Plan for the Sarcoma Data Extraction Project

Naihe Xiao

1 Background

In the realm of medical analysis, manually identifying and extracting the dimensions of sarcomas (tumours) in patients from medical reports emerges as a laborious and time-consuming task. Addressing this challenge, this project is unwaveringly dedicated to the automation of tumour size extraction from medical reports. By harnessing the capabilities of advanced large language models, the project aspires to ensure a more efficient and streamlined analysis of sarcoma data, thereby augmenting the speed and accuracy of diagnosis and treatment planning.

2 Previous Achievements

A prior research team, led by Conor and Yucheng, has explored various strategies to address the issues associated with manually identifying and extracting tumour sizes from medical reports. These strategies included employing the LSTM (long short-term memory networks) model and developing two distinct Named Entity Recognition (NER) models: Yucheng's model and the Stanza NER model. Despite these commendable efforts, numerous challenges still persist. These include the unintended extraction of non-tumour sizes and the occurrence of hallucinations, posing significant obstacles to achieving fully automated and accurate extraction of tumour sizes from medical reports.

3 Aims

3.1 Primary Aims

1. To update and use the latest Vicuna model to better pull tumour sizes from medical reports. This involves tweaking prompts through several tries and might mean re-training the last model layer or adding and training a new one.

2. To test how well the model works on unlabelled observations. This means picking some reports randomly from the dataset and figuring out a solid way to label them. Right now, we're thinking about using ChatGPT-4 to process the data directly.

3.2 Secondary Aims

1. To refine prompts for each report type through trial and error.
2. To figure out the optimal tuning for the Vicuna model.
3. To check if a patient has at least one tumour.
4. To determine the most accurate method for categorizing report types.