Overview of the 2025 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 [2]University of Pittsburgh

{jiarui.yao, wonjin.yoon, eli.goldner, guergana.savova}@childrens.harvard.edu
harryh@pitt.edu

Abstract

Extracting patient treatment timelines from clinical notes is a complex task involving identification of relevant events, temporal expressions, and temporal relations in individual documents and developing cross-document summaries. The 2025 Shared Task on Chemotherapy Treatment Timeline Extraction builds upon the initial 2024 challenge, using data from 57,530 breast and ovarian cancer patients and 15,946 melanoma patients. Participants were provided with a subset annotated for treatment entities, temporal expressions, temporal relations, and timelines for each patient. This training data was used to addressed two subtasks. **Subtask 1** focused on extracting temporal relations and creating timelines, given documents and gold-standard events and temporal expressions. **Sutask 2** involved development of an end-to-end system involving extraction of entities, temporal expressions, and relations, and construction of timelines, given only the Electronic Health Record notes. Five teams participated, submitting eight entries for Subtask 1 and twelve for Subtask 2. Supervised fine-tuning remains a productive approach albeit with a shift of supervised fine-tuning of very large language models compared to the 2024 task edition. Even with the much more "strict" evaluation metric, the best results are comparable to the best less strict 2024 relaxed-to-month results.

1 Introduction

As cancer patients are generally treated through detailed protocols involving multiple drugs administered in well-defined patterns over extended periods of time (Warner et al., 2019), identification of specific protocols administered to patients is a critical step in retrospective cancer research. Unfortunately, extraction of this information from real-world data is often challenging, as notes and medication administration records may indicate which treatments are given and when, but specific protocols are rarely named. Extraction of these details from clinical notes is a challenging task, involving multiple steps. First, mentions of administrations of chemotherapeutic agents must be extracted and normalized. The date of the notes and any temporal modifiers must then be used to assign a temporal extent to the medication event (Laparra et al., 2018). Finally, all events must be assembled into an ordered timeline. Several of these tasks have been the focus of previous SemEval challenges (Elhadad et al., 2015; Laparra et al., 2018; Bethard et al., 2017).

Temporal relations extraction challenges including temporal relation extraction, time expression normalization, and domain adaptation, were the focus of the 2015-2021 SemEval shared tasks (Bethard et al., 2015, 2016, 2017; Laparra et al., 2018, 2021) based on the THYME and THYME2 corpora (Styler IV et al., 2014; Wright-Bettner et al., 2020). To facilitate a focus on temporal relation extraction, these tasks provided the gold event and time expressions. Methodological advances resulting from these challenges enabled initial applications to real world biomedical use cases.

The 2024 Chemotherapy Treatment Timeline Extraction shared task* built on this experience with two subtasks aimed at capturing the difficulty of extracting cancer treatment information. Subtask 1 asked participants to assemble provided individual events and temporal expressions into timelines, while Subtask 2 called for the development of an end-to-end system including extraction of mentions and timeline assembly. Nine participating teams used a data set of more than 73,000 cancer patients from 2004-2020 from University of Pittsburgh Medical Center (UPMC) to complete these tasks, using a variety of models and approaches. Although most teams used deep-learning approaches,

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

Proceedings of the 7th Clinical Natural Language Processing Workshop, pages 1–10

entries were divided in their specific approaches, with some using prompting approaches for large-language models (LLMs) and others relying on fine-tuning of smaller models, with the best fine-tuned smaller models outperforming the larger models. Not surprisingly, Subtask 2 was significantly more difficult than Subtask 1 (Yao et al., 2024).

The 2025 edition of the shared task repeats the structure of the early task [†], with the expectation that substantial methodological advances in the field would encourage experimentation and yield insights into the application of state-of-the-art tools to these challenging tasks.

The next sections described the shared task in detail, including the dataset, the evaluation methodology, the and baseline system. Approaches used by each of the teams are described along with results. Additional details are provided in companion papers by the participating teams.

2 Description of the Shared Task and Subtasks

Like the 1st edition of the shared task – the 2024 Chemotherapy Timeline Extraction Shared task – the overall goal of the 2025 shared task is to create patient-level timelines of systemic anticancer therapies (SACT), which we refer to as *chemotherapy treatment events*, from all the notes in the Electronic Health Records (EHRs) available for a given patient. SACT include traditional cytotoxic chemotherapy, endocrine therapy, targeted therapy, and immunotherapy. Clinical narrative texts from the EHR often contain extensive descriptions of the temporal sequencing of SACT, presenting a valuable opportunity for automated extraction methods.

Clinical timelines require structured representation for computational processing. Following established temporal relation frameworks (Wright-Bettner et al., 2020; Styler IV et al., 2014), we model chemotherapy treatment timeline using six core temporal relations: BEFORE, CONTAINS, CONTAINS-1 (inverse containment), OVERLAP, NOTED-ON, BEGINS-ON, and ENDS-ON. Following the 2024 shared task, we formalize treatment timelines as structured triplets: <*chemotherapy_event, temporal_relation, time_expression*>, enabling direct computational analysis of SACT treatment sequences.

Thus, a sentence "*2 cycles Carboplatin and Taxol, 9/30/13, 10/20/13*" in a clinical note can be modeled as:

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

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

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

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

This representation enables a modular pipeline approach comprising chemotherapy event extraction, temporal expression (TIMEX3) identification, temporal relation classification, time normalization, and patient-level timeline assembly. The 2024 and 2025 editions of Chemotherapy Treatment Timeline Extraction shared task both contain two subtasks. Subtask 1 provides gold-standard chemotherapy events and temporal expressions alongside EHR notes, focusing participants on temporal relation extraction and timeline construction given perfect entity recognition. Subtask 2 presents the realistic scenario where only raw EHR notes are available, requiring end-to-end systems that jointly perform entity extraction and timeline reconstruction. Figure 1 illustrates the overall task framework.

2.1 Data

The 2025 edition uses the same dataset as in the previous year. We provide a brief description below, and refer readers to the 2024 overview paper (Yao et al., 2024) for further details.

We included all available EHR notes for each patient, regardless of their direct relevance to the patient's cancer. A subset of patients' EHRs was annotated with <chemotherapy_event, temporal_relation, time_expression> triplets to create the gold-standard dataset, following the THYME2 annotation schema (Wright-Bettner et al., 2020; Styler IV et al., 2014), which is widely used in the clinical temporal relation extraction research community (Bethard et al., 2015, 2016, 2017; Lin et al., 2019, 2021). The final gold-standard patient-level timelines were automatically generated by merging all instance-level annotations, followed by deduplication and collapsing of temporal relations. The gold-standard dataset was then divided 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 an *Unlabeled Dataset* containing EHR notes from UPMC for 57,530

[†] https://sites.google.com/view/chemotimelines2025/

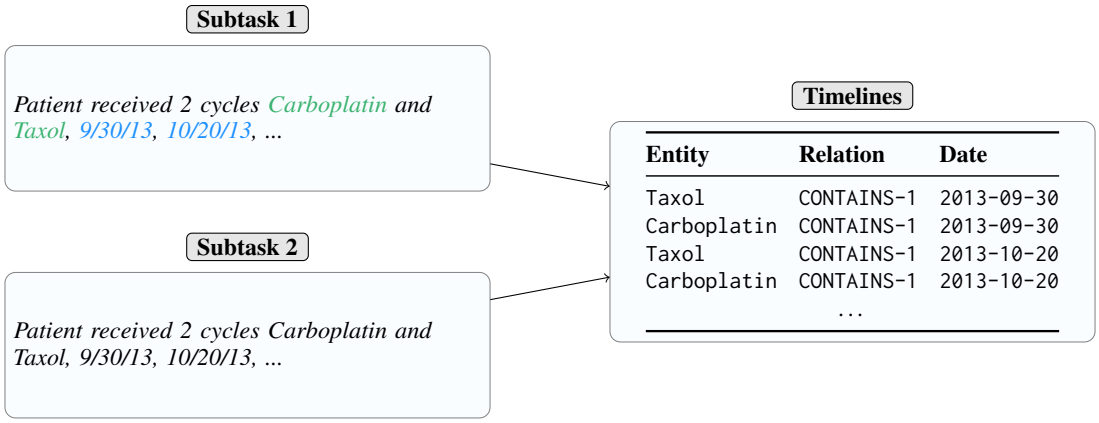

Figure 1: Illustration of the two subtasks in the 2024 and 2025 Chemotherapy Treatment Timeline Extraction shared task. The input of Subtask 1 is patient notes with gold events (highlighted in green) and time expressions (highlighted in blue). The input of Subtask 2 is patient notes only. The output of both subtasks is a list of chemotherapy treatment timelines with normalized time expressions. See details in section 2. (Figure is a re-print of Yao et al. (2024) with slight modifications).

	Train			Dev			Test		
	Patients	Notes	Words	Patients	Notes	Words	Patients	Notes	Words
Ovarian cancer	26	1,675	1,183,632	8	562	308,814	8	559	257,116
Breast cancer	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.

breast/ovarian cancer patients (2004–2020) and 15,946 melanoma patients (2010–2020). This dataset has no gold-standard annotations and may be used for continued training or pretraining of language models.

Each team's Principal Investigator (PI) was required to execute a Data Use Agreement (DUA) with the University of Pittsburgh to get access to the *Labeled* and *Unlabeled datasets*. Once the DUA was executed, data were distributed via Globus[‡].

3 Evaluation

We used the standard F1 metric to evaluate system performance. Unlike the 2024 edition of the shared task (Yao et al., 2024), we used the "strict" evaluation metric for this year's evaluation. That is, all elements in a triplet must match the corresponding ones in the gold annotations to count as a match. The 2024 edition used "relaxed-to-month" matches, where the month and year must match to count as matches. Thus, results reported in the 2025 edition are not directly comparable to the 2024 edition. The "strict" metric used in the 2025 edition presents an increased level of challenge for the participants.

We used two types of metrics to account for chemotherapy treatment patterns. Although most patients are treated with chemotherapy agents, some are not. This is particularly the case for melanoma patients, who are often treated surgically. To handle these differences, we used two types of scores based on results as motivated above:

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

Type A scores are intended to capture false positives for patients without chemotherapy treatments, while Type B score measure 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

[‡] https://www.globus.org

	Train			Dev			Test	
	EVENT	TIMEX3	TLINK	EVENT	TIMEX3	TLINK	EVENT	TIMEX3
Ovarian cancer	1,168	597	494	790	312	226	664	381
Breast cancer	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.

script[§] is posted on the shared task website. Each team was allowed to upload up to five submissions for each task.

4 Baseline System

The baseline system for both Subtask 1 and Subtask 2 is the same as in the 2024 edition; we provide a high-level description below. A detailed description can be found in Yao et al. (2024).

The baseline system is a pipeline system utilizing Apache cTAKES[¶] (Savova et al., 2010) and its python wrapper (Python bridge to Java `ctakes-pbj`) . To fine-tune PubMedBERT (Gu et al., 2020) first on the THYME2 clinical temporal relation dataset (Wright-Bettner et al., 2020) and then on the shared task dataset, the baseline system uses Huggingface Transformers (Wolf et al., 2019). Temporal expression normalization to ISO standard is done with CLUlab Timenorm's synchronous context free grammar module (Bethard, 2013). The final timeline is a summarization where the most specific temporal relation between a chemotherapy event and a temporal expression is represented following a temporal relation hierarchy. The system processes all notes for a given patient without any type of document pruning. The baseline system is available as a docker container on the shared task website [‖].

5 Participating Systems

In this section, we briefly describe the approaches of participating systems. Each team was asked to submit short system descriptions along with their official system submissions as outlined on the shared task website [**]. The system summaries below are based on these descriptions. Details of each system can be found in the separate papers by each of the team. Table 3 provides a high level summary description of the participating systems.

Brim utilized the Brim platform [††] and LLMs to extract relevant elements for Subtask 1. They applied GPT-4o (OpenAI et al., 2024) and GPT-4.1 [‡‡] to perform hierarchical question answering for this task. The Brim team received the dataset one week before the run submission deadline, therefore they only submitted one system for Subtask 1.

NLP4Health submitted 3 systems for Subtask 1. They finetuned small to mid-size opens LLMs (i.e. Llama3.2-3B, Llama3.1-8B, Grattafiori et al. 2024) for temporal relation extraction using the gold standard dataset. For the TIMEX3 normalization step, they prompted the Qwen3-14B model (Qwen Team, 2025) with zero-shot chain-of-thought (CoT) (Wei et al. 2022) prompting. In two of their submissions, they also conducted a rule-based postprocessing step. They found the model tended to incorrectly predict CONTAINS rather than ENDS-ON if there were "being on" or "was on" preceding the TIMEX3 in the text. Therefore, they used Regular Expression to adjust the final predictions for those situations. They also used the medSpaCy (Eyre et al., 2021) package for sentence segmentation during preprocessing.

NousTime participated in both Subtask 1 and Subtask 2. They prompted GPT-4o for both the temporal relation extraction step and TIMEX normalization step. Their prompt encoded extensive rule logic for inclusion/exclusion criteria, TIMEX3 normalization, and TLINK label assignment.

UAB submitted 3 systems for Subtask 1 and 1 system for Subtask 2. Their main technique was zero-shot prompting of LLMs. They used Phi4:14B (Abdin et al., 2024) and GPT-4.1-mini for their experiments. Unfortunately, in their final submissions, the team missed to submit the output with

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

[¶] https://ctakes.apache.org

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

[**] https://sites.google.com/view/chemotimelines2025/submission-of-test-output

[††] https://www.brimanalytics.com/

[‡‡] https://openai.com/index/gpt-4-1/

Team	Approach	Model	Task
Brim	LLM with the Brim platform	GPT-4o, GPT-4.1	2
NLP4Health_submission1	Supervised fine-tuning LLMs, Zero-shot Chain-of-Though prompting	Llama 3.1-8B, Qwen3-14B	1
NLP4Health_submission2,3	Supervised fine-tuning LLMs, Zero-Shot Chain-of-Thought prompting, rule-based post-processing	Llama 3.1-8B, Llama 3.2-3B, Qwen3-14B	1
NousTime	Prompting LLMs	GPT-4o	1,2
UAB_submission1,2,3	Zero-shot prompting	phi4:14b, GPT-4.1-mini	1
UAB_submission4	Zero-shot prompting	phi4:14b	2
UW-BioNLP_submission1,2	Supervised fine-tuning LLMs, Direct Preference Optimization	Qwen3-14B	2
UW-BioNLP_submission3	Lookup table for entity extraction	Qwen3-14B	2
UW-BioNLP_submission4	Prompting LLMs with thinking mode	Qwen3-30B-A3B	2

Table 3: Characteristics of participating systems. NOTE: not enough information for UW-BioNLP_submission3 provided by the participating team; no description of UW-BioNLP_submission5 provided by the participating team.

the last step of summarization, thus their results might be lower.

UW-BioNLP participated in Subtask 2. They surpervised finetuned (SFT) Qwen3-14B using the gold standard dataset. In one of their systems, they continued to train the model after SFT with Direct Preference Optimization (DPO, Rafailov et al. 2023) to align the outputs with human preferences. They also explored prompting Qwen3-30B-A3B (Qwen Team, 2025) with thinking mode.

6 Results and Discussion

Table 4 and 5 present the average F1 scores across three cancer types with the strict evaluation metric, which we use as the main metric in the Leader Board. Results per type of cancer are presented in Table 6.

Subtask 1 In Subtask1, all participating systems underperformed the baseline system, a finetuned model using PubMedBERT (Gu et al., 2020) and described in detail in Yao et al. (2024). This indicates that in well-defined information extraction biomedical tasks, domain-specific pretrained language models retain a competitive edge over general-purpose LLMs. Among the 8 participating systems, both the NousTime team and UAB team used the "prompting LLM" method, the performance discrepancy shows that prompting LLMs is far from a uniform strategy. Larger models such as GPT-4o appear more successful in a prompting setup as compared to the smaller models such as Llama 3.1/3.2 and Qwen. Model capacity, prompt design, inclusion of few-shot examples, decoding strategy, and post-processing choices all potentially contribute to the final outcome.

Subtask 2 In Table 5 which presents Subtask 2 results, we notice the big performance drop from the best Subtask 1 result (73.01 F1 for Subtask 1 and 67.81 for Subtask 2). The same is observed across team, for example, NousTime's best result in Subtask 2 is 62.41 F1, about 10 points lower than its performance in Subtask 1 (73.01 F1), showing that the end-to-end timeline extraction is a much harder task (as is expected to be the case). Ten out of the twelve Subtask 2 submissions beat the baseline system by 5.35 - 21.9 F1 points. All of those ten systems employ LLMs in various ways (SFT or prompting), which suggests that when entity recognition is part of the task, LLMs' ability to jointly extract and reason the timeline and its necessary components is a big advantage over smaller biomedical pretrained language models. Relying on the LLM's learned knowledge appears a better strategy than explicitly representing events in lookup tables (UW-BioNLP submission 1/2/4 v. UW-BioNLP submission 3). We also notice that within LLM-based strategies, fine-tuned and alignment-optimized models (e.g. UW-BioNLP submission 1/2/4) clearly outperform simple prompting-based systems.

Team	Submission	Type A	Type B	Official
NousTime	submission1	82.20	63.81	73.01
NousTime	submission2	81.77	63.05	72.41
NLP4Health	submission2	74.01	45.32	59.66
NLP4Health	submission1	73.99	45.30	59.64
NLP4Health	submission3	73.66	44.01	58.84
UAB*	submission3	47.80	20.30	34.05
UAB*	submission2	40.12	15.60	27.86
UAB*	submission1	32.62	13.18	22.90
Baseline	-	85.73	68.73	77.23

Table 4: Subtask 1 evaluation results. We report the average F1 scores across three cancer types (breast cancer, ovarian cancer, melanoma) in the dataset. Scores are with the strict evaluation metric, thus not comparable to the results from the 2024 Chemotherapy shared task which included relaxed-to-month evaluation. *: potentially without timeline summarization at the time of submission, thus results are likely lower.

Team	Submission	Type A	Type B	Official
UW-BioNLP	submission1	74.81	60.81	67.81
UW-BioNLP	submission2	74.70	58.50	66.60
UW-BioNLP	submission4	72.24	56.58	64.41
NousTime	submission4	73.02	51.80	62.41
NousTime	submission3	73.14	50.10	61.62
NousTime	submission5	72.69	49.22	60.96
UW-BioNLP	submission5	68.20	52.34	60.27
UW-BioNLP	submission3	64.85	44.10	54.48
NousTime	submission1	52.83	51.34	52.09
NousTime	submission2	55.26	47.25	51.26
Brim	submission1	51.48	29.22	40.35
UAB*	submission4	42.25	3.38	22.82
Baseline	-	59.79	32.03	45.91

Table 5: Subtask 2 evaluation results. We report the average F1 scores across three cancer types (breast cancer, ovarian cancer, melanoma) in the dataset. Scores are with the strict evaluation metric, thus not comparable to the results from the 2024 Chemotherapy shared task which included relaxed-to-month evaluation. *: potentially without timeline summarization at the time of submission, thus results are likely lower.

6.1 Comparison of Systems and Results – 2024 and 2025 ChemoTimelines Shared Task

In the 2025 shared task edition, there are more submissions for Subtask 2, the end-to-end timeline extraction task which is the more difficult albeit realistic task – 8 submissions for Subtask 1 and 12 submissions for Subtask 2. In the 2024 shared task edition (Yao et al., 2024), there were 18 submissions for Subtask 1 and 9 submissions for Subtask 2.

A comparison between the 2024 and 2025 shared task Subtask 2 results reveals the substantial impact of the evaluation metric strictness and likely genuine system improvements over the intervening year. Under the relaxed-to-month evaluation in 2024, the baseline system achieved an official F1 score of 58, while the same baseline under the strict evaluation in 2025 dropped to 45.91. The top-performing 2025 systems (UW-BioNLP at 67.81 F1 under strict evaluation) would likely achieve substantially higher scores if evaluated under 2024's relaxed-to-month metric, likely significantly outperforming the best 2024 systems LAILab (Haddadan et al., 2024) at 70. This suggests that while the stricter 2025 evaluation exposes remaining challenges in precise temporal boundary detection, the underlying systems have indeed made considerable advances in temporal reasoning capabilities.

Methodswise, supervised fine-tuning remains

BREAST CANCER				
Team	**Submission**	**Type A**	**Type B**	**Official**
NousTime	submission1	79.31	65.32	72.31
NousTime	submission2	78.90	64.53	71.72
NLP4Health	submission3	71.02	43.65	57.34
NLP4Health	submission2	70.10	41.87	55.98
NLP4Health	submission1	70.06	41.78	55.92
UAB*	submission3	45.36	38.21	41.78
UAB*	submission1	36.13	25.81	30.97
UAB*	submission2	32.66	19.06	25.86
Baseline	-	86.85	74.44	80.64
MELANOMA				
Team	**Submission**	**Type A**	**Type B**	**Official**
NousTime	submission2	83.56	58.90	71.23
NousTime	submission1	83.11	57.76	70.43
NLP4Health	submission1	80.56	51.41	65.99
NLP4Health	submission2	80.56	51.41	65.99
NLP4Health	submission3	76.90	42.25	59.58
UAB*	submission2	54.73	11.83	33.28
UAB*	submission3	53.30	8.26	30.78
UAB*	submission1	30.57	1.43	16.00
Baseline	-	82.22	55.54	68.88
OVARIAN CANCER				
Team	**Submission**	**Type A**	**Type B**	**Official**
NousTime	submission1	84.18	68.36	76.27
NousTime	submission2	82.85	65.70	74.28
NLP4Health	submission3	73.07	46.14	59.61
NLP4Health	submission1	71.35	42.70	57.02
NLP4Health	submission2	71.35	42.70	57.02
UAB*	submission3	44.72	14.44	29.58
UAB*	submission2	32.96	15.91	24.43
UAB*	submission1	31.15	12.30	21.72
Baseline	-	88.11	76.21	82.16

(a) Subtask 1

BREAST CANCER				
Team	**Submission**	**Type A**	**Type B**	**Official**
UW-BioNLP	submission2	74.79	67.64	71.22
UW-BioNLP	submission4	70.89	71.18	71.04
UW-BioNLP	submission1	72.11	67.99	70.05
UW-BioNLP	submission5	63.10	67.14	65.12
NousTime	submission4	73.65	54.31	63.98
UW-BioNLP	submission3	62.25	59.93	61.09
NousTime	submission1	63.02	55.87	59.45
NousTime	submission3	72.15	45.85	59.00
NousTime	submission2	62.49	54.85	58.67
NousTime	submission5	70.80	43.23	57.01
Brim	submission1	47.10	41.59	44.34
UAB*	submission4	38.43	8.05	23.24
Baseline	-	54.04	43.96	49.0
MELANOMA				
Team	**Submission**	**Type A**	**Type B**	**Official**
NousTime	submission1	64.42	61.04	62.73
NousTime	submission4	70.45	51.12	60.78
NousTime	submission3	69.81	49.52	59.66
NousTime	submission5	69.81	49.52	59.66
UW-BioNLP	submission1	69.59	48.97	59.28
UW-BioNLP	submission2	68.46	46.14	57.30
NousTime	submission2	60.70	51.75	56.23
UW-BioNLP	submission3	65.53	38.83	52.18
UW-BioNLP	submission5	63.82	34.55	49.19
UW-BioNLP	submission4	63.80	34.51	49.16
Brim	submission1	62.84	32.11	47.48
UAB*	submission4	50.83	2.08	26.46
Baseline	-	52.94	7.34	30.14
OVARIAN CANCER				
Team	**Submission**	**Type A**	**Type B**	**Official**
UW-BioNLP	submission1	82.73	65.45	74.09
UW-BioNLP	submission4	82.02	64.05	73.04
UW-BioNLP	submission2	80.86	61.72	71.29
UW-BioNLP	submission5	77.67	55.34	66.50
NousTime	submission3	77.46	54.92	66.19
NousTime	submission5	77.46	54.92	66.19
NousTime	submission4	74.98	49.96	62.47
UW-BioNLP	submission3	66.78	33.55	50.16
NousTime	submission2	42.58	35.16	38.87
NousTime	submission1	31.06	37.12	34.09
Brim	submission1	44.49	13.98	29.24
UAB*	submission4	37.50	0.00	18.75
Baseline	-	72.40	44.79	58.59

(b) Subtask 2

Table 6: Evaluation results for each cancer type. Scores are with the strict evaluation metric (F1 score), thus not comparable to the results from the 2024 Chemotherapy shared task which included relaxed-to-month evaluation. *: potentially without timeline summarization at the time of submission, thus results are likely lower.

a productive approach albeit with a shift of SFT to very large language models. For example, the 2024 team LAILab (Haddadan et al., 2024) finetuned flan-T5-XXL which has 11B parameters (Chung et al., 2022), while the 2025 team UW-BioNLP finetuned Qwen3-14B (Qwen Team, 2025). Prompting techniques evolved as well – the UW-BioNLP_submission4 applied prompting LLMs with the thinking mode. Classic machine learning techniques were not explored in the 2025

edition unlike in the 2024 edition.

Even with the "strict" evaluation metric, the best 2025 results for Breast Cancer and Ovarian Cancer are either better or on par with the best 2024 relaxed-to-month results. 2025 results per type of cancer are presented in Table 6. Unlike the 2024 shared task where for Subtask 2 melanoma and breast cancer achieved better results, the 2025 results are slightly reversed – results for ovarian and breast cancer are better than for melanoma. The best results for breast cancer for Subtask 2 are 71.22 F1 for 2025 strict Official score v. 68 F1 for 2024 relaxed-to-month Official score. The best results for ovarian cancer for Subtask 2 are 74.09 F1 for 2025 strict Official score v. 74 for 2024 relaxed-to-month Official score. These results are encouraging as they are approaching the human-in-the-loop performance target as suggested by the US National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) program – end-to-end applications target of at least 0.75 F1 for human-in-the-loop setting which allows corrections by experts. However, reaching the NCI SEER recommendation for automated systems target performance of at least 0.95 F1 would require more methodological research.

While multiple factors could have contributed to the performance improvements in the 2025 shared task, a likely explanation is the fundamental advancement in LLM technology – including architectural refinements, enhanced pretraining data quality and diversity, and improved training paradigms.

7 Conclusion

The 2025 Shared Task on Chemotherapy Treatment Timeline Extraction builds upon the initial 2024 challenge but raised the bar by applying the strict F1 metric where exact matches of normalized dates and treatments are needed. The participating systems employed a variety of methods with a distinct shift towards exploiting very large language models – either through supervised fine-tuning, alignment or prompt engineering. As we point out in the Discussion section, the results are encouraging as they are approaching the human-in-the-loop performance target recommendation by the NCI SEER program – end-to-end applications target of at least 0.75 F1 for human-in-the-loop setting which allows corrections by experts. However, reaching the NCI SEER recommendation for automated systems target performance of at least 0.95 F1 would require more methodological research on this complex task that remains unsolved even with the current technological advances.

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. Piet de Groen, Danielle Bitterman, Elizabeth Buchbinder and Jeremy Warner for guiding us through the thickness of the oncology domain. Funding is provided by the United States National Institutes of Health (grants U24CA248010 and a supplement to it, 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

The data used in this shared task consisted of notes for patients with breast cancer, ovarian cancer, and melanoma, from a single health care system (UPMC), and NLP efforts focused solely on SACT administration. Results may not generalize to other types of cancers and treatments (radiation therapy, surgery, etc.), or to data from other health care providers.

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 patient privacy.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

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.

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, and 16 others. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

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.

H. Eyre, A. B. Chapman, K. S. Peterson, J. Shi, P. R. Alba, M. M. Jones, T. L. Box, S. L. DuVall, and O. V. Patterson. 2021. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA Annu Symp Proc*, 2021:438–447.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

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.

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.

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.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Qwen Team. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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.

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.

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. *Journal of Biomedical Informatics*, 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, and 1 others. 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.

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*, pages 557–569, Mexico City, Mexico. Association for Computational Linguistics.