Clinical Language Understanding

2024 Shared Task on Chemotherapy Treatment Timeline Extraction

Xiran Hu, Hiu Lam Choy, Mengtong Guo, Zhaokun Wang, Raziye Sari

Part 1

Introduction and Overview

Overview of the 2024 Shared Task on Chemotherapy Treatment Timeline Extraction
(Jiarui Yao, Harry Hochheiser, WonJin Yoon, Eli Goldner, Guergana Savova)

Introduction

What is the Shared Task?

- The task focuses on extracting chemotherapy treatment timelines from clinical narratives, enabling machine-readable patient histories to support oncology research.
- Participants aim to develop methods for processing unstructured clinical notes into structured timelines that link events, temporal expressions, and relations.

Why is it important?

- Structured timelines enable precise treatment planning and large-scale retrospective studies, particularly in cancer care.
- Automated extraction reduces manual annotation burden and errors.

Introduction

Challenges:

- **Unstructured Data**: Electronic Health Records (EHRs) often lack consistent formatting, with highly variable language use.
- **Temporal Complexity**: Relations between events and time expressions require subtle contextual understanding.
- **Data Sparsity**: Annotated data is limited, complicating model training.

Goals:

- Build robust models for temporal reasoning in clinical NLP.
- Support downstream tasks like patient monitoring and outcome prediction.

Key Components of the Shared Task

Objectives:

1. Extract chemotherapy-related events from clinical notes.
2. Normalize temporal expressions into standard formats (e.g., ISO dates).
3. Predict temporal relations (e.g., `BEFORE`, `CONTAINS`).

Dataset:

- Cancer types: Breast, ovarian, melanoma.

Temporal Relations

Relation Types:

- `BEFORE`, `CONTAINS`, `CONTAINS-1` (inverse of `CONTAINS`), `OVERLAP`, `NOTED-ON`, `BEGINS-ON`, `ENDS-ON`.

Example:

- "Cycle 1 contains the treatment in January." → *<Cycle 1, `CONTAINS`, January>.*

Challenges:

- Variability in clinical text structure and language.
- Ambiguity in time expressions.

Subtask 1 - Temporal Relation Classification

Goal:

- Predict relations between annotated chemotherapy events and time expressions.

Input-Output:

- **Input:** Gold chemotherapy event mentions and time expressions are provided (along with the EHR notes).
- **Output:** <*chemotherapy, temporal_relation, time_expression*> triplets.

Key Focus:

- Operates on gold-standard annotated data.
- Tests temporal reasoning in an ideal scenario.

Subtask 2 - End-to-End Timeline Extraction

Goal:

- Extract events, temporal expressions, and relations from **raw** clinical text.

Workflow:

1. Detect chemotherapy events (e.g., *Taxol*).
2. Normalize temporal expressions (e.g., *October 20, 2013 → 2013-10-20*).
3. Predict relations to construct triplets.

Input-Output:

- **Input:** only EHR notes are provided
- **Output:** <*chemotherapy, temporal_relation, time_expression*> triplets.

Challenges:

- End-to-end extraction introduces potential error propagation.
- Real-world variability in text complicates performance.

Overview of the 2 subtasks

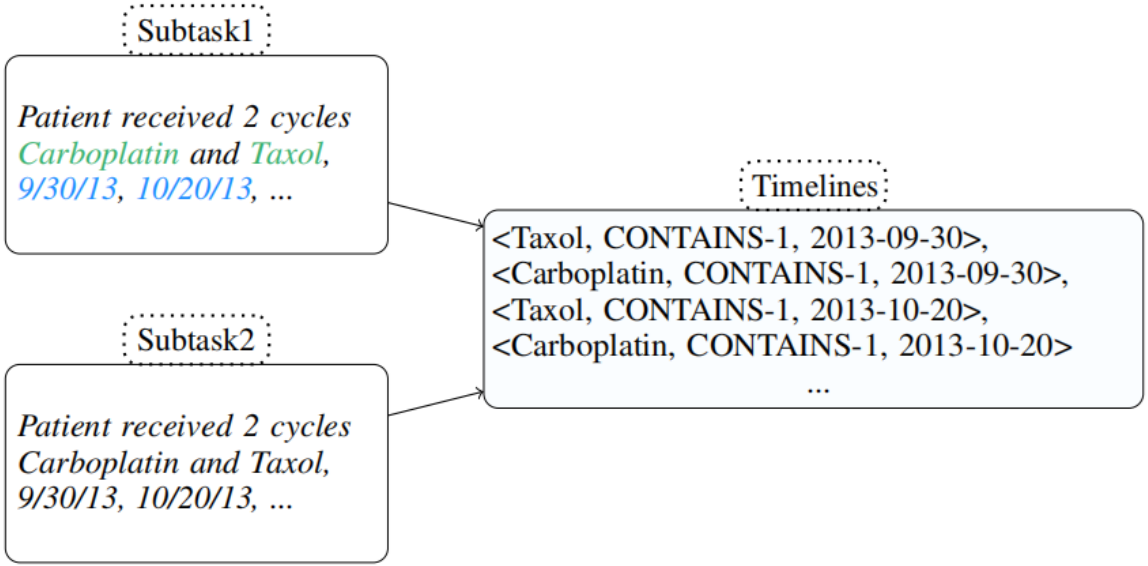


Figure 1: Illustration of the two subtasks in the 2024 Chemotherapy Treatment Timeline Extraction shared task. The input of Subtask1 is patient notes with gold events (highlighted in green) and time expressions (highlighted in blue). The input of Subtask2 is patient notes only. The output of both subtasks is a list of chemotherapy treatment timelines with normalized time expressions. See details in section 2.

Data

The Gold dataset (the *Labeled Dataset*):

	Train			Dev			Test		
	Patients	Notes	Words	Patients	Notes	Words	Patients	Notes	Words
Ovary	26	1,675	1,183,632	8	562	308,814	8	559	257,116
Breast	33	1,002	465,644	16	499	225,588	35	1,333	786,896
Melanoma	10	233	124,924	3	211	178,308	10	229	156,083

Table 1: Gold labeled dataset: number of patients, notes, and words across train/dev/test sets. "Words" denotes the tokens delimited by white spaces.

	Train			Dev			Test	
	EVENT	TIMEX3	TLINK	EVENT	TIMEX3	TLINK	EVENT	TIMEX3
Ovary	1,168	597	494	790	312	226	664	381
Breast	1,023	576	455	279	146	113	2,560	1,118
Melanoma	147	78	48	789	261	201	398	193

Table 2: Gold labeled dataset: EVENTs/ TIMEX3s/ TLINKs distribution in the labeled dataset. TIMEX3 and TLINK refer to time expressions and temporal relations respectively.

Data

Unlabeled Dataset:

- Consists of the UPMC EHR notes for 57,530 patients with breast and ovarian cancer, collected between 2004-2020.
- 15,946 patients with melanoma, collected between 2010-2020.
- Does not have any gold annotations.
- Potentially used for continued training of pretrained language models or for pretraining a language model.

Evaluation Overview

- **Goal:** Evaluate system performance for chemotherapy treatment timeline extraction using metrics tailored for clinical applicability.
- **Official Metric:** Relaxed-to-**month** F1 score is used for ranking systems. It reflects practical use cases by prioritizing month/year granularity.
- **Evaluation Granularity**
  - **Strict:** Exact match for all elements in triplets.
  - **Relaxed (day, month, year):** Focuses on broader temporal matches (e.g., *2013-02* instead of *2013-02-13* for relaxed-to-month metric).

Types of Scores

- **Type A (With Non-Chemotherapy Patients):**
  - Includes all patients in F1 calculation.
  - Aims to catch false positives for non-chemotherapy patients.
- **Type B (Chemotherapy Patients Only):**
  - Excludes non-chemotherapy patients, focusing on chemotherapy-specific cases.
  - Measures the effectiveness of the methods on patients with confirmed chemotherapy treatments.

Evaluation Methodology in Practice

Steps:

1. Systems upload results to Globus for secure processing.
2. Organizers run scripts to calculate F1 scores across metrics.
3. Teams allowed three submissions per subtask.

Score Aggregation:

- The F1 score for each patient was computed and the final F1 score for each type is the average across all patients.
- **Final Official Score** = Average of Type A and Type B relaxed-to-month F1 scores.
- Ensures balanced evaluation between general and chemotherapy-focused scenarios.

Baseline Systems Overview

- **Purpose:** Provide baseline results for Subtask 1 (Temporal Relation Classification) and Subtask 2 (End-to-End Timeline Extraction).
- **Baseline Models:**
  - Use Apache **cTAKES** for preprocessing and Timenorm for temporal normalization.
  - Employ pre-trained **PubMedBERT** for temporal relation classification.
- **Evaluation:**
  - Baselines are tested on both subtasks to establish reference performance metrics.

Subtask 1 Baseline

Pipeline:

1. Tokenization and sentence splitting using **cTAKES**.
2. Identify chemotherapy events and TIMEX3 annotations (gold-standard input).
3. Normalize time expressions using **Timenorm**.
4. Classify temporal relations using fine-tuned **PubMedBERT** on **THYME2** and task-specific datasets.

Performance:

- Baseline achieves reasonable F1 scores by leveraging clean, gold-standard inputs.

Subtask 2 Baseline

Pipeline:

1. Use **cTAKES** with a custom dictionary to detect chemotherapy mentions.
2. Extract and normalize time expressions with Timenorm.
3. Classify temporal relations using the same **PubMedBERT** model as Subtask 1.

Performance:

- Baseline F1 scores are significantly lower compared to Subtask 1.
- Error propagation across detection, normalization, and classification stages reduces accuracy.

Participating Systems

Teams	Approach	LM or Algorithm	Task
BioCom_submission1	Machine Learning	Logistic Regression	1
ClinicalRXMiners_submission1	Machine Learning	Soft voting classifier	1
ClinicalRXMiners_submission2	Deep Learning	GLiNER Base	1
KCLab_submission1	Finetuned LM	PubMedBert	1, 2
LAILab_submission1,2,3	Finetuned LM	flan-T5-xxl, bart-large	1, 2
Lexicans-submission1,2,3	Zero-shot Prompting	Llama2, Mistral, Zephyr, Meditron, and Mixtral	1
NLPeers_submission1	Finetuned LM	deberta-v3-base	1
NLPeers_submission2	Few-shot Prompting	Mixtral-8X7B-Instruct-v0.1	1
NYULangone_submission1	Zero-shot prompting	Mixtral 8x7B	2
UTSA-NLP_submission1,2,3	Instruction tuning LM, continued pretraining LM	OpenChat-3.5-7B	1, 2
Wonder_submission1,2,3	Finetuned LM	Bio-LM	1, 2

Table 3: Characteristics of participating systems.

Overall results for subtask1

Submission	Type A	Type B	Official Score
LAILab_submission1	0.94	0.86	0.90
LAILab_submission2	0.94	0.86	0.90
LAILab_submission3	0.94	0.86	0.90
Baseline_subtask1	0.93	0.85	0.89
Wonder_submission2	0.90	0.78	0.84
Wonder_submission1	0.89	0.77	0.83
Wonder_submission3	0.88	0.73	0.80
NLPeers_submission1	0.85	0.70	0.77
BioCom_submission1	0.84	0.64	0.74
Lexicans_submission1	0.81	0.61	0.71
UTSA-NLP_submission3	0.80	0.58	0.69
UTSA-NLP_submission1	0.80	0.58	0.69
Lexicans_submission2	0.79	0.57	0.68
UTSA-NLP_submission2	0.80	0.56	0.68
NLPeers_submission2	0.76	0.52	0.64
KCLab_submission1	0.76	0.49	0.63
Lexicans_submission3	0.75	0.47	0.61
ClinicalRXMiners_submission1	0.51	0.28	0.40
ClinicalRXMiners_submission2	0.56	0.21	0.38

Top Performers:

- **LAILab**: Achieved the highest F1 score

Key Observations:

- Transformer-based systems performed well due to clean gold-standard inputs.
- **Finetuned LM** performs overall better(LAILab, Wonder, NLPeers).
- Type A Scores are higher than Type B for all submissions.

Overall results for subtask2

Submission	Type A	Type B	Official Score
LAILab_submission2	0.76	0.63	0.70
Baseline_subtask2	0.67	0.48	0.58
LAILab_submission1	0.65	0.47	0.56
KCLab_submission1	0.63	0.45	0.54
Wonder_submission3	0.59	0.46	0.53
Wonder_submission2	0.59	0.46	0.52
Wonder_submission1	0.58	0.46	0.52
LAILab_submission3	0.47	0.47	0.47
NYULangone_submission1	0.26	0.21	0.23
UTSA-NLP_submission1	0.22	0.22	0.22

Top Performers:

- **LAILab**: Achieved the highest F1 score

Key Observations:

- A comparison of the scores between Subtask1 and Subtask2 shows a substantial drop of at least **0.2** F1 Official Score
- Event and time expression extraction is not a solved problem while the task of temporal relation extraction holds strong.
- **Finetuned LM** performs overall better(LAILab, Wonder, KCLab).
- Type A Scores are higher than Type B for most submissions.

Fine-Tuning LMs

- **LAILab**
  - Fine-tuned **Flan-T5-xxl** (11B parameters) and **Bart-large** (400M parameters).
  - Achieved top results across most subtasks.
- **Wonder**
  - Fine-tuned **Bio-LM**, consistently ranking in the top 3 for both subtasks.
- **Other teams:**
  - **NLPeers** and **KCLab** fine-tuned `deberta-v3-base` and **PubMedBERT**, achieving commendable performance.
- **Key Insights:**
  - Finetuning LMs remains the optimal approach for optimizing system performance if gold labeled data and computing resources are available.

Prompting LLMs

Teams:

- **Lexicans Team:** Experimented with zero-shot prompting using five LLMs, including Llama2 and Mistral.
- **NYULangone Team:** Used **Mixtral** for zero-shot prompting.
- **NLPeers Submission 2:** Applied **few-shot prompting** using `Mixtral-8X7B-Instruct-v0.1`.

Performance:

- Prompting methods performed worse than fine-tuned LMs.
- Challenges:
  - Alignment of general-purpose LLMs to clinical tasks.
  - Handling complex temporal relations and ambiguous expressions.

Conclusion & Limitation

- The **2024 Shared Task on Chemotherapy Treatment Timeline Extraction** stands out for:
  - Tackling a highly complex, clinically relevant task.
  - Providing a large EHR dataset for robust benchmarking.
- Fine-tuned smaller LMs (e.g., **PubMedBERT, Flan-T5**) consistently outperformed large LLMs in this domain.
- Highlighted the need for more sophisticated LLMs or task-specific prompting techniques to address challenges in clinical timeline extraction.
- **Limitation:**
  - The shared task focuses solely on chemotherapy treatments, leaving timeline construction for other cancer therapies for future research.

Questions?

Subtask 1

Team NLPeers at Chemotimelines 2024:
Evaluation of two timeline extraction methods, can generative LLM do it all or is smaller model fine-tuning still relevant ?

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

Recap: Objectives

- Temporal relations extraction and patient-level timeline creation
- Given gold *(event, time)* input
- Timeline format: *(chemotherapy, temporal_relation, time_expression)* triplets

Team NLPeers

Approaches:

1. MLM fine-tuning
2. Automated few-shot LLM prompting

Submission 1: MLM fine-tuning approach

- Model: `DeBERTa-v3 base` (85M parameters)
- Multi-class classification task
- Classify labels as {`BEGINS-ON, ENDS-ON, CONTAINS-1, no_link`}
- Discard `no_link`

Hyperparameters:

- Learning rate: 2e-5
- Weight decay: 0.01
- Epoch: max 10 (with evaluation strategy)

Data pre-processing

- Add prefixes (TIME=) and (EVENT=) to the respective entities
- Highlights the terms for easier classification
- Limit the number of pairs considered
    - Threshold of character distance (between event and time relation) ≤ 300 characters

Post-processing

- Date normalisation
- To produce a normalised *TIMEX3* expression

1. HeidelTime
   - Using document creation time (DCT)
   - Normalise "*currently*" as DCT
1. LLM-based query
   - Model: `OpenChat 3.5`
   - Few-shot with 6 synthetic examples in format: *(time_expression, doctime/None, answer_date/error_string)*

Post-processing

LLM-based query normalisation prompt:

“Please normalise the following string to a date format YYYY-MM-DD or, if you can’t to a YYYY-MM format (the time at which the document is redacted is <doc_time_input>): <time expression>”

Submission 2:
Automated few-shot prompting LLM approach

- Generation task
- Model generates relation triplets *(event, relation_type, time)*
- Date normlisation: HeidelTime only

Model:

- `Mixtral-8X7B-Instruct-v0.1`
- Temperature: 0 (low randomness)

DSPy framework

- Declarative Self-improving Python
- Optimises prompts to match changes in code, data, or metric

- Adds chain of thought reasoning statement

Signature:

- Task and input/output description

Teleprompters:

- `BootstrapFewShotWithRandomSearch`
- Self-generates demonstrations few times
- Randomly search over these demonstrations to select the best prompt

Prompts

Respond to the question based on the given text.
The possible answers are: 'CONTAINED-BY', 'BEGINS-ON', 'ENDS-ON'.

Follow the following format.

Question: ${question}

Text: ${text}

Reasoning: Let's think step by step in order to ${produce the answer}. We …

Answer: a list containing only the relation. If no relation is found, the answer is solely an empty list.

Question: Given this chemotherapy event: ${EVENT} and this temporal expression: ${TIMEX}, which is the relation between these entities, if any?

Text: ${text}

Output: relation type (unofficial)

Respond to the question based on the given text.
The possible answers are: 'CONTAINED-BY', 'BEGINS-ON', 'ENDS-ON'.

Follow the following format

Question: ${question}

Text: ${text}

Reasoning: Let's think step by step in order to ${produce the answer}. We …

Answer: Each answer is an ordered list, containing the chemotherapy event, then the corresponding answer then the temporal expression. If no relation is found, the answer is an empty list.

Question: Given this chemotherapy event: ${EVENT} and this temporal expression: ${TIMEX}, which is the relation between these entities, if any ?

Text: ${text}

Output: relation triplet (event, relation,

Patient-level summarisation

- To create a timeline from the triplets

1. Discard triplets when *time* does not match the pattern YYYY-MM-DD
2. Only keep the more precise triplets
    - When only relation types differ (same date & event)
    - `CONTAINS-1` vs a more precise type (`BEGINS-ON`, `ENDS-ON`)
3. De-duplicate
4. Sort

Evaluation metric

- Type A: includes patients with no chemotherapy
- Type B: only patients with chemotherapy
- relaxed-to-month F1 for submission
- Strict F1 for LLM prompting approach
  - To optimise the model and ensure quality
  - If the output format does not follow *(event, relation, time)* → not matched

Results

Approach	Average Score	Breast cancer	Melanoma	Ovarian
Fine-tuned MLM + HeidelTime & OC normalization *(NLPeers 1)*	0.77	0.72	0.84	0.75
Automated few-shot LLM **(Relation triplet)** + HeidelTime normalization *(NLPeers 2)*	0.64	0.49	0.81	0.63
Baseline system	0.89	0.93	0.87	0.88

Table 1: The official results on the test set. OC refers to the LLM-based normalization using the `OpenChat` model.

Results

Approach	Average Score	Breast	Melanoma	Ovarian
Fine-tuned MLM				
Relation type (classification) + HeidelTime & OC normalization *(official submission, NLPeers 1)*	0.85	0.84	0.81	0.88
Relation type (classification) + HeidelTime normalization *(non official submission)*	0.74	0.61	0.85	0.76
Automated few-shot LLM				
Relation triplet (generation) + HeidelTime & OC normalization *(non official submission)*	0.72	0.70	0.74	0.71
Relation type (generation) + HeidelTime normalization *(non official submission)*	0.61	0.57	0.78	0.48
Relation triplet (generation) + HeidelTime normalization *(official submission, NLPeers 2)*	0.56	0.53	0.70	0.47

Table 2: The results on the development set. OC refers to the LLM-based normalization using the `OpenChat` model.

Discussion

Errors of MLM fine-tuning approach:

- `no_link` mislabelled as `CONTAINS`
  - Due to imbalance class
- `ENDS-ON` mislabelled as `BEGINS-ON`
  - But never the other way around (2/103 vs 30/83)
- Relatively worse performance on the melanoma subset
  - Due to less melanoma examples in the training set

Gold	BEGINS-ON	CONTAINS	ENDS-ON	no_link
BEGINS-ON	52	18	30	0
CONTAINS	49	328	30	68
ENDS-ON	2	1	11	0
no_link	0	7	12	587
Total	103	354	83	655

Table 3: Confusion matrix for the MLM fine-tuning approach applied on the development set.

Discussion

Errors of LLM prompting approach:

- Semantic error: A relation type that is not in the pre-defined set of relation types
- Model tends to generate large texts containing explanations and hallucinations

	Relation Type Error	Semantically incorrect samples
Automated few-shot LLM **Relation triplet** + Heideltime *(official submission, NLPeers 2)*	43	occurs on, occurs-on, contained-in, not going to occur, not related, duration, ended-on
Automated few-shot LLM **Relation type** + Heideltime *(non official submission)*	16	answer, be, beg, begins, conta, during, every-on, happening-on, happens-on, lasts-for, planned-for

Table 4: Semantic errors and semantically incorrect samples on the development set.

Discussion

Normalisation

- Improved performance for both approaches
- Without needing specific background knowledge
  → Does not requrie complex task description or prompt search strategy

- Did not tested organiser’s normalisation approach
  → Incomparable and must discard terms that could not be normalised
  → Potentially correct time expressions and relations as incorrect

Lexicans

Approach:

- Zero-shot prompting

Zero-shot Prompting

- Without requiring specific training data
- Only considers {`BEGINS-ON, ENDS-ON, CONTAINS`}

Model

- LLaMA 2, Mistral 7B
- Fully auto-generate a chemotherapy timeline
  - Minimised human intervention

Hyperparameters:

- Chunk size: 256
- Temperature: 0.1

Data Pre-processing

Document chunking

- Divide into paragraphs and sections

Paragraph detection

- Only focus on entities within each paragraph
  → To increase precision
- An overlap parameter
  - Concatenate if entities span across paragraphs
  → Prevent separation of information across paragraphs

Prompt

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.

Desired output format:
JSON with list of triplet tuples

Post-processing and Evaluation

- Date normalisation
- Directed Acyclic Graph (DAG)
  - Ensures validation functions are executed in a specific order
  - Resolves conflicting relations
- Aggregation of parsed subgraphs into a timeline

To get the precisions and recall:

- Two physicians validating the accuracy of the timelines

Results

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.

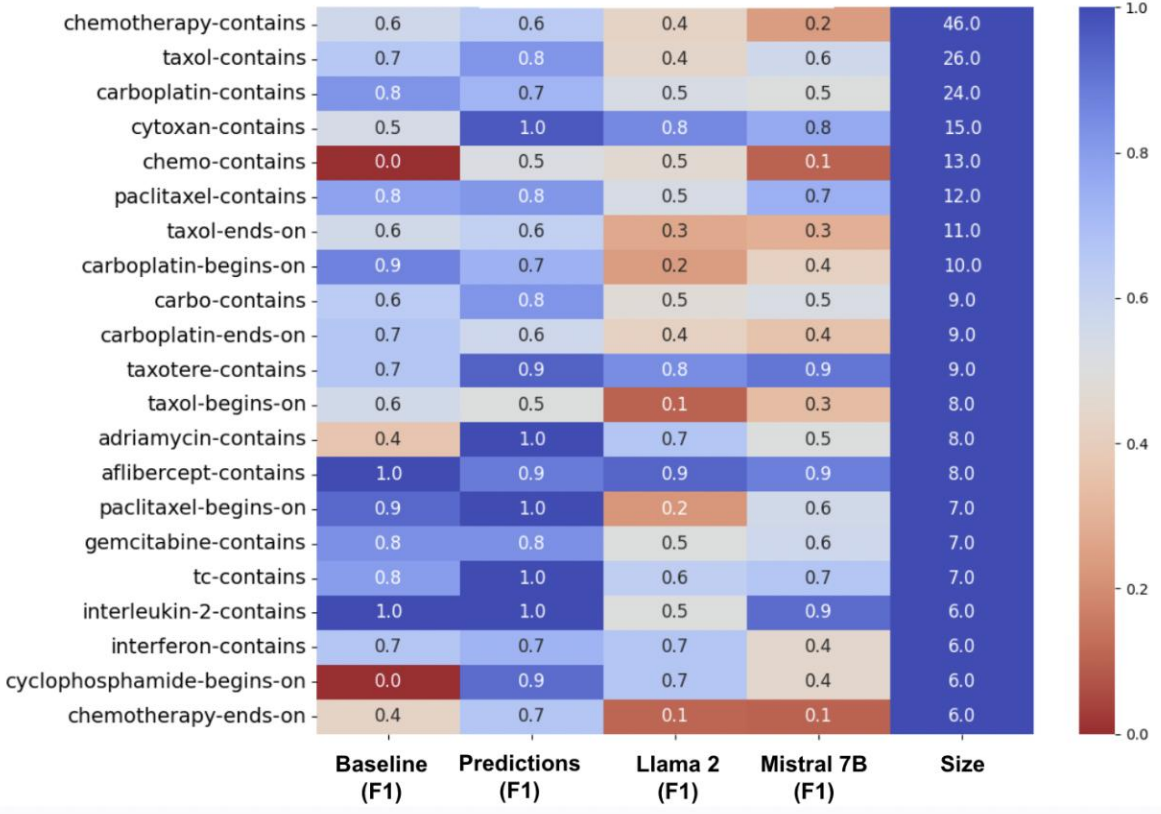
Runs	LLMs	Average Score	Breast	Melanoma	Ovarian
Run 1	Llama 2	0.71	0.68	0.83	0.61
Run 2	Llama 2	0.68	0.66	0.80	0.59
Run 3	Mistral 7B	0.61	0.62	0.59	0.62

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

Issues

- Oversight of relations by LLM
- Dates misinterpreted as a future date
- LLM predicted all instances as timeline-related events (many are in fact not)
- Over-classification of `CONTAINS`
  → Low precision
  → Low recall on other classes

Error analysis



Overall results for subtask 1

Submission	Type A	Type B	Official Score
LAILab_submission1	0.94	0.86	0.90
LAILab_submission2	0.94	0.86	0.90
LAILab_submission3	0.94	0.86	0.90
Baseline_subtask1	0.93	0.85	0.89
Wonder_submission2	0.90	0.78	0.84
Wonder_submission1	0.89	0.77	0.83
Wonder_submission3	0.88	0.73	0.80
NLPeers_submission1	0.85	0.70	0.77
BioCom_submission1	0.84	0.64	0.74
Lexicans_submission1	0.81	0.61	0.71
UTSA-NLP_submission3	0.80	0.58	0.69
UTSA-NLP_submission1	0.80	0.58	0.69
Lexicans_submission2	0.79	0.57	0.68
UTSA-NLP_submission2	0.80	0.56	0.68
NLPeers_submission2	0.76	0.52	0.64
KCLab_submission1	0.76	0.49	0.63
Lexicans_submission3	0.75	0.47	0.61
ClinicalRXMiners_submission1	0.51	0.28	0.40
ClinicalRXMiners_submission2	0.56	0.21	0.38

Limitations of subtask 1

- Potential biases in training data
  - From diverse populations
- Black-box
- Computationally expensive to train and fine-tune
- Requires human oversight

Questions?

Subtask 2 zero-shot prompting

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

Overview

- **Introduction and Related Work:** Extract chemotherapy timelines without domain-specific training.
- **System Description:** Process clinical notes, extract events, structure in JSON.
- **Results:** F1score- 0.35 (Dev), 0.23 (Validation), below baseline.
- **Discussion:** Summarize System error type.
- **Future Work:** Improve using RAG for knowledge retrieval and Tree of Thought for better reasoning.

Introduction

- **Background:** Extracting structured information from unstructured clinical narratives is essential for healthcare informatics, enabling better patient care and clinical decision-making.
- **Research Objective:** The team aimed to use LLMs to extract chemotherapy-related events without domain-specific training.
- **Significance:** This paper explores whether LLMs can process medical text effectively, even without specialized medical training.

Related Work

Traditional Methods

- **Rule-Based Systems**: Use predefined rules (e.g., regex, UMLS) but lack flexibility.
- **Machine Learning Models**: Need domain-specific data and work well with structured input but are less adaptable.

Modern NLP Approaches

- **Transformer-Based Models**: Improve general-purpose language understanding (e.g., BERT, GPT).
- **Large Language Models (LLMs)**: Show promise in handling unstructured text (e.g., GPT-4, Mixtral 8x7B).

LLM Limitations: still in early stages for chemotherapy timeline extraction.

- Lack of domain-specific training– LLMs are mostly trained on general text.
- Complex medical language– Dense terminology and inconsistent formatting.
- High accuracy required– Errors in medical data can have serious.

These challenges highlight the need for further research in applying LLMs.

System Description

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
```

- **System Architecture:** builds upon a locally deployed instance of Mixtral, an open-weights LLM.
- **Two-stage inference process:**
  - Extract chemotherapy events from individual medical notes.
  - Aggregate extracted events into a single patient-level timeline.
- **Advantages:**
  - Stepwise processing reduces computational complexity and improves accuracy.
  - Local LLM deployment ensures data security and privacy.

System Description- Architecture

LLM Choice

- Mixtral 8x7B v0.1 (open-weights LLM)
- No domain-specific fine-tuning

Processing Workflow

- Input: Raw clinical text (EHR notes)
- Extracts events, dates
- Output: structures them into the required JSON

Advantages

- Open-source, avoids proprietary dependence
- Direct text processing, minimal preprocessing
- JSON format for storage and analysis

System Description- Prompts

- **GOAL and PURPOSE:** Let the LLM play the role of an experienced medical annotator with special expertise in natural language processing of oncology documents.
- **INSTRUCTIONS:**
  - Read the patient's note under "# PATIENT NOTE".
  - Follow THYME guidelines to extract "events", every mention of a chemotherapeutic drug or component should have:
    - **Chemotherapy Drug Name**: The name of the drug
    - **Associated Date**
    - **Temporal Relation**: The temporal relation between the use of that drug and the associated date (option:["contains-1", "begins-on", "ends-on", "before"])
  - 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.

System Description- Prompts(continue)

- **EXAMPLES:**

  [’herceptin’, ’begins-on’,’2013-06-17’] [’taxol’, ’contains-1’, ’2013-09’]

- **OUTPUT FORMAT:**
  - Only output well-formatted JSON under ‘TIMELINE‘.
  - No additional notes or comments beyond structured JSON.

Summary:

- This step extracts chemotherapy events from individual notes.
- However, the events are not organized by patient yet.
- Next step: Aggregating multiple notes into a comprehensive patient timeline.

System Description- Prompts(continue)

- **GOAL and PURPOSE:** Continue to simulate a medical annotator. The task is to output a list of lists for each patient (First will be given a JSON list of lists).
- **EXAMPLE OUTPUT:**

```
patient_01:
['taxol', 'begins-on', '2013-06-17']
['taxol', 'ends-on', '2013-09']
...
patient_02:


[/INST]
```

Summary:

- **Structured Data:** Supports tracking drug usage and predicting treatment progress.
- **Automated Parsing:** Extracts key medical data without domain-specific training.

Results

Average Scores		Breast Cancer		Melanoma		Ovarian	
Team	Score	Team	Score	Team	Score	Team	Score
LAILab 2	0.70	KCLab 1	0.68	LAILab 2	0.74	LAILab 2	0.74
LAILab 1	0.56	Wonder 2	0.64	LAILab 1	0.57	LAILab 1	0.59
KCLab 1	0.54	Wonder 1	0.63	KCLab 1	0.49	Wonder 3	0.55
Wonder 3	0.53	Wonder 3	0.63	Wonder 3	0.39	Wonder 2	0.55
Wonder 2	0.52	LAILab 2	0.62	Wonder 1	0.39	Wonder 1	0.53
Wonder 1	0.52	LAILab 3	0.53	Wonder 2	0.39	LAILab 3	0.49
LAILab 3	0.47	LAILab 1	0.52	LAILab 3	0.38	KCLAb 1	0.45
NYULangone	0.23	UTSA-NLP 1	0.25	NYULangone	0.32	UTSA-NLP 1	0.19
UTSA-NLP 1	0.22	NYULangone	0.19	UTSA-NLP 1	0.21	NYULangone	0.18
Baseline	0.58	Baseline	0.59	Baseline	0.43	Baseline	0.71

- On the **dev** set, this 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. Results are shown in Table of the competition results.

Discussion

- Performance below baseline but provides insights into locally hosted LLMs for medical NLP.
- System error type:
  - **Confabulation:** Extracting nonexistent drugs (e.g., “herceptin” from a radiology report).
  - **Inclusion of Non-Chemotherapy Drugs:** Extracting unrelated medications (e.g., “prednisone” for immunosuppression).
  - **Omission of Clearly Mentioned Drugs:** Missing documented drugs (e.g., failing to extract “aflibercept”).
- Despite the objectively poor performance, the results prove: local LLMs, like Mixtral, currently perform at a GPT-3 level but have potential for improvement.

Conclusion and Future Work

Key Contributions

- Explore the use of local LLMs (e.g., Mixtral) for extracting chemotherapy timelines.
- Show that domain-specific fine-tuning is not required for basic medical text extraction.
- Build a baseline for future improvements in medical NLP.

Future Improvements

- **Enhancing Context Understanding**: Use retrieval-augmented generation (RAG) to incorporate external medical data.
- **Improving Prompt Strategies**: Implement ensemble prompting techniques like "Tree of Thought."
- **Expanding Real-World Applications**: Optimize local LLMs for secure and private hospital environments.

Questions?

Thank you!