Lexicans at Chemotimelines 2024: Chemotimeline Chronicles - Leveraging Large Language Models (LLMs) for Temporal Relations Extraction in Oncological Electronic Health Records

Vishakha Sharma[1], Andres Fernandez[2], Andrei Ioanovici[2], David Talby[2], Frederik Buijs[3]
[1]Roche Diagnostics, California, USA
[2]John Snow Labs, Delaware, USA
[3]F. Hoffmann-La Roche Ltd, Basel, Switzerland

Abstract

Automatic generation of chemotherapy treatment timelines from electronic health records (EHRs) notes not only streamlines clinical workflows but also promotes better coordination and improvements in cancer treatment and quality of care. This paper describes the submission to the Chemotimelines 2024 shared task that aims to automatically build a chemotherapy treatment timeline for each patient using their complete set of EHR notes, spanning various sources such as primary care provider, oncology, discharge summaries, emergency department, pathology, radiology, and more. We report results from two large language models (LLMs), namely Llama 2 and Mistral 7B, applied to the shared task data using zero-shot prompting.

1 Introduction

Electronic Health Records (EHRs) are a rich repository of patient information, encompassing a wide array of formats and sources including physician notes, laboratory results, radiology images, and pathology reports. Due to the heterogeneous and unstructured nature of clinical data, it is cumbersome to visualize patient journeys or extract meaningful information from Electronic Health Records (EHRs) to help guide clinical decision making (Anand and Sadhna, 2023; Najafabadipour et al., 2020). EHR data is often dispersed, recorded in free text with substantial variability in terminology, and embedded in narrative formats that are not easy to process or normalize across healthcare settings and systems. In addition, privacy concerns further limit the use of clinical data across hospitals and geographical borders further compounding complexity (Reisman, 2017; Kehl et al., 2020; Levine et al., 2019; Banerjee et al., 2019) and difficulty to leverage EHR data for insights generation.

Large Language Models (LLMs), with their advanced natural language (Guevara et al., 2024; Chen et al., 2023a,c; Hochheiser et al., 2023; Bitterman et al., 2023) understanding capabilities, offer a transformative solution to these challenges. They can be trained to interpret complex language found in EHRs, extracting relevant clinical events and concepts, and mapping these onto a coherent information or treatment timelines which can be difficult to realize manually by humans. LLMs are appropriate for handling the variability and ambiguity that arise in medical documentation, enabling them to identify and organize critical information such as chemotherapy treatments, such as drug names, dosages, administration dates, and associated clinical outcomes (Jahan et al., 2024).

Moreover, by leveraging the latest advancements in transfer learning and domain-specific fine-tuning, LLMs can be programmed in such a way to understand the specific lexicon and data structures unique to domains as complex as oncology and chemotherapy treatment regimes (Chen et al., 2023b).

All in all, this can help with the generation of comprehensive, accurate, and personalized chemotherapy treatment timelines that are an essential component for advancing precision oncology, and also supporting the development and assessment of patient-centric therapeutic strategies (Levine et al., 2019; Banerjee et al., 2019).

To better understand the impact of various factors on tumor behavior and responsiveness, particularly in the context of precision oncology, the Chemotimelines 2024 shared tasks has been proposed (Yao et al., 2024). In this work, we describe our submission to Subtask 1, which aims to build timelines of chemotherapy treatments for individual patients using their Electronic Health Records (EHR) notes. We achieved a 5th place ranking in Subtask 1, with an averaged accuracy across breast, ovarian, and melanoma indications.

The contributions of our paper can be outlined as follows:

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 394–405
June 21, 2024 ©2024 Association for Computational Linguistics

1. We developed a Large Language Model (LLM)-based system customized for description and exploration, providing substantial value in tasks related to natural language understanding.

2. We employed multiple LLMs and prompts across diverse development and training datasets, our approach aimed to improve performance and enhance generalization.

3. We introduced a framework that presents a modular strategy for zero-shot relation extraction, leveraging well-established LLMs.

2 Related Work

In recent years, there has been a growing body of research demonstrating the effectiveness of Large Language Models (LLMs) in comprehending medical text data and extracting valuable insights from Electronic Health Records (EHRs) across various clinical domains (Beam et al., 2019; Van Veen et al., 2024; Wong et al., 2023; Eriksen and Ryg, 2023). Prior investigations have shown the application of Natural Language Processing (NLP) in healthcare, encompassing tasks such as clinical text classification, medical entity recognition, and patient risk prediction. Efforts to construct clinical timelines from EHR data have predominantly focused on structured data such as procedure codes, diagnosis codes, and laboratory results (Rajkomar et al., 2018; Mullenbach et al., 2018).

Within oncology, NLP methodologies have been employed to analyze cancer-related textual data, including pathology reports, clinical notes, and research articles (Bodenreider, 2004; Meystre et al., 2008). Researchers have investigated the utility of NLP in extracting treatment regimens, identifying adverse drug events, and predicting treatment outcomes among cancer patients (Savova et al., 2010; Xu et al., 2019). Techniques for temporal event extraction and sequence modeling have been explored extensively to develop patient timelines for disease progression tracking and treatment monitoring (Ebadi et al., 2021). Temporal reasoning techniques have found applications in healthcare to analyze the temporal associations between clinical events, treatments, and patient outcomes (Sun et al., 2013). Studies have explored temporal logic, temporal abstraction, and probabilistic models to represent and analyze temporal data in healthcare contexts (Orphanou et al., 2014).

Transformer based large language models have demonstrated remarkable performance improvements across various NLP benchmarks (Devlin et al., 2018; Chiu and Nichols, 2016). Furthermore, healthcare-specific models (Lee et al., 2020) have exhibited state-of-the-art accuracy in biomedical entity recognition (Kocaman and Talby, 2020) and relation extraction (Kocaman and Talby, 2021).

The current state of the art lies in several notable Large Language Models (LLMs), each featuring distinct model architectures and sizes (Pan et al., 2024). Prominent examples include Llama 2 (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), Zephyr (Tunstall et al., 2023), MEDITRON-70B (Chen et al., 2023d), and Mixtral of Experts (Jiang et al., 2024). LLMs possess the capability to analyze extensive textual data, and the task of summarizing crucial information from electronic health records (Van Veen et al., 2024) can significantly impact how clinicians manage their time, enabling them to dedicate more time to interacting with patients (Khairat et al., 2018) and improve quality of care.

3 Task and Dataset Details

3.1 Task Details

Chemotimelines 2024 at NAACL-ClinicalNLP Workshop is a shared task (Yao et al., 2024) that focuses on building a timeline of chemotherapy treatment for each patient given all the available Electronic Health Records (EHRs) notes of that patient. The shared task has 2 subtasks. Subtask 1 involves using provided gold annotations of chemotherapy events (EVENTs) and time expressions (TIMEX3s) along with Electronic Health Record (EHRs) notes to predict temporal relations between them and generate patient-level timelines. This task requires deduplicating and resolving conflicts in pairwise temporal relations, with the option to derive timelines without relying on pairwise relations. Additionally, attributes such as modality and relation to document creation time are included. Subtask 2 entails building an end-to-end system for chemotherapy timeline extraction using only patient EHR notes. Both subtasks are evaluated against gold patient-level timelines. We submitted the results of Subtask 1. The submission scripts for evaluation can be found here[1].

[1] https://github.com/HealthNLPorg/chemoTimelinesEval

Indications (Train, Dev and Test)	**# Patients**	**# Reports**	**# Entities**	**# Relations**
Breast (Train)	23	236	1599	455
Melanoma (Train)	3	32	225	48
Ovarian (Train)	14	273	1765	494
Breast (Dev)	10	61	425	113
Melanoma (Dev)	2	99	1050	20
Ovarian (Dev)	8	138	1102	226
Breast (Test)	25	379	3678	0
Melanoma (Test)	4	77	591	0
Ovarian (Test)	6	143	1045	0
Total	95	1438	11480	1356

Table 1: Summary of Dataset Statistics: Number of Patients, Reports, Entities, and Relations across Training (Train), Development (Dev) and Testing (Test) Sets for different indications (Breast, Melanoma and Ovarian).

Indications (Train and Dev)	**BEGINS-ON**	**CONTAINS**	**ENDS-ON**
Breast (Train)	131	298	26
Melanoma (Train)	10	37	1
Ovarian (Train)	101	327	66
Breast (Dev)	27	57	29
Melanoma (Dev)	42	157	2
Ovarian (Dev)	34	140	52

Table 2: Summary of Dataset Statistics: Indications (breast, melanoma, and ovarian) across training and development sets, including the three types of temporal relations.

3.2 Dataset

The dataset comprises 95 patients with 1438 reports. Table 1 summarizes dataset statistics, including indications (breast, melanoma and ovarian) for training, development, and testing sets, along with the number of patients, reports, entities, and relations. The annotated dataset has been using THYME ontology (Styler IV et al., 2014) and temporal relation annotations (Wright-Bettner et al., 2020) with three different temporal relations used for `TLINKs` (temporal links): `BEGINS-ON`, `CONTAINS` and `ENDS-ON`. Table 2 presents summarized statistics for indications (breast, melanoma, and ovarian) across training and development sets, including the three types of temporal relations.

4 Approach

We aimed to significantly contribute to the development of advanced cutting-edge methodologies and techniques for automatically constructing chemotherapy treatment timelines from Electronic Health Records (EHRs) clinical notes of individual patients. We leveraged current state of the art Large Language models (LLMs) for this shared task. We tested various LLMs with different sizes and architectures to determine which model works best for relation extraction (See Figure 1).

4.1 Natural Language Processing (NLP) Pipeline with Language Representations

4.1.1 Document Chunking

We divided the documents into paragraphs or groups of paragraphs (sections) to facilitate manageable processing units.

Sequence Length The experiments involved evaluating various sequence lengths, which determine the number of words or tokens processed by the model at once. Assessing lengths of 1024, 512, and 256 tokens provides insights into how input length impacts the system's accuracy in extracting relations.

Paragraph and Sentence Detection Paragraph Detection NLP plays a crucial role in enhancing contextual understanding within Electronic Health Records (EHRs) by segmenting the text into meaningful units. By identifying paragraphs, NLP models can discern distinct sections of the EHRs, such as patient history, symptoms, diagnoses, and treatment plans. This segmentation enables the model

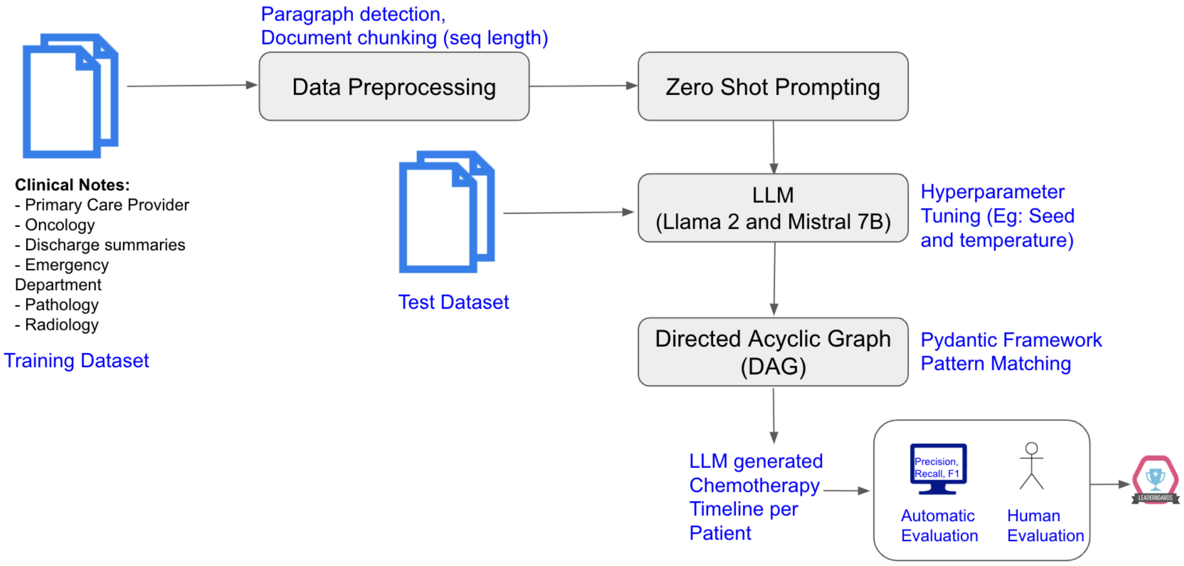


Figure 1: NLP Pipeline for Subtask 1

to focus on specific aspects of the patient's medical information, facilitating more accurate analysis and interpretation. We have incorporated section chunking and paragraph detection techniques into our system. This involves identifying individual sentences within the text data. By isolating paragraphs, the system can focus on extracting relations specifically from relevant pairs of entities within each paragraph, which enhances precision. We have incorporated section chunking and paragraph detection techniques into our system. This involves identifying individual sentences and paragraphs within the text data. By isolating paragraphs, the system can focus on extracting relations specifically from relevant pairs of entities within each paragraph (Kocaman and Talby, 2020), which enhances precision.

- In terms of extracting relations from various document paragraphs, our sequences already extend beyond a single paragraph, as our sequence length is configured to accommodate 256 tokens. Nonetheless, such occurrences are rare within this dataset. If necessary, we can concatenate adjacent or contiguous paragraphs or clusters of paragraphs to enable the extraction of relations spanning multiple paragraphs.

- To address chunking concerns, we implemented an overlap parameter for enhanced performance. This parameter prevents the inadvertent separation of essential information by preserving sentence integrity, even without overlap. It facilitates the reconciliation of fragmented data, mitigating the risk of context loss and preserving predictive accuracy. The risk of reduced recall arises from potential pairs not being prompted for relation classification. Encouragingly, the model's metrics exhibit no specific recall-related issues, signaling positive performance in this regard.

4.1.2 Zero-Shot Prompting for Related Pairs

We developed structured prompts to guide the system in identifying and extracting relations between pairs of entities. These prompts serve as cues for the system to recognize and analyze relevant information in the text pertaining to the specified entities (See Figure 2).

This process involved prompt engineering techniques aimed at refining the instructions within the relation extraction pipeline, optimizing them to extract more precise and relevant information during subsequent stages. The zero-shot prompt gave us a reasonably high precision by leveraging the prompt templates that guided the LLMs to generate responses that closely match the desired output without requiring explicit training data for each class or category. Prompt 1 was used for the submission and for evaluation. We tried Prompt 2 but encountered challenges in labeling the relations from distinct lists. (See Figure 2)

The input to the LLMs involves combining the prompt with the paragraph or groups of paragraphs (sections).

4.1.3 Tokenization and Embedding

Each paragraph is tokenized, and the tokens are encoded using the tokenizer specific to the chosen Large Language Model (LLM).

4.1.4 Embedding Decoding

The encoded tokens were fed to the LLM, resulting in serialized outputs.

4.1.5 Semantic Object Construction

Using the outputs from the LLMs and predefined validation classes for each of the prompts, we construct semantic-rich objects that encapsulate the information extracted from the text.

Directed Acyclic Graph (DAG) We constructed a simplified DAG to outline the logical framework guiding the construction of the output taxonomy, enabling a structured representation of the reasoning process. (See Figure 1)

After establishing the relations with the output from the LLMs, we leveraged Pydantic (Colvin and contributors, 2024), a Python library for data validation and settings management. Pydantic (Colvin and contributors, 2024) facilitates data parsing and validation, ensuring that the data adheres to the expected types specified using Python's standard type hints. A Directed Acyclic Graph (DAG) can impact model accuracy positively by ensuring that validation functions are executed in a specific, predictable order. This helped in maintaining data integrity and correctness, thereby reducing the likelihood of errors or inconsistencies in the model's predictions. Additionally, DAGs prevent cyclic dependencies, which led to more stable and reliable model behavior.

Date Normalization We normalized the data using both the natural language representation for the temporal entity and the document time as a reference. We then transformed the temporal entity to an absolute datetime.

- The date normalization process is integrated into the validation procedure through a dedicated class field validator. It involves multiple steps to handle various date formats and potential failure scenarios, including cases where external services like Duckling (Rasa, 2024) may not parse the input successfully.

- Initially, the validator attempts to parse the raw date string using the `parse_timex` function, which sends the string to Duckling (Rasa, 2024) and, if unsuccessful, to the SparkNLP date normalizer annotator (John Snow Labs, 2024). These tools excel at interpreting natural language and complex date expressions, providing robust initial parsing. If successful, the parsed value undergoes further processing with `dateutil` to ensure compatibility with Python's datetime object format.

- In case of failure with Duckling (Rasa, 2024) and SparkNLP parsing (John Snow Labs, 2024), the validator employs fallback strategies. It checks for ISO week format dates and `year-month-only` strings, attempting to convert them into complete dates. If these strategies fail, the validator employs a battery of parsers (e.g., `dateutil_parser.parse`, `pd.to_datetime`, `arrow.get`) in a loop until successful parsing occurs.

- Throughout the process, detailed logging captures various states and errors, aiding in debugging and understanding parsing issues. Finally, if a valid date is obtained through any of these strategies, it is stored as the normalized value in the model, which may represent a full date or just the year and month, depending on the input string and specified context.

4.1.6 Serialization for Submission

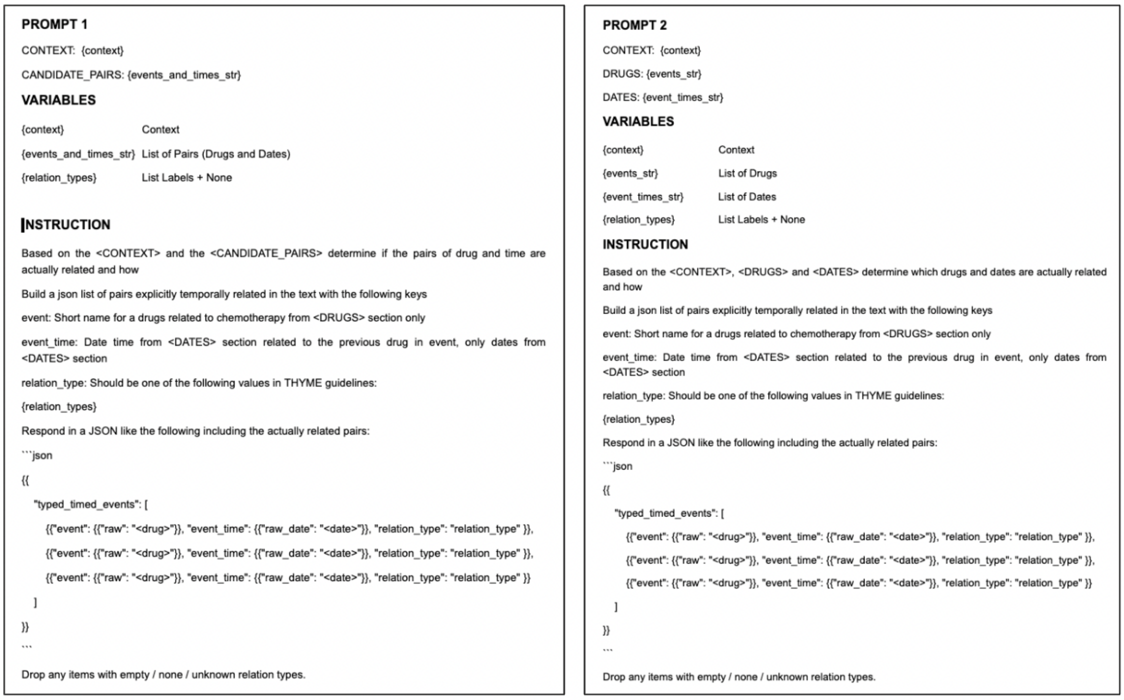
Finally, we aggregated and serialized these semantic objects into the submission format specified by the competition guidelines.

4.2 Baseline Models

We fine-tined pre-trained Llama 2 (Touvron et al., 2023) and Mistral 7B (Jiang et al., 2023) for our submission to this shared task.

Llama 2 Llama 2 (Touvron et al., 2023) is a collection of large language models (LLMs) ranging from 7 billion to 70 billion parameters. They are fine-tuned LLMs optimized for dialogue applications.

Mistal 7B Mistral 7B (Jiang et al., 2023) is a language model consisting of 7 billion parameters designed to deliver superior performance and efficiency. Mistral 7B demonstrates superior performance compared to the best open 13B model (Llama 2) (Touvron et al., 2023) across all assessed benchmarks and outperforms the leading released



~~~~
PROMPT 1

CONTEXT:  {context}

CANDIDATE_PAIRS: {events_and_times_str}

VARIABLES

{context}                 Context
{events_and_times_str}    List of Pairs (Drugs and Dates)
{relation_types}          List Labels + None

INSTRUCTION

Based on the <CONTEXT> and the <CANDIDATE_PAIRS> determine if the pairs of drug and time are actually related and how

Build a json list of pairs explicitly temporally related in the text with the following keys

event: Short name for a drugs related to chemotherapy from <DRUGS> section only

event_time: Date time from <DATES> section related to the previous drug in event, only dates from <DATES> section

relation_type: Should be one of the following values in THYME guidelines:

{relation_types}

Respond in a JSON like the following including the actually related pairs:

```json

{{

   "typed_timed_events": [

      {{"event": {{"raw": "<drug>"}}, "event_time": {{"raw_date": "<date>"}}, "relation_type": "relation_type" }},

      {{"event": {{"raw": "<drug>"}}, "event_time": {{"raw_date": "<date>"}}, "relation_type": "relation_type" }},

      {{"event": {{"raw": "<drug>"}}, "event_time": {{"raw_date": "<date>"}}, "relation_type": "relation_type" }}

   ]

}}
```

Drop any items with empty / none / unknown relation types.
~~~~

~~~~
PROMPT 2

CONTEXT:  {context}

DRUGS: {events_str}

DATES: {event_times_str}

VARIABLES

{context}            Context
{events_str}         List of Drugs
{event_times_str}    List of Dates
{relation_types}     List Labels + None

INSTRUCTION

Based on the <CONTEXT>, <DRUGS> and <DATES> determine which drugs and dates are actually related and how

Build a json list of pairs explicitly temporally related in the text with the following keys

event: Short name for a drugs related to chemotherapy from <DRUGS> section only

event_time: Date time from <DATES> section related to the previous drug in event, only dates from <DATES> section

relation_type: Should be one of the following values in THYME guidelines:

{relation_types}

Respond in a JSON like the following including the actually related pairs:

```json

{{

   "typed_timed_events": [

      {{"event": {{"raw": "<drug>"}}, "event_time": {{"raw_date": "<date>"}}, "relation_type": "relation_type" }},

      {{"event": {{"raw": "<drug>"}}, "event_time": {{"raw_date": "<date>"}}, "relation_type": "relation_type" }},

      {{"event": {{"raw": "<drug>"}}, "event_time": {{"raw_date": "<date>"}}, "relation_type": "relation_type" }}

   ]

}}
```

Drop any items with empty / none / unknown relation types.
~~~~

Figure 2: Prompt template: Prompt 1 for the relation label from the pairs (left) and Prompt 2 for the relation label from the separate lists of drugs and dates (right).

34B model (Llama 1) in tasks such as reasoning, mathematics, and code generation.

4.2.1 Validation and Quality of LLM Response

We rewrote the THYME ontology on top of a typed validation framework based on (pydantic (Colvin and contributors, 2024)) library. We binded every result from a prompted task to one of these objects: Thyme; in Subtask 1 the validation class for the LLM response is the graph representation defined in `TypedTimedEvents:List[Tuple[Event, Timex, TLinkType]]`. We forced the output from the LLM to conform to this type, and if not we kept refining the prompt. After obtaining accurately processed outputs from the LLM, the next step involved aggregation. This entails concatenating the parsed subgraphs from each chunk of the LLM output into a deduplicated timeline at the patient level.

During the inference phase, we focused on the post processing techniques, such as parsing and refining, applied to the output generated by the Large Language Models (LLMs). These techniques aim to enhance the quality and accuracy of the extracted information, ensuring its suitability for downstream analysis and applications.

4.3 Evaluation Metrics

Models were evaluated with the official evaluation script[2] on the test set. The following metrics were used: Precision, Recall and F-score (Hossin and Sulaiman, 2015). We reported performance as the arithmetic mean of F-score.

4.4 Human Evaluation

In our study, two medical professionals conducted a comparative analysis of chemotherapy timelines generated by LLMs, specifically using the Llama 2 model for our initial submission, against a ground truth established by the dataset (train and dev set) provided by the challenge. The dataset combines training, development, and testing sets, encompassing a total of ninety five (n=95) patients. Train and dev set contain sixty patients (n=60) patients and the test set contains thirty five (n=35) patients. The gold standard for the test set of thirty five (n=35) patients was not released. Therefore, the two medical professionals randomly selected five patients (n = 5) from each indication (breast, melanoma and ovarian) and manually reviewed the predictions generated by the LLMs, performed a qualitative evaluation.

The LLMs demonstrated a tendency to misclas-

[2] https://github.com/HealthNLPorg/chemoTimelinesEval

Indications (Train and Dev)	Baseline	Predictions	Llama 2	Mistral 7B
Breast (Train)	0.427713	0.800827	0.695125	0.606543
Breast (Dev)	**0.863988**	**0.888878**	0.768916	0.723611
Melanoma (Dev)	0.455782	0.797009	0.633271	0.767574
Melanoma (Train)	0.765196	0.842803	**0.882037**	**0.799432**
Ovarian (Dev)	0.715926	0.607934	0.561085	0.625625
Ovarian (Train)	0.715137	0.816064	0.647571	0.595842

Table 3: Performance on relation extraction by approach.

Parameters	Value	Description
Chunk Size	256	number of tokens or words processed at a time during training or inference
Temperature	0.1	controls the randomness of the generated output
Seed	123	predefined starting point for the random no. generator used during training

Table 4: Hyperparameters used with LLama 2 (Touvron et al., 2023) and Mistral 7B (Jiang et al., 2023) for the Chemotimelines 2024 Subtask 1.

sify CONTAINS relation over the BEGINS-ON and ENDS-ON, resulting in low recall for BEGINS-ON and ENDS-ON, and low precision for CONTAINS. For instance, in one case, where the actual relationship indicated Taxotere ENDS-ON at a specific date, the model incorrectly predicted it as a CONTAINS relation.

Another noteworthy observation was the occasional complete oversight of a relation by the LLMs. Additionally, discrepancies arose when the year was occasionally misinterpreted as a future date due to errors in the dates mentioned in the reports.

5 Results and Discussion

As previously stated, our study utilizes the two large language models, Llama 2 (Touvron et al., 2023) and Mistral 7B (Jiang et al., 2023). Table 3 shows the performance metrics on the relation extraction NLP task for the training and development set across three indications (Breast, Melanoma and Ovarian). We evaluated the performance of both Llama 2 and Mistral 7B against the baseline. Notably, we attained the highest performance on the Melanoma training set with both Llama 2 and Mistral 7B.

We utilized the default parameters for both Llama 2 and Mistral 7B, except for the chunk size, which was set to 256, temperature set to 0.1, and seed set to 123 (Refer to Table 4). Chunk size refers to the number of tokens or words processed at a time during training or inference. Temperature regulates the randomness of the generated output, while the seed serves as the predefined starting point for the random number generator used during model training.

Table 5 illustrates the results of three test data runs utilizing Llama 2 and Mistral 7B for Subtask 1. Our highest performing model was Llama 2, achieving an F1 average score of 0.71, while Mistral 7B attained an average F1 of 0.61. Llama 2 exhibited superior performance compared to Mistral 7B, resulting in a higher rank. Specifically, Llama 2 secured the 5th position in the average score for Subtask 1, the 4th position for the Melanoma indication, and the 7th position for Breast and Ovarian indications.

Error Analysis Error analysis in Large Language Models (LLMs) involves scrutinizing the model's prediction errors to discern their types, frequency, and underlying causes. This entails evaluating the model's performance on a test dataset and categorizing errors into various types, including false positives, false negatives, ambiguous cases, out-of-distribution errors, and conceptual errors. By analyzing these errors, insights can be gleaned regarding patterns and areas for improvement in the model. This analysis guides strategies for enhancing the model's performance through fine-tuning, refining training data, and optimizing input representations. Furthermore, error analysis is crucial for establishing confidence in the model's predictions and comprehending its limitations in real-world scenarios.

Figure 3 shows the error analysis presented compares Llama 2 and Mistral for the baseline established by the organizers, as well as prediction

Runs	LLMs	Average Score	Breast	Melanoma	Ovarian
Run 1	Llama 2	0.71	0.68	0.83	0.61
Run 2	Llama 2	0.68	0.66	0.80	0.59
Run 3	Mistral 7B	0.61	0.62	0.59	0.62

Table 5: Results of Runs on Test Data for Subtask 1.

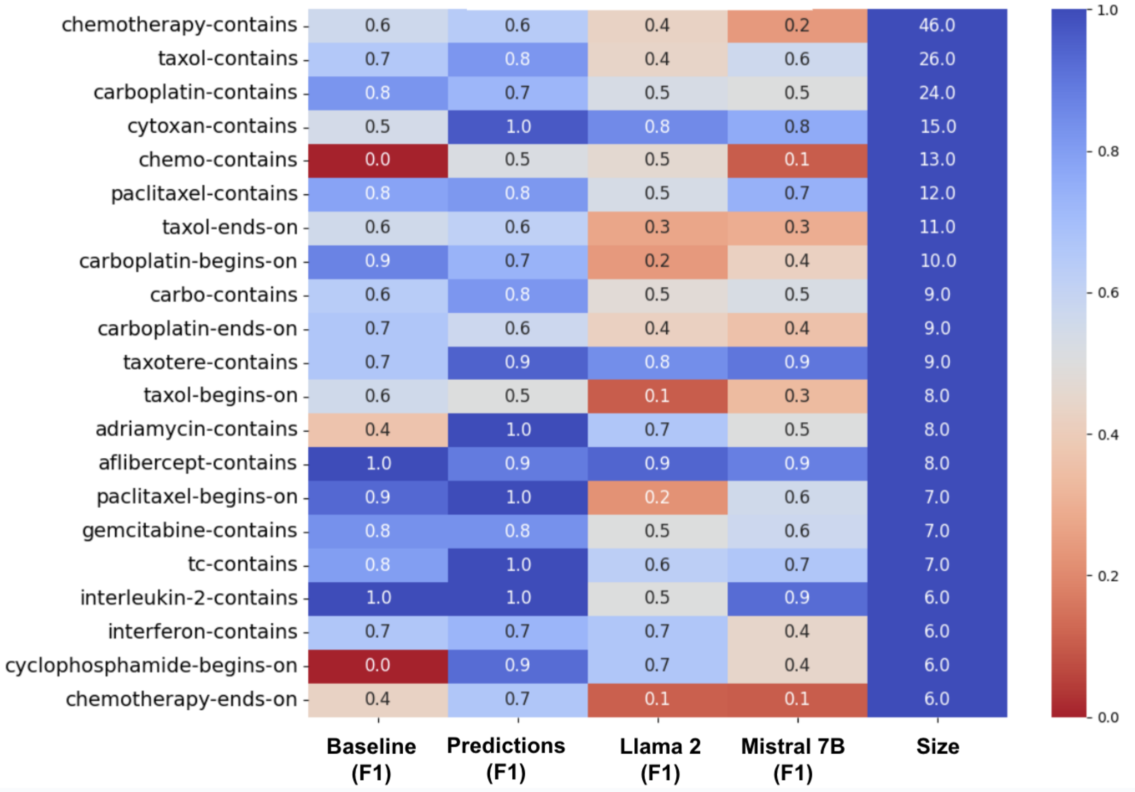


Figure 3: Error Analysis for Subtask 1.

scores derived by directly utilizing the golden relations as the timeline. The results from Llama 2 and Mistral 7B are based on the question answering prompting approach used to generate our timelines.

The metrics reveal lower precision within the system, characterized by exceptionally high recall. Further investigation into the distribution of false positives across event types or relation categories may unveil discernible patterns. It appears that the Large Language Model (LLM) is indiscriminately predicting all instances as if they are related events to timelines.

6 Conclusion and Future Work

In this paper, we present our submission to the Chemotimelines 2024 shared tasks (Yao et al., 2024) to build chemotherapy treatment timelines using Electronic Health Records (EHRs) notes from various sources, such as primary care providers, oncology departments, discharge summaries, emergency department, pathology, radiology, and more. We used zero shot prompted relation extraction (Wang et al., 2023; Jun and et al., 2022) driven by the THYME ontology (Styler IV et al., 2014) and temporal relation annotations (Wright-Bettner et al., 2020).

We evaluated pre-trained Large Language Models (LLMs) like Llama 2 (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023), Zephyr (Tunstall et al., 2023), MEDITRON-70B (Chen et al., 2023d), and Mixtral of Experts (Jiang et al., 2024) with different sizes and architectures. We only reported results on Llama 2 (Touvron et al., 2023) and Mistral 7B (Jiang et al., 2023). We conducted a series of experiments with different setups to improve the system's performance. From our analysis, we conclude that our approach helped us determine which model works best for this shared task. We conclude that LLMs provide a promising path forward for extracting timelines that contextualize cancer treatment, which were previously unavailable. We also show that our model provides high recall that is beneficial for instances where high sensitivity is required such as with output-sensitive predictions like cancer prediction models. However, due to the auto-generation approach and minimizing human intervention, the models we developed demonstrated relative low precision. Precision was evaluated with two physicians validating the accuracy of the generated chemotherapy timelines.

Our proposed methodology represents a significant advancement in the field, providing a flexible and efficient solution for relation extraction tasks in natural language processing. Large Language Models (LLMs) offer a promising approach for auto-generating chemotimelines from Electronic Health Records (EHRs) due to their advanced natural language understanding capabilities, contextual understanding, and semantic representation of medical information. LLMs can comprehend complex medical texts, capture contextual relationships between different medical events, and generate rich semantic representations of medical concepts and events mentioned in EHRs. As we see in our current study, our effort to attempt fully auto-generated chemotherapy timelines have shown great promise in terms of recall but have a negative impact on precision. In future studies we will explore further training rounds or human-in-the-loop models to explore the right balance between automation and human guided outputs. Nevertheless, our study demonstrates great promise in integrated LLM-generated chemotherapy timelines that have the potential to alleviate documentation and data harmonization burdens, potentially easing clinician workload and enhancing quality of patient care.

Exploring the potential of LLMs is an emerging area in research. We have experimented with two state-of-the-art LLMs (Llama 2 and Mistral 7B) for this task, comparing each with the gold standard for various cancer types. Our approach was to maintain a broad, domain-agnostic perspective, treating it as a high-level NLP relation detection task. We assumed that the underlying LLMs were general-purpose. In the future, we aim to explore domain-specific LLMs tailored for biomedical texts, such as JSL-MedMNX-7B (JSL-Med-Sft-Llama-3-8B, 2024), which could offer improved accuracy by better handling specialized language and data structures inherent in this domain.

Furthermore, we aim to validate the effectiveness of our LLM-based system across diverse healthcare datasets to enhance its performance. Additionally, we intend to conduct comprehensive analysis of the generated chemotherapy timelines to fine-tune them further and improve precision. This includes conducting in-depth error analyses to pinpoint the root causes of false positives. Our goal is to identify any consistent patterns, words, or phrases that the model may misinterpret, facilitating targeted improvements to enhance its accuracy.

Limitations

While leveraging Large Language Models (LLMs) for creating chemotherapy timelines from clinical notes offers numerous benefits, it also presents several limitations: 1. The accuracy and reliability of generated timelines heavily depend on the quality and consistency of input clinical notes, potentially leading to inaccuracies or omissions. 2. LLMs may exhibit biases inherent in the training data, leading to disparities, inaccuracies or generalization in the generated timelines, especially when applied to diverse patient populations. 3. LLMs are complex models with billions of parameters, making it challenging to interpret their decision-making processes, limiting clinicians' ability to trust and validate the generated outputs. 4. Training and fine-tuning LLMs for healthcare applications, including generating chemotherapy timelines, require significant computational resources, expertise, and time. Due to time constraints, we investigated a narrow range of models and hyperparameter configurations. Given their demonstrated proficiency in natural language processing, these models serve as an ideal starting point for extracting pertinent clinical events and concepts essential for constructing treatment timelines. 5. Despite the automation capabilities of LLMs, human oversight and validation are still essential to ensure the accuracy and relevance of the generated chemotherapy timelines. Clinicians must review and validate the outputs to identify and correct any inaccuracies or inconsistencies. In our study, two medical professionals compared chemotherapy timelines generated by LLMs, particularly the Llama 2 (Touvron et al., 2023) model, with a ground truth dataset provided by the challenge.

Ethics Statement

Leveraging Large Language Models (LLMs) for constructing timelines of chemotherapy treatments using Electronic Health Records (EHR) notes raises numerous ethical considerations. Foremost among these is the imperative to safeguard patient privacy and confidentiality, given the sensitive nature of personal health information stored in EHRs. By leveraging openly available LLMs, physicians can inadvertently expose patient data to private companies (Blease, 2024). Robust data security measures, and digital literacy training is essential to thwart unauthorized patient data exposure to LLMs or data breaches, thereby averting potential cyber threats. Additionally, obtaining informed consent from patients regarding the utilization of their health data is paramount to uphold patient autonomy and foster transparency. Ensuring the accuracy and integrity of the data is vital to mitigate risks of erroneous treatment timelines that could lead to patient harm. Moreover, LLMs may perpetuate biases inherent in the data, thereby introducing disparities or unfairness in the generated timelines (Singh et al., 2023). Prioritizing algorithmic transparency and accountability is imperative to identify and mitigate biases in LLM decision-making processes. Furthermore, granting patients control over their health data, including access and consent for research or analytical purposes, is fundamental in upholding patient autonomy and fostering trust in the healthcare system. The organizers of the Chemotimelines 2024 at NAACL-ClinicalNLP Workshop shared tasks (Yao et al., 2024) have provided a de-identified dataset.

In leveraging Large Language Models (LLMs) for Open Book Question Answering (QA), it's crucial to address the potential ethical concerns surrounding the minimization of generation divergence risk. This entails ensuring that the responses generated by LLMs align closely with the intended context and accurately reflect the information available in the open book. By minimizing generation divergence risk, we aim to uphold the integrity of the QA process, promote transparency, and mitigate the dissemination of misinformation or biased responses. Additionally, efforts should be made to continually evaluate and refine LLMs to enhance their reliability and trustworthiness in providing accurate and contextually appropriate answers.

It is noteworthy that LLMs often demonstrate a propensity to produce hallucinations when generating coherent answers, underscoring the necessity for human supervision in their utilization. Ensuring human supervision during the deployment of LLMs in healthcare contexts is crucial to validate the accuracy, appropriateness and potential harmfulness of the generated outputs and to mitigate potential risks or errors (Chen et al., 2023a). Moreover, it is crucial to recognize that the present system serves as an experimental tool intended to catalyze further research, including additional fine-tuning and model explainability studies. Such endeavors are indispensable before these systems can be safely incorporated into clinical settings, ensuring their reliability and efficacy in supporting clinical decision-making processes. Additionally, another critical aspect deserving careful consideration is the explainability and interpretability of Language Models (LLMs) when deployed in healthcare contexts.

References

Gaurav Anand and Divya Sadhna. 2023. Electronic health record interoperability using fhir and blockchain: A bibliometric analysis and future perspective. *Perspectives in Clinical Research*.

Imon Banerjee, Selen Bozkurt, Jennifer Lee Caswell-Jin, Allison W Kurian, and Daniel L Rubin. 2019. Natural language processing approaches to detect the timeline of metastatic recurrence of breast cancer. *JCO clinical cancer informatics*, 3:1–12.

Andrew L Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. 2019. Clinical concept embeddings learned from massive sources of multimodal medical data. In *Pacific Symposium on Biocomputing 2020*, pages 295–306. World Scientific.

Danielle S Bitterman, Eli Goldner, Sean Finan, David Harris, Eric B Durbin, Harry Hochheiser, Jeremy L Warner, Raymond H Mak, Timothy Miller, and Guergana K Savova. 2023. An end-to-end natural language processing system for automatically extracting radiation therapy events from clinical texts. *International Journal of Radiation Oncology* Biology* Physics*, 117(1):262–273.

Charlotte Blease. 2024. Open AI meets open notes: surveillance capitalism, patient privacy and online record access. *Journal of Medical Ethics*, 50(2):84–89.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Shan Chen, Benjamin H Kann, Michael B Foote, Hugo JWL Aerts, Guergana K Savova, Raymond H Mak, and Danielle S Bitterman. 2023a. Use of artificial intelligence chatbots for cancer treatment information. *JAMA oncology*, 9(10):1459–1462.

Shan Chen, Benjamin H Kann, Michael B Foote, Hugo JWL Aerts, Guergana K Savova, Raymond H Mak, and Danielle S Bitterman. 2023b. The utility of chatgpt for cancer treatment information. *MedrXiv*, pages 2023–03.

Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo JWL Aerts, Guergana K Savova, and Danielle S Bitterman. 2023c. Evaluation of chatgpt family of models for biomedical reasoning and classification. *arXiv preprint arXiv:2304.02496*.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. 2023d. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Jason P Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Samuel Colvin and contributors. 2024. Pydantic: Data validation and settings management library for python. https://github.com/samuelcolvin/pydantic.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ashkan Ebadi, Pengcheng Xi, Stéphane Tremblay, Bruce Spencer, Raman Pall, and Alexander Wong. 2021. Understanding the temporal evolution of covid-19 research through machine learning and natural language processing. *Scientometrics*, 126:725–739.

Sören Möller Eriksen, Alexander V. and Jesper Ryg. 2023. Use of gpt-4 to diagnose complex clinical cases. *NEJM AI*.

Marco Guevara, Shan Chen, Spencer Thomas, Tafadzwa L Chaunzwa, Idalid Franco, Benjamin H Kann, Shalini Moningi, Jack M Qian, Madeleine Goldstein, Susan Harper, et al. 2024. Large language models to identify social determinants of health in electronic health records. *NPJ digital medicine*, 7(1):6.

Harry Hochheiser, Sean Finan, Zhou Yuan, Eric B Durbin, Jong Cheol Jeong, Isaac Hands, David Rust, Ramakanth Kavuluru, Xiao-Cheng Wu, Jeremy L Warner, et al. 2023. Deepphe-cr: Natural language processing software services for cancer registrar case abstraction. *JCO Clinical Cancer Informatics*, 7:e2300156.

Mohammad Hossin and Md Nasir Sulaiman. 2015. A review on evaluation metrics for data classification evaluations. *International journal of data mining & knowledge management process*, 5(2):1.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, page 108189.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

John Snow Labs. 2024. Spark nlp - date normalizer annotator documentation. https://nlp.johnsnowlabs.com/licensed/api/python/modules/sparknlp_jsl/annotator/normalizer/date_normalizer.html.

JSL-Med-Sft-Llama-3-8B. 2024. John Snow Labs jsl-med-sft-llama-3-8b. https://huggingface.co/johnsnowlabs/JSL-Med-Sft-Llama-3-8B.

Zhao Jun and et al. 2022. An exploration of prompt-based zero-shot relation extraction method. In *Proceedings of the 21st Chinese National Conference on Computational Linguistics*.

Kenneth L Kehl, Wenxin Xu, Eva Lepisto, Haitham Elmarakeby, Michael J Hassett, Eliezer M Van Allen, Bruce E Johnson, and Deborah Schrag. 2020. Natural language processing to ascertain cancer outcomes from medical oncologist notes. *JCO Clinical Cancer Informatics*, 4:680–690.

Saif Sherif Khairat, Aniesha Dukkipati, Heather Alico Lauria, Thomas Bice, Debbie Travers, and Shannon S Carson. 2018. The impact of visualization dashboards on quality of care and clinician satisfaction: Integrative literature review. *JMIR Hum Factors*, 5(2):e22.

V Kocaman and D Talby. 2020. Biomedical named entity recognition at scale. *Pattern Recognition. ICPR International Workshops and Challenges*.

Veysel Kocaman and David Talby. 2021. Spark nlp: Natural language understanding at scale. *Software Impacts*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Mark N Levine, Gordon Alexander, Arani Sathiyapalan, Anjali Agrawal, and Greg Pond. 2019. Learning health system for breast cancer: pilot project experience. *JCO clinical cancer informatics*, 3:1–11.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics*, 17(01):128–144.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. *arXiv preprint arXiv:1802.05695*.

Marjan Najafabadipour, Massimiliano Zanin, Alejandro Rodríguez-González, Maria Torrente, Beatriz Nuñez García, Juan Luis Cruz Bermudez, Mariano Provencio, and Ernestina Menasalvas. 2020. Reconstructing the patient's natural history from electronic health records. *Artificial intelligence in medicine*, 105:101860.

Kalia Orphanou, Athena Stassopoulou, and Elpida Keravnou. 2014. Temporal abstraction and temporal bayesian networks in clinical domains: A survey. *Artificial intelligence in medicine*, 60(3):133–149.

Shirui Pan, Jie Wang, Chuxu Zhang, and Jiawei Han. 2024. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*.

Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M Dai, Nissan Hajaj, Michaela Hardt, Peter J Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, et al. 2018. Scalable and accurate deep learning with electronic health records. *NPJ digital medicine*, 1(1):1–10.

Rasa. 2024. Duckling: Open-source library for parsing structured information from text. https://github.com/facebook/duckling.

Miriam Reisman. 2017. Ehrs: the challenge of making electronic data usable and interoperable. *Pharmacy and Therapeutics*, 42(9):572.

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.

Nina Singh, Katharine Lawrence, Safiya Richardson, and Devin M Mann. 2023. Centering health equity in large language model deployment. *PLOS Digital Health*, 2(10):e0000367.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the association for computational linguistics*, 2:143–154.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Temporal reasoning over clinical text: the state of the art. *Journal of the American Medical Informatics Association*, 20(5):814–819.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, pages 1–9.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers. *arXiv preprint arXiv:2312.01044*.

Cliff Wong, Sheng Zhang, Yu Gu, Christine Moung, Jacob Abel, Naoto Usuyama, Roshanthi Weerasinghe, Brian Piening, Tristan Naumann, Carlo Bifulco, et al. 2023. Scaling clinical trial matching using large language models: A case study in oncology. In *Machine Learning for Healthcare Conference*, pages 846–862. PMLR.

Kristin Wright-Bettner, Guergana Savova, and Steven Bethard. 2020. Defining and learning refined temporal relations in the clinical narrative. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*.

Yanjun Xu, Qun Dong, Feng Li, Yingqi Xu, Congxue Hu, Jingwen Wang, Desi Shang, Xuan Zheng, Haixiu Yang, Chunlong Zhang, et al. 2019. Identifying subpathway signatures for individualized anticancer drug response by integrating multi-omics data. *Journal of translational medicine*, 17:1–16.

Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli 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.