RADAR: Enhancing Radiology Report Generation with Supplementary Knowledge Injection

Wenjun Hou[1,2], **Yi Cheng**[1*], **Kaishuai Xu**[1*], **Heng Li**[2], **Yan Hu**[2†], **Wenjie Li**[1], **Jiang Liu**[2,3†]
[1]Department of Computing, The Hong Kong Polytechnic University
[2]Research Institute of Trustworthy Autonomous Systems and
Department of Computer Science and Engineering,
Southern University of Science and Technology
[3]School of Computer Science, University of Nottingham Ningbo China
houwenjun060@gmail.com

Abstract

Large language models (LLMs) have demonstrated remarkable capabilities in various domains, including radiology report generation. Previous approaches have attempted to utilize multimodal LLMs for this task, enhancing their performance through the integration of domain-specific knowledge retrieval. However, these approaches often overlook the knowledge already embedded within the LLMs, leading to redundant information integration. To address this limitation, we propose RADAR, a framework for enhancing radiology report generation with supplementary knowledge injection. RADAR improves report generation by systematically leveraging both the internal knowledge of an LLM and externally retrieved information. Specifically, it first extracts the model's acquired knowledge that aligns with expert image-based classification outputs. It then retrieves relevant supplementary knowledge to further enrich this information. Finally, by aggregating both sources, RADAR generates more accurate and informative radiology reports. Extensive experiments on MIMIC-CXR, CHEXPERT-PLUS, and IU X-RAY demonstrate that our model outperforms state-of-the-art LLMs in both language quality and clinical accuracy[1].

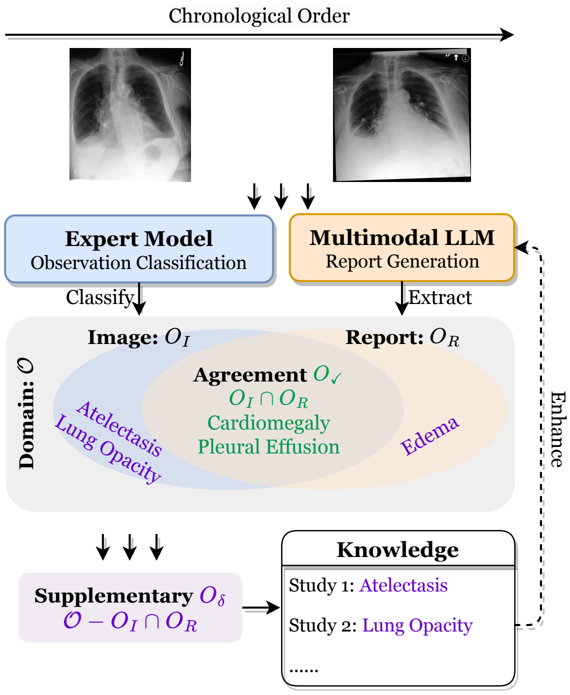


Figure 1: A motivating example. The report directly generated by the multimodal LLM showcases its knowledge regarding several findings (O_R) but can contain hallucinations and overlook some other findings. To address this, we regard the part that aligns with another expert model ($O_R \cap O_I$) as trustworthy and we incorporate supplementary knowledge for the remaining part ($\mathcal{O} - O_R \cap O_I$) to enhance the report generation.

1 Introduction

Radiology report generation (Chen et al., 2020, 2021) plays a crucial role in chest X-ray interpretation, requiring highly specialized domain knowledge (Jain et al., 2021; Irvin et al., 2019). Recent advances in foundation models (Pellegrini et al., 2023; Chen et al., 2024; Hyland et al., 2024), which leverage large language models (LLMs) for enhanced medical image analysis, have demonstrated remarkable potential in generating fluent and cohesive clinical text, aiding radiologists in their diagnostic workflow.

Despite their ability to generate highly readable and clinically plausible report content, LLMs still face persistent challenges in ensuring clinical accuracy. One major challenge lies in the knowledge gap between the medical and general domains. Many studies have attempted to bridge this disparity by augmenting models with retrieved domain-specific knowledge (Yang et al., 2021; Liu et al.,

*Equal contribution.
†Corresponding authors.
[1]Our code is available at: https://github.com/wjhou/Radar

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26366–26381
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

2021; Li et al., 2023c; Ranjit et al., 2023; Sun et al., 2025). However, these approaches often overlook the knowledge LLMs have already acquired. That is, much of the retrieved information is often duplicate knowledge already encoded within the model's parameters, leading to redundant information retrieval. Moreover, the knowledge learned by LLMs (Liu et al., 2024) is not always trustworthy, as hallucinations frequently occur (Huang et al., 2025). For instance, in Figure 1, the LLM correctly identifies *Cardiomegaly*, making the retrieval of additional knowledge about this observation unnecessary. Additionally, the generated *Pleural Effusion* is highly credible, as it aligns with the expert model, whereas *Edema* remains uncertain. Thus, balancing learned and retrieved knowledge in radiology report generation is crucial to address these challenges.

In this paper, we propose RADAR, a framework for RADiology report generation that integrates both the internal knowledge of LLMs and external supplementARy knowledge. Our framework primarily consists of two stages: preliminary findings generation and supplementary findings augmentation. In the first stage, RADAR generates an initial report from the input images. Subsequently, an expert model processes the images for observation classification. The overlapping information between the generated report and the classified observations is identified as high-confidence internal knowledge. In the second stage, RADAR additionally retrieves new knowledge to supplement the internal knowledge. Finally, both internal and supplementary knowledge sources are aggregated to enhance the report generation process. Our main contributions can be summarized as follows:

- We propose RADAR, a novel framework that enhances the clinical accuracy of radiology report generation by effectively integrating both the internal knowledge of LLMs and externally retrieved domain-specific knowledge.
- To optimize knowledge utilization, we introduce a knowledge extraction method that identifies and retains non-overlapping information from the model's learned knowledge, reducing redundancy and bridging the knowledge gap.
- We conduct extensive experiments on three benchmark datasets: MIMIC-CXR, CHEXPERT-PLUS, and IU X-RAY, demonstrating the effectiveness of RADAR.

2 Preliminary

2.1 Problem Formulation

A multimodal LLM (MLLM) generally consists of a vision encoder, a vision connector that transforms visual signals into the language space (e.g., MLP (Liu et al., 2023), Q-Former (Li et al., 2023b), or Perceiver Resampler (Xue et al., 2024)), and an LLM, as illustrated in the left part of Figure 2. For radiology report generation[2], the MLLM takes a radiograph X, its prior X_p (if available), and the clinical context C (e.g., *Indication* or *Prior Findings*) as input and generates the report $Y = \{y_1, \ldots, y_T\}$. The probability of the t-th token is computed as follows:

$$p(y_t) = \mathrm{MLLM}(X, X_p, C, y_{<t}),$$

where the MLLM is optimized using the negative log-likelihood loss:

$$\mathcal{L} = -\sum_{t=1}^{T} \log p(y_t).$$

2.2 Semi-Structured Report as Knowledge

In this paper, the training set of MIMIC-CXR serves as the knowledge source for radiology report generation. To effectively leverage the knowledge encoded in each report, we convert it into semi-structured data. Specifically, given a report consisting of N sentences, $Y = \{S_1, \ldots, S_N\}$, we annotate each sentence using the 14-category CheXpert observations (Irvin et al., 2019) with the CheXbert model (Smit et al., 2020). Each observation falls into one of four classes: *Positive*, *Negative*, *Uncertain*, or *Blank*. To ensure conciseness, we retain only sentences annotated with *Positive* observations. These selected sentences collectively represent the knowledge extracted from the report, as illustrated in the top-right part of Figure 2. Note that we annotate and process Preliminary Findings (§3.1) and Supplementary Findings (§3.2) in the same manner.

3 The RADAR Framework

3.1 Stage I: Preliminary Findings Generation

We illustrate the Stage I process in the left part of Figure 2. To assess the learned knowledge of an

[2] In this paper, "report" typically refers to "findings," and we use these two terms interchangeably.

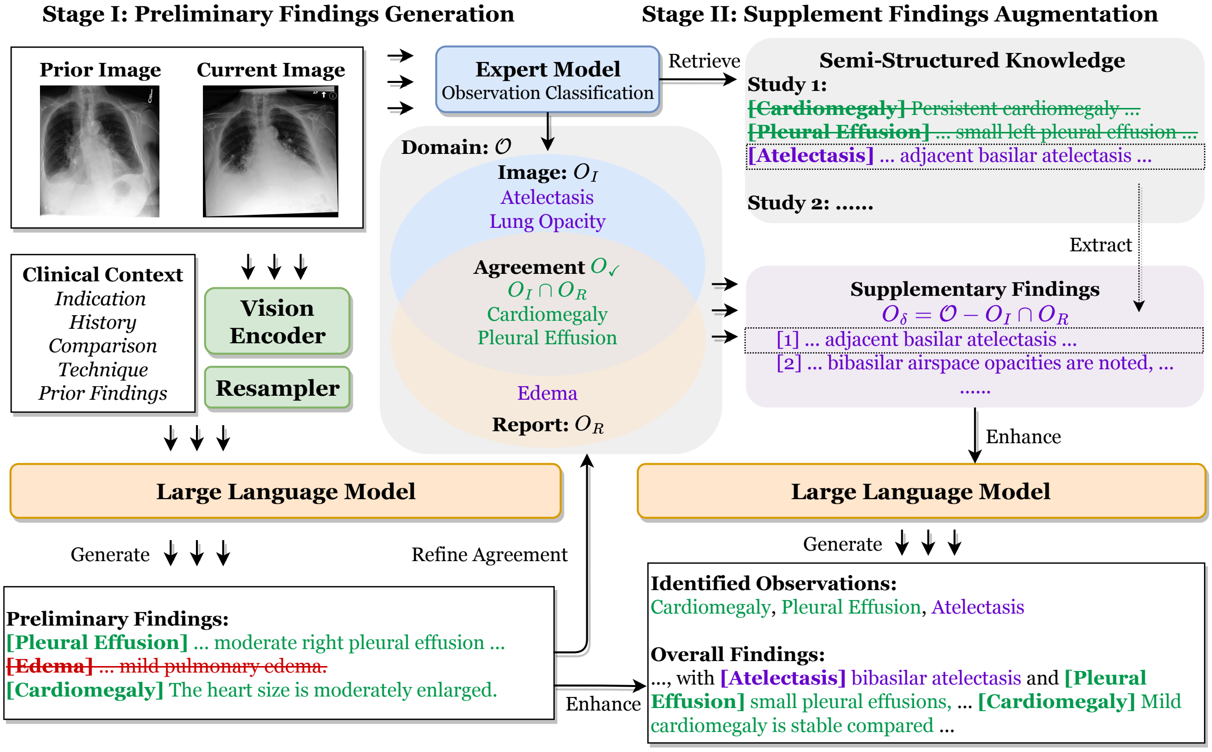


Figure 2: Overview of the RADAR. In Preliminary Findings, only sentences that reach agreement are retained, whereas in Supplementary Findings, only sentences that supplement the Preliminary Findings are preserved.

LLM, we first feed the input (X, X_p, and C) into RADAR to generate a report $\hat{Y}$:

$$\hat{Y} = \underset{\hat{Y} \in \mathcal{Y}}{\text{argmax}} \prod_{t=1}^{T} \text{MLLM}(X, X_p, C, \hat{y}_{<t}),$$

where $\mathcal{Y}$ represents the set of possible reports. Note that exact maximization is intractable and we employ an approximate decoding algorithm for generation. Next, we convert the findings into semi-structured knowledge, as described in §2.2, and denote the observations of $\hat{Y}$ as O_R.

To extract credible knowledge from $\hat{Y}$ while filtering out untrustworthy information, we train an expert model that predicts observations for the image. Unlike previous works (Hou et al., 2023b; Pellegrini et al., 2023), which consider only the image as input, we incorporate the clinical context to enhance performance. Specifically, the expert model $f(X)$ encodes X and C using an image encoder Encoder_v and a text encoder Encoder_t, respectively, and then processes their outputs through an MLP for observation classification:

$$\mathbf{h}_v = \text{Encoder}_v(X), \quad \mathbf{h}_t = \text{Encoder}_t(C),$$

$$p(O_i) = \sigma(\text{MLP}([\mathbf{h}_v; \mathbf{h}_t])),$$

where $[;]$ is the concatenation function, $\mathbf{h}_v$ and $\mathbf{h}_t$ are the pooled outputs of the image and text encoders, respectively, and $p(O_i)$ represents the probability of the i-th observation. We denote the observations derived from $f(X)$ as O_I, and the credible and high-confidence observations, $O_\checkmark$, are then obtained by intersecting O_I and O_R, as follows:

$$O_\checkmark = O_I \cap O_R.$$

Finally, we refine $\hat{Y}$ by removing sentences that do not correspond to $O_\checkmark$, yielding the Preliminary Findings (PF).

To train the expert model, we collect observations from each report as image annotations and optimize the expert model using binary cross-entropy loss. Following Pellegrini et al. (2023), we address data imbalance by re-weighting the positive observations with a log-scale weight, defined as $\alpha_i = \log\left(1 + \frac{|\mathcal{D}_{\text{train}}|}{w_i}\right)$, where $|\mathcal{D}_{\text{train}}|$ is the total number of training samples and w_i denotes the frequency of observation O_i.

3.2 Stage II: Supplementary Findings Augmentation

Supplementary Knowledge Retrieval. We follow the retrieval process of Yang et al. (2021) to search for domain knowledge. Specifically, the expert model described in §3.1 produces probabilities for 14 observations, and we compute the similarity between different samples using KL-divergence:

$$\hat{z} = \mathrm{Normalize}(f(X)),$$

$$\mathrm{Sim}(X, X_i) = -\sum_{j=1}^{|\mathcal{O}|} \hat{z}_j \log \frac{\hat{z}_j}{\hat{z}_{i,j}},$$

where $\mathrm{Normalize}(\cdot)$ normalizes $f(X)$ to 1, $\hat{z}$ represents the normalized scores for (X), and $\hat{z}_{i,j}$ denotes the score of the j-th observation in the i-th sample from the database (i.e., the training set of the MIMIC-CXR dataset). We then rank the samples based on their similarity scores, $\mathrm{Sim}(X, X_i)$, and retrieve the top-K reports, denoted as $\mathcal{Y}^S = \{Y_1^S, \ldots, Y_K^S\}$.

Supplementary Knowledge Extraction. Since the retrieved information may overlap with the knowledge learned by LLMs, we extract only supplementary knowledge based on two principles: (1) it should be concise and relevant, and (2) it should complement, rather than duplicate, the preliminary findings. Thus, for each supplementary report Y_i^S with its corresponding observations O^S, we retain only the following observations:

$$O_\delta = \mathcal{O} - O_\checkmark.$$

Next, we convert Y_i^S into semi-structured knowledge and remove sentences that do not correspond to O_δ, referring to these findings as Supplementary Findings (SF). Notably, all sentences corresponding to negative observations are removed, ensuring that SF remains concise and clinically relevant.

3.3 Enhanced Radiology Report Generation

We integrate both PF and SF into the clinical context C to form the augmented context C^A, from which the final report Y is generated as:

$$Y = \underset{Y \in \mathcal{Y}}{\mathrm{argmax}} \prod_{t=1}^{T} \mathrm{MLLM}(X, X_p, C^A, y_{<t}).$$

Since PF and SF contain information from various studies, summarizing high-level information before generating the report is necessary. Thus, we include the observations of Y as part of the training targets. Specifically, during training, Y is converted into a structured format:

$$Y^O = \{O_1, \ldots, O_N, y_1, \ldots, y_L\},$$

where $\{O_1, \ldots, O_N\}$ represents the observations in Y, and $\{y_1, \ldots, y_L\}$ corresponds to the tokens of the report. We refer to this process as Observation Identification (OI). During inference, we extract the final report from the generated output for evaluation.

4 Experiments

4.1 Datasets

We evaluate our model using three publicly available radiology report generation datasets: MIMIC-CXR[3] (Johnson et al., 2019), CHEXPERT PLUS[4] (Chambon et al., 2024), and IU X-RAY[5] (Demner-Fushman et al., 2016):

- MIMIC-CXR contains 377,110 chest radiographs and 227,827 reports. We use this dataset for fine-tuning, and we include only frontal images in our experiments. The number of samples in the train, validation, and test sets is 162,955, 1,286, and 2,461, respectively.
- CHEXPERT PLUS comprises 223,462 unique radiology reports and chest X-ray pairs from 187,711 studies. We evaluate our model using only frontal images from the validation set, which includes 62 samples.
- IU X-RAY is a dataset collected by Indiana University. Following Bannur et al. (2024), we use all frontal images for evaluation, totaling 3,199 studies.

4.2 Evaluation Metrics

Lexical Metrics. Following previous research (Chen et al., 2020; Li et al., 2023c), BLEU-1/4 (Papineni et al., 2002), ROUGE-L (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) are adopted for evaluating the languages of generated outputs.

Clinical Metrics. We evaluate the factual accuracy using several metrics. Specifically, RG-F_1 and $\mathrm{RG}_{\overline{\mathrm{ER}}(\mathrm{ER})}$ (Jain et al., 2021) evaluate the entity-level factuality and RadCliQ$_0$ (Yu et al., 2022), denoted as CliQ$_0$, aligns with the preference of radiologists. For observation evaluation, ^{14}Macro-F_1 (^{14}Ma-F_1) and ^{14}Micro-F_1 (^{14}Mi-F_1) evaluate the macro and micro F_1 of 14 observations (refers to Table 7), respectively. In addition, ^{5}Macro-F_1

[3] https://physionet.org/content/mimic-cxr-jpg/2.0.0/
[4] https://aimi.stanford.edu/datasets/chexpert-plus
[5] https://openi.nlm.nih.gov/

Dataset: MIMIC-CXR (*Training and Evaluation*)												
Model	**Lexical Metrics**				**Clinical Metrics** (*CheXpert: Uncertain as Negative / Positive*)							
	B-1	**B-4**	**MTR**	**R-L**	**RG-F$_1$**	**RG$_{ER}$**	**CliQ$_0$** (↓)	**^{14}Ma-F$_1$**	**^{5}Ma-F$_1$**	**^{14}Mi-F$_1$**	**^{5}Mi-F$_1$**
RadFM	–	0.128	–	0.182	–	–	–	–	–	–	–
XrayGPT	0.128	0.004	0.079	0.111	–	–	–	–	–	–	–
R2GenGPT	0.411	0.134	0.160	0.297	–	–	–	0.389	–	–	–
R2-LLM	0.402	0.128	0.175	0.291	–	–	–	–	–	–	–
RaDialog	0.346	0.095	0.140	0.271	–	–	–	0.394	–	–	–
LLaVA-Med	0.354	0.149	0.353	0.276	0.191	0.238	3.30	0.269	0.363	0.427	0.439
CheXagent	0.169	0.047	–	0.215	–	0.205	–	0.247	0.345	0.393	0.412
GPT-4V	0.164	0.178	–	0.132	–	0.132	–	0.204	0.196	0.355	0.258
Med-PaLM	0.323	0.115	–	0.275	0.267	–	–	0.398	0.516	0.536	0.579
LLaVA-Rad	0.381	0.154	–	0.306	–	0.294	–	0.395	0.477	0.573	0.574
MAIRA-1	0.392	0.142	0.333	0.289	0.243	0.296	3.10	0.386 0.423	0.477 0.517	0.557 0.553	0.560 0.588
MAIRA-2	0.465	0.234	0.420	0.384	**0.346**	**0.396**	2.64	0.416	0.504	0.581	0.591
MedVerse	–	0.178	–	–	0.280	–	2.71	–	–	–	–
M4CXR	0.339	0.103	–	–	0.218	0.285	–	0.400	0.495	0.606	0.618
Libra	**0.513**	0.245	**0.489**	0.367	0.329	0.376	2.70	0.404	0.538	0.559	0.601
RADAR (Ours)	0.509	**0.262**	0.450	**0.397**	**0.346**	0.393	**2.61**	**0.460** 0.497	**0.567** 0.602	**0.627** 0.627	**0.653** 0.674

Table 1: Evaluation results of our model and baseline methods on the MIMIC-CXR dataset. Baseline results are cited from their respective literature. The best results are shown in **bold**, while underlined values indicate the second-best results. ↓ denotes that lower values are better. Results of CheXpert treat *Uncertain* labels as *Positive* when compared with MAIRA-1. Comparisons with SOTA specialists are provided in Table 8.

Dataset: IU X-RAY (*Evaluation Only*)						
Model	**Lexical**		**Clinical**			
	B-4	**R-L**	**RG-F$_1$**	**CliQ$_0$** (↓)	**^{14}Ma-F$_1$**	**^{14}Mi-F$_1$**
LLaVA-Rad	–	0.253	–	–	–	0.535
MAIRA-2	0.117	0.274	0.271	2.68	0.319	0.525
RADAR (Ours)	0.116	0.276	0.237	2.78	0.325	0.546
BACKBONE	0.112	0.275	0.236	2.79	0.269	0.514

Table 2: Evaluation on the IU X-RAY dataset. Results of LLaVA-Rad and MAIRA-2 are cited from Bannur et al. (2024).

Dataset: CHEXPERT PLUS (*Evaluation Only*)						
Model	**Train**	**Lexical**		**Clinical**		
		B-4	**R-L**	**RG$_{\overline{ER}(ER)}$**	**$^{14(5)}$Ma-F$_1$**	**$^{14(5)}$Mi-F$_1$**
SWIN$_{V2}$-BERT	M*	0.034	0.191	0.136 (0.198)	0.268 (0.383)	0.410 (0.423)
	C	0.057	0.228	0.183 (0.250)	0.331 (0.401)	0.508 (0.432)
	M&C	0.056	0.234	0.201 (0.277)	0.366 (0.495)	0.560 (0.532)
RADAR (Ours)	M	0.076	0.203	0.143 (0.216)	0.362 (0.417) 0.401 (0.540)	0.541 (0.524) 0.554 (0.608)
BACKBONE	M	0.073	0.203	0.143 (0.206)	0.282 (0.437) 0.317 (0.502)	0.477 (0.466) 0.492 (0.552)

Table 3: Evaluation on the CHEXPERT PLUS dataset. The results for SWIN$_{V2}$-BERT are cited from Chambon et al. (2024), and we primarily compare RADAR with its ⋆ variant. The "Train" column indicates the training datasets, where M and C denote the MIMIC-CXR and CHEXPERT PLUS datasets, respectively.

(^{5}Ma-F$_1$) and ^{5}Micro-F$_1$ (^{5}Mi-F$_1$) measure the performance of 5 common observations (*Atelectasis*, *Cardiomegaly*, *Consolidation*, *Edema*, and *Pleural Effusion*). Two lines of CheXpert results are reported, i.e., *Uncertain as Negative* and *Uncertain as Positive*.

4.3 Baselines

On the MIMIC-CXR dataset, we compare our models with the state-of-the-art (SOTA) MLLMs, including RadFM (Wu et al., 2023), XrayGPT (Thawakar et al., 2024), LLaVA-Med (Li et al., 2023a), R2GenGPT (Wang et al., 2023b), R2-LLM (Liu et al., 2024), RaDialog (Pellegrini et al., 2023), CheXagent (Chen et al., 2024), GPT-4V (OpenAI, 2023), LLaVA-Rad (Chaves et al., 2024), Med-PaLM (Singhal et al., 2022), MAIRA-1 (Hyland et al., 2024), MAIRA-2 (Bannur et al., 2024), MedVerse (Zhou et al., 2024), M4CXR (Park et al., 2024), and Libra (Zhang et al., 2024). Other SOTA specialists are in the Appendix A.1. We also compare RADAR with LLaVA-Rad and MAIRA-2 on the IU X-RAY dataset. On the CHEXPERT-PLUS dataset, we compare RADAR with the baseline SWIN$_{V2}$-BERT (Chambon et al., 2024) consisting of a Swin Transformer V2 (Liu et al., 2022) and a BERT decoder (Devlin et al., 2019). The SWIN$_{V2}$-BERT model has three variants, each trained on a different dataset: the MIMIC-CXR dataset, the CHEXPERT PLUS dataset, and a combined version

Model	Modules			Lexical Metrics				Clinical Metrics (*CheXpert: Uncertain as Negative*)						
	PF	SF	OI	B-1	B-4	MTR	R-L	RG-F_1	RG_{ER}	$CliQ_0$ (↓)	^{14}Ma-F_1	^{5}Ma-F_1	^{14}Mi-F_1	^{5}Mi-F_1
RADAR	✓	✓	✓	0.509	0.262	0.450	0.397	0.346	0.393	2.61	0.460	0.567	0.627	0.653
BACKBONE	✗	✗	✗	0.497	0.259	0.444	0.396	0.343	0.387	2.67	0.402	0.495	0.565	0.581
$\text{RADAR}_{w/o\ F}$	✗	✗	✓	0.506	0.260	0.448	0.396	0.343	0.391	2.63	0.442	0.545	0.624	0.651
$\text{RADAR}_{w/o\ SF}$	✓	✗	✓	0.508	0.262	0.451	0.398	0.346	0.394	2.62	0.447	0.543	0.626	0.650
$\text{RADAR}_{w/o\ PF}$	✗	✓	✓	0.508	0.261	0.450	0.396	0.344	0.389	2.63	0.456	0.559	0.623	0.652

Table 4: Ablation results of RADAR with different modules. Per-observation results of BACKBONE, $\text{RADAR}_{w/o\ F}$, $\text{RADAR}_{w/o\ SF}$, $\text{RADAR}_{w/o\ PF}$, and RADAR are provided in Appendix, Table 7.

Model	Modules			Lexical Metrics				Clinical Metrics (*CheXpert: Uncertain as Negative*)						
	Vision	Resampler	LLM	B-1	B-4	MTR	R-L	RG-F_1	RG_{ER}	$CliQ_0$ (↓)	^{14}Ma-F_1	^{5}Ma-F_1	^{14}Mi-F_1	^{5}Mi-F_1
BACKBONE	✓	✓	✓	0.497	0.259	0.444	0.396	0.343	0.387	2.67	0.402	0.495	0.565	0.581
BACKBONE-V1	✗	✓	✗	0.430	0.183	0.359	0.318	0.245	0.296	3.15	0.284	0.415	0.476	0.508
BACKBONE-V2	✗	✓	✓	0.483	0.246	0.428	0.381	0.321	0.368	2.78	0.361	0.465	0.532	0.550

Table 5: Ablation results of fine-tuning different modules of BACKBONE.

Hyperparameters	Stage I	Stage II
Trainable Module	Vision Encoder (LoRA) Perceiver Resampler (Full) LLM (LoRA)	LLM (LoRA)
Training Epoch	3	2
Learning Rate	$1e-4$	
Optimizer	AdamW	
LR Scheduler	Cosine	
Warmup Ratio	0.03	
LoRA Config	$r = 64, \alpha = 128$	
Batch Size	32	

Table 6: Detailed hyperparameters for training RADAR. LoRA is used to fine-tune both the vision encoder and the LLM, while the Perceiver Resampler is fully fine-tuned.

of both.

4.4 Implementation Details

Training and Inference. We implement RADAR using BLIP-3[6] (Xue et al., 2024) as the backbone, which comprises a SigLIP (Zhai et al., 2023) vision encoder, a Perceiver Resampler, and a Phi-3-mini$_{3.8B}$ (Abdin et al., 2024) language model. Our implementation is based on Hugging Face's `Transformers` library (Wolf et al., 2020). The expert model consists of a Swin Transformer V2[7] (Liu et al., 2022) and a BioClinicalBERT[8] (Alsentzer et al., 2019). Top-2 reports are selected as knowledge. The hyperparameters used for training RADAR are provided in Table 6. During inference, we employ beam search with a beam width of 5 for report generation and set the length penalty to 2.0. As proposed by Xue et al. (2024), BLIP-3 samples vision tokens using a Perceiver Resampler with learned queries and supports images of any resolution, resulting in significant performance gains across multiple tasks. In this paper, we use only the base resolution (384 × 384) with 128 learned query tokens to ensure a fair comparison with other baselines. For training, in Stage I, we fine-tune all three components (i.e., the vision encoder, the Perceiver Resampler, and the LLM) in BLIP-3 since the model is not specifically designed for medical tasks. In Stage II, we further fine-tune only the LoRA of the LLM to enhance performance.

Data Preprocessing. Following previous research (Hyland et al., 2024; Bannur et al., 2024; Zhang et al., 2024), we incorporate *Indication*, *History*, *Comparison*, *Technique*, and *Prior Findings* as clinical context for the MIMIC-CXR and CHEXPERT PLUS datasets, when available. Since the IU X-RAY dataset does not include follow-up studies, we extract only *Indication*, *Comparison*, and *Technique* as clinical context. For a better illustration, we provide the prompt template in Table 9.

5 Results and Analyses

5.1 Quantitative Analysis

Comparison with MLLMs. As shown in Table 1, RADAR achieves SOTA performance compared to other MLLM baselines. In terms of lexical metrics, RADAR outperforms the best baselines (i.e., Libra and MAIRA-2) with absolute improvements of 1.7% in BLEU-4 and 1.3% in ROUGE-L, while maintaining competitive performance of 0.509 in BLEU-1 and 0.450 in METEOR. Regarding entity-

[6] The model card is "Salesforce/xgen-mm-phi3-mini-instruct-interleave-r-v1.5."

[7] The model card is "microsoft/swinv2-large-patch4-window12to16-192to256-22kto1k-ft."

[8] The model card is "emilyalsentzer/Bio_ClinicalBERT."

level clinical metrics, our model achieves the best performance on RG-F_1 and RadCliQ$_0$, attaining scores of 0.346 and 2.61, respectively. Additionally, RADAR surpasses the top baselines, achieving improvements across multiple observation-level clinical metrics, with ^{14}Macro-F_1 increasing to 0.460, ^{5}Macro-F_1 to 0.567, ^{14}Micro-F_1 to 0.627, and ^{5}Micro-F_1 to 0.653, respectively. Notably, the smallest gain over the second-best model is 2.1%, underscoring RADAR's effectiveness. Furthermore, we provide an additional set of CheXpert results using the *Uncertain as Positive* policy and compare RADAR with MAIRA-1. We observe that the improvements under this setting follow a similar trend to those obtained with the *Uncertain as Negative* policy. These results collectively demonstrate the effectiveness of RADAR in generating coherent and clinically accurate radiology reports.

Comparison with SOTA Specialists. The results of other specialists are shown in Table 8. We find that models incorporating clinical context (e.g., *Indication*) as input generally achieve better performance than others. For example, the Controllable model significantly outperforms other baselines across both lexical and clinical metrics. This trend also holds for MLLMs, as shown in Table 1. Moreover, benefiting from the strong contextual comprehension and language generation capabilities of LLMs, RADAR further improves linguistic quality, which requires models to integrate diverse information sources. However, we observe that the ^{14}Macro-F_1 score of our model still lags behind that of the Controllable baseline (0.497 vs. 0.553). This discrepancy may stem from differences in learning objectives, as this baseline treats *Uncertain* cases as *Positive*.

Model Generalization. Following prior research (Bannur et al., 2024), we further evaluate RADAR on the CHEXPERT PLUS and IU X-RAY datasets to assess its generalization capability. The results are presented in Table 2 and Table 3. On the IU X-RAY dataset, RADAR outperforms MAIRA-2 in terms of CheXpert metrics, achieving a ^{14}Macro-F_1 of 0.325 and a ^{14}Micro-F_1 of 0.546. However, a performance gap remains in RG-F_1 and RadCliQ$_0$, which may be attributed to differences in training data, as MAIRA-2 is trained with the additional USMix dataset. Meanwhile, RADAR demonstrates comparable performance to the baselines in terms of lexical metrics. On the CHEXPERT PLUS dataset, our model significantly outperforms SWIN$_{\mathrm{V2}}$-BERT trained on the MIMIC-CXR dataset, across both lexical and clinical metrics. Furthermore, RADAR surpasses the baseline that is trained on CHEXPERT PLUS alone as well as the one trained on a combination of both datasets. These results demonstrate the strong generalization ability of RADAR across different datasets. Additionally, RADAR significantly outperforms the BACKBONE, underscoring the effectiveness of the integrated knowledge.

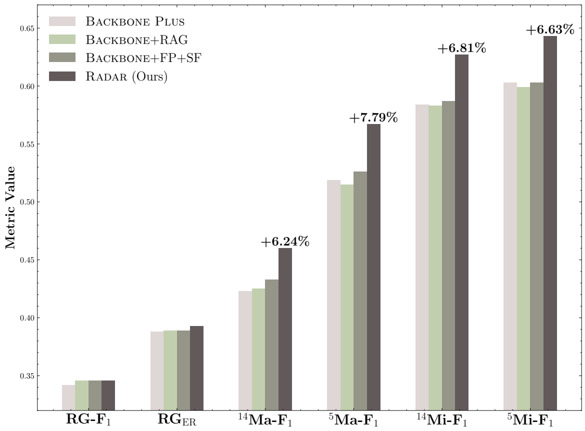
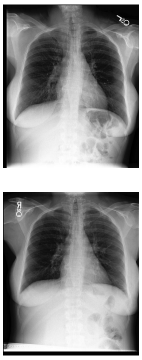


Figure 3: Comparisons among BACKBONE+RAG, BACKBONE+FP+SF, and RADAR on six clinical metrics.

Analysis of PF, SF, and OI. We analyze the impact of PF, SF, and OI on the performance of RADAR, with results summarized in Table 4. RADAR$_{w/o\ \mathrm{F}}$, which first identifies observations before report generation without incorporating knowledge, significantly improves the CheXpert metrics, particularly ^{14}Macro-F_1 and ^{5}Macro-F_1, as observation information captures high-level abstractions of reports and aligns closely with the objectives of these metrics. This highlights the crucial role of OI in enhancing clinical accuracy, independent of other components. When PF and SF are introduced individually with OI, introducing PF alone helps preserve the knowledge embedded in the LLM, resulting in comparable performance across both lexical and clinical metrics. In contrast, introducing SF alone substantially improves $^{14/5}$Macro-F1, but negatively impacts RG$_{\mathrm{ER}}$ and RadCliQ$_0$. Moreover, combining both PF and SF leverages the strengths of each, leading to further improvements in the clinical metrics while maintaining comparable performance across the other metrics. We notice that BACKBONE tends to retain easily acquired knowledge (i.e., PF) and that selectively supplementing it with external information (i.e., SF) is crucial for bridging the remaining knowledge gaps.

Case A



Preliminary Findings: O_R **[Atelectasis]**
Heart size is normal. ...... ~~[Atelectasis] Mild atelectasis is noted in the left lung base.~~ No focal consolidation, pleural effusion or pneumothorax is present. ......
Expert Observations: $O_I = \emptyset$
Agreement: $O_\checkmark = \emptyset$
Overall Findings: $O = \emptyset$
The cardiac, mediastinal and hilar contours are normal. Lungs are clear. ......
Reference: $O = \emptyset$
Cardiac, mediastinal and hilar contours are normal. ......

Preliminary Findings: O_R **[Edema, Cardiomegaly]**
[Edema] ..... there is still evidence of mild-to-moderate pulmonary edema. [Cardiomegaly] Moderate cardiomegaly persists.
Supplementary Findings:
[Atelectasis] There are areas of left mid-to-lower lung linear atelectasis/scarring.
[Atelectasis] Mild areas of atelectasis at the left lung bases.
Expert Observations: O_I **[Cardiomegaly, Atelectasis, Edema]**
Agreement: $O_\checkmark$ **[Edema, Cardiomegaly]**
Overall Findings: O
...... [Edema] Moderate pulmonary edema with [Cardiomegaly] moderate cardiomegaly and [Atelectasis] bilateral areas of atelectasis.
......
Reference:
[Edema] There are still signs indicative of mild pulmonary edema. [Atelectasis] In addition, there is a small right medial basal atelectasis. [Cardiomegaly] Moderate cardiomegaly. ......

Case B

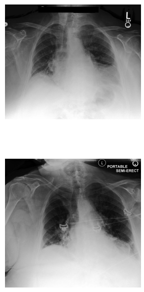


Figure 4: Two cases generated by RADAR, where false positive observation appears in the PF of case A and false negative observation shows in the PF of case B.

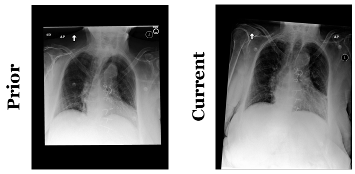
Analysis of RADAR versus RAG. To evaluate the effectiveness of knowledge integration in RADAR, we conduct experiments comparing our model against three baselines: (1) BACKBONE PLUS, (2) BACKBONE+RAG, and (3) BACKBONE+PF+SF. The results are presented in Figure 3. Note that these baselines do not include the OI. Since RADAR undergoes two-stage training (i.e., two additional epochs), we apply the same extended training to BACKBONE, referring to this variant as BACKBONE PLUS. In addition, we introduce a standard RAG baseline (BACKBONE+RAG), which utilizes the same retrieved findings as RADAR. Building upon this baseline, BACKBONE+PF+SF further includes PF as context. Our findings reveal that while all four models achieve comparable performance on lexical metrics (e.g., 50%/26% B-1/4), they differ in clinical metrics. Specifically, BACKBONE+RAG and BACKBONE PLUS show similar performance, and BACKBONE+FP+SF outperforms these two baselines on CheXpert metrics and exhibits similar performance on RadGraph metrics. This demonstrates that incorporating credible knowledge can effectively enhance report generation even without OI. Moreover, RADAR demonstrates a relative improvement of over 6% across four key CheXpert metrics. This suggests that structured integration of internal and external knowledge contributes to its enhanced clinical accuracy.

Analysis of Fine-tuning Different Modules in BACKBONE. To assess the contributions of different components in the base model (i.e., BLIP-3), we conduct an ablation study on the impact of fine-tuning the vision encoder, the Resampler, and the LLM. The results are summarized in Table 5. By comparing BACKBONE and BACKBONE-V2, we find that fine-tuning the vision encoder to incorporate domain-specific knowledge is crucial for achieving high clinical accuracy, even though both configurations exhibit strong language coverage in lexical metrics. Furthermore, fine-tuning the LLM (i.e., Phi-3) results in substantial improvements in both lexical and clinical metrics, as evidenced by the comparison between V1 and V2. This highlights the importance of adapting the LLM to the clinical domain for optimal performance. Notably, RADAR utilizes a 3.8B LLM as the decoder and outperforms many larger models (e.g., LLaVA-Med and MAIRA-1).

5.2 Qualitative Analysis

Case Study. We conduct a case study to illustrate the advantages of incorporating both internal knowledge and retrieved information, as shown in Figure 4. In Case A, RADAR initially generates a report that includes the finding *Atelectasis*. However, expert assessment indicates the image shows no positive findings. As a result, their intersection is $\emptyset$, and by removing this incorrect observation, RADAR ultimately produces an accurate report. This example highlights the model's ability to refine its predictions when guided by expert constraints, effectively eliminating unnecessary or incorrect findings. Another more complex case is presented on the right side of this Figure. Specifically, RADAR initially identifies findings related to *Edema* and *Cardiomegaly*, which the expert model also notes. However, the observation of *Atelectasis* is omitted from the preliminary findings. By incorporating retrieved evidence such as *"... linear atelectasis ..."* and *"Mild areas of atelectasis ..."*, RADAR successfully corrects the omission and generates a complete and accurate report. This case demonstrates the model's capability to lever-

Indication: History of chest pain x20 minutes, please evaluate for pneumonia or other acute process.



Preliminary Findings: O_R **[Cardiomegaly]**
[Cardiomegaly] Mild cardiomegaly has been stable compared to exams dated back to at least ___.
Supplementary Findings:
[Edema] Mild pulmonary edema is improved from ___.
[Lung Opacity] ......, there is an ill-defined somewhat rounded opacity ......
[Fracture] Left lateral rib fractures are again noted.
Expert Observations: O_I **[Edema, Fracture, Cardiomegaly]**
Agreement: $O_{\checkmark}$ **[Cardiomegaly]**
Overall Findings:
[Cardiomegaly] Mild cardiomegaly is stable compared to exams dated back to at least ___.
~~[Edema] There is mild pulmonary vascular congestion with mild pulmonary edema.~~ ......
Reference:
[Cardiomegaly] There is stable mild cardiomegaly.
[Lung Opacity] There appears to be a subtle increase in opacification in the retrocardiac region, ...

Figure 5: Error case generated by RADAR, where ~~spans~~ and spans indicate incorrect and correct observations.

age external knowledge to recover missing findings, thereby improving factual completeness.

Error Analysis. We conduct an error analysis to gain deeper insights, as shown in Figure 5. RADAR initially generates a report containing the observation *Cardiomegaly*, which is also present in the expert model's output. In this case, the observation reflects credible knowledge possessed by the LLM and should be preserved. Subsequently, RADAR produces a false positive finding, *Edema*, which aligns with the retrieved supplementary findings. This error may result from the model's overreliance on external knowledge. Moreover, since *Edema* is clinically associated with *Cardiomegaly*, it is possible that RADAR has learned only superficial correlations between them. To address these issues, potential solutions include refining the expert model and expanding the training dataset.

6 Related Works

Radiology report generation (Jing et al., 2018; Li et al., 2018) is a valuable yet challenging task. Numerous research efforts have been dedicated to improving clinical accuracy, employing diverse approaches such as memory-based neural models (Chen et al., 2020, 2021), planning-based methods (Nishino et al., 2022; Hou et al., 2023b), and reinforcement learning-optimized techniques (Lovelace and Mortazavi, 2020; Miura et al., 2021; Qin and Song, 2022). Additionally, several studies (Ramesh et al., 2022; Bannur et al., 2023; Hou et al., 2023a; Dalla Serra et al., 2023; Hou et al., 2024) have addressed the issue of hallucination, particularly in the absence of prior studies. Given the critical role of domain knowledge in this field, researchers have leveraged knowledge graphs to enhance report generation (Yang et al., 2021; Li et al., 2023c; Huang et al., 2023; Yan et al., 2023).

With the emergence of MLLMs (Li et al., 2023b; Liu et al., 2023), which demonstrate exceptional capabilities in image understanding and captioning, many studies (Singhal et al., 2022; Wu et al., 2023; Thawakar et al., 2024; Li et al., 2023a) have explored their application in the medical domain. Chen et al. (2024) introduced a foundation model for chest X-ray interpretation, while Chaves et al. (2024) developed a lightweight MLLM tailored for radiology. Park et al. (2024) investigated the multi-task potential of LLMs, and Zhang et al. (2024) incorporated temporal information to enhance chest X-ray analysis.

7 Conclusion

In this paper, we introduce RADAR, a novel approach designed to enhance radiology report generation by leveraging both the internal knowledge of an LLM and externally retrieved information. Our model first generates a report and subsequently classifies the image based on observations, with their shared components regarded as internal knowledge. It then retrieves supplementary information to further refine and complement this knowledge. Extensive experiments on three public datasets demonstrate that RADAR achieves SOTA performance in both language quality and clinical accuracy, highlighting the effectiveness of integrating internal and external knowledge for more accurate and coherent radiology report generation.

Limitations

Our experiments are conducted using a single backbone architecture. While this choice provides a controlled evaluation, the performance of alternative architectures remains unexplored. Future work should investigate whether different model architectures can achieve comparable or better results. In addition, our study focuses exclusively on a single imaging modality (e.g., Chest X-ray). The model's effectiveness in other imaging modalities, such as CT scans or MRI, has not been evaluated. Extending our approach to multiple imaging modalities would be an important direction for future research to enhance its clinical utility and generalizability.

Ethical Considerations

This study utilizes the MIMIC-CXR (Johnson et al., 2019), IU X-RAY (Demner-Fushman et al., 2016), and CHEXPERT PLUS (Chambon et al., 2024) datasets, all of which are publicly available and have been automatically de-identified to mitigate privacy risks. Our primary objective is to improve the clinical accuracy of reports generated by LLMs in medical imaging. However, despite our efforts, the generated reports may contain inaccuracies or omissions. Therefore, these outputs should not be used as a substitute for expert medical judgment. We strongly advocate for thorough validation by qualified radiologists or healthcare professionals before any clinical or diagnostic application.

Acknowledgments

This work was supported in part by the National Key R&D Program of China (2024YFE0198100, 2024YFC2510800), the Natural Science Foundation of China (82272086), and the Research Grants Council of Hong Kong (15207920, 15207821, 15207122).

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, Misha Bilenko, Johan Bjorck, Sébastien Bubeck, Martin Cai, Qin Cai, Vishrav Chaudhary, Dong Chen, Dongdong Chen, Weizhu Chen, Yen-Chun Chen, Yi-Ling Chen, Hao Cheng, Parul Chopra, Xiyang Dai, Matthew Dixon, Ronen Eldan, Victor Fragoso, Jianfeng Gao, Mei Gao, Min Gao, Amit Garg, Allie Del Giorno, Abhishek Goswami, Suriya Gunasekar, Emman Haider, Junheng Hao, Russell J. Hewett, Wenxiang Hu, Jamie Huynh, Dan Iter, Sam Ade Jacobs, Mojan Javaheripi, Xin Jin, Nikos Karampatziakis, Piero Kauffmann, Mahoud Khademi, Dongwoo Kim, Young Jin Kim, Lev Kurilenko, James R. Lee, Yin Tat Lee, Yuanzhi Li, Yunsheng Li, Chen Liang, Lars Liden, Xihui Lin, Zeqi Lin, Ce Liu, Liyuan Liu, Mengchen Liu, Weishung Liu, Xiaodong Liu, Chong Luo, Piyush Madan, Ali Mahmoudzadeh, David Majercak, Matt Mazzola, Caio César Teodoro Mendes, Arindam Mitra, Hardik Modi, Anh Nguyen, Brandon Norick, Barun Patra, Daniel Perez-Becker, Thomas Portet, Reid Pryzant, Heyang Qin, Marko Radmilac, Liliang Ren, Gustavo de Rosa, Corby Rosset, Sambudha Roy, Olatunji Ruwase, Olli Saarikivi, Amin Saied, Adil Salim, Michael Santacroce, Shital Shah, Ning Shang, Hiteshi Sharma, Yelong Shen, Swadheen Shukla, Xia Song, Masahiro Tanaka, Andrea Tupini, Praneetha Vaddamanu, Chunyu Wang, Guanhua Wang, Lijuan Wang, Shuohang Wang, Xin Wang, Yu Wang, Rachel Ward, Wen Wen, Philipp Witte, Haiping Wu, Xiaoxia Wu, Michael Wyatt, Bin Xiao, Can Xu, Jiahang Xu, Weijian Xu, Jilong Xue, Sonali Yadav, Fan Yang, Jianwei Yang, Yifan Yang, Ziyi Yang, Donghan Yu, Lu Yuan, Chenruidong Zhang, Cyril Zhang, Jianwen Zhang, Li Lyna Zhang, Yi Zhang, Yue Zhang, Yunan Zhang, and Xiren Zhou. 2024. Phi-3 technical report: A highly capable language model locally on your phone. Preprint, arXiv:2404.14219.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical bert embeddings. Preprint, arXiv:1904.03323.

Satanjeev Banerjee and Alon Lavie. 2005. 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, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, Felix Meissen, Mercy Ranjit, Shaury Srivastav, Julia Gong, Noel C. F. Codella, Fabian Falck, Ozan Oktay, Matthew P. Lungren, Maria Teodora Wetscherek, Javier Alvarez-Valle, and Stephanie L. Hyland. 2024. Maira-2: Grounded radiology report generation. Preprint, arXiv:2406.04449.

Shruthi Bannur, Stephanie Hyland, Qianchu Liu, Fernando Pérez-García, Maximilian Ilse, Daniel C. Castro, Benedikt Boecking, Harshita Sharma, Kenza Bouzid, Anja Thieme, Anton Schwaighofer, Maria Wetscherek, Matthew P. Lungren, Aditya Nori, Javier Alvarez-Valle, and Ozan Oktay. 2023. Learning to exploit temporal structure for biomedical vision-language processing. Preprint, arXiv:2301.04558.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Chu The Chuong, and Curtis P. Langlotz. 2024. Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats. Preprint, arXiv:2405.19538.

Juan Manuel Zambrano Chaves, Shih-Cheng Huang, Yanbo Xu, Hanwen Xu, Naoto Usuyama, Sheng Zhang, Fei Wang, Yujia Xie, Mahmoud Khademi, Ziyi Yang, Hany Awadalla, Julia Gong, Houdong Hu, Jianwei Yang, Chunyuan Li, Jianfeng Gao, Yu Gu, Cliff Wong, Mu Wei, Tristan Naumann, Muhao Chen, Matthew P. Lungren, Akshay Chaudhari, Serena Yeung-Levy, Curtis P. Langlotz, Sheng Wang, and Hoifung Poon. 2024. Towards a clinically accessible radiology foundation model: open-access and lightweight, with automated evaluation. Preprint, arXiv:2403.08002.

Zhihong Chen, Yaling Shen, Yan Song, and Xiang Wan. 2021. 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021, pages 5904–5914. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 2020. Generating radiology reports via memory-driven transformer. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1439–1449, Online. Association for Computational Linguistics.

Zhihong Chen, Maya Varma, Jean-Benoit Delbrouck, Magdalini Paschali, Louis Blankemeier, Dave Van Veen, Jeya Maria Jose Valanarasu, Alaa Youssef, Joseph Paul Cohen, Eduardo Pontes Reis, Emily B. Tsai, Andrew Johnston, Cameron Olsen, Tanishq Mathew Abraham, Sergios Gatidis, Akshay S. Chaudhari, and Curtis Langlotz. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. Preprint, arXiv:2401.12208.

Francesco Dalla Serra, Chaoyang Wang, Fani Deligianni, Jeff Dalton, and Alison O'Neil. 2023. Controllable chest X-ray report generation from longitudinal representations. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 4891–4904, Singapore. Association for Computational Linguistics.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. Journal of the American Medical Informatics Association, 23(2):304–310.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Yan Hu, Wenjie Li, and Jiang Liu. 2024. ICON: Improving inter-report consistency in radiology report generation via lesion-aware mixup augmentation. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 9043–9056, Miami, Florida, USA. Association for Computational Linguistics.

Wenjun Hou, Yi Cheng, Kaishuai Xu, Wenjie Li, and Jiang Liu. 2023a. RECAP: Towards precise radiology report generation via dynamic disease progression reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 2134–2147, Singapore. Association for Computational Linguistics.

Wenjun Hou, Kaishuai Xu, Yi Cheng, Wenjie Li, and Jiang Liu. 2023b. ORGAN: Observation-guided radiology report generation via tree reasoning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8108–8122, Toronto, Canada. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2):1–55.

Zhongzhen Huang, Xiaofan Zhang, and Shaoting Zhang. 2023. Kiut: Knowledge-injected u-transformer for radiology report generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 19809–19818.

Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, Noel Codella, Matthew P. Lungren, Maria Teodora Wetscherek, Ozan Oktay, and Javier Alvarez-Valle. 2024. Maira-1: A specialised large multimodal model for radiology report generation. Preprint, arXiv:2311.13668.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn L. Ball, Katie S. Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In The Thirty-Third AAAI Conference on Artificial Intelligence, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Q. H. Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P. Lungren, Andrew Y. Ng, Curtis P. Langlotz, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. CoRR, abs/2106.14463.

Haibo Jin, Haoxuan Che, Yi Lin, and Hao Chen. 2024. Promptmrg: Diagnosis-driven prompts for medical report generation. Preprint, arXiv:2308.12604.

Baoyu Jing, Pengtao Xie, and Eric P. Xing. 2018. On the automatic generation of medical imaging reports. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers, pages 2577–2586. Association for Computational Linguistics.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. Preprint, arXiv:2306.00890.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Preprint, arXiv:2301.12597.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. 2023c. Dynamic graph enhanced contrastive learning for chest x-ray report generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3334–3343.

Yuan Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. 2018. Hybrid retrieval-generation reinforced agent for medical image report generation. In Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada, pages 1537–1547.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chang Liu, Yuanhe Tian, Weidong Chen, Yan Song, and Yongdong Zhang. 2024. Bootstrapping large language models for radiology report generation. In AAAI, pages 18635–18643.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021. Exploring and distilling posterior and prior knowledge for radiology report generation. In IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021, pages 13753–13762. Computer Vision Foundation / IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. Preprint, arXiv:2304.08485.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. 2022. Swin transformer v2: Scaling up capacity and resolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 12009–12019.

Justin Lovelace and Bobak Mortazavi. 2020. Learning to generate clinically coherent chest X-ray reports. In Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1235–1243, Online. Association for Computational Linguistics.

Yasuhide Miura, Yuhao Zhang, Emily Tsai, Curtis Langlotz, and Dan Jurafsky. 2021. 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, pages 5288–5304, Online. Association for Computational Linguistics.

Toru Nishino, Yasuhide Miura, Tomoki Taniguchi, Tomoko Ohkuma, Yuki Suzuki, Shoji Kido, and Noriyuki Tomiyama. 2022. Factual accuracy is not enough: Planning consistent description order for radiology report generation. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Online. Association for Computational Linguistics.

Farhad Nooralahzadeh, Nicolas Perez Gonzalez, Thomas Frauenfelder, Koji Fujimoto, and Michael Krauthammer. 2021. Progressive transformer-based generation of radiology reports. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2824–2832, Punta Cana, Dominican Republic. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4v(ision) system card.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jonggwon Park, Soobum Kim, Byungmu Yoon, Jihun Hyun, and Kyoyun Choi. 2024. M4cxr: Exploring multi-task potentials of multi-modal large language models for chest x-ray interpretation. Preprint, arXiv:2408.16213.

Chantal Pellegrini, Ege Özsoy, Benjamin Busam, Nassir Navab, and Matthias Keicher. 2023. Radialog: A large vision-language model for radiology report generation and conversational assistance. Preprint, arXiv:2311.18681.

Han Qin and Yan Song. 2022. Reinforced cross-modal alignment for radiology report generation. In Findings of the Association for Computational Linguistics: ACL 2022, Dublin, Ireland, May 22-27, 2022, pages 448–458. Association for Computational Linguistics.

Vignav Ramesh, Nathan Andrew Chi, and Pranav Rajpurkar. 2022. Improving radiology report generation systems by removing hallucinated references to non-existent priors. Preprint, arXiv:2210.06340.

Mercy Ranjit, Gopinath Ganapathy, Ranjit Manuel, and Tanuja Ganu. 2023. Retrieval augmented chest x-ray report generation using openai gpt models. Preprint, arXiv:2305.03660.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Scharli, Aakanksha Chowdhery, Philip Mansfield, Blaise Aguera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2022. Large language models encode clinical knowledge. Preprint, arXiv:2212.13138.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 2020. 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), pages 1500–1519, Online. Association for Computational Linguistics.

Xiao Song, Xiaodan Zhang, Junzhong Ji, Ying Liu, and Pengxu Wei. 2022. Cross-modal contrastive attention model for medical report generation. In Proceedings of the 29th International Conference on Computational Linguistics, pages 2388–2397, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Liwen Sun, James Jialun Zhao, Wenjing Han, and Chenyan Xiong. 2025. Fact-aware multimodal retrieval augmentation for accurate medical radiology report generation. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 643–655, Albuquerque, New Mexico. Association for Computational Linguistics.

Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. 2023. Interactive and explainable region-guided radiology report generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7433–7442.

Omkar Chakradhar Thawakar, Abdelrahman M. Shaker, Sahal Shaji Mullappilly, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, Jorma Laaksonen, and Fahad Khan. 2024. XrayGPT: Chest radiographs summarization using large medical vision-language models. In Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, pages 440–448, Bangkok, Thailand. Association for Computational Linguistics.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023a. Metransformer: Radiology report generation by transformer with multiple learnable expert tokens. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 11558–11567.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023b. R2gengpt: Radiology report generation with frozen llms. Preprint, arXiv:2309.09812.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. Preprint, arXiv:2308.02463.

Le Xue, Manli Shu, Anas Awadalla, Jun Wang, An Yan, Senthil Purushwalkam, Honglu Zhou, Viraj Prabhu, Yutong Dai, Michael S Ryoo, Shrikant Kendre, Jieyu Zhang, Can Qin, Shu Zhang, Chia-Chih Chen, Ning Yu, Juntao Tan, Tulika Manoj Awalgaonkar, Shelby Heinecke, Huan Wang, Yejin Choi, Ludwig Schmidt, Zeyuan Chen, Silvio Savarese, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. 2024. xgen-mm (blip-3): A family of open large multimodal models. Preprint, arXiv:2408.08872.

Benjamin Yan, Ruochen Liu, David Kuo, Subathra Adithan, Eduardo Reis, Stephen Kwak, Vasantha Venugopal, Chloe O'Connell, Agustina Saenz, Pranav Rajpurkar, and Michael Moor. 2023. Style-aware radiology report generation with RadGraph and few-shot prompting. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 14676–14688, Singapore. Association for Computational Linguistics.

Shuxin Yang, Xian Wu, Shen Ge, Shaohua Kevin Zhou, and Li Xiao. 2021. Knowledge matters: Radiology report generation with general and specific knowledge. CoRR, abs/2112.15009.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. 2022. Evaluating progress in automatic chest x-ray radiology report generation. medRxiv.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. Preprint, arXiv:2303.15343.

Xi Zhang, Zaiqiao Meng, Jake Lever, and Edmond S. L. Ho. 2024. Libra: Leveraging temporal images for biomedical radiology analysis. Preprint, arXiv:2411.19378.

Hong-Yu Zhou, Subathra Adithan, Julián Nicolás Acosta, Eric J. Topol, and Pranav Rajpurkar. 2024. A generalist learner for multifaceted medical image interpretation. Preprint, arXiv:2405.07988.

A Appendix

A.1 Full List of Specialists

In addition to specialist baselines in Table 1, the following baselines are included: R2GEN (Chen et al., 2020), R2GENCMN (Chen et al., 2021), $\mathcal{M}^2$TR (Nooralahzadeh et al., 2021), KNOWMAT (Yang et al., 2021), CMM-RL (Qin and Song, 2022), CMCA (Song et al., 2022), KiUT (Huang et al., 2023), DCL (Li et al., 2023c), METrans (Wang et al., 2023a), RGRG (Tanida et al., 2023), RECAP (Hou et al., 2023a), Controllable (Dalla Serra et al., 2023), PromptMRG (Jin et al., 2024), and ICON (Hou et al., 2024).

Observation	**P**	**R**	**F_1**
Atelectasis	0.518	0.645	0.574
Cardiomegaly	0.656	0.783	0.713
Consolidation	0.370	0.174	0.237
Edema	0.518	0.610	0.560
Pleural Effusion	0.695	0.800	0.744
[5]Macro Average	0.551	0.602	0.567
[5]Micro Average	0.607	0.707	0.653
Enlarged Card.	0.277	0.204	0.235
Lung Opacity	0.644	0.496	0.561
Lung Lesion	0.492	0.207	0.291
Pneumonia	0.283	0.232	0.255
Pneumothorax	0.407	0.636	0.496
Pleural Other	0.333	0.173	0.228
Fracture	0.421	0.244	0.309
Support Devices	0.823	0.866	0.844
No Finding	0.302	0.569	0.395
[14]Macro Average	0.481	0.474	0.460
[14]Micro Average	0.614	0.640	0.627

Table 7: Experimental results of RADAR for each observation on the MIMIC-CXR dataset.

Dataset: MIMIC-CXR *(Compared with SOTA Specialists)*									
Model	**Lexical Metrics**						**CE (^{14}Macro) Metrics**		
	B-1	**B-2**	**B-3**	**B-4**	**MTR**	**R-L**	**P**	**R**	**F$_1$**
R2Gen	0.353	0.218	0.145	0.103	0.142	0.270	0.333	0.273	0.276
R2GenCMN	0.353	0.218	0.148	0.106	0.142	0.278	0.344	0.275	0.278
$\mathcal{M}^2$Tr	0.378	0.232	0.154	0.107	0.145	0.272	0.240	0.428	0.308
KnowMat	0.363	0.228	0.156	0.115	–	0.284	0.458	0.348	0.371
CMM-RL	0.381	0.232	0.155	0.109	0.151	0.287	0.342	0.294	0.292
CMCA	0.360	0.227	0.156	0.117	0.148	0.287	0.444	0.297	0.356
KiUT	0.393	0.243	0.159	0.113	0.160	0.285	0.371	0.318	0.321
DCL	–	–	–	0.109	0.150	0.284	0.471	0.352	0.373
METrans	0.386	0.250	0.169	0.124	0.152	0.291	0.364	0.309	0.311
RGRG	0.373	0.249	0.175	0.126	0.168	0.264	0.380	0.319	0.305
ORGan	0.407	0.256	0.172	0.123	0.162	0.293	0.416	0.418	0.385
Recap	0.429	0.267	0.177	0.125	0.168	0.288	0.389	0.443	0.393
Controllable	0.486	0.366	0.295	0.246	0.216	**0.423**	**0.597**	**0.516**	**0.553**
PromptMRG	0.398	–	–	0.112	0.157	0.268	0.396	0.393	0.381
ICon	0.429	0.266	0.178	0.126	0.170	0.287	0.445	0.505	0.464
Radar (Ours)	**0.509**	**0.390**	**0.315**	**0.262**	**0.450**	0.397	**0.481** 0.523	**0.474** 0.500	**0.460** 0.497

Table 8: Experimental results of our model and SoTA specialists on the MIMIC-CXR dataset. Results denotes *Uncertain* as *Positive*.

Role	**Prompt**
SYSTEM	<|system|> You are an assistant in radiology, responsible for analyzing medical imaging studies and generating detailed, structured, and accurate radiology reports. <|end|>
USER	<|user|> <prior image> (*If prior available*) <current image> *Indication*: ...... *History*: ...... *Comparison*: ...... *Technique*: ...... *Prior Findings*: ...... (*If prior available*) *Preliminary Findings*: ...... (*If available*) *Supplementary Findings*: ...... (*If available*) Generate a comprehensive and detailed description of findings based on this chest X-ray image. Include a thorough comparison with a prior chest X-ray, emphasizing any significant changes, progression, or improvement. (*If prior available*) Before this, systematically identify all observations. <|end|>
ASSISTANT	<|assisitant|> *Identified Observations*: ...... *Overall Findings*: *(e.g., the target)* ...... <|end|>

Table 9: The prompt template for RADAR and its variants, consisting of three roles: System, User, and Assistant.