Overview of the 2024 Shared Task on Chemotherapy Treatment Timeline Extraction

*Jiarui Yao[1], *Harry Hochheiser[2], WonJin Yoon[1], Eli Goldner[1], Guergana Savova[1]
[1]Boston Children's Hospital and Harvard Medical School
{jiarui.yao, wonjin.yoon, eli.goldner, guergana.savova}@childrens.harvard.edu
[2]University of Pittsburgh
harryh@pitt.edu

Abstract

The 2024 Shared Task on Chemotherapy Treatment Timeline Extraction aims to advance the state of the art of clinical event timeline extraction from the Electronic Health Records (EHRs). Specifically, this edition focuses on chemotherapy event timelines from EHRs of patients with breast, ovarian and skin cancers. These patient-level timelines present a novel challenge which involves tasks such as the extraction of relevant events, time expressions and temporal relations from each document and then summarizing over the documents. De-identified EHRs for 57,530 patients with breast and ovarian cancer spanning 2004-2020, and approximately 15,946 patients with melanoma spanning 2010-2020 were made available to participants after executing a Data Use Agreement. A subset of patients is annotated for gold entities, time expressions, temporal relations and patient-level timelines. The rest is considered unlabeled data. In **Subtask1**, gold chemotherapy event mentions and time expressions are provided (along with the EHR notes). Participants are asked to build the patient-level timelines using gold annotations as input. Thus, the subtask seeks to explore the topics of temporal relations extraction and timeline creation if event and time expression input is perfect. In **Subtask2**, which is the realistic real-world setting, only EHR notes are provided. Thus, the subtask aims at developing an end-to-end system for chemotherapy treatment timeline extraction from patient's EHR notes. There were 18 submissions for Subtask 1 and 9 submissions for Subtask 2. The organizers provided a baseline system. The teams employed a variety of methods including Logistic Regression, TF-IDF, n-grams, transformer models, zero-shot prompting with Large Language Models (LLMs), and instruction tuning. The gap in performance between prompting LLMs and finetuning smaller-sized LMs indicates that for a challenging task such as patient-level chemotherapy timeline extraction, more sophisticated LLMs or prompting techniques are necessary in order to achieve optimal results as finetuing smaller-sized LMs outperforms by a wide margin.

1 Introduction

Cancer treatment is rarely simple. Complex protocols involving multiple drugs, given over extended period of times in specified orders, are the norm (Warner et al., 2019). This poses a challenge for clinical researchers. Ideally, real-world studies of the impact of specific protocols would require to identify which patients have been given which protocols. In practice, this task is complicated by a dearth of detailed information: although medication records and clinical notes might indicate the administration of a given chemotherapeutic agent to a patient, they rarely, if ever, name specific protocols. Furthermore, structured medication administration records are insufficient, as clinical notes may contain mentions of medications in the context of reasons for discontinuing treatment, prior treatments given at differing institutions, or reactions to treatment.

Extracting chemotherapy timelines from clinical notes involves a series of challenges. Individual mentions of relevant drug administrations (chemotherapy events) must be extracted and mapped to appropriate medication terminologies. Each event must then be assigned a time extent, based on the date of the note and any temporal modifiers and indicators (e.g. time expressions) identified alongside the medication event (Laparra et al., 2018). Finally, these individual instances must be ordered into a timeline. Each of these tasks involves substantial challenges, several of which have been the focus of previous SemEval challenges (Elhadad et al., 2015; Laparra et al., 2018; Bethard et al., 2017).

* indicates co-first authors.

Proceedings of the 6th Clinical Natural Language Processing Workshop, pages 557–569
June 21, 2024 ©2024 Association for Computational Linguistics

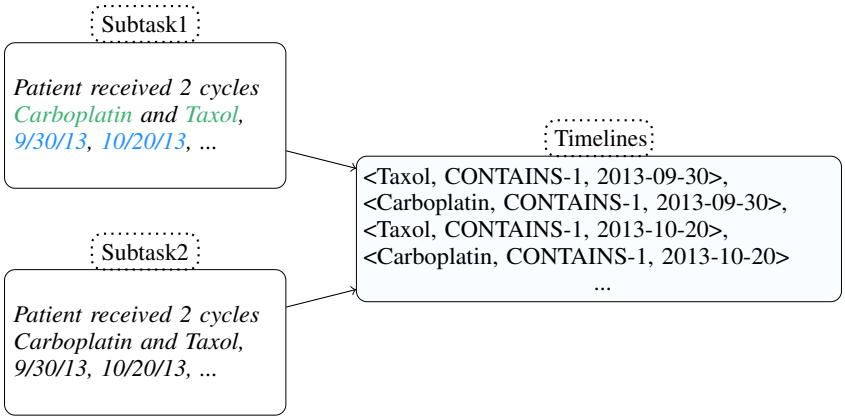

Figure 1: Illustration of the two subtasks in the 2024 Chemotherapy Treatment Timeline Extraction shared task. The input of Subtask1 is patient notes with gold events (highlighted in green) and time expressions (highlighted in blue). The input of Subtask2 is patient notes only. The output of both subtasks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

The 2015-2021 SemEval shared tasks (Bethard et al., 2015, 2016, 2017; Laparra et al., 2018, 2021) on temporal relation extraction from the clinical narrative used the THYME and THYME2 corpora (Styler IV et al., 2014; Wright-Bettner et al., 2020), each with a separate focus on one of the following tasks – pairwise temporal relation extraction, time expression normalization, and domain adaptation. The SemEval shared tasks provided the gold event and time expressions so that the teams focus on the temporal relation extraction to advance approach development. The state-of-the-art methodologies and results they established allowed the community to start exploring applications to real world biomedical use cases.

The 2024 Chemotherapy Treatment Timeline Extraction shared task* elevates the technical challenges to a new level by presenting participants with two challenges: assembling timelines from individual event mentions and temporal/time expressions provided as input (Subtask1), and building timelines directly from clinical notes, thus the real-world task of end-to-end extraction (Subtask2). Both subtasks go beyond the 2015-2021 Semeval shared tasks, however they build on the community knowledge advanced through them. For the 2024 Chemotherapy Treatment Timeline Extraction shared task the organizers provided a dataset of the Electronic Health Records (EHRs) of more than 73,000 cancer patients from 2004-2020 from University of Pittsburgh Medical Center (UPMC).

In the next sections, we describe the shared task, its substasks, the dataset, the evaluation methodology, the baseline system, the teams with highlights of their approaches, and finally the results. Details of each team's approach is described in a separate paper by the team.

2 Description of the Shared Task and Subtasks

The overall goal of the task was to create patient-level timelines of *chemotherapy treatment events* from all the notes in the EHR available for a given patient. In general, timelines can be represented in different formats. We can describe a patient's treatment timeline in natural language, such as "*2 cycles Carboplatin and Taxol, 9/30/13, 10/20/13*" which is easy to understand by humans, however, it cannot be "understood" directly by machines. Over the years, the research community has developed a parsimonious set of relations to express temporality between two events or between an event and temporal/time expression (Wright-Bettner et al., 2020; Styler IV et al., 2014). We adopt these conventions where an event is any occurrence that can be positioned on a timeline (in our case chemotherapy events) and the set of temporal relations are defined as BEFORE, CONTAINS (with inverse CONTAINS-1 which is the equivalent of CONTAINED-BY), OVERLAP, NOTED-ON, BEGINS-ON, ENDS-ON. We limit events to only chemotherapy treatment events. Therefore, for the shared task we represent the chemotherapy treatment timelines in a computable format as a list of <*chemotherapy, temporal_relation, time_expression*> triplets. Thus, the previous ex-

*https://sites.google.com/view/chemotimelines2024

	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.

ample can be converted to:

<Carboplatin, CONTAINS-1, 2013-09-30>,

<Taxol, CONTAINS-1, 2013-09-30>,

<Carboplatin, CONTAINS-1, 2013-10-20>,

<Taxol, CONTAINS-1, 2013-10-20>.

With this representation, the construction of chemotherapy treatment timelines can be naturally decomposed into the following stages: chemotherapy event extraction, time expression extraction, temporal relation classification, time expression normalization and patient-level timeline refinement. Time expressions are also referred to as temporal expressions and TIMEX3.

The shared task defined two subtasks. In **Subtask1**, gold chemotherapy event mentions and time expressions are provided (along with the EHR notes). Participants were asked to build the patient-level timelines using gold annotations as input. Thus, the subtask sought to explore the topics of temporal relation extraction and timeline creation if event and time expression input is perfect. In **Subtask2**, which is the realistic real-world setting, only EHR notes are provided. Thus, the subtask aimed at developing an end-to-end system for chemotherapy treatment timeline extraction from patient's EHR notes. Figure 1 is an overview of this task.

2.1 Data

The EHR for each patient included all types of available notes regardless of their relevance to the patient's cancer, e.g. radiology reports, pathology notes, clinical notes, oncology notes, discharge summaries, progress reports, etc. We sampled a subset of patients to create the gold annotations. For the gold annotations, we follow the THYME2 annotation schema (Wright-Bettner et al., 2020; Styler IV et al., 2014) as it is widely used in the clinical temporal relation classification community (Bethard et al., 2015, 2016, 2017; Lin et al., 2019, 2021). Two domain experts created gold annotations of the chemotherapy events, time expressions, and temporal relations. These represent instance-level annotations. These pairwise gold annotations are in the Anafora [†] (Chen and Styler, 2013) xml format. The final gold patient-level timeline was created automatically by merging all instance-level annotations followed by deduplicating and collapsing temporal relations. The gold dataset was split into training, development (dev) and test sets. Table 1 and Table 2 present the distributions of the gold dataset (*the Labeled Dataset*).

Additionally, we provided the *Unlabeled Dataset* which consists of the UPMC EHR notes for 57,530 patients with breast and ovarian cancer, collected between 2004-2020, and 15,946 patients with melanoma, collected between 2010-2020. As implied by its name, this dataset does not have any gold annotations. The *Unlabeled dataset* could potentially be used for continued training of pre-trained language models or for pretraining a language model.

To access both *Labeled* and *Unlabeled* datasets, the PI (Principal Investigator) of each team was

[†] https://github.com/weitechen/anafora

required to execute a Data Use Agreement (DUA) with University of Pittsburgh. The process took 3-4 weeks on the average. Upon execution of DUAs, the data were distributed to the teams through Globus [‡] with gated Collections for each split and dataset. Globus provides a secure way of sharing the sensitive patient EHR data.

3 Evaluation

We used the standard F1 metric to evaluate system performance, with variations to reflect the real world use case of chemotherapy treatment timelines. In consultation with our oncology domain experts it was determined that the level of granularity most useful for both point of care and translational studies is the month and the year for the chemotherapy treatment; the exact date was not deemed critical.

Therefore, we designed four evaluation strategies with different levels of granularity: strict, relaxed-to-day, relaxed-to-month and relaxed-to-year. Strict evaluation requires all elements in a triplet to match the corresponding ones in the gold annotations to count as a match. In all relaxed evaluations, we consider certain temporal relations interchangeable, and only compare the predicted month (relaxed-to-month) or year (relaxed-to-year) with the gold ones. For instance, under relaxed-to-month evaluation, we consider <TC, BEGINS-ON, 2013-02> correct if the gold timeline is <TC, BEGINS-ON, 2013-02-13>. In this shared task, based on our consultations with our oncology domain experts as described above we use the relaxed-to-month metric as the official score for the leader board and rankings.

Our scoring metrics account for differences in patterns of chemotherapy treatments. Most, but not all patients have chemotherapy. Some melanoma patients, for example, are treated surgical with no chemotherapy. To handle these differences, we used two types of scores based on relaxed-to-month results as motivated above:

- Type A: F1 where all patients are included regardless of whether they have chemotherapy gold timelines.
- Type B: F1 where patients with no chemotherapy timelines are excluded.

Type A score aims to catch false positives for these patients. Type B score measures the effectiveness of the methods on patients with confirmed chemotherapy treatments. The F1 score for each patient was computed and the final F1 score for each type is the average across all patients. The Official score used for the rankings in the Leader Board is the average of Type A and Type B. A link to the evaluation script[§] is posted on the shared task website.

Teams uploaded their systems output into their gated Globus collection and the organizers ran the evaluation script to produce the results posted on the Leader Board on the shared task website. Each team was allowed to upload up to three submissions for each task.

4 Baseline Systems

The shared task organizers provide baseline results for Subtask1 and Subtask2.

For both subtasks we used Apache cTAKES[¶] (Savova et al., 2010) for sentence boundary detection, tokenization, and pipelining of software components via the Python bridge to Java (`ctakes-pbj`) module. We use Huggingface Transformers (Wolf et al., 2019) for model training and inference, and CLUlab Timenorm's synchronous context free grammar module (Bethard, 2013) for normalizing time expressions to ISO standard. The system processes all the patients and notes for a given cancer type and split of the dataset. We processed patients by cancer type and dataset split since there are overlapping patient identifiers across different cancer types and splits (although the patients are different).

4.1 Subtask1

We used cTAKES' default tokenization and sentence splitting stack, then loaded chemotherapy event mentions and time expressions from the annotated gold data. We normalized as many time expressions as possible using Timenorm. Taking all the relevant pairs of chemotherapy event mentions and normalized time expressions, i.e. within a certain number of tokens from each other, we generated instances for classification by our temporal relation model (described below). Following (Lin et al., 2021), we used tags to distinguish the chemotherapy event mentions from the time expressions, e.g. *The patient received <e> paclitaxel*

[‡] https://www.globus.org

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

[¶] https://ctakes.apache.org

</e> *on* <t> *February 2nd, 2011* </t>. Note, in the generated instance we used the original text of the time expression, not its normalized form from Timenorm (i.e. `2011-02-02`). The normalized form is associated with its source time expression in a data structure within cTAKES and is used later when collecting instances for summarization and scoring.

For the temporal relation classification model we used Microsoft Research's PubMedBERT (Gu et al., 2020), and first fine-tuned on the THYME2 clinical temporal relation dataset (Wright-Bettner et al., 2020), then continue fine-tuned on the shared task training set to produce the type of temporal relation. Finally, when all the pairs have been classified, we generated a text table, with a row for each classified pair. Each row contains the original text of the chemotherapy mention, the normalized form of its paired time expression, their predicted temporal relation, and the identifier of the patient with whom this instance is associated. We then processed this table into a collection of summarized patient-level timelines for each patient.

To derive the patient-level timelines, we refined the pairwise temporal relations by 1) deduplication, and 2) choosing the most specific temporal relation between a chemotherapy treatment and a time expression following a predefined label hierarchy (BEGINS-ON/ENDS-ON > CONTAINS/CONTAINS-1 > BEFORE). In addition, for generic chemotherapy mentions such as "chemo" and "chemotherapy", we added them to the final timelines only if there was not a more specific chemotherapy treatment (e.g. Taxol) having the same temporal relation with the exact same time expression.

4.2 Subtask2

Here we also used cTAKES' default sentence detection and tokenization stack. For detecting chemotherapy mentions, we used cTAKES' dictionary lookup module with a customized dictionary of common chemotherapy terms collected from the training split of the shared task gold annotated corpus to identify potential chemotherapy mentions in each note. For detecting time expressions, we used the SVM-based tagger in the cTAKES' temporal module to identify potential time expressions, then normalize as many potential time expressions with Timenorm as possible. As in Subtask1, we generated instances for temporal relation classification from all relevant pairs of chemotherapy mentions and normalized time expressions, along with a table of the classified instances and relevant associated information for further summarization and evaluation. We used the same model for temporal relation classification as in Subtask1. We provided a docker implementation [‖] of the baseline system for Subtask 2 as a resource on the shared task website.

5 Participating Systems

In this section, we briefly describe the approaches of participating systems. Details of each system can be found in the separate papers by each of the team.

The participants explored a variety of methods, including Logistic Regression, TF-IDF, n-grams, transformer models, zero-shot prompting with Large Language Models (LLMs), and instruction tuning. Table 3 summarizes all teams' approaches.

BioCom participated in Subtask 1. They utilized SciSpacy for Named Entity Extraction (NER) and Logistic Regression to classify temporal relations. They used unigram Term Frequency-Inverse Document Frequency (TF-IDF) to get features from the input text.

ClinicalRxMiners submitted two systems for Subtask 1. In submission 1, ClinicalRxMiners utilized a machine learning (non-deep learning) approach and employed n-grams as features of the input, with a soft voting classifier as the model for making predictions. In submission 2, ClinicalRxMiners utilized a pretrained Language Model (LM) named GLiNER (Zaratiana et al., 2023), which is specialized for NER.

KCLab (Tan et al., 2024) utilized a hybrid method, employing cTAKES (Savova et al., 2010) for preprocessing and PubMedBERT (Gu et al., 2020) for post-processing. Their system was built on top of the baseline model provided by the organizers. Additionally, KCLab used the UMLS (Bodenreider, 2004) database. KCLab participated in both Subtask1 and Subtask2.

LAILab (Haddadan et al., 2024) utilized two approaches: supervised fine-tuning of language models and a pipeline approach combining rule-based NER with deep learning based relation classification. For Subtask 1, they finetuned

[‖] https://github.com/HealthNLPorg/chemoTimelinesBaselineSystem

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.

`flan-T5-XXL` (Chung et al., 2022). For Subtask 2, they used a sequence-to-sequence approach in the first two submissions, and a lookup table for chemotherapy event extraction with a deep learning method for temporal relation classification in the third submission.

Lexicans (Sharma et al., 2024) used LLMs with zero-shot prompting to extract relations. They also utilized the THYME ontology to formalize the representation of entities and their relationships. A few LLMs such as Llama2, Mistral, Zephyr, Meditron, and Mixtral (Touvron et al., 2023; Jiang et al., 2023; Tunstall et al., 2023; Chen et al., 2023) were tested under various settings. Additionally, a data normalization step was performed to transform time entities into absolute date-time formats.

NLPeers (Bannour et al., 2024) developed two systems, both submitted for Subtask1. For submission 1, NLPeers fine-tuned the `microsoft/deberta-v3-base` model and used it for temporal relation classification. Additionally, the Heideltime library[**] (Strötgen and Gertz, 2010) and an LLM-based prompt with the OpenChat 3.5 model (Wang et al., 2024a) were used to normalize time expressions. For submission 2, the NLPeers team applied few-shot prompting with the `Mixtral-8X7B-Instruct-v0.1` model (Jiang et al., 2023), the prompt was chosen by DSPy (Khattab et al., 2023), a framework for algorithmically optimizing LM prompts. A Chain-Of-Thought (Wei et al., 2022) approach was integrated during the prompt searching step by DSPy. For time expression normalization, Heideltime was also used in submission 2.

NYULangone employed an LLM-based prompt approach with minimal pre- and post-processing. NYULangone participated only in Subtask2, which means the team did not use the gold annotation provided in Subtask1.

UTSA-NLP (Zhao and Rios, 2024) presented an instruction-tuning based approach. The UTSA-NLP team reformulated the task into a question-answering (QA) dataset for both the entity extraction step and temporal relation classification step, then instruction-tuned an LLM, `OpenChat-3.5-7B`, on the QA dataset. The team continued pre-training the instruction-tuned model on a portion of the *Unlabeled* dataset in one of their submissions. For the temporal relation classification step, they used an open-sourced LLM to generate reasoning for the answer.

Wonder (Wang et al., 2024b) participated in Subtasks 1 and 2. They employed a supervised fine-tuning approach, formulating the task as a multi-class sentence classification task, where the input was the text between the event and time expression. For Subtask 2, MedTagger [††] was used to identify all the potential EVENT-TIMEX3 pairs. Time ex-

[**] https://github.com/HeidelTime/heideltime

[††] https://github.com/OHNLP/MedTagger

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

Table 4: Evaluation results of Subtask1 (test set). All scores are macro-F1 of relaxed-to-month setting. We compute two types of scores: F1 with patients with no gold timelines (Type A) and F1 without patients with no gold timelines (Type B). Official score is the average of Type A and Type B, which is used for the rankings in the leader board. See details in section 3.

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

Table 5: Evaluation results, Subtask 2 (test set). All scores are macro-F1 of relaxed-to-month setting. We compute two types of scores: F1 with patients with no gold timelines (Type A) and F1 without patients with no gold timelines (Type B). Official score is the average of Type A and Type B, which is used for the rankings in the leader board. See details in section 3.

pressions were normalized with MedTime (Sohn et al., 2013).

6 Results and Discussion

Overall results are presented in Table 4 and 5. Results per type of cancer are presented in Table 6 and 7 in the Appendix.

Most teams employed deep-learning-based methods for this shared task. Two teams used non-deep-learning models: ClinicalRXMiners submission 1 used a machine learning model, BioCom trained a Logistic Regression system. For the event mention extraction step, the Wonder team and the baseline system used off-the-shelf tools for time expression extraction and normalization. LAILab used a lookup table for chemotherapy event identification

in one of their submissions. The other approaches employed in this shared task include end-to-end timeline building (meaning no separate steps for event mention extraction), supervised event mention extraction model, and zero-shot prompting.

Finetuning LMs: For both subtasks, the top teams, i.e. LAILab and Wonder, employed finetuned pretrained language models as the core technology. LAILab finetuned `Flan-T5-xxl` (Chung et al., 2022) and `Bart-large` (Lewis et al., 2020a), which have 11B and 400M parameters respectively. They achieve best performance on all subtasks (overall and per type of cancer) except for Subtask2, breast cancer. The Wonder team finetuned Bio-LM (Lewis et al., 2020b), yielding top 3 results across all subtasks (excluding the baseline system). The other two teams with good results are NLPeers and KCLab, who finetuned `deberta-v3-base` (He et al., 2023) and PubMedBert (Gu et al., 2020) respectively. Overall, the commendable performances of those teams suggest that finetuning LMs remains the optimal approach for optimizing system performance if gold labeled data and computing resources are available.

Prompting LLMs: A few teams took the approach of prompting LLMs. The Lexicans team experimented with zero-shot prompting of 5 different LLMs, namely LLAMA2 , Mistral, Zephyr, Meditron, and Mixtral (Touvron et al., 2023; Jiang et al., 2023; Tunstall et al., 2023; Chen et al., 2023). NYULangone applied zero-shot prompting with the Mixtral model. Submission 2 from the NLPeers team prompted the `Mixtral-8X7B-Instruct-v0.1` model in a few-shot fashion.

The gap in performance between prompting LLMs and finetuning smaller-sized LMs indicates that for a challenging task such as patient-level chemotherapy timeline extraction, more sophisticated LLMs or prompting techniques are necessary in order to achieve optimal results. The state-of-the-art results for the 2024 Chemotherapy Treatment Timeline Extraction shared task are established by fine-tuning smaller LMs.

A comparison of the scores between Subtask1 and Subtask2 shows a substantial drop of at least 0.2 F1 Official Score when gold event and time expressions (thus perfect input) are provided. This gap, surprisingly, implies that what is considered the easier task of event and time expression extraction is not a solved problem while the task of temporal relation extraction holds strong.

7 Conclusion

The 2024 Shared Task on Chemotherapy Treatment Timeline Extraction is unique in both (1) focusing on a highly complex task, and (2) providing a large corpus of EHR data to the participants. The community embraced the task with enthusiasm and employed diverse methodologies, thus enabling robust comparison of approaches. Perhaps surprising in our current era of very large LMs, fine-tuned smaller LMs achieved superior performance. This discrepancy between prompting LLMs and finetuning smaller-sized LMs suggests that more sophisticated LLMs or prompting techniques are necessary in order to achieve optimal results for challenging tasks such as patient-level chemotherapy timeline extraction.

8 Acknowledgements

We are very grateful for our annotators David Harris and Gabrielle Dihn who spent days creating the gold annotations. We are grateful for our oncology domain experts Drs. Danielle Bitterman, Jeremy Warner, Piet de Groen and Elizabeth Buchbinder for guidng us through the thickness of the oncology domain. Funding is provided by the United States National Institutes of Health (grants U24CA248010, R01LM010090, R01LM013486, R01LM012973, R01MH126977). The content is solely the responsibility of the authors and does not necessarily represent the official views of the United States National Institutes of Health.

Limitations

There are different types of cancer treatments, such as Immunotherapy, Radiation Therapy, Surgery and Targeted Therapy. In this shared task, we only focus on chemotherapy treatments. We leave the timeline construction of other types of therapy for future research.

Ethics Statement

All the data used in this shared task are de-identified patient notes. The access the data, the PI of each team was required to execute a Data Use Agreement with University of Pittsburgh. The data were distributed through Globus, which provides a secure way of sharing senstitive data such as patient EHRs. Participants were also required to

submit the final timelines via Globus, to protect the patient privacy.

References

Nesrine Bannour, Judith Jeyafreeda Andrew, and Marc Vincent. 2024. 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? In *Proceedings of the 6th Clinical NLP Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Steven Bethard. 2013. A synchronous context free grammar for time normalization. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 821–826, Seattle, Washington, USA. Association for Computational Linguistics.

Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, Denver, Colorado. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1052–1062, San Diego, California. Association for Computational Linguistics.

Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32 Database issue:D267–70.

Wei-Te Chen and Will Styler. 2013. Anafora: A web-based general purpose annotation tool. In *Proceedings of the 2013 NAACL HLT Demonstration Session*, pages 14–19, Atlanta, Georgia. Association for Computational Linguistics.

Zeming Chen, Alejandro Hern'andez Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Kopf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *ArXiv*, abs/2311.16079.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. SemEval-2015 task 14: Analysis of clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310, Denver, Colorado. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

Shohreh Haddadan, Tuan-Dung Le, Thanh Duong, and Thanh Q Thieu. 2024. Lailab at chemotimelines 2024: Finetuning sequence-to-sequence language models for temporal relation extraction towards cancer patient undergoing chemotherapy treatment. In *Proceedings of the 6th Clinical NLP Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.

Egoitz Laparra, Xin Su, Yiyun Zhao, Özlem Uzuner, Timothy Miller, and Steven Bethard. 2021. SemEval-2021 task 10: Source-free domain adaptation for semantic processing. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 348–356, Online. Association for Computational Linguistics.

Egoitz Laparra, Dongfang Xu, Ahmed Elsayed, Steven Bethard, and Martha Palmer. 2018. SemEval 2018 task 6: Parsing time normalizations. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 88–96, New Orleans, Louisiana. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020b. Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2021. EntityBERT: Entity-centric masking strategy for model pretraining for the clinical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 191–201, Online. Association for Computational Linguistics.

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.

Vishakha Sharma, Andres Fernandez, Andrei Constantin Ioanovici, David Talby, and Frederik Buijs. 2024. Lexicans at chemotimelines 2024: Chemotimeline chronicles - leveraging large language models (llms) for temporal relations extraction in oncological electronic health records. In *Proceedings of the 6th Clinical NLP Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Sunghwan Sohn, Kavishwar B. Wagholikar, Dingcheng Li, Siddhartha R. Jonnalagadda, Cui Tao, K. E. Ravikumar, and Hongfang Liu. 2013. Comprehensive temporal information detection from clinical text: medical events, time, and tlink identification. *Journal of the American Medical Informatics Association : JAMIA*, 20 5:836–42.

Jannik Strötgen and Michael Gertz. 2010. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324, Uppsala, Sweden. Association for Computational Linguistics.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Yukun Tan, Merve Dede, and Ken Chen. 2024. Kclab at chemotimelines 2024: End-to-end system for chemotherapy timeline extraction - subtask2. In *Proceedings of the 6th Clinical NLP Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2024a. Openchat: Advancing open-source language models with mixed-quality data. In *The Twelfth International Conference on Learning Representations*.

Liwei Wang, Qiuhao Lu, Rui Li, Sunyang Fu, and Hongfang Liu. 2024b. Wonder at chemotimelines 2024: Medtimeline: An end-to-end nlp system for timeline extraction from clinical narratives. In *Proceedings of the 6th Clinical NLP Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

Jeremy L. Warner, Dmitry Dymshyts, Christian G. Reich, Michael J. Gurley, Harry Hochheiser, Zachary H. Moldwin, Rimma Belenkaya, Andrew E. Williams, and Peter C. Yang. 2019. HemOnc: A new standard vocabulary for chemotherapy regimen representation in the OMOP common data model. 96:103239.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Kristin Wright-Bettner, Chen Lin, Timothy Miller, Steven Bethard, Dmitriy Dligach, Martha Palmer, James H. Martin, and Guergana Savova. 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*, pages 104–114, Online. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. Gliner: Generalist model for named entity recognition using bidirectional transformer.

Xingmeng Zhao and Anthony Rios. 2024. Utsa-nlp at chemotimelines 2024: Evaluating instruction-tuned language models for temporal relation extraction. In *Proceedings of the 6th Clinical NLP Workshop*, Mexico City, Mexico. Association for Computational Linguistics.

9 Appendix

Submission (Breast)	Type A	Type B	Official Score
LAILab_submission1	0.97	0.94	0.96
LAILab_submission3	0.97	0.94	0.95
LAILab_submission2	0.97	0.94	0.95
Baseline_subtask1	0.95	0.91	0.93
Wonder_submission1	0.94	0.87	0.90
Wonder_submission2	0.93	0.87	0.90
Wonder_submission3	0.93	0.87	0.90
BioCom_submission1	0.92	0.85	0.88
KCLab_submission1	0.84	0.68	0.76
NLPeers_submission1	0.79	0.66	0.72
UTSA-NLP_submission1	0.79	0.60	0.70
UTSA-NLP_submission3	0.79	0.60	0.69
UTSA-NLP_submission2	0.79	0.59	0.69
Lexicans_submission1	0.78	0.58	0.68
Lexicans_submission2	0.77	0.55	0.66
Lexicans_submission3	0.74	0.49	0.62
NLPeers_submission2	0.63	0.34	0.49
ClinicalRXMiners_submission1	0.49	0.39	0.44
ClinicalRXMiners_submission2	0.49	0.18	0.33

Submission (Melanoma)	Type A	Type B	Official Score
LAILab_submission1	0.93	0.81	0.87
Baseline_subtask1	0.92	0.81	0.87
LAILab_submission2	0.91	0.79	0.85
NLPeers_submission1	0.91	0.78	0.84
Wonder_submission2	0.91	0.78	0.84
Wonder_submission1	0.91	0.78	0.84
LAILab_submission3	0.91	0.77	0.84
Lexicans_submission1	0.90	0.76	0.83
NLPeers_submission2	0.89	0.73	0.81
Lexicans_submission2	0.88	0.71	0.80
Wonder_submission3	0.86	0.65	0.76
UTSA-NLP_submission1	0.82	0.55	0.68
UTSA-NLP_submission3	0.82	0.54	0.68
UTSA-NLP_submission2	0.80	0.51	0.65
BioCom_submission1	0.78	0.45	0.61
KCLab_submission1	0.77	0.42	0.60
Lexicans_submission3	0.77	0.42	0.59
ClinicalRXMiners_submission2	0.70	0.24	0.47
ClinicalRXMiners_submission1	0.67	0.17	0.42

Submission (Ovarian)	Type A	Type B	Official Score
LAILab_submission3	0.93	0.86	0.89
LAILab_submission2	0.93	0.85	0.89
LAILab_submission1	0.92	0.84	0.88
Baseline_subtask1	0.92	0.83	0.88
Wonder_submission2	0.84	0.69	0.77
Wonder_submission3	0.83	0.67	0.75
NLPeers_submission1	0.83	0.66	0.75
Wonder_submission1	0.83	0.66	0.74
BioCom_submission1	0.82	0.63	0.72
UTSA-NLP_submission2	0.80	0.59	0.70
UTSA-NLP_submission3	0.80	0.59	0.70
UTSA-NLP_submission1	0.79	0.58	0.69
NLPeers_submission2	0.75	0.50	0.63
Lexicans_submission3	0.74	0.49	0.62
Lexicans_submission1	0.74	0.48	0.61
Lexicans_submission2	0.73	0.46	0.59
KCLab_submission1	0.68	0.37	0.53
ClinicalRXMiners_submission2	0.48	0.21	0.34
ClinicalRXMiners_submission1	0.39	0.27	0.33

Table 6: Subtask 1, per type of cancer (test set). All scores are macro-F1 of relaxed-to-month setting. We compute two types of scores: F1 with patients with no gold timelines (Type A) and F1 without patients with no gold timelines (Type B). Official score is the average of Type A and Type B, which is used for the rankings in the leader board. See details in section 3

Submission (Breast)	Type A	Type B	Official Score
KCLab_submission1	0.71	0.65	0.68
Wonder_submission2	0.70	0.57	0.64
Wonder_submission1	0.70	0.57	0.63
Wonder_submission3	0.69	0.57	0.63
LAILab_submission2	0.68	0.55	0.62
Baseline_subtask2	0.61	0.57	0.59
LAILab_submission3	0.47	0.58	0.53
LAILab_submission1	0.54	0.49	0.52
UTSA-NLP_submission1	0.32	0.18	0.25
NYULangone_submission1	0.17	0.21	0.19

Submission (Melanoma)	Type A	Type B	Official Score
LAILab_submission2	0.78	0.70	0.74
LAILab_submission1	0.68	0.45	0.57
KCLab_submission1	0.64	0.35	0.49
Baseline_subtask2	0.60	0.26	0.43
Wonder_submission3	0.37	0.42	0.39
Wonder_submission1	0.37	0.42	0.39
Wonder_submission2	0.37	0.41	0.39
LAILab_submission3	0.43	0.33	0.38
NYULangone_submission1	0.40	0.25	0.32
UTSA-NLP_submission1	0.12	0.30	0.21

Submission (Ovarian)	Type A	Type B	Official Score
LAILab_submission2	0.83	0.65	0.74
Baseline_subtask2	0.80	0.61	0.71
LAILab_submission1	0.73	0.46	0.59
Wonder_submission3	0.70	0.40	0.55
Wonder_submission2	0.70	0.39	0.55
Wonder_submission1	0.69	0.38	0.53
LAILab_submission3	0.49	0.49	0.49
KCLab_submission1	0.55	0.35	0.45
UTSA-NLP_submission1	0.21	0.17	0.19
NYULangone_submission1	0.21	0.16	0.18

Table 7: Subtask 2, results for each type of cancer (test set). All scores are macro-F1 of relaxed-to-month setting. We compute two types of scores: F1 with patients with no gold timelines (Type A) and F1 without patients with no gold timelines (Type B). Official score is the average of Type A and Type B, which is used for the rankings in the leader board. See details in section 3