SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Structuring Free-text Abdominal CT Reports using Privacy-preserving LLMs

Tanvir Agnihotri, MD, Mount Sinai; Dogan Polat, MD; Mark Finkelstein, MD; David Mendelson, MD; Yan Zhuang, PhD, MS; Neil M. Rofsky, MD

Introduction/Background

Existing medical VLMs have been trained using whole radiology reports, which are semantically dense and often ambiguous to learn from. Several recent studies indicated that fine-grained VLMs, trained by region-specific image-text pairs, can provide improved representation granularity. However, few fine-grained VLMs have been investigated for abdominal radiology. In this study, we aimed to transform free-text abdominal CT reports into organ-wise structured reports using a locally-run, privacy-preserving LLM that allow the creation of localized organ-wise image-text pairs for training fine-grained abdominal VLMs.

Methods/Intervention

225 free-text reports were randomly selected from the publicly available Merlin abdominal CT dataset. We designed the prompts to guide the LLM model (Llama 3.3) to generate structured organ-wise descriptions for 10 abdominal organs from free-text reports, including spleen, right kidney, left kidney, gallbladder, liver, stomach, pancreas, right adrenal gland, left adrenal gland, and gastrointestinal tract (Fig. 1). It results in a total of 2250 cases for assessment. The structured reports were graded by a PGY-5 radiology resident (grader 1) and a PGY-3 radiology resident (grader 2). Each grader provided a binary evaluation for each organ, indicating whether the extraction for that organ was successful or not. The accuracy score, defined as the proportion of correct extractions out of the total number of extractions, is used to evaluate the quality of structured reports. Inter-rater reliability was quantified using Cohen's Kappa coefficient.

Results/Outcome

The LLM achieved a mean accuracy of 0.956 when using grader 1’s ratings as a reference standard, and a mean accuracy of 0.951 when using grader 2’s ratings as a reference standard. Organ-wise accuracy is detailed in Fig. 2 (a). The Cohen’s Kappa coefficient was 0.614, at the lower bound of substantial agreement. Given that the Cohen’s Kappa coefficient may be affected by an unbalanced distribution of correct and incorrect extractions, we additionally present the confusion matrix in Fig. 2(b) to provide detailed breakdown of agreement and disagreement patterns among the graders.

Conclusion

The study demonstrated that the LLM can accurately structure free-text abdominal CT reports.

Statement of Impact

Transforming free-text reports into structured ones allows to the creation of organ-wise image-text pairs for training fine-grained VLMs with improved supervision granularities.

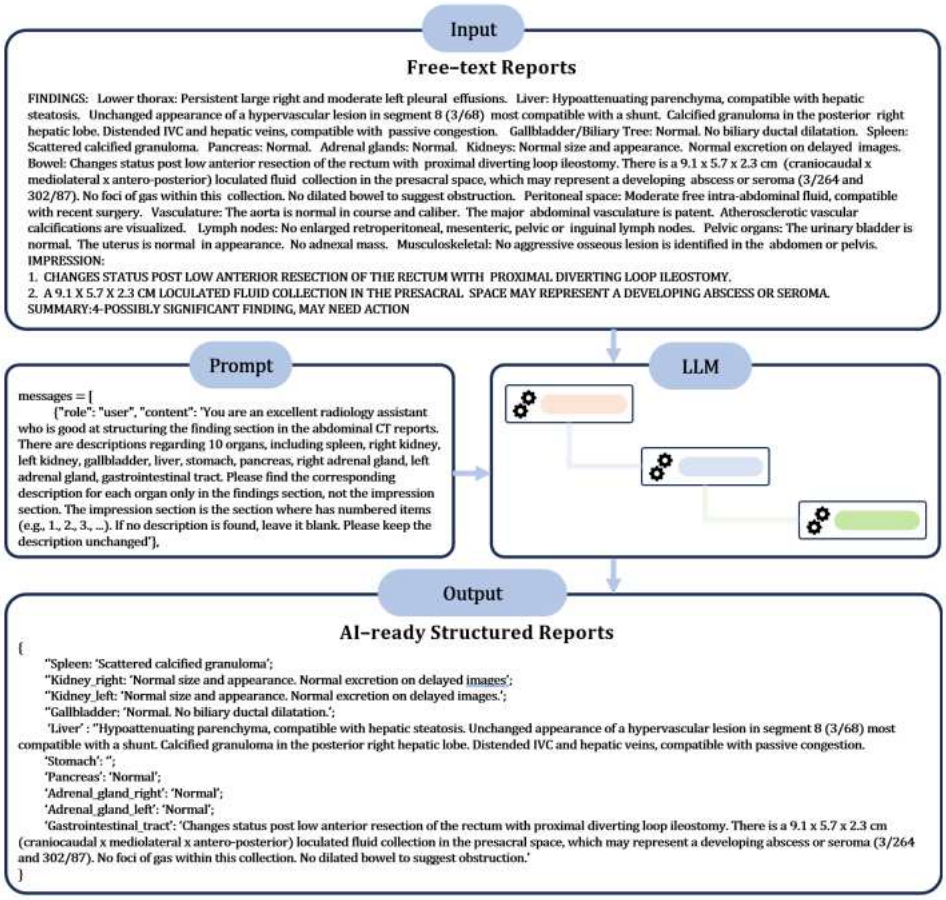

Figure 1. The overall framework to convert free-text reports to structured ones.

Figure 1. The overall framework to convert free-text reports to structured ones.

Accuracy	Spleen	Right kidney	Left kidney	Gallbladder	Liver	Stomach	Pancreas	Right adrenal gland	Left adrenal gland	Gastrointestinal tract	Average
Grader 1	0.987	0.898	0.929	0.991	0.991	0.978	0.991	0.893	0.902	1.0	0.956
Grader 2	0.996	0.929	0.938	0.996	0.996	0.996	0.987	0.858	0.844	0.973	0.951

(a)

		Grader 1	
		Correctly structured	Incorrectly structured
Grader 2	Correctly structured	2107	44
	Incorrectly structured	33	66

(b)

Figure 2. (a) The organ-wise accuracy and overall mean accuracy. The first row shows the LLM accuracy using Grader 1's ratings as the reference standard. The second row shows the LLM accuracy using Grader 2's ratings as the reference standard. **(b)** The confusion matrix for graders 1 and 2, where they agree that 2107 cases were correctly structured and 66 cases were incorrectly structured, while they disagreed on 77 cases in total.

Keywords
Fine-grained large-language model; Structured reports; Region-specific image-text pairs