A. Multilabel NER Evaluation

In a multi-label multi-class classification context, we are able to introduce the errors Superset (SUP): system produces all labels in the golden annotation plus additional labels, and Subset (SUB): system produces a subset of golden annotation. As shown in Table A, these errors can be further simplified into COR, SPU, MIS for each label, which naturally correspond to TP, FP, FN and allows the calculation of metrics such as precision, recall, and F1. If we additionally measure TN, we are able to calculate Matthew's correlation coefficient. Since the Type schema only requires overlap between the system and predicted surface strings, with the alignment algorithm this practically corresponds to our segment-level classification evaluation.

For segmentation, we can also adapt the Partial schema to better capture the types of errors in our task and dataset. Specifically, we allow one-to-many and many-to-one alignments (spurious and missed boundaries respectively) to count towards the Partial schema score. Each mistake within the same segment results in a penalty of 0.25, down to a minimum of 0.5 (the same as that would be awarded for a partial match). In order for the partial match to count as subsets or supersets, we require that the boundaries of the parent segment match the outer boundaries of the child segments. Table 9 shows examples of the additional error types.

Finally, the Exact schema is unchanged, and Strict requires boundary string match as well as full multi-label match.

B. Pretraining Parameters

For the unsupervised pre-training of the language model, we used all available reports as detailed in section 3.1 for a total of 1.53m unique reports. We split these into train and validation sets with a ratio of 99:1 and saved them as txt files in the language modelling format (one report per line). We then continued pre-training from the roberta-base checkpoint with 18 samples per batch with 6 Nvidia V100 GPUs on a DGX-1 server and a gradient accumulation step of 18 for a total of 1966 samples per gradient update. The learning rate is then set to 0.0005 in accordance with fairseq recommendations. The model ran for 100 epochs and the best model achieved an evaluation loss of 0.2952 at epoch 90.

C. Finetuning Parameters

We use a 72/18/10 train/validation/test split stratified at the report level, so that each split contains approximately the same prevalence of each finding when aggregated to the report level. Batch sizes were optimised in conjunction with max length for 12GB GPUs (Nvidia Titan V, Nvidia GTX Titan X). This resulted in a batch size of 30 and max length of 256 for report-level tasks, and a batch size of 128 and max length of 64 for the segment-level classification task. Where intermediate layers are required between the transformer embedding and final output logits (e.g., segmentation LSTM, classification pooling), the hidden dimensions are consistently set to 768. We used a learning rate of 5e-6 for the segmentation experiments as it is faster to converge, and 5e-5 for the classification and joint experiments. We did not use warmup schedules for the learning rate. Early stopping is based on boundary match rate for segmentation and macro-averaged segment-level F1 score for classification and joint training. We set min delta to 0.001 and patience to 8, with four validation loops per epoch, which is equivalent to stopping training after 2 epochs of no improvement in the monitored metric. The random seed is set to 768 throughout.

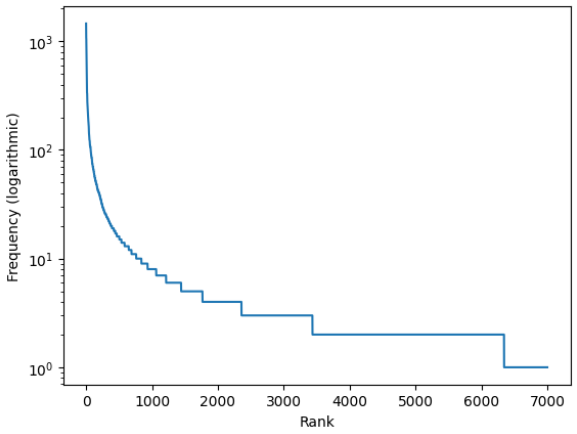

Figure 3: Distribution of segment frequencies in the annotated reports. The horizontal axis is truncated as the remainder of the 77,829 segments occurred only once.

D. Unseen Segments

Due to the templated nature of chest X-ray reports, the number of occurrences of segments follows a power law distribution as shown in Figure D. The most common phrases typically refer to normal imaging findings since their descriptions do not require elaboration, e.g. "no pneumothorax" (n = 1,456), "normal heart size" (n = 939). Of the 127,809 annotated segments, 71,486 occur only once. This makes it important to consider the effect of data contamination between train/test splits, as it is likely that some segments will have been present in both, artificially inflating the evaluation

Poor inspiratory effort. (technical issue) No focal active lung lesion. (normal) CTR equals 15/26 (cardiomegaly) probably artifactual given the suboptimal inspiration. (technical issue)
The lungs and pleural spaces are clear. (normal) Several densely calcific foci adjacent to the trachea may represent calcified mediastinal lymph nodes. (paratracheal hilar enlargemnet, possible diagnosis) Otherwise normal mediastinum. (normal)
Right upper nodule noted with background of COPD changes. (emphysema, hyperexpanded lungs, parenchymal lesion, possible diagnosis) Minor congestive changes also noted in the lung fields in the form of prominent upper lobe veins. (upper lobe blood diversion) CTR is 11/23. (normal)
Large right pneumothorax with collapsed lung. (airspace opacification, pneumothorax, volume loss) Shift of the mediastinum to the left. (mediastinum displaced) Some increased shadowing in the left lung base which may be due to infection. (airspace opacification, possible diagnosis) This report is transcribed using voice recognition software. (unlabelled)

Table 7: Additional example data. Each line corresponds to one labelled segment, and each section is one report.

true	pred	type	counts
y	z	INC	y-MIS, z-SPU
xyz	xy	SUB	xy-COR, z-MIS
xy	xyz	SUP	xy-COR, z-SPU
xy	xz	PAR	x-COR, y-MIS, z-SPU

Table 8: Entity type errors in multi-class multi-label classification.

true	pred	partial	score
abcde	abcd	PAR	0.5
abcde	ab|cde	SUB	0.75
abcde	ab|cd|e	SUB	0.5
ab|cde	abcde	SUP	0.75

Table 9: Partial segmentation errors. In row 1, e was not predicted as a segment, resulting in PAR. Contrasted with row 3 where the prediction contains two spurious boundaries, resulting in SUB and a score of 0.5.

metrics of models trained on segments. However, it is difficult to perfectly identify, or even define, unseen segments. For example, consider cases such as varying severity ("mild cardiomegaly" vs "cardiomegaly"), the aforementioned differing but not incorrect segmentations, or small variations in language ("lungs are clear" vs "lungs clear"). On the other hand, it may be the case that truly difficult/different segments do not follow these patterns and instead contain more detailed information.

With these caveats in mind, we make a best effort to select the unseen segments by filtering for the segments in the test set which do not match any segments in the train and validation sets when both converted to lower case. Our test set consisted of 2,846 reports, resulting in 12,979 predicted segments of which 7,688 were unseen during training and validation. The macro-averaged F1 scores for the seen, unseen, and all segments were 0.8513, 0.7567, and 0.8103 respectively. Similarly, we can calculate the proportion of unseen segments present within a report and observe worse classification performance for reports with the highest proportions of unseen segments.

unseen	count	prec	recall	f1	mcc
50%	570	0.83	0.81	0.82	0.80
60%	504	0.86	0.84	0.85	0.83
67%	990	0.84	0.84	0.84	0.82
71%	221	0.81	0.86	0.82	0.80
75%	119	0.79	0.83	0.80	0.78
100%	400	0.77	0.73	0.74	0.73

Table 10: Effect of the proportion of unseen segments on report-level classification performance. Only showing unseen proportions for which there are over 100 reports.

Despite the apparent trend here, we would like to note two potential confounding factors to the difference in performance on seen and unseen segments. Segments selected in this way are by definition infrequent, as any common phrases would

likely exist in the train/validation sets and filtered out as a result. To be precise, 4,782 of the 7,688 unseen segments occurred only once within the entire annotated set, with 73 additional segments occurring twice or thrice. The remaining 2,833 unseen predicted segments did not exist in the annotations and may be result of the imperfect nature of the segmentation predictions, where segmentation errors may negatively impact the subsequent pooling and classification steps.

E. Full Classification Results

Full classification results for each label are given in table E.

Label	Prevalence	Precision	Recall	F1	MCC
Airspace Opacification	935	0.8417	0.7508	0.7937	0.7035
Aortic Calcification	99	0.9789	0.9394	0.9588	0.9575
Apical Fibrosis	81	0.6500	0.8025	0.7182	0.7133
Atelectasis	361	0.8520	0.8449	0.8484	0.8265
Bone Lesion	154	0.7453	0.7792	0.7619	0.7482
Bronchial Changes	104	0.8824	0.8654	0.8738	0.8691
Bulla	45	0.9149	0.9556	0.9348	0.9339
Cardiac Calcification	13	0.3810	0.6154	0.4706	0.4813
Cardiomegaly	859	0.9714	0.9884	0.9798	0.9710
Cavitating Lung Lesion	55	0.8644	0.9273	0.8947	0.8932
Clavicle Fracture	44	0.8958	0.9773	0.9348	0.9346
Dextrocardia	24	1.0000	0.9583	0.9787	0.9788
Dilated Bowel	44	0.8000	0.9091	0.8511	0.8504
Emphysema	190	0.8608	0.8789	0.8698	0.8604
Extrapleural Soft Tiss. Abn.	28	0.7391	0.6071	0.6667	0.6670
Ground Glass Opacification	55	0.9057	0.8727	0.8889	0.8869
Hemidiaphragm Elevated	124	0.9154	0.9597	0.9370	0.9344
Hernia	106	1.0000	0.9245	0.9608	0.9601
Hyperexpanded Lungs	179	0.9086	0.8883	0.8983	0.8916
Interstitial Shadowing	812	0.7311	0.8202	0.7731	0.6773
Mediastinum Displaced	98	0.9062	0.8878	0.8969	0.8933
Medical Devices	796	0.9674	0.9686	0.9680	0.9555
Paraspinal Mass	5	0.5000	0.4000	0.4444	0.4463
Paratracheal Hilar Enlarg.	299	0.6295	0.4716	0.5392	0.4999
Parenchymal Lesion	202	0.7857	0.7079	0.7448	0.7275
Pleural Abnormality	227	0.8908	0.8987	0.8947	0.8856
Pleural Effusion	684	0.9500	0.9722	0.9610	0.9486
Pneumomediastinum	24	0.9200	0.9583	0.9388	0.9384
Pneumoperitoneum	26	0.8929	0.9615	0.9259	0.9259
Pneumothorax	101	0.9479	0.9010	0.9239	0.9214
Rib Fracture	151	0.9662	0.9470	0.9565	0.9542
Scoliosis	98	0.9697	0.9796	0.9746	0.9737
Subcutaneous Emphysema	57	0.9167	0.9649	0.9402	0.9392
Unfolded Aorta	154	0.9742	0.9805	0.9773	0.9761
Upper Lobe Blood Diversion	267	0.7738	0.6404	0.7008	0.6766
Volume Loss	298	0.8194	0.8523	0.8355	0.8161
Widened Mediastinum	84	0.6329	0.5952	0.6135	0.6024
Abn. Non Clinic. Important	460	0.7458	0.7652	0.7554	0.7076
Comparison	1258	0.9251	0.9428	0.9339	0.8807
Normal	1842	0.9715	0.9794	0.9754	0.9298
Other Non-Findings	357	0.7188	0.7087	0.7137	0.6730
Possible Diagnosis	866	0.8101	0.8372	0.8234	0.7445
Recommendation	407	0.9070	0.9582	0.9319	0.9206
Technical Issue	450	0.8877	0.9311	0.9089	0.8917
Undefined Sentence	90	0.5556	0.5000	0.5263	0.5124
Findings Macro Average	213.1	0.8454	0.8474	0.8441	0.8330
Non-findings Macro Average	716.2	0.8152	0.8278	0.8211	0.7825
Overall Macro Average	302.5	0.8401	0.8439	0.8400	0.8240

Table 11: Report-level classification results of the proposed joint model by label, separated into findings (n = 37) and non-findings (n = 8), and macro-averaged for each category.