LlamaRG: A Multi-View Large Language Model for Radiology Report Generation

Supplementary Material

A.1. Dataset overview and pre-processing

The details of the MIMIC-CXR dataset, which has been used in this study, are presented in Table A5. This includes the number of study samples we have used. To avoid any inconsistency with previous works [3, 4, 45], we have followed a similar pre-processing protocol. In this study, we filtered out samples with incomplete findings in their reports. For the remaining samples, a cleaning pipeline was used to reduce noise and standardize all the samples before training. Reports were converted to lowercase, and all special characters and punctuation marks were removed. Numbered lists and redundant formatting were also removed to avoid artifacts. Furthermore, an 'end token' is added to the reports to tell the model where the reports end.

Table A5. The statistics of MIMIC-CXR dataset including the number of patients, studies, CXR-images and average length (Avg Len) of reports (in tokens).

	Train	Validation	Test
Subjects	64,586	500	293
Studies	222,758	1,808	3,269
CXRs	368,960	2,991	5,159
Avg Len	53.00	53.05	66.40

A.2. Performance of llamaRG in generating findings

The primary sources of abnormality descriptions are 'FINDINGS' and 'IMPRESSION' in a radiology report. The 'FINDINGS' section describes all the observations regarding anatomical systems and any abnormalities present in detail from different views of X-ray images, and the 'IMPRESSION' section provides a summary and diagnostic interpretation based on the findings. Table A6 summarizes the performance of our model llamaRG in generating the 'FINDINGS' section, along with the 'FINDINGS' and 'IMPRESSION' sections together. The results indicate that our model llamaRG is capable of generating both sections.

A.3. Clinical Accuracy across different labels

Table A7 summarizes the CE metrics for clinical label extraction obtained by applying the CheXpert labeller to the generated reports. We evaluate Precision, Recall, and F1-score for the 14 labels of CheXpert. Compared to the baseline R2GenGPT, llamaRG (both stages of training) achieved higher results—especially F1-score— on majority of the clinical labels, including cardiomegaly, lung opacity, atelectasis, pneumothorax, and pleural effusion. Overall, the macro-average trends confirm that llamaRG generates more clinically accurate reports that are better aligned with the CheXpert labeller.

A.4. Additional Examples

Figure A4 presents additional examples of reports—with 'FINDINGS' and 'IMPRESSION' sections— generated by both stages of llamaRG compared with ground truth and baseline model, R2GenGPT. These examples illustrate that our model continuously identifies the correct position of the views across multi-view studies. In contrast, the R2GenGPT baseline often misattributes view information, sometimes reporting two views despite being trained as a single input method. Also, as shown in Figure A4(a), R2GenGPT generates several clinically incorrect and hallucinated findings—including missed pulmonary edema and cardiomegaly. Furthermore, our model consistently captures key abnormalities, device position, and locates pleural densities with higher clinical specificity. However, one limitation of llamaRG is that it over-reports subtle ancillary findings when visual cues are ambiguous. Despite this, llamaRG improves clinical specificity and view-awareness, generating more reliable reports than prior methods.

A.5. Comparison between different Reinforcement Learning techniques

To evaluate the effect of different RL algorithms during the second stage of training, we also compared the SCST method with Proximal Policy Optimization (PPO) [33], a commonly used reinforcement learning algorithm for sequence generation. PPO uses a critic model to update token level policy which can be noisy and unstable, whereas SCST uses its own greedy baseline to anchor learning and reduce reward variance. This is also shown in Table A8, SCST with a hybrid reward function demonstrates superior performance over PPO, yielding better performance across both NLG metrics (BLEU-1, BLEU-4, METEOR) and CE metrics (Precision and F1 score). These results indicate that SCST when paired with clinically oriented rewards, provides more stable optmization and generated reports with clinical correctness and textual coherence.

Table A6. Performance on MIMIC-CXR dataset for generating "FINDINGS" and "IMPRESSION" sections in the report. 'B-1', 'B-2', 'B-3', 'B-4', 'RG', 'MTR', 'Cr', 'P', 'R', refers to BLEU-1, BLEU-2, BLEU-3, BLEU-4, ROUGE-L, METEOR, CIDEr, Precision, and Recall, respectively.

Generated Section(s)	**NLG Metrics**							**CE Metrics**		
	B-1	B-2	B-3	B-4	RG	MTR	Cr	P	R	F1
FINDINGS	0.390	0.239	0.160	0.115	0.277	0.152	0.148	0.454	0.367	0.379
FINDINGS + IMPRESSION	0.417	0.270	0.178	0.130	0.302	0.162	0.169	0.448	0.431	0.414

Table A7. Clinical accuracy on the MIMIC-CXR dataset. "ECM" refers to Enlarged Cardiomedastinum. "P", "R", and "F1" represent Precision, Recall, and F1 score, respectively.

CheXpert Label	**llamaRG (Stage 1)**			**llamaRG (Stage 2)**			**R2GenGPT [45]**		
	P↑	R↑	F1↑	P↑	R↑	F1↑	P↑	R↑	F1↑
ECM	0.098	0.026	0.042	**0.263**	**0.033**	**0.059**	0.059	0.022	0.032
Cardiomegaly	**0.570**	0.486	0.525	0.545	**0.532**	**0.539**	0.536	0.460	0.495
Lung Opacity	0.579	0.153	0.242	**0.621**	**0.186**	**0.287**	0.577	0.115	0.192
Lung Lesion	0.400	**0.017**	**0.032**	**1**	**0.008**	**0.016**	0.300	0.013	0.026
Edema	0.574	0.351	0.435	0.472	**0.448**	**0.460**	**0.579**	0.115	0.191
Consolidation	0.158	0.031	0.052	**0.472**	**0.448**	**0.460**	0.250	0.083	0.125
Pneumonia	**0.440**	0.066	0.115	0.388	**0.156**	**0.222**	0.392	0.151	0.218
Atelectasis	**0.444**	0.255	0.324	0.376	**0.732**	**0.497**	0.400	0.434	0.416
Pneumothorax	0.222	0.035	0.061	**0.250**	**0.088**	**0.130**	0.179	0.060	0.089
Pleural Effusion	**0.805**	0.390	0.525	0.619	**0.827**	**0.708**	0.703	0.619	0.658
Pleural Other	**0.500**	**0.015**	**0.029**	0.000	0.000	0.000	0.167	0.007	0.014
Fracture	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Support Devices	0.756	**0.720**	**0.737**	**0.804**	0.652	0.720	0.676	0.709	0.692
No Finding	0.128	**0.496**	0.203	**0.220**	0.424	**0.290**	0.139	0.410	0.207
macro avg	0.405	0.217	0.283	**0.415**	**0.300**	**0.348**	0.354	0.228	0.278

Table A8. Comparision between basic RL technique, PPO and SCST with hybrid reward function. 'B-1', 'B-4', 'MTR', and 'P', represent BLEU-1, BLEU-2, METEOR and Precision, respectively.

RL technique	BL-1	BL-4	MTR	P	F1
PPO	0.397	0.119	0.154	0.446	0.382
SCST	0.417	0.130	0.162	0.448	0.414

A.6. External data validation

We further evaluated the generalization capability of our method through external data validation on IU-Xray and RexGradient datasets. We additionally included RateScore(RS) as a complementary metric to assess clinical correctness alongside CE metrics. These results, presented in Table A9, demonstrate that llamaRG shows consistent improvement compared to MedGemma, demonstrating robust generalization across different datasets.

Table A9. External data validation comparison between MedGemma and llamaRG. "P", "R", "F1", "RS" represent Precision, Recall, F1 score, and RateScore respectively.

Dataset	Method	P	R	F1	RS
RexGrad	MedGemma	0.239	0.213	0.214	0.477
	llamaRG-SCST	**0.250**	**0.251**	**0.244**	**0.483**
IU-Xray	MedGemma	0.144	0.141	0.139	0.509
	llamaRG-SCST	**0.497**	**0.495**	**0.494**	**0.579**

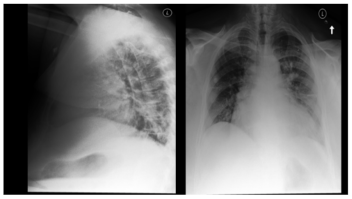
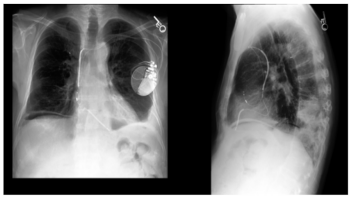
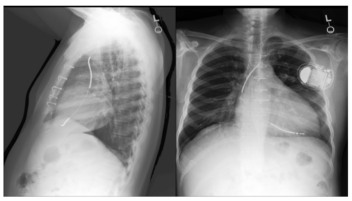
	Original Input	Ground Truth	llamaRG - stage1	llamaRG - stage2	R2GenGPT
a		impression : moderate pulmonary edema overall increased compared to the prior exam from. findings : moderate cardiomegaly is all stable compared to the prior exams dated back to at least. there has been an interval increase in bilateral moderate pulmonary edema with interstitial thickening and perihilar vascular congestion compared to the prior exam from. there may be small bilateral pleural effusions. there is no evidence of pneumothorax. the visualized osseous structures are unremarkable.	impression : cardiomegaly with mild pulmonary edema. findings : pa and lateral views of the chest provided. cardiomegaly is again noted with hilar congestion and mild pulmonary edema. no large effusion or pneumothorax. no convincing signs of pneumonia. mediastinal contour is stable. bony structures are intact. no free air below the right hemidiaphragm.	impression : cardiomegaly with mild pulmonary edema. no focal consolidation. findings : frontal and lateral views of the chest. the cardiac and mediastinal silhouettes are stable with the cardiac silhouette enlarged. there is mild pulmonary edema. no focal consolidation is seen. there is no pleural effusion or pneumothorax.	impression : no evidence of acute cardiopulmonary process. findings : lateral view somewhat limited due to overlying motion artifact. the lungs are low in volume. there is no focal airspace consolidation to suggest pneumonia. a 12-cm calcified granuloma just below the medial aspect of the right hemidiaphragm is unchanged from prior study. no pleural effusions or pulmonary edema.
b		impression : stable chest findings . persistent loculated pleural density on the left base and parenchymal density occupying posterior portions of the left lower lobe . findings : pa and lateral chest views were obtained with the patient in upright position . analysis is performed in direct comparison with the next preceding pa and lateral chest examination of . previously described heart size mediastinal structures and permanent pacer with dual electrode system remain unchanged. the same holds also with the previously described loculated pleural effusion that blunts the left-sided lateral pleural sinus.	impression : small left pleural effusion. findings : there is a small left pleural effusion. there is no focal consolidation or pneumothorax. the cardiomediastinal silhouette is within normal limits. left chest wall dual lead pacing device is seen with leads terminating in the right atrium and right ventricle.	impression : small bilateral pleural effusions with adjacent atelectasis. no evidence of pneumonia. findings : frontal and lateral views of the chest demonstrate low lung volumes which accentuate bronchovascular markings. there are small bilateral pleural effusions with adjacent atelectasis. there is no focal consolidation or pneumothorax. the cardiomediastinal and hilar contours are unremarkable. there is a left-sided pacemaker with leads terminating in the right atrium and right ventricle. there is no pulmonary edema.	impression : no acute cardiopulmonary abnormalities. findings : left-sided pacemaker with leads terminating in the right atrium and right ventricle is unchanged. there is no pneumothorax or pleural effusion. there is no focal consolidation or pulmonary edema. the cardiomediastinal and hilar contours are within normal limits.
c		impression : mild cardiomegaly without superimposed acute cardiopulmonary process. findings : the lungs are clear without focal consolidation effusion or edema. left chest wall single lead pacing device is noted. mild cardiomegaly is noted. median sternotomy wires and mediastinal clips are seen. prior endotracheal and enteric tubes are no longer visualized.	impression : no acute cardiopulmonary process. findings : the lungs are clear without focal consolidation. no pleural effusion or pneumothorax is seen. the cardiac and mediastinal silhouettes are stable with the cardiac silhouette top-normal to mildly enlarged. left chest wall dual lead pacing device is again seen. no pulmonary edema is seen.	impression : cardiomegaly without superimposed acute cardiopulmonary process. findings : frontal and lateral views of the chest. the lungs are clear without consolidation effusion or pulmonary vascular congestion. the cardiac silhouette is enlarged but stable. left chest wall dual lead pacing device is again seen. median sternotomy wires and mediastinal clips are again noted. no acute osseous abnormality is identified.	impression : no acute cardiopulmonary process. findings : pa and lateral views of the chest provided. left chest wall single lead pacing device is again seen. midline sternotomy wires and mediastinal clips are again noted. there is no focal consolidation effusion or pneumothorax. the cardiomediastinal silhouette is stable. no acute osseous abnormalities .

Figure A4. Additional examples of generated reports by two stages of llamaRG compared with the ground truth and R2GenGPT. The highlighted texts represent phrases that overlap with the ground truth report. The wrong generated phrases are shown in red.