NYULangone at Chemotimelines 2024: Utilizing Open-Weights Large Language Models for Chemotherapy Event Extraction

Jeff Zhang[1], Yindalon Aphinyanaphongs[1], Anthony Cardillo[1],
[1]NYU Langone Health, MCIT Department of Health Informatics,
Correspondence: Yin.A@nyulangone.org

Abstract

The extraction of chemotherapy treatment timelines from clinical narratives poses significant challenges due to the complexity of medical language and patient-specific treatment regimens. This paper describes the NYULangone team's approach to Subtask 2 of the Chemotimelines 2024 shared task, focusing on leveraging a locally hosted Large Language Model (LLM), Mixtral 8x7B (MistralAI, France), to interpret and extract relevant events from clinical notes without relying on domain-specific training data. Despite facing challenges due to the task's complexity and the current capacity of open-source AI, our methodology highlights the future potential of local foundational LLMs in specialized domains like biomedical data processing.

1 Introduction

The extraction of structured information from unstructured clinical narratives is a crucial task in healthcare informatics, enabling better patient care and clinical decision-making. The Chemotimelines 2024 shared task focuses on extracting chemotherapy treatment timelines from clinical narratives, a challenging task for understanding oncology patients' treatment paths. Our team, NYULangone, participated in Subtask 2, aiming to leverage the general reasoning capabilities of large language models (LLMs) for this purpose.

2 Related Work

Clinical narrative processing traditionally relies on rule-based systems or machine learning models trained on domain-specific annotated data. Recent advances in NLP have seen the rise of transformer-based models and LLMs, offering powerful general-purpose language understanding capabilities. However, their application in domain-specific tasks like chemotherapy timeline extraction remains in the infancy of exploration.

3 System Description

Our system builds upon a locally deployed instance of Mixtral, an open-weights LLM. The system comprises two rounds of text inference: the first round is an extraction of chemotherapy events from individual notes, and the second round is the aggregation of events from multiple notes to a single timeline.

Algorithm 1 Patient Chemotherapy Summary Algorithm

```
for each patient do
    for each note of the patient do
        Prompt Mixtral to read the note and extract chemotherapies
    end for
end for
Prompt Mixtral to combine the extracted chemotherapies from every note to create a patient-level summary of all chemotherapies
```

3.1 Architecture

We employed Mixtral 8x7B v0.1, an open-weights LLM originally published by Mistral AI in December 2023. The system leverages its pretrained weights without further domain-specific fine-tuning. The system processes clinical narratives as raw text files, uses the LLM to extract relevant events and dates, and structures them into the required JSON format for output.

3.2 Implementation

The system was hosted on NYU Langone's high-performance cluster "Ultraviolet." Using SLURM, a compute instance was requisitioned using three NVIDIA A100s with 128GB of system RAM. The model weights for Mixtral 8x7B were downloaded from Hugging Face, and inference was performed with the Transformers library for Python.

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 428–430

3.3 Prompts

For the first inference used to extract chemotherapy events from notes, we used the following Markdown-style prompt:

> [INST] **GOAL and PURPOSE:** You are an experienced medical annotator with special expertise in natural language processing of oncology documents. You will be given a list of JSON objects to turn into a list of lists.
>
> **INSTRUCTIONS:** Read the patient's note in its entirety, given in the section "# PATIENT NOTE" below. Use THYME guidelines to create "events"; every mention of a chemotherapeutic drug or component should have: the name of the drug, an associated date, the temporal_relation between the use of that drug and the associated date. Each event must be in the form ['chemo drug name', 'temporal_relation', 'YYYY-MM-DD']. If a drug is associated with multiple dates, or a date is associated with multiple drugs, break them into separate events. 'temporal_relation' must be one of ["contains-1", "begins-on", "ends-on", "before"].
>
> **EXAMPLES:** ['herceptin', 'begins-on', '2013-06-17'], ['taxol', 'contains-1', '2013-09']
>
> **OUTPUT:** Use only well-formatted JSON. Only output the timeline of chemotherapy events; place it under "# TIMELINE". Do not make any additional notes or comments, only JSON under "# TIMELINE". [/INST] **PATIENT NOTE** <insert patient note here> **TIMELINE**

This first inference accomplishes the extraction of each chemotherapy event in each note. However, the events are not organized by patient yet. For the second inference used to aggregate chemotherapy events from multiple notes into patient timelines, we used the following prompt:

> [INST] **GOAL and PURPOSE:** You are an experienced medical annotator with special expertise in natural language processing of oncology documents. You will be given a JSON list of lists. Your job is to output a list of lists for each patient.
>
> **EXAMPLE OUTPUT:**
>
> ```
> patient_01:
> ['taxol', 'begins-on', '2013-06-17']
> ['taxol', 'ends-on', '2013-09']
> ...
> patient_02:
> ```
>
> [/INST]

4 Results

On the dev set, our system achieved an average F1 score of 0.35. On the validation set, our system achieved an average F1 score of 0.23 across different cancer types, as shown in Table 2 of the competition results.

5 Discussion

While our performance was well below the baseline and leading teams, it provided valuable insights into the challenges and potential of using locally hosted LLMs in clinical NLP tasks without domain-specific training.

The opaque inner workings of LLMs preclude an exact understanding of why certain chemotherapy events are more easily extracted than others. The errors our system demonstrates could largely be grouped into several types:

- Confabulation of drugs not mentioned (e.g. extracting "herceptin" from a patient radiology report without any mention of chemotherapy)
- Inclusion of non-chemotherapeutic drugs, especially steroids (e.g. extracting "prednisone" for a patient on immunosuppression)
- Failure to include clearly mentioned drugs (e.g. failing to extract "aflibercept" when it was well documented in a patient note)

Despite the objectively poor performance, our results highlight a future potential for LLMs to be used in biomedical NLP tasks. Local LLMs that can competently perform general reasoning are still a new technology, with expert opinion suggesting that local models like Mixtral currently perform at a GPT-3 (OpenAI, United States) level of performance.

6 Conclusion and Future Work

Our exploration into using local LLMs for chemotherapy treatment timelines extraction offers a starting point for further research in this area. Future work will focus on enhancing model understanding of clinical contexts through retrieval augmented generation (RAG) and ensemble prompting techniques such as "tree of thought."

7 Acknowledgments

We thank NYU Langone Health for providing the computational resources for this project.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Tom Brown and et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jacob Devlin and et al. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yonghui Zhang, Gerardo Flores, Gavin E Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H Shah, Atul J Butte, Michael D Howell, Claire Cui, Greg S Corrado, and Jeff Dean. 2018. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Ashish Vaswani and et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Junyuan Yao, Harry Hochheiser, Woosub Yoon, Erin Goldner, and Guergana Savova. 2024. Overview of the 2024 shared task on chemotherapy treatment timeline extraction. In *Proceedings of the 6th Clinical Natural Language Processing Workshop, NAACL June 2024*, Mexico City, Mexico.