

Supplementary Materials: In-context Learning for Zero-shot Medical Report Generation

Anonymous Authors

In the appendix, we will introduce how we select demonstrations to guide the zero-shot learning and detailed performances on each disease.

1 DEMONSTRATION SELECTION

For the cross-center evaluation, we used the samples from the validation set as demonstrations. We conducted the experiments on the IU-Xray to MIMIC adaptation. As previously mentioned, there are 14 disease categories in the MIMIC dataset. Consequently, we randomly selected one demonstration from each category, totaling 14 demonstrations. We also compared the performance with randomly selected demonstrations from the overall samples. Furthermore, we tested the performance using 28 and 42 demonstrations and found that increasing the number of samples does not significantly improve the NLG metrics. Notably, these demonstrations were sourced from the MIMIC validation subset. In Table 1, we presented the detailed performances. It was observed that using 14 evenly distributed demonstrations across diseases, our model can achieve peak performance. Moreover, increasing the number of demonstrations only slightly improves the results. Therefore, we opted to use 14 demonstrations to optimize the query generation process.

2 CROSS-DISEASE EVALUATION

In our cross-disease evaluation, we meticulously maintained a balanced distribution by utilizing 14 demonstrations with equal representation across diseases. This strategy ensures that our model

captures multi-modal contextual information from categories it has not previously learned. By employing this approach, we aimed to enhance the model’s generalizability and robustness across various unencountered medical conditions, thereby demonstrating its efficacy in handling diverse diagnostic scenarios to describe novel diseases.

3 DETAILED PERFORMANCE ON EACH DISEASE

In Fig. 1, we present the detailed performance analysis based on BLEU-4 and Precision metrics for cross-disease evaluation. It is observed that the term ‘No Finding’ achieves the highest values for both metrics, indicating a strong correlation between the generated reports and the reference standards in cases where no disease is observed. Conversely, the category ‘Lung Lesion’ proves to be the most challenging, showing lower scores in both metrics, which may suggest that this condition is more difficult to describe accurately. Furthermore, we also note that the generated reports for ‘Edema’ exhibit high BLEU-4 scores but lower Precision. This implies that while the reports may follow a similar structure or template, they might include varied terminology, which could affect the accuracy of the diagnostic terms used to describe ‘Edema.’

Table 1: Ablation study of demonstration selection in the IU-Xray to MIMIC adaptation across both NLG and CE metrics, where rd and ed refer to randomly distributed and equally distributed.

Setting	NLG Metrics				CE Metrics		
	BLEU-4	ROUGE	METOER	CIDEr	Precision	Recall	F1-score
14 (rd)	0.071	0.219	0.106	0.073	0.125	0.118	0.121
14 (eq)	0.106	0.256	0.128	0.142	0.249	0.235	0.237
28 (eq)	0.108	0.257	0.128	0.142	0.251	0.233	0.236
42 (eq)	0.110	0.258	0.130	0.144	0.249	0.238	0.240

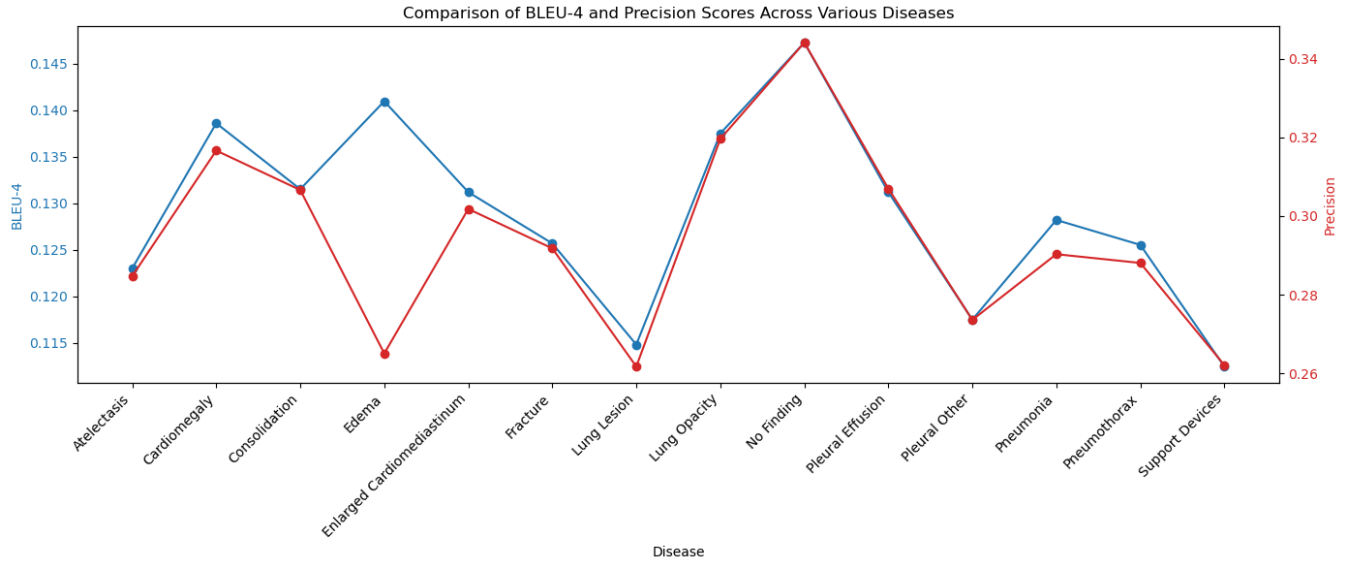


Figure 1: Detailed performance on each disease when conducting cross-disease report generation.