Automated Structured Radiology Report Generation

Jean-Benoit Delbrouck, Justin Xu, Johannes Moll, Alois Thomas, Zhihong Chen, Sophie Ostmeier, Asfandyar Azhar, Kelvin Zhenghao Li, Andrew Johnston, Christian Bluethgen, Eduardo Reis, Mohamed Muneer, Maya Varma, Curtis Langlotz

Stanford AIMI

jbdel@stanford.edu github.com/jbdel/StructEval huggingface.co/collections/StanfordAIMI/structured-radiology-reports

What is structured reporting?

Generating radiology reports using strict, template-driven rules to ensure clear, consistent, and standardized clinical documentation.

EXAMPLE

Findings:

Lungs and Airways:	- Mild pulmonary edema, slightly increased.
Pleura:	- No significant pleural effusion.
Cardiovascular:	- Moderate enlargement of the cardiac silhouette, stable.

Impression: 1. Slight increase in mild pulmonary edema.
2. Stable moderate enlargement of the cardiac silhouette.

Motivation

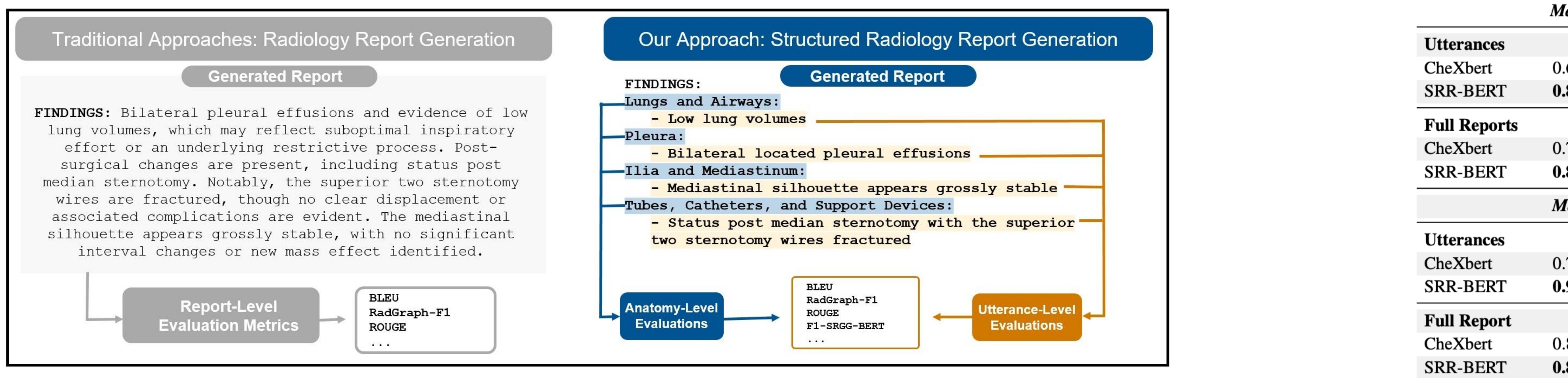
- Free-form reports make it hard for AI systems to learn accurate generation
- Generated free-form reports are difficult to evaluate
- Structured reports are easier to read and understand
- Structured reporting can benefit downstream machine learning applications

Our Contributions

1. We create a **novel dataset** by restructuring reports using large language models (LLMs) following strict structured reporting desiderata
2. We introduce SRR-BERT, a **fine-grained disease classification model** trained on 55 labels.
3. We introduce a **new evaluation framework** for structured radiology reporting.
4. We **fine-tune popular RRG systems** (CheXagent, MAIRA-2, RaDialog, and CheXpert-Plus) on the restructured findings and impressions.

Our Approach: SRRG

Traditional methods generate unstructured reports that vary in style and clarity, making automated evaluation challenging. In contrast, **SRRG** enforces a **standardized format with anatomical section headers**. This structured format enables more granular anatomy-level and utterance-level evaluations, including our proposed F1-SRR-BERT metric, which complements traditional report-level evaluation metrics.



Traditional Approaches: Radiology Report Generation
Generated Report
Report-Level Evaluation Metrics
Our Approach: Structured Radiology Report Generation
Generated Report
Anatomy-Level Evaluations
Utterance-Level Evaluations

FINDINGS: Bilateral pleural effusions and evidence of low lung volumes, which may reflect suboptimal inspiratory effort or an underlying restrictive process. Post-surgical changes are present, including status post median sternotomy. Notably, the superior two sternotomy wires are fractured, though no clear displacement or associated complications are evident. The mediastinal silhouette appears grossly stable, with no significant interval changes or new mass effect identified.

```
BLEU
RadGraph-F1
ROUGE
...
```

```
FINDINGS:
Lungs and Airways:
  - Low lung volumes
Pleura:
  - Bilateral located pleural effusions
Ilia and Mediastinum:
  - Mediastinal silhouette appears grossly stable
Tubes, Catheters, and Support Devices:
  - Status post median sternotomy with the superior
  two sternotomy wires fractured
```

```
BLEU
RadGraph-F1
ROUGE
F1-SRGG-BERT
...
```

Dataset Generation

We structured our dataset to align with the Radiology Report Generation (RRG) task by specifically mapping chest X-ray images to:
Findings (X-ray → Findings)
Impressions (X-ray →Impression).

Dataset	Split	Num. Examples
SRRG-Impression	Train	405,972
	Validate	1,505
	Test	2,219
	Test Reviewed	231
	Total	409,927
SRRG-Findings	Train	181,874
	Validate	976
	Test	1,459
	Test Reviewed	233
	Total	184,542

Dataset	Split	Num. Examples
StructUtterances	Train	1,203,332
	Validate	150,417
	Test	150,417
	Test Reviewed	1,609
	Total	1,506,158

We then annotate all utterances in our SRRG datasets, where an utterance is defined as either a single-sentence finding or a numbered impression. This process results in 1,506,158 utterances annotated with 1,782,983 labels.

Our reader study, conducted by board-certified radiologists, confirms the quality of both the structured reports and annotated disease labels.

Experiments

Disease Classification Model: To benchmark disease classification, we fine-tune CXR-BERT (Boecking et al., 2022) on weakly-labeled utterances in the StructUtterances dataset and compare it to CheXbert.

	Precision	Recall	F1-Score	Support
		Mapped with Leaves		
Utterances				
CheXbert	0.69	0.64	0.65	1,759
SRR-BERT	**0.88**	**0.82**	**0.84**	1,759
Full Reports				
CheXbert	0.73	**0.59**	**0.62**	260
SRR-BERT	**0.84**	0.48	0.58	260
		Mapped with Upper		
Utterances				
CheXbert	0.70	0.48	0.50	2,004
SRR-BERT	**0.90**	**0.84**	**0.86**	2,004
Full Report				
CheXbert	0.80	0.49	0.56	278
SRR-BERT	**0.89**	**0.60**	**0.70**	278

Structured RRG Benchmark: We benchmark four distinct models: **MAIRA-2** (Bannur et al., 2024), **CheXagent** (Chen et al.,2024), **CheXpert-Plus** (Chambon et al., 2024), and **RaDialog** (Pellegrini et al., 2023).

SRRG-Impression (unaligned)		Traditional Metrics				F1-SRR-BERT		
Model	**Split**	**BLEU**	**ROUGE-L**	**BERTScore**	**F1-RadGraph**	**Precision**	**Recall**	**F1-Score**
CheXagent	Validate	7.86	28.94	60.55	20.62	50.02	56.32	50.60
CheXagent	Test	**6.95**	**27.18**	**61.51**	**19.70**	**49.78**	56.47	**50.63**
CheXagent	Test Reviewed	4.68	26.10	59.70	18.33	45.24	**56.70**	48.64
CheXpert-Plus	Validate	16.86	33.42	62.74	27.74	54.40	51.26	50.26
CheXpert-Plus	Test	**14.84**	**28.01**	**60.76**	**22.14**	**48.74**	47.60	**46.48**
CheXpert-Plus	Test Reviewed	14.07	26.79	59.21	18.89	43.46	**48.15**	44.56
MAIRA-2	Validate	9.66	31.50	62.84	23.21	52.53	61.16	54.46
MAIRA-2	Test	**8.12**	**27.82**	**62.30**	**20.37**	**48.72**	**57.91**	**50.36**
MAIRA-2	Test Reviewed	5.28	26.61	60.79	19.08	44.80	57.69	47.97
RaDialog	Validate	5.35	23.93	57.74	15.27	39.80	52.41	40.70
RaDialog	Test	3.32	**21.59**	**57.48**	**12.32**	**37.30**	50.59	**39.22**
RaDialog	Test Reviewed	**3.33**	19.95	54.82	10.26	33.65	**50.71**	36.39

Takeaway: SRRG improves consistency compared to existing free-form generation methods.

Additional Information

Paper

Code