2022/12/ai-will-start-fact-checking-we-may-not-like-the-results/
9. Lin, L.I.K.: A concordance correlation coefficient to evaluate reproducibility. Biometrics **45**, 255 (3 1989). https://doi.org/10.2307/2532051, https://pubmed.ncbi.nlm.nih.gov/2720055/
10. Mahmood, R., Wang, G., Kalra, M., Yan, P.: Fact-checking of ai-generated reports. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **14349 LNCS**, 214–223 (7 2023)
11. Mahmood, R., Yan, P., Reyes, D.M., Wang, G., Kalra, M.K., Kaviani, P., Wu, J.T., Syeda-Mahmood, T.: Evaluating automated radiology report quality through fine-grained phrasal grounding of clinical findings (12 2024), https://arxiv.org/abs/2412.01031v2
12. Nguyen, H.Q., et al.: Vindr-cxr: An open dataset of chest x-rays with radiologist's annotations. Scientific Data 2022 9:1 **9**, 1–7 (7 2022)
13. Nicolson, A., Dowling, J., Koopman, B.: Improving chest X-ray report generation by leveraging warm starting. Artificial Intelligence in Medicine **144**, 102633 (2023)
14. Pang, T., Li, P., Zhao, L.: A survey on automatic generation of medical imaging reports based on deep learning. BioMedical Engineering OnLine **22**, 48 (2023), https://doi.org/10.1186/s12938-023-01113-y
15. Passi, K., Shah, A.: Distinguishing fake and real news of twitter data with the help of machine learning techniques. ACM International Conference Proceeding Series pp. 1–8 (8 2022)
16. Radford, A., et al.: Learning transferable visual models from natural language supervision. Proceedings of Machine Learning Research **139**, 8748–8763 (2 2021), https://arxiv.org/abs/2103.00020v1
17. Rafailov, R., Sharma, A., Mitchell, E., Manning, C.D., Ermon, S., Finn, C.: Direct preference optimization: Your language model is secretly a reward model. In: Thirty-seventh Conference on Neural Information Processing Systems (2023), https://arxiv.org/abs/2305.18290
18. Ramesh, V., Chi, N.A., Rajpurkar, P.: Improving radiology report generation systems by removing hallucinated references to non-existent priors. Proceedings of Machine Learning Research **193**, 456–473 (9 2022)
19. Ranjit, M., Ganapathy, G., Manuel, R., Ganu, T.: Retrieval augmented chest x-ray report generation using openai gpt models. Proceedings of Machine Learning Research **219**, 650–666 (5 2023), https://arxiv.org/abs/2305.03660v1
20. Suprem, A., Pu, C.: Midas: Multi-integrated domain adaptive supervision for fake news detection (2022), https://arxiv.org/pdf/2205.09817.pdf
21. Syeda-Mahmood, T., et al.: Extracting and learning fine-grained labels from chest radiographs. In: Proc. American Medical Association Annual Symposium (AMIA). p. 1190–1199 (Nov 2020)
22. Syeda-Mahmood, T., Wong, K.C.L., Gur, Y., Wu, J.T., Jadhav, A., Kashyap, S., Karargyris, A., Pillai, A., Sharma, A., Syed, A.B., Boyko, O., Moradi, M.: Chest x-ray report generation through fine-grained label learning. In: MICCAI-2020 (2020)
23. Tanida, T., Müller, P., Kaissis, G., Rueckert, D.: Interactive and explainable region-guided radiology report generation. In: CVPR (2023)

24. Thawkar, O., Shaker, A., Mullappilly, S.S., Cholakkal, H., et al.: Xraygpt: Chest radiographs summarization using medical vision-language models (6 2023), https://arxiv.org/abs/2306.07971v1
25. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray: Hospital-scale chest x-ray database and benchmarks on weakly supervised classification and localization of common thorax diseases. In: Deep Learning and Convolutional Neural Networks for Medical Imaging and Clinical Informatics (2019)
26. Wang, Z., Liu, L., Wang, L., Zhou, L.: R2gengpt: Radiology report generation with frozen llms. Meta-Radiology **1**, 100033 (11 2023)
27. Wu, J., et al.: Ai accelerated human-in-the-loop structuring of radiology reports. In: Proc. American Medical Association Annual Symposium (AMIA). p. 1305–1314 (Nov 2020)
28. Wu, J., et al.: Automatic bounding box annotation of chest x-ray data for localization of abnormalities. Proceedings - International Symposium on Biomedical Imaging **2020-April**, 799–803 (4 2020)
29. Wu, J.T., Agu, N.N., Lourentzou, I., Sharma, A., Paguio, J.A., Yao, J.S., Dee, E.C., Mitchell, W., Kashyap, S., Giovannini, A., Celi, L.A., Moradi, M.: Chest imagenome dataset for clinical reasoning (7 2021), https://arxiv.org/abs/2108.00316v1
30. Yu, F., Endo, M., Krishnan, R., Langlotz, C.P., Venugopal, V.K., Correspondence, R.: Evaluating progress in automatic chest x-ray radiology report generation. Patterns **4**, 100802 (2023). https://doi.org/10.1016/j.patter.2023.100802, https://doi.org/10.1016/j.patter.2023.100802
31. Zhang, S., Sambara, S., Banerjee, O., Acosta, J., Fahrner, L.J., Rajpurkar, P.: Radflag: A black-box hallucination detection method for medical vision language models. Proceedings of Machine Learning Research (10 2024), http://arxiv.org/abs/2411.00299
32. Zheng, R., et al.: Secrets of rlhf in large language models part i: Ppo (2023)
33. Zhou, Y., Cui, C., Yoon, J., Zhang, L., Deng, Z., Finn, C., Bansal, M., Yao, H.: Analyzing and mitigating object hallucination in large vision-language models. arXiv preprint arXiv:2310.00754 (2023)
34. Ziegler, et al.: Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593 (2019), https://arxiv.org/abs/1909.08593