RadGraph-XL: A Large-Scale Expert-Annotated Dataset for Entity and Relation Extraction from Radiology Reports

Anonymous ACL submission

Abstract

In order to enable extraction of structured clinical data from unstructured radiology reports, we introduce RadGraph-XL, a large-scale, expert-annotated dataset for clinical entity and relation extraction. RadGraph-XL consists of 2,300 radiology reports, which are annotated with over 410,000 entities and relations by board-certified radiologists. Whereas previous approaches focus solely on chest X-rays, RadGraph-XL includes data from four anatomy-modality pairs - chest CT, abdomen/pelvis CT, brain MR, and chest X-rays. Then, in order to automate structured information extraction, we use RadGraph-XL to train transformer-based models for clinical entity and relation extraction. Our evaluations include comprehensive ablation studies as well as an expert reader study that evaluates trained models on out-of-domain data. Results demonstrate that our model surpasses the performance of previous methods by up to 52% and notably outperforms GPT-4 in this domain. We release RadGraph-XL as well as our trained model to foster further innovation and research in structured clinical information extraction.

1 Introduction

Radiology reports, which are critical for patient care, present a challenge for clinical research and applications due to their unstructured format and complex language. To address this, various methods have been developed to automatically extract important information from these reports (Langlotz and Meininger, 2000; Savova et al., 2010; Sugimoto et al., 2021). This is essential for tasks like training medical imaging models and monitoring diseases (Johnson et al., 2019; Irvin et al., 2019; Reis et al., 2022). However, the effectiveness of these methods is often limited by the specific types of information they are designed to extract and the scarcity of densely annotated datasets, which are costly to produce due to the need for expert input.

(a) Anatomy and Modality	
RadGraph-1.0	Chest X-ray
RadGraph-XL	**Chest CT, Abdomen/Pelvis CT, Brain MR, Chest X-ray**

(b) Annotation Complexity	Sample Length	Expert Knowledge
CoNLL04	29.0	✗
RadGraph-1.0	111.3	✓
RadGraph-XL	**409.8**	✓

(c) Dataset Scale	# Sentences	# Annotations
CoNLL04	1.4k	5.9k
RadGraph-1.0	3.7k	30.2k
RadGraph-XL	**68.7k**	**409.0k**

(d) Comparison w/ GPT-4 (Why so "old-school"?)	Entity F1	Relation F1
GPT-4 (0-shot)	0.158	0.012
GPT-4 (10-shot)	0.203	0.024
BERT (RadGraph-1.0)	0.744	0.453
BERT (RadGraph-XL)	**0.863**	**0.691**

Table 1: Illustrations of our motivation and contribution: (a) RadGraph-XL extends RadGraph-1.0 to other anatomies and modalities; (b) The long radiology reports and the requirements of expert knowledge pose a significant challenge to the annotation; (c) The scale of RadGraph-XL is much larger than existing general-domain and medical-domain datasets; (d) We show the performance comparisons with GPT-4 to demonstrate why we are so "old-school" and why we need RadGraph-XL in the large language model (LLM) era.

Recent initiatives have been directed towards addressing the complexities of deriving structured clinical data from the unstructured text of radiology reports. One such approach is RadGraph[1] (Jain et al., 2021), which comprises a dataset and schema that aim to capture a wide array of clinically relevant information, such as observation and anatomical details. This schema is designed to streamline and standardize the annotation process, thereby

[1]We denote it as RadGraph-1.0 in our paper.

Dataset	Anatomy and Modality	Report Level			Annotation Level	
		MIMIC	Hosp. A	$\bar{w}$	Σa	$\bar{a}$
Radgraph-1.0	Chest X-ray	550	50	111.3	30.2k	50.5
	Total	550	50	-	30.2k	-
Radgraph-XL (**Ours**)	Chest CT	100	500	502.1	115.7k	192.9
	Abdomen/Pelvis CT	100	500	576.8	169.8k	283.1
	Brain MR	100	500	352.3	95.2k	158.7
	Chest X-ray	-	500	167.7	28.5k	57.0
	Total	300	2000	-	409.0k	-

Table 2: We provide an overview of the RadGraph-XL annotations, encompassing 2,300 reports, in comparison to RadGraph-1.0 (Jain et al., 2021). Additionally, we highlight the total number of annotations Σa, as well as the average number of words $\bar{w}$ and annotations $\bar{a}$ **per report**, underscoring the significant expansion our annotations contribute to the existing RadGraph-1.0 dataset. This involves adding annotations to reports from both new types of imaging and body regions, as well as those originating from a different institution.

facilitating the extraction of meaningful insights from radiology narratives. However, its application has been primarily confined to chest X-ray reports, which limits its utility across the diverse spectrum of radiology documentation (as shown in Table 1(a)). In parallel, there is a burgeoning interest within the medical AI community to extend beyond chest X-rays, exploring a wider variety of imaging modalities and anatomical regions. This expansion is evident in recent advancements in report summarization (Chen et al., 2023; Delbrouck et al., 2023), report generation (Li et al., 2022; Zhang et al., 2023a), and the development of foundation models (Wu et al., 2023b; Tu et al., 2023). These advancements underscore the necessity for innovative methodologies capable of interpreting a broader range of radiology reports.

Additionally, Large Language Models (LLMs) have also been explored to extract information with various prompting strategies (Liu et al., 2023a), beginning with a single example and expanding up to 200-shot examples to maximize the GPT-4 model's context window. In addition to the impracticalities of the approach, including issues with access, costs, and inference time. Moreover, there is no certainty that LLMs will perform equally well across different modalities and anatomical studies, and this hypothesis remains untestable due to the lack of annotated data in these areas.

Motivated by these limitations, we introduce RadGraph-XL, a large-scale dataset featuring 2,300 radiology reports with approximately 410,000 expert annotations by radiologists (as shown in Table 1(b)(c)). These annotations cover a range of entities, relationships, and measurements across four different modality-anatomy pairs, aimed at significantly enhancing the precision and richness of data extracted from radiology texts. Leveraging our annotations, we train a transformer-based model tailored for the automatic annotation of radiology reports using proven frameworks for entity and relation extraction. Our evaluation encompasses a series of ablation studies and a reader study focused on out-of-domain data, providing a thorough assessment of the model's capabilities. Our model not only surpasses the performance benchmarks set by previous methodologies (up to 52%) but also demonstrates a significant edge over GPT-4's capabilities in this domain (as shown in Table 1(d)).

The structure of the paper is organized as follows. First, we introduce the RadGraph-XL dataset in Section 3.1 and discuss its differences from RadGraph-1.0. Next, we outline the process of annotating the dataset and the challenges encountered in Section 3.2, and present some key statistics in 3.3. Our focus then shifts to experiments in Section 4, where we elaborate on our model's training process (Section 4.1), our methodology for selecting the best-performing model (Section 4.2), and its comparison with a solution that employs a Large Language Model (LLM) as the transformer backbone (Section 4.3). Importantly, we highlight our model's performance on entities defined as measurements (Section 4.4)—a novel aspect of our annotations—and compare our model's performance with that of models from previous studies (Section 4.5). The experiments section concludes with a brief evaluation of GPT-4 against our reference

test set (Section 4.6). Following this, Section 5 presents the outcomes of our reader study, and the paper concludes with Section 6.

2 Related work

2.1 Extracting Information from Radiology Reports

In the field of chest x-rays, traditional automated radiology report labelers, used in datasets like MIMIC-CXR (Johnson et al., 2019) and CheXpert (Irvin et al., 2019), categorize reports for common medical conditions but miss finer details like specific entities and their relationships. More detailed approaches use entity extraction schemas (Bustos et al., 2020) and focus on facts and spatial relations (Datta et al., 2020a,b), but these require dense annotation by experts. The most advanced work intended to cover most clinically relevant information within the report on chest x-ray is RadGraph-1.0 as discussed in Section 1. New annotations or information extraction approaches are proposed on modalities and anatomies beyond chest x-rays, such as head CT (Jantscher et al., 2023) or chest CT (Lau et al., 2023), but remain coarse and scarce.

2.2 Downstream Tasks

Downstream tasks often leverage structured clinical data to enhance model performance. RadGraph-1.0 (Jain et al., 2021) annotations, for example, have been utilized to boost the quality of radiology report generation by using annotations as a form of reward (Delbrouck et al., 2022), as an indicator of style (Yan et al., 2023), or to eliminate hallucinated references (Ramesh et al., 2022). They are also used in pretraining (Zhang et al., 2023b; Wu et al., 2023a), to augment the performance of fine-grained image-text self-supervised models (Varma et al., 2023), and to assess the capabilities of Large Language Models (Liu et al., 2023a; Tu et al., 2023).

3 RadGraph-XL

3.1 Overview

RadGraph-XL aims to enhance the capabilities of RadGraph-1.0 (Jain et al., 2021) by expanding its application across different medical imaging modalities, anatomical regions, and healthcare institutions. The proposed extensions include:

- **New Modality:** Annotating Computed Tomography (CT) reports for the chest, moving beyond the initial focus on Chest X-ray reports.
- **New Anatomy:** Expanding the scope to include CT reports for the abdomen and pelvis, based on the experience with Chest CT reports.
- **New Modality and Anatomy:** To evaluate the model's performance on data that is significantly different from the training set, the proposal includes annotating Brain Magnetic Resonance (MR) imaging reports, which represents a new imaging modality and anatomical region.
- **New Institution:** Broadening the data source to include reports from a new institution, Hospital A (Hosp. A), in addition to the previously used MIMIC-CXR reports from RadGraph-1.0.

We select reports based on the following criteria in an effort to curate a diverse dataset, prioritized as follows: (i) We select reports with annotated disease labels and aim for a balanced selection to ensure an even distribution across different conditions, (ii) We employ unsupervised semantic clustering (Universal Sentence Encoder (USE) (Cer et al., 2018)) to group the reports and then select samples from each cluster, and (iii) we cluster the remaining reports by their length and sample from each cluster. A semantic projection of the USE embeddings using t-SNE is proposed in Figure 3.

3.2 Annotations

Each report is annotated by two board-certified radiologists. To ensure that there is a baseline level of concordance in the clinical judgments made by the two radioligists, we require the average agreement to be equal to or exceed a threshold of 50%. The average agreement rates for different imaging studies are 53.58% for Chest X-ray, 59.26% for Chest CT, 59.44% for Abdomen Pelvis CT, and 55.55% for Brain MR. If there is no consensus between the two radiologists, one judge is called upon to make a decision. In total, 406,141 annotations have been validated.

We use the same schema as RadGraph-1.0 to extract **entities** and **relations** from radiology reports: **entities** can be labeled as 'Observation definitely present', 'Observation definitely absent', 'Observation uncertain', 'Anatomy definitely absent', 'Anatomy definitely present' or 'Anatomy uncertain' and **relations** between entities can be labeled as 'Located At', 'Modify', 'Suggestive Of'. For a detailed explanation of what we consider to be an entity or a relation, please refer to Appendix A. In addition to the established schema, we've introduced a post-processing step that identifies entities related to **measurements**. This effort

Type	Label	Anatomy and Modality				Total
		Chest CT	Abdomen/Pelvis CT	Brain MR	Chest X-ray	
Entity	Anatomy	33,976	46,326	25,104	7,715	113,121
	Observation definitely present	22,425	35,595	18,033	6,469	82,522
	Observation: Definitely Absent	5,705	8,975	7,215	987	22,882
	Observation: Uncertain	2,104	2,867	2,121	946	8,038
	Total	64,210	93,763	52,473	16,117	226,563
Relation	Modify	29,892	46,708	29,608	7,471	113,679
	Located at	18,313	25,333	11,345	4,163	59,154
	Suggestive of	2,081	2,443	1,504	717	6,745
	Total	50,286	74,484	42,457	12,351	179,578

Table 3: Overview of the 406,141 RadGraph-XL annotations categorized by entity and relations across various modalities and anatomies, detailing the different distributions per labels. Additionally, a subset of 3,297 measurements have been identified. We show the details in Table 13 in Appendix.

is geared towards encouraging future research to create new models designed for dealing with or forecasting measurements, an area within radiology AI that, based on our informed understanding, presents unique challenges and has not been extensively addressed. This additional step, detailed in Appendix C, allowed us to annotate a subset of 3,297 entities in RadGraph-XL and 65 in RadGraph-1.0.

Finally, the task of annotating new modality-anatomy pairs presented significant challenges that are quite distinct from those encountered with chest X-rays, which were the focus of RadGraph-1.0. These complexities are detailed in Appendix B. In particular, we note that chest X-ray reports are considerably shorter than the reports for other modality-anatomy pairs, as shown in Figure 1.

3.3 Statistics

Table 3 provides a detailed breakdown of the annotations collected, organized by type of imaging study and annotation categories. It is important to highlight that the dataset is evenly balanced, with anatomical annotations comprising 49.92% and observations making up 50.08%. The abdomen/pelvis CT reports, which are the longest reports in our collection as depicted in Figure 1, account for 41.38% of all annotations. This is followed by chest CTs at 28.34%, brain MRs at 23.16%, and chest x-rays at 7.11%. Regarding the types of relations annotated, 63.30% are classified as 'modify', 32.94% as 'located at', and 3.75% 'suggestive of'. For entities, we identify 19,772 unique (entity, label) pairs; for relations, we find 67,323 (source entity, target entity, label) unique triplets. The 10 most common unique entity pairs and relation triplets are presented in Table 14.

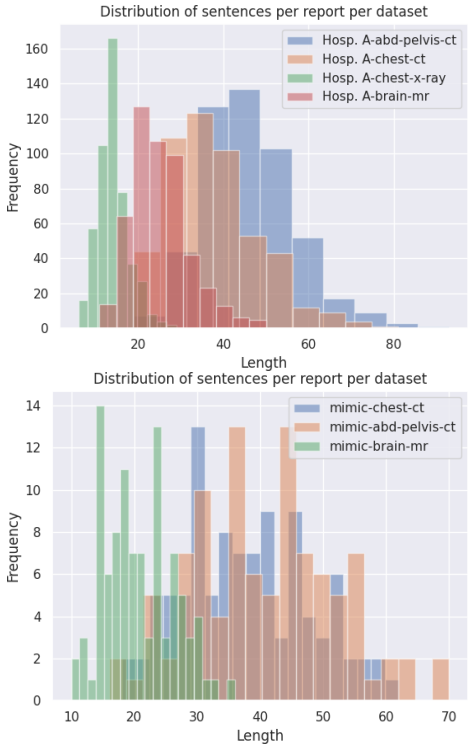

Figure 1: Statistics of RadGraph-XL on the Hosp. A (Top) and MIMIC-CXR (Bottom), where the distributions of the number of sentences per report, per imaging studies, and per institution are shown.

4 Experiments

4.1 Baseline training

In this section, we aim to develop a predictive model using both the newly annotated dataset and the data from RadGraph-1.0. To achieve

Approach	Entity		Relations	
	Micro F1	Macro F1	Micro F1	Macro F1
SpERT				
BERT	$0.844_{\pm 0.001}$	$0.707_{\pm 0.033}$	$0.638_{\pm 0.016}$	$0.513_{\pm 0.023}$
BiomedBERT	$0.741_{\pm 0.002}$	$0.646_{\pm 0.020}$	$0.535_{\pm 0.007}$	$0.463_{\pm 0.018}$
BiomedVLP-CXR-BERT	$0.743_{\pm 0.003}$	$0.642_{\pm 0.033}$	$0.538_{\pm 0.007}$	$0.431_{\pm 0.025}$
DYGIE++				
BERT	$0.877_{\pm 0.002}$	$0.758_{\pm 0.005}$	$0.729_{\pm 0.016}$	$0.664_{\pm 0.018}$
BiomedBERT	$0.880_{\pm 0.000}$	$0.785_{\pm 0.002}$	$0.725_{\pm 0.014}$	$0.671_{\pm 0.017}$
BiomedVLP-CXR-BERT	$0.889_{\pm 0.000}$	$0.796_{\pm 0.001}$	$0.737_{\pm 0.015}$	$0.689_{\pm 0.023}$
BiomedVLP-CXR-BERT(b)	0.889	0.797	0.739	0.691

Table 4: Aggregated results from the 10-folds for the DYGIE++ and SpERT framework. In this context, a true positive is defined as an instance where the prediction of an entity or relation is completely accurate. This accuracy encompasses correctly identifying the span and label of an entity. For relations, it involves correctly determining the spans of both the source and target entities involved in the relation, as well as accurately identifying the label of the relation.

this, we evaluate two transformer-based libraries under MIT License for Entity and Relation Extraction: DyGIE++ (Wadden et al., 2019) and SpERT (Eberts and Ulges, 2020) with three proven transformer architectures, namely BERT (Kenton and Toutanova, 2019), BiomedBERT Gu et al. (2020) and BiomedVLP-CXR-BERT Boecking et al. (2022). We assess the effectiveness of our training through a 10-fold cross-validation process. Considering that the majority of the models discussed in this paper handle sequences of no more than 512 tokens, we ensure that both the reports and annotations are divided appropriately.

The summarized results are presented in Table 4, where we demonstrate that the DYGIE++ framework delivers the best performance overall. In terms of comparing various transformer architectures, the differences observed between them are minimal. The highest-scoring transformer model is BiomedVLP-CXR-BERT.

4.2 Selecting the Best Model

From the 10-fold cross-validation process, we identified the training, validation, and testing splits based on the fold where BiomedVLP-CXR-BERT achieved its highest performance, recording scores of 0.889 and 0.797 for Entity F1 Micro and F1 Macro, and 0.739 and 0.691 for Relations F1 Micro and F1 Macro, respectively. This top-performing model is now referred as to *BiomedVLP-CXR-BERT(b)*. These splits will be used as the standard for our subsequent ablation studies; for context, the selected splits for the training, validation, and test sets include 2320, 290, and 290 reports, respectively. The comprehensive results of *BiomedVLP-CXR-BERT(b)* for entities and relationships in this specific split are detailed in Table 5.

Category		F1 Score	Precision	Recall
NER Label Metrics				
Anatomy	definitely present	0.93	0.92	0.93
Observation	definitely absent	0.90	0.90	0.91
Observation	definitely present	0.85	0.85	0.85
Observation	uncertain	0.77	0.78	0.77
Anatomy	definitely absent	0.53	0.57	0.50
Relations Label Metrics				
modify	-	0.74	0.74	0.74
located at	-	0.75	0.74	0.76
suggestive of	-	0.58	0.60	0.55

Table 5: Comprehensive results of *BiomedVLP-CXR-BERT(b)* for entities and relationships in the chosen split from the 10-fold cross-validation process, based on its peak performance. This split, marked by Entity F1 Micro and Macro scores of 0.889 and 0.797, and Relations F1 Micro and Macro scores of 0.739 and 0.691 respectively, will serve as the official split.

4.3 Scaling with LLMs

The baseline architectures we selected are relatively small by current standards, each having a total of 0.11 billion parameters. In addition, we investigated transformer models with varying numbers of parameters, specifically XLM-Roberta (Conneau et al., 2019), which has 0.5 billion parameters and was trained on 2.5TB of filtered CommonCrawl

data. We also looked at Pythia (Biderman et al., 2023), with 1 billion parameters trained on the Pile, and StableLM2 (StabilityAI, 2024), which has 1.6 billion parameters and was trained on a dataset of 2 trillion tokens.

Approach	Entity	Relations
	Macro F1	Macro F1
BiomedVLP-CXR-BERT	0.796	0.689
XLM-Roberta	0.702	0.650
Pythia 1B	0.650	0.632
StableLM2 2.7B	0.789	0.656

Table 6: Comparison of BiomedVLP-CXR-BERT backbone against larger models on a 10-fold cross validation experiment.

XLM-Roberta, Pythia, and StableLM2 reports micro F1 scores that are closely matched with those of BiomedVLP-CXR-BERT, with values for entities between 0.87 and 0.88 and for relations between 0.71 and 0.73. However, they fall short in performance for certain under-represented labels, as indicated by the macro F1 scores presented in Table 6. Particularly, XLM-Roberta reports an F1 score of 0 for 'Anatomy definitely absent', 0.68 for 'Observation uncertain', and 0.47 for the relation 'suggestive of'.

4.4 Performance on Measurements

We found that a small number of outlier labels were identified as measurements, specifically 'Observation definitely absent', 'Observation uncertain', 'Anatomy definitely absent', and 'Anatomy uncertain', with occurrence totals of 11, 7, 4, and 3, respectively.[2] In our official test set, only four labels are included as measurements. Table 7 presents the performance metrics for these labels as evaluated by our top-performing model, *BiomedVLP-CXR-BERT(b)*.

Measurements	Entity		
	F1 Score	Precision	Recall
Obs. definitely present	0.820	0.860	0.780
Anat. definitely present	0.630	0.580	0.700
Obs. definitely absent	0.660	1.000	0.500
Obs. uncertain	0.660	1.000	0.500

Table 7: Performance on measurements entities by our best model *BiomedVLP-CXR-BERT (b)* on the test-set of our official split.

[2]We provide details in Table 12 in the Appendix.

4.5 Comparison to RadGraph-1.0

To assess the value of our new annotations, we conducted two experiments.

The first experiment involves testing the model trained on RadGraph-1.0 (chest X-rays only) on our official test split, ensuring we excluded annotations labeled as 'Anatomy Uncertain' and 'Anatomy definitely absent' since they do not exist in the RadGraph-1.0 schema.

Approach	Entity	Relation
	Macro F1	Macro F1
BiomedVLP-CXR-BERT(b)	0.863	0.691
RadGraph-1.0	0.744	0.453

Table 8: Comparison of the model provided by RadGraph-1.0 with our top-performing model on our official test-set. The presented results were obtained by excluding the categories 'Anatomy Uncertain' and 'Anatomy definitely absent', since these are not included in RadGraph-1.0.

As seen in Table 8, the *BiomedVLP-CXR-BERT(b)* model significantly outperforms RadGraph-1.0 in both categories. Specifically, in the Entity category, our model achieves a Macro F1 score of 0.863, which is approximately 16.0% higher than RadGraph-1.0's score of 0.744. In the Relations category, the improvement is even more pronounced, with our model attaining a score of 0.691, which surpasses RadGraph-1.0's score of 0.453 by 52.5%. These results suggest that *BiomedVLP-CXR-BERT(b)* provides a significantly more effective approach for recognizing entities and their relations on reports from various imaging studies. The detailed results of RadGraph-1.0 on our test-set are presented in Table 9. A significant discrepancy is observed in the category 'Observation definitely present', where the performance of RadGraph-1.0's model is 20 f1-score points inferior compared to *BiomedVLP-CXR-BERT(b)*.

In the second experiment, we trained the BiomedVLP-CXR-BERT backbone using all available data except for the official test set from RadGraph-1.0, which includes only annotations for chest X-rays. The outcomes of this experiment are detailed in Table 10. We observe that our BiomedVLP-CXR-BERT model, despite being trained on a large, diverse dataset, can match the performance of the RadGraph-1.0 model on the RadGraph-1.0 test-set. It's also worth mentioning

Category		F1 Score	Precision	Recall
Anatomy	definitely present	0.83	0.80	0.86
Observation	definitely absent	0.71	0.65	0.77
Observation	definitely present	0.61	0.67	0.56
Observation	uncertain	0.81	0.80	0.83
Relations Label			**Metrics**	
modify	-	0.48	0.44	0.52
located at	-	0.52	0.54	0.50
suggestive of	-	0.35	0.45	0.29

Table 9: Detailed results of the RadGraph-1.0 model tested on our RadGraph-XL official test-split. These results can be directly compared to Table 5 as they are computed on the same test-set.

that the test set is relatively small, consisting of 100 reports focused on chest X-rays. These reports are typically brief and offer limited semantic variety compared to other types of imaging studies found in our RadGraph-XL dataset.

Approach	Entity	Relation
	Macro F1	Macro F1
BiomedVLP-CXR-BERT	0.862	0.694
RadGraph-1.0	0.862	0.692

Table 10: Comparison of the model provided by RadGraph-1.0 with our top-performing model on RadGraph-1.0 test-set.

4.6 Comparisons with GPT-4

Recent work has demonstrated the utility of GPT-4, a task-agnostic foundation model, in effectively performing a variety of natural language tasks (Achiam et al., 2023; Liu et al., 2023b). In order to compare state-of-the-art task-agnostic models with our task-specific approach, we benchmark performance of GPT-4 on the RadGraph-XL test set. Given an input radiology report, we use GPT-4 to extract entities and relations. We evaluate performance of zero-shot GPT-4, where no in-context examples are provided, and few-shot GPT-4, where between one and ten in-context examples are included in the prompt. In-context examples are sampled randomly from the RadGraph-XL training set. Our results are summarized in Table 11.

We find that performing entity and relation extraction on the RadGraph-XL dataset is challenging for GPT-4, with macro-F1 scores observed to be significantly lower (0.594 F1-points on entity extraction and 0.667 F1-points on relation extraction)

Approach	Entity	Relation
	Macro F1	Macro F1
GPT4 (0-shot)	0.158	0.012
GPT4 (1-shot)	0.173	0.010
GPT4 (5-shot)	0.203	0.020
GPT4 (10-shot)	0.203	0.024
BiomedVLP-CXR-BERT(b)	0.797	0.691

Table 11: We compare our top-performing model with GPT-4 on the official RadGraph-XL test set.

than our task-specific approach. Few-shot GPT-4 with in-context examples exhibit slight improvements in performance over zero-shot GPT-4 (0.045 F1 points on entity extraction and 0.012 F1 points on relation extraction). In line with prior work (Liu et al., 2023a), we find that the key source of GPT-4 errors comes from incorrect understanding of the annotation schema, even in few-shot settings.

Overall, our experiments show that GPT-4 requires substantial manual prompt tuning, generates outputs that do not adequately align with the annotation schema, and requires significant post-processing of generated outputs. Additionally, evaluations with GPT-4 are expensive, which is a particular concern on the RadGraph-XL dataset where reports are lengthy with a large number of entities and relations. Our results demonstrate (i) the need for task-specific models like our *BiomedVLP-CXR-BERT(b)* model, which are capable of performing specialized tasks with high accuracy, and (ii) that RadGraph-XL can serve as a useful and challenging test-bed for future foundation models.

5 Reader Study

We conduct a reader study on out-of-domain data, namely Deep Vein Thrombosis (DVT) ultrasound reports, in order to evaluate the ability of our model to generalize to new radiological text. We chose 20 reports with semantic diversity, extracted the impressions section, and ran our top-performing model *BiomedVLP-CXR-BERT(b)* to predict entities and relations. Our model generated 265 entities (13.25 per report) and 207 relations (10.35 per report). A board-certified radiologist was tasked to detect critical errors, imprecise or ambiguous classifications and unclear labels, and provide a subjective overview summary.

Critical errors Three critical errors were detected. First, 'deep' in 'deep veins' was twice labeled as an observation, though it should be anatomy. This is a surprising edge case because i) our RadGraph-XL training set contains 51 'deep' annotations, 40 of which are labeled as Anatomy: definitely present ii) the other 'deep' words were labeled correctly. Secondly, in one case, 'some areas' was labeled as an observation instead of anatomy (referring to some areas of the blood vessel). Finally, in one impression, 'loss of phasicity' and 'loss of normal response' were labeled as 'definitely present', but should have been labeled as 'definitely absent'.

Imprecise or ambiguous classifications A few awkward labels have been predicted: 'Thrombosis' incorrectly modified 'venous' instead of indicating location. The entities 'baker cyst' and 'color flow' were wrongly marked as 'present' instead of 'uncertain', while 'infection' and 'focal' were mistakenly labeled as 'uncertain' rather than 'definitely present'.

Overview summary RadGraph-XL can effectively generalize to an unknown modality and anatomic terms. For example, it was able to show that 'spectral doppler imaging' modifies 'flow', a combination of entities that is non-existent in our training dataset. Although the overwhelming majority of anatomic terms were classified correctly, there is opportunity for improvement in classifying anatomic terms, in this case "deep" as an anatomic modifier of deep vein thrombosis, that were frequently misclassified.

In summary, 5 entities out of 265 were critical errors (1.8%) and 4 entities were subjectively flagged as imprecise (1.5%). Only one relation was subjectively flagged as imprecise. Despite this study being carried on a small sample focused on ultrasound done for deep venous thrombosis, the results are encouraging for the broader use of our RadGraph-XL model for radiological information extraction.

6 Conclusion

We introduced RadGraph-XL, an expansive dataset comprising 2,300 radiology reports enriched with over 410,000 expert annotations from radiologists (Section 3). This dataset spans a variety of entities, relations, and measurements across multiple modality-anatomy pairs, enriching the data extracted from radiological texts with unprecedented precision and depth. We have conducted experiments (Section 4) using transformer-based models trained for automatic annotation of radiology reports, employing state-of-the-art frameworks for entity and relation extraction. Through comprehensive ablation studies (Section 4.2, 4.3, and 4.4) and a reader study (Section 5) that extends to out-of-domain data, we meticulously evaluated our model's performance. The results reveal that our model not only sets a new benchmark, but also outperforms previous methods by as much as 52% (Section 4.5) and notably outperforms GPT-4 (Section 4.6) in this specific field. To encourage further innovation and research, we release the reports, the annotations, and our trained model.

7 Limitations

We denote three limitations to our work. First and foremost, the experiments have been conducted on the raw annotations without further post-processing. The annotations could be refined by implementing various heuristics to identify and address outliers in the dataset. For instance, entities with unusually long spans could be flagged for review, as these may indicate potential mislabeling or annotation errors. Similarly, entities that appear to be mislabeled could be systematically identified and corrected; Those that lack any annotations might be removed to ensure the dataset's consistency and relevance.

Secondly, the selected transformer architectures have a maximum input size of 512 tokens, but many reports in our dataset are longer than that. It's uncertain if dividing a report into several parts affects the model's effectiveness due to the loss of context. Additionally, expanding the model to a size comparable to 'Large Language Models' and fine-tuning all its parameters hasn't led to any enhancements. More advanced techniques, referred to as Parameter-Efficient Fine-Tuning (PEFT), might allow for more consistent training and the scaling up to larger models that are capable of more sophisticated reasoning.

Finally, our reader study indicates that while our model generally produces good annotations on unseen datasets, it is not immune to significant errors when dealing with out-of-distribution data. It remains uncertain how effectively our model handles unseen modalities and anatomies, and whether it can be considered reliable for annotating data in such contexts for subsequent tasks.

Ethical considerations

The reports subject to the annotations have been automatically deidentified with human review and approved for release by the institution. To guarantee patient safety when deploying clinical models in practice, it's crucial for researchers training models on our datasets to rigorously audit for performance disparities across key demographic attributes, such as sex, age, and race. This involves a proactive approach to identifying and addressing potential distribution shifts that may occur when these models are applied across diverse patient populations, ensuring equitable and effective healthcare outcomes for all individuals.

We publicly release of our top-performing model, *BiomedVLP-CXR-BERT(b)*, which is capable of automatically annotating radiology reports. Along with the model, we are also sharing our annotations with the public.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer.

Aurelia Bustos, Antonio Pertusa, Jose-Maria Salinas, and Maria De La Iglesia-Vaya. 2020. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174.

Zhihong Chen, Maya Varma, Xiang Wan, Curtis Langlotz, and Jean-Benoit Delbrouck. 2023. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 469–484, Toronto, Canada. Association for Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Surabhi Datta, Yuqi Si, Laritza Rodriguez, Sonya E Shooshan, Dina Demner-Fushman, and Kirk Roberts. 2020a. Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest x-ray reports using deep learning. *Journal of biomedical informatics*, 108:103473.

Surabhi Datta, Morgan Ulinski, Jordan Godfrey-Stovall, Shekhar Khanpara, Roy F Riascos-Castaneda, and Kirk Roberts. 2020b. Rad-spatialnet: a frame-based resource for fine-grained spatial relations in radiology reports. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation*, volume 2020, page 2251. NIH Public Access.

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.

Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. 2023. Overview of the radsum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 478–482.

Markus Eberts and Adrian Ulges. 2020. Span-based joint entity and relation extraction with transformer pre-training. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI)*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Michael Jantscher, Felix Gunzer, Roman Kern, Eva Hassler, Sebastian Tschauner, and Gernot Reishofer. 2023. Information extraction from german radiological reports for general clinical text and language understanding. *Scientific Reports*, 13(1):2353.

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.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2.

Curtis P Langlotz and Lee Meininger. 2000. Enhancing the expressiveness and usability of structured image reporting systems. In *Proceedings of the AMIA symposium*, page 467. American Medical Informatics Association.

Wilson Lau, Kevin Lybarger, Martin L Gunn, and Meliha Yetisgen. 2023. Event-based clinical finding extraction from radiology reports with pre-trained language model. *Journal of Digital Imaging*, 36(1):91–104.

Mingjie Li, Wenjia Cai, Karin Verspoor, Shirui Pan, Xiaodan Liang, and Xiaojun Chang. 2022. Cross-modal clinical graph transformer for ophthalmic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20656–20665.

Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel Castro, Maria Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, Pranav Rajpurkar, Sameer Khanna, Hoifung Poon, Naoto Usuyama, Anja Thieme, Aditya Nori, Matthew Lungren, Ozan Oktay, and Javier Alvarez-Valle. 2023a. Exploring the boundaries of GPT-4 in radiology. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14414–14445, Singapore. Association for Computational Linguistics.

Qianchu Liu, Stephanie Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C Castro, Maria Teodora Wetscherek, Robert Tinn, Harshita Sharma, Fernando Pérez-García, Anton Schwaighofer, et al. 2023b. Exploring the boundaries of gpt-4 in radiology. *arXiv preprint arXiv:2310.14573*.

Vignav Ramesh, Nathan A Chi, and Pranav Rajpurkar. 2022. Improving radiology report generation systems by removing hallucinated references to non-existent priors. In *Machine Learning for Health*, pages 456–473. PMLR.

Eduardo P Reis, Joselisa PQ de Paiva, Maria CB da Silva, Guilherme AS Ribeiro, Victor F Paiva, Lucas Bulgarelli, Henrique MH Lee, Paulo V Santos, Vanessa M Brito, Lucas TW Amaral, et al. 2022. Brax, brazilian labeled chest x-ray dataset. *Scientific Data*, 9(1):487.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

StabilityAI. 2024. Stable lm 2 1.6b.

K Sugimoto, T Takeda, JH Oh, S Wada, S Konishi, A Yamahata, S Manabe, N Tomiyama, T Matsunaga, K Nakanishi, et al. 2021. Extracting clinical terms from radiology reports with deep learning. *Journal of Biomedical Informatics*, 116:103729–103729.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Chuck Lau, Ryutaro Tanno, Ira Ktena, et al. 2023. Towards generalist biomedical ai. *arXiv preprint arXiv:2307.14334*.

Maya Varma, Jean-Benoit Delbrouck, Sarah Hooper, Akshay Chaudhari, and Curtis Langlotz. 2023. Villa: Fine-grained vision-language representation learning from real-world data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22225–22235.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. 2019. Entity, relation, and event extraction with contextualized span representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5784–5789.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023a. Medklip: Medical knowledge enhanced language-image pre-training. *medRxiv*, pages 2023–01.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023b. Towards generalist foundation model for radiology. *arXiv preprint arXiv:2308.02463*.

Benjamin Yan, Ruochen Liu, David Kuo, Subathra Adithan, Eduardo Reis, Stephen Kwak, Vasantha Venugopal, Chloe O'Connell, Agustina Saenz, Pranav Rajpurkar, et al. 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.

Shujun Zhang, Liwei Tan, Qi Han, Hongyan Wang, and Jianli Meng. 2023a. Automatic report generation on a large-scale stroke mri dataset. In *2023 IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, pages 123–128.

Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Weidi Xie, and Yanfeng Wang. 2023b. Knowledge-enhanced visual-language pre-training on chest radiology images. *Nature Communications*, 14(1):4542.

A Information Schema

In RadGraph-1.0 (Jain et al., 2021), Entities and Relations are defined as such:

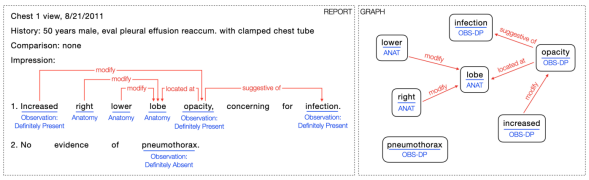

Figure 2: Example of annotations for an impression. Figure taken from Jain et al. (2021).

Entities: We categorize text into units called 'entities', which are spans of text that might be just one word or a string of words stuck together. These entities fall into two big buckets: 'Anatomy', which is about body parts like the lung you might read about in a medical report, and 'Observation', which is about words associated with visual features, identifiable pathophysiologic processes, or diagnostic disease classifications.

Relations: We look at how these entities relate to each other, which are like arrows that connect one entity to another in a specific way. We use three types of relations: 'Suggestive Of', which connects two Observations when one might imply the other; 'Located At', which links an Observation to an Anatomy to show where something's happening or to describe their relationship in other ways; and 'Modify', which can connect two Observations or two Anatomies to show how one changes or adds detail to the other.

B Labeling challenges

In the complex landscape of radiology reports, accurately identifying and annotating anatomical terms and their associated modifiers presents a significant challenge. This challenge is not only important to create high-quality labels but also crucial for maintaining consistency across reports. The nuances involved in this process can lead to variability in interpretations, which, in turn, may affect patient care and outcomes.

Anatomical Term Identification A primary concern in anatomical term identification is distinguishing between the main anatomical regions or organs and the modifiers that specify their exact locations or characteristics. An illustrative example can be seen in the description of lung scarring: 'The lung bases are clear with the exception of some scarring in the right lung base.' Here, 'right', 'lung', and 'base' are all anatomical terms. The ambiguity arises in determining whether 'right' modifies 'lung' or 'base', or if 'right lung base' should be collectively annotated as a singular anatomical entity. To mitigate such ambiguities, it is recommended that the major anatomic region or organ, in this case, 'lung', be labeled as the primary anatomical term. The terms 'right' and 'base' should then be annotated as modifiers that delineate the specific location within the lung.

Modifier Identification Another layer of complexity is introduced when considering how to accurately label modifiers, particularly in phrases where multiple anatomical terms are present. For instance, the phrase "There is moderate intrahepatic biliary duct dilatation" contains "intrahepatic," "biliary," and "duct" as anatomical terms. The challenge here is to ascertain whether "intrahepatic" modifies "duct" or "biliary." Consistency can be achieved by identifying the duct as the primary anatomical term and treating "intrahepatic" and "biliary" as modifiers that provide additional specificity.

Measurements A common question that arises in this context is how to handle phrases that include qualifiers such as "up to," "less than," or "greater than," which provide crucial information about the measurements being reported. Consider the sentence: "The CBD (Common Bile Duct) itself measures up to 3 cm in diameter." The use of "up to" may not be the most precise phrasing for a radiology report, where the exact measurement is typically preferred. However, the reality of clinical practice often involves approximations and ranges, particularly when exact measurements are challenging to obtain. Given their significance, it is recommended that qualifiers such as "up to," "less than," and "greater than" be labeled as observation modifiers.

Qualitative Modifiers Annotating qualitative modifiers such as 'extensive', 'some', and 'clear' in radiology reports presents a notable challenge. These terms significantly impact the clinical interpretation by modifying observations (e.g., 'extensive diverticulosis') or indicating uncertainty (e.g., 'grossly unremarkable'). The complexity arises from their dual role in describing the severity of findings and spatial relationships between anatomical entities. Our approach recommends labeling terms that alter the interpretation of findings as observations and utilizing a generalized 'located_at' relation for spatial descriptors to simplify the annotation process. Terms that introduce ambiguity,

like ‘clear’ and ‘grossly’, are best represented by annotating the corresponding observations as ‘uncertain’.

Contextual Modifiers Phrases like ‘in the setting of recent surgical procedure’ or ‘hematocrit drop’ provide essential clinical context but do not directly describe imaging findings. Our guideline suggests excluding these terms from annotation, as they do not describe the radiological findings.

Compound words Determining whether to split or merge terms for annotation, such as in "hiatal hernia" or "focal pancreatitis," can be perplexing. The rule of thumb is to label words individually to maintain clarity, especially since compound terms might not always appear together in the text. However, it’s crucial to identify the primary entity in each compound term, which typically represents the main anatomy or observation. For example, "hernia" in "hiatal hernia" is the observation, with "hiatal" specifying the anatomical location. Similarly, "pancreatitis" is the observation in "focal pancreatitis," with "focal" indicating the observation’s nature.

C Measurements

The following code snippet was used to detect measurements.

Algorithm 1 Check measurement in an entity

```
# entity is a list of words
# e.g. ["5","x","5","mm"]
if "mm" in entity or "cm" in entity or \
   "MM" in entity or "CM" in entity or \
   ("x" in entity and any(w.isdigit()
    for w in entity)):
        # entity is considered a measure
```

Captured measurements are highly diverse, such as ’approximately a 4.6 cm’, ’advanced by at least 11 cm’, ’measuring slightly less than 6 mm’ or ’smaller in size compared to the prior study measuring 1.5 cm in the largest dimension’. Measurements are distributed across labels as such:

Category	Count
Observation definitely present	3212
Anatomy definitely present	125
Observation definitely absent	11
Observation uncertain	7
Anatomy definitely absent	4
Anatomy uncertain	3

Table 12: Distribution of measurements per label

The measurements are distributed as follows between imaging studies:

Imaging Study	Count
Hosp A. Abdomen/Pelvis CT	1421
Hosp A. Chest CT	1035
Hosp A. Brain MR	241
MIMIC Chest CT	225
MIMIC Abdomen/Pelvis CT	199
Hosp A. Chest X-ray	96
MIMIC Brain MR	80
MIMIC Chest X-ray	65

Table 13: Distribution of measurements per imaging study

D Dataset

The figure below illustrates the process we used to select the reports, as detailed in Section 3.1.

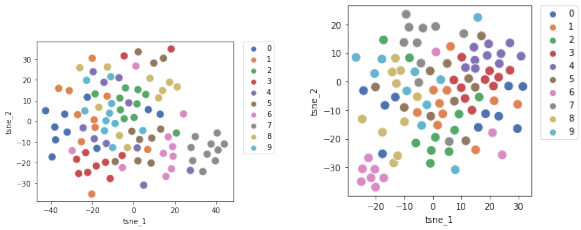

Figure 3: t-SNE representation of the embeddings generated by the Universal Sentence Encoder for CT abdomen/pelvis (left) and MR Brain (right). We use the automatic topic modeling LDA algorithm (Blei et al., 2003) to generate ten clusters.

Entities	Label	Count
Right	Anatomy: DP	4078
Left	Anatomy: DP	3652
Normal	Observation: DP	3619
Unremarkable	Observation: DP	1840
Lobe	Anatomy: DP	1572
Pulmonary	Anatomy: DP	1553
Artery	Anatomy: DP	1402
Size	Anatomy: DP	1222
Small	Observation: DP	1193
Pleural	Anatomy: DP	1118
Source => Target	**Label**	**Count**
Right => Lobe	Modify	818
Normal => Caliber	Located At	706
Normal => Size	Located At	695
Effusion => Pericardial	Located At	665
Left => Lobe	Modify	564
Lower => Lobe	Modify	501
Effusion => Pleural	Located At	462
Small => Bowel	Modify	414
Size => Heart	Modify	378
Caliber => Aorta	Modify	376
Adrenal => Glands	Modify	375

Table 14: Most common entities and relations in the dataset

E Training details

Our best model is trained using the Entity and Relation Extraction framework DyGIE++ (Wadden et al., 2019). The parameters are defined in Table 15.

Parameter	Value
max_span_width	8
initializer	xavier_normal
Loss Weights - ner	0.2
Loss Weights - relation	1.0
Feedforward Params - num_layers	2
Feedforward Params - hidden_dims	768
Feedforward Params - dropout	0.4
Data Loader - sampler_type	random
Data Loader - batch_size	8
num_epochs	100
grad_norm	5.0
Optimizer (classifier) - lr	1e-3
Optimizer (classifier) - weight_decay	0.0
Optimizer (transformer) - lr	5e-5
Optimizer (transformer) - weight_decay	0.1
Learning Rate Scheduler - type	slanted_triangular

Table 15: Hyperparameters

F GPT-4 Evaluations

We provide the prompt used for GPT-4 evaluations in Figure 4.

GPT-4 prompt

Prompt:
Your task is to extract medical entities and relations from a given radiology report. I'll provide you with 1) the problem setup, 2) the radiology report, and 3) the output format.

1) Problem setup: For each report, you will be asked to identify 7 types of medical entities:
(1) `observation::present`, which is used for visual features, pathophysiologic processes, or diagnosable diseases that are present;
(2) `observation::absent`, which is used for visual features, pathophysiologic processes, or diagnosable diseases that are absent;
(3) `observation::uncertain`, which is used for visual features, pathophysiologic processes, or diagnosable diseases where you are uncertain about presence or absence;
(4) `observation::measurement::present`, which refers to a measurement associated with visual features, pathophysiologic processes, or diseases;
(5) `anatomy::present`, which refers to an anatomical body part that is present;
(6) `anatomy::absent`, which refers to an anatomical body part that is absent;
(7) `anatomy::measurement::present`, which refers to a measurement associated with an anatomical body part;
For each report, you will also be asked to identify 3 types of relations between entities:
(1) `suggestive_of`, which is a relation between two Observation entities indicating that the presence of the second Observation is inferred from the first Observation.
(2) `located_at`, which is a relation between an Observation entity and an Anatomy entity indicating that the Observation is related to the Anatomy
(3) `modify`, which is a relation between two Observation entities or two Anatomy entities indicating that the first entity modifies the scope of or quantifies the degree of the second entity.

2) Radiology report:
`Report`

3) Output format:
Please strictly follow this output format. Entities must be short substrings (often just 1 word) from the radiology report with no changes to formatting. Each relation exists between a pair of identified entities. Please list entities and relations in the order they appear in the radiology report.

```
[Entities]:
[[<entity>, <entity type>], [<entity>, <entity type>], ..., [<entity>, <entity
type>]]

[Relations]:
[[<entity 1>, <entity 2>, <relation type>], [<entity 1>, <entity 2>, <relation
type>], ..., [<entity 1>, <entity 2>, <relation type>]]
```

Figure 4: Here, we provide the input prompt used by GPT-4 in order to extract entities and relations from RadGraph-XL. Definitions for entities and relations are adapted from (Jain et al., 2021). For few-shot prompting, we append example reports and example outputs to the end of this prompt.