SIIM CMIMI24
Oct 21-22 | Boston, MA
Conference On Machine Intelligence In Medical Imaging

Abdominal Organ Labeling for Abnormality in CT Reports Using a Large Language Model

Ricardo Bigolin Lanfredi, PhD, National Institutes of Health; Yan Zhuang, PhD; Luke Krembs, DO; Brandon Khoury, MD; Pritam Mukherjee, PhD; Ronald M. Summers, MD, PhD

Introduction/Background

Medical report labelers capable of handling various abnormalities, such as CheXpert and CheXbert, have mainly targeted chest X-ray (CXR) reports. However, compared with chest X-ray reports, computed tomography (CT) reports, particularly in the abdominal region, are more complex and cover a broader range of organs and abnormalities. These challenges make abnormality labeling for abdominal organs underexplored.

Methods/Intervention

To address this challenge, we propose a large language model (LLM) labeler, MAPLEZ-CT, to annotate whether major abdominal organs are abnormal. MAPLEZ-CT is an adaptation to CT reports of the previously published MAPLEZ (Medical report Annotations with Privacy-preserving Large language model using Expeditious Zero shot answers) LLM prompt system. The employed zero-shot prompt, which uses the publicly available Meta-Llama-3-70B-Instruct model, run locally to preserve privacy, is displayed in Figure 1. A key feature of the prompt was the inclusion of an extensive definition of abnormalities: any unusual findings the radiologist deems worth mentioning for a specific organ, including atypical anatomical variations, postsurgical changes, and findings in subparts organs. This definition excludes findings indicating limited evaluation, normal organs, adjacent structures, or broad anatomical areas. Additional modifications included LLM pre-extraction of relevant sentences and chain-of-thought reasoning, whose computational complexity was partially offset by the vLLM library, which reduced the processing time by around 92%.

Results/Outcome

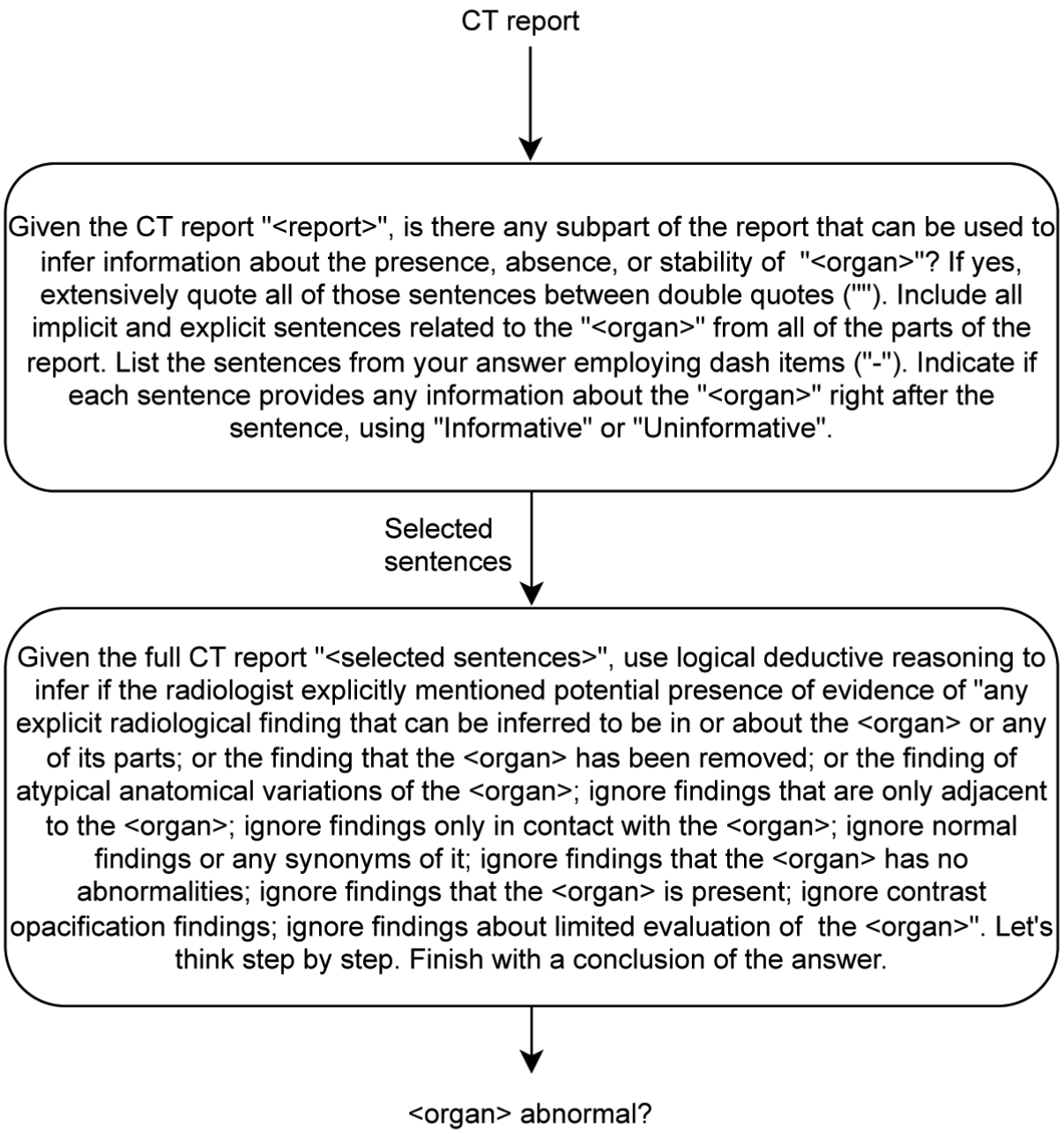
One research fellow and two radiology residents annotated the test set for five major abdominal organs (the spleen, liver, kidneys, gallbladder, and intestines) using 100 private reports randomly sampled from the publicly available Deep Lesion dataset. The final labels were decided through majority voting. The proposed method was compared to MAPLEZ and the rule-based SARLE (Sentence Analysis for Radiology Label Extraction) labeler. It achieved a median F1 score of 0.954 [0.927, 0.975], with an improvement ranging from 0.135 to 0.403 over the median scores of the baseline models. Table 1 shows the results for each organ. We calculated 95% confidence intervals through bootstrapping.

Conclusion

MAPLEZ-CT can reliably label abnormalities for major organs in abdominal CT, outperforming alternatives. It has the potential to create large-scale annotated CT datasets for abnormalities detection.

Statement of Impact

Zero-shot privacy-preserving LLMs can successfully label abnormal organs for CT reports.



Representation of the main parts of the prompt system employed with the Meta-Llama-3-70B-Instruct LLM to extract labels about the abnormality of organs from CT reports

Organ	SARLE	MAPLEZ (Llama2)	MAPLEZ (Llama3)	MAPLEZ-CT (Ours - Llama3)
Spleen	0.349 [0.197, 0.500]	0.606 [0.333, 0.780]	0.875 [0.741, 0.971]	**0.958 [0.872, 1.000]**
Liver	0.692 [0.574, 0.800]	0.921 [0.847, 0.974]	0.968 [0.921, 1.000]	**0.989 [0.963, 1.000]**
Kidney	0.740 [0.632, 0.827]	0.833 [0.725, 0.918]	0.884 [0.794, 0.947]	**0.977 [0.938, 1.000]**
Gallbladder	0.333 [0.185, 0.486]	0.400 [0.100, 0.640]	0.560 [0.273, 0.769]	**0.884 [0.741, 0.978]**
Intestine	0.578 [0.378, 0.732]	0.400 [0.160, 0.615]	0.750 [0.571, 0.889]	**0.902 [0.800, 0.980]**
Median	0.568 [0.511, 0.630]	0.743 [0.679, 0.801]	0.860 [0.812, 0.903]	**0.954 [0.927, 0.975]**

F1 scores of the different CT report labelers applied to find all abnormalities mentioned in the reports for different organs. In bold, we highlight the best model in each row.

Keywords

Large-language models; Abdominal CT; medical reports; Abnormality labels