Using Open-Source Large Language Models to Extract Labels from Radiology Reports for Machine Learning Applications: A Proof-of-Concept Study

Michael Fei, Medical Student, Creighton School of Medicine
Satvik Tripathi; Krishnaveni Parvataneni; Felix Dorfner; Christopher Bridge, PhD; Dania Daye, MD, PhD

Introduction

When developing deep learning models, labeled data is expensive and/or time-consuming to obtain, presenting as one of the largest barriers. Open-source Large Language Models (LLMs) present as a tool to extract labels from radiology reports cheaply and efficiently. This study assesses the efficacy of using open-source LLMs to extract extravasation injury labels from radiology reports.

Hypothesis

LLMs can efficiently extract extravasation injury labels from radiology reports.

Methods

This was an IRB approved study with 6024 radiology reports of abdominal CTs from Massachusetts General Hospital and Brigham and Women's Hospital with the indication of abdominal trauma. The Meta-Llama-3.1-70B, Qwen2.5-72B-Instruct, and Mistral-7B-Instruct-v0.3 models were run locally. The models were prompted using a zero-shot prompt including the report impressions and instructions to evaluate if active extravasation was present and output either: "No", "Yes", or "Undefined". Ground truth was manually generated for 300 reports for statistical analysis. Accuracy, precision, recall, and F1-score with respect to the ground truth were calculated for each model and an ensemble of all three models. On the reports where the Llama model identified active extravasation, the LLM models were further prompted to classify where the extravasation was located between bowel, liver, kidney, abdominal wall, gluteal/thigh, spleen, retroperitoneal, or other.

Results

A total of 125 of the 300 reports contained active extravasation. The Llama model presented with an F1-score of .990, followed by the Ensemble (0.980), Qwen (0.973), and Mistral (.932) models. Looking at localizing the extravasation, the LLM performed strongly in identifying extravasation in the bowel with the highest F1-score of .966.

Conclusion

Small open-source LLM models prove to be an effective tool for labeling radiology reports with high accuracy. This proof of concept yields promising potential to label other pathologies and extract other free texts from radiology reports, greatly reducing cost, labeling time, and burden.

Figure(s)

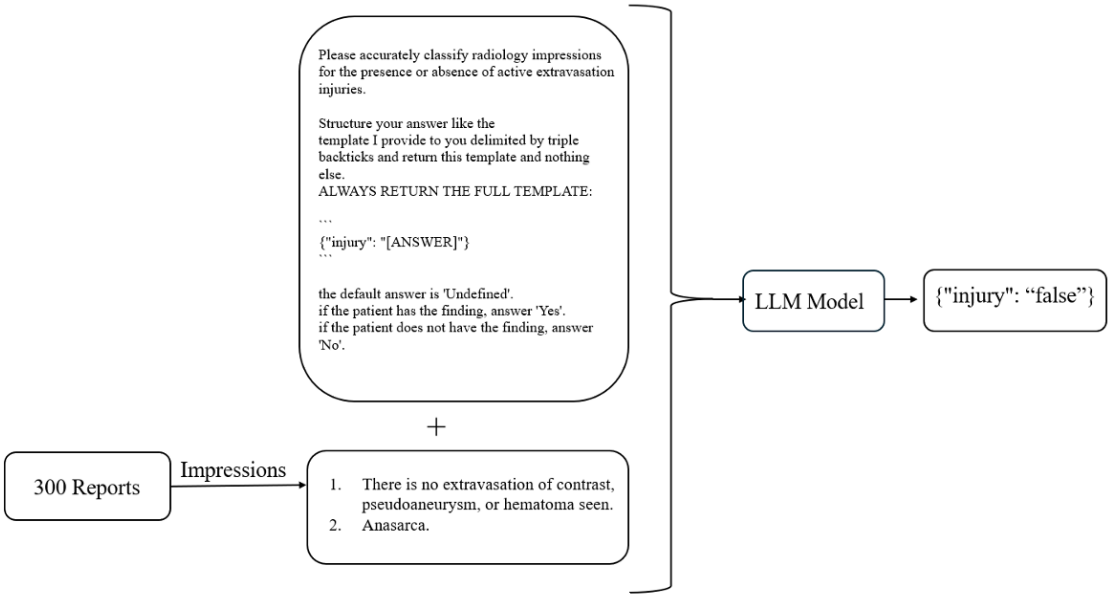

Table 1A. Precisions, Recall, F-score, and accuracy of Llama, Qwen, and Mistral, and ensemble model. **Table 1B.** F1-score of LLM models to localize where the active extravasation from the radiology report.

A.

Model	Precision	Recall	F1-score	Accuracy
Llama-3.1-70B	0.976	1	0.988	0.990
Qwen2.5-72B	0.983	0.951	0.967	0.973
Mistral-7B	0.981	0.854	0.913	0.932
Ensemble	0.983	0.967	0.975	0.980

B.

	Llama-3.1-70B	Qwen2.5-72B	Mistral-7B	Ensemble
Bowel	0.950	0.966	0.788	0.943
Liver	0.714	0.727	0.364	0.706
Spleen	1.00	.667	.250	.667
Abdominal wall	0.562	0.571	0.278	0.562
Kidney	0.76	0.705	0.429	0.667
Gluteal/Thighs	0.683	0.765	0.742	0.722
Retroperitoneal	0.344	0.347	0.140	0.348

Figure 1. Pipeline overview of prompting and output of the LLM model.

Keywords

Clinical Workflow & Productivity; Emerging Technologies; Imaging Research; Quality Improvement & Quality Assurance