ARIA
APPLIED RESEARCH IN ACTION

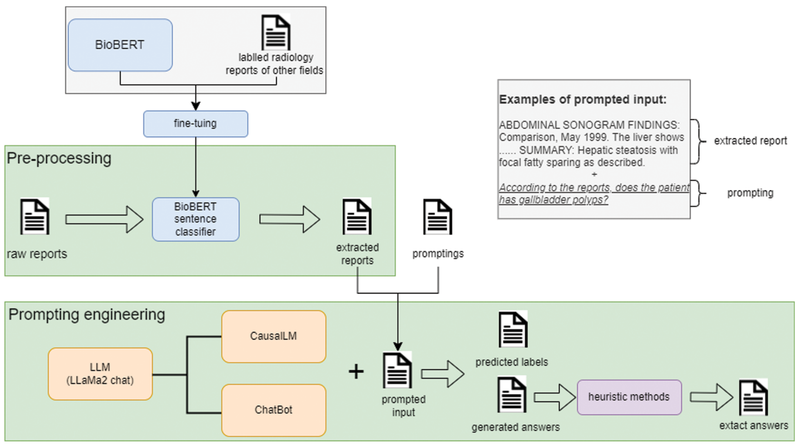
Unsupervised Data Extraction from Radiology Reports with LLMs

Building an unsupervised NLP pipeline to extract nuggets of data from a vast 20 year archive of unstructured radiology reports.

Haoran Sheng

Gerald Penn
ACADEMIC SUPERVISOR

Korosh Khalili and Anne Martel
INDUSTRY SUPERVISOR



PROJECT SUMMARY

Radiology reports are filled with data regarding diseases and incidental findings. Unfortunately, these data are locked in unstructured and non-standardized format/language. To extract from the data and/or label the data manually is time consuming and requires medical training. The purpose of this project was building an NLP pipeline that automatically extracts specified information from purely textual reports.

The pipeline was a combination of fine-tuned BERT, prompt engineering and heuristic methods. Unsupervised methods were developed because of the lack of labels. Due to the diversity of medical terms' descriptions, the model needed to model rich semantic meanings. Thus, the pipeline heavily relied on the outstanding capacity of large language models to decipher everyday and medical semantics.

At the beginning of the pipeline, the original text was processed by fine-tuned BioBERT to extract related sentences. Subsequently, an open-source LLM (LLaMa2) was used to process the selected sentences with the help of prompt-engineering methods, following heuristic methods to get exact results. The LLM was used as a casual language model or chatbot and deployed locally.

Experiments on our radiology reports showed that our method, despite the lack of field-specified knowledge, achieved a quite usable accuracy of higher than 90% on most tasks.

UHN
Toronto General
Toronto Western
Princess Margaret
Toronto Rehab
Michener Institute

Computer Science
UNIVERSITY OF TORONTO
Master of Science in Applied Computing