Leveraging Large Language Models and Transformer Architectures for Data Extraction from Unstructured Clinical Notes: Experiences and Challenges

Soujanya Samineni*, Michael Rosenthal*, Travis Zack**
Dana-Farber Cancer Institute (DFCI)*, UCSF**
soujanya_samineni@dfci.harvard.edu, michael_rosenthal@dfci.harvard.edu,
travis.zack@ucsf.edu

Introduction:

In the realm of healthcare, extracting structured data from unstructured clinical notes is essential for various tasks, including clinical decision support, research, and quality improvement initiatives. Large Language Models (LLMs) and Transformer architectures have emerged as promising tools for this purpose, offering the ability to understand and process natural language text effectively. This abstract explores our experiences and the challenges encountered when using LLMs versus custom Transformer models for data extraction tasks in clinical settings.

Experiences:

LLMs, such as ChatGPT and Mistral, have demonstrated remarkable performance in capturing contextual information and extracting key data elements from unstructured clinical notes. These models offer the advantage of pre-trained language representations and can be fine-tuned on domain-specific datasets to further enhance their performance. Prompt engineering plays a pivotal role in optimizing the performance of LLMs for clinical data extraction tasks. By strategically designing prompts, we could guide the models to focus on specific information, thereby improving extraction accuracy. In contrast, custom Transformer architectures provide flexibility and scalability for designing task-specific models tailored to the unique requirements of clinical data extraction tasks. In one case, the LLM identified previously overlooked patterns in patient histories that correlated with early pancreatic cancer, leading to a 20% improvement in early symptom extraction.

Challenges:

Clinical data presents unique challenges compared to public datasets, as there are no set rules for sectioning notes, and notes often contain copy forwards and repeated text. Moreover, symptoms or clinical conditions may be mentioned in multiple sections, such as the History of Present Illness (HPI) and medication records, leading to ambiguity. LLMs address these challenges by employing context reasoning via prompt-based approaches, enabling them to accurately identify and extract relevant information from clinical notes.

Conclusion:

Leveraging LLMs and Transformer architectures for data extraction from unstructured clinical notes presents exciting opportunities but also entails significant challenges. Successful application of these models in clinical settings requires careful consideration of the unique characteristics of clinical data, including variability in note formatting and the presence of repeated text. Addressing these challenges through context reasoning and prompt-based approaches will be essential for advancing data extraction capabilities and unlocking the full potential of natural language processing in healthcare applications.