Quality Control for Radiology Report Generation Models via Auxiliary Auditing Components

Hermione Warr[1], Yasin Ibrahim[1], Daniel R. McGowan[2,3], Konstantinos Kamnitsas[1,4,5]

[1] Department of Engineering Science, University of Oxford, Oxford, UK
{first_name.last_name}@eng.ox.ac.uk
[2] Department of Oncology, University of Oxford, UK
[3] Department of Medical Physics and Clinical Engineering, Oxford University Hospitals NHS FT, Oxford, UK
[4] Department of Computing, Imperial College London, London, UK
[5] School of Computer Science, University of Birmingham, Birmingham, UK

Abstract. Automation of medical image interpretation could alleviate bottlenecks in diagnostic workflows, and has become of particular interest in recent years due to advancements in natural language processing. Great strides have been made towards automated radiology report generation via AI, yet ensuring clinical accuracy in generated reports is a significant challenge, hindering deployment of such methods in clinical practice. In this work we propose a quality control framework for assessing the reliability of AI-generated radiology reports with respect to semantics of diagnostic importance using modular auxiliary auditing components (ACs). Evaluating our pipeline on the MIMIC-CXR dataset, our findings show that incorporating ACs in the form of disease-classifiers can enable auditing that identifies more reliable reports, resulting in higher F1 scores compared to unfiltered generated reports. Additionally, leveraging the confidence of the AC labels further improves the audit's effectiveness. Code will be made available at: https://github.com/hermionewarr/GenX_Report_Audit

Keywords: Radiology Report Generation · Error Detection · Vision Language Modeling · Reliability

1 Introduction

Radiologists are facing mounting challenges managing the growing volume of medical imaging data under resource constraints, necessitating more efficient and innovative solutions for image analysis and interpretation. Automated (or partly automated) radiology report generation is one such solution. Since the introduction of transformers [28], which significantly enhanced language generation, numerous studies have investigated radiology report generation from medical images, particularly chest X-rays (CXR) [4, 7, 9, 15, 22, 26]. However, deployment of language models to medical applications faces a variety of challenges, with a key area being the need for factual accuracy. Various methods

have been developed to address this - particularly useful have been the introduction of clinical metrics to supplement the standard natural language processing (NLP) metrics used to evaluate report generation quality [4,18]. Previous efforts to improve the quality of the generated reports have incorporated extra information during training, such as bounding boxes for improved localisation, or have aimed to refine diagnostic accuracy by including loss functions that reward clinical accuracy [5,7,21,26]. Additionally, more complex architectures, such as memory-augmented attention, have been developed to facilitate better image understanding [6,21].

Besides efforts to improve report generation accuracy, another line of work is error detection in language modeling. Most work in this area has targeted the natural language domain. The two main types of uncertainty estimation in large language models (LLMs) are statistical uncertainty (SU) and in-dialogue uncertainty (IDU) [27]. The latter relies on large pretrained models to admit when they don't know an answer [13,16]. SU relates certainty to the model's output entropy, though in language modelling, it must address the coherence of entire sentences, as well as on a per token basis [14,20]. These methods often depend on extensive labelled datasets in the natural language domain, which do not adequately capture the specialised terminology and context of clinical settings [27]. Few works have investigated the possibility of error detection in reports written by radiologists [8,12,29,30]. To our knowledge, no previous work has investigated error detection for *automated* report generation, which is the focus of our study.

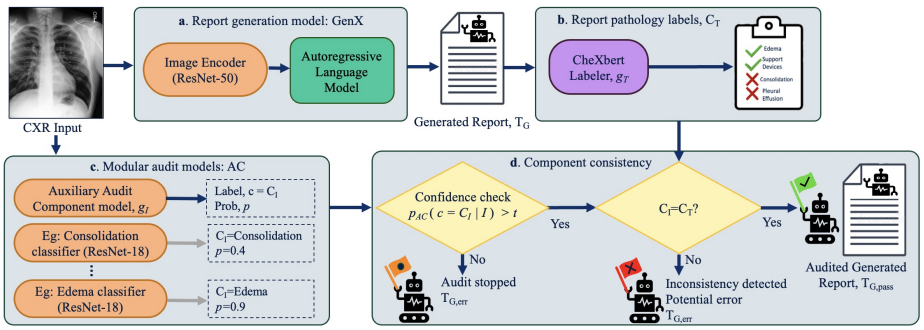
We introduce an auditing framework for identifying potential errors in AI-based radiology report generation, using modular auxiliary auditing components (AC) for quality control. Our study focuses on chest-Xray (CXR) imaging, defining ACs as image-based disease classifiers to extract semantics relevant to diagnosis. We develop a report generation model, GenX, which performs competitively with previous works and serves as basis for testing our auditing framework. Reports are audited using ACs that provide disease classification and confidence levels. Reports with consistent diagnoses across the pipeline meet our criteria, while inconsistencies flag potential errors. Additionally, by leveraging the ACs' classification confidence, we enforce stricter quality control, requiring consistency with the highest confidence predictions. Our evaluation on MIMIC-CXR shows a significantly higher F1 score for reports meeting the audit's criteria. Results demonstrate that consistency between report and auxiliary components is a promising auditing mechanism for automated report generation.

2 Methodology

2.1 Report Generation Model - GenX

The proposed framework for auditing radiology report generation is shown in Fig. 1. The backbone of the framework is the report generation model. The generalised vision language modeling problem of generating a textual sequence

Fig. 1: Proposed error detection pipeline of radiology report generation using auxiliary auditing components. The standard report generation pipeline (a) is followed by the CheXbert labeler, g_T (b) that extracts pathology labels, C_T, from the reports, that are semantically meaningful for clinical diagnosis. (c) Modular image-based audit models, AC, here disease classifiers, predict disease class labels, C_I, based on image, I. (d) If the labels predicted based on image (C_I) and report (C_T) are consistent, the report's contents are deemed likely reliable. In case of inconsistency, $C_I \neq C_T$, the report is flagged as less reliable, potentially containing an error. If the image-based classifiers have predictive confidence below threshold t, auditing can be deferred to user due to uncertainty.



T given an input image I can be framed as estimating the conditional distribution $p(T|I)$, as a product of the conditional probabilities:

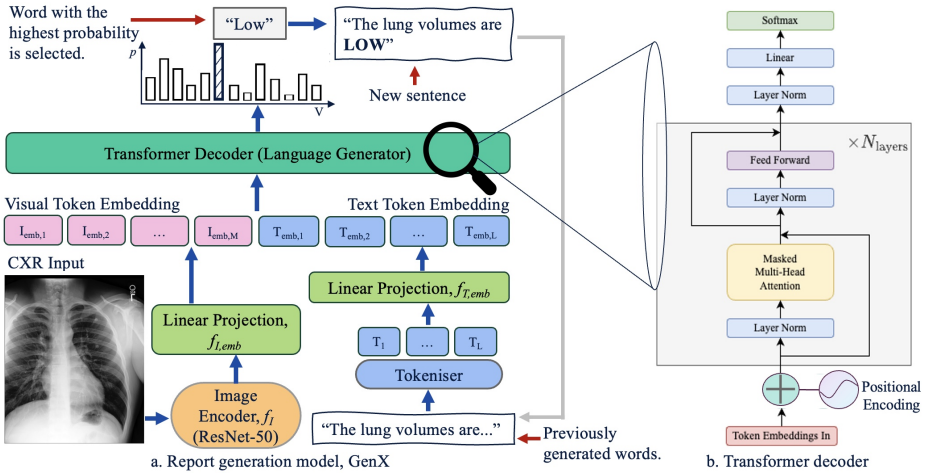
$$p(T|I) = \prod_{l=1}^{L} p(T_l \mid T_{<l},\ I;\ \theta), \tag{1}$$

where T is modelled as a sequence of word tokens $\{T_1, ..., T_L\}$, with $T_{<l}$ being the set of tokens preceding T_l, where L is the number of tokens in the sequence, and θ the model parameters [1, 22].

An overview of our report generation model, *GenX*, is shown in Fig. 2. The model is trained to maximise the probability of generating the target sequence from the image, $p(T|I)$ from eq 1. Image features $I_{feat} = f_I(I) \in \mathbb{R}^{D_I}$, with dimensionality D_I, are extracted by image encoder f_I. A linear layer, $f_{I,emb}$, projects I_{feat} to the space of token embeddings that will be input to the language generator, giving $I_{emb} = \{I_{emb,1} \ldots I_{emb,M}\} = f_{I,emb}(I_{feat}) \in \mathbb{R}^{D_T \times M}$, a set of M image token embeddings with dimensionality D_T, similarly to LLaVA [19]. A tokeniser maps text to a set of L token vectors $\{T_l\}_{1:L} \in \mathbb{R}^{L \times V}$, where V is the vocabulary size. Each T_l is embedded by a linear layer to $T_{l,emb} = f_{T,emb}(T_l) \in \mathbb{R}^{D_T}$. The concatenated embeddings of image and text tokens $\{I_{emb,1} \ldots I_{emb,M}, T_{emb,1} \ldots T_{emb,L}\}$ are passed to a Transformer decoder (Fig. 2b) which predicts a probability distribution over V vocabulary tokens, expressing which is more probable in the sequence. The decoder has N_{layers} with Attention blocks that each has N_{heads}

parallel heads [28]. The model is trained using the common paradigm of diagonal attention masking [28]. A report T_G is generated with auto-regressive inference, starting only with image token embeddings I_{emb} as input, then predicting the first and each following word token as the one with highest predicted probability.

Fig. 2: Illustration of one inference step for radiology report generation (a). Embeddings of image and previously generated text tokens are passed to the autoregressive language model (b) which outputs a probability distribution over the vocabulary, to predict the next word in the sequence.



2.2 Auditing Generated Reports via Auxiliary Classifiers

While some types of mistakes in AI-generated radiology reports, such as grammatical errors, are merely inconvenient, semantic mistakes related to disease diagnosis can be detrimental for patient care. Consider a semantic concept of interest, denoted c. We focus on the case where c represents a specific disease, with the class label $C \in \{1, 0\}$ indicating the presence or absence of that disease. To audit whether a generated report T_G is semantically correct about c, we need to extract the value C_T described in text T_G via a mapping function $g_T : T_G \rightarrow c$. This can be any text-based classifier, or prompt-based LLM that can answer questions in the form "Does this text suggest evidence or not for disease c?", which have become widely available. Herein we use a pretrained CheXbert text-classifier [25] as g_T, obtaining $C_T = g_T(T_G)$ (Fig 1b).

We then introduce to the framework an independent model $g_I : I \rightarrow c$, which infers value C_I for concept of interest c from image I (Fig. 1c). We term such models auxiliary auditing components (AC). In the examined case when c is existence of specific disease, g_I is a classifier predicting $C_I = 1$ if image I shows

the disease or $C_I = 0$ otherwise, with model confidence $p_{AC}(c = C_I|I)$. If multiple concepts c are of interest (e.g. different diseases), an independent AC model can be used per c, for ease of modular development, or a multi-label AC model.

Auditing is performed by comparing values C_T and C_I about the concept c of interest (e.g. disease) extracted from the report or the image respectively (Fig 1d). To consider the contents of the report reliable with respect to c, agreement between the report and AC is required, i.e. $C_T = C_I$. We further extend this by taking into consideration the confidence p_{AC} of the AC. We only consider the auditing successful if it satisfies the additional requirement that the value $C_T = C_I$ has been inferred by the AC with confidence p_{AC} over a pre-determined threshold t. Formally, for a successfully completed auditing we require:

$$(C_I = C_T) \wedge (p_{AC}(c = C_I \mid I) \geq t), \tag{2}$$

where $\wedge$ is the logical *and* operator. The set of generated reports that pass the audit for a particular concept c is denoted by $T^{(c)}_{G,pass} = \{T_G|$ Eq. 2 is $True\}$. The remainder, $T^{(c)}_{G,err} = \{T_G|$ Eq. 2 is $False\}$, can be deferred to the human user for review as potentially inaccurate for that disease.

The framework is designed based on ACs that are modular, independent models. The design draws inspiration from N-Version programming [3] and the principle of redundancy for reliability in software engineering. Assuming independence of components ($GenX$, g_T, g_I) with label error rates e_{C_T} and e_{C_I}, the error rate of the auditing framework is $e = e_{C_T} \cdot e_{C_I}$, which is lower than the individual error rates of C_T and C_I, as failure of auditing requires both components to fail simultaneously to incorrectly satisfy Eq. 2. Modular ACs enable independent component training and validation prior to integration into the audit pipeline.

3 Experiments and Results

3.1 Data

Experiments are performed using MIMIC-CXR [11], containing chest X-rays (CXR), associated radiology reports, and disease labels for ∼228,000 studies of over 65,000 patients. As common practice, we consider only the frontal-view images, and only the "findings" section of the reports, which describes evidence of pathologies and support devices. The labels include 14 classes (12 diseases, 1 support devices, 1 no findings) and can be positive, uncertain, negative, or not mentioned. We report 5 out of 14 classes separately (Atelectasis, Cardiomegaly, Consolidation, Edema, Pleural Effusion) to facilitate comparison with literature [10]. We divide data into splits for training (226,261), validating hyperparameters (1,864), and testing performance (3,595). To train and evaluate the report generation, we further exclude studies that do not include a "findings" section, resulting in 155,322 training, 1,231 validation and 2,607 testing samples. Finally, to help address class imbalance and noisy labels, following past works [9, 10, 21, 26], we consider uncertain disease labels as positive and class no-mention as negative.

3.2 Evaluating the Report Generation Model

Experimental settings: We here show that our report generation model, *GenX*, is of representative quality in comparison to previous work to serve as effective backbone for the study of our auditing framework. To build GenX we use ResNet-50 as the image encoder f_I, which we pre-train as a multi-label classifier of the 14 classes described above. Its penultimate layer of dimension $D_I = 512$ is then projected to $M = 10$ image token embeddings with $D_T = 512$. These are concatenated with up to $L = 512$ text embeddings, resulting from a GPT-2 tokeniser ($V \approx 50,257$) [24], followed by a projection, $D_T = 512$. The Transformer has $N_{layers} = 8$, $N_{heads} = 8$, and dimension $D_{ff} = 2048$ of its feed-forward layer. GenX is then trained end-to-end (Fig. 2), learning parameters of the randomly initialised projections and Transformer, and fine-tuning the f_I encoder.

Table 1: Report generation performance of our model, GenX, compared to previous works. To assess reports' clinical correctness, we report macro and micro averages of F1 score, recall (R) and precision (P), averaged over the 5 most commonly reported classes (Atelect., Cardiom., Consolid., Edema, Pleural Eff. cf Sec. 3.1) and over all 14 classes, in format (5 / 14 classes), to facilitate comparison with literature. Our model performs well in terms of F1 that quantifies clinical correctness, surpassed only by more complex frameworks using reinforcement [7] or additional labels [21], demonstrating that GenX is an effective backbone for our auditing study.

Model	Disease Classif. Metrics (Report)						NLP Metrics		
	macro average			*micro average*					
	F1(5/14)	**R(5/14)**	**P(5/14)**	**F1(5/14)**	**R(5/14)**	**P(5/14)**	**BL-1**	**M**	**R-L**
R2Gen [4]	-/27.6	-/27.3	-/33.3	-/-	-/-	-/-	35.3	14.2	27.7
Ratchet [9]	38.8/-	-/-	-/-	-/-	-/-	-/-	23.2	10.1	24.0
$\mathcal{M}^2$ Trans [21]	47.3/30.2	55.1/36.0	45.8/40.5	56.7/-	65.1/-	50.3/-	-	-	-
RGRG [26]	-/-	-/-	-/-	54.7/-	61.7/-	49.1/-	37.3	16.8	26.4
A_t+TE+RG [7]	-/-	-/-	-/-	-/53.7	-/49.6	-/58.5	49.0	21.3	40.6
GenX	46.8/30.0	43.1/31.9	48.0/37.1	56.6/46.2	58.3/43.1	55.0/49.8	20.8	10.9	22.2

Results: Table 1 reports test performance by our report generation model, along with previous works. NLP metrics BLEU (BL) [23], METEOR (M) [2] and ROUGE-L (R-L) [17] assess generated language quality. More important for this study is the assessment of diagnosis-related semantics in generated reports, using disease-classification metrics such as F1 score, by comparing agreement of class labels, C_T, extracted from generated and reference reports via CheXbert [4]. Our report generator, GenX, achieves NLP metrics similar to Ratchet [9], which has the most comparable architecture. Importantly, GenX achieves a high F1 score in comparison to Ratchet, R2Gen [4], even RGRG [26] that needs additional labelled bounding boxes for training. It approaches F1 score by $\mathcal{M}^2$ Trans [21] that trains using a more complex framework and reinforcement, and is surpassed by A_t+TE+RG [7] which uses extra labels for training. We conclude that GenX produces reports of sufficient quality to serve as effective backbone for

the following study of our auditing method. Two examples of generated reports are shown in Table 2. The generated reports use realistic language, though they can occasionally miss a diagnosis. Thus a framework that can extract and audit report semantics is crucial.

Table 2: Examples of CXRs and their associated radiology report alongside the AI generated report from GenX.

X-ray	Reference Report	Generated Report
	The lungs are low in volume but without focal consolidation. Diffuse opacities likely reflect mild pulmonary edema. There is no pleural effusion or pneumothorax. The cardiac size and cardiac silhouette are obscured by low lung volumes. The mediastinal and hilar contours appear unremarkable.	The lung volumes are low. There is mild pulmonary edema. There is no focal consolidation, pleural effusion, or pneumothorax. The cardiomediastinal silhouette is normal.
	The ET and NG tubes have been removed. A right PICC line terminates in the low SVC. Calcified left lung nodules are unchanged. The lungs are otherwise clear except for left basilar atelectasis. A small left pleural effusion has developed. Moderate cardiomegaly is unchanged.	The heart is mildly enlarged. There is mild unfolding of the thoracic aorta. There is a moderate-sized hiatal hernia. Streaky left basilar opacities suggest minor atelectasis. Otherwise, the lungs appear clear. There is no definite pleural effusion or pneumothorax. Mild rightward convex curvature is centered along the mid thoracic spine.

3.3 Evaluating the Auditing Framework for Generated Reports

Experimental settings: We analyse whether including auxiliary auditing components, ACs, enable us to quality control the generated reports and filter out those that are more likely to contain errors. We evaluate on the 5 diseases most commonly reported in the literature (Sec. 3.1), from the CheXpert competition [10]. For this, we first train 5 independent disease-specific AC image classifiers (AC_1), where each is a ResNet-18. We then apply the framework as described in Sec. 2.2 to the test-split. For each image, the 5 AC_1 models predict 5 disease labels $C_I^{(c)}$, one per disease c. GenX generates a report for each test image and we extract 5 report-based class labels $C_T^{(c)}$ via Chexbert, for each disease. For a given scan, per disease c, we compare image and report based labels, $C_I^{(c)}$ and $C_T^{(c)}$, to check if they fulfil Eq. 2. We perform the check in two settings: First, without considering a confidence threshold for p_{AC}, satisfying only the first condition in Eq. 2; Second, with a confidence threshold $t = 0.8$ where Eq. 2 is fulfilled if $p_{AC} \geq 0.8$, to investigate whether stricter auditing criteria results in more accurate reports For each disease, the images that fulfil the above criteria are considered to have successfully passed the audit. For these images, we calculate average F1 score by comparing their report-based labels $C_T^{(c)}$ with the reference report labels, to determine whether the average score is higher than that of all unfiltered reports generated by GenX. Finally, we repeat these experiments using a bigger ResNet-50 AC, trained as a multi-label image classifier

of all 14 classes (AC_{14}). This is to evaluate whether it's beneficial to integrate a larger AC that learns from all 14 labels available per scan, or independent, disease-specific ACs.

Table 3: Evaluation on test split of our auditing framework. For the 5 most commonly reported classes in literature, $GenX$ column reports F1 score of all generated reports (C_T vs true label). AC_1 assesses image-based classification by 5 disease-specific ACs (C_I vs true label). $GenX + AC_1^{t=0}$ and $GenX + AC_1^{t=0.8}$ show semantic factuality of generated reports that passed auditing for the specific disease against AC_1, satisfying Eq. 2 without the requirement for confidence $p_{AC} \geq t$ and when confidence over $t = 0.8$ is required, respectively. Bold shows improvement by auditing over baseline GenX. Parentheses show percentage of reports that satisfy the auditing per disease. We also report results when auditing with a single multi-label AC trained for all 14 classes (AC_{14} and $GenX{+}AC_{14}^{t=0}$). Developing independent ACs per disease is easier and more effective in practice.

	F1 score (AC_1)				**F1 score (AC_{14})**	
Disease	**AC_1**	**GenX**	**GenX+$AC_1^{t=0}$**	**GenX+$AC_1^{t=0.8}$**	**AC_{14}**	**GenX+$AC_{14}^{t=0}$**
Atelectasis	55.6	42.2	**52.6** (72)	**51.7** (30)	37.1	31.8 (77)
Cardiomegaly	57.6	66.8	**73.7** (78)	**81.3** (34)	40.3	53.3 (64)
Consolidation	23.5	5.6	**5.8** (76)	1.6 (65)	15.9	2.1 (89)
Edema	62.1	55.6	**64.6** (80)	**74.1** (45)	53.1	**60.2** (75)
Pleural Effusion	71.9	63.8	**74.7** (80)	**83.5** (57)	61.8	**64.1** (82)
Average	54.1	46.8	**54.3**	**<u>58.4</u>**	41.6	42.3

Results: Table 3 shows the results. Class labels C_T extracted from reports generated by GenX achieve an average F1 score of 46.8 over the 5 diseases. Reports that agree with AC_1, fulfilling the auditing without the extra requirement for confidence $p_{AC} \geq t$, achieve average F1 score 54.3. These benefits are realised over a significant percentage of the reports (72-80% across the diseases), while the rest would be deferred for user inspection and error correction in an actual workflow. Reports that pass the stricter auditing and additionally fulfil the confidence requirement $p_{AC} \geq 0.8$ present even less errors, with average F1 score 58.4. This restricts the number of reports that pass the audit (30-65%), in return for higher reliability. In comparison to the unfiltered reports of GenX, the reports that fulfil the audit show increased F1-scores for 4 out of 5 studied diseases. The exception is Consolidation, the class where both GenX and AC image classifiers have very low performance. Interestingly, for 3 out of 5 diseases, the reports that successfully pass the audit show F1-scores that are even higher than the auditing AC_1 classifiers. These results show that measuring consistency between semantics of reports and auditing classifiers is a promising framework.

Finally, we find that a single, multi-label classifier, trained for all 14-classes as AC_{14}, achieves significantly lower F1 scores for image classification, which also leads to lower effectiveness of the auditing. Although AC_{14} is a ResNet-50, larger

than ResNet-18s used for the 5 AC_1 classifiers, it is difficult to develop it at the level of disease-specific classifiers. This demonstrates the benefits of modularity, with independently developed AC for each concept of interest.

4 Conclusion

We introduced a framework to quality control AI-generated radiology reports with respect to semantics of interest via auxiliary auditing components. Our study focused on the task of report generation from chest-Xrays, detecting potential semantic errors in diagnosis-related report content, using auxiliary disease classifiers and assessing consistency between labels inferred from image and report. Experiments show that reports fulfilling the auditing criteria exhibit fewer errors, with average F1 score that can even exceed the score of the auditing components. The framework is generic and future work could explore it for auditing semantic concepts beyond disease classification, such as for description of pathology location or volume, using regression models as ACs for such properties.

The limitation of the framework, as with any filtering pipeline for quality-control, is the inherent trade-off between reducing the number of reports that successfully pass the audit and can be used for down-stream workflow, in exchange for their higher reliability. Future work could improve performance in both aspects by integrating more potent models for inferring image-based and report-based labels, thereby ensuring greater consistency and reliability from their comparison.

5 Acknowledgements

Hermione Warr and Yasin Ibrahim are supported by the EPSRC Centre for Doctoral Training in Health Data Science (EP/S02428X/1).

References

1. Alayrac, J.B., et al.: Flamingo: a visual language model for few-shot learning. In: Advances in Neural Information Processing Systems. vol. 35, pp. 23716–23736. Curran Associates, Inc. (2022) 3
2. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72. Association for Computational Linguistics, Ann Arbor, Michigan (2005), https://aclanthology.org/W05-0909 6
3. Chen, L., Avizienis, A.: N-version programming: A fault-tolerance approach to reliability of software operation. In: Proc. 8th IEEE Int. Symp. on Fault-Tolerant Computing (FTCS-8). vol. 1, pp. 3–9 (1978) 5
4. Chen, Z., et al.: Generating radiology reports via memory-driven transformer. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1439–1449. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.emnlp-main.112 1, 2, 6

5. Chen, Z., et al.: Cross-modal memory networks for radiology report generation. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. vol. 1, pp. 5904–5914. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.acl-long.459 2
6. Cornia, M., et al.: Meshed-memory transformer for image captioning (2020). https://doi.org/10.48550/arXiv.1912.08226 2
7. Dalla Serra, F., et al.: Finding-aware anatomical tokens for chest x-ray automated reporting. In: Machine Learning in Medical Imaging. MLMI 2023. Lecture Notes in Computer Science. Springer, Cham. (2024). https://doi.org/10.1007/978-3-031-45673-2_41 1, 2, 6
8. Gertz, R.J., et al.: Potential of GPT-4 for detecting errors in radiology reports: Implications for reporting accuracy. Radiology **311**(1) (2024) 2
9. Hou, B., et al.: Ratchet: Medical transformer forcchest x-ray diagnosis and reporting. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. p. 293–303. Lecture Notes in Computer Science, Springer International Publishing (2021). https://doi.org/10.1007/978-3-030-87234-2_28 1, 5, 6
10. Irvin, J., et al.: CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. Proceedings of the AAAI Conference on Artificial Intelligence pp. 590–597 (2019), https://doi.org/10.1609/aaai.v33i01.3301590 5, 7
11. Johnson, A., et al.: MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. Sci Data **6**, 317 (2019), https://doi.org/10.1038/s41597-019-0322-0 5
12. Kim, Y., et al.: Enhancing human-computer interaction in chest x-ray analysis using vision and language model with eye gaze patterns (arXiv:2404.02370) (2024). https://doi.org/10.48550/arXiv.2404.02370 2
13. Kostumov, V., et al.: Uncertainty-aware evaluation for vision-language models (arXiv:2402.14418) (2024), http://arxiv.org/abs/2402.14418 2
14. Kuhn, L., Gal, Y., Farquhar, S.: Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation (arXiv:2302.09664) (2023), http://arxiv.org/abs/2302.09664 2
15. Li, C., et al.: LLaVA-med: Training a large language-and-vision assistant for biomedicine in one day (2023). https://doi.org/10.48550/arXiv.2306.00890 1
16. Li, M., et al.: Think twice before assure: Confidence estimation for large language models through reflection on multiple answers (arXiv:2403.09972) (2024), http://arxiv.org/abs/2403.09972 2
17. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04). pp. 605–612 (2004). https://doi.org/10.3115/1218955.1219032 6
18. Liu, G., et al.: Clinically accurate chest x-ray report generation. In: Proceedings of the 4th Machine Learning for Healthcare Conference. Proceedings of Machine Learning Research, vol. 106, pp. 249–269. PMLR (2019) 2
19. Liu, H., et al.: Visual instruction tuning (2023). https://doi.org/10.48550/arXiv.2304.08485 3
20. Malinin, A., Gales, M.: Uncertainty estimation in autoregressive structured prediction (arXiv:2002.07650) (2021), http://arxiv.org/abs/2002.07650 2
21. Miura, Y., et al.: Improving factual completeness and consistency of image-to-text radiology report generation. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human

Language Technologies. pp. 5288–5304. Association for Computational Linguistics (2021). https://doi.org/10.18653/v1/2021.naacl-main.416 2, 5, 6

22. Moor, M., et al.: Med-flamingo: a multimodal medical few-shot learner. https://doi.org/10.48550/arXiv.2307.15189 1, 3
23. Papineni, K., et al.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (2002). https://doi.org/10.3115/1073083.1073135 6
24. Radford, A., et al.: Language models are unsupervised multitask learners (2019), OpenAI 6
25. Smit, A., et al.: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). p. 1500–1519. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.emnlp-main.117 4
26. Tanida, T., et al.: Interactive and explainable region-guided radiology report generation. In: CVPR (2023) 1, 2, 5, 6
27. Tomani, C., et al.: Uncertainty-Based Abstention in LLMs Improves Safety and Reduces Hallucinations (arXiv:2404.10960) (2024), http://arxiv.org/abs/2404.10960 2
28. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30 (2017) 1, 4
29. Wu, J., et al.: Exploring multimodal large language models for radiology report error-checking (arXiv:2312.13103) (2024), http://arxiv.org/abs/2312.13103 2
30. Zech, J., et al.: Detecting insertion, substitution, and deletion errors in radiology reports using neural sequence-to-sequence models. Annals of Translational Medicine **7**(1111), 233–233 (2019). https://doi.org/10.21037/atm.2018.08.11 2