LLM-Powered Radiology Impression Automation

By

Bimalsen Rajbhandari

Supervisor: Utku Pamuksuz

A Capstone Project

Submitted to the University of Chicago in partial fulfillment of the Requirements
for the degree of
Master of Science in Applied Data Science

Division of Physical Sciences
December 2024

Overview

This documentation outlines the steps taken during my capstone project, which involved developing a fine-tuned language model to generate radiology impressions from clinical and findings data. Below are the detailed steps and methodologies employed

Introduction

In a radiology report, the "Findings" section presents the observations made by the radiologist during the examination of the imaging studies. This section provides a detailed description of the abnormalities, if any, detected in the images. It typically includes information such as the size, location, shape, and characteristics of any lesions, masses, fractures, or other abnormalities found in the images. The "Impression" section offers a synthesized interpretation of these findings, aiding in clinical decision-making and patient management, which is why it is considered the most important part of the report.

This project proposes a methodological approach centered on fine-tuning a selection of open-sourced pre-trained Large Language Models (LLMs) using radiology reports sourced from the University of Chicago (UC) Medicine. The objective is to generate impressions of high quality that are factually accurate.

This project leverages open-source LLMs to ensure the accessibility of our research framework to a broad audience. We explore the utilization of LLMs with varying sizes and architectures, to strike a balance between computational efficiency and accuracy.

The adoption of LLMs promises manifold benefits. Firstly, it streamlines the radiology reporting process, enabling radiologists to redirect their attention towards direct patient care, thereby potentially expanding patient outreach. Additionally, by alleviating the burden associated with impression creation, the proposed methodology holds promise in mitigating radiologist burnout with its attendant negative ramifications.

Methodology

1. Data Preparation

- **Data Segmentation**: The dataset was divided based on different radiology modalities (e.g., MRI, CT, X-Ray).
- This ensured modality-specific characteristics were retained, enabling targeted analysis and modeling.

2. Exploratory Data Analysis (EDA)

- Focused EDA was performed on the MRI dataset to:
    - Identify common patterns in findings and impressions.

    - Understand the distribution of word count in findings, clinical information, and impressions
    - Detect potential outliers or inconsistencies in the dataset.
    - Handle missing values

3. Fine-Tuning Language Models

- Fine-tuned several pre-trained LLMs:
    - **Mistral Models**: 7B, 22B
    - **Llama Models**: 1B, 3B, 8B
- **Hyperparameter Tuning:** Experimented with over 50 variations of training parameters to optimize performance.
- Explored different input configurations:
    - **Findings Only**: Using findings as input.
    - **Findings + Clinical Information**: Using findings and clinical data as input.
- Also experimented with **instruction-tuning** to improve the models' ability to follow specific prompts effectively

4. Selecting the Best Models

- After extensive iterations, the 5 best models were selected based on multiple evaluation metrics:
    - **ROUGE Score**
    - **BLEU Score**
    - **BERT Score**
    - **Clinical BERT Similarity**
    - **RadBERT Similarity**
- All top models utilized **findings + clinical information** as input and generated impressions as output and had no instruction tuning.

5. Factual Correctness Evaluation

- Reviewed 100 random samples from the top models' outputs to evaluate factual correctness manually.
- Developed a rigorous factual correctness scoring method. In addition to main definition of factual correctness (stating a patient does not have cancer when the original impression indicated cancer)

- Penalized models for:
  - Adding extra clinical information not present in the findings.
  - Missing critical clinical information.

6. Fine-Tuning a Factual Correctness Model

- Used the manually scored 100 samples to fine-tune an additional LLM.
- Input: Original impression and generated impression.
- Output: Factual correctness metric.
- The fine-tuned model was then employed to generate factual correctness scores for all five top models.

7. Final Evaluation and Analysis

- Generated factual correctness scores for all five models using the fine-tuned LLM.
- Compiled a table comparing all five models' scores, revealing:
  - **Mistral 22B** achieved the best scores in terms of metrics but was computationally the most expensive.

8. Manual Comparison of Top Models

- Manually reviewed 50 random samples from Llama 3B and Mistral 7B, where they outperformed Mistral 22B:
  - **Mistral 7B and Llama 3B**: Performed better with longer original impressions, capturing detailed medical findings more effectively.
  - **Mistral 22B**: Provided more general overviews but occasionally lacked detail for comprehensive impressions.

Final Results

LLM	Input	Training Time	Generation Time	FactualCorrectness	(BERTScore Precision/Recall/F1)			ClinicalBERT Similarity	RadBERT Similarity
LLama 1B	Clinical + Findings	16min 27	1h 42min	5	0.66	0.64	0.64	0.89	0.94
LLama 3B	Clinical + Findings	37min	2h 57min	5	0.67	0.65	0.66	0.89	0.94
LLama 8B	Clinical + Findings	49min	2h 53min	5	0.70	0.68	0.68	0.90	0.95
Mistral 7B	Clinical + Findings	1h 16min	4h 3min	5	0.69	0.68	0.68	0.90	0.95
Mistral 22B	Clinical + Findings	2h 38min	10h 14min	6	0.70	0.69	0.69	0.90	0.95

LLM	ROUGE 1 Precision	ROUGE 1 Recall	ROUGE 1 F1	ROUGE 2 Precision	ROGUE 2 Recall	ROUGE 2 F1	ROUGE L Precision	ROUGE L Recall	ROUGE L F1	BLEU Score
LLama 1B	0.40	0.36	0.33	0.18	0.17	0.15	0.30	0.28	0.25	0.07
LLama 3B	0.43	0.39	0.37	0.21	0.19	0.18	0.32	0.30	0.28	0.08
LLama 8B	0.48	0.43	0.41	0.26	0.23	0.22	0.37	0.34	0.32	0.11
Mistral 7B	0.46	0.44	0.41	0.25	0.24	0.23	0.37	0.35	0.32	0.11
Mistral 22B	0.48	0.47	0.43	0.27	0.26	0.24	0.38	0.37	0.34	0.12

Key Findings

- **Mistral 22B**: Best-performing model by metrics but computationally expensive.
- **Mistral 7B, Llama 3B**: They have competitive scores, but are computationally more efficient and might be better for detailed and longer impressions.

Conclusion

The capstone project demonstrates the potential of fine-tuned LLMs in generating accurate and detailed radiology impressions, highlighting the nuanced performance of different LLMs across varying input and training configurations and evaluation metrics. While Mistral 22B demonstrated superior performance in terms of overall metrics, the specific strengths of other models, such as Mistral 7B's ability to handle detailed information, should be considered when selecting the most suitable model for a particular application.

Recommendations for Future Work

- Improve factual correctness generation methodology
- Train model to create personalized impression
- Explore more efficient model architectures