Automated Structured Radiology Report Generation

Jean-Benoit Delbrouck[♠♡] jbdel@stanford.edu
Justin Xu[♠]
Johannes Moll[♠]
Alois Thomas[♠]
Zhihong Chen[♠]
Sophie Ostmeier[♠]
Asfandyar Azhar[♠]
Kelvin Zhenghao Li[♠]
Andrew Johnston[♠]
Eduardo Reis[♠]
Christian Bluethgen[♠]
Mohamed Muneer[♠]
Maya Varma[♠]
Curtis Langlotz[♠]

[♠] Stanford AIMI [♡] HOPPR

🤗 https://huggingface.co/StanfordAIMI
https://stanford-aimi.github.io/srrg.html

Abstract

Automated radiology report generation from chest X-ray (CXR) images has the potential to improve clinical efficiency and reduce radiologists' workload. However, most datasets, including the publicly available MIMIC-CXR and CheXpert Plus, consist entirely of free-form reports, which are inherently variable and unstructured. This variability poses challenges for both generation and evaluation: existing models struggle to produce consistent, clinically meaningful reports, and standard evaluation metrics fail to capture the nuances of radiological interpretation. To address this, we introduce Structured Radiology Report Generation (SRRG), a new task that reformulates free-text radiology reports into a standardized format, ensuring clarity, consistency, and structured clinical reporting. We create a novel dataset by restructuring reports using large language models (LLMs) following strict structured reporting desiderata. Additionally, we introduce SRR-BERT, a fine-grained disease classification model trained on 55 labels, enabling more precise and clinically informed evaluation of structured reports. To assess report quality, we propose F1-SRR-BERT, a metric that leverages SRR-BERT's hierarchical disease taxonomy to bridge the gap between free-text variability and structured clinical reporting. We validate our dataset through a reader study conducted by five board-certified radiologists and extensive benchmarking experiments.

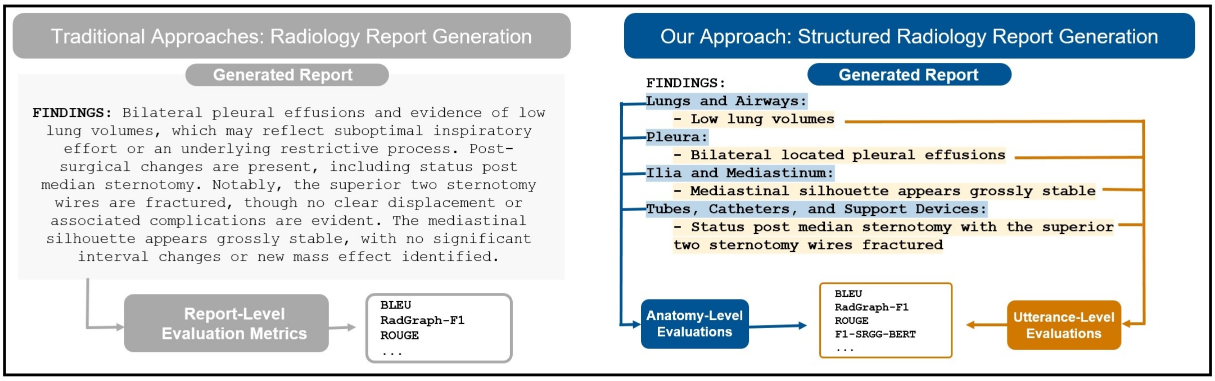
1 Introduction

An important medical application of natural language generation (NLG) is the construction of assistive systems that take X-ray images of a patient and generate a textual report describing clinical observations in the images. This is a clinically important task, offering the potential to reduce the repetitive workload of radiologists and generally improve clinical communication (Dunnick and Langlotz, 2008; Kahn Jr et al., 2009).

Since this task was first explored on chest X-ray (CXR) images, much of the related work, including exploring vanilla transformers (Chen et al., 2020), reinforcement learning algorithms (Miura et al., 2021; Delbrouck et al., 2022), and foundation models (Chen et al., 2024; Bannur et al., 2024), has been conducted on two primary datasets: MIMIC-CXR (Johnson et al., 2019) and CheXpert Plus (Chambon et al., 2024). These datasets share notable similarities in terms of size, population diversity, and reporting style.

However, it is important to note that CXR reports themselves are typically free-form rather than structured by organ systems, primarily due to protocols, workflow efficiency, and the holistic nature of the necessary image interpretation (Weiss and Langlotz, 2008; Bosmans et al., 2012). This free-form style can pose unique challenges for automated report generation and clinical decision support as the variability in reporting styles often leads to inconsistencies in the way findings are described.

The need for more consistent, structured, or template-based radiology reporting is further reinforced by the difficulty faced by all proposed metrics in evaluating automated radiology report generation. Existing evaluation methods, ranging from standard NLG metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) to clinical factuality-based metrics such as F1-RadGraph (Delbrouck et al., 2022), RadFact (Bannur et al., 2024), or GREEN (Ostmeier


Traditional Approaches: Radiology Report Generation
Generated Report
Report-Level Evaluation Metrics
BLEU
RadGraph-F1
ROUGE
...
Our Approach: Structured Radiology Report Generation
Generated Report
Anatomy-Level Evaluations
BLEU
RadGraph-F1
ROUGE
F1-SRGG-BERT
...
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.
```

```
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
```

Figure 1: Comparison between traditional free-text radiology report generation (left) and our proposed Structured Radiology Report Generation (SRRG) approach (right). 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.

et al., 2024), may struggle to capture the nuances of radiological interpretation due to the inherent diversity in reporting styles.

Given these limitations and observations, we introduce a new task, Structured Radiology Report Generation (SRRG, Section 2), aimed at transforming free-text radiology reports into a standardized format that enhances clarity and consistency through structured clinical documentation. To support this task, we present a new dataset derived from MIMIC-CXR and CheXpert Plus, where reports have been reformulated using large language models (LLMs) following strict desiderata for structured reporting. Additionally, we introduce SRR-BERT (Section 3), a novel disease classification model with 55 labels, designed to enable fine-grained automated evaluation. To further enhance the assessment of generated structured reports, we propose F1-SRR-BERT, a new metric that leverages SRR-BERT's hierarchical disease taxonomy alongside a more precise evaluation paradigm made possible by the structured design of our task (Section 4.2.1). We validate our new datasets through a reader study (Section B) conducted by five board-certified radiologists, along with extensive experiments (Section 4).

2 Structured Radiology Reporting

2.1 Desiderata

We define a structured radiology report as a report that follows a standardized format to ensure clarity and consistency. Such a report consists of distinct sections, each introduced by a section header followed by a colon, ensuring uniformity in presentation. The required sections include **Exam Type, History, Technique, Comparison, Findings, and Impression**.

The Findings section is organized under predefined anatomical headers, which are strictly limited to the following categories: **Lungs and Airways, Pleura, Cardiovascular, Hila and Mediastinum, Tubes, Catheters, and Support Devices, Musculoskeletal and Chest Wall, Abdominal, and Other**. Within each category, observations should be clearly listed using bullet points, and include all relevant positive and negative findings.

The Impression section summarizes the key findings in a numbered list, ranked from most to least clinically significant, ensuring that the most critical observations are highlighted effectively.

To maintain clarity and relevance, strict content guidelines need to be applied. References to previous studies or historical comparisons should be excluded, ensuring that the report reflects only the current examination. Identifiable information, including dates, surnames, first names, healthcare providers, vendors, and institutions, must be removed, although patient sex and age should be retained when provided. The content must strictly adhere to the defined structured sections, without extrapolating interpretations or introducing unrelated details. Additionally, only the specified anatomical headers may be used, ensuring a standardized report. The full prompt is available in Prompt 2.

2.2 Dataset Creation

Previous research has shown that GPT models can outperform traditional fine-tuned models in general summarization tasks by offering better factual consistency and reducing hallucinations (Pu et al., 2023), achieve human-level performance in medical summarization of findings (Van Veen et al., 2024), and demonstrate strong capabilities in radiological error categorization (Ostmeier et al., 2024). Motivated by this, as well as by GPT-4's "exceptional" performance across various medical benchmarks (Nori et al., 2023), we leverage LLMs to restructure the two largest publicly available chest X-ray datasets: MIMIC-CXR (Johnson et al., 2019) and CheXpert Plus (Chambon et al., 2024). The prompt used to rephrase the reports in accordance with our desiderata is provided in Prompt 4. This prompt was executed using GPT-4 "Turbo 1106 preview" via Azure services, with the account explicitly opted out of human review.

2.3 Dataset Statistics

We structured our dataset to align with the Radiology Report Generation (RRG) task by specifically mapping X-ray images to Findings (X-ray → Findings) and Impressions (X-ray → Impression). These setups correspond to our datasets, SRRG-Findings and SRRG-Impression, respectively. To construct the SRRG dataset, we combined MIMIC-CXR and CheXpert Plus and pooled them together to create our splits. Notably, SRRG-Impression is larger than SRRG-Findings, primarily because CheXpert predominantly contains Impression sections while often lacking Findings sections.
Lastly, we conducted a human review of 464 reports, sampled from the MIMIC-CXR test set and the CheXpert Plus validation set, with evaluations performed by five board-certified radiologists (Appendix B). Statistics of our datasets and splits are highlighted in Table 1.

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

Table 1: Dataset statistics for SRRG-Impression and SRRG-Findings.

3 Disease Classification Models

In this section, we introduce SRR-BERT, a novel model for fine-grained disease prediction that builds upon CheXbert to provide a more detailed assessment. Our approach extends the traditional set of 14 CheXbert disease labels to a set of 55 labels, covering a more granular hierarchy of pulmonary, pleural, cardiac, mediastinal, musculoskeletal, and abdominal findings, as well as more detailed support devices. This expanded taxonomy allows for more precise classification and evaluation of radiological abnormalities, enhancing the depth and accuracy of disease prediction.

3.1 Desiderata

To ensure **clarity, consistency, and clinical relevance**, our disease annotation framework follows the following key principles. Each finding must be mapped to **all relevant diseases** from a predefined list, allowing for zero, one, or multiple conditions. If no disease is present, the annotation explicitly states *"No Finding"* to ensure systematic coverage. Every disease is assigned a **status**—*Present*, *Absent*, or *Uncertain*—capturing clinical uncertainty and preventing over-assumptions. For example:

> *Right perihilar consolidation, likely atypical edema, with pneumonia as a differential diagnosis.*

is annotated as:

```
=> Perihilar airspace opacity
(Present)
=> Edema (Uncertain)
=> Pneumonia (Uncertain)
```

The selected diseases and their hierarchical structure are detailed in Prompt 4. This disease tree has been validated by a board-certified radiologist. While the first level of the hierarchical structure corresponds to the Anatomical Headers / Category, the lowest level is referred to as tree "leaves", and "upper" labels denote the item one-level above "leaves". Appendix D shows a dataset breakdown of each of the "upper" labels.

3.2 Dataset Creation

We annotate all utterances in our SRRG dataset, where an utterance is defined as either a single-sentence finding or a numbered impression. This process results in 1,562,277 unique impressions. To ensure consistency in annotation, we follow the guidelines outlined in Section 3.1 and craft the structured annotation template accordingly provided in Prompt 3.

To validate the correctness of the assigned labels, we employ both automated and human reviews. The automated review follows a mixture-of-experts approach, where each utterance is processed using three different GPT models: GPT-4 Turbo (2024-04-09), GPT-4 Turbo 1106 Preview, and GPT-4o (2024-08-06). The final labels for each utterance are determined by selecting the diseases that appear in at least two out of the three model outputs. This ensures robustness and reduces inconsistencies in the predictions. If an utterance has no labels, we discard it. We ultimately obtain a total of 1,506,158 valid utterances (as detailed in Section 3.3)

3.3 Dataset Statistics

The dataset comprises 1,506,158 utterances annotated with 1,782,983 labels, averaging 1.18 labels per utterance. Among all utterances, 905,764 correspond to positive findings (i.e., not labeled as "No Finding"), with these having an average of 1.31 labels per utterance.

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

Table 2: Dataset statistics for StructUtterances.

The test-reviewed split was evaluated by five board-certified radiologists (Appendix B) and includes utterances extracted from the reports in the test-reviewed split of our SRRG dataset (Table 1).

4 Benchmarking

4.1 Disease Classification Models

To benchmark disease classification, we fine-tune CXR-BERT (Boecking et al., 2022) on weakly-labeled utterances in the StructUtterances dataset under four experimental settings. First, we set aside the status annotations (i.e., Present, Absent, Uncertain) and only classify the "leaves" and "upper" labels. We then integrate the three statuses by creating a separate class for each combination, yielding "leaves with statuses" and "upper with statuses".
The benchmarking results for the disease classification models demonstrate strong overall performance on the reviewed test split, with F1 scores exceeding 0.75 for most classes. However, as is typical in classification tasks, rare labels posed a challenge. For the model operating at the "leaves" level, the overall F1 score was 0.836, with the three best-performing labels being "No Finding" (F1 = 0.83, n=452), "Simple Pleural Effusion" (F1 = 0.93, n=174), and "Atelectasis" (F1 = 0.94, n=131). Noticeably poor-performing classes include "Air space opacity-multifocal" (F1 = 0.62, n=60) and "Suboptimal central line" (F1 = 0.19, n=29). At the "upper" level with reduced granularity, our model achieved an overall F1 score of 0.839, with top-three performing labels being "No Finding" (F1 = 0.82, n=452), "Consolidation" (F1 = 0.89, n=215), and "Pleural Effusion" (F1 = 0.94, n=185).

When incorporating status annotations, performance declined slightly due to the number of labels effectively being tripled. The "leaves with statuses" model yielded an F1 score of 0.794, while the "upper with statuses" model achieved an F1 score of 0.795. In both cases, "No Finding" remained a strong performer (F1 = 0.82), while disease-specific labels such as "Simple Pleural Effusion (Present)" (F1 = 0.91, n=96) and "Cardiomegaly (Present)" (F1 = 0.98, n=82) performed very well. However, some uncertain findings, such as "Consolidation (Uncertain)" (F1 = 0.82, n=95), demonstrated slightly lower scores, reflecting the intrinsic difficulty of differentiating between ambiguous disease states.

4.1.1 Comparison to CheXbert

We compare our models to CheXbert as they both aim to accomplish the same task of disease

	Precision	Recall	F1-Score	Support
		Leaves		
Micro Avg	**0.85**	0.82	**0.84**	1,644
Macro Avg	0.63	0.53	0.55	1,644
Weighted Avg	**0.85**	0.82	0.82	1,644
Samples Avg	0.84	**0.84**	**0.84**	1,644
		Upper		
Micro Avg	0.85	0.83	**0.84**	1,588
Macro Avg	0.70	0.62	0.65	1,588
Weighted Avg	**0.87**	0.83	0.83	1,588
Samples Avg	0.85	**0.84**	**0.84**	1,588
		Leaves with Statuses		
Micro Avg	**0.81**	0.78	**0.80**	1,644
Macro Avg	0.31	0.27	0.28	1,644
Weighted Avg	0.79	0.78	0.77	1,644
Samples Avg	0.80	**0.80**	0.79	1,644
		Upper with Statuses		
Micro Avg	**0.81**	0.79	**0.80**	1,574
Macro Avg	0.41	0.38	0.38	1,574
Weighted Avg	0.79	0.79	0.78	1,574
Samples Avg	0.80	**0.80**	**0.80**	1,574

Table 3: Benchmark results for disease classification on the test_reviewed split. Highest scores are in bold.

classification. Given the more restricting label set of CheXbert, we first filter the reviewed test set to only include utterances with a label that is mappable to CheXbert classes. This mapping between label spaces was conducted after consulting a combination of web sources, a clinician, and GPT-4o. However, some degree of overlap and ambiguity remains (Section 7).

Using structured utterances as input, we first derive CheXbert labels using the author-provided CheXbert model checkpoint. Using SRR-BERT, we then compute labels at both the "leaves" level and the "upper" level and map them to the 14 classes used by CheXbert. Table 4 illustrates the direct comparison of model performances, where SRR-BERT outperformed CheXbert in both settings (0.80 vs. 0.61 when "leaves" were used for the mapping, and 0.83 vs. 0.47 when "upper" labels were used for the mapping).

We acknowledge that SRR-BERT was trained on structured utterances while CheXbert was not, which may skew the comparison. Hence, we also leverage the unstructured full-length reports as input. SRR-BERT outperforms CheXbert when using "upper" labels to map to CheXbert classes, and exhibits only slightly lower F1 when using "leaves". This demonstrates the robustness of SRR-BERT models as they can accommodate texts of varying lengths and complexity, from short utterances to full-length reports.

	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

Table 4: Weighted average performance comparison for CheXbert and SRR-BERT using “leaves” and “upper” mappings to 14 CheXbert classes on the test_reviewed split. Highest scores are in bold.

4.2 Structured RRG

4.2.1 Evaluation Metrics

To ensure consistency with prior work in "traditional" RRG, we report BLEU (Papineni et al., 2002), BERTScore (Zhang et al., 2019), ROUGE-L (Lin, 2004), and F1-RadGraph (Delbrouck et al., 2022). Additionally, we introduce F1-SRR-BERT, a new metric leveraging our SRR-BERT model (Section 3), which is trained to predict abnormalities across 55 diseases based on CXR utterances.

F1-SRR-BERT measures the F1-Score between SRR-BERT’s predictions on the generated structured report and the corresponding reference structured reports. This score has two variants: (1) *leaves prediction*, which classifies diseases at the finest granularity (55 labels from the disease tree in Prompt 4), and (2) *upper-level prediction*, which groups diseases into 25 broader categories for a coarser classification. These broader categories are the level right above the "leaves".

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

Table 5: Model scores on different splits of our **SRRG-impression** dataset. Traditional metrics (BLEU, ROUGE-L, BERTScore, F1-RadGraph) are shown as percentages. F1-SRR-BERT scores (weighted averages for utterance-level diseases Precision, Recall, and F1-Score). Bold indicates the best score per model group on the Test vs. Test Reviewed splits.

An additional consideration in our evaluation is that utterances can be assessed in either an *aligned* or *unaligned* setting across all previously mentioned metrics. In the *aligned* setting, utterances are evaluated in the order they appear under an organ system header or by their numerical order in the impression section (i.e., generated impression one is compared to reference impression one). In contrast, the *unaligned* setting evaluates utterances as a set—comparing all findings under an organ system or all numbered impressions as a block against the reference. This unaligned approach allows us to assess whether the model prioritizes findings and impressions from the most to the least clinically relevant. Finally, we assign a score of 0 for missing references sections and extra predicted sections in findings.

4.2.2 Results

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). These models vary in size, architecture, and reported performance.

Impression. Table 5 shows the performance of various models in generating impressions (evaluated without alignment), revealing that models tend to score higher in this task than in free-form impression generation. Notably, CheXpert-Plus stands out as the best performer on the SRRG-Impression dataset. On the test split, it achieves the highest traditional metric scores, with a BLEU of 14.84, ROUGE-L of 28.01, and F1-RadGraph of 22.14, while also registering the highest utterance-level precision at 58.99. Although CheXagent and MAIRA-2 excel in BERTScore and Recall respectively, CheXpert-Plus consistently delivers superior performance across both traditional and SRRG metrics.

Split	**BLEU**	**ROUGE-L**	**BERTScore**	**F1-RadGraph**
		SRRG-Impression		
Validate	7.61 ↓9.25	23.35 ↓10.07	39.95 ↓22.79	16.68 ↓11.06
Test	3.78 ↓11.06	16.77 ↓11.24	36.35 ↓24.41	10.23 ↓11.91
Test Reviewed	3.63 ↓**10.44**	16.89 ↓**9.90**	38.82 ↓**20.39**	10.42 ↓**8.47**
		SRRG-Findings		
Validate	3.77 ↓0.35	19.23 ↓1.67	26.81 ↓4.77	14.23 ↓2.72
Test	3.21 ↓**0.30**	16.89 ↓**2.08**	25.83 ↓**5.67**	12.31 ↓**2.68**
Test Reviewed	3.45 ↓0.51	16.27 ↓2.45	24.93 ↓6.40	11.68 ↓3.21

Table 6: Updated scores for the CheXpert-Plus model using the "**aligned**" settings. The differences from the unaligned settings (Tables 5 and 7) are shown in red. For each section, the smaller drop between the Test and Test Reviewed splits is highlighted in bold.

Findings. In the SRRG-Findings (unaligned) setting (Table 7), traditional metric scores are generally lower than in the SRRG-Impression setting, indicating that generating structured findings is more challenging than producing impressions.

SRRG-Findings (unaligned)		Traditional Metrics				F1-SRR					
						F1-SRR-BERT			Category		
Model	**Split**	**BLEU**	**ROUGE-L**	**BERTScore**	**F1-RadGraph**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
CheXagent	Validate	1.93	19.72	29.58	15.35	42.86	44.04	41.88	75.98	77.16	74.70
CheXagent	Test	1.80	19.65	31.65	15.41	43.22	42.07	41.13	**77.12**	82.56	**77.90**
CheXagent	Test Reviewed	**2.38**	**19.88**	**32.48**	**16.04**	**44.56**	**42.53**	**41.73**	75.26	**85.22**	77.40
CheXpert-Plus	Validate	4.12	20.90	31.58	16.95	44.28	43.19	42.08	72.10	85.45	76.52
CheXpert-Plus	Test	3.51	**18.97**	**31.50**	**14.99**	**42.79**	**40.08**	**39.85**	**72.84**	86.17	77.18
CheXpert-Plus	Test Reviewed	**3.96**	18.72	31.33	14.89	42.78	39.10	39.28	71.63	**88.71**	**77.24**
MAIRA-2	Validate	6.32	29.00	39.38	25.62	49.66	49.66	49.66	78.21	86.24	80.52
MAIRA-2	Test	**3.39**	**23.15**	**35.44**	**19.03**	**43.65**	**43.65**	**43.65**	**75.64**	86.23	**79.03**
MAIRA-2	Test Reviewed	2.26	20.55	32.87	16.90	42.36	42.36	42.36	72.25	**88.90**	77.79
RaDialog	Validate	1.47	18.23	28.67	13.92	40.15	39.63	39.08	70.12	70.48	69.33
RaDialog	Test	1.28	17.53	**29.07**	**13.82**	38.42	38.10	37.89	69.48	70.12	**69.76**
RaDialog	Test Reviewed	**1.42**	**17.60**	28.90	13.75	**38.95**	**38.30**	**38.05**	**69.90**	**70.22**	69.85

Table 7: Model scores on different splits of our **SRRG-Findings** dataset. Traditional Metrics include BLEU, ROUGE-L, BERTScore, and F1-RadGraph. F1-SRR-BERT metrics (weighted averages) are evaluated for Diseases and for Category (organ section headers). Bold indicates the best score per model group on the Test vs. Test Reviewed splits.

For findings, CheXpert-Plus achieves moderate scores on validation (e.g., BLEU 4.12, ROUGE-L 20.90, BERTScore 31.58, F1-RadGraph 16.95), while CheXagent and MAIRA-2 show similar patterns with slight drops from validation to test splits. Category scores—reflecting the correct prediction of organ section headers—are consistently high (around 75–78%) across models. In contrast, the impression results reveal substantially higher traditional metrics, with CheXagent and CheXpert-Plus achieving BLEU scores above 14 and BERTScores in the low 60s, suggesting that the impression task yields more polished, concise outputs. Overall, these results highlight that while all models struggle with the detailed nature of findings, they perform significantly better when generating shorter, impression-style summaries.

Alignment. As expected, generating impressions and findings that align with the ground-truth is challenging, as demonstrated by CheXpert-Plus' scores (Table 6). This challenge is even more pronounced in the impression setting, which typically contains more utterances than organ sections.

Table 8 reveals marked heterogeneity in CheXpert-Plus's organ-specific performance. The model is most reliable for cardiovascular structures, with an F1 score of roughly 60, and for hardware-related findings ("Tubes, Catheters, and Support Devices"), where the score is about 51; pleural and musculoskeletal regions follow, each in the mid-40s. Performance drops substantially for lung parenchyma and airways, which score around 38, and is weakest for abdominal findings (about 11) and the miscellaneous "Other" category (around 3). These disparities suggest that CheXpert-Plus excels when imaging cues are distinct or well-represented in the training data, but struggles with rarer or more heterogeneous organ systems.

Organ	**Precision**	**Recall**	**F1-Score**
Pleura	54.53	40.28	44.23
Abdominal	10.53	10.53	10.53
Hila and Mediastinum	22.26	21.58	21.69
Other	3.69	3.42	3.39
Lungs and Airways	41.85	40.41	38.32
Cardiovascular	63.78	58.73	59.78
Musculoskeletal and Chest Wall	45.99	43.91	44.29
Tubes, Catheters, and Support Devices	51.27	54.94	50.56

Table 8: Organ-level F1-SRRG-BERT weighted-average scores for CheXpert-Plus on the test-reviewed split.

OOD. Finally, we perform an out-of-distribution (OOD) evaluation using the HOPPR test set, which consists of 1,300 samples sourced from the HOPPR Platform. These samples come from two imaging providers across eight U.S. states. Each report in the set contains at least one confirmed positive finding, including conditions such as Acute Rib Fracture, Air Space Opacity, Cardiomegaly, Lung Nodule or Mass, Non-Acute Rib Fracture, Pleural Fluid, Pneumothorax, or Pulmonary Artery Enlargement. When tested on this new, out-of-distribution dataset, all three public models exhibit

SRRG-Findings (unaligned)		Traditional Metrics				F1-SRR					
						F1-SRR-BERT			Category		
Model	Split	BLEU	ROUGE-L	BERTScore	F1-RadGraph	Precision	Recall	F1-Score	Precision	Recall	F1-Score
CheXagent	Test (OOD)	3.90 ↑2.10	16.50 ↓3.15	28.10 ↓3.55	13.70 ↓1.71	42.70 ↓0.52	44.10 ↑2.03	43.38 ↑2.25	77.70 ↑0.58	87.50 ↑4.94	82.30 ↑4.40
CheXpert-Plus	Test (OOD)	6.10 ↑2.59	15.84 ↓3.13	28.00 ↓3.50	13.31 ↓1.68	42.28 ↓0.51	42.28 ↑2.20	42.28 ↑2.43	73.50 ↑0.66	91.56 ↑5.39	80.91 ↑3.73
MAIRA-2	Test (OOD)	5.90 ↑2.51	20.00 ↓3.15	31.90 ↓3.54	17.30 ↓1.73	43.10 ↓0.55	45.90 ↑2.25	44.45 ↑0.80	76.20 ↑0.56	91.80 ↑5.57	82.80 ↑3.77

SRRG-Impression (unaligned)		Traditional Metrics				F1-SRR-BERT		
Model	Split	BLEU	ROUGE-L	BERTScore	F1-RadGraph	Precision	Recall	F1-Score
CheXagent	Test (OOD)	3.00 ↓3.95	13.50 ↓13.68	46.00 ↓15.51	4.50 ↓15.20	30.50 ↓19.28	40.00 ↓16.47	33.00 ↓17.63
CheXpert-Plus	Test (OOD)	7.00 ↓7.84	14.78 ↓13.23	45.73 ↓15.03	5.25 ↓16.89	30.11 ↓18.63	44.59 ↓3.01	33.15 ↓13.33
MAIRA-2	Test (OOD)	3.50 ↓4.62	14.50 ↓13.32	47.00 ↓15.30	4.80 ↓15.57	29.50 ↓19.22	42.00 ↓15.91	32.50 ↓17.86

Table 9: CheXpert-Plus, CheXagent, and MAIRA-2 performance on the out-of-distribution HOPPR test set, showing deltas relative to their original Test results from Tables 5 and 7.

a typical domain shift: their lexical metrics—such as BLEU, ROUGE-L, and BERTScore (drop by 3 to 15 points). However, structure-aware metrics remain much more stable. RadGraph F1 decreases by only about 1.5 points, and interestingly, disease-level F1 using SRR-BERT for the Findings section actually increases by 0.8 to 2.4 points. Performance on organ-category labels also improves, rising by 3 to 4 points. The main weakness lies in generating the Impression section, where models lose between 13 and 18 points.

5 Conclusion

We presented Structured Radiology Report Generation (SRRG), a new task reformulating free-text CXR reports into standardized templates to improve clarity and enable more precise evaluation. To support SRRG, we introduce a large-scale dataset with clinically validated structured reports and SRR-BERT, a 55-label disease classifier trained on fine-grained radiological findings. We further propose F1-SRR-BERT, a metric leveraging SRR-BERT's hierarchical labels to capture clinically meaningful variations. Our reader study, conducted by board-certified radiologists, confirms the quality of both the structured reports and annotated disease labels. Benchmark experiments show that SRRG improves consistency compared to existing free-form generation methods.

6 Related Work

Structured Reporting Chest X-ray reporting has long been characterized by a free-text narrative style, which while flexible, can lack clarity and consistency (Weiss and Langlotz, 2008; Bosmans et al., 2012). The lack of widespread standardization further reinforces this approach, as structured reporting templates, such as RSNA's RadLex or BI-RADS for breast imaging, have not been universally adopted for CXRs. Studies have shown that even though structured reporting can improve completeness and diagnostic clarity (Schwartz et al., 2011; Bosmans et al., 2012), many radiologists perceive it as rigid and less efficient compared to narrative reporting (Bosmans et al., 2015). Consequently, structured reporting remains underutilized, in part because CXRs require simultaneous assessment of multiple structures in context rather than in isolation (Langlotz, 2002).

Given these challenges, efforts to standardize CXR reporting continue to face resistance, balancing the need for consistency with the flexibility required for nuanced clinical communication (Dunnick and Langlotz, 2008; Kahn Jr et al., 2009). For systems aiming to generate automated or semi-automated reports from medical images, addressing this variability is crucial. Recent works in natural language processing and computer vision have attempted to handle the complexity of unstructured radiology reports, either by adopting standardized label sets derived from clinical knowledge bases or by using large-scale language models to learn patterns in free-text narratives. However, the gap between free-form clinical practice and structured data requirements remains a major challenge in achieving both clinical relevance and interoperability.

Automated Radiology Reporting Prior work in radiology report generation has explored architectural innovations, reinforcement learning, and retrieval-based approaches. Architectural novelties include memory-driven transformers to retain key generation details (Chen et al., 2020), cross-modal memory networks to align images and text (Chen et al., 2021), and models incorporating prior medical knowledge graphs for structured report generation (Liu et al., 2021a,b). Reinforcement learning has also been used to optimize factual correctness (Liu et al., 2019; Miura et al., 2021; Delbrouck et al., 2022). Recently, larger models have been employed for radiology report generation. Notable examples include RaDialog (Pellegrini et al., 2023), which integrates visual features and structured pathology findings with an LLM through parameter-efficient fine-tuning, and RGRG (Tanida et al., 2023), a region-guided model that detects and describes anatomical regions to enhance transparency, interactivity, and explainability. Additionally, "LLM-sized" models such as MAIRA-2 (Bannur et al., 2024), CheXagent (Chen et al., 2024), and MedVersa (Zhou et al., 2024) have also been introduced to further advance the field.

7 Limitations

Despite the promising results of our proposed Structured Radiology Report Generation (SRRG) framework, several limitations remain:

Synthetic Dataset & Annotations Our SRRG dataset was produced by reformulating free-form radiology reports into a structured format using LLMs. Although our methodology enforces strict desiderata to avoid hallucinations and preserve factual content, it remains challenging to verify all generated samples at scale. To mitigate inaccuracies, we conducted a comprehensive reader study involving five board-certified radiologists, as described in Appendix B. Nevertheless, the possibility of subtle inconsistencies or biases introduced by the LLMs cannot be fully excluded.

Fine-tuning Approaches The range of model sizes and different training strategies used in our experiments (e.g., LoRA-based parameter-efficient fine-tuning for large models such as MAIRA-2 vs. full fine-tuning for smaller models) may affect the comparability of results. While these choices were made to accommodate computational feasibility, a standardized fine-tuning scheme across all models might yield a more uniform assessment of performance and could be explored in future work.

Reader Study Constraints Our reader study focused on validating both structured reports and fine-grained disease labels derived from the SRR-BERT model. Although board-certified radiologists reviewed a representative sample of utterances, they occasionally encountered ambiguous cases where the available clinical context did not suffice to differentiate among closely related conditions (e.g., pneumonia, atelectasis, or aspiration). Additionally, rare findings not covered by our disease taxonomy were annotated under an *"Other"* category, potentially oversimplifying certain nuanced clinical observations. Expanding the taxonomy or incorporating additional clinical context (e.g., lab values or clinical notes) may address these ambiguities in future iterations.

F1-SRR-BERT vs. F1-CheXbert Directly comparing F1-Scores of SRR-BERT (with 55 disease labels) and CheXbert (with 14 labels) remains inherently imperfect due to the many-to-many relationship in label mapping. A single CheXbert class can correspond to multiple labels in our hierarchical disease ontology, and vice versa. Although we attempted a best-effort alignment, the lack of a one-to-one mapping between the label spaces makes straightforward performance comparisons challenging. Future work could improve this alignment by exploring probabilistic approaches or expert-guided hierarchical restructuring to reconcile label disparities.

8 Contributions

JBD led the project, curated the datasets, supervised the experimental workflow, and was the primary author of the manuscript. JX conducted the experiments related to the Disease Classification Models. JM, AT, and ZC were responsible for fine-tuning the models on the SRRG dataset. SO and AA contributed through early-stage brainstorming. KZI, AJ, ER, CB and MM served as reviewing radiologists, providing expert evaluation of the results. MV and CL offered guidance during the project.

References

Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. 2024. Maira-2: Grounded radiology report generation. *arXiv preprint arXiv:2406.04449*.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C. Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, Hoifung Poon, and Ozan Oktay. 2022. *Making the Most of Text Semantics to Improve Biomedical Vision–Language Processing*, page 1–21. Springer Nature Switzerland.

Jan ML Bosmans, Emanuele Neri, Osman Ratib, and Charles E Kahn Jr. 2015. Structured reporting: a fusion reactor hungry for fuel. *Insights into Imaging*, 6:129–132.

JML Bosmans, Lieve Peremans, Maurizio Menni, AM De Schepper, PO Duyck, and PM Parizel. 2012. Structured reporting: if, why, when, how—and at what expense? results of a focus group meeting of radiology professionals from eight countries. *Insights into Imaging*, 3:295–302.

Pierre Chambon, Jean-Benoit Delbrouck, Thomas Sounack, Shih-Cheng Huang, Zhihong Chen, Maya Varma, Steven QH Truong, Curtis P Langlotz, et al. 2024. Chexpert plus: Hundreds of thousands of aligned radiology texts, images and patients. *arXiv e-prints*, pages arXiv–2405.

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 (Volume 1: Long Papers)*, pages 5904–5914, Online. 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.

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, et al. 2024. Chexagent: Towards a foundation model for chest x-ray interpretation. *arXiv preprint arXiv:2401.12208*.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

N Reed Dunnick and Curtis P Langlotz. 2008. The radiology report of the future: a summary of the 2007 intersociety conference. *Journal of the American College of Radiology*, 5(5):626–629.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856.

Curtis P Langlotz. 2002. Automatic structuring of radiology reports: harbinger of a second information revolution in radiology. *Radiology*, 224(1):5–7.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Fenglin Liu, Xian Wu, Shen Ge, Wei Fan, and Yuexian Zou. 2021a. Exploring and distilling posterior and prior knowledge for radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13753–13762.

Fenglin Liu, Chenyu You, Xian Wu, Shen Ge, Xu Sun, et al. 2021b. Auto-encoding knowledge graph for unsupervised medical report generation. *Advances in Neural Information Processing Systems*, 34:16266–16279.

Guanxiong Liu, Tzu-Ming Harry Hsu, Matthew McDermott, Willie Boag, Wei-Hung Weng, Peter Szolovits, and Marzyeh Ghassemi. 2019. Clinically accurate chest x-ray report generation. In *Proceedings of the 4th Machine Learning for Healthcare Conference*, volume 106 of *Proceedings of Machine Learning Research*, pages 249–269. PMLR.

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.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson Md, Michael Moseley, Curtis Langlotz, Akshay S Chaudhari, and Jean-Benoit Delbrouck. 2024. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 374–390, Miami, Florida, USA. Association for Computational Linguistics.

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.

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. *arXiv preprint arXiv:2311.18681*.

Xiao Pu, Mingqi Gao, and Xiaojun Wan. 2023. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*.

Lawrence H Schwartz, David M Panicek, Alexandra R Berk, Yuelin Li, and Hedvig Hricak. 2011. Improving communication of diagnostic radiology findings through structured reporting. *Radiology*, 260(1):174–181.

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*, pages 7433–7442.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

David L Weiss and Curtis P Langlotz. 2008. Structured reporting: patient care enhancement or productivity nightmare? *Radiology*, 249(3):739–747.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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

A Potential Risks

All experiments in this study are conducted using publicly available chest X-ray datasets (MIMIC-CXR and CheXpert Plus) that are fully deidentified, thereby minimizing risks related to patient privacy and data confidentiality. The text restructuring and disease label generation steps use GPT-4 deployed via Azure services, with the account explicitly configured to opt out of human data review.

While we believe releasing our models and code is valuable for advancing research, we emphasize that these models are for investigational and educational purposes only. They have not received regulatory approval for clinical deployment, and medical professionals must retain ultimate responsibility for diagnosis and patient management. As with all machine learning models, there is an inherent risk of errors or hallucinations, and predictions should be verified by qualified clinicians. We strongly encourage the community to apply robust validation, audits, and clinical oversight when exploring or extending our work.

B Reader Study

The reader study has been carried out by five board-certified radiologists from our institution on the annotation platform detailed in Appendix E. The following examples and statistics summarize the textual changes between the original and edited impression sections. For each report pair, differences were quantified by counting word-level insertions, deletions, and replacements. The **similarity ratio** was computed using Python's `difflib.SequenceMatcher` via

$$\text{Similarity Ratio} = \frac{2 \times \text{Matches}}{\text{Total Tokens in Original and Edited}}$$

yielding a value between 0 (completely different) and 1 (identical).

Example 1: `mimic-53235571`

```
Original Impression:
1. Bibasilar opacities that may be
    related to atelectasis, with a
    differential
   including underlying infection,
    pneumonia, or aspiration.
2. New opacity in the lateral left mid
    lung, nonspecific but potentially
   representing additional consolidation
     or pulmonary infarct.

Edited Impression:
1. Bibasilar opacities may be related to
     atelectasis, although underlying
   infection, pneumonia, and/or
    aspiration is of concern.
2. New opacity in the lateral left mid
    lung, nonspecific but potentially
   representing additional consolidation
     or pulmonary infarct.

Diff Stats:
Insertions: 0, Deletions: 1,
    Replacements: 9, Similarity Ratio:
    0.82
```

Example 2: `mimic-59654440`

```
Original Impression:
1. Resolving consolidation at the right
    lung base, likely due to dependent
   edema or combined dependent edema and
     atelectasis.
2. Mild to moderate enlargement of the
    heart.
3. No pneumothorax.
4. Dual-channel dialysis catheter in
    situ with the tip in the right
    atrium.

Edited Impression:
1. Resolving consolidation at the right
    lung base with minimal residual
   interstitial edema.

Diff Stats:
Insertions: 0, Deletions: 0,
    Replacements: 35, Similarity Ratio:
    0.29
```

Impression Statistics

```
Total studies reviewed: 233
Studies with changes: 130 (55.79%)
Average insertions per study: 0.42
Average deletions per study: 4.16
Average replacements per study: 4.50
Average similarity ratio: 0.77
```

Although 55.79% of the impression exhibited changes, many modifications are subtly reflected by a relatively high overall similarity ratio. However, some reports demonstrate significant edits, underlining the need for enhanced clarity and precise clinical communication in the impression sections of CXR reports.

Findings Statistics

```
Total studies reviewed: 233
Studies with changes: 164 (70.39%)
Average insertions per study: 4.97
Average deletions per study: 3.46
Average replacements per study: 4.64
Average similarity ratio: 0.88
```

The analysis reveals that a significant portion of the studies (70.39%) underwent modifications, indicating that changes were applied in the majority of the cases. However, the higher average similarity ratio of 0.88 may suggest that these edits are relatively minor. On average, the modifications involved about 4.97 insertions, 3.46 deletions, and 4.64 replacements per study, which implies that while the impression sections were updated, the overall content remains largely consistent with the original. This balance indicates that the editing process likely focused on refining clarity and precision without altering the fundamental diagnostic information conveyed in the reports.

Utterance Label Consistency

In this experiment, we assess the consistency of utterance labels extracted from the GPT models and compare them with manually reviewed labels. Two metrics are computed:

1. **Exact match** GPT's labels and reviewed labels are the same.
2. **Jaccard Similarity:** The ratio of the size of the intersection to the size of the union of the GPT's and reviewed label sets.

The overall statistics from the evaluation are as follows:

```
Total utterances reviewed: 1609
Matched utterances: 1339
Exact Match Rate: 0.72
Average Jaccard Similarity: 0.74
```

These results indicate that, on average, 72% of the consensus labels are present in the reviewed labels, and there is a 74% overlap between the two label sets. The high similarity metrics suggest that the consensus approach is effective for capturing the expected labels across different sources, thereby validating our methodology for robust label extraction in utterances.

C Model Sizes and Hyperparameters

MAIRA-2 uses an 87M-parameter ViT model, with its language model initialized from Vicuna 7B v1.5. We evaluated the 3B version of CheXagent-2. CheXpert-Plus is a SwinV2-based model with a BERT decoder (2 layers), while RaDialoG is a 7B-parameter model. For fine-tuning SRRG, we trained all the weights of CheXpert-Plus and CheXagent, using the default LoRA parameters from the Hugging Face PEFT library.

D Dataset Breakdown of Diseases

Table 10: Dataset Breakdown for Upper Labels

Anatomical Header / Category	Upper Levels	Num. Examples
Lungs and Airways	Consolidation	340,867
	Diffuse air space opacity	100,154
	Lung Finding	95,122
	Air space opacity	47,921
	Solitary masslike opacity	40,831
	Focal air space opacity	14,222
	Segmental collapse	10,685
	Multiple masslike opacities	547
	Total	650,349
Pleura	Pleural Effusion	173,883
	Pneumothorax	56,706
	Pleural Thickening	31,210
	Pleural finding	7,734
	Total	269,533
Cardiovascular	Widened cardiac silhouette	58,189
	Vascular finding	20,480
	Total	78,669
Hila and Mediastinum	Widened aortic contour	17,513
	Mediastinal finding	13,779
	Mediastinal mass	5,922
	Total	37,214
Musculoskeletal and Chest Wall	Fracture	34,192
	Chest wall finding	11,614
	Musculoskeletal finding	617
	Total	46,423
Abdominal	Subdiaphragmatic gas	3,475
Support Devices	Support Devices	96,274
No Finding	–	600,328

Structuring Prompt

Your task is to improve the formatting of a radiology report, ensuring it is **clear, concise, and well-structured** with appropriate section headings.
Guidelines:

1. **Section Headers:** Each section should begin with a section header followed by a colon. Include only the relevant information as specified.
2. **Identifiers:** Remove any sentences containing identifiers such as dates, surnames, first names, healthcare providers, vendors, or institutions. **Important:** Retain sex and age information if present.
3. **Findings and Impression Sections:** Focus exclusively on the **current examination results**. Do not reference previous studies or historical data.
4. **Content Restrictions:** Strictly include only content relevant to the structured sections provided. Do not add or extrapolate beyond the original report.

Sections to Include (if applicable):

1. **Exam Type:** Specify the type of examination conducted.
2. **History:** Provide a brief clinical history and state the clinical question or suspicion prompting the imaging.
3. **Technique:** Describe the examination technique and any specific protocols used.
4. **Comparison:** Indicate prior imaging studies reviewed for comparison.
5. **Findings:** List all positive and relevant negative observations for each organ system under structured headers.

Template for Findings:

```
Header 1:
- Observation 1
- ...
Header 2:
- Observation 1
- Observation 2
- ...
...
```

Use only the following headers for organ systems:

- Lungs and Airways
- Pleura
- Cardiovascular
- Hila and Mediastinum
- Tubes, Catheters, and Support Devices
- Musculoskeletal and Chest Wall
- Abdominal
- Other

Important: *Do not use any headers other than those listed above. Only use the specified headers corresponding to the organ systems mentioned in the original radiology report.*
6. Impression: Summarize the key findings in a numbered list, ranking them from most to least clinically relevant.
The radiology report to improve is the following:

```
{}
```

Diseases prompt

Your task is to identify the diseases discussed in chest X-ray findings. You will be provided with:
1) Instructions
2) A list of possible diseases
3) A list of chest X-ray findings

1) Instructions: Your task is to provide the following:

a) The diseases that are present as a numbered list. There can be zero, one, or multiple diseases discussed. If no disease is present or discussed in a finding, answer: `"1. No Finding"` for that finding.

b) The status of the disease discussed. The status can be:

   - **Present**: The disease is confirmed to be present in the patient.
   - **Absent**: The disease is confirmed to be not present in the patient.
   - **Uncertain**: It is unclear whether the disease is present or absent, often due to inconclusive test results or insufficient information.

Below is the template to provide your answer. You must respect this format and not provide any explanations or additional content:

```
<finding 1> => 1. <disease 1> (Present) 2. <disease 2> (Uncertain)
<finding 2> => 1. <disease 1> (Absent)
...
```

2) List of possible diseases:

- No Finding
- Lung Lesion
- Edema
- Pneumonia
- Atelectasis
- Lung collapse
- Perihilar airspace opacity
- Air space opacity–multifocal
- Mass/Solitary lung mass
- Nodule/Solitary lung nodule
- Cavitating mass with content
- Cavitating masses
- Emphysema

...

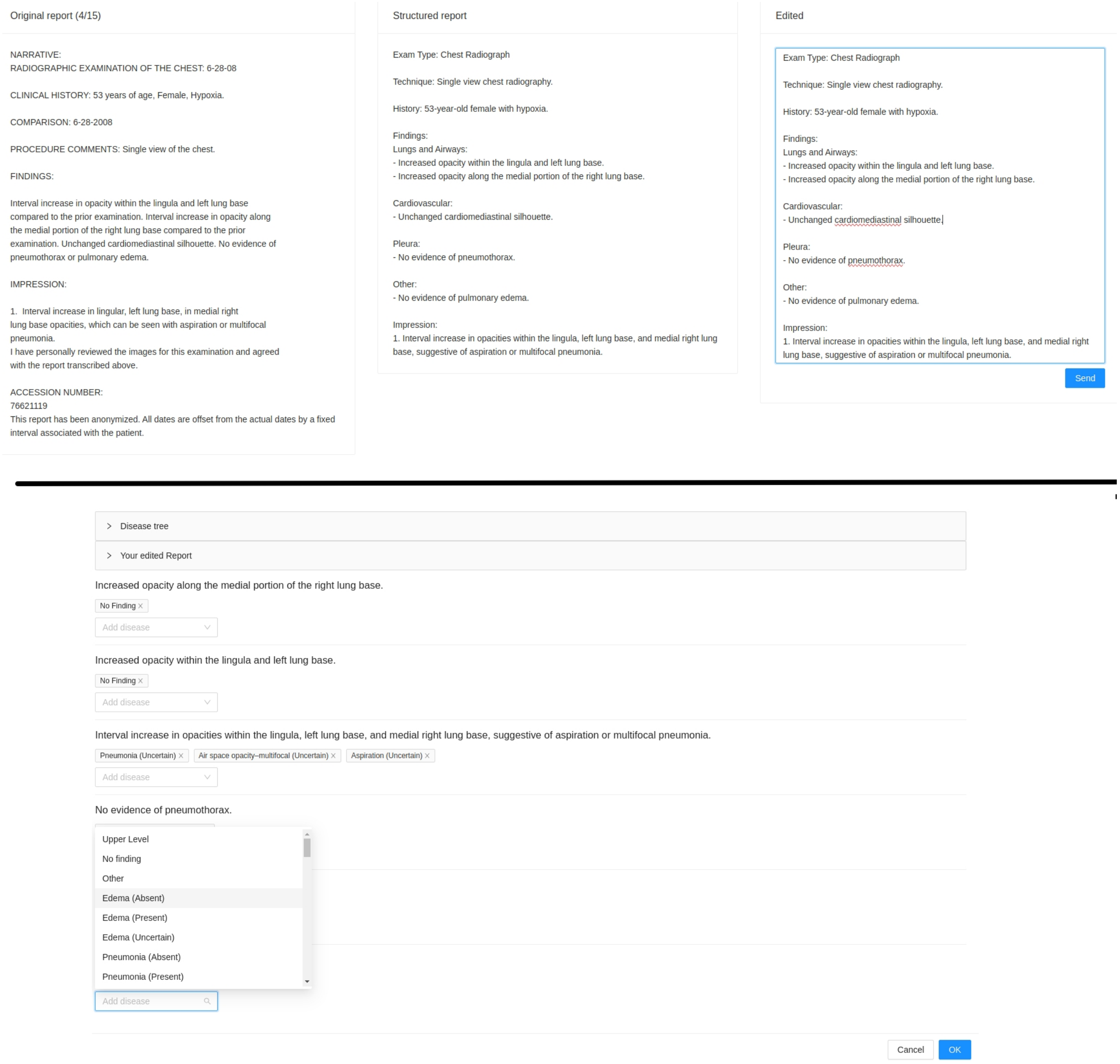
3) List of chest X-ray findings (one per line):

```
{}
```

Diseases Tree

```
1. No Finding
2. Lung Finding
   2.1. Lung Opacity
       2.1.1. Air space opacity
           2.1.1.1. Diffuse air space opacity
               2.1.1.1.1. Edema
           2.1.1.2. Focal air space opacity
               2.1.1.2.1. Consolidation
                   2.1.1.2.1.1. Pneumonia
                   2.1.1.2.1.2. Atelectasis
                   2.1.1.2.1.3. Aspiration
               2.1.1.2.2. Segmental collapse
                   2.1.1.2.2.1. Lung collapse
               2.1.1.2.3. Perihilar airspace opacity
           2.1.1.3. Air space opacity–multifocal
       2.1.2. Masslike opacity
           2.1.2.1. Solitary masslike opacity
               2.1.2.1.1. Mass/Solitary lung mass
               2.1.2.1.2. Nodule/Solitary lung nodule
               2.1.2.1.3. Cavitating mass with content
           2.1.2.2. Multiple masslike opacities
               2.1.2.2.1. Cavitating masses
   2.2. Emphysema
   2.3. Fibrosis
   2.4. Pulmonary congestion
   2.5. Hilar lymphadenopathy
   2.6. Bronchiectasis
3. Pleural Finding
   3.1. Pneumothorax
       3.1.1. Simple pneumothorax
       3.1.2. Loculated pneumothorax
       3.1.3. Tension pneumothorax
   3.2. Pleural Thickening
       3.2.1. Pleural Effusion
           3.2.1.1. Simple pleural effusion
           3.2.1.2. Loculated pleural effusion
       3.2.2. Pleural scarring
   3.3. Hydropneumothorax
   3.4. Pleural Other
4. Widened Cardiac Silhouette
   4.1. Cardiomegaly
   4.2. Pericardial effusion
5. Mediastinal Finding
   5.1. Mediastinal Mass
       5.1.1. Inferior mediastinal mass
       5.1.2. Superior mediastinal mass
   5.2. Vascular Finding
       5.2.1. Widened aortic contour
           5.2.1.1. Tortuous Aorta
       5.2.2. Calcification of the Aorta
       5.2.3. Enlarged pulmonary artery
   5.3. Hernia
   5.4. Pneumomediastinum
   5.5. Tracheal deviation
6. Musculoskeletal Finding
   6.1. Fracture
       6.1.1. Acute humerus fracture
       6.1.2. Acute rib fracture
       6.1.3. Acute clavicle fracture
       6.1.4. Acute scapula fracture
       6.1.5. Compression fracture
   6.2. Shoulder dislocation
   6.3. Chest wall finding
       6.3.1. Subcutaneous Emphysema
7. Support Devices
   7.1. Suboptimal central line
   7.2. Suboptimal endotracheal tube
   7.3. Suboptimal nasogastric tube
   7.4. Suboptimal pulmonary arterial catheter
   7.5. Pleural tube
   7.6. PICC line
   7.7. Port catheter
   7.8. Pacemaker
   7.9. Implantable defibrillator
   7.10. LVAD
   7.11. Intraaortic balloon pump
8. Upper Abdominal Finding
   8.1. Subdiaphragmatic gas
       8.1.1. Pneumoperitoneum
```

E Reader Study Platform



NARRATIVE:
RADIOGRAPHIC EXAMINATION OF THE CHEST: 6-28-08

CLINICAL HISTORY: 53 years of age, Female, Hypoxia.

COMPARISON: 6-28-2008

PROCEDURE COMMENTS: Single view of the chest.

FINDINGS:

Interval increase in opacity within the lingula and left lung base compared to the prior examination. Interval increase in opacity along the medial portion of the right lung base compared to the prior examination. Unchanged cardiomediastinal silhouette. No evidence of pneumothorax or pulmonary edema.

IMPRESSION:

1. Interval increase in lingular, left lung base, in medial right lung base opacities, which can be seen with aspiration or multifocal pneumonia.
I have personally reviewed the images for this examination and agreed with the report transcribed above.

ACCESSION NUMBER:
76621119
This report has been anonymized. All dates are offset from the actual dates by a fixed interval associated with the patient.

Exam Type: Chest Radiograph

Technique: Single view chest radiography.

History: 53-year-old female with hypoxia.

Findings:
Lungs and Airways:
- Increased opacity within the lingula and left lung base.
- Increased opacity along the medial portion of the right lung base.

Cardiovascular:
- Unchanged cardiomediastinal silhouette.

Pleura:
- No evidence of pneumothorax.

Other:
- No evidence of pulmonary edema.

Impression:
1. Interval increase in opacities within the lingula, left lung base, and medial right lung base, suggestive of aspiration or multifocal pneumonia.

Exam Type: Chest Radiograph

Technique: Single view chest radiography.

History: 53-year-old female with hypoxia.

Findings:
Lungs and Airways:
- Increased opacity within the lingula and left lung base.
- Increased opacity along the medial portion of the right lung base.

Cardiovascular:
- Unchanged cardiomediastinal silhouette

Pleura:
- No evidence of pneumothorax.

Other:
- No evidence of pulmonary edema.

Impression:
1. Interval increase in opacities within the lingula, left lung base, and medial right lung base, suggestive of aspiration or multifocal pneumonia.

Increased opacity along the medial portion of the right lung base.

Increased opacity within the lingula and left lung base.

Interval increase in opacities within the lingula, left lung base, and medial right lung base, suggestive of aspiration or multifocal pneumonia.

No evidence of pneumothorax.

Figure 5: This figure illustrates our reader study annotation workflow. At the top, the radiologist sees the original report (left), the GPT-generated structured report (middle), and an editable text box (right). At the bottom, after validating the structured report, the radiologist annotates each utterance. The labels for these utterances are pre-filled based on the GPT model's consensus. Throughout this process, the radiologist can consult both the edited report and a disease tree to guide the labeling.